Université de Montréal

Vocabulaire employé pour l'accès thématique aux documents d'archives patrimoniaux : étude linguistique exploratoire de termes de recherche, de description, d'indexation

par

Laure Amélie Guitard

École de bibliothéconomie et des sciences de l'information

Faculté des arts et des sciences

Thèse présentée

en vue de l'obtention du grade de Philosophia Doctor (Ph.D.)

en sciences de l'information

Avril 2018

Université de Montréal
Faculté des études supérieures

Cette thèse intitulée

Vocabulaire employé pour l'accès thématique aux documents d'archives patrimoniaux : étude linguistique exploratoire de termes de recherche, de description, d'indexation

présentée par

Laure Amélie Guitard

a été évaluée par un jury composé des personnes suivantes

Yvon Lemay, président-rapporteur
Lyne Da Sylva, directrice de recherche
Sabine Mas, co-directrice de recherche
Marcel Caya, examinateur externe
Éric Leroux, examinateur interne
François Lareau, représentant du doyen

Résumé

Les usagers recherchent des documents d'archives par sujet mais l'accès offert dans les services d'archives est principalement par provenance : il réside un écart entre l'accès recherché et l'accès offert. Pour répondre à la demande des usagers, certains services d'archives fournissent déjà des accès thématiques. Les instruments de recherche de ces précurseurs servent de base à notre recherche. Nous avons analysé le vocabulaire que les archivistes emploient pour décrire et indexer par sujet, c'est-à-dire le choix des mots et les relations sémantiques que ces mots entretiennent les uns avec les autres dans les documents dont ils sont issus (notices descriptives, index). Parallèlement, nous avons analysé le vocabulaire des usagers dans les questions envoyées par courriel à la référence avec la réponse de l'archiviste. Ainsi, notre étude couvre une large partie de la chaîne communicationnelle entre les usagers et les documents d'archives qu'ils recherchent par sujet. La comparaison de ces deux vocabulaires a fait émerger l'écart sémantique qui les sépare. Par l'étude des termes de recherche, de description, d'indexation, nous souhaitons contribuer à l'avancement des connaissances sur le vocabulaire employé pour l'accès thématique aux archives patrimoniales (VATAP). Notre méthode de recherche est l'étude de corpus. Nous étudions les relations sémantiques entre les termes d'un corpus de termes provenant de quatre sources de données (questions d'usagers à la référence, réponses d'archivistes, notices descriptives, index), collectées dans trois milieux. Nous circonscrivons la recherche à des données de centres ou services d'archives de grande envergure parce qu'ils ont les moyens de développer des outils pour encadrer et faciliter l'accès. L'originalité de notre recherche réside dans l'application de la linguistique à l'étude du VATAP. Par cette recherche, nous souhaitons décrire le VATAP et contribuer à l'amélioration de l'accès thématique aux archives patrimoniales en émettant des recommandations à l'usage des archivistes à partir de l'étude des relations sémantiques les plus fréquentes dans notre corpus. La formalisation linguistique nécessaire à notre recherche pourrait servir de prémisses à une automatisation de l'indexation. Cette étude exploratoire du vocabulaire des usagers permettra aux établissements participants de mieux connaître cet aspect du profil de leur clientèle.

Mots-clés : accès thématique, relations sémantiques, archives patrimoniales, analyse linguistique, vocabulaire, documents d'archives, indexation, recherche par sujet, référence archivistique, usager

Abstract

Users carry out searches for archives by subject, yet the access offered in the archives is mainly by provenance resulting in a gap between the desired access and the offered access. To meet users' demand, some archival services already offer access by subject. Research instruments used by these services provide the basis for our research. We analyzed the vocabulary used by archivists to describe and to index archival documents by subject. That is to say, the choice of words and the semantic relationships that exist between the meanings of these words in the documents from which they come from (descriptions, index). We analyzed both the users' vocabulary in the questions asked to the reference archivist by email and the answers offered by the latter. Thus, our study covers a large part of the communicative chain between users and archives they are looking for by subject. The comparison of these two vocabularies has brought out the semantic gap that separates them. By studying the terms used for research, for description, and for indexing, we wish to contribute to the advancement of knowledge on the vocabulary used for thematic access to heritage archives (VATAP). We have chosen corpus studies as our research methodology. We study the semantic relationships between the terms in a corpus from four data sources (user-to-reference questions, reference archivists' answers, descriptions, indexes), collected from three different archives services. We limit our research to data obtained from large archival repositories as they have the means to develop tools to guide and facilitate access. The originality of our research lies in the application of linguistics to the study of VATAP. Our study describes VATAP and contributes to a better accessing of heritage archives by subject. We accomplish this by formulating recommendations for the archivists' use based on the study of the most frequent semantic relationships in our corpus. The linguistic formalization needed for our research could serve as a premise for automatic indexing. This exploratory study of users' vocabulary allows participating institutions to better understand this aspect of their users' profile.

Keywords: subject access, semantic relationships, heritage archives, linguistic analysis, vocabulary, indexing, subject research, archival reference, user

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et acronymes

ADBS : Association des professionnels de l'information et de la documentation

AFNOR : Association française de normalisation

AAQ : Association des archivistes du Québec

AM : Archives (de) Montréal

ARK (modèle) : Connaissances pour la référence archivistique (*Archives Reference Knowledge model*)

BAC : Bibliothèque et Archives Canada

BAnQ : Bibliothèque et Archives nationales du Québec

BCA : Bureau canadien des archivistes

BRDV : Bureau Recherche Développement Valorisation

CERAS : Comité d'éthique de la recherche en arts et en sciences

N : 1. Notice(s) descriptive(s) ; 2 [dans la colonne *Précision* des tableaux] Catégorie grammaticale du nom

NC : nom commun

NP : nom propre

ICA : Conseil international des archives (*International Council of Archives*)

ISAD(G) : *Norme générale et internationale de description archivistique*

ISAAR (CPF) : *Norme Internationale sur les notices d'autorité utilisées pour les Archives relatives aux collectivités, aux personnes ou aux familles*

ISDF : *Norme de description des fonctions*

ISDIAH : *Norme internationale pour la description des institutions de conservation des archives*

ISO : Organisation internationale de normalisation

LRC : Lois révisées du Canada

LRQ : Lois refondues du Québec

OQLF : Office québécois de la langue française

PMEST : Personnalité, matière, énergie, espace et temps (Personality Matter Energy Space Time)Q : question(s) d'usager(s)

QRN : Question Réponse Notice

QRNTix : Question Réponse Notice Terme d'indexation

R : réponse(s) d'archiviste(s)

RDDA : Règles pour la description des documents d'archives

RDAQ : Réseau des archives du Québec

RAMEAU : *Répertoire d'autorité-matière encyclopédique et alphabétique unifié*

RAR : Regroupement des archivistes religieux

RVM : Répertoire de vedettes-matière

SI : sciences de l'information

TAL : traitement automatique de la langueTix : Terme(s) d'indexation

TLFi : *Trésor de la langue française informatisé*

V : verbe

VATAP : vocabulaire pour l'accès thématique des archives patrimoniales

Liste des abréviations et symboles

c.-à-d. : c'est-à-dire

cf. : *confer*

etc. : *et cætera*

id. : *idem*

i. e. : *in extenso*

maj : majuscule

min : minuscule

nb : nombre

n. m. : nom masculin

n. t. : notre traduction

p. ex. : par exemple

pl : pluriel

sg : singulier

% : pourcentage ou pourcent

Remerciements

Merci à vous, mes directrices, pour votre accompagnement académique et un doux réconfort dans les épreuves de la vie. Merci à Lyne Da Sylva pour nos passionnantes discussions linguistiques et l'encouragement à persévérer dans l'application de la linguistique en sciences de l'information. Merci à Sabine Mas pour ta vision méthodologique et enjouée de la recherche avec toujours une nouvelle piste à envisager quand j'étais dans une impasse. Merci pour votre confiance en mon projet et en mes capacités de chercheuse, le soutien financier dispensé au besoin et votre grande disponibilité. Les mots me manquent, ce n'est pas facile pour une linguiste de l'avouer, mais je vous remercie sincèrement.

Merci aux archivistes qui ont pris le temps de lire mes courriels d'invitation à participer et le temps de me recevoir pour que je leur expose ma recherche. Parmi ceux-ci, je remercie particulièrement les centres et services d'archives participants : vous m'avez épaulée dans ma démarche et avez gentiment patienté que je développe mon projet qui a finalement abouti. Vous m'avez accueillie dans vos locaux de quelques heures à plusieurs semaines, mis à disposition du matériel et vous n'avez pas compté votre temps pour me permettre de faire ma collecte de données dans de bonnes conditions. C'est grâce à vous si j'ai pu avoir des données aussi riches, merci.

Merci aux membres du jury d'évaluation de la thèse pour leur lecture attentive et leurs suggestions.

Merci au Fonds québécois pour la recherche en sciences sociales (FQRSC) pour le soutien financier qui m'a permis de ne pas abandonner le doctorat pour des raisons financières.

Merci à l'École de bibliothéconomie et des sciences de l'information (EBSI) pour les bourses départementales en début et en fin de doctorat, l'attribution de charges de cours, l'excellent soutien logistique : laboratoire de doctorat, postes informatiques, logiciels toujours à jour. Une belle équipe nous accompagne et nous offre d'excellentes conditions pour étudier. Merci à mon comité de recherche de m'avoir poussée à mener mes réflexions toujours plus loin. Merci aux professeurs, chargés de cours et professionnels de l'EBSI : en particulier, James Turner, Yvon Lemay et François Cartier pour leur soutien relatif à l'archivistique,

Christine Dufour et ses pompons dans les couloirs de l'EBSI, Alain Tremblay et sa rigueur administrative exemplaire et Sarah Pasutto : aucun problème logistique ne te résiste. À Aminata Keita, bibliothécaire des sciences de l'information, du plan de concepts à la commande de ressources en ligne ou imprimées, tu as toujours répondu à mes besoins documentaires au maximum de ton budget, merci.

Merci aux archivistes du groupe Arédiq (Archivistes étudiants et récents diplômés du Québec) et de l'AAQ (Association des archivistes du Québec), notamment Catherine D, Catherine F, Mylène, Myriam, Noura : discuter des défis que les archivistes rencontrent sur le terrain et mener des projets pour la communauté des archivistes avec enthousiasme et dynamisme ont enrichi mon cheminement.

Merci aux étudiants au doctorat, ceux - je devrais dire surtout *celles* - du début et ceux de chaque nouvelle cohorte enthousiaste et ceux de la fin, si sensibles et encourageants dans les épreuves. Une mention spéciale pour Claire : quelle précieuse entraide à chaque étape du doctorat ou de la vie et pour Anne : j'aime nos longues conversations archivistiques… ou pas.

Merci au groupe de recherche en histoire de la traduction en Amérique latine (Histal) : vous êtes des amis avec qui je pouvais parler de tout, de rien, de l'université, de terminologie, de traduction ou de restaurants. Des pauses midi qui se finissaient par un bon café afin d'avoir toute l'énergie nécessaire pour continuer. À M. Bastin, Mayra, Àlvaro, Àngela, Aura, Marc, Sônia, Malka, Jonathan, Cristina, Patrìcia, Paula, Irène et tous les autres : merci pour votre soutien au quotidien.

À mes amis, de près ou de loin (la liste serait trop longue, je suis choyée par la vie), merci de m'avoir offert autant de bouffées d'air frais que j'en avais besoin et à ceux qui sont passés par le doctorat, merci en plus d'avoir été autant d'exemples de thésards qui sont devenus docteurs… désormais encore plus occupés que pendant leur thèse : vous avez été inspirants (et vous l'êtes encore!).

À ma famille, ma mère, mon père et mes frères et sœur, qui ont cru en moi dès le début et m'ont toujours soutenue; vos encouragements discrets ou explicites me redonnaient force à chaque fois. À mon mari, soutien indéfectible, roc sur lequel je m'appuie avec une confiance totale, vent qui me porte et me rappelle le cap à tenir, moitié douce, aimante et attentionnée au

quotidien, tu m'as appris à devenir qui je suis. À ma fille Héloïse, qui m'a appris que lorsque l'on tombe, on se relève, autant de fois qu'il le faut pour apprendre à marcher, courir et danser même, ta persévérance et ton regard plein de ton amour incommensurable font de moi une meilleure personne.

Merci à Dieu de m'avoir donné la force, la persévérance et de m'avoir si bien entourée à chaque étape importante de ma vie et le doctorat en fut une.

À Maman,

À Héloïse,

Oncques ne cède

Introduction

Notre recherche doctorale porte sur le vocabulaire employé pour l'accès thématique aux archives patrimoniales (VATAP) formulé par des termes de description, d'indexation et de recherche qui sont des noms communs.

Notre recherche part du constat que les archivistes devraient offrir davantage d'accès par sujet pour mieux répondre aux besoins des usagers et améliorer les accès par sujet existants (Duff 2010), ce que montrent les études d'usagers (Daniels et Yakel 2010; Schaffner 2015). En outre, le numérique amplifierait le besoin d'accès par sujet (Duff, Yakel et Tibbo 2013; Giuliano 2012; Zhang 2012) en augmentant le nombre d'usagers ayant accès directement, sans l'intermédiaire d'un archiviste, à des ressources archivistiques – documents d'archives et instruments de recherche. Mais hormis le rapport du Groupe de travail sur l'indexation par sujet au sein du Bureau canadien des archivistes (BCA 1992), l'accès par sujet des documents d'archives avec des noms communs n'est pas encadré par une norme répondant au contexte spécifique des documents d'archives au Canada.

La question du vocabulaire est centrale dans la recherche d'information et dans l'accès aux documents. Selon Salaba (2005), il s'agit du principal problème d'accès des catalogues en ligne; elle compare le vocabulaire des usagers et celui des systèmes et constate qu'ils correspondent peu. Pour Hennicke (2011), il existerait un « fossé sémantique » entre la formulation des besoins des usagers et les documents d'archives. Par exemple, il y aurait un fossé sémantique entre les expressions linguistiques porteuses de sujet (thémanymes)[1] présentes dans les notices descriptives et celles que l'usager mentionne dans sa question de recherche sans savoir qu'elles renvoient à des choses différentes selon les époques, ce qui arrive très souvent dans les archives. En effet, les collections couvrent des périodes de quelques décennies à plusieurs siècles. La polysémie temporelle cause donc de l'ambigüité

[1] Nous employons *expressions linguistiques*, *mots* et *termes* de façon interchangeable. Notons tout de même que les termes et les expressions linguistiques sont possiblement formés de plusieurs mots (p. ex. *anatomie, pomme de terre, recette de cuisine*).

pour le repérage. Pourtant, l'ampleur ni même l'existence de cet écart sémantique entre les diverses étapes de la chaîne communicationnelle des usagers aux documents d'archives n'a, à notre connaissance, pas été vérifiée empiriquement.

Notre recherche a justement pour but d'étudier l'écart sémantique présupposé entre le vocabulaire des usagers qui recherchent par sujet des documents d'archives patrimoniaux et celui employé dans les instruments de recherche pour l'accès thématique à ces mêmes documents. Nous voulons vérifier empiriquement l'existence de l'écart sémantique présupposé entre ces deux vocabulaires, le qualifier et le quantifier. Ces vocabulaires sont présents dans quatre sources de données qui forment le vocabulaire pour l'accès thématique aux archives patrimoniales (VATAP) : les courriels d'usagers à la référence, les notices descriptives, les index contenant des termes d'indexation professionnelle et les nuages d'étiquettes contenant les termes d'indexation collaborative (ou étiquettes). Notre étude s'est centrée sur un corpus constitué d'expressions provenant de trois des sources mentionnées ci-dessus, que nous avons analysé en comparant les expressions qui s'y trouvaient.

L'analyse s'est effectuée en trois temps. Dans un premier temps, nous avons comparé les chaînes de caractères des expressions linguistiques porteuses de sujet (à base nominale) dans les sources de données de notre corpus pour vérifier empiriquement l'écart présupposé entre ceux-ci. Dans un deuxième temps, dans les cas où il résidait un écart entre deux expressions de sources différentes, nous avons nommé la relation sémantique qui sous-tend cet écart (p. ex. la synonymie) en utilisant la méthodologie classique de la sémantique lexicale. Nous avons ainsi qualifié l'écart sémantique entre les expressions linguistiques porteuses de sujet (thémanymes) du vocabulaire employé dans les sources de données retenues. Dans un troisième temps, nous avons recherché des récurrences dans les relations sémantiques identifiées précédemment, afin de quantifier l'écart observé dans le corpus.

La thèse suit le plan suivant. Nous présentons d'abord la problématique de notre recherche suivie du but, des objectifs et des questions de recherche. S'ensuit une revue de la littérature sur les concepts majeurs de notre recherche. Vient ensuite l'exposition de la méthodologie employée : la présentation des données et de leur prétraitement, la méthode d'analyse et les critères de qualité de la recherche. L'analyse des résultats et la discussion constituent les deux dernières parties. La conclusion résume la présente étude et inclut des recommandations et des pistes de recherche futures.

1. Contexte et problématique

Au début de 2015, la Table de concertation des archives religieuses de Montréal (TCARM) diffuse le projet d'établissement d'un centre du patrimoine religieux à Montréal sur le site web du Regroupement des archivistes religieux (RAR). Ce projet comporte deux volets : la sauvegarde et la mise en valeur du patrimoine archivistique religieux de la région de Montréal. « Cette Table de concertation a comme objectif principal de mettre sur pied un centre de conservation et de diffusion regroupant des documents d'archives, des publications et des biens mobiliers qui témoignent de l'engagement de l'Église dans le diocèse de Montréal et éventuellement dans les diocèses en périphérie. » (Billaudele 2015) Les documents d'archives historiques ou archives définitives sont des documents d'archives n'ayant généralement plus de valeur administrative ni légale qui sont conservés pour leur valeur d'information et leur potentiel de réutilisation. Ces documents sont aussi appelés *patrimoniaux* (Cardin 1999; Frey et Treleani 2013) et c'est l'expression que nous emploierons désormais (voir Annexe 1 – Définitions).

Le regroupement de plusieurs organisations créatrices de documents d'archives patrimoniaux et d'autres types d'objets informationnels culturels nécessite de développer des outils adéquats pour retracer correctement et rapidement ceux-ci dans une collection de gros volume et hétéroclite. Le repérage rapide de l'information est justement le rôle de l'indexation (Weinberg 2009, 2287).

L'indexation (par sujet[2]) est « l'opération qui consiste à décrire et à caractériser un document à l'aide de représentations des concepts évoqués dans ce document » (AFNOR 1993, 10). Elle a pour fonction de repérer des documents (connus ou inconnus) ou de l'information qu'ils contiennent dans une collection et de discriminer des documents au contenu similaire en fonction du sujet (adapté de Da Sylva 2015). L'indexation humaine, bien

[2] La définition de l'AFNOR s'applique tout spécialement à l'indexation par sujet. Nous employons l'adjectif *thématique* et la locution *par sujet* de manière indifférente.

que produite aujourd'hui dans la plupart des cas grâce à des outils informatiques (Hudon 2013, 31), repose sur le jugement humain. L'indexation automatique, elle, résulte de l'application d'algorithmes informatiques basés de manière importante sur la fréquence des mots présents dans les documents. Bien que les progrès dans l'automatisation de l'indexation soient indéniables, ils n'atteignent pas (encore) la qualité de l'indexation humaine (Hudon 2013, 31). Dans le contexte archivistique, « la création et l'extraction automatique de métadonnées n'en est encore qu'à ses balbutiements » (Gilliland 2014, 119, notre traduction[3]). L'indexation automatique des notices descriptives présente des défis relatifs à la nature des documents d'archives (voir section 3.1). L'indexation peut être basée uniquement sur les mots employés dans le document à indexer (indexation par extraction), il y a dans ce cas reprise de l'expression linguistique exacte du document source. Celle-ci est généralement réalisée automatiquement. Il n'existe alors a priori pas d'écart sémantique entre le document et les termes d'indexation. Par opposition, l'indexation par assignation recourt à des termes attribués à un document à partir d'une autre source que le document lui-même (Lancaster 2003, 18, n. t.). Ces termes ont généralement deux sources possibles. Soit ils sont tirés d'un langage documentaire, c'est-à-dire « un sous-ensemble structuré de termes extraits de la langue naturelle [*sic*] dont on a contrôlé la forme et la signification » (Hudon *et al.* 2009, 85). Soit ils sont tirés d'un langage d'indexation non contrôlé, le vocabulaire libre. Dans ce dernier cas, quand les termes d'indexation sont tirés « de la tête de l'indexeur » (Lancaster 2003, 18, n. t.), l'indexation par assignation est basée sur le jugement d'un indexeur qui décide comment traduire le concept retenu et comment choisir le mot à retenir, présent ou non dans le texte à indexer. Or, qui dit jugement d'un être humain, dit subjectivité. L'indexation par assignation humaine en vocabulaire libre souffre de problèmes bien connus tels que la polysémie et le manque de rappel (Arsenault 2006; Salaün et Arsenault 2009). Nous pouvons prendre la mesure de cette faiblesse du vocabulaire libre dans les études sur les folksonomies (p. ex. Lezcano, Garcia-Barriocanal et Sicilia 2012, voir section 3.4.3.1). L'indexation en vocabulaire libre peut nuire à la cohérence inter- et intra-indexeurs (Arsenault 2006, 141) et connaît un

[3] Désormais n. t.

manque de prédictibilité des termes d'indexation. C'est pourquoi on cherche généralement à l'encadrer par un langage documentaire ou une politique d'indexation.

Nous avons pu constater dans nos échanges avec des archivistes en prévision de notre collecte que, généralement, dans les services d'archives qui la pratiquent, l'indexation est humaine, par assignation et en vocabulaire libre. Les quelques publications sur l'indexation professionnelle par sujet de documents d'archives traitent de tentatives d'appliquer un vocabulaire contrôlé aux archives patrimoniales, mises en place par des institutions nationales dans le monde (voir section 3.3.2). Ceci nous laisse penser que les autres services d'archives procèdent à une indexation thématique en vocabulaire libre ou n'indexent pas les sujets de leurs documents d'archives. À notre connaissance, il n'existe pas de lignes directrices sur l'indexation par sujet adaptées aux documents d'archives. Et, il est rare de trouver une politique d'indexation dans les services d'archives – nous n'en avons trouvé qu'une seule au Québec, à Bibliothèque et Archives nationales du Québec (BAnQ), diffusée uniquement à l'interne. Or, une politique d'indexation a pour but d'augmenter la cohérence inter- et intra-indexeurs (Da Sylva 2015). L'indexation par sujet des archives patrimoniales est donc prédisposée à l'incohérence. Ce dernier fait risque d'avoir pour conséquence de réduire l'accès thématique[4] des documents aux usagers qui peinent à trouver sous quel(s) terme(s) les documents d'archives ont été indexés.

L'accès par sujet aux documents d'archives, longtemps délaissé par les archivistes, bénéficie d'un intérêt croissant dans la profession. En 2008, une enquête lancée au Royaume-Uni auprès des archivistes a montré que ceux-ci étaient intéressés par l'accès thématique, qu'ils l'avaient implanté ou voulaient l'implanter prochainement dans leur service d'archives (Fenton 2010, 194-195). L'accès par sujet aux documents d'archives s'envisage par la rédaction de guides thématiques ou, ce qui nous occupe ici, par l'indexation thématique. Plus récemment, Vilar et Šauperl (2015) ont démontré le besoin de lignes directrices pour l'indexation thématique de documents d'archives en Slovénie et Bosnie Herzégovine; ils ont

[4] Pour un approfondissement des notions d'accès thématique et d'indexation thématique en archivistique, voir section 3 Revue de la littérature.

initié un projet d'écriture de telles lignes directrices, subventionné par les gouvernements de ces deux pays. Kyriaki-Manessi et Dendrinos (2014) rendent compte d'un projet sur l'accès thématique à des données de documents d'archives numériques conservés à la bibliothèque du *Technological Educational Institute of Athens*, subventionné par l'Europe de 2009 à 2013. La seconde partie de leur projet consiste à connecter cet ensemble de données au Web sémantique, en ayant recours aux données liées. Cette recherche établit clairement le lien entre l'indexation par sujet et le Web sémantique. Afin de permettre aux services d'archives de faire figurer leurs instruments de recherche dans le Web sémantique dans une optique de *découvrabilité* (Schaffner 2015), il est considéré nécessaire pour la profession de développer en premier lieu l'indexation par sujet. Ceci améliorerait l'accès par sujet aux archives.

L'accès par sujet est l'accès principal que les usagers utilisent lors d'une recherche de documents ou d'informations (BCA 1992, 35; Gagnon-Arguin 1998, 92; Lévesque 2001-2002, 36; Nahuet 2009, 55; Pugh 2005). Ils mènent leurs recherches de la même manière quelle que soit l'institution conservatrice – bibliothèque, archives, musée, etc. – en reproduisant le modèle de la boîte de recherche simple « à la Google » (Yu et Young 2004; Zhang 2012) où les types de recherche sont indifférenciés. Le numérique amplifie ce besoin d'accès par sujet aux documents d'archives (Duff, Yakel et Tibbo 2013; Giuliano 2012). Selon Zhang (2012, 54), la représentation du contenu thématique des collections archivistiques est incontournable pour la découverte de l'information archivistique dans l'environnement numérique.

Pourtant, dans les services d'archives, l'accès se fait traditionnellement par provenance, c'est-à-dire en fonction du nom du créateur des documents. La provenance sert à établir l'unité de traitement (le fonds d'archives), à bâtir la classification (à partir des fonctions et activités du créateur) et à créer les principales clés d'accès de l'indexation. Ce choix a été fait pour rendre compte du contexte de création des documents d'archives (Duff 2010; Zhang et Mauney 2013)[5] afin notamment de préserver la valeur de témoignage et d'assurer l'authenticité des documents, caractéristique essentielle des documents d'archives (Lemay et

[5] La délimitation de l'unité de traitement archivistique s'effectue en fonction du principe du respect de provenance qui s'oppose historiquement au classement méthodique par matières (Héon 1999, 225, Nahuet 2009).

Klein 2012; MacNeil 2005; Zhang 2012, 60). Les noms propres de personnes physiques, familles ou personnes morales forment les accès de provenance classiques. Mais les noms propres – de personnes physiques, familles ou personnes morales, de lieux, d'événements, etc. – peuvent aussi constituer des accès par sujet de documents d'archives.

Le sujet de documents d'archives peut ainsi s'exprimer soit par des noms propres, soit par des noms communs. Les noms propres servent souvent de points d'accès thématique parce que les fonds d'archives ont la particularité de porter *ipso facto* sur leur créateur. Sémantiquement, ils renvoient à des segments de la réalité qui sont connus, généralement non ambigus; ils constituent ainsi des accès extrêmement efficaces. La formalisation et la désambigüisation éventuelle des points d'accès par noms propres de créateur – personnes physiques, familles et personnes morales – est prise en charge par les normes nationale (voir les chapitres 22 *Vedettes de personnes physiques* et 24 *Vedettes de personnes morales* des *Règles pour la description des documents d'archives* (*RDDA*) (BCA 2008) et internationale (voir la *Norme Internationale sur les notices d'autorité utilisées pour les Archives relatives aux collectivités, aux personnes ou aux familles* (*ISAAR-CPF*) ICA 2004). La seconde partie des *RDDA* porte sur la formulation très précise des points d'accès qui sont des noms propres. Ainsi, non seulement les noms propres de personnes physiques, de familles et de personnes morales sont pris en charge, mais aussi les noms de lieux (voir le chapitre 23 *Noms de lieux* dans les *RDDA* 2e partie et les *Règles de catalogage anglo-américaines* (*RCAA2*) Asted 1998-2005). Dans le chapitre 21 des *RDDA* intitulé « Choix des accès à la description des documents », la première règle qui indique sa portée générale stipule que « [l]es règles de ce chapitre servent à déterminer le choix des accès indépendants du sujet qui permettent d'identifier et de rechercher une description. » (BCA 2008, 21-3) Le Conseil international des archives a créé une norme spécifique sur les accès de fonctions des producteurs, étant donné leur importance dans le contexte de production des archives (*ISDF*, ICA 2011, 4, section P.2). Cette norme favorise au sein d'un établissement, d'un pays ou à l'international la normalisation des dénominations de fonctions des créateurs des documents d'archives et « toute subdivision d'une fonction telle que sous-fonction, procédure opérationnelle, activité, tâche, transaction ou tout autre terme employé au niveau international, national ou local » (*ISDF*, ICA 2011, 7, section 1.2). Hormis les accès de fonctions de producteurs – ce qui constitue une infime partie des accès par sujet possibles, l'accès par sujet des documents

d'archives[6] dont le sujet est formulé par des noms communs reste quant à lui sans directives claires, et ce manque nous interpelle particulièrement, vu l'intérêt de l'accès thématique en général pour les usagers.

Les principes de l'indexation des documents d'archives, en particulier ceux de l'indexation avec des noms communs, sont peu connus des archivistes, relèvent Cadieux et Charbonneau (2002, 82). La connaissance générale sur les accès par sujet dans les archives est à développer (Duff 2010). C'est pourquoi notre recherche porte sur les noms communs employés pour l'accès par sujet aux archives patrimoniales.

L'indexation est difficile à appréhender dans son ensemble car c'est une opération intellectuelle : une partie s'effectue mentalement par les indexeurs. C'est pour cela que notre recherche s'intéresse plutôt au produit de cette opération. Ce produit a une forme concrète : l'ensemble des termes d'indexation figurant dans les index. Ainsi, nous considérons à la suite de Deweze (1981), Amar (2000) et Lancaster (2003) que l'indexation est une opération linguistique et que son produit est analysable par la linguistique.

Dès lors, adoptant un point de vue linguistique, nous considérons que l'indexation des documents d'archives patrimoniaux peut être étudiée par le biais des expressions linguistiques retenues dans les termes d'indexation, choisis par l'indexeur sur la base des expressions linguistiques présentes dans les notices descriptives. Il existe alors des relations sémantiques entre ces deux types d'expressions linguistiques, qui varient du changement de casse à la paraphrase ou la synonymie et qui peuvent inclure maintes autres relations sémantiques d'intérêt pour un linguiste et déterminantes pour l'interprétation de l'index et pour l'accès aux documents par les usagers.

---

[6] Une quatrième norme a été créée par l'ICA : ISDIAH (ICA 2008). Cet appareillage de quatre normes internationales (ISAD(G), ISAAR(CPF), ISDF et ISDIAH) forme le socle d'un modèle conceptuel (*RiC-CM*) associé à une ontologie (*RiC-O*, pas encore disponible) proposés par le Groupe d'experts sur la description archivistique du Conseil international des archives (ICA 2016). Le modèle conceptuel fractionne les éléments de la notice pour leur permettre de devenir chacun des clés d'accès à la notice descriptive ainsi qu'à l'unité de description.

L'indexation est tributaire de la prise en compte non seulement des documents mais également des usagers auxquels elle est destinée. Les questions que les usagers posent à la référence révèlent des besoins en accès thématique (Duff 2010; Gagnon-Arguin 1998; Mas 2013-2014). Également, les étiquettes (*tags* en anglais) qu'ils attribuent à des documents d'archives ou leur notice descriptive expriment ce que représentent ces documents pour eux, dans leurs mots. Les courriels d'usagers à la référence et les étiquettes sont des expressions linguistiques librement choisies par les usagers; ces expressions linguistiques révèlent le vocabulaire qu'ils emploient pour accéder aux documents d'archives. Les archivistes développent les instruments de recherche pour que les usagers puissent accéder aux documents d'archives, pour leur en faciliter l'accès (Light 2008; Nougaret, Galland et Direction des archives de France 1999). On s'attendrait donc à ce qu'il y ait une correspondance entre le vocabulaire employé dans les instruments de recherche (notice descriptive et termes d'indexation professionnelle) et le vocabulaire employé par les usagers lors de leur recherche, notamment dans les courriels d'usagers à la référence et dans les étiquettes qu'ils attribuent en ligne à des notices descriptives ou des documents d'archives. Pourtant, les études précitées évoquent des différences importantes entre les deux vocabulaires.

En matière de recherche documentaire, le principal problème dans les catalogues en ligne repose sur le vocabulaire (Salaba 2005). Il existe un « fossé sémantique » (Hennicke 2011) entre l'expression des besoins des usagers et les instruments de recherche. Il s'agit même d'un « golfe » (Zhang 2011) quand on parle de l'écart sémantique entre les documents d'archives et leurs représentations en ligne. Cependant, la nature et l'ampleur de cet écart sont inconnues. C'est pourquoi notre recherche veut vérifier empiriquement l'existence de l'écart sémantique présupposé entre, d'une part, le vocabulaire employé dans les instruments de recherche - saillant dans les notices descriptives et les termes d'indexation professionnelle - et, d'autre part, le vocabulaire des usagers - saillant dans les courriels d'usagers à la référence et les étiquettes attribuées par les usagers en ligne. Celles-ci pourraient également être examinées. Cet écart sera ensuite qualifié par une analyse linguistique des relations sémantiques et quantifié, dans la perspective de faire des recommandations relatives au vocabulaire pour l'accès thématique aux archives patrimoniales (VATAP) pour les relations sémantiques les plus fréquentes. Nous espérons ainsi contribuer à comprendre l'écart

sémantique entre l'usager et les représentations des documents afin que des études ultérieures puissent proposer des façons de le réduire et favoriser un meilleur accès thématique aux documents patrimoniaux.

2. But, objectifs et questions de recherche

La présente section présente le but, la question générale, les objectifs et les questions de recherche de notre projet.

2.1. But et question générale de la recherche

Notre recherche porte sur le vocabulaire employé pour l'accès thématique aux documents d'archives patrimoniaux (VATAP). Elle a pour but d'étudier l'écart sémantique présupposé entre le vocabulaire des usagers qui recherchent des documents d'archives patrimoniaux et celui employé par les archivistes dans les instruments de recherche. Ces deux vocabulaires sont constitués d'expressions linguistiques porteuses de sujet; nous appelons ces expressions des *thémanymes*[7]. Nous voulons vérifier empiriquement l'existence de l'écart sémantique présupposé entre les thémanymes de ces deux vocabulaires, qualifier cet écart selon une approche linguistique et le quantifier.

La question générale qui sous-tend notre démarche est la suivante : quelles sont la nature et l'ampleur de l'écart sémantique présupposé entre le vocabulaire des usagers qui recherchent des documents d'archives patrimoniaux et le vocabulaire employé par les archivistes dans les instruments de recherche pour l'accès thématique à ces documents ? Notre recherche montre à partir de quelles relations sémantiques on pourrait privilégier un enrichissement lexical des instruments de recherche en vue de favoriser un meilleur accès thématique par sujet aux documents d'archives patrimoniaux.

2.2. Objectifs de recherche (OR)

Pour atteindre ce but, notre recherche vise les objectifs suivants.

---

[7] Le mot *thémanyme* est un néologisme que nous avons créé pour les besoins terminologiques de notre recherche. Il vient de *théma–* (sujet, comme dans thématique) et de *–nyme* (nom, au sens de syntagme ou unité linguistique de forme et de sens) et signifie 'expression linguistique porteuse de sens'. Le mot est masculin. L'adjectif *thémanymique* signifie 'relatif au(x) thémanyme(s)'.

OR. 1. Valider empiriquement l'existence d'un écart sémantique présupposé entre le vocabulaire des usagers qui recherchent des documents d'archives patrimoniaux par sujet et le vocabulaire employé par les archivistes dans les instruments de recherche pour l'accès thématique à ces documents.

OR. 2. Qualifier l'écart sémantique entre le vocabulaire des usagers qui recherchent des documents d'archives patrimoniaux par sujet et le vocabulaire employé dans les instruments de recherche pour l'accès thématique à ces documents.

a. Caractériser les relations sémantiques entre les thémanymes des deux vocabulaires.

OR. 3. Quantifier l'écart sémantique entre le vocabulaire des usagers qui recherchent des documents d'archives patrimoniaux par sujet et le vocabulaire employé dans les instruments de recherche pour l'accès thématique à ces documents.

a. Calculer la fréquence de correspondance et de non-correspondance entre les thémanymes des deux vocabulaires.
b. Calculer la fréquence de chaque type d'écart sémantique entre les thémanymes des deux vocabulaires.

2.3. Questions de recherche (QR)

À partir des objectifs de notre recherche, nous nous sommes donné des questions de recherche.

OR 1 (Vérifier empiriquement l'existence de l'écart sémantique présupposé)

- QR 1. Lors de la comparaison du vocabulaire des usagers et celui employé dans les instruments de recherche, des paires de thémanymes entretiennent-elles une relation d'identité ?

OR 2 (Qualifier l'écart sémantique)

- QR 2. Quelles sont les relations sémantiques qui existent entre les thémanymes non identiques appartenant au vocabulaire des usagers et ceux appartenant au vocabulaire des instruments de recherche ?

OR 3 (Quantifier l'écart sémantique)

- QR 3. Quelle est la fréquence de la relation d'identité entre les thémanymes appartenant au vocabulaire des usagers et ceux appartenant au vocabulaire employé dans les instruments de recherche ?
- QR 4. Quelle est la fréquence de chacune des relations sémantiques qui caractérisent un écart sémantique entre les thémanymes appartenant au vocabulaire des usagers et ceux appartenant au vocabulaire employé dans les instruments de recherche ?

Nous présentons en annexe un tableau méthodologique récapitulant le titre de notre recherche, le but, la question générale, les objectifs et questions de recherche ainsi que les moyens mis en œuvre pour les atteindre et y répondre : la méthode d'analyse, les outils de collecte et d'analyse ainsi que les résultats obtenus (voir Annexe 3 – Tableau méthodologique).

3. Revue de la littérature

Notre sujet de recherche sur le vocabulaire employé pour l'accès thématique aux archives patrimoniales (VATAP) relève de deux grandes disciplines : l'archivistique et la linguistique. Notre revue de la littérature reflète cette double appartenance. Nous exposons d'abord les notions archivistiques pertinentes à notre sujet. Celles-ci sont présentées à partir de l'indexation thématique des documents d'archives, concept qui a été le point de départ de notre réflexion. La section sur les notions linguistiques présente les diverses notions linguistiques nécessaires à la compréhension de notre méthodologie. Nous recourons à l'analyse linguistique de corpus textuel, notamment en ce qui a trait aux phénomènes de sens en jeu dans des instruments de recherche tels que l'index thématique de documents d'archives ou un entretien de référence différé tel qu'un échange de courriels entre un usager et un archiviste de référence.

Notions archivistiques

Notre recherche porte sur le vocabulaire employé pour l'accès thématique aux archives patrimoniales (VATAP). Le Conseil international des archives (ICA) amorce le document intitulé *Principes relatifs à l'accès aux archives* par la phrase suivante : « L'accès est le processus qui rend possible la consultation des archives parce qu'il existe d'une part une autorisation légale, d'autre part des instruments de recherche. » (2012, 3) Dans son ouvrage sur la référence et l'accès archivistiques, Pugh (2009) définit l'accès de manière intellectuelle, légale et physique. Une étude du vocabulaire pour l'accès thématique aux documents d'archives se place dans l'étude de l'accès intellectuel. L'auteure identifie plusieurs composantes dans une situation de référence archivistique : les documents d'archives, les usagers et, pour assurer la médiation entre les deux, les archivistes et les instruments de

recherche[8]. *Instrument de recherche* est un « terme générique pour [désigner] tout outil de description ou de référence élaboré ou reçu par un service d'archives dans l'exercice de son contrôle administratif ou intellectuel sur les documents d'archives » (ICA 2012, 15). Les instruments de recherche sont nombreux et variés. Selon le manuel *La Pratique archivistique française* (Direction des archives de France 1993, 154-161), ils peuvent être destinés soit à l'usage interne, soit à l'usage du public. Ces derniers peuvent être synthétiques (l'état des fonds ou le guide d'archives, des outils plus ou moins détaillés, par provenance ou thématiques, au niveau du fonds, du dépôt ou transversaux à plusieurs dépôts) ou analytiques (l'inventaire et le répertoire, des listes descriptives plus ou moins détaillées au niveau respectivement de la pièce ou du dossier, ainsi que l'index). Cardinal *et al*. (1984) proposent une typologie similaire. Ils ajoutent une distinction entre les instruments de recherche de premier niveau, rattachés aux unités de description et ceux de second niveau, qui « permettent de repérer plus rapidement une information spécifique à l'intérieur d'une masse de documents. Ce sont les index alphabétiques ou chronologiques, ainsi que les guides par sujets de recherche ou guides thématiques. » (Cardinal *et al*. 1984, 21). Les instruments de recherche sont généralement rédigés, qu'ils soient ou non imprimés (Direction des archives de France 1993, 154). Ils incluent aussi des outils de recherche informatisés, qui, même s'ils sont exclus explicitement de la typologie principale, sont jugés complémentaires des instruments de recherche papier (Direction des archives de France 1993, 154). Ainsi, nous considérons comme un instrument de recherche à part entière une base de données de description archivistique, telle que la proposent sur place ou en ligne certains centres et services d'archives de grande envergure[9].

---

[8] Pour une adaptation de la situation de communication au milieu archivistique, voir notamment Senécal 1998 et pour une représentation graphique de la situation de communication d'un usager en recherche, voir notamment Pugh 2009, 166, fig. 2 et Guitard 2014.

[9] Par exemple, « La banque de données Pistard (Programme informatisé servant au traitement des archives et à la recherche documentaire) donne accès à la description des fonds et collections conservés par Bibliothèque et Archives nationales du Québec (BAnQ). » (BAnQ 2018) Également, « [l]e catalogue BnF archives et manuscrits

Le VATAP est envisagé dans notre recherche à partir des instruments de recherche institutionnels suivants : les notices descriptives, les termes d'indexation professionnelle qui leur sont associés. À cela, nous ajoutons les questions que les usagers posent par courriel aux archivistes de référence et les étiquettes qu'ils formulent pour décrire le contenu de documents en ligne. Dans notre recherche, la notion d'accès thématique regroupe les notions de recherche, de description et d'indexation. L'indexation étant à la fois reliée à la description et à la recherche, notre revue de la littérature a été bâtie sur les principales caractéristiques de l'indexation. Nous les présentons une à une en les reliant à notre recherche ci-dessous.

L'entretien de référence archivistique peut être vu comme une situation de communication où est présentée la chaîne communicationnelle entre d'un côté les documents d'archives et de l'autre les usagers dont l'archiviste essaie de favoriser la rencontre. De ce fait, nous débutons par la nature des documents d'archives sur lesquels portent la description, l'indexation et la recherche. Puis, nous abordons les besoins et pratiques des usagers auxquels la description et l'indexation sont destinées. Ayant rappelé la prévalence de l'accès par sujet dans les recherches des usagers, nous présentons ensuite le sujet en archivistique, notion centrale de l'accès thématique aux archives patrimoniales. Nous évoquons le flou définitionnel de cette notion, le manque de lignes directrices pour l'indexation thématique adaptées aux documents d'archives et l'apport de la notion de facettes dans l'approche du sujet en archivistique.

Ainsi, notre revue de la littérature sur les notions archivistiques s'articule autour des axes de lecture suivants : la nature des documents d'archives (section 3.1), la référence archivistique (section 3.2) et l'indexation thématique en archivistique (section 3.3).

---

contient les descriptions de manuscrits, dessins, fonds d'archives sur différents supports (par exemple : audiovisuels) conservés aux départements des Manuscrits, des Arts du spectacle, de la Musique, des Monnaies, médailles et antiques (site Richelieu), de l'Audiovisuel (site François-Mitterrand), et à la Bibliothèque de l'Arsenal. On y trouve également les descriptions d'archives administratives de la Bibliothèque depuis le XVII^{e} siècle. » (Bibliothèque nationale de France 2018).

3.1. Nature des documents d'archives

L'indexation est tributaire de la nature des documents sur lesquels elle porte (Da Sylva 2015). En effet, les normes ISO 5963 et NF Z 47-102 présentent comme étape préliminaire à l'indexation la lecture du document total, son appréhension (voir le tableau récapitulatif des étapes de l'indexation présentées dans les normes et les principaux textes de référence dans Guitard 2013). En 1994, Fidel distingue l'indexation centrée sur le document de l'indexation centrée sur l'usager. Cette distinction transparaît sur le vocabulaire. L'indexation centrée sur le document reprendra exclusivement la terminologie du document alors que l'indexation centrée sur l'usager adaptera le vocabulaire aux pratiques et besoins des usagers. Aujourd'hui, l'approche se mitige; les normes sur l'indexation recommandent de prendre en compte à la fois le vocabulaire des documents et le vocabulaire des usagers, en adéquation avec la politique d'indexation en vigueur.

L'indexation archivistique est tributaire de la singularité des documents d'archives et de la hiérarchie propre aux fonds d'archives (BCA 1992, Guitard 2013). La description et la recherche sont également influencées par ces caractéristiques des documents d'archives et nous verrons comment dans cette section. Cette dernière porte sur la nature des documents d'archives et les spécificités de leur accès thématique. Les documents d'archives ont plusieurs caractéristiques qui les distinguent des autres documents. D'abord, ils sont uniques, les doublons étant éliminés au cours du traitement et il est nécessaire d'assurer leur authenticité pour qu'ils puissent servir de preuve, par exemple. Ensuite, leur nature organique et leur caractère historique sont à prendre en compte dans l'élaboration d'un système de repérage (BCA 1992, 31-32). La singularité (voir section 3.1.1.) et la hiérarchie d'un fonds d'archives (voir section 3.1.2.) sont deux caractéristiques que nous avons retenues parce qu'elles ont des répercussions sur l'indexation thématique des documents d'archives et sur la recherche des usagers, comme nous le décrivons ci-dessous.

3.1.1. Singularité d'un fonds d'archives

Au Québec, la *Loi sur les archives* définit les archives comme « l'ensemble des documents, quelle que soit leur date ou leur nature, produits ou reçus par une personne ou un organisme pour ses besoins ou l'exercice de ses activités et conservés pour leur valeur

d'information générale » (*Loi sur les archives*, art. 2). Les fonds d'archives patrimoniales ne représentent que 5 à 10 % de la masse documentaire originale d'une personne physique ou morale ou d'une famille (Lambert 1999, 167). En effet, un tri est effectué en fonction de critères d'évaluation. Parmi ces critères figure l'élimination des doublons. Généralement, un fonds d'archives patrimonial est composé de documents existant en un seul exemplaire[10], et chaque document a une seule place dans l'organisation hiérarchique du fonds. Chaque fonds d'archives patrimonial est unique. Ainsi, un archiviste n'aura pas à indexer deux documents ou ensembles de documents identiques. Par ailleurs, le coût de traitement (description et indexation) des documents d'archives est très élevé car il nécessite une analyse unique qui n'est jamais réutilisable. Pour chaque fonds, l'archiviste doit procéder à une nouvelle indexation : par l'analyse, il obtient une combinaison de concepts unique, qu'il doit ensuite traduire en termes d'indexation. Selon les services d'archives, l'indexation peut porter sur les documents du fonds d'archives ou sur les notices de description archivistique, ce qui modifie ce qui est couvert par l'indexation, par le volume et par le niveau de l'analyse. D'après nos observations de plusieurs milieux, l'indexation des notices est très fréquente et s'explique par le fait qu'il est plus simple d'indexer une notice descriptive que plusieurs boîtes ou mètres linéaires de documents. Les termes d'indexation consignés par des archivistes dans des instruments de recherche mènent rarement à un document particulier mais plutôt à un ensemble de documents déjà regroupés selon les principes archivistiques tels qu'un dossier, une série ou un fonds (Hudon 1997-1998, 95; Hennicke 2011, 511). La notice descriptive est le résultat d'une première analyse qui est faite par un archiviste, une analyse parfois différente de l'archiviste qui indexe. En effet, la lecture documentaire est orientée en fonction de l'activité globale effectuée par le lecteur (Waller et Masse 1999) : ici, la condensation ou l'indexation. Les concepts auxquels l'archiviste indexeur accède sont de deux types, les uns directement issus des documents s'il les consulte pour les indexer, les autres extraits d'une première analyse des documents, la notice descriptive. Cette différence influence le choix des

[10] En réalité, le fonds (surtout les fonds d'archives publiques des ministères et des grandes organisations) peut contenir de multiples copies d'un même document dans plusieurs dossiers portant sur des affaires différentes.

termes d'indexation. La norme NF ISO 999:1996 indique qu'« il convient de choisir les points d'accès [c'est-à-dire les termes d'indexation] à partir de la terminologie employée dans le document » (10, section 7.2.1.2). Dans la notice descriptive, il y a déjà des termes qui sont employés pour dénoter les concepts thématiques principaux du fonds, ce qui peut influencer les choix terminologiques effectués par l'archiviste lors de l'indexation. Mais nous ne connaissons pas la proportion de reprise des termes entre la notice descriptive d'un document et les termes d'indexation, en archivistique ou en bibliothéconomie (Bertrand-Gastaldy *et al.* 1994, 32). Les données de notre étude pourraient servir à identifier cette proportion au sein du corpus. La divergence ou la reprise des termes de description comme termes d'indexation pourrait avoir un impact sur la recherche de l'usager. Par exemple, un usager peut chercher précisément un terme (p. ex. *prison*) et le trouver dans l'index (p. ex. *prisons*) mais ne pas le retrouver par la suite dans la notice descriptive (les termes reliés au concept évoqué par le terme d'indexation pourraient être p. ex. *milieu carcéral*, *peine*, *détention*).

La singularité d'un fonds d'archives influe sur l'indexation et plus largement sur la recherche de l'usager : la variation terminologique pour la description et l'indexation d'un fonds d'archives à un autre, d'un service d'archives à un autre oblige l'usager à effectuer plusieurs requêtes. L'usager souhaite des résultats pertinents et exhaustifs *tout de suite* (Schaffner 2015), Cela se produit quand il interroge Internet (Yu et Young 2014) et non une base de données spécifique où le langage d'indexation pourrait ne pas être contrôlé, ou partiellement. L'impact de la singularité des documents d'archives sur l'indexation thématique est dû à l'utilisation largement répandue du vocabulaire libre pour indexer les fonds d'archives. Nous traitons de ce type de langage d'indexation et de langage de recherche dans la section 3.4 ci-après.

À cause de la potentielle variation terminologique entre deux expressions linguistiques qu'elle entraîne, la singularité des fonds d'archives est une caractéristique des documents d'archives dont on doit tenir compte dans une étude du vocabulaire de l'accès thématique aux documents d'archives.

3.1.2. Hiérarchie d'un fonds d'archives

La définition de *fonds d'archives* dans les *Règles pour la description des documents d'archives* (*RDDA*, BCA 2008) donne une caractéristique distinctive des documents d'archives par rapport aux objets documentaires généralement conservés dans les bibliothèques et les musées. Il est fait mention de la structure interne du fonds qui prend la forme d'une hiérarchie ayant pour principe de division les activités et les fonctions pour rendre compte de la manière dont le fonds a été constitué :

> Ensemble de documents de toute nature *réunis automatiquement et organiquement*, créés et/ou accumulés et utilisés par une personne physique ou morale ou par une famille *dans l'exercice de ses activités ou de ses fonctions*. (BCA 2008, D5, nous soulignons)

Cette idée de hiérarchie est également exprimée dans l'énoncé de principes (P) des *RDDA* : « P5.1 Les niveaux de classification et de description forment un système hiérarchique » (BCA 2008, xxvi). L'archiviste indexe non seulement des documents (niveau physique de la description), mais des regroupements de documents (niveaux de synthèse de la description). Ces regroupements sont basés sur l'organisation interne du fonds, sur sa classification. Les niveaux de classification constituent également les niveaux de description, comme le préconise le principe archivistique mentionné par les *RDDA* : P5.0 « La description reflète la classification (c'est-à-dire que les niveaux de description sont établis en fonction des niveaux de classification) » (BCA 2008, xxv-xxvi). Le niveau le plus englobant est le fonds qui se divise en séries, elles-mêmes divisées en dossiers, eux-mêmes rassemblant des pièces. Si la complexité du fonds l'exige et pour faciliter son appréhension intellectuelle, on ajoute des niveaux intermédiaires : les sous-séries, sous-sous-séries, etc. Les documents physiques sont les pièces ; les niveaux supérieurs appartiennent à la classification, l'organisation intellectuelle appliquée à ces documents physiques. Le regroupement des documents est différent selon le niveau de la classification, qui est aussi le niveau de description et d'indexation. Les concepts principaux de chacun des rassemblements (fonds, série, etc.) sont de ce fait différents. La structure hiérarchique typique d'un fonds d'archives – autrement dit sa classification – n'est pas d'origine philosophique (comme cela est le cas de la classification bibliographique), mais plutôt organique : elle découle principalement des activités et des fonctions du créateur.

La présence d'une hiérarchie dans un fonds d'archives a une répercussion sur le vocabulaire employé pour l'indexation. Les normes d'indexation mentionnent généralement que l'on doit choisir le niveau de vocabulaire le plus spécifique : « Lorsqu'on utilise un thésaurus, il faut sélectionner le terme le plus spécifique existant pour représenter une notion donnée. » (ISO 5963:1985) En archivistique, à chaque niveau de la hiérarchie, on adapte la spécificité du vocabulaire en fonction de l'analyse. Prenons l'exemple d'un fonds portant sur plusieurs écoles. Le terme « établissement scolaire », s'il se trouve à un niveau élevé de la hiérarchie, tel que le fonds par exemple, dénotera un sujet général, les établissements scolaires en général. À un niveau inférieur, on retrouvera plutôt « école » ou « collège » ou encore « séminaire ».

En outre, la spécificité du vocabulaire est à pondérer avec la mission de l'établissement. « Elle [la spécificité de l'indexation] est très variable […]. Le niveau d'indexation restera plus général pour un catalogue de bibliothèque que pour un outil de recherche documentaire destiné à une recherche rétrospective fine. » (AFNOR 1993) Par exemple, le service d'archives d'une communauté religieuse ne va pas retenir pour terme d'indexation de ses fonds le mot « religion », cela serait inutile car il serait employé pour de trop nombreux documents dans l'ensemble de ses fonds et collections. Par contre, dans le vocabulaire religieux, les mots « novice » ou « vœux temporaires » seraient des termes que l'on pourrait envisager et dans le vocabulaire général, les mots « coiffe » ou « recette de cuisine » pourraient être pertinents. Mais si les notices de cet établissement sont versées dans une base de données plus large, telle que celle du Réseau de diffusion des archives du Québec (RDAQ) ou si un projet de repérage est mis en place au sein du projet de la TCARM (cf. début de la section 1), alors il pourrait tout de même être pertinent de retenir « religion » comme terme d'indexation, afin de discriminer les fonds de cette communauté religieuse de ceux d'autres établissements.

En corollaire à l'existence de la hiérarchie d'un fonds d'archives, il existe le principe du général au particulier : P5.2 « La description doit procéder du général au particulier » (BCA 2008, xxvi). À chaque niveau de la hiérarchie, on peut procéder à l'indexation. La notice descriptive présente des termes d'indexation pour le niveau supérieur d'un fonds d'archives ou pour les niveaux inférieurs. Il existe généralement au moins trois à quatre

niveaux de description dits de synthèse avant d'arriver à la pièce, au document d'archives lui-même, situé au niveau physique. À moins d'une politique d'indexation qui indique le contraire, les niveaux inférieurs ne sont indexés que si cela est nécessaire ; dans la pratique, il arrive qu'un dossier ou une pièce aient été indexés parce qu'ils sont très demandés par les usagers ou parce qu'ils entrent dans une exposition virtuelle par exemple et qu'il faut pouvoir retrouver ce document rapidement s'il est demandé. Mais ces cas particuliers dérogent au principe du général au particulier selon lequel l'archiviste analyse et décrit en premier le fonds d'archives.

La norme ISO 999:1996 parle des « points d'accès des différents niveaux utilisés dans les entrées d'index » (1996, partie 1, section 1) et non des niveaux de l'indexation. L'indexation qui porte sur chacun des niveaux doit s'adapter à son degré de précision. Il est question de la spécificité de l'indexation (qualité des termes) et de son exhaustivité (quantité des termes). Ces deux critères de l'indexation doivent être adaptés au niveau de description choisi (fonds, série, dossier, pièce). À partir du principe du général au particulier, il est courant de commencer par indexer le niveau du fonds puis les niveaux inférieurs. Plus le niveau est élevé dans la hiérarchie (fonds ou série), plus on retiendra des termes généraux. Plus le niveau est bas dans la hiérarchie (dossier ou pièce), plus on retiendra des termes spécifiques. Les *RDDA* recommandent de ne pas répéter une information à plusieurs niveaux : P5.3 « L'information fournie à chaque niveau de description doit être pertinente à ce niveau » (BCA 2008, xxvi). L'archiviste doit alors chercher le niveau de la hiérarchie approprié pour indiquer tel concept afin que l'information ne soit pas redondante. Ainsi, les liens hiérarchiques qui participent de la nature du fonds d'archives sont à rapprocher des relations sémantiques qui sont en jeu entre les termes d'indexation qui décrivent les documents et regroupements de documents des divers niveaux de la hiérarchie. Finalement, tout cela est à pondérer avec la mission de l'établissement conservateur.

L'indexeur s'adapte à la nature des documents qu'il indexe. L'indexeur de documents d'archives tient compte de leur singularité et de la hiérarchie présente dans un fonds. Ainsi, l'unicité des documents d'archives et les liens hiérarchiques des fonds d'archives rendent les sujets eux-mêmes complexes c'est-à-dire renfermant de nombreuses relations sémantiques imbriquées les unes dans les autres, ce que notre étude illustre (voir section 5). L'archivistique

dispose de peu d'outils propres pour l'indexation thématique (Pugh 2005; Duff 2010). Les archivistes ont envisagé d'utiliser les outils développés pour des documents d'une autre nature. Dans son essai de recourir aux vedettes-matière de la Bibliothèque du Congrès des États-Unis pour indexer des documents d'archives, Smiraglia (1990) constate que la spécificité des documents d'archives rend les langages documentaires qui n'ont pas été créés spécialement pour eux difficiles à appliquer, tant pour l'indexation que pour la recherche. Cette faiblesse des langages documentaires pour l'indexation thématique des documents d'archives montre une force du vocabulaire libre. Il a cela de précieux qu'il permet d'être en adéquation totale avec le sujet dont il est question dans les documents d'archives qui portent sur des sujets uniques et hiérarchisés d'une manière unique, organique pour chaque fonds.

3.2. Référence archivistique

« Le service de référence est considéré comme étant une partie intégrante de la pratique professionnelle archivistique et centrale dans la relation usager-archiviste » (Oliver, Jamieson et Daniel 2017, 1). La référence archivistique est la rencontre entre les archives et les usagers (Bédard et Morel 2013-2014). Elle est considérée par plusieurs archivistes comme une fonction à part entière (Couture et Lajeunesse 2014, 161-167; Munn et Rioux 1998). Elle s'intègre à une vision large de la diffusion qui englobe également la valorisation (Lemay et Klein 2012, 17). Issus du traitement des documents d'archives, les instruments de recherche – par exemple, l'état général des fonds, les guides thématiques, les répertoires et inventaires, ou encore les notices descriptives et les termes d'indexation en eux-mêmes – sont les outils privilégiés pour aider l'archiviste de référence à répondre aux questions thématiques et non thématiques des usagers. Ils sont qualifiés d'« essentiels, tant pour l'archiviste que le chercheur » par Cardinal *et al*. (1984, 21). Si la description suit les normes nationale (*RDDA*) et internationale (ISAD(G)) en vigueur, l'indexation doit s'établir en fonction des besoins des usagers auxquels elle est destinée (Da Sylva 2015, Hudon 2013, 35). Les usagers des archives – i. e. les personnes qui font un usage de documents d'archives – sont désormais connus et les archivistes peuvent tirer profit des études d'usagers (Schaffner 2015). Ces dernières ont identifié les principaux obstacles que rencontrent les usagers pour accéder aux archives, parmi lesquels le vocabulaire occupe une place prépondérante.

Cette section présente la référence archivistique en tant que situation de communication; nous illustrons nos propos à partir d'exemples de notre corpus. Nous présentons ensuite des études d'usagers relatives au vocabulaire employé par les usagers dans les établissements patrimoniaux et, finalement, la notion de fossé sémantique propre à notre recherche.

3.2.1. La référence archivistique comme situation de communication

Pour reprendre les mots de Bédard et Morel, la référence est une « rencontre entre les archives et l'usager » grâce à l'archiviste, « un intervenant au cœur d'une médiation réussie » (2013-2014, 54). Giuliano s'étonne que, dans la formation des archivistes, seulement quelques heures soient consacrées à la référence au sein de plusieurs cours alors que l'usager a, plus que jamais dans le monde numérique, dit-il (2012, 3), besoin d'un guide. Duff et Fox (2006), après avoir sondé deux services d'archives, appuient cette idée que l'archiviste de référence est davantage un guide qu'un expert. Selon leur étude, s'il a besoin de bases solides en histoire puisqu'il doit avoir une excellente connaissance des fonds sous sa garde (leur provenance, les types de documents, leur contenu), il doit aussi avoir une bonne compréhension des besoins des usagers et savoir comment fonctionne le système d'accès aux documents (par contenu et par provenance, pour reprendre la dichotomie de Lytle 1980a et b). En tant qu'expert, l'archiviste donnerait une réponse toute faite à l'usager; en tant que guide, il oriente l'usager dans les fonds et il propose des éléments de signification à l'usager qui les interprète en fonction de ses besoins.

3.2.1.1. Le processus de référence[11]

Le processus de référence archivistique a surtout été étudié du point de vue de l'usager et non de celui de l'archiviste (Oliver, Jamieson et Daniel 2017), il s'agit d'un champ de

[11] Nous remercions François Cartier, chargé de cours à l'École de bibliothéconomie et des sciences de l'information de l'Université de Montréal, qui nous a le premier montré une schématisation du processus de référence archivistique dans son recueil de cours (Cartier 2008).

recherche habituellement négligé (Duff et Fox 2006). Pourtant l'archiviste de référence occupe une place cruciale dans la chaîne communicationnelle entre l'usager et les documents étant donné les notions relatives à la nature des documents d'archives (voir section 3.1) :

> due to the complex arrangement of archival records and the procedural difficulties associated with accessing archival materials, reference archivists provide the sole link between users and records, and, therefore, perform a crucial role within archival institutions. (Duff et Fox 2006)

La nature des documents d'archives ainsi que les délais de traitement habituellement longs et l'arrérage dans le traitement poussent l'usager à recourir aux services de l'archiviste de référence. Les documents d'archives sont

> "non-browsable and non-circulating," direct interaction between the archivist and the researcher is essential, especially when catalogs provide inadequate access to the repository's holdings. (Ruth 1988, 268-269)

Ainsi, l'usager doit acquérir un ensemble des connaissances archivistiques pour repérer les documents d'archives dont le fonctionnement des instruments de recherche. L'ensemble des connaissances archivistiques qu'un usager expert possède après de multiples consultations d'archives a été décrit par Yakel et Torres ; il s'agit du modèle de l'expertise de l'usager dans les archives (Yakel et Torres 2003, 53[12]). Ce modèle inclut la connaissance du domaine général du sujet de la recherche, la littératie de l'objet qui consiste à repérer la forme, la structure, le type de document et à évaluer son authenticité[13] et l'entendement archivistique (notre traduction de *Archival Intelligence, AI*). Ce dernier concept est décrit selon trois composantes : 1) la connaissance des principes archivistiques, pratiques et institutions, 2) l'habileté à développer des stratégies de recherche et 3) comprendre le lien qui unit les documents d'archives et leur représentation. Les auteurs soulignent que l'acquisition de

[12] Ces concepts n'ayant pas encore été cités par des auteurs en français, nous les traduisons au mieux de notre compréhension de ce qu'ils représentent.

[13] Les auteures expliquent que la littératie de l'objet consiste à « lire un texte comme un objet et un objet comme un texte » (Yakel et Torres 2003, 76, n. t.). Nous rapprochons cette connaissance de l'approche de la diplomatique moderne (voir Chabin *et al*.2016; Delmas et Blouin 1996).

l'entendement archivistique devrait être considéré par les archivistes comme de leur responsabilité dans l'éducation aux usagers (2003, 52, n. t.).

Duff et Fox (2006) se sont intéressées au processus de référence archivistique du point de vue de l'archiviste. Elles précisent que l'archiviste de référence met en relation les usagers d'archives et l'information et que, par consequent, « fielding research requests and meeting the complex and diverse needs of users requires a unique combination of customer service and research skills, a mix of broad historical and organisational knowledge, and a significant amount of training and expertise » (Duff et Fox 2006, 129).

Ainsi le processus de référence archivistique est propre à cette discipline parce qu'il repose en partie sur la nature même des documents. Comme l'identifie le modèle *Archival Reference Knowledge* (*ARK*, Duff, Yakel et Tibbo 2013), la référence archivistique nécessite des connaissances particulières de la part de l'usager et de l'archiviste de référence, les deux principaux acteurs de cette situation de communication.

3.2.1.2. La référence comme situation de communication

La situation qui fait intervenir un archiviste de référence, un usager et des documents d'archives avec leurs représentations est une situation de communication. De ce fait, nous l'avons considérée comme telle, à partir d'un schéma de la communication. Parmi les schémas de la communication, nous avons retenu celui de Jakobson (1963, voir Figure 1), largement utilisé en linguistique parce qu'il intègre la dimension du contexte de la communication. Et ce point, justement, nous intéresse pour rendre compte de la spécificité de la référence archivistique.

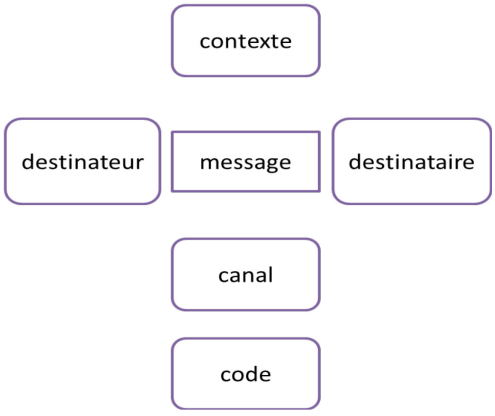

Figure 1. Schéma de la communication (Jakobson 1963)

Les six facteurs inaliénables de la communication verbale proposés par Jakobson sont les suivants : le destinateur (autre nom de l'émetteur), le destinataire (autre nom du récepteur), le message, le code, le canal et le contexte. Nous avons adapté ce schéma à la situation de communication qu'est la référence archivistique lors d'un entretien de référence différé (voir Figure 2).

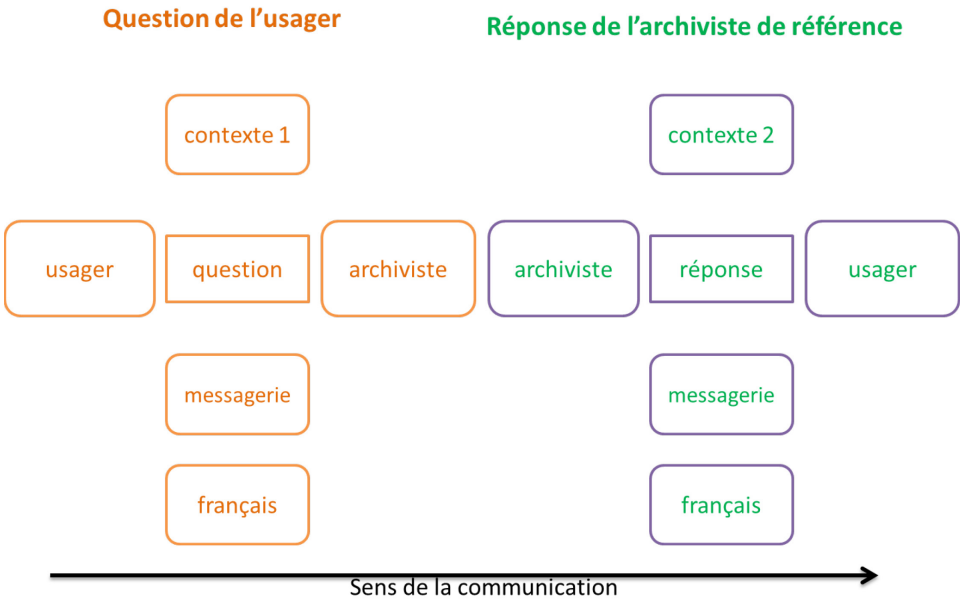

Figure 2. Schéma de la communication pour la référence archivistique (à partir de Jakobson 1963)

Dans la situation de communication particulière qu'est un entretien de référence différé par courriel, le *message* est la question posée. Le *code* est la langue, le français dans notre cas, ainsi que la netiquette (Office québécois de la langue française 2013). Le *canal* est à la fois le réseau Internet et le logiciel de courrier électronique (avec des adresses électroniques valides). Le *destinateur* est l'usager qui pose la question à un archiviste de référence qui est donc le *destinataire*. Dans le schéma original de Jakobson, le *contexte* recouvre l'ensemble des éléments disponibles au destinateur pour ancrer son message dans le réel, mais ces éléments doivent pouvoir être raisonnablement accessibles au destinataire. Ici, le *contexte* est l'ensemble des éléments qui construisent la situation de recherche de l'usager que celui-ci partage avec le destinataire, l'archiviste : ses anciennes recherches, sa motivation à effectuer

cette recherche, son environnement professionnel ou familial, etc. Pour l'archiviste qui répond, le contexte est essentiellement son contexte institutionnel (p. ex. les habitudes de description et de mise en valeur auprès des usagers internes, externes, les pratiques organisationnelles, la terminologie spécifique ou encore les compétences relationnelles). C'est pourquoi nous avons répété l'élément *archiviste* comme *destinateur* du second schéma, positionné à la suite du premier selon une ligne du temps qui sert à indiquer le sens de lecture du schéma (c'est-à-dire de gauche à droite).

Le schéma indique à gauche le destinateur et à droite le destinataire. Nous avons indiqué le sens de la communication pour conserver ces rôles, même lors de la réponse de l'archiviste. Il est possible que l'archiviste qui a reçu la question ne soit pas l'archiviste qui y répondra, à cause de la spécificité du domaine par exemple (p. ex. le patrimoine bâti, la jurisprudence) ou les types de support (p. ex. documents textuels, iconographiques, cartes et plans, etc.) ou encore une période (p. ex. la Nouvelle-France). L'entretien de référence est une situation de communication qui met en relation un usager et un archiviste de référence, à propos de documents d'archives.

Dans sa réponse à l'usager, l'archiviste de référence fait intervenir les instruments de recherche existant dans le service d'archives. L'instrument de recherche joue un « rôle vital » pour l'archiviste de référence (Duff et Fox 2006, 149, Ruth 1988). Dans notre recherche, les archivistes de référence proposaient à l'usager parfois la consultation d'instruments de recherche « rédigés » (Direction des archives de France 1993, 154), de « premier niveau » (Cardinal *et al*. 1984, 19-20), c'est-à-dire en fonction du niveau de description, du dépôt à la pièce (p. ex. état général des fonds, guide par fonds, répertoire numérique, etc.). Mais les archivistes de référence proposaient le plus souvent des cotes archivistiques et des titres d'unités de description de divers niveaux ou bien des liens vers des résultats de recherche listant directement des notices descriptives de plusieurs niveaux. Dans notre recherche, les instruments de recherche que nous avons retenus et observés sont la notice descriptive et les termes d'indexation présents dans les portails archivistiques accessibles en ligne. À partir de nos observations, nous avons schématisé la chaîne communicationnelle des usagers aux documents d'archives à partir d'une question d'usager posée par courriel à la référence (voir Figure 3).

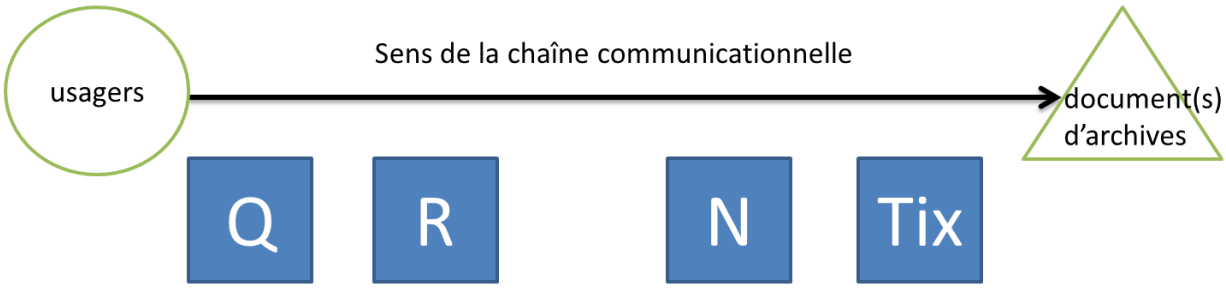

Figure 3. Chaîne communicationnelle des usagers aux documents d'archives à partir d'une question d'usager posée par courriel à la référence

Cette chaîne communicationnelle inclut les questions d'usagers posées par courriel à la référence (Q), les réponses d'archivistes de référence (R), les notices descriptives des unités archivistiques recommandées par les archivistes de référence (N) et éventuellement les termes d'indexation qui sont rattachés à chaque notice (Tix). Ces quatre éléments (QRNTix) ont formé les sources des termes étudiés dans notre recherche.

3.2.1.3. L'archiviste de référence, médiateur sémantique[14]

La fonction de référence investit l'archiviste d'un rôle de médiateur de sens.

> (…) [Le] médiateur incarne, puis co-fabrique des mises en circulation de savoirs, de contenus, de formes, d'œuvres avec d'autres formes, d'autres espaces, d'autres relations, d'autres citoyens. Le médiateur culturel devient le déclencheur de cette chaîne opératoire d'actions invisibles et décisives, le catalyseur d'une réaction chimique – celle de l'ouverture au monde, à soi, à l'autre (Camart *et al*. 2015, 10).

L'archiviste de référence joue le rôle de médiateur culturel, mais également de médiateur sémantique entre les documents d'archives et les usagers (voir Figure 4). « L'archiviste-référencier joue le rôle de médiateur entre les usagers et la chaîne de traitement archivistique. » (Giuliano 2012, 5)

[14] Cette section est tirée d'un chapitre de livre accepté : Guitard (2018).

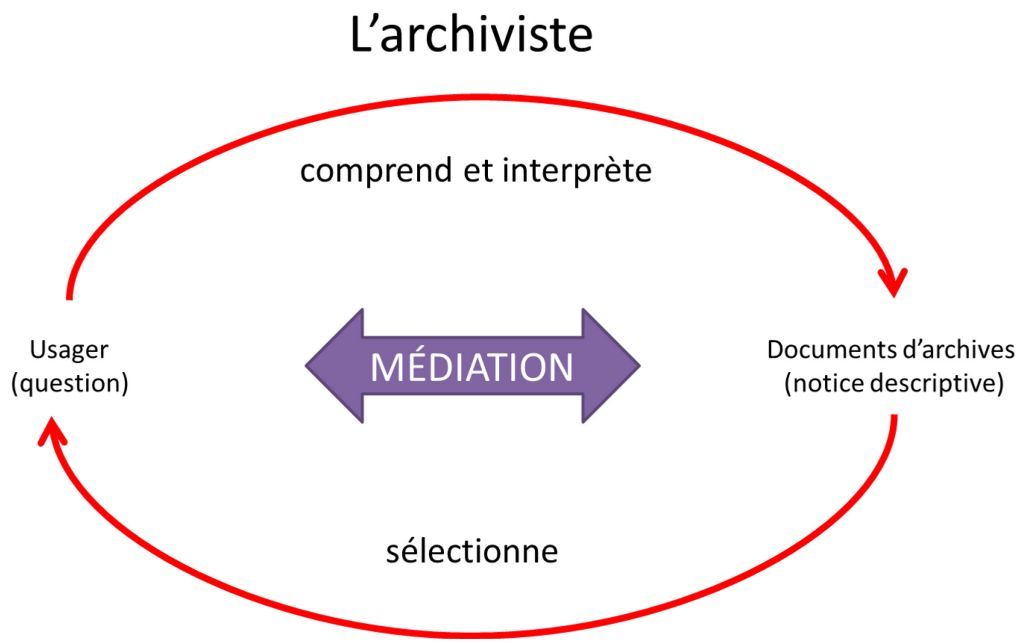

Figure 4. L'archiviste de référence, médiateur sémantique

Le premier rôle de l'archiviste de référence est donc d'être un agent cognitif : il est capable de comprendre et interpréter la question de l'usager (le message) sur le plan formel, sémantique et référentiel. Ensuite, il sélectionne les documents qui répondent à cette question. Cette sélection s'opère notamment en fonction du contexte de recherche exposé par l'usager dans sa question.

En interprétant le message de l'usager, en y répondant par une sélection de documents, l'archiviste ajoute une couche de sens aux documents, une « surcouche informationnelle » (Gauthier 2014). Cet ajout d'information pourrait se concrétiser par l'ajout de termes d'indexation. La médiation sémantique opère en réduisant le fossé sémantique entre le vocabulaire de l'usager et celui des instruments de recherche. Nous l'observerons en qualifiant l'écart sémantique entre les deux. Nous y reviendrons plus loin (voir section 6.4.3).

La référence a été largement étudiée à partir du point de vue des usagers. Même si l'on ne connaît pas encore le processus avec précision, grâce aux études d'usagers, nous sommes en mesure d'identifier les obstacles qui peuvent entraver la référence et d'identifier ce que les usagers désirent (Schaffner 2015).

3.2.2. Études d'usagers relatives à leur vocabulaire

Selon Schaffner (2015), après 30 ans d'études d'usagers menées selon de nombreuses méthodes différentes (*focus group*, entrevue, questionnaire, jeu de rôle, etc.), nous avons les moyens de savoir ce qu'attendent les usagers des archives patrimoniales. D'après une recension de ces études, elle résume les pratiques et besoins des usagers qui mènent une recherche. Les usagers des services d'archives :

- recherchent par sujet et par mots-clés;
- aimeraient des résultats classés par pertinence;
- sont capables de parcourir de longs instruments de recherche s'ils sont assurés de trouver l'information qu'ils recherchent;
- n'utilisent pas de portails archivistiques ni de catalogues de bibliothèque pour commencer leur recherche (ils n'en connaissent pas l'existence);
- ne viennent dans les établissements culturels tels que les bibliothèques et les services d'archives que pour des items connus (Schaffner 2015, 86, n. t.).

« Unfortunately, there is a gap between the expectations of users and historical descriptive practices in archives and special collections. » (Schaffner 2015, 91) L'auteure préconise dès lors d'améliorer la description des documents d'archives, quitte à « tordre les règles de description » (*bending rules for description* p. 91, n. t.), afin de rendre les notices descriptives davantage visibles sur Internet et compréhensibles pour l'usager. Elle propose par exemple d'inclure des synonymes dans les notices descriptives. L'ensemble lexical se verrait alors enrichi et permettrait moins de silence dans les résultats. Notre recherche montre à partir de quelles relations sémantiques nous pourrions enrichir le vocabulaire présent dans les instruments de recherche pour qu'il s'apparie au vocabulaire des usagers.

La description et l'indexation sont effectuées par l'archiviste afin de faciliter la recherche de documents d'archives par les usagers. Light « posit[s] that we do archival description in order to facilitate user's discovery of materials. (…) We create finding aids to help people find things. » (2008) Il est incontournable dans une étude sur l'accès aux documents d'archives de prendre en considération les besoins et habitudes des usagers. C'est pourquoi, pour étudier le vocabulaire employé pour l'accès thématique aux archives patrimoniales (VATAP), notre corpus inclut les courriels échangés entre les usagers et les archivistes de référence en plus des notices descriptives et des termes d'indexation. La chaîne

communicationnelle (voir Figure 3) étudiée a pour point de départ les courriels envoyés par les usagers.

Le vocabulaire employé par des usagers pour accéder à des archives patrimoniales n'est pas seulement saillant dans des courriels envoyés à l'archiviste de référence, mais aussi dans les étiquettes, des termes d'indexation librement générés par des usagers à propos de documents en ligne. Theimer (2011) et Duff, Johnson et Cherry (2013) soulignent la valeur ajoutée d'une contribution des usagers : en plus de mieux les connaître pour mieux les servir, l'indexation collaborative permet aux archivistes de développer une communauté d'usagers contributeurs autour d'un service d'archives. Golder et Hubermann (2006) ont étudié les étiquettes lors des premiers engouements pour l'étiquetage. Ils ont montré qu'elles présentent une stabilité inattendue et en ont dégagé un modèle qui permet de prédire certains schémas récurrents dans le partage de connaissances. C'est pourquoi, pour rendre compte du vocabulaire employé par les usagers, nous avons inclus dans notre conception du VATAP non seulement des courriels d'usagers à la référence mais aussi des étiquettes qu'ils ont générées et rattachées à des documents d'archives ou, plus rarement, à des notices descriptives. Cependant, bien qu'elles fassent partie du VATAP, les étiquettes n'ont pas pu entrer dans le corpus de la présente recherche (voir section 4.1.2.4).

Contrairement à Schaffner qui préconise de modifier seulement la manière de décrire, Pugh (2005) et Duff (2010) invitent les archivistes à repenser leurs instruments de recherche en promouvant l'accès thématique, essentiellement en développant l'indexation par sujet. Dans notre propre travail, nous avons comparé les termes de recherche avec les termes de description et les termes d'indexation. Nous avons pu alors identifier la proportion de reprise entre le vocabulaire des usagers et celui employé dans les instruments de recherche. Comparer les termes de description et les termes d'indexation pourrait alors permettre de rendre compte de l'utilité de l'indexation thématique dans les instruments de recherche actuels, par rapport à la simple existence de la notice descriptive dans laquelle l'usager peut chercher des mots en texte intégral.

Notre étude, incluant des données directement issues des recherches des usagers, vise à apporter des éléments pour améliorer les résultats de recherche tant dans la description que

dans l'indexation, sur la base de l'étude linguistique du vocabulaire employé pour l'accès aux documents d'archives.

Parmi les études d'usagers que nous avons repérées, plusieurs proposent une méthodologie dont nous avons pu nous inspirer pour notre recherche. Certaines études fournissent un cadre d'analyse de contenu de courriels d'usagers à la référence.

Gagnon-Arguin a dressé une typologie des questions d'usagers, une typologie des stratégies de recherche et une classification des sujets de questions (Gagnon-Arguin 1998, 91-99), dans une étude sur les archivistes utilisant la méthodologie du *sense-making*. Mais les indicateurs permettant de dresser les typologies ne sont pas détaillés, il s'agit *a priori* du seul jugement de la chercheuse. Cependant l'auteure souligne que les questions dans les courriels d'usagers à la référence sont un « matériau d'étude des usagers » (Gagnon-Arguin 1998, 86).

Veal (2011) a, dans sa thèse de doctorat, étudié 762 questions de référence d'une bibliothèque universitaire en ligne et en a analysé le contenu. Son but était d'identifier les thèmes et les mots exprimant des émotions, en relation avec les phases du modèle de recherche d'information de Kuhlthau, afin de permettre aux bibliothécaires de mieux servir les étudiants. Sa méthodologie, trop spécifique à ce modèle de recherche d'information, n'a finalement pas pu nous être utile.

Un des problèmes de sens récurrent dans les questions d'usagers est l'ambigüité. Kim établit trois sortes d'ambigüité dans les questions de référence posées en contexte numérique : ambigüité sémantique, syntaxique et pragmatique (2014, 316 L'ambigüité sémantique inclut principalement la polysémie (p. ex. *chien* : chien de révolver et chien de porte) et l'homonymie (*voler* : voler en avion et voler un avion). L'ambiguïté syntaxique couvre des phénomènes de structure de phrase vague, dont le sens peut changer selon le choix d'une structure grammaticale ou son alternative (p. ex. *la belle porte le voile* : une jolie femme cache quelqu'un ou une jolie porte le dissimule à peine; *belle* est respectivement un nom ou un adjectif qualificatif, *porte* est un nom ou un verbe, *voile* est un nom ou un verbe)). Enfin, l'ambigüité pragmatique repose sur l'absence ou le manque de certains éléments contextuels dans la question pour qu'on puisse lui attribuer une seule réponse, les éléments sémantiques et syntaxiques, eux ne faisant pas défaut. Kim donne l'exemple suivant d'une question d'usager

ambigüe sur le plan pragmatique, *i. e*. sur le plan de la réalité : « Quelle est la plus brillante étoile visible de la Terre ? » (2014, 316). L'adjectif *visible* est relié à un contexte d'observation à l'œil nu ou bien avec un instrument. Cet élément du contexte change la réponse. Les exemples que Kim donne et sa méthodologie nous ont servi à développer notre propre méthodologie de l'analyse des questions dans les courriels d'usagers à la référence. Dans notre étude, il serait particulièrement intéressant de relier les phénomènes d'ambiguïté à la spécificité des informations comme Kim le fait. Par exemple, une question sur le temps peut recevoir une réponse sur l'année, ou l'année et le mois, ou bien l'année, le mois et le jour, ou encore l'année, le mois, le jour, l'heure et ainsi de suite en évoquant des informations de plus en plus précises. Nous voyons dans la section méthodologie comment ces études ont contribué à l'élaboration de notre cadre d'analyse de notre corpus (voir section 4.2.1.1).

Les études d'usagers révèlent un fossé sémantique entre le vocabulaire qu'ils utilisent et celui employé dans les instruments de recherche. Cette notion de fossé sémantique fait l'objet de la prochaine section.

3.2.3. Notion de fossé sémantique

Jörgensen (2007) retrace l'origine de l'expression *semantic gap* dans le milieu informatique ; elle était employée pour identifier une différence entre deux descriptions d'un objet utilisant deux langages différents, en particulier la différence entre une description lisible par un être humain et une représentation informatique. L'expression est reprise notamment par Oomen *et al.* (2010) à propos des archives. Ils proposent de combler ce fossé sémantique en archivistique par le recours aux étiquettes générées par les usagers. Celles-ci offrent des interprétations différentes et de l'information additionnelle selon Choi et Syn (2016). Les étiquettes correspondent alors tout à fait à la « surcouche informationnelle » dont parle Gauthier (2014) à propos de la médiation à opérer entre usagers et ressources dans le milieu muséal. Pour Hennicke (2011), il existerait un « fossé sémantique » entre la formulation des besoins des usagers et les documents d'archives. Nous nous appuyons précisément sur Hennicke (2011) pour considérer ce fossé d'un point de vue linguistique en l'appelant « écart sémantique » (voir section 3.5.4) puisqu'il ressort de la comparaison d'expressions linguistiques (voir section 4.3.1).

Dans notre recherche, nous adoptons le point de vue de l'archiviste lors de la référence, ce qui a été assez peu étudié jusqu'à présent (Oliver, Jamieson et Daniels 2017). Les études d'usagers sur la référence ont donné les indices qu'il existait probablement un fossé entre les formulations des usagers et celles existant dans les instruments de recherche, relativement à l'accès thématique.

3.3. Indexation thématique en archivistique

L'indexation thématique en archivistique consiste en la représentation des sujets de documents d'archives. La notion de sujet est donc au cœur de l'accès thématique aux documents d'archives patrimoniales. Pourtant, un flou définitionnel entoure cette notion et ne permet pas d'identifier clairement, autrement qu'au cas par cas, à partir d'une lecture humaine, ce que sont les sujets présents dans un courriel d'usagers à la référence, une notice descriptive, un index ou un nuage d'étiquettes. Cette section fait un bref tour d'horizon de la notion pour en définir des paramètres opérationnels dans notre recherche. Ainsi, nous présentons dans un premier temps la notion spécifique de sujet en archivistique. Dans un deuxième temps, après avoir rappelé le besoin d'accès thématique particulièrement criant en archivistique, nous présentons les lignes directrices utiles pour l'indexation thématique archivistique. Finalement, nous aborderons la notion de facettes comme méthode d'appréhension du sujet en sciences de l'information.

3.3.1. Notion de sujet en archivistique

À la suite des sections 3.1 et 3.2, nous abordons la notion de sujet en archivistique par le biais de l'indexation. Voici la description qu'en font Maurel et Champagne :

> Rappelons brièvement que l'indexation non thématique permet d'extraire des points d'accès indépendants du sujet (principalement reliés à la provenance des documents), tandis que l'indexation thématique permet d'extraire des sujets et d'autres éléments de « contenu » qui peuvent couvrir plus large – ou différemment – que le seul sujet (par exemple une fonction administrative, une activité, un type de document, etc.). (Maurel et Champagne 1999, 324)

À partir de cette citation, nous remarquons que l'indexation thématique couvre non seulement le sujet mais aussi d'autres *éléments de « contenu »*. L'analyse de contenu permet d'extraire plus que des sujets. L'indexation thématique s'oppose à l'indexation onomastique qui couvre

autre chose que le sujet. Les éléments qui sont autres que le sujet sont couverts à la fois par l'indexation non thématique et thématique, ce qui cause un flou définitionnel. La différence entre les deux indexations porte généralement sur le fait que l'indexation onomastique (ou non-thématique) porte sur des noms propres. Tout ce qui n'est pas un nom propre ne pouvant être présenté dans un index onomastique se retrouve dans un index thématique. Maurel et Champagne (1999) poursuivent :

> Comme nous le mentionnons précédemment, la notion de « sujet » en archivistique couvre une réalité assez vaste considérant les diverses manifestations possibles du « contenu » dans les documents. À tenter de circonscrire ce que comporte la notion de « sujet » dans l'ensemble du cycle de vie des documents, il est possible d'avancer qu'il représente généralement :
>
> - un thème comme tel;
> - une fonction, une activité (par exemple une activité administrative) ou une action du créateur du fonds ou d'un auteur spécifique;
> - une occupation, une profession ou un corps de métier;
> - un lieu (de naissance, de décès, d'emploi, de résidence, etc.) où s'est déroulée une action ou une activité;
> - une donnée chronologique (une date de naissance, de décès, d'un événement, d'un fait historique), le plus souvent utilisée comme sujet secondaire en relation avec une action ou une activité;
> - un type ou un support de document utilisé le plus souvent comme sujet secondaire en relation avec une action ou une activité;
> - un nom d'individu, de famille ou d'organisme en relation avec les actions ou les activités, ou dont traitent les documents, pour autant que ce nom ne soit pas relié, ici, à la provenance[15]. (Maurel et Champagne 1999, 330)

Nous remarquons dans cette citation qu'un *thème* – pour reprendre la terminologie employée ici – n'est qu'un des *sujets* – *idem* – généralement inclus dans un index thématique et ne forme que le premier des sept points. Les autres points peuvent-ils réellement être appelés sujets ? À quoi correspond le sujet qui est un thème si ce n'est pas une fonction, une occupation, un lieu,

[15] Le texte renvoie ici à une communication de 1997 dont la version écrite a été publiée l'année suivante (Gagnon-Arguin 1998).

une donnée chronologique, un type de document[16] ni le nom d'un individu ? Il semble que « thème » soit défini par la négative. Nous ne pouvons toujours pas donner une définition théorique positive de ce qu'est un sujet en archivistique. Nous nous tournons alors vers la pratique archivistique et examinons où le sujet de documents est décrit dans une notice descriptive, selon la norme canadienne (*RDDA*).

Dans la notice descriptive, quelles sont les zones de description renfermant les sujets d'un document ? Le sujet relève du contenu thématique des documents et non de l'histoire administrative ni de la notice biographique qui peuvent dépasser (avant ou après) les dates extrêmes de l'unité archivistique décrite (ces informations se trouvent généralement au niveau du fonds). Une recherche en texte intégral dans la notice entière plutôt que seulement dans ces zones pourrait donc retourner des résultats inappropriés et causer du bruit. La pertinence des résultats basés sur l'indexation par sujet dépend donc des zones indexées. Le sujet des documents d'archives est décrit dans la partie « Portée » de la zone « Portée et contenu », telle que décrite dans les *RDDA* : « L'élément Portée et contenu renseigne sur la structure interne ou la classification et sur le contenu des documents. » (*RDDA* 2008, 1-52). Contrairement à ce que pourrait laisser penser l'ordre des constituants de la définition des *RDDA*, la *portée* décrit ce sur quoi porte l'unité archivistique et le *contenu* présente les catégories de documents inclus, les documents d'exception ou remarquables et indique les unités archivistiques inférieures dans la hiérarchie classificatoire, le cas échéant. Or, la portée et contenu n'est pas nécessaire lorsque le titre de la pièce est suffisamment explicite (*RDDA* 2008, I-64, règle 1.7D4). Le sujet est ainsi décrit dans la portée et contenu ou dans le titre. Ceci nous amènera, lors de notre analyse, à considérer ces deux éléments dans une notice descriptive pour identifier le sujet des documents décrits (voir section 4.2.1.2).

Dans le cadre de notre recherche, le sujet d'un document est le contenu thématique auquel peut correspondre le sujet d'une question de recherche thématique d'un usager. Cette

[16] Nous ne savons pas si le « type de document » correspond à la « catégorie de document » dont parlent les *RDDA* (BCA 2008). Nous préférons employer l'expression utilisée dans les actuelles *RDDA*.

définition n'est pas opérationnelle car comment savoir ce qu'est le sujet d'une question de recherche d'un usager et l'identifier clairement ? Afin de ne pas se lancer dans des considérations philosophiques qui sortiraient de notre propos principal, nous déterminons la présence d'un sujet en fonction des éléments linguistiques qui l'introduisent. Les expressions linguistiques qui forment les termes d'indexation sont considérés comme un sujet à partir du moment où il est indiqué que l'index est thématique. Il en va de même pour les étiquettes. Dans une notice, le sujet est exprimé de manière directe dans un titre et introduit dans le texte de la portée par des expressions linguistiques telles que *l'unité archivistique porte sur, traite de, évoque*. Pour les sujets de courriels d'usagers à la référence, nous avions imaginé qu'ils seraient précédés d'expressions linguistiques telles que *porter sur, aborder, parler de, traiter de, relatif à, s'intéresser à*. Nous présentons une liste exhaustive des formulations d'introduction du sujet dans les notices descriptives et les courriels d'usagers à la référence après l'examen de nos données (voir section 5.5). L'expression formelle du sujet nous a servi d'indice pour repérer concrètement le sujet dans les données textuelles de notre corpus. Quelques travaux de détection du sujet en linguistique ou en traitement automatique de la langue nous ont aidée à identifier le sujet, par exemple Marandin (1988) et Longo et Todirascu (2010) par l'explicitation de critères formels identifiés en français (voir section 4.2.4).

Dans notre recherche, la notion de sujet en archivistique est définie surtout formellement. Dans la prochaine section, nous nous demandons quelles sont les lignes directrices qui encadrent l'accès thématique et plus précisément l'indexation thématique.

3.3.2. Lignes directrices archivistiques pour l'indexation thématique

L'accès thématique fait l'objet de plusieurs normes nationales et internationales en documentation / pour les documents publiés dont les deux principales sont ISO 5963 sur l'analyse des documents, la détermination de leur contenu et la sélection des termes d'indexation (ISO 1985) et NF Z 47-102 sur les principes généraux pour l'indexation des documents (AFNOR 1993). Ces normes présentent trois étapes pour indexer : l'appréhension du contenu total du document, l'identification des concepts représentant ce contenu et la sélection des concepts nécessaires à la recherche ultérieure. En archivistique, la norme nationale canadienne, les *RDDA*, traite de l'accès onomastique, c'est à dire exprimé par des

noms propres (de personnes ou de lieux), que ceux-ci soient employés pour représenter un sujet ou la provenance ou un autre métadonnée. Cette norme présente uniquement la manière de les écrire et non des critères de sélection. Le Bureau canadien des archivistes (BCA) avait rédigé un rapport sur l'indexation thématique des documents d'archives en 1992 et se reporte à l'analyse thématique pratiquée dans le monde de la documentation tout en attirant l'attention sur certains éléments différents tels que la nature des documents d'archives (voir section 3.1.).

Adoptant une vue d'ensemble, Pugh (2005) et Duff (2010) redisent le manque d'indexation par sujet dans les services d'archives à partir d'une réflexion sur le service de référence qui y est offert, service chargé de répondre aux courriels d'usagers. Elles constatent le manque d'outils thématiques à la disposition tant des usagers que des archivistes eux-mêmes, à commencer par des politiques d'indexation ou d'accès par sujet. Vilar et Šauperl (2015) participent à un projet d'écriture de lignes directrices pour l'indexation thématique de documents d'archives en Slovénie et Bosnie Herzégovine. Il est intéressant de voir que la première étape de leur recherche a consisté en une étude d'usagers et des compétences professionnelles dont les archivistes ont besoin pour les servir (voir Duff, Yakel et Tibbo 2013 pour une description de ces compétences aux États-Unis). Nous voyons ici le lien indéfectible entre recherche et représentation thématique (description et indexation) qui est à la base de l'accès thématique en archivistique. Le besoin de lignes directrices pour l'accès par sujet aux documents d'archives se fait sentir dans la profession. La rédaction de lignes directrices sur l'indexation thématique de documents d'archives précède normalement la rédaction d'instruments de recherche thématiques adaptés au contexte archivistique. Ce fait contribue à démontrer l'intérêt de notre recherche sur l'accès thématique aux documents d'archives. Et nous tenterons de contribuer à la réponse à ce besoin de la communauté archivistique par notre recherche.

3.3.3. Notion de facettes

Le sujet de documents a été abordé en documentation à partir des besoins liés au catalogage des documents publiés. En archivistique, il a été conseillé de ne pas recourir à un classement thématique pour préserver le lien organique entre les documents par l'application du principe de respect des fonds. Dans notre étude, nous avons eu besoin de procéder à une classification thématique des expressions linguistiques porteuses de sujet (thémanymes). Nous

appelons cette opération la caractérisation sémantique. Plutôt qu'un étiquetage sémantique fin et propre à chaque thémanyme du corpus qui ne serait pas transférable, nous avons décidé de nous reporter à une caractérisation plus générale qui permet de regrouper les éléments s'apparentant sémantiquement d'une section du corpus à l'autre, c'est-à-dire les éléments se répondant l'un à l'autre entre la question, la réponse et les termes d'indexation.

Étudiant dans un département de sciences de l'information s'inscrivant en sciences sociales, nous ne pouvons ignorer la méthode d'analyse qui a été utilisée initialement par plusieurs grands noms de la bibliothéconomie tels que Bliss, Otlet ou Ranganathan dès les années 1930, mais formalisée par ce dernier en 1957 (La Barre 2010b, 106). La pertinence des facettes a été redécouverte dans l'environnement numérique. Elle a été appliquée notamment en archivistique aux plans de classification d'organisations (Mas 2011). Il s'agit d'un « concept incontournable » (Hudon et El Hadi 2017, 12).

3.3.3.1. Définition d'une facette

Il n'y a pas de consensus sur la définition de *facette*. Cette notion est comprise différemment selon la région du monde (les États-Unis, le Royaume-Uni, l'Inde) et cela s'est cristallisé dans des traditions différentes. Selon Ranganathan, une facette est « a generic term used to denote any component of a compound subject, also its ranked forms, terms and numbers » (1967, 88 cité dans La Barre 2010b, 106). Broughton et Slavic (2004) indiquent que les facettes sont des ensembles homogènes de termes ou de concepts qui sont le résultat de la technique de l'analyse en facettes. La Barre (2010b, 106) reconnaît que les facettes sont généralement utilisées pour représenter une partie de sujet, ou l'attribut d'un objet et peuvent être combinées (réunies) pour représenter des sujets complexes. Elles [ont la capacité d'] « intégrer différentes dimensions d'analyse mutuellement exclusives sur des sujets ou objets informationnels, à caractériser et à rendre l'accès à l'information plus facile en offrant de multiples voies de navigation vers n'importe quel document ou notice descriptive correspondant à ces sujets » (Marleau, Mas et Zacklad 2008).

Broughton indique que l'analyse en facettes est une technique qui s'opère par l'application de « catégories fonctionnelles et/ou de nature linguistique (p. ex. entités, processus, propriétés, opérations, agents) » (2003, n. t.).

3.3.3.2. Catégories fondamentales

Proposant une classification à facettes pour classer l'ensemble des connaissances dans une bibliothèque, Ranganathan a déterminé cinq facettes qui permettent de couvrir l'ensemble du savoir. Le groupe de recherche sur la classification (CRG) – dont fait partie Vickery dans les années 1960 – propose rapidement d'élargir le nombre de facettes, mais celles-ci ont une pertinence relative au domaine des sciences naturelles. Nous nous reportons alors uniquement aux cinq catégories originelles, qu'il s'agit dès lors de bien comprendre.

> Misunderstanding or confusion often arises when attempting to understand the nature of several of the original PMEST categories […]. Vickery (1966, 25) [17] offers clarification: "*Space and Time* are relatively straightforward geographical and chronological schedules. *Energy* covers categories such as problem, method, process, operation, handling and technique. *Matter* comprises constituent materials of all kinds. *Personalities* (the most difficult for many to grasp) include libraries, numbers, equations, wavelengths of radiation, engineering works, organisms, crops, religions, art styles, literary works, languages, social groups, communities and states." Many have observed that these vague terms offer a robust framework for the creation of more precise categories. » (La Barre 2010a, 252)

Bellec rappelle en une courte glose entre parenthèses ce que couvre chacune des cinq facettes de Ranganathan : « personnalité (le concept principal du document) ; matière (une substance ou une propriété) ; énergie (l'opération ou action subie par l'objet) ; espace (localisation géographique) ; temps (localisation chronologique et temporelle) » (Bellec 2016, 28).

On reconnaît aux cinq facettes fondamentales une grande polyvalence et une remarquable adaptabilité à chaque nouveau contexte. « [A]ll subjects can be analyzed in a way that fits these [Ranganathan's] five categories » (Hjørland 2008, 90). Cette dernière caractéristique entérine notre choix de recourir aux facettes dans notre méthode d'analyse du sujet de nos données : cela facilitera le rapprochement d'expressions linguistiques porteuses de

---

[17] La référence citée par l'auteure est la suivante : Vickery, B. C. 1966. *Faceted classification schemes* (Rutgers Series on Systems for the Intellectual Organization of Information, vol. 5). New Brunswick, NJ : Greaduate School of Library Service, Rutgers, the State University.

sujet (thémanymes) des quatre sources de données de notre corpus, *a priori* hétéroclite thématiquement.

3.3.3.3. Analyse en facettes

L'analyse en facettes est une manière d'approcher un sujet et de le décomposer en éléments de sens appartenant à des groupes homogènes et mutuellement exclusifs, pour reprendre la formulation de Vickery (1960). La Barre (2010b, 107) distingue, à la suite de Vickery (1960, 12-13), l'analyse en facettes et la classification à facettes : la classification est le fruit de l'analyse. « The technique of facet analysis relies upon a list of fundamental categories or distinguishing characteristics. » (La Barre 2010b, 107) L'auteure ne précise pas la nature de ces catégories fondamentales ni celle des caractéristiques distinctives.

Les étapes de l'analyse en facettes ont déjà été identifiées. « The technique of facet analysis follows seven basic steps. The first three steps are pure facet analysis; the last four demonstrate how facet analysis may assist in the creation of the structure of a faceted classification » (Vickery 1960, 12-13 cité par La Barre 2010b, 107). Nous ne nous intéressons pas à la classification par facettes mais à l'analyse en facettes, alors nous examinons seulement les trois premières étapes.

Les étapes liées à l'analyse en facettes sont :

- Defines the domain and interests of domain participants
- Formulates facets by examining representative domain material by:
  - collecting terms that reflect domain participant interests, and
  - sorting terms into homogeneous, mutually exclusive groupings (facets).
- Structures each facet in a hierarchical order to coalesce synonyms, identify misaligned terms, and gasps in the system. (Vickery 1960, 12-13[18] cité par LaBarre 2010b, 107)

Dans notre recherche, nous identifions à la suite de ces auteurs deux étapes de l'analyse en facettes qui pourraient convenir à une caractérisation sémantique ayant peu de niveaux hiérarchiques, au sein de notre corpus de langue générale (sans domaine défini) qui

[18] La référence citée par l'auteure est la suivante : Vickery B. (1960). *Faceted classification: A guide to construction and use of special schemes*. Londres, UK : Aslib.

sert – nous le rappelons – à rapprocher les quatre types de données de notre corpus incluant des mots différents mais évoquant le même sujet :

a. collecter des termes porteurs de sujet;

b. regrouper des termes dans des ensembles thématiquement homogènes et mutuellement exclusifs.

Cette analyse en facettes opère selon l'analyse logique qui permet d'identifier des catégories abstraites (La Barre 2010b, 110-111), à partir des éléments qui apparaissent dans le corpus. Certaines classifications à facettes incluent plusieurs types de facettes. Ainsi, Ménard, Mas et Alberts (2010) ont recours à trois types de facettes, des facettes contextuelles, des facettes de propriétés physiques et des facettes thématiques dans un modèle de taxinomie pour Artefacts Canada. En effet, « [l]e processus d'analyse conceptuelle permet de déterminer à partir de la collection de termes qui reflètent les intérêts d'un groupe d'usagers, quelles entités ou facettes sont pertinentes à ce groupe. » (Mas, 2013-2014, 85) Nous recourons non seulement à des facettes thématiques, relatives au contenu exprimé par les données, mais aussi à d'autres types de facettes de nature non thématique (relative au document archivistique, au contexte et à la stratégie de recherche, ou encore à la métalinguistique) (voir arbre de codage section 4.1.3.1).

3.3.3.4. Autres méthodes d'analyse du sujet similaire en linguistique : le champ sémantique

Murray (2003[19], cité par La Barre 2010b, 111) se demandait quelles étaient les approches similaires à l'analyse en facettes. Formée en linguistique, nous pouvons rapprocher l'analyse de facettes de l'analyse sémique qui détermine les plus petites unités de sens distinctives (des sèmes ou traits sémantiques) ce qui rappelle la caractérisation donnée par La Barre ci-dessus. Ce ne sont pas les plus petites unités de sens, mais les plus petites unités de

[19] Murray P. et Faceted Classification List. (2003). « Viewpoint (scope of FC list) », message 97, 10 Janvier 2003, disponible à : http://www.oclc.org/research/activities/fast/default.htm.

sens distinctives. Pour se distinguer les unes des autres, il est nécessaire de les analyser dans un bassin d'unités au sein duquel elles prennent place. Par exemple, Pottier (1963) recourt à l'analyse sémique pour décrire le champ sémantique du siège à partir de six traits sémantiques (p. ex. *avec bras, pour une personne*) pour distinguer *chaise, fauteuil, pouf, sofa*, etc.[20]. Dans notre recherche, le bassin de mots est constitué par les expressions linguistiques porteuses de sujet (thémanymes) d'une question, une réponse et une notice descriptive.

L'analyse en facettes est en fait un regroupement sémantique, à un niveau très général. Chaque facette comprend un trait sémantique général[21] propre au regroupement de thémanymes, par exemple, les activités ou les entités. Le vocabulaire contient un grand nombre de noms ayant pour trait sémantique *activité* ou *entité*. À un niveau plus spécifique, les traits sémantiques sont plus précis, plus proches du sens tel que décrit dans un dictionnaire d'usage général. Ce sont ces derniers qui sont le point commun entre des expressions linguistiques d'un champ sémantique. « Le champ sémantique de S – champ sémantique de la cuisine, du sport, des sentiments…- est un regroupement de lexies[22] dont les définitions ont en commun le sens de 'S' - 'cuisine', 'sport', 'sentiment'… » (Polguère 2016, 228) Ainsi, le champ sémantique regroupe des mots appartenant à plusieurs facettes, mais plus proches du point de vue du sens (p. ex. le champ sémantique de la cuisine regroupe *cuisine, cuisinier, casserole, brûlé, à point*, etc.). Nous utilisons la notion de champ sémantique pour raffiner la caractérisation sémantique suite à l'analyse en facettes (voir section 4.2.5). Un champ sémantique est, dans notre recherche, un regroupement de thémanymes ayant en commun dans leur définition un élément de sens.

L'évocation de ces quelques notions linguistiques d'ores et déjà dans la section consacrée aux notions archivistiques nous permet de faire la transition vers la section suivante.

---

[20] Pour un autre exemple d'analyse sémique, voir p. ex. Mounin 1965.

[21] Bien que ce trait participe du sens du thémanyme inclus sous une facette (Polguère 2016 parle plutôt de *classes de noms*), il opère à un niveau très générique.

[22] Dans le cadre de cette recherche, nous préférons *expression linguistique* à *lexie*, même si ce dernier terme a un sens bien particulier en linguistique (voir Polguère 2013 et 2016).

L'imbrication de ces deux champs disciplinaires dans une recherche sur l'indexation par sujet est inexorable.

Notions linguistiques

Les références qui suivent appartiennent pour la plupart à la linguistique. Elles présentent le fonctionnement de la langue indépendamment d'un usage particulier. Les normes documentaires et autres lignes directrices des sciences de l'information proposent quant à elles des manières de faire et des codifications qui partent de l'outil (l'index ou le terme d'indexation) et de son usage (la recherche) pour restreindre l'usage de la langue dans ce contexte ainsi défini. Notre recherche explore les rouages de la langue connus des linguistes pour améliorer l'usage de la langue que font les archivistes et les usagers dans le cadre de l'accès aux archives patrimoniales. C'est pourquoi nous recourons à ces deux types de littérature. La présente section regroupe des notions linguistiques qui traitent, d'une part, des langages et vocabulaires employés pour favoriser la mise en relation des usagers avec les documents ou l'information dont ils ont besoin et, d'autre part, de la sémantique des termes employés pour l'accès : le sens en linguistique, le sens en sciences de l'information et les modèles de règles pour créer des termes d'indexation.

Le type d'indexation est le résultat de la combinaison d'un langage d'indexation et du statut d'un indexeur. En effet, il existe plusieurs types de langages d'indexation – essentiellement les langages documentaires et le vocabulaire libre – et plusieurs types d'indexeurs – principalement professionnel et amateur ou bien archiviste et usager. Nous examinons les contraintes langagières qui pèsent sur les combinaisons de langage d'indexation et d'indexeurs. Plus largement, l'étude des contraintes sur les langages employés pour l'accès constitue l'objet de la section 3.4.3.

Le produit de l'indexation, à savoir les termes d'indexation, constitue un ensemble de mots qui intéresse la linguistique en tant qu'ils forment une manifestation de la compétence langagière d'êtres humains. Les principaux problèmes relatifs aux langages d'indexation ont rapport au sens des mots, principalement la synonymie (plusieurs formes pour un sens) et la polysémie (plusieurs sens pour une forme). L'ensemble des phrases qui composent les notices descriptives et les échanges de courriels d'usagers à la référence constituent des textes,

également objets d'étude pour la linguistique. La composante sémantique des mots – autrement dit la composante lexico-sémantique – sert d'assise à la sélection d'un mot plutôt qu'un autre par l'archiviste qui décrit ou indexe manuellement ou par l'usager qui indexe ou pose une question, actions pour lesquelles ils exercent leur jugement. L'étude des phénomènes de sens dans les index, ou plus largement dans les instruments de recherche (description, indexation professionnelle, indexation collaborative), et sur les termes de recherche employés dans des courriels d'usagers à la référence, est alors incontournable dans une étude du vocabulaire employé pour l'accès aux documents d'archives.

3.4. Vocabulaire employé pour l'accès thématique aux documents d'archives (VATAP)

La présente section traite du vocabulaire employé pour l'accès thématique aux documents d'archives (VATAP), c'est-à-dire du vocabulaire employé en recherche, en description et en indexation. Nous reprenons l'ordre identifié dans la chaîne communicationnelle de la référence archivistique allant des usagers aux documents d'archives (voir Figure 3, section 3.2.1.2). Nous nous penchons d'abord sur la recherche, puis sur la description et finalement sur l'indexation, section largement plus approfondie que les deux autres puisque l'indexation thématique a justement pour fonction principale l'accès thématique, comme déjà évoqué au début de la section 3 Revue de la littérature.

3.4.1. Termes de recherche

Dans notre étude, les termes de recherche forment un ensemble de mots et expressions linguistiques au sein de phrases et de textes librement formulés par les usagers dans un courriel envoyé à un archiviste de référence. Dans le cadre de cette production linguistique, les seules règles qui pourraient s'appliquer sont celles, implicites, des échanges épistolaires et celles de la nétiquette. Ainsi, les usagers ont peu ou pas de contraintes dans leur rédaction. Quelques tendances dans leurs habitudes langagières ont déjà été répertoriées. Dans cette section, nous survolons les études d'usagers principalement en archivistique afin de relever les habitudes langagières déjà identifiées et de cerner certains de leurs besoins en vocabulaire pour l'accès thématique aux documents d'archives.

Dans le but d'améliorer les instruments de recherche, Duff et Johnson (2001) ont étudié 375 questions de référence posées par courriel à divers types de services d'archives américains (fédéral, provincial, universitaire, municipal, etc.) et ont observé en quels mots les usagers exprimaient leur besoin informationnel. Les auteures ont catégorisé le type de question puis le type d'information. Selon cette double catégorisation, elles ont regroupé les questions de référence des usagers en, respectivement, trois et huit catégories telles que *formulaire, information générale, sujet, événement, lieu, date*. Les auteures s'attachent à catégoriser le besoin informationnel et n'analysent pas les mots employés d'un point de vue linguistique. Mais leurs catégories nous seront utiles pour déterminer quels sont les courriels d'usagers qui comportent une question thématique en tant que telle et quelles sont les demandes d'un autre type, p. ex., une demande d'une reproduction d'un document d'archives (voir section 4.2.1.1.).

Actuellement, les études du vocabulaire des usagers portent davantage sur les étiquettes qu'ils génèrent que sur les courriels qu'ils envoient à la référence. Nous n'avons trouvé jusqu'à présent que peu de références sur la question du vocabulaire dans les questions de référence posées par courriel et seulement deux études qui touchent d'une manière ou d'une autre notre recherche. Veal (2011) a mené une analyse qualitative du contenu de courriels d'usagers à la référence. Elle s'est intéressée à l'expression des sentiments des usagers en vue de les rapporter au modèle de recherche d'information de Kuhlthau. Bien qu'ancienne, l'étude de Gagnon-Arguin (1998) est intéressante pour notre recherche. L'auteure a étudié les questions d'usagers posées par courriel à la référence dans le but de mieux connaître les usagers d'un service d'archives québécois, ce qu'ils cherchaient et la manière dont ils formulaient leur besoin. Il est intéressant de noter qu'elle recense 41% de questions d'usagers ayant une teneur thématique exprimée par des noms communs (Gagnon-Arguin 1998, 92). Par ailleurs, l'étude des courriels d'usagers à la référence peut servir à « ajuster, d'une part, le traitement à accorder aux fonds d'archives décrits et, d'autre part, [à] créer des structures d'accès propres aux recherches menées dans les fonds d'archives » (Gagnon-Arguin, 1998, 87). L'étude des courriels d'usagers à la référence pourrait ainsi permettre une adéquation – ou du moins un arrimage – entre la réalité des recherches effectuées dans les fonds d'archives par les usagers et les instruments de recherche proposés par les archivistes. L'auteure établit un

lien entre la catégorie d'usagers du service d'archives et le niveau de la réponse à donner (Gagnon-Arguin, 1998, 93). Nous ajoutons qu'il serait intéressant de corroborer ce fait à, par exemple, la précision sémantique des termes dans la réponse de l'archiviste, ce que notre analyse linguistique de contenu des données tentera de montrer. Cette étude s'est reportée à la réponse de l'archiviste pour établir les « stratégies de recherche » (Gagnon-Arguin, 1998, 94). Elle est remontée jusqu'à la notice descriptive pour identifier la relation entre le type de questions, la stratégie de recherche et la description (Gagnon-Arguin, 1998, 96). Mais l'auteure a mené une étude sur la présence *potentielle* du type d'information en fonction des recommandations des *RDDA* sur le niveau hiérarchique où il doit se trouver. Elle n'a pas mené une étude sur la correspondance *effective et réelle*, dans les notices descriptives, des concepts et sujets, ce que notre recherche se propose de faire, à partir de l'étude du vocabulaire employé par les usagers et dans les instruments de recherche. Une autre similitude de l'étude de Gagnon-Arguin avec la nôtre est qu'elle traite de données en français. Cependant, elle a une portée archivistique et non linguistique. Ce point renforce l'idée de la complémentarité des deux disciplines pour arriver à des conclusions plus précises sur la nature sémantique du vocabulaire pour l'accès aux archives patrimoniales.

Engerer (2017) a comparé le vocabulaire employé par les usagers et celui employé pour l'indexation grâce à un thésaurus. Cette étude rend saillants divers points sur la comparaison du lexique (la langue générale employée par les usagers) et sur la terminologie (la langue propre à un domaine particulier comme le reflète un thésaurus élaboré par des professionnels de l'information). Bien qu'elle se place dans un contexte bibliothéconomique, les notions présentées dans cette étude nous ont été utiles pour l'analyse de nos données. Ainsi, nous comprenons que le degré de contrôle et l'agencement des mots dans les expressions linguistiques de recherche, de description ou d'indexation jouent un rôle de premier plan pour l'intercompréhension de l'usager et de l'archiviste de référence.

L'étude des termes de recherche a cela de précieux qu'elle pourrait aider à améliorer la description afin de mieux correspondre aux besoins des usagers, comme le souligne Light dans ses conseils aux archivistes pour que leur collection soit visible des moteurs de recherche : « [w]hen summarizing the content of the collection, use meaningful key words that users would likely search for. » (Light 2008, 3) L'auteur invite les archivistes à faire correspondre le

vocabulaire employé dans les instruments de recherche au vocabulaire des usagers (et non l'inverse).

La question de la diachronie dans la langue et dans le traitement des archives est très intéressante. Elle pose un défi de taille aux usagers et de ce fait aux archivistes. Une langue vivante évolue avec le temps; d'une part, les documents d'archives ont été produits à un moment (parfois sur une longue période), d'autre part, les instruments de recherche ont été produits à un ou plus généralement plusieurs moments. Nous pensons que les archivistes ne peuvent pas demander aux usagers d'adapter leur vocabulaire à celui de divers instruments de diverses époques. Par ailleurs, les archivistes ne peuvent pas sans cesse ajuster le vocabulaire de tous les instruments de recherche. Il serait bon de se suivre les développements de la recherche sur cette question, notamment en terminologie.

3.4.2. Termes de description

Nous appelons les termes de description un ensemble d'expressions linguistiques au sein de phrases et de textes librement formulés par les archivistes quand ils rédigent la notice descriptive. La description archivistique est normée par les *Règles pour la description des documents d'archives* (*RDDA*, BCA 2008) qui décrivent les éléments à inclure dans une notice descriptive ainsi que leur ordre de présentation. Dans la notice descriptive, la zone qui présente le plus de sujets est, selon le rapport du BCA sur l'indexation par sujet (BCA 1992), la zone Portée et contenu (cf. section 3.3.1.). Cette zone contient un résumé de ce dont traitent les documents (la portée) et de ce que contient l'unité de description, à savoir les types de documents et éventuellement le détail de la section subordonnée de la classification (le contenu). Ce rapport de 1992 – toujours cité comme référence dans les *RDDA* (2008) – indique également les différents types de résumés que l'archiviste peut produire, en fonction de l'utilité des documents pour les usagers et de la politique de description de l'établissement (si elle existe). Selon le type de résumé (annotation, résumé descriptif, sélectif ou

informatif)[23], il est recommandé à l'archiviste de reprendre ou non la terminologie employée dans les documents. S'il n'est pas contraint de reprendre le vocabulaire des documents, l'archiviste peut alors adapter les termes de description au vocabulaire des usagers.

Il n'existe pas d'autres lignes directrices pour la rédaction de notices descriptives archivistiques au Canada. Ceci a pour répercussion une plus grande variété de styles dans les ensembles de phrases qui forment la Portée et contenu puisque l'archiviste est libre d'adopter le style qu'il préfère, en l'absence de lignes directrices institutionnelles.

Nous ne connaissons pas la proportion des termes directement repris de la notice descriptive (Bertrand-Gastaldy *et al*. 1994, 32) ou des documents d'archives pour constituer des termes d'indexation, tel que mentionné ci-dessus. De plus, lorsque les termes ne sont pas repris à l'identique, nous ne savons pas quelle relation sémantique ils entretiennent avec la représentation dans les instruments de recherche. Cette question trouve un début de réponse dans notre travail.

3.4.3. Termes d'indexation

Après les termes de description et les termes de recherche, nous considérons les termes d'indexation, soit ceux que l'on trouve dans les index ou dans les nuages d'étiquettes.

D'un point de vue opérationnel, nous entendons par *indexation professionnelle* l'assignation de termes d'indexation à des documents d'archives ou à des notices descriptives par une personne œuvrant dans un service d'archives (un professionnel, un technicien ou un bénévole d'un service d'archives). Le résultat est donc sous l'autorité du service d'archives qui a la garde des documents et est utilisé dans la recherche de documents par ce service. Quant à l'*indexation collaborative*, elle consiste en de courts segments linguistiques, souvent des noms communs (Kakali et Papatheodorou 2010, 192), attribués par des usagers de manière collaborative.

---

[23] Pour la description complète des quatre types de résumés employés en archivistique pour la Portée et contenu, voir BCA 1992. Nous les présenterons brièvement dans la thèse.

La recension des études du type d'indexation en fonction du type de langage d'indexation (libre ou contrôlé) et du statut de l'indexeur (professionnel ou amateur, archiviste ou usager) permet d'avoir un panorama des recherches ayant un intérêt pour l'étude des termes d'indexation. Nous avons retenu trois combinaisons de ces deux types : l'indexation collaborative par les usagers en vocabulaire libre, l'indexation classique par les archivistes avec un langage documentaire et l'indexation par les archivistes en vocabulaire libre. La quatrième combinaison possible – les usagers qui utiliseraient un langage documentaire – n'a pas été retenue pour plusieurs raisons : d'abord, nous n'avons pas trouvé de références sur des tentatives l'implantant et, ensuite, « One cannot help but wonder whether such enthusiasm for metadata would be the same if people were asked to use only prescribed and standardized vocabularies. » (Spiteri 2005, 85[24], cité par Peters 2009, 1).

3.4.3.1. Usagers et vocabulaire libre

La première combinaison concerne les usagers qui indexent en vocabulaire libre. À l'ère du Web 2.0, les usagers ajoutent aux documents une strate d'information ou « surcouche informationnelle » (Gauthier 2014) qui leur est utile pour les retrouver. Peters (2009) décrit les *folksonomies*, c'est-à-dire l'ensemble des termes résultant de l'indexation collaborative; elle consacre un livre à l'étude des étiquettes d'usagers pour l'indexation et la recherche. Elle mentionne notamment que les folksonomies adoptent une démarche différente de la classification et de la structuration des collections numériques. « Instead of choosing a classification criterion and filling it with resources, it is now the resources that are allocated the criteria. » (Peters 2009, 3) En archivistique, les usagers choisissent les mots qui forment les étiquettes en fonction des documents présents dans la collection et non sur un découpage prédéfini du monde. Lezcano, Garcia-Barriocanal et Sicilia (2012) rapprochent automatiquement des étiquettes de favoris et des ontologies et dressent des parallèles sémantiques; ils appliquent ces rapprochements dans un système de recommandation à

[24] Il semble s'agir de la référence suivante : Spiteri, L. (2006). The use of folksonomies in public library catalogues. *The Serials Librarian*, *51*(2): 75-89.

l'usager pour une navigation hybride entre les deux systèmes. Cette étude met en lumière les apports d'une formalisation sémantique stricte à la souplesse des étiquettes. Lu, Park et Hu (2010) comparent les étiquettes librement générées par des usagers aux termes d'indexation par sujet tirés d'un langage documentaire (les vedettes-matière de la Bibliothèque du Congrès) choisis par des professionnels et créent des ponts sémantiques entre les deux vocabulaires. Choi et Syn (2016) présentent les caractéristiques du comportement d'usagers qui étiquettent des collections patrimoniales en ligne. Si leur méthodologie consiste en catégorisation des étiquettes et en comparaison des étiquettes avec les termes du titre ou de la description qui correspond à la ressource étiquetée, les auteurs ne précisent pas sur quels éléments (de la forme, du sens ou autres) ils se basent pour effectuer la comparaison ou la catégorisation; ils parlent seulement d'une méthode « manuelle » (Choi et Syn 2016, 1092). L'article de Broudoux (2013) retrace quelques études récentes sur l'étiquetage (le « tagging ») et en fait une lecture en sciences de l'information. Elle montre les tendances en recherche, y figure la comparaison du vocabulaire des étiquettes avec des langages documentaires.

De telles études sont des exemples des nombreuses études du vocabulaire libre généré par des usagers. Adoptant un point de vue linguistique ou terminologique (c'est-à-dire linguistique dans un domaine particulier), ces études traitent de phénomènes linguistiques tels que la forme des termes d'indexation (question de la morphologie) ou la différenciation de plusieurs sens pour une forme (question de la polysémie). Mais elles n'ont pas pour objet de pousser l'analyse linguistique jusqu'à la nature sémantique des mots. Ce travail reste à faire et pourrait révéler des phénomènes récurrents qui serviraient à éclairer certaines décisions des usagers qui étiquettent les documents d'archives. La présente recherche ne participe pas à l'étude du problème pour proposer des pistes de solutions mais suggère une méthodologie linguistique pour ce faire.

Chen, Liu et Qin (2008) ont étudié les relations sémantiques au sein de paires d'étiquettes générées par des usagers dans *Flickr*. Déplorant le manque de contexte pour identifier la relation sémantique entre deux étiquettes, ils ont créé une collection de contextes contenant les paires d'étiquettes, à partir des résultats de recherche d'un moteur de recherche dans lequel ils avaient lancé les deux étiquettes étudiées. Ils étudient la correspondance caractère pour caractère, toute variation dans la chaîne de caractères constituant une différence.

Margaret E. I. Kipp travaille depuis plusieurs années sur l'indexation collaborative. Elle décrit la mesure de la cohérence *intertaggueurs* : « Some of the consistency in tagging on delicious.com may be influenced by the incorporation of suggested terms into the interface, acting as a form of semi-controlled vocabulary. » (Kipp 2009) Le vocabulaire peut connaître plusieurs influences et non seulement deux états : contrôlé ou non; la suggestion d'expressions linguistiques comme étiquettes est un des moyens d'influencer le choix terminologique (lexical et syntaxique) de l'usager.

3.4.3.2. Archivistes et langage documentaire

La deuxième combinaison concerne les professionnels qui indexent avec un langage documentaire. Force est de constater que les cas d'archivistes recourant à un langage documentaire pour indexer des fonds d'archives sont rares ou, du moins, peu documentés dans la littérature à laquelle nous avons eu accès. Smiraglia (1990) évalue le recours aux vedettes-matière de la bibliothèque du Congrès (LCSH) pour indexer par sujet des documents d'archives et montre que ceux-ci nécessitent des sujets plus précis que les publications. Cadieux et Charbonneau (2002) traitent de l'indexation des sujets à Bibliothèque et Archives nationales du Québec (BAnQ). Ils indexent avec un vocabulaire contrôlé, sur plusieurs niveaux. Et, ils procèdent selon une approche inverse au principe archivistique, en allant du particulier au général, de la pièce ou du dossier avec une indexation spécifique pour aller au fonds à un niveau général. Cela correspond bien à la structure du lexique où les niveaux inférieurs sont plus précis. Ils ont choisi des critères d'indexation qui, les auteurs le soulignent bien, conviennent à la particularité de leur institution. Cette mise en œuvre coûteuse n'est pas à la portée de services d'archives de moins grande envergure, pourtant le besoin est là. Un sondage anglais montre que les archivistes recherchent des ressources terminologiques pour indexer les documents d'archives, mais plusieurs estiment que les ressources existantes ne conviennent pas à leurs besoins et ils n'ont pas les moyens (temps, finances, personnel qualifié) d'en créer une sur mesure (Fenton 2010). L'adéquation d'un langage documentaire aux sujets traités dans un service d'archives en particulier est donc un problème récurrent pour les archivistes. Nous tenterons par notre recherche de contribuer à la connaissance des relations sémantiques entre deux des vocabulaires en jeu et de relever les éventuelles raisons

qui indiqueraient pourquoi les langages documentaires sont si difficilement utilisables dans les services d'archives.

3.4.3.3. Archivistes et vocabulaire libre

La troisième combinaison concerne les professionnels que plusieurs contraintes poussent à choisir d'indexer en vocabulaire libre plutôt que de ne pas indexer du tout. Le manque de temps et de ressources, d'une part, et la pression de la recherche en ligne par un accès par sujet à la pièce, d'autre part, poussent les archivistes à utiliser le vocabulaire libre pour indexer. Cette tendance ne touche pas que le milieu archivistique. Le vocabulaire libre généré par des professionnels est étudié entre professionnels de plusieurs milieux du patrimoine culturel ou en comparaison avec le recours à un langage documentaire (Angel 2012; Bertrand, Célier et Giroux 1994; Rorissa 2010; Smith-Yoshimura 2012; Stewart 2013). Bertrand, Célier et Giroux (1994) ont étudié l'indexation par des professionnels avec un langage documentaire contrôlé (le langage d'indexation précoordonné RAMEAU[25]) et en vocabulaire libre par des indexeurs experts ou novices. Les auteurs redisent le principal avantage du vocabulaire libre pour un domaine particulier, à savoir l'adéquation à la terminologie employée dans les documents, ce qui est recherché pour les archives dont la terminologie varie en fonction de l'espace, du temps et des créateurs. Angel, dans sa thèse soutenue à l'Université de Caroline du Sud en 2012, compare les étiquettes descriptives apportées par des professionnels à des ressources appartenant aux trois institutions que sont les bibliothèques, archives et musées. Elle a mesuré leur cohérence inter-indexeur pour les mêmes documents. Le nombre des étiquettes ne variait pas mais la nature des étiquettes variait selon l'institution et selon la catégorie de document (objet *versus* photographie). Pour une meilleure recherche en ligne qui ne tient pas compte de l'institution d'origine des documents, l'auteure suggère de mieux contrôler la qualité des termes d'indexation utilisés lors de l'indexation en vocabulaire libre par des professionnels. Les faiblesses de l'indexation en vocabulaire libre –

[25] RAMEAU : Répertoire d'autorité-matière encyclopédique et alphabétique unifié; un répertoire de vedettes-matières (Bibliothèque nationale de France, 2017)

l'étiquetage en particulier – sont connues. Il s'agit d'une indexation fragmentée, où foisonnent les synonymes non regroupés, où le manque de cohérence est grand, où les homographes ne sont pas distingués, etc. (voir notamment Arsenault 2006). Pourtant, de Kayser (2012) compile plusieurs études sur la cohérence inter- et intra-indexeurs avec langage documentaire ou en vocabulaire libre et constate que les résultats de taux de cohérence vont de 10 à 80%. Par ailleurs, selon Wolfram, Olson et Bloom (2009), l'étiquetage – à savoir l'indexation par de non-professionnels sans le bénéfice d'un langage documentaire – est plus cohérent que l'indexation traditionnelle. Adoptant le point de vue de la linguistique, notre étude pourrait apporter un éclairage intéressant à l'étude du type d'indexation qu'est l'indexation collaborative.

Comme mentionné plus tôt (section 1), nombreux sont les services d'archives qui ont recours à cette dernière combinaison, volontairement ou par la simple recherche en texte intégral, qui équivaut à une indexation en vocabulaire libre sur chacun des mots présents dans les notices descriptives de documents d'archives par exemple. Le vocabulaire libre est souvent utilisé pour des documents d'un domaine particulier ou d'une thématique spécifique afin, par exemple, de pouvoir rendre compte des innovations les plus récentes et d'être en accord avec la recherche (Salaün et Arsenault 2009, 90). Par contre, il est peu utilisé pour des documents touchant à la vie en général (la langue générale dans son entier) à cause de ses points faibles liés au fait qu'il est proche de la langue (ambigüité et redondance de l'information). Notre recherche aura pour effet de témoigner du recours au vocabulaire libre dans des milieux archivistiques pour des archives patrimoniales aux thématiques variées, au sein de notre échantillon. En outre, l'émergence prévisible de patrons devrait nous permettre d'extraire des règles de ce qui paraît être pour l'instant un chaos lexical.

Les termes de recherche, de description et d'indexation jalonnent le cheminement du sujet à partir des usagers jusqu'aux documents, selon la chaîne communicationnelle de la référence archivistique (voir Figure 3, section 3.2.1.2). Ainsi, le vocabulaire des usagers est caractérisé par les termes de recherche issus des questions de recherche et les termes d'indexation collaborative que sont les étiquettes, alors que le vocabulaire des archivistes est employé dans les instruments de recherche par les termes de description issus des notices descriptives et les termes d'indexation professionnelle. Dans notre recherche, la réponse de

l'archiviste est tantôt considérée comme un bassin de termes de recherche, tantôt un bassin de mots appartenant au vocabulaire des archivistes, nous y reviendrons lorsque nous décrirons notre méthodologie (voir section 4).

3.5. Vocabulaire : aspect sémantique

Cette section de notre revue de la littérature présente les notions linguistiques nécessaires à la compréhension de l'analyse qui sera menée dans notre projet de recherche. Elle présente en particulier les aspects sémantiques traités dans les index et plus généralement dans le vocabulaire employé pour l'accès aux archives patrimoniales. Ces aspects sémantiques sont relatifs aux termes d'indexation, de description et de recherche. Le sens est une composante essentielle des index parce que c'est sur la base de la compréhension du sens d'un terme qui figure dans l'index que l'usager discrimine quelle entrée d'index correspond à son besoin d'information, quelle cote archivistique consulter et ainsi quel(s) document(s) d'archives pourrai(en)t permettre de répondre à sa question. D'ailleurs, les problèmes de sens (essentiellement synonymie et polysémie) préoccupent les professionnels de l'information depuis longtemps (Chaumier 1988; Deweze 1981; Hutchins 1975).

Quelques études du sens en linguistique nous permettent de comprendre comment le sens des mots et leurs relations sont abordés en linguistique (voir section 3.5.1). Suit une section sur le sens mais cette fois-ci appliqué en sciences de l'information (voir section 3.5.2). Ensuite, nous exposons un panorama des relations sémantiques étudiées en linguistique et surtout en sciences de l'information (voir section 3.5.3). Finalement, nous décrivons la notion d'écart sémantique (voir section 3.5.4).

3.5.1. Aspect sémantique des mots en linguistique

Nous nous intéressons à l'aspect sémantique du message véhiculé dans le schéma de la référence archivistique vue en tant que situation de la communication (voir Figure 2). La sémantique lexicale, la branche de la linguistique, traite des phénomènes lexico-sémantiques, c'est-à-dire ceux qui concernent le sens de mots. Elle s'intéresse d'une part aux propriétés des mots et d'autre part aux relations sémantiques. Cette section présente les principales propriétés linguistiques auxquelles s'intéressent des travaux récents en linguistique et surtout en

sémantique lexicale (pour les relations sémantiques, voir la section 3.5.3). Parmi tous les mots, nous nous intéressons principalement aux noms, puisque les termes d'indexation sont des groupes nominaux, comme exigé par les normes d'indexation. La linguistique distingue deux typologies des noms selon que les noms « sont ou non conçus sur le mode d'une partition du réel en catégories fondamentales » (Huyghe 2015, 7). « La typologie opère sur un mode « ontologique » lorsque le typage vise à faire coïncider les propriétés référentielles des noms, selon le type d'entités dénotées, et leurs spécificités de construction linguistique » (*id.*) (par exemple, les noms d'objet, noms d'événement, noms de propriété). « La typologie opère sur un mode « fonctionnel » (ou « relationnel ») lorsque le typage repose sur la description d'une relation ou d'une fonction référentielles » (*id.*) (par exemple, nom relationnel, nom partitif, nom collectif). Puisque notre recherche n'est pas destinée à des linguistes conscients des phénomènes en jeu dans la langue mais à des archivistes et des usagers *a priori* non formés en linguistique, nous retenons la première typologie qui sera – nous espérons – moins opaque ou plus accessible. En effet, « les types nominaux « ontologiques » sont définis par la dénotation de différents segments de la réalité, tels qu'ils sont communément appréhendés par les locuteurs » (Huyghe 2015, 7).

Fasciolo et Lammert (2015) mentionnent des critères définitoires pour certains noms. Dans un numéro qu'ils dirigent consacré à la typologie des noms, les auteurs relèvent deux perspectives de l'étude des noms : « soit en considérant un type de noms particulier (noms d'idéalités, prédicatifs, de couleurs, de bruits) et en examinant quels sont les critères qui permettent de circonscrire leur sous-classe, soit en appliquant un critère déterminé à une ou plusieurs sous-catégories de noms (massif/ comptable, lecture taxinomique/ floue, prédication) » (2015, 8). La lecture taxinomique (p. ex. *X est un Y*) et la lecture floue (p. ex. *X est une sorte/espèce de Y*) s'exprime dans des phrases, du texte; nous ne la retenons pas puisque nous nous intéressons aux propriétés linguistiques touchant possiblement les termes d'indexation. Les deux autres critères (opposition massif/ comptable et prédication) sont particulièrement intéressants car ils traversent un grand nombre de classes de noms, si ce n'est la majorité.

Ainsi, nous pensions nous reporter à une typologie de type ontologique, en appliquant des critères linguistiques déterminés, à savoir l'opposition massif/comptable, l'opposition

concret/abstrait, l'opposition singulier/pluriel, la prédicativité sémantique, *i. e.* la relation prédéterminée et encodée dans la langue entre des éléments de sens. Nous avions envisagé également le prototype comme meilleur exemplaire d'une classe puisque cette notion a déjà été étudiée en sciences de l'information (notamment Yeo 2008). Ce critère définitoire des noms appartient à la sémantique du prototype qui est basée sur un critère préliminaire pour l'application du prototype : l'opposition dénomination/désignation. Cette dernière aurait constitué le dernier critère retenu pour caractériser les expressions linguistiques dans notre recherche. Finalement, nous brossons un rapide comparatif des notions de nom commun, nom propre et entité nommée, utiles dans notre analyse, puis nous terminons par la notion principale de notre recherche, à savoir celle d'écart sémantique, résultant de l'application en linguistique de la notion de fossé sémantique vue précédemment dans les notions archivistiques (voir section 3.2.2).

3.5.1.1. L'opposition massif/comptable

Par exemple, BEURRE ou SABLE sont des massifs : *du beurre* et *du sable.* L'ouvrage de Flaux et Van de Velde *Les noms en français : esquisse de classement* (2000) constitue « la » référence pour la classification sémantique des noms communs en français. Les auteures indiquent en introduction qu'elles classent les noms à partir de propriétés syntactico-morphologiques ; elles établissent une gradation ou un « spectre » des noms concrets aux noms abstraits en fonction de la forme des mots et de leur combinatoire. Les auteures soulignent en introduction que leur classification n'est pas exhaustive ni un modèle à suivre. Puisque l'opposition massif/comptable traverse les typologies nominales (Fasciolo et Lammert 2015, 8) et comme nous ne savons pas par avance quels types de noms nous allons trouver dans notre corpus, nous l'avons retenue dans notre grille d'analyse. L'approche générale des auteures et plus précisément leurs tests linguistiques pourraient être utiles pour aider à classer certains termes dans l'une ou l'autre catégorie.

3.5.1.2. L'opposition concret/abstrait

Par exemple, TOMATE et CHAT sont des comptables : *une tomate et deux chats.* Les noms concrets renvoient à des « objets concrets, – disons, pour aller vite, ceux perçus par les sens » et les noms abstraits, des « objets abstraits sociaux » (Kleiber 2003, 97). Cette

différence est évidente pour certains mots (p. ex. *nounours* est concret, *idée* est abstrait) et moins claire pour d'autres (p. ex. *changement*, *écriture*). En cas de polysémie (*i. e.* lorsque le mot a plusieurs sens), un sens peut être concret et un autre abstrait (p. ex. le document-*procédure* est concret alors que la marche à suivre-*procédure* est abstraite). Nombreux sont ces derniers : les actions en particulier ont souvent un résultat concret qui correspond à l'application de leur concept abstrait (p. ex. *publication, climatisation*).

Pourtant ce trait fait partie des traits lexico-sémantiques essentiels à la description du sens des noms (Huyghe 2015; Fasciolo et Lammert 2015). C'est pourquoi il serait intéressant pour la description des propriétés des expressions linguistiques d'un corpus.

3.5.1.3. L'opposition singulier/pluriel

La variation en nombre (ou opposition entre le singulier et le pluriel) se marque généralement morphologiquement par l'ajout d'un –s aux noms communs en français. Cette variation peut rendre compte du fait qu'un nom est dénombrable-comptable plutôt que non dénombrable-massif ou encore concret ou abstrait (voir sections précédentes). Ainsi, cette marque peut être interprétée différemment selon le contexte d'apparition du nom. Le passage du singulier au pluriel et inversement, d'une source à l'autre (p. ex. de la notice descriptive au terme d'indexation ou de la question de l'usager à la notice) ne s'avère pas anodin. Ainsi, nous retenons que ce trait morphologique pourrait avoir d'éventuelles répercussions sémantiques dans l'étude des expressions linguistiques d'un corpus.

3.5.1.4. La prédicativité sémantique

Dans son manuel de sémantique lexicale, Polguère présente une classification du sens des mots (2008, 131-134). Il indique qu'il existe deux types de sens lexicaux – les sens liants et les sens non liants (ces derniers sont autonomes référentiellement, *i. e*. ils n'ont pas besoin de contexte pour savoir à quoi ils renvoient). Il en dégage une typologie des sens : les prédicats sémantiques, les noms sémantiques et les quasi-prédicats. Par exemple, le sens de TOMATE est non liant et se comprend sans aucune autre référence; c'est un nom sémantique. AMOUR est un mot pour lequel on a besoin de savoir ce qui est aimé (Roméo, son chat, chocolat, faire du sport) et éventuellement par qui (Juliette) pour être bien compris : c'est un prédicat sémantique. La troisième classe est composée d'entités qui, en même temps qu'elles

sont entités, possèdent un sens liant. Par exemple, NEZ dans *le nez de Cléopâtre* est une entité ayant nécessairement un lien avec la personne ou la chose qui possède le nez : c'est un quasi-prédicat. Les noms sémantiques se comprennent instantanément et de manière isolée, deux qualités pour figurer dans un index. C'est pourquoi cette classification tripartite de Polguère nous semble intéressante à examiner dans l'étude des propriétés linguistiques des thémanymes dans un corpus.

3.5.1.5. Le prototype

Nous reprenons la définition d'un prototype tel qu'envisagé en linguistique par Kleiber : « le prototype est l'exemplaire qui est reconnu comme étant le meilleur par les sujets (…) d'une même communauté » (1990, 48). Plus couramment, « a prototype is usually envisaged as a mental mapping of typical features » (Yeo 2008, 119). La notion de prototype a d'abord été développée en psychologie cognitive par Rosch (1975) et Rosch et Mervis (1975). Par exemple, le moineau est l'oiseau prototypique dans les régions d'Europe de l'ouest et d'Amérique du nord. Pour connaître l'historique de la notion, voir Fortis (2010); pour une définition plus fine et des distinctions avec des termes proches en linguistique, voir Poitou (2001).

Flaux et Van de Velde (2000) évoquent le critère lexico-sémantique de l'appartenance au prototype ou le degré d'éloignement par rapport à celui-ci. Ce trait, qu'elles ne retiennent pas dans leur étude, est présenté initialement dans la sémantique du prototype développée par Kleiber (1990). Pour identifier le prototype, les auteures précisent qu'il faudrait mener une enquête sur les représentations psychiques, ce qui ne correspond pas à notre méthodologie. Ces représentations pourraient être disponibles, éventuellement mais pas systématiquement, dans *le Larousse* (2018) ou un autre dictionnaire encyclopédique. Puisque cette méthode n'est pas systématique, nous ne retenons pas ce trait (appartenance ou degré d'éloignement au prototype). Mais la théorie du prototype dont plusieurs travaux en sciences de l'information ont montré l'importance dans la recherche d'information (voir section suivante 3.5.1.6) a attiré notre attention sur l'importance du statut dénominatif des mots et expressions nominales. Il s'agit du prochain critère envisagé pour notre étude.

3.5.1.6. L'opposition dénomination/désignation

La dénomination est la qualité d'un terme d'évoquer non seulement son concept, son sens, mais aussi le fait que ce terme est un signe linguistique autonome (exemple de dénomination : « Analyse des besoins »). Par opposition, la désignation est compositionnelle[26] et dépend du contexte dans lequel elle est énoncée (exemple de désignation : « Analyse préalable à la classification »). L'opposition dénomination/désignation réside dans la reconnaissance de l'expression linguistique en tant qu'unité lexicale au sens codé dans la langue – la dénomination – ou simple agrégat de mots dont le sens est compositionnel, c'est-à-dire déductible par la somme des sens de chaque mot – la désignation.

La notion de dénomination a été relativement peu étudiée en linguistique (Petit 2012a, 4), elle est même marginale dans les recherches en linguistique (Mejri 2000, 609). Sans reprendre tout le panorama historique que Petit (2009) fait de cette notion, notons qu'elle a deux origines possibles : la philosophie du langage et la logique (Mill 1988; Kripke 1972; Strawson 1973, 1974; Geach 1962 : cités par Petit 2012b, 27[27]). Elle est traitée différemment dans la linguistique anglo-saxonne où la distinction entre désignation et dénomination n'apparaît pas aussi clairement. En France, elle est abordée depuis la fin des années 1970 et a été rattachée à la sémantique lexicale dans les années 1980. Nous pouvons considérer que « la dénomination est une fonction centrale du lexique et qu'à ce titre, la notion de dénomination appartient de plein droit à l'appareil conceptuel de la sémantique lexicale » (Petit 2012b, 28).

> [...] si la dénomination ressortit au codage, à l'arbitraire, nécessite un apprentissage et une mémorisation, la désignation relève, pour sa part, de la syntagmatique libre et de l'interprétation compositionnelle, en dehors de toute rigidité. (Petit 2012a, 4)

---

[26] Un sens compositionnel est un sens dont le tout est le résultat de l'addition des parties. Par exemple, « ruban rouge » désigne un ruban qui est rouge alors que « cordon bleu » désigne selon le *Grand Robert* (2017) une « cuisinière très habile ».

[27] Les références données par l'auteur sont les suivantes : Geach, P. T. (1962). *Reference and Generality*. Ithaca/London : Cornell University Press; Kripke, S. (1972). « Naming and necessity ». Dans D. Davidson & G. Harman (Dir.). *Semantics of Natural Language.* Dordrecht : Reidel, 253-355; Mill, J. S. (1988). *Système de logique*. Bruxelles : Mardaga; Strawson, P.-F. (1973). *Les individus.* Paris : Le Seuil.

Les unités monolexicales (ou mots simples) sont généralement des dénominations (Mejri 2000, 612). La question délicate est de savoir distinguer dans une unité polylexicale (ou ensemble de plusieurs mots formant une seule unité lexicale) une dénomination d'une désignation. Cette distinction nous apparaît utile pour l'indexation parce que la dénomination a reçu un acte de baptême antérieur à la constitution de l'index, ce qui permet à l'usager de comprendre son sens non pas compositionnellement mais pour ce qu'elle veut dire réellement dans un domaine.

Georges Kleiber (p. ex. 1984, 1990, 1996, 2003, 2012) est probablement le linguiste actuel qui a le plus travaillé sur la dénomination. Il a introduit la notion de dénomination en sémantique lexicale en France. Par sa thèse (1981), il a montré que la dénomination avait sa place en linguistique. Dans son article « Dénomination et relations dénominatives » (1984), il est le premier à définir les contours de la notion en linguistique et plus précisément en sémantique lexicale. Ses travaux ultérieurs s'appuient sur cette notion qu'il continue à délimiter. Il reprend lui aussi les travaux de Rosch en psychologie cognitive et applique les notions de prototype (présentée dans la section 3.5.1.1.5) et de niveau de base dans une hiérarchie lexicale aux unités lexicales assez stables pour cela, autrement dit, les dénominations. Le niveau de base est un concept tiré de la théorie cognitive développée par Rosch à la fin des années 1970 (Rosch 1975; Rosch et Mervis 1975), selon laquelle il existe une hiérarchie cognitive où le niveau de base se place entre le niveau superordonné et le niveau subordonné. Par exemple, le terme *chat* qui appartient au niveau de base a pour superordonnés les termes *félin* et *animal* et pour subordonné le terme *chat siamois*. Plus récemment, Gérard Petit, avec une approche toute lexicale (2001a et b), a rassemblé les différents éléments portant sur la notion de dénomination dans un ouvrage volumineux intitulé « La dénomination : approches lexicologique et terminologique » (2009). Il continue d'approfondir la notion et d'en délimiter les applications en linguistique et en terminologie (2012a et b). Les travaux de Kleiber et de Petit sont à notre sens particulièrement intéressants

pour l'évaluation du statut dénominatif des expressions linguistiques dans un corpus comme le nôtre.

La dénomination installe un lien référentiel stable, intersubjectivement partagé, appris et contractuel[28], qui permet à tout locuteur de comprendre la dénomination hors contexte ou dans un contexte de communication aussi épuré qu'un index. Cette opposition définit deux ensembles de mots au comportement différent linguistiquement. Dans un index, cette caractéristique sémantico-référentielle a des répercussions sur la lecture de l'index et son utilisation par l'usager. L'indexeur devrait donc la prendre en compte. Dans un travail d'exploration, cette opposition a été appliquée à des termes figurant dans des index de publications de la Collection des sciences de l'information des Presses de l'Université du Québec, par Guitard et Da Sylva (2014; recherche en cours).

3.5.1.7. Le nom commun, le nom propre et les entités nommées

En règle générale, le nom propre se distingue du nom commun en français par la présence d'une majuscule (Kleiber 1996). Les noms propres désignent des entités uniques (principalement être physique, famille, personne morale, lieu géographique). Les noms communs, par opposition désignent des classes d'entités. Le mot *chien* peut s'appliquer à tous les canins (y compris les chiwawas), alors que le nom propre *Rantanplan* désigne un seul chien qui appartient à la réalité (littéraire).

Précisons ce qu'est une entité nommée. On l'associe souvent aux noms propres. Mais « la notion d'entité nommée représente une catégorisation bien plus large que celle des noms propres (...), puisqu'elle inclut des expressions temporelles ou numériques, des maladies ou

[28] *Contractuel* signifie ici que la dénomination relève du contrat tacite entre les locuteurs de la langue, comme l'a souligné Fernand de Saussure dans son cours d'introduction à la linguistique (1916), après un « acte de baptême », comme l'évoque Kleiber (notamment 1990).

des drogues. » (Daille et Morin 2000, cités par Ehrmann 2008, 257[29]). Selon Ehrmann (2008, 169), une entité nommée est une expression linguistique monoréférentielle (*i. e.* qui ne renvoie qu'à un seul élément dans le monde concret) et autonome au sein d'un corpus (*i. e.* le lecteur sait de quoi il s'agit au sein du contexte défini par le corpus et le milieu dont il est tiré). Le contexte est couvert par plusieurs des éléments suivants : l'institution créatrice et conservatrice des documents, le pays des créateurs, la langue des documents, l'époque de création, de description, de recherche des documents). Par exemple, le nom commun *pont* renvoie à la classe des ponts (y compris les ponts couverts, les ponts tunnels et autres). Si on associe un nom propre tel que *Jacques Cartier* (nom propre d'une personne) à ce nom commun, on obtient une entité nommée : le *pont Jacques-Cartier*. Cette entité nommée est unique dans le contexte auquel elle appartient. Nous nous appuyons sur la reconnaissance (humaine) des entités nommées dans notre recherche mais nous ne les analysons pas, tout comme les noms propres, puisqu'elles désignent des segments de la réalité tout à fait identifiables. En outre, les entités nommées sont peu à peu prises en charge par la linguistique computationnelle et le traitement automatique de la langue (Nouvel, Ehrman et Rosset 2016), comme le sont les noms propres. Nous nous concentrons sur les noms communs qui couvrent des classes génériques et qui, de ce fait, requièrent une interprétation. C'est le flou sémantique généré par les noms communs employés par les usagers (p. ex. *pont*) qui permet aux archivistes de leur proposer des documents avec des entités nommées ou des noms propres (p. ex. *pont Jacques-Cartier*), comme exemplaires d'une classe. Il s'agit de la « traduction » effectuée par l'archiviste de référence, des accès thématiques en accès de provenance.

3.5.2. Aspect sémantique des termes en sciences de l'information

En sciences de l'information (SI), nous nous sommes reportée à deux types de littérature pour cerner les notions relatives à l'aspect sémantique des termes : la littérature

[29] La référence donnée par l'auteure est la suivante : Daille, B. et Morin, E. (2000). « Reconnaissance automatique des noms propres de la langue écrite : Les récentes réalisations ». *Traitement Automatique des Langues* 41(3):601–621.

scientifique (les travaux de chercheurs) et la littérature normative (les normes nationales et internationales).

3.5.2.1. Littérature scientifique

Diverses théories linguistiques ont été utilisées en sciences de l'information pour étudier le sens. Lee et Neal (2010) ont développé un modèle permettant de vérifier le niveau sémantique souhaité par les usagers pour leurs étiquettes apposées à des photographies personnelles. On peut considérer ces dernières comme des documents d'archives, bien qu'elles n'aient pas été envisagées comme telles par les chercheurs de cette étude. Le niveau souhaité par les usagers est bel et bien le niveau de base. Chung et Yoon (2009) ont étudié le lien entre les questions de recherche et les étiquettes dans la recherche d'images; ils confirment que le niveau de base est à préférer dans un système de repérage de l'information. Green (2006) a généré automatiquement des termes du niveau de base au sein du réseau lexical anglophone WordNet développé par Princeton[30] et essaye de créer des passerelles entre des systèmes classificatoires à partir des termes du niveau de base. Les auteurs ont remarqué que certains domaines de la connaissance ne se prêtent pas à leur analyse, mais dans l'ensemble le test est concluant quand il s'agit seulement de noms concrets (voir section 3.5.1.2 pour l'explication de cette notion). Rorissa et Iyer (2008) ont étudié les noms donnés aux catégories de groupes d'images dans deux études qui ont impliqué 100 images et 105 participants. Ils ont montré que les usagers novices qui indexent des groupes d'images ont tendance à utiliser des termes superordonnés, contrairement au niveau de la pièce qui est du niveau de base. Le niveau de base a une saillance cognitive. Dans notre étude du vocabulaire pour l'accès thématique des archives patrimoniales (VATAP), nous en avons tenu compte, dans notre observation de l'emploi de tel ou tel terme dans la formulation des sujets dans les courriels d'usagers à la référence.

Les index en tant que documents à part entière ainsi que les termes d'indexation qui les constituent ont fait l'objet de quelques recherches linguistiques. Lepage et de Schaetzen

[30] Disponible à https://wordnet.princeton.edu/ ; voir notamment Fellbaum 1998.

(1997) ont étudié les index pour leur valeur terminologique. Ils ont identifié plusieurs caractéristiques des index qui formaient leur corpus. Les auteurs insistent sur le fait qu'ils sont les premiers à mener une étude sur les index et leurs termes. Ils adoptent une perspective terminologique et de traitement automatique de la langue (extraction de vocabulaire spécialisé à partir des termes d'index). Ils n'analysent pas les particularités des termes d'indexation comme leurs propriétés lexico-sémantiques. Da Sylva (2009b) étudie les propriétés morphologiques, syntaxiques et sémantiques des termes d'indexation présents dans des index de fin de livre afin de mettre au jour une classe de vocabulaire entre le lexique général et le lexique savant qu'elle appelle le *vocabulaire savant de base* (VSB), grâce à des outils de traitement automatique de la langue. Elle étudie notamment le lien sémantique entre les vedettes principales et les vedettes secondaires à partir de la Sémantique des cadres (Fillmore 1976). Ces travaux démontrent l'intérêt d'étudier les termes d'indexation sous un angle linguistique. Cependant, ils ne se rapportent pas aux index de documents d'archives qui diffèrent par la nature des documents indexés (Da Sylva 2015). En outre, ils n'apportent pas de résultats clairs sur les phénomènes sémantiques en jeu dans la chaîne communicationnelle de la référence archivistique[31] (voir Figure 3).

3.5.2.2. Littérature normative

Outre les études menées par des chercheurs, nous nous sommes reportée à plusieurs normes nationales et internationales sur le sens et les propriétés linguistiques en SI. La norme Z 47-100 (*Règles d'établissement des thésaurus monolingues* AFNOR, 1981) indique qu'un descripteur doit couvrir une notion et doit s'exprimer en un minimum de mots, généralement

[31] Ce jeu de relations sémantiques entre les documents (courriel d'usager à la référence, notice descriptive, index) au cœur duquel l'index joue un rôle prépondérant s'étudie en lien avec la notion d'intertextualité que nous n'abordons pas dans la thèse. On attribue à Kristeva l'invention même du mot *intertextualité*, phénomène qu'elle décrit ainsi : « tout texte se construit comme une mosaïque de citations, tout texte est absorption et transformation d'un autre texte » (Kristeva 1969, 85). L'intertextualité est en jeu dans un réseau de textes liés entre eux par des éléments de sens communs (Allen 2000). L'intertextualité est une des caractéristiques de l'indexation (Bertrand-Gastaldy *et al*. 1994).

un seul; cet élément est à mettre en lien avec le statut de dénomination puisque les mots simples sont tous des dénominations (voir section 3.5.1.6). Elle recommande aussi des éléments quant à la forme des termes. La norme ISO 2788 (1986) : *Principes directeurs pour l'établissement et le développement de thésaurus monolingues* décrit la forme des termes également. Cette norme est remplacée par la norme ISO 25964-1 (2011) *Thésaurus et interopérabilité avec d'autres vocabulaires. Partie 1 : Thésaurus pour la recherche d'information*[32]. Suite à sa consultation, nous constatons qu'elle ajoute peu d'éléments différents de la synthèse des éléments issus des autres normes. Principalement, nous relevons la mention du bas de casse pour les entrées qui ne requièrent pas de majuscules, le choix de formes familières ou argotiques comme termes préférés si le thésaurus est destiné à une communauté pour laquelle elles sont l'usage principal ainsi que la prise en compte d'un usage du singulier pour les entrées par exemple en français, afin de les rapprocher de l'usage du singulier dans les dictionnaires. La norme ANSI/NISO Z 39.19 (2005) : *Guidelines for the construction, format and management of monolingual thesauri* complète les indications que la norme ISO donne sur la forme grammaticale des termes. La norme *BS 8723*:2005 : *Structured vocabularies for information retrieval – Guide. Part 2* reprend de nombreux éléments de la norme ANSI. Seul le critère du nombre est différent, c'est pourquoi nous le reproduisons ici : *(entités dénombrables ou non) « how much ? » ou « how many ? »*. La mention d'une différence selon la langue, le français ou l'allemand utilisant la forme « singulier », contrairement à l'anglais, est la seule que nous ayons trouvée (*BS 8723* :2005).

Le Tableau I Critères linguistiques dans les principales normes documentaires reprend les éléments cités dans le texte récapitulatif de Chichereau *et al*. (2007) et ceux d'une analyse préalable du Groupe ADBS Web sémantique en 2005-2007, consultée sur le site de l'ADBS[33]. Précisons que la formulation des éléments est celle qu'ont donnée les auteurs.

---

[32] La norme ISO 25964-1 : 2011 annule et remplace les normes ISO 2788 : 1986 et ISO 5964 : 1985.

[33] Le document a été consulté à l'adresse suivante : http://www.adbs.fr/servlet/com.univ.collaboratif.utils.LectureFichiergw?ID_FICHE=376&OBJET=0017&ID_FICHIER=1380

Tableau I.Critères linguistiques dans les principales normes documentaires

Norme / Élément	Z 47-100 (1981)	ISO 2788 (1986)	ANSI/NISO Z 39.19 (2005)
Noms	forme substantive à utiliser	Noms et locutions nominales (avec adjectif ou préposition)	substantifs, phrases nominales (avec adjectif ou préposition dans les cas où il n'est pas possible de faire autrement)
Adjectifs	exceptionnellement autorisés	à éviter mais peuvent être utiles pour faire des renvois d'un adjectif vers les termes composés qui le contiennent	en descripteurs seuls pour limiter le nombre de termes composés, mais à usage limité car attention aux fausses combinaisons à la recherche, utilisation des adjectifs pour limiter les renvois vers les termes qui les contiennent
Adverbes	pas autorisés seuls	pas autorisés seuls, expressions commençant par un adverbe	ne pas utiliser seuls
Verbes	pas autorisés, voir la forme substantive	pas autorisés, voir la forme substantive	ne pas les utiliser seuls
Article	*Néant*	*Néant*	à éviter au début d'un terme mais à garder si partie intégrante d'un nom propre (El Nino)
Nombre	singulier	lié à la notion d'index pré-coordonné (le pluriel pouvant apporter du sens), aux facteurs culturels (pays anglophones ou non), à la notion d'entités concrètes et abstraites, dénombrables (how many ?) ou non (how much ?). Remarque : l'exemple « bois » n'est pas bon en français car c'est une traduction de l'exemple anglais : woods et wood).	(entités dénombrables ou non : idem à la norme ISO). Selon le principe de « how many ?» et « how much ? » Mais voir les exemples des exceptions du pluriel, en médecine l'anatomie : « lung » et « ear » alors que dans l'ISO, ces deux termes sont au pluriel (poumons et oreilles) et les objets de musée sont au singulier.
Abréviations	à éviter, synonymie avec la forme développée	à éviter, synonymie avec la forme développée	
Sigles ou acronymes	à éviter, synonymie avec la forme développée	à éviter, synonymie avec la forme développée	oui si connue universellement
Orthographe	si différentes formes les indiquer en synonymes		Orthographe courante, de référence
Termes composés	de préférence sans mots de liaison		
Polysémie et homographes	lever l'ambiguïté par un qualificatif sans parenthèses	qualificatif entre parenthèses	
Traductions	faire des synonymies		utilisés si usage courant
Termes d'emprunt	faire des synonymies		utilisés si usage courant
Formes familières ou argotiques	à éviter		
Néologisme, jargon			oui si pas d'alternative largement acceptée

Norme / Élément	Z 47-100 (1981)	ISO 2788 (1986)	ANSI/NISO Z 39.19 (2005)
Noms scientifiques et vernaculaires			selon les utilisateurs
Marques déposées	à éviter, utiliser plutôt les noms communs		de préférence nom commun, si c'est une marque il faut l'indiquer dans le libellé du terme (contraintes juridiques)
Noms commerciaux	à éviter, utiliser plutôt les noms communs		
Noms propres (géographiques et de personnes)			importance de leur contrôle, soit dans le thésaurus (noms de lieux), soit dans une liste d'autorité (personnes), établir des synonymes entre appellation officielle et commune
Casse	Majuscules pour les descripteurs		
Étendue	limitation du nombre de caractères pour les descripteurs		
Signes diacritiques	À exclure		
Ponctuation	à éviter		
Caractères spéciaux	à éviter		
Note d'application	importance des notes pour préciser un terme (sens ou usage)		
Critères de sélection de la forme préférentielle			selon l'usage, littérature courante, référentiels, autres langages d'indexation, etc.

Les normes qui s'appliquent initialement à l'anglais traitent en priorité des spécificités de l'anglais, telles que la différence dénombrables et indénombrables. Pourtant, ce trait lexico-sémantique, plus volontiers appelé en français la massivité ou l'opposition massif/comptable (voir section 3.5.1.1) existe bel et bien en français mais ne s'applique pas systématiquement aux mêmes mots qu'en anglais. Ceci explique la remarque des auteurs en ce qui concerne le nombre dans la norme ANSI/NISO Z 39.19 en français à propos de l'exemple traduit mot à mot qui ne fonctionne pas réellement en français pour illustrer la différence massif/comptable (*bois/woods*).

Les normes ne parlent que de peu de propriétés lexico-sémantiques : l'opposition massif/comptable, l'opposition concret/abstrait (en rapport avec le nombre). Les autres propriétés sont d'ordre morphosémantique (i.e. qui est relatif à la forme et au sens intimement liés) ou concernent carrément d'autres considérations que la description du sens des mots (par

exemple, la note d'application). Pour le français, les propriétés lexico-sémantiques ne sont ainsi pas toutes relevées par les normes. C'est pourquoi nous nous sommes reportée à la littérature linguistique pour trouver d'autres propriétés linguistiques utiles à notre analyse (voir section 3.5.1).

3.5.3. Les relations sémantiques

Cette section présente ce que sont les relations sémantiques et comment elles sont abordées en sciences de l'information et en linguistique.

Hjørland explique que « les relations entre concepts indiqués dans un thésaurus sont des relations sémantiques » (2007, 368, n. t.), mais il ne les définit pas explicitement. « Les relations sémantiques sont des associations porteuses de sens entre deux ou plusieurs concepts, entités ou ensemble d'entités. Elles peuvent être vues comme des liens orientés entre les concepts ou entités qui participent de la relation » (Khoo et Na 2006, 159, n. t.) Dans notre recherche, une relation sémantique est un lien de sens entre deux expressions linguistiques porteuses de sujet (thémanymes), comme bateau et navire, liés par une relation sémantique de spécificité. Nous étudions des paires de thémanymes orientées en fonction de la chaîne communicationnelle de la référence archivistique (voir Figure 3). Ce lien se perçoit aisément pour un locuteur natif de la langue, il est évident, même s'il est généralement difficile à expliciter (Mortureux 1993; Kleiber 2009).

Certaines relations sémantiques représentent un lien sémantique stable, connu et entériné par des ouvrages de référence tels que le manuel de lexicologie de Polguère (2016), celui de Mortureux (2008) ou celui de Lehmann et Martin-Berthet (2013). Polguère (2016, 119-128 et 147-158) présente les six relations lexicales suivantes en les qualifiant de fondamentales : hyperonymie et hyponymie (relation de spécificité), synonymie, antonymie, conversivité, homonymie, polysémie. Ainsi, *chat* est l'hyponyme ou le spécifique d'*animal* et inversement *animal* est l'hyperonyme ou le générique de *chat*; *automobile* et *voiture* ont un

sens similaire (et sont généralement interchangeables[34] dans la plupart des contextes) et sont donc synonymes; *gentil* et *méchant* ont des sens contraires, ils sont antonymes; *voler* (en avion) et *voler* (de l'argent) sont homonymes et même homographes; *tasse* a au moins deux sens, la pièce de vaisselle et l'unité de mesure, ce mot est polysémique. Deux termes sont conversifs s'ils sont des prédicats sémantiques ou quasi-prédicats et qu'au moins deux de leurs actants sont inversés (Polguère 2008, 154). L'exemple *X emploie Y* et *Y travaille pour X* montre que *employer* et *travailler (pour)* sont des conversifs l'un de l'autre. Bien que Polguère place la conversivité dans les relations fondamentales, il s'agit d'une relation moins courante, absente de Lehmann et Martin-Berthet (2013) et Mortureux (2008). Par contre, les cinq autres relations sont communes aux trois ouvrages lexicologiques : hyperonymie/hyponymie, synonymie, antonymie, homonymie, polysémie. Nous avons évalué au cours de notre analyse la pertinence de cette relation – la conversivité – dans le vocabulaire employé pour l'accès aux archives patrimoniales et nous ne l'avons pas retenue en tant que telle mais plutôt en tant que relation référentielle, nous y reviendrons (voir section 5.1.1).

Plusieurs auteurs des sciences de l'information se sont intéressés aux relations sémantiques. Clarke (2001) traite des relations sémantiques en jeu dans les thésaurus, qu'elle définit comme un outil pour contrôler le vocabulaire utilisé tant à l'indexation qu'à la recherche sous la forme d'une liste de concepts que des termes permettent d'identifier et de rendre saillants (2001, 38). Clarke rappelle que les relations sémantiques sont classiquement[35] regroupées en trois types distincts et mutuellement exclusifs : relations d'équivalence, relations hiérarchiques et relations associatives. Un thésaurus est une structure de connaissance incluant des termes et des relations entre ces termes (2001, 37, n. t.). Dans un

[34] Kleiber (2009) montre que l'interchangeabilité entre deux unités lexicales (une seule et même acception par unité lexicale) n'est pas un critère définitoire de la synonymie absolue. Mais pour notre étude, ce critère moins rigoureux est suffisant.

[35] Clarke précise que ces trois types de relations sont connus depuis au moins trois décennies; elle renvoie notamment à Vickery (1971) et Lancaster (1986). Pour une référence plus récente, nous ajoutons Hudon (2009).

thésaurus unilingue[36], la relation d'équivalence concerne en théorie les termes synonymes ou quasi-synonymes. Mais dans la pratique, elle couvre de nombreux phénomènes observés par Clarke (2001, 39-40). Précisant qu'il ne s'agit pas d'une liste exhaustive, elle relève treize phénomènes qu'elle exemplifie en anglais :

1. Nom commun/nom scientifique : *mad cow disease/bovine spongiform encephalopathy*
2. Nom générique/marque de commerce : *adhesive plaster/band-aid/elastoplast*
3. Nom standard/argot (variation diastratique, nous ajoutons) : *supplementary earnings/perks*
4. Forme au long/abréviations, acronymes : *auto-immune deficiency syndrome/AIDS*
5. Variantes lexicales (variation orthographique, nous ajoutons) : *color/colour; edema/oedema; databases/data-bases/data bases*
6. Entrées complexes inversées : *electric cables/cables, electric*
7. Termes de différentes cultures partageant un langage commun : *elevator/lifts; splinter/sliver/skelf*
8. Termes d'origines linguistiques différentes : *buying/purchasing; thermal resistance/heat resistance*
9. Termes en compétition pour des concepts ou des technologies émergents (néologie, nous ajoutons) : *phasmid/phagemid; lap-top computers/notebook computers*
10. Termes actuels/désuets (variation diachronique, nous ajoutons) : *capacitors/(electric) condensers*
11. Pluriels irréguliers : *mouse/mice*
12. Quasi-synonymes : *perfume/eau de cologne; smoothness/roughness*
13. Concepts spécifiques englobés dans un concept plus large : *flavour/bitterness/sweetness; handling machinery/cranes/extractors* (Clarke 2001, 40)

Il est à remarquer que certains points paraissent évidents pour l'anglais (p. ex. 8, 11). Il est implicite que les quasi-synonymes intègrent les antonymes; en effet, le couple *douceur/rugosité* illustre l'antonymie. L'auteure traite la variation de langue en fonction de l'aire géographique où elle est parlée (variation diatopique). Les usagers posant des questions aux archivistes au Québec peuvent provenir de diverses régions du monde et s'exprimer dans

[36] Notre recherche porte uniquement sur le français et nous ne prenons pas en compte la réalité différente d'un contexte multilingue.

un français légèrement différent. Clarke traite également d'équivalence partielle : il s'agit de la couverture d'un concept par deux entrées du thésaurus. Étant donné qu'il s'agit d'une stratégie d'indexation ou de recherche propre aux thésaurus, nous n'en rendons pas plus compte ici.

La relation hiérarchique est assignée à une paire de termes dont l'un a une portée qui englobe totalement celle d'un autre, plus spécifique (Clarke 2001, 42). Par exemple, la portée de *fleur* englobe celle de *rose*. On reconnaît trois types de relations hiérarchiques : la relation générique/spécifique (hyperonymie/hyponymie), la relation partie-tout (méronymie/holonymie) et la relation instance/classe. Par exemple, *fleur* est l'hyperonyme de *rose*, *manche* (n. m.) est une partie du tout *raquette*. Dans notre corpus, *navire échoué en 1912* est une instance de la classe *bateau*. Clarke mentionne que la plupart des concepteurs de thésaurus estiment qu'il n'est pas nécessaire d'expliciter cette typologie. Par ailleurs, un terme plus générique peut l'être de manière directe ou indirecte : *chat* est le spécifique de *félin*, mais aussi d'*animal* ou d'*être vivant* ou encore d'*entité* qui représente un niveau très élevé dans la hiérarchie des concepts. Le niveau maximal entre un terme et un terme qui lui est générique (*broader term*, BT) est parfois noté dans les thésaurus en tant que « terme racine » (*top term*, n. t., noté TT plutôt que BT). Certaines relations hiérarchiques sont invariablement valides et d'autres sont contextuelles ou bien dépendent d'un point de vue (Svenonius 1997; Clarke 2001). Nous sommes restée attentive à cette distinction et nous y revenons dans la présentation des résultats et dans la discussion. Étant donné que nous ne travaillons pas dans la perspective de créer un thésaurus mais de rendre compte des liens sémantiques en jeu dans la référence archivistique, nous ne nous attarderons pas sur les hiérarchies complexes.

La méronymie (relation partie-tout) repose sur une connaissance encyclopédique du mot; à ce titre elle aurait pu être incluse dans les relations associatives ou référentielles. Cependant, étant donné qu'elle est fréquente et utilisée dans les thésaurus, elle est généralement incluse dans les relations de hiérarchie (Hudon 2013, 57-59) Il en va de même pour la relation classe/instance, qui s'apparente à une relation hyperonyme/hyponyme au niveau le plus bas du lexique, à savoir le nom propre (Hudon 2013, 59).

Le troisième des types de relations sémantiques classiques concerne les relations associatives (appelées *associative* ou parfois *affinitive relations* en anglais). Il s'agit de relations qui ne sont ni hiérarchiques ni d'équivalence, selon Clarke (2001, 46). Ces relations

sont censées rendre compte d'une association intellectuelle forte entre les deux concepts évoqués (*ISO 2788* :1986 et *ANSI/NISO Z 39.19* :1993 cités par Clarke 2001, 46[37]). Dans un thésaurus, elles ont une fonction d'aide à l'usager, que ce soit dans un contexte d'indexation ou de recherche, pour préciser sa pensée ou l'élargir. Le but est de lui permettre d'envisager des termes, voire des concepts auxquels il n'a pas pensé de lui-même de prime abord. Cette fonction prend place dans un contexte informationnel; elle éloigne ce type de relations d'une approche linguistique/sémantique autonome (en langue), stricte. Elle a recours aux connaissances encyclopédiques partagées du concepteur de l'outil comme de ses usagers, ce qui rejoint bien la perspective de notre recherche qui a pour cible la profession archivistique.

Clarke souligne la zone grise relative à la distinction entre ces deux types de relations – relations hiérarchiques et associatives – et que les discussions peuvent être longues à ce sujet (2001, 46). L'auteure a recensé dans six études des années 1960 à 2000 le nombre de relations associatives dans un ou plusieurs thésaurus; ce nombre va d'une dizaine à 120. « There is no consistent pattern of relations in use » (Willetts 1975, cité par Clarke 2001, 47[38]). Clarke a tenté de dégager une certaine régularité dans l'attribution des relations associatives : « Attempts to be consistent and comprehensive in applying semantic rules can lead to lists of RT[39]s so long that they actually can impede users in finding the related terms they really need help with » (Clarke 2001, 47).

---

[37] Les références données par l'auteure sont les suivantes : International Organization for Standardization. (1986). *Documentation_Guidelines for the Establishment and Development of Monolingual Thesauri* (2nd ed.). [Geneva:] ISO. (ISO 2788-1986(e)) et National Information Standards Organization. (1994). *Guidelines for the Construction, Format, and Management of Monolingual Thesauri.* Bethesda (MD): NISO Press. (ANSI/NISO Z 39.19-1993).

[38] La référence donnée par l'auteure est la suivante : Willetts, M. (1975). « An investigation of the nature of the relation between terms in thesauri ». *Journal of Documentation* (31): 158-184.

[39] RTs pour *Related Terms*, soit *termes associés*.

Hjørland (2007), dans un chapitre sur la sémantique et l'organisation des connaissances, relève dix-sept relations sémantiques dans la littérature relative aux relations associatives :

1. **Active relation**: A semantic relation between two concepts, one of which expresses the performance of an operation or process affecting the other. The inverse of the passive relation.
2. **Antonymy**: A semantic relation in which A is the opposite of B (e.g., cold is the opposite of hot).
3. **Associative relation**: A semantic relation defined psychologically as the mental association of concepts (i.e., A is mentally associated with B by somebody). Often, associative relations are simply unspecified relations. In thesauri, antonyms are not usually specified but may be listed along with terms representing other kinds of relations under "associative relations."
4. **Causal relation**: A semantic relation in which A is the cause of B (e.g., a lack of vitamin C causes scurvy).
5. **Homonymy**: A semantic relation in which two concepts, A and B, are expressed by the same symbol (e.g., both a financial institution and the edge of a river are expressed by the word "bank"; i.e., the word has two senses).
6. **Hyponymous relations (hyponym-hyperonym)**: Relations in which A is a kind of B; A is subordinate to B; A is narrower than B; B is broader than A. Also known as generic relation, genus-species relation, or hierarchical subordinate relation.
7. **Is-a relation**: A semantic relation between a general concept and individual instances of that concept; that is, A is an example, or instance, of B (e.g., Copenhagen is an instance of the general concept "capital").
8. **Locative relation**: A relation in which a concept indicates a location of a thing designated by another concept: that is, A is located in B (e.g., minorities in Denmark).
9. **Paradigmatic relation**: As defined by Wellisch (2000, p. 501, "a semantic relation between two concepts, that is considered to be either fixed by nature, self-evident, or established by convention. Examples: mother/child; fat/obesity; a state/its capital city.')
10. **Partitive (i.e, part-whole) relation (meronymy)**: a relationship between the whole and its parts; that is, A is part of B. A meronym is the name of a constituent part of, the substance of, or a member of something. Meronymy is the opposite of holonymy (i.e., B has A as part of itself).
11. **Passive relation**: A semantic relation between two concepts, one of which is affected by, or subjected to, an operation or process expressed by the other. The inverse of the active relation.
12. **Polysemy**: A mode of semantic relation in which a word has several subsenses that are related with one another (i.e., concepts Al, A2, and A3 are all expressed by the word "A"). Such a word is termed "polysemous" or "polysemantic."
13. **Possessive relation**: a semantic relation between a possessor and what is possessed (i.e., A belongs to B; B possesses A).

14. **Related term**: A term that is semantically related to another term. In thesauri, related terms are often coded RT and used for kinds of semantic relations other than synonymy (USE, UF), homonymy (separated by parenthetical qualifier), and generic relations and/or partitive relations (BT, NT). Related terms may, for example, express antagonistic relations, active/passive relations, causal relations, locative relations, or paradigmatic relations.
15. **Synonymy**: A semantic relation in which A denotes the same as B; A is equivalent with B.
16. **Temporal relation**: A semantic relation in which a concept indicates a time or period of an event designated by another concept (e.g., Second World War, 1939-1945).
17. **Troponymy**: According to WordNet 2.1 (2005), "the semantic relation of being a manner of [doing] something." (Hjørland 2007, 404-405, notre emphase)

Le classement alphabétique des éléments de la liste nuit à sa clarté ; certaines oppositions telles que relation active / relation passive ou encore homonymie / polysémie sont difficiles à cerner. Cette liste contient des relations sémantiques mais également des termes (*related terms*, termes reliés), nous y reviendrons. Nous ne passerons pas en revue chacune des relations mentionnées. Nous remarquons que sont présentes les relations sémantiques de base que sont l'hyperonymie-hyponymie, la classe-instance, la méronymie-holonymie, la synonymie ainsi que la polysémie. Certaines relations n'appartiennent pas au même niveau que les autres; nous nous expliquons : nous voyons l'appellation « relation associative » qui peut inclure plusieurs autres relations présentées ici (l'antonymie est citée) et qui est « souvent non spécifiée » mais qui « vont de soi ». De même, la relation paradigmatique pourrait, nous semble-t-il, inclure d'autres relations; l'exemple donné *un état/sa capitale* pourrait être inclus dans la relation de lieu (*locative relation*).

Milstead (2001) étudie les relations sémantiques répertoriées dans les normes documentaires. Elle décrit essentiellement trois types de relations : les relations d'équivalence, hiérarchiques et associatives (Milstead 2001, 55-56). Ce sont classiquement des relations thésaurales. L'équivalence correspond principalement à la synonymie et la variation orthographique; la hiérarchie correspond à un changement de spécificité dans le vocabulaire. Si les deux premières sont bien définies dans la littérature, la troisième laisse place à l'interprétation dans de nombreux cas. Ce flou définitionnel nous a poussée à investigation étant donné que nous souhaitions identifier ce type de relations dans notre corpus de manière claire et solide. Milstead en cite douze suivies d'exemples en anglais. Il est intéressant de

noter que Milstead qualifie ces relations de *Relations entre des termes reliés*, dans une section intitulée *Relations associatives*. Nous nous demandons alors quelle différence opère Hjørland (2007) entre les deux éléments de sa liste *Relations associatives* et *Termes reliés*. Voici la liste des relations associatives de Milstead :

> Siblings with overlapping meanings: boats/ships
> Familial/derivational: children/parents
> Discipline and object of study: zoology/animals
> Operation/process and its agent/instrument: heating/furnaces
> Thing and its counter agent: plants/herbicides
> Action and its product: weaving/cloth
> Action and its target: harvesting/crops
> Concepts and their unique properties: perception/acuity
> Concepts related to their origins (ISO only): Dutch/Netherlands
> Concepts liked by causal dependence: bereavement/death
> Concept and its unit of measurement: length/meter(measure)
> Phrases in which the noun is not a true broader term: heart/artificial hearts. (Milstead 2001, 61)

La première relation (*Siblings with overlapping meanings*) semble correspondre à de la synonymie : le sens des mots se chevauchent. Quant à la deuxième relation (*Familial/derivational*), elle est assez large; elle serait à rapprocher d'une part de la relation paradigmatique de Hjørland (2007) en ce qui concerne les mots de la même famille et d'autre part des relations morpho-lexicales de Ménard (1989); nous le verrons dans le prochain paragraphe, en ce qui concerne la dérivation. À la manière des relations de temps et de lieu de Hjørland (2007), l'avant-dernière relation concerne les mesures et leur unité de mesure et sont très spécifiques : elles contraignent sémantiquement les éléments qu'elle relie. Finalement, la dernière relation de Milstead repose davantage sur un critère de forme que de sens : un mot de l'expression est repris, mais dans une expression où il change d'acception (*phrases in which the noun is not a true boarder term*).

Soergel (1974) propose une définition opérationnelle des relations associatives : « Concept A is related to concept B (has an associative relationship to concept B) if the following holds: an indexer or searcher weighing the use of A should be reminded of the existence of B (and there is no hierarchical relationship between A and B) ». Et immédiatement après, il ajoute : « The associative relationships complement the system of hierarchical relationships. » (1974, 107). Cette approche se base clairement sur l'usage des

relations associatives dans un thésaurus. Clarke et Lee rapportent cette définition qu'ils jugent suffisamment ouverte pour être adaptable à de nombreux contextes et en même temps trop large pour servir de guide (2015, n. t.). Quant à elle, Hudon décrit cette relation de la manière suivante : « La relation associative lie des concepts qui sont souvent associés mentalement par les spécialistes d'un domaine. Le lien associatif reste mal défini et très subjectif. » (Hudon 2013, 68) Ainsi, ce lien reste flou. Nous nous sommes alors reportée à des études linguistiques.

Dans son étude des textes scientifiques, Ménard (1989, 475-477) présente un tableau des relations lexico-sémantiques présentes dans son corpus :

1. ÉQUIVALENCE (même catégorie syntaxique)
1.1. Permutables
1.1.1. Anaphores et cataphores (pronoms, possessifs, déictiques)
1.1.2. Synonymes et coréférents lexicaux Ex. : Matière noire / masse manquante
1.2. Substituables (asymétriques)
1.2.1. Hyperonyme / hyponyme Ex. : bête / mouton ; étoile / soleil ; naines brunes
1.2.2. Base / base+ modificateur Ex. : salut / salutation ; camion / camionnette ; mille / millier, milliard
2. RELATIONS MORPHO-LEXICALES (avec changement de classème)
2.1. Dérivés (avec ou sans changement de catégorie syntaxique)
2.1.1. De même base Ex. : voir / visiblement ; lumière / lumineux, luminosité ; lait / voie lactée
2.1.2. De bases différentes (mais même sémème) Ex. : eau / hydraulique ; lactée / galaxie ; loin / télescope
2.2. Composés
2.2.1. Élément / lexie composée Ex. : dynamique / électro-dynamique ; année / année lumière
2.2.2. Élément / lexie complète Ex. : galaxie / année galactique
3. CONTRASTE
3.1. Binaires
3.1.1. Graduables Ex. : lentement / vivement ; général / restreint
3.1.2. Non graduables
3.1.2.1. En opposition privative Ex. : naître / mourir ; mouvement / inertie
3.1.2.2. En opposition équipollente Ex. : à pied / à cheval ; nocturne / diurne ; théorique / réel
3.1.3. Converses Ex. : question / réponse
3.1.4. En opposition directionnelle et antipodale Ex. : en arrière / en face
3.1.5. Tout / partie Ex. : atome / électron, noyau
3.2. X-naire (x>2)
3.2.1. De même niveau taxinomique Ex. : divan / chaise ; colline / plateau ; nébuleuse d'Andromède / voie lactée

3.2.2. Sériés Ex. : méthane / éthane / propane
3.2.3. En opposition cyclique et orthogonale Ex. : en arrière / à côté ; parallèle / perpendiculaire

4. COLLOCATIONS ET IMPLICTIONS DIVERSES

4.1. Prédicat / argument, régissant / subordonné

4.1.1. Collocation contraignante Ex. : germain / cousin ; nucléique / acide
4.1.2. Collocation à forte probabilité Ex. : broncher / cheval ; champ / (de) gravitation ; théorie / (de la) relativité

4.2. Implication par définition (certaines relations de présupposition lexicale) Ex. : recommencer / arrêt ; naseau / cheval ; estuaire / fleuve

4.3. Implication indirecte, « médiatisée » Ex. : savoir / journal ; caillou / parler (?) ; éclair / foudre (Ménard 1989, 475-477, emphase de l'auteur)

Cette grille est trop développée pour nos besoins et présente des phénomènes linguistiques fins qu'il n'est pas nécessaire d'identifier dans notre recherche à vocation archivistique. En outre, les relations sont identifiées dans le contexte particulier d'un texte. Les relations sont donc internes au texte et une chaîne de référence – c'est-à-dire le suivi du sujet – se construit au fur et à mesure de la lecture. Notre recherche, quant à elle, porte sur plusieurs textes ou segments de textes reliés par le sujet et entre lesquels un agent cognitif peut suivre ce sujet de l'un à l'autre. Nous retenons de Ménard (1989) la souplesse dans l'attribution des étiquettes de relations sémantiques. Ainsi, dans le contexte d'un texte en particulier, deux expressions seront coréférentielles (renvoient à la même entité) et sont donc qualifiées de synonymes. Le passage du plan sémantique au plan référentiel est intéressant, nous en reparlerons en discussion.

Stock (2010, 1957) présente sous forme de figure les relations sémantiques qu'il traite.

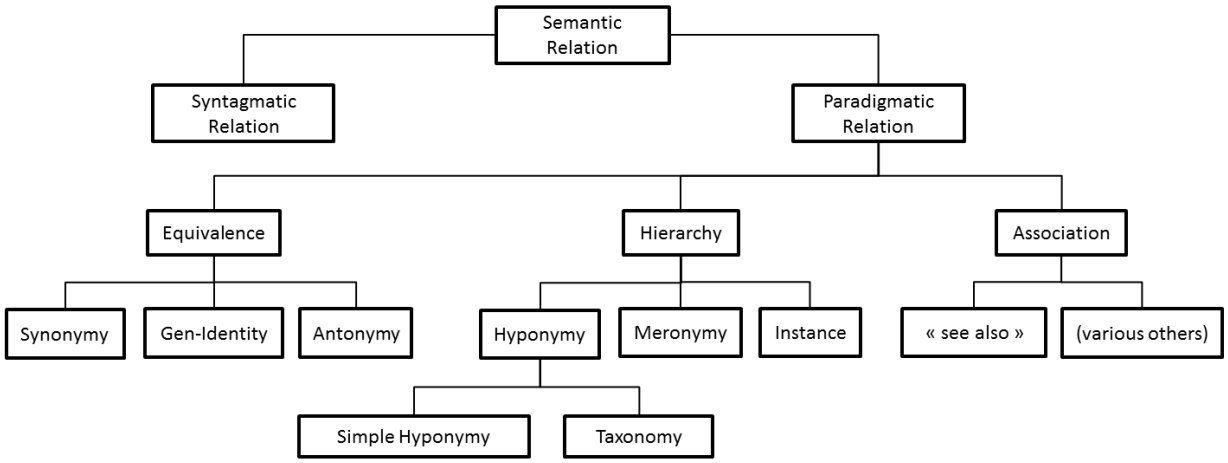

Figure 5. Relations sémantiques (Stock 2010, 1957)

Dans la Figure 5, nous remarquons que les relations syntagmatiques et les relations paradigmatiques sont considérées toutes deux en tant que relations sémantiques. Les relations syntagmatiques s'effectuent sur l'axe syntagmatique (traditionnellement en linguistique, l'axe horizontal) qui est relatif à la morphosyntaxe, à l'agencement des éléments dans la phrase. Quant aux relations paradigmatiques, elles s'effectuent sur l'axe paradigmatique (traditionnellement en linguistique, l'axe vertical) qui est relatif au sens d'éléments de la phrase et permet la permutation ou la substitution (asymétrique) qu'évoque Ménard (1989). Ainsi, nous retenons de Stock (2010) qu'une différence de forme fait partie des phénomènes qui entraînent une différence de sens.

Buckland (1999) adopte une approche de vulgarisation de la langue. Il étudie la nature et le rôle du vocabulaire dans les systèmes d'information. Il aborde la question du vocabulaire en identifiant cinq vocabulaires, en réalité, cinq *discours* au sens linguistique, directement en rapport avec un schéma de la communication. Il propose une échelle de variation de la forme et du sens :

> Minimally the relationship between any pair of terms in just two vocabularies involves four contingencies:
> Same form, same meaning: Same word.
> Same form, different meaning: Homograph.
> Different form, same meaning: Synonym.
> Different form, different meaning: Different word. (Buckland 1999)

L'auteur précise que « same » ici n'indique pas l'identité au sens strict mais, concrètement, un équivalent acceptable pour le but recherché (Buckland 1999, n. t.). Notons d'abord qu'une même forme avec un sens différent peut rendre compte d'une relation non seulement d'homonymie, mais aussi de polysémie. La différence entre les deux n'est pas toujours claire et souvent l'on se reporte aux décisions prises par les dictionnaires pour savoir s'ils ont établi deux entrées distinctes (cas d'homonymie) ou bien deux sections au sein d'une même entrée (cas de polysémie)[40].

Dans notre analyse, nous nous sommes reportée à des dictionnaires de la langue française. Cette méthode a notamment été éprouvée par Mortureux (Mortureux et Petit 1989) et Cheminée (1990, sous la direction de Mortureux) dans son analyse des reformulations : « En examinant l'ensemble des informations disponibles dans quelques dictionnaires de langue pris en référence, on peut échelonner la distance qui sépare le lexème *timbre* de plusieurs de ses substituts » (Mortureux 1993). Les objets analysés dans ce travail ne sont pas très différents de ceux de Mortureux : celle-ci étudie les reformulations qui peuvent assurer la cohérence du texte par la thématisation et la conceptualisation de notions dans le discours. De notre côté, nous nous penchons sur les expressions linguistiques porteuses de sujet (thémanymes) d'une recherche d'information par un usager auprès d'un archiviste dans la chaîne documentaire archivistique. Nous revenons sur l'usage des dictionnaires dans la méthodologie (voir section 4.2.5.3).

Pour les besoins de notre recherche, nous avons élaboré une échelle de valeurs plus précise, tirée en partie de Jacquemin (2001). Celui-ci étudie la variation par la forme et le sens d'expressions pouvant être couvertes par un même terme d'indexation mais il ne propose pas de grille en tant que telle dans son étude. Nous avons développé une grille et avons ajouté à celle-ci des éléments de variation entre deux expressions à partir des phénomènes classiques en linguistique tels que la dérivation, la paraphrase, etc. Nous décrivons et exemplifions chacun des éléments de la grille dans la section 5.1.1.

[40] Nous ne traitons pas ici de la distinction entre l'homonymie et la polysémie. Voir par exemple Gross (2015).

Le lien sémantique qui existe entre deux expressions linguistiques peut être une relation sémantique connue parce que récurrente dans la langue. Par exemple, la relation entre *animal* et *chat* est claire, il s'agit d'une relation d'hyperonymie où *animal* est un terme plus générique que *chat*; le sens d'*animal* est contenu dans celui de *chat* et le chat appartient à la classe des animaux. Mais certaines relations sémantiques peuvent être *ad hoc* ou simplement moins fréquentes si bien qu'elles n'auront pas de nom en tant que tel. À ce moment, à la suite de Kister, Jacquey et Gaiffe, nous usons de « verbalisation de liens sémantiques » (2011, paragraphe 44) et nous apposons un nom explicite à des relations sémantiques que nous décrivons (par des précisions). Les liens sémantiques ayant un élément de sens commun forment un champ sémantique (voir section 3.3.3). Cette notion développée en premier par Trier (1894-1970) consiste en un « Ensemble des mots, des notions se rapportant à un même domaine conceptuel ou psychologique » (*Trésor de la langue française informatisé* 2017). Certains linguistes établissent des différences entre le champ lexical et le champ sémantique (notamment Polguère 2013). Mortureux parle d'« isotopie sémantique portée par le lexique » (1993, paragraphe 26). Nous nous reportons à la définition originelle du terme et nous l'intégrons à notre échelle d'écart sémantique (voir section 4.3.1). Nous présentons plus loin quelques exemples présents dans notre corpus à partir de la comparaison des thémanymes (voir section 5.1.1).

Les relations sémantiques sont des liens entre des concepts ou des expressions linguistiques, ces dernières peuvent être de longueur plus ou moins étendue allant du mot au texte en passant par la phrase ou ses segments (Khoo et Na 2006, 160). Dans notre recherche, nous les envisageons principalement entre des expressions linguistiques porteuses du sujet (des thémanymes) relativement courtes : mot, groupe de mots ou syntagme[41], segment de phrase qui dépasserait le syntagme. Une des propriétés des relations sémantiques (Murphy

[41] « Syntagme : unité syntaxique intermédiaire entre le mot et la phrase. Aussi appelé *groupe*, le syntagme constitue une unité de sens dont chaque constituant conserve sa signification et sa syntaxe propre. » (Bougoin 2015, 19)

2003, cité par Khoo et Na 2006, 160-161[42]) est la prédictibilité : les relations sémantiques suivent certains patrons généraux et des règles. Cet élément nous intéresse particulièrement dans la visée de recommandations relatives au vocabulaire employé pour l'accès thématique aux archives patrimoniales (VATAP, voir section 7.2).

3.5.4. La notion d'écart sémantique

La dernière notion linguistique relative à l'aspect sémantique tel qu'il est traité en linguistique et que nous abordons dans ce texte parce qu'elle touche l'indexation concerne l'écart sémantique. La variation de forme induit généralement une variation de sens. Ainsi, le guide d'indexation de BAnQ (2012) recommande d'utiliser le pluriel pour les noms concrets et le singulier pour les noms abstraits. Le trait concret/abstrait permet parfois de distinguer les sens de certains mots. Ainsi, un même terme peut apparaître deux fois dans l'index, une fois au singulier en tant que nom abstrait (p. ex. *prison* au sens de la peine, l'état punitif) et au pluriel en tant que nom concret (p. ex *prisons* au sens de bâtiment, établissement carcéral).

Dans notre recherche, un écart sémantique est un éloignement relatif au sens entre deux expressions linguistiques. Par exemple, quand le terme qui figure dans un index n'est pas exactement identique aux expressions linguistiques dénotant le concept retenu pour l'indexation (exemples de différences : singulier/pluriel, synonymes, paraphrases ou définition, hyperonyme ou hyponyme), alors il y a nécessairement un écart. Nous présupposons que tout changement dans la forme induit un changement dans le sens de l'expression linguistique. L'écart sémantique est plus ou moins grand selon que la variation de forme est plus ou moins importante. Pour identifier les variations de forme plus finement, nous nous appuyons notamment sur Jacquemin (2001) qui détaille la variation de formes en vue d'un traitement automatique de la langue, de la plus insignifiante à la plus complexe, en partant de la forme pour aller vers le sens. Le dernier écart de notre grille (l'échelle d'écart

[42] La référence donnée par les auteurs est la suivante : Murphy, M. L. (2003). *Semantic relations and the lexicon: Antonymy, synonymy, and other paradigms*. Cambridge, UK : University Press.

sémantique, voir section 4.3.1) est le champ sémantique que nous avons déjà présenté en tant qu'autre méthode d'analyse du sujet (voir section 3.3.3.4).

3.6. Conclusion du chapitre

En conclusion de cette revue de littérature, nous pouvons dire que l'indexation thématique en archivistique a été peu étudiée que ce soit de manière générale ou bien selon un point de vue linguistique. Nous avons vu que l'accès par sujet des documents d'archives est tributaire de la nature des documents d'archives et des habitudes des usagers. Cependant, il réside un flou dans la définition du sujet de manière théorique en archivistique qui se définit plutôt de manière pragmatique, en fonction des besoins des usagers, quand on étudie leurs questions posées par courriel à des archivistes de référence. Il apparaît alors pertinent de remonter la chaîne communicationnelle de la référence archivistique de l'usager au document d'archives pour mieux rende compte des paramètres applicables à l'indexation thématique des archives patrimoniales et à l'accès thématique en général. Plusieurs types d'indexation existent, mais le recours à des langages d'indexation contrôlés est largement encouragé par les normes ou lignes directrices. Pourtant, le vocabulaire libre retrouve actuellement une certaine légitimité dans les services d'archives, qu'il soit généré par des usagers ou par des professionnels. C'est pourquoi notre étude se concentre sur les termes du VATAP en vocabulaire libre, employé pour indexer mais aussi décrire et rechercher. Côté linguistique, les normes ont déjà relevé certaines propriétés linguistiques, mais ce sont les ouvrages de lexicologie qui sont les plus complets. Les termes de recherche, de description et d'indexation entretiennent des relations sémantiques les uns avec les autres qu'il reste à identifier, à partir de notre grille appelée « échelle d'écart sémantique », afin notamment de dégager des récurrences. Il découlera de notre étude des recommandations pour favoriser l'accès thématique aux archives patrimoniales, c'est-à-dire des recommandations sur l'indexation ainsi que la description en tant qu'outils de repérage.

4. Méthodologie

Notre recherche doctorale est une application de la linguistique à l'archivistique. Nous avons appliqué des principes d'analyse sémantique au vocabulaire pour l'accès thématique des archives patrimoniales (VATAP). Ce vocabulaire est une représentation des documents d'archives, issue à la fois de la recherche, de la description et de l'indexation – opérations effectuées par des archivistes et des usagers – et manifestée par des courriels d'usagers à la référence, des notices descriptives archivistiques et des termes d'indexation professionnelle ou des étiquettes (*tags* en anglais). Le VATAP est constitué d'expressions linguistiques porteuses de sujet, que nous appelons des thémanymes.

Cette recherche s'inscrit globalement dans une approche qualitative (Creswell 2009; Fortin 2010). Bien que notre troisième objectif soit de nature quantitative (*calculer la fréquence*, voir section 2.2 Objectifs), il ne nous sert qu'à obtenir des mesures indicatives. Il permet d'identifier la part des reprises exactes, entre le vocabulaire des usagers et celui employé dans les instruments de recherche créés par les archivistes. Il permet ainsi de quantifier l'écart sémantique entre les deux vocabulaires et ce, en dénombrant des occurrences de codes et en appréciant l'émergence de patrons récurrents parmi ceux-ci (Miles et Huberman 1994, 252-254). Le devis de cette recherche est exploratoire, en raison du peu d'études sur le sujet, et descriptif, pour nous permettre d'établir des critères de description d'un plus grand échantillon lors de recherches futures (Fortin 2010, 170-186).

Nous avons procédé à une analyse de type linguistique, principalement lexico-sémantique (relative au sens des mots) sur un corpus de termes. Un tableau méthodologique présenté en annexes (voir Annexe 3) récapitule la démarche méthodologique globale de notre recherche. Dans la section 4.1, nous présentons la collecte en décrivant la méthode de collecte, les centres et services d'archives participant à notre recherche et la méthode d'échantillonnage de ces milieux, le corpus d'étude ainsi que la méthode d'échantillonnage du corpus et, finalement, les données elles-mêmes telles qu'elles se présentaient lors de la collecte. Dans la section 4.2, nous détaillons le prétraitement des données nécessaire à leur analyse. Dans la section 4.3, nous expliquons la méthode d'analyse. Dans la section 4.4, nous présentons la qualité de la recherche.

4.1. Collecte de données

Dans cette section, nous décrivons les milieux de notre collecte, les procédures de sélection de notre corpus et finalement, nous décrivons les données collectées.

4.1.1. Description des milieux

Dans cette section, nous décrivons les critères de collecte de nos données, à savoir les critères de sélection des services d'archives susceptibles d'avoir les données que nous cherchions pour notre étude, le mode de sollicitation des archivistes dans ces établissements (*i. e.* la prise de contact) et la description des services d'archives prenant part à notre recherche, c'est-à-dire les milieux de notre étude.

Pour étudier le VATAP, nous avons envisagé l'ensemble des établissements archivistiques patrimoniaux du Québec (Canada). Nous avons restreint cet ensemble selon un échantillonnage non probabiliste, à choix raisonné (Fortin 2010).

4.1.1.1. Critères de sélection des services d'archives

Dans cette section, nous exposons et justifions les critères de l'échantillonnage retenu. Puisque nous étudions les termes en français, nous avons sélectionné des établissements francophones ou bilingues. Nous avons ajouté un critère de convenance, à savoir des établissements situés dans la région où l'on parle le *français laurentien* (LeBlanc, Martineau et Frénette 2010), aux alentours de Montréal. Nous visions des services d'archives de grande envergure (p. ex. Bibliothèque et Archives Canada, Bibliothèque et Archives nationales du Québec), parce qu'ils ont les moyens (financiers, humains, etc.) de procéder à l'indexation professionnelle ou de gérer une indexation en collaboration avec des usagers par des albums *Flickr* (p. ex. Archives de la Ville de Montréal, Archives/UdeM (Direction des archives et des documents de l'Université de Montréal), Musée McCord). Les indexations professionnelle et collaborative sont définies plus haut (voir section 3.4.3 Termes d'indexation).

Dans un premier temps, nous souhaitions évaluer la faisabilité de notre étude et nous rendre compte des quantités de termes du VATAP éventuellement disponibles et accessibles. Nous avons envoyé une courte demande par courriel à un groupe pilote d'archivistes[43] de notre réseau. Ainsi, nous avons discuté avec des archivistes de cinq services d'archives de grande envergure à propos de leur pratique de l'indexation par sujet, les outils utilisés lors de la référence et la gestion des demandes des usagers dans le service. Nous avons interrompu ce premier survol de la quantité et de la qualité des données disponibles au moment où nous nous sommes rendue compte que les données de ces cinq services étaient déjà trop volumineuses pour envisager un traitement linguistique humain. Nous avons alors posé un autre critère de sélection des milieux : avoir le plus grand nombre de données issues du plus grand nombre de sources envisagées pour notre recherche. Les sources correspondent au bassin de mots dont nous avons tiré les termes de recherche, de description et d'indexation, à savoir les questions d'usagers, les notices descriptives, les termes d'indexation et les étiquettes.

Malgré leurs riches données, deux milieux répondaient moins bien à nos critères, nous ne les avons pas retenus. Sur les cinq services contactés, trois milieux ont ainsi répondu au mieux à nos critères et ont accepté de participer à notre recherche : nous les nommons ci-après établissement 1, établissement 2, établissement 3.

Dans l'établissement 1, le responsable de l'indexation nous a envoyé par courriel le fichier contenant tous les termes d'indexation par sujet avec des noms communs, utilisés par les archivistes pour indexer les notices descriptives de la base de données archivistique (interne et externe). Concernant la référence, nous avons demandé une autorisation à l'établissement de consulter les courriels reçus par les archivistes de la référence archivistique des dix centres régionaux. Compte tenu du nombre important de courriels, nous nous sommes limitée, par convenance, à cinq centres régionaux (les plus importants et les plus accessibles géographiquement à partir de Montréal). Nous avons sélectionné les questions des usagers

[43] Dans ce texte, le terme *archiviste* est employé au masculin dans le seul but d'alléger la lecture.

selon notre grille (voir section 4.1.3.1), en incluant les réponses d'archivistes. Nous avons rendu les courriels anonymes.

Dans l'établissement 2, l'indexation est faite avec des vedettes-matière. Mais les archivistes ont également l'option de remplir un champ d'indexation libre. Nous avons récupéré ce fichier de termes. Celui-ci inclut des noms communs et des noms propres, en anglais et en français. En outre, nous avons obtenu de la part des archivistes de référence un tableau Excel contenant les demandes des usagers avec la réponse du ou des archivistes et les liens vers les notices descriptives suggérées à l'usager pour répondre à son besoin. Nous avons rendu les courriels anonymes. Finalement, les notices descriptives sont accessibles en ligne par le Web ou par le moteur de recherche interne auquel nous avons eu accès sur place. Le moteur de recherche de la page Web de l'établissement est plus englobant et ne retourne pas les mêmes résultats que le moteur de recherche interne pour lequel les résultats étaient plus limités (moins de bruit, plus de précision).

Dans l'établissement 3, nous avons rencontré un archiviste de référence. Pour renseigner les usagers, sont utilisés plusieurs instruments de recherche papier et informatisés (numériques et numérisés), mais surtout informatisés. Pour la recherche, les questions d'usagers étaient disponibles dans une boîte de courriel dédiée à la référence, seulement depuis le printemps 2015. Nous avons vérifié chacun des courriels un à un avec l'archiviste, lors de notre rencontre en novembre 2015. Après entente sur la confidentialité, nous avons copié les fichiers et les avons rendus anonymes. Pour la description, les notices descriptives sont accessibles dans leur catalogue en ligne.

L'archiviste de référence de chacun des deux premiers établissements nous a donné accès non pas à toutes les questions posées à la référence qui sont extrêmement nombreuses, mais seulement aux questions qui demandent une certaine connaissance des instruments de recherche archivistiques. Un premier tri des questions de référence a été fait par un professionnel qui assigne les questions à la référence selon le niveau de difficulté à y répondre et selon la discipline invoquée (archivistique ou bibliothéconomie), puisque ces deux établissements ont une mission double (voir section 4.1.1.2 Description des trois établissements participants). L'archiviste de référence de chacun des établissements nous a expliqué que les demandes archivistiques auxquelles les personnes de la référence « générale »

(archivistes ou non) répondent ne nécessitent pas de connaissances archivistiques poussées telles que la connaissance du fonctionnement des instruments de recherche ou des instruments de recherche eux-mêmes. Voici des exemples de questions catégorisées « simples » repris tels que nous les avons reçus :

> Bonjour, nous recevons une demande de notre bureau législatif afin de recevoir copie du décret 1971-2477 ainsi que toutes annexes et records au dossier .s. v. p,. (Usager 1245)
>
> Je cherche le titre français, s'il existe, d'une photo dont la légende anglaise dit : PANNING GOLD ON THE FRASER, B. C. On voit sur cette photo un chercheur d'or à genoux dans le lit d'un cours d'eau, avec un tamis dans les mains. La référence est la suivante : Library and Archives Canada / PA-021422. Pourriez-vous me dire aussi comment on indique cette référence en français, s'il y a lieu. (Usager 1276)
>
> Je suis à la recherche d'un timbre illustrant la Forteresse de Louibourg et n'ayant pas de valeur nominal [*sic*], mais je crois qu'il pourrait avoir été émis en 1993 et avoir une valeur de .43 cents. (Usager 1850)

Voici des exemples de questions catégorisées « complexes » :

> Je suis à la recherche de documents d'archives pour documenter la construction d'une patinoire dans la Ville de Salaberry-de-Valleyfield. La construction a eu lieu à partir de l'automne 1894. Les documents intéressants se trouveraient durant la période 1890 à 1920. La compagnie responsable de la construction est la société d'amusement Ste-Cécile et le nom de la patinoire est Le palais des amusements. La particularité est que la patinoire a été construite le long du Canal de Beauharnois. Puisque le nom d'NOM, ingénieur-surintendant des canaux, est mentionné, je pense qu'il pourrait y avoir des informations, correspondance, cartes et plans ou entente dans les fonds gouvernementaux. Est-ce qu'il y a une méthode simple pour retrouver les fonds concernés? (Usager 36737)
>
> Je suis présentement à la recherche de documents photographiques, audiovisuels et sonores issues ou concernant les pensionnats autochtones canadiens. Il peut s'agir de photographies personnelles, de négatifs, de documentaires, etc. Exclusivement, pour les documents visuels, je cherche des présences humaines (ainsi les photographies des bâtiments ne sont pas nécessaires). J'ai trouvé 32 photographies dans votre banque d'archives en ligne. Je souhaiterais en trouver davantage. (Usager 473)
>
> Mon aïeul NOM (1795-1873) était avocat, député en 1936 du Saguenay et patriote. Il demeurait à Québec. J'écris un livre sur l'histoire et la généalogie de ma famille Drolet. Je consacre plusieurs pages à NOM. Il fomente et dirige l'évasion de Dodge et Theller de la Citadelle de Québec le 16 octobre 1838. Mon livre va bon train, mais je cherche depuis plusieurs années la photographie de NOM. J'ai en ma possession de la correspondance entre NOM et Louis Joseph Papineau. Ces lettres proviennent du fonds de la famille Papineau à Ottawa, numéro MG24,B2. Dans une lettre datée du 5 octobre 1864 NOM écrit « J'accède à votre demande, et je vous envoie ma photographie: heureux de l'espoir de la voir placé dans l'album qui, dites-vous devra contenir les

> portraits des braves et désintéressés patriotes qui se glorifiaient de vous avoir pour guide ». Cette lettre de trois pages est signée par NOM et envoyé [*sic*] Louis-Joseph Papineau à Montebello. J'espère que l'album existe et qu'après toutes ces années de recherches et de trouvailles je pourrai enfin voir la photo de NOM. (Usager 35636)

Nous ne traitons donc dans cette étude que les questions complexes auxquelles ont répondu des archivistes. Parmi les trois questions ci-dessus, le sujet de la dernière est formulé par provenance. La première question contient un sujet exprimé par des noms communs mais le travail de traduction en accès de provenance a été amorcé par l'usager et son besoin exprimé concerne davantage la stratégie de recherche. La deuxième constitue un sujet exprimé par des noms communs tel qu'on l'entend dans notre recherche.

4.1.1.2. Description des trois établissements participants

Pour participer à notre recherche, les centres et services d'archives devaient répondre aux critères qui suivent :

- francophones ou bilingues, pour que les données soient en français, critère d'inclusion dans le corpus (voir section 4.1.2.3);
- à proximité de Montréal (critère de convenance);
- de grande envergure, c'est-à-dire qui détenait les moyens de procéder à l'indexation par sujet des archives;
- offrant l'accès au plus grand nombre de sources de termes (questions d'usagers avec réponse d'archivistes, notices descriptives, termes d'indexation, étiquettes).

Après une entente formelle de consentement à participer à notre recherche, nous avons recueilli des données de trois établissements. Dans cette section, nous décrivons, en fonction des informations qui nous ont été accessibles, ces trois établissements, essentiellement la mission propre à chacun d'eux, leur pratique relative à la description et à l'indexation thématique et finalement, de manière conjointe, les données de leur présence sur la plateforme *Flickr*.

Établissement 1

1. La mission

L'établissement 1 est issu de la fusion de trois établissements à vocation patrimoniale. Selon la loi constitutive, « Établissement 1 a pour mission de rassembler, de conserver de manière permanente et de diffuser le patrimoine documentaire québécois publié et tout document qui s'y rattache et qui présente un intérêt culturel, de même que tout document relatif au Québec et publié à l'extérieur du Québec » (loi constitutive). Son rapport annuel 2015-2016 souligne le rôle essentiel qu'il se donne dans la société du savoir : « (…) l'établissement 1 rassemble, conserve et diffuse le patrimoine documentaire québécois ou relatif au Québec. Véritable carrefour culturel, l'établissement 1 œuvre à la démocratisation de l'accès à la connaissance à titre d'acteur clé de la société du savoir (…). » (*Rapport annuel 2015-2016*, Établissement 1 2016). En ce qui concerne spécifiquement les archives, la loi précise que la mission de l'établissement 1 consiste à soutenir les organismes publics en matière de gestion de leurs documents, conserver et faciliter l'accès aux archives publiques et promouvoir la conservation et l'accessibilité des archives privées. Le site Internet de l'établissement 1 est plus précis quant aux divers types d'archives qu'il a sous sa garde : « l'établissement 1 conserve et diffuse les archives les plus représentatives de l'histoire du Québec. Les ministères et les organismes gouvernementaux ainsi que les tribunaux versent à l'établissement 1 leurs documents destinés à une conservation permanente. Les archives judiciaires et civiles (notaires et état civil) consignent divers aspects de la vie sociale, familiale, foncière et économique des personnes, permettant à ces dernières de connaître leur histoire ou de faire valoir leurs droits. Pour compléter ce patrimoine « officiel », l'établissement 1 acquiert aussi des archives privées de personnes, de familles, d'organismes et d'entreprises d'horizons variés et de multiples sphères d'activité. » (*site Internet de l'établissement 1, 2017*). Cette description donne un aperçu de l'éventail des nombreux sujets potentiels.

2. La description

L'établissement 1 suit la norme de description des *RDDA*. Le site Internet de l'établissement 1 présente l'instrument de recherche général du service d'archives : « la banque de données NOM (…) donne accès à la description des fonds et collections conservés

à l'établissement 1 » (site Internet de l'établissement 1 2017). D'après notre observation, dans la base de données NOM, une notice descriptive, quel que soit son niveau hiérarchique, peut contenir les rubriques présentées dans la Figure 6 site Internet de l'établissement 1 2017).

- la cote;
- le centre, soit le centre qui conserve les documents;
- le titre officiel de l'unité de description;
- les dates permettant d'identifier la période chronologique couverte par les documents;
- la quantité de documents associés à l'unité de description ainsi que leurs caractéristiques physiques;
- les restrictions, s'il y a lieu. Les détails concernant ces restrictions pourront être consultés dans l'écran d'affichage des contenants;
- la notice biographique/histoire administrative de la personne, de la famille ou de l'organisme assumant la responsabilité principale ou entière de la création, du rassemblement et de l'accumulation des documents associés à l'unité de description sélectionnée;
- le champ historique de la conservation/source immédiate de l'acquisition identifie les propriétaires et les gardiens successifs de l'unité de description. Il indique aussi les dates de transfert de propriété ou de responsabilité de conservation. Le nom du cédant, le mode (versement, don, vente, prêt ou dépôt) et la date d'acquisition sont aussi précisés;
- les éléments d'information concernant portée et contenu ont trait aux fonctions (pouvoirs, devoirs, responsabilités) et aux activités (actes, tâches) à l'origine de la création des documents, aux sujets qu'ils abordent, aux personnes physiques, aux familles et aux personnes morales qu'ils mentionnent, à la période qu'ils couvrent, aux bâtiments et aux lieux géographiques auxquels ils se rapportent. On peut retrouver également la classification interne de l'unité de description;
- d'éventuels compléments d'information;
- les instrument(s) de recherche, soit le titre de l'instrument de recherche, son auteur, l'année de sa production et le nombre de pages. Sont aussi précisées : l'existence ou non d'un index faisant partie de l'instrument de recherche, la nature de son support de consultation (papier, microfiches, etc.) et sa localisation (numéro de six chiffres);
- les termes rattachés, soit les termes associés à une unité de description (indexation)

Figure 6. Éléments d'une notice descriptive – Établissement 1

Lors de l'un de nos entretiens, le responsable de l'indexation a mentionné que les notices descriptives étaient rédigées « en langage courant » et que ceci rendait leur indexation automatique plus difficile (communication personnelle, automne 2015). Par exemple, la pêche peut être commerciale ou sportive, un père et un fils portant le même nom ont des dates de vie qui se chevauchent, un toponyme peut renvoyer à plusieurs zones délimitées selon des normes différentes (municipalité de village, de canton, etc.). La langue – le « langage courant » ou le vocabulaire libre – présente de nombreux défis pour l'indexation automatique des notices descriptives.

3. L'indexation

Dans l'établissement 1, l'indexation est contrôlée, mais les suggestions d'ajout étaient fréquentes tant que la base était en cours de constitution. Les termes d'indexation rattachés à

une notice descriptive de cet établissement sont issus de plusieurs « registres ». Les sources des termes d'indexation sont au nombre de six. Le *Répertoire des vedettes-matière* (*RVM*) de l'Université Laval est la première, elle contient environ 400 000 termes et constitue la première source d'autorité. Sont consultés aussi le *Thésaurus de l'activité gouvernementale* (*TAG*), utile pour faire le pont entre la langue courante et la terminologie du gouvernement et une base de données des parlementaires (voir notamment Assemblée nationale du Québec 2018). Pour les toponymes, la *Banque de noms de lieux du Québec* de la Commission de toponymie du Québec (2018) est incontournable. *La mémoire du Québec* (2018) est un dictionnaire de noms propres relatifs au Québec, pour les lieux et les personnes. Enfin, le fichier d'autorité local de noms communs contient environ 45 000 termes. Ces divers « registres » sont intégrés à la base de données interne.

Le responsable de l'indexation dans l'établissement 1 nous a fourni la politique d'indexation révisée en décembre 2012 ainsi que le fichier des termes d'indexation qui comprend 44 924 termes d'indexation (en français), mis à jour en décembre 2012, dont 17 644 termes acceptés. Parmi ces termes, certains prennent la forme d'une vedette-matière : deux noms ou groupes nominaux sont séparés par deux tirets. La seconde partie du terme ainsi composé revêt diverses natures : complément d'information sur le concept (*Achigans -- Pêche aux filets* et *Achigans -- Pêche sportive*), l'identification de l'aspect traité (*Acide sulfurique -- Industrie*). D'autres termes d'indexation peuvent prendre la forme de descripteurs de thésaurus, avec parfois des qualificateurs entre parenthèses. Ces qualificateurs ont diverses fonctions : désambigüisation (p. ex. *Jersey (Race bovine)*), précision de la nature du terme (p. ex. Laura Secord *Marque de commerce*), précision du domaine d'appartenance du concept (*Musique* pour *Accompagnateur*), complément d'information sur le concept (*Secondaire* ou *Primaire* pour *Étude et enseignement*). Certains termes sont hybrides et présentent un double tiret et des parenthèses introduisant un qualificateur sur la première ou la seconde partie du terme (*Acquisitions (Bibliothèques) – Manuscrits*; *Actes notariés -- Dépôt (Droit)*). Certains peuvent ainsi être des vedettes-matière. Le fichier de termes d'indexation de l'établissement 1 présente trois colonnes : le terme, le statut accepté ou rejeté et parfois une remarque pouvant inclure la date de modification, l'auteur de la modification, la source (p. ex. *RVM*), la cote de documents auxquels il correspond. À peu près la moitié des remarques comprennent une cote

archivistique. Le Tableau II présente un court extrait du fichier de termes d'indexation de l'établissement 1.

Tableau II. Extrait du fichier de termes d'indexation (établissement 1)

Nom commun	Accepté/ Refusé	Note sur le terme
Aberdeen-Angus (Race bovine) -- Vente	Terme accepté	X NOTE : AUCUNE ; SOURCE : RVM ; REMARQUE : 06M E6,S7,SS1,D630898 À 630912 ; DATE : 2003-01-22
Emballages-cadeaux	Terme accepté	2015-2016 Montréal vmil1411; NOTE : AUCUNE ; SOURCE : RVM; REMARQUE : 06M,P48,S1,P3670 + 06M,P48,S1,P3671; RESPONSABLE : *NOM*; DATE : 99/07/23
Embarcations de sauvetage	Terme accepté	NOTE : SOUS CE TERME, ON TROUVE DES DOCUMENTS SUR LES EMBARCATIONS DESTINÉES À PORTER SECOURS AUX NAUFRAGÉS ET QUI SE TROUVENT À BORD DES BATEAUX, DES NAVIRES OU DES AVIONS. LES DOCUMENTS SUR LES EMBARCATIONS DESTINÉES AUX MÊMES FINS MAIS QUI PARTENT DES PORTS SE TROUVENT SOUS LE TERME « CANOTS DE SAUVETAGE » ; SOURCE : RVM ; RESPONSABLES : *NOM, NOM* ; DATE : 99/11/26
Zouaves pontificaux	Terme accepté	NOTE : AUCUNE ; SOURCE : REPERTOIRE DES VEDETTES-MATIERES ; RESPONSABLE : PROJET ETUDIANT ÉTÉ 2000 ; DATE : 2000-06-12
Monoxyde de carbone	Terme refusé	2014-2015 MONTRÉAL VMIL1259
Morphogénèse (Géomorphologie)	Terme refusé	
Moulage (Argile, plâtre, etc.) -- Maquettes	Terme refusé	
Navigation (Aéronautique) -- Contrôle	Terme refusé	
Normes internationales du travail	Terme refusé	NOTE : CE RENVOI DOIT ÊTRE PRIS EN CONSIDÉRATION DANS LE CAS OÙ LE CONCEPT QU'IL EXPRIME EST LE SUJET DONT TRAITE L'UNITÉ DE DESCRIPTION ET NON PAS COMME UNE FORME DOCUMENTAIRE

La cote archivistique qui se trouve parfois dans la note sur le terme pourrait servir à retrouver les notices indexées par le terme en question.

Établissement 2

1. La mission

L'établissement 2 est le résultat de la fusion de deux établissements patrimoniaux. Selon sa loi constitutive, l'établissement 2 a notamment pour mission de constituer, préserver et diffuser le patrimoine canadien. Cette vaste mission se reflète dans les archives conservées et dans la grande variété des questions de référence. Dans son rapport annuel de 2015-2016, l'établissement 2 souligne justement que : « [e]n 2004, le Parlement du Canada a confié un mandat très large à l'établissement 2 : constituer, traiter et préserver le patrimoine

documentaire du Canada, et le rendre accessible. L'établissement 2 est également la mémoire permanente de l'administration fédérale et de ses institutions. » (Rapport annuel 2015-2016, Établissement 2).

Dans l'établissement 2, nous avons obtenu en un seul fichier l'ensemble des questions de référence sur la période voulue. Il y a 658 questions « simples » indifféremment en bibliothéconomie et en archivistique et 518 questions « complexes » en archivistique uniquement. Un rapide survol des questions de niveau de complexité faible a confirmé qu'elles ne seraient pas intéressantes pour notre recherche; en effet, très peu étaient des questions thématiques et le peu qui en était ne nécessitait effectivement pas de recherche approfondie dans les archives.

Le cas de la description dans l'établissement 2 est un peu complexe étant donné que les deux langues officielles se côtoient. Mais la politique interne préconise de rédiger la notice dans la langue majoritaire des documents ou dans les deux langues s'il y a égalité dans la proportion des deux langues officielles dans les documents; elle préconise que les accès (notamment les termes d'indexation) se fassent dans les deux langues officielles (Politique sur la langue de description des documents d'archives, approuvée par le Comité de la haute direction, Établissement 2, 1989, révisé en 2000). Il va sans dire que l'anglais est présent majoritairement, la majorité des documents étant en anglais. Selon le directeur du département de recherche stratégique, le français couvrirait environ 25% des fonds (communication personnelle, février 2016). Les archivistes qui rédigent les notices descriptives et indexent par sujet sont généralement des francophones bilingues. Les notices descriptives sont disponibles à l'interne et à l'externe dans une base de données; elles sont identifiées par un numéro unique d'identification. Le champ d'indexation libre inclut des termes de toutes sortes qu'il a fallu discriminer en fonction de la langue et de la catégorie grammaticale. Nous avons commencé manuellement afin d'avoir un échantillon test pour rechercher les notices qui correspondaient à ces termes. Du moment où une partie du terme d'indexation parfois long et segmenté par des barres obliques et des traits d'union était en français, nous le retenions pour notre corpus. Puis nous avons automatisé cette discrimination en fonction du type de données (voir section 4.1.3.3).

2. La description

L'établissement 2 suit la norme de description des *RDDA*. Les notices de description (des fonds et collections traités) et les notices d'acquisition (notice descriptive temporaire réalisée à l'acquisition et permettant de donner un premier accès sommaire aux usagers[44]) sont regroupées dans NOM, une base de données accessible sur place (avec une interface de recherche interne) et en ligne (avec un moteur de recherche puissant). Auparavant, il existait plusieurs bases de données selon les types de documents (p. ex. films, images, textes). Ces outils ont été fusionnés dans les dernières années, mais nous trouvons dans les réponses d'archivistes des liens vers ces autres bases de données, encore disponibles, dont les formulaires de recherche contiennent des champs spécifiques adaptés au type de document (p. ex. « le genre de document » tel que « cartes et documents cartographiques », « timbres-poste et documents philatéliques »; site Internet de l'établissement 2).

- le titre de l'unité archivistique (UA) avec généralement l'indication générale du genre de document prévue aux *RDDA* dans la section 1.1C page 1-22 (BCA 2008),
- la structure de classement (un lien permet de visualiser où se trouve l'UA dans l'arborescence classificatoire,
- le type d'UA (p. ex. dossier) suivi du titre du fonds auquel elle appartient,
- les dates de création,
- le lieu de création (il est parfois indiqué qu'il est inconnu ou indéterminé),
- l'étendue (appelée dans les *RDDA collation*),
- la langue de description,
- la portée et contenu (s'il y en a une),
- les conditions d'accès (mention de la catégorie de document et la nature de la restriction d'accès, p. ex. « Documents textuels Volume 5 Numéro de dossier 49 : 99 : fermé pour fins de traitement »),
- les modalités d'utilisation (p. ex., mention de la présence de photocopies pour préserver les originaux),
- un instrument de recherche (nature et localisation, éventuellement en ligne, s'il y en a un)
- le créateur ou la provenance,
- des informations additionnelles (ancienne cote, historique de la conservation, emplacement des originaux)
- les vedettes-matière,
- la nature de la source (p. ex. *privé*, *gouvernement*),
- le numéro de contrôle d'autres systèmes
- le numéro d'identification.

Figure 7. Éléments d'une notice descriptive – Établissement 2

[44] Cet élément s'inscrit dans le courant *More Product Less Process* (MPLP) : voir Greene et Meissner (2005) pour une introduction au concept.

À partir de notre observation de plusieurs centaines d'enregistrements de l'établissement 2, une notice descriptive, quel que soit son niveau, peut contenir les rubriques identifiées dans la Figure 7. Le numéro d'identification est un numéro unique attribué à chaque unité archivistique décrite dans une langue et donc présent sur chaque notice descriptive. Un numéro d'identification différent est attribué à l'unité archivistique selon la langue. Les fonds d'archives sont décrits dans la langue majoritaire des documents.

Certaines informations dans ces rubriques sont notées en hyperlien et permettent ainsi une recherche rapide. Par exemple, le nom du créateur est un hyperlien qui permet d'obtenir tous les résultats de ce créateur dans ce champ.

La cote archivistique est difficile à retrouver car deux systèmes de cotes archivistiques, l'un ancien, l'autre nouveau coexistent et ne se recouvrent pas à l'identique. Il est préférable de se fier au numéro d'identification, unique et propre à chaque notice descriptive. Le moteur de recherche de l'établissement permet de rechercher exclusivement ce champ.

L'établissement 2 étant bilingue, la description est faite dans la langue majoritaire des documents sauf pour les fonds bilingues pour lesquels la description est faite dans les deux langues (voir section 4.1.1.2). Ainsi, certaines notices descriptives, bien qu'existantes et indiquées à des usagers dans un courriel rédigé en français, n'ont pas été retenues dans notre recherche parce qu'elles ne satisfaisaient pas au critère de langue (voir section 4.1.2.3).

3. L'indexation

L'établissement 2 procède à l'indexation des notices de description de documents d'archives sur une base volontaire de la part de l'archiviste rédacteur. Si bien que l'indexation est très variable d'un archiviste à l'autre, en quantité (le nombre de termes attribués à une unité archivistique est variable) et en nature (les concepts varient en profondeur et les groupes de mots varient en forme, nous les décrivons ci-après). L'établissement 2 a recours à des vedettes-matière bilingues pour indexer ses ressources archivistiques. Par contre, il est toujours possible à l'archiviste rédacteur d'ajouter des termes libres. Cette indexation en vocabulaire libre prend parfois la forme d'une vedette-matière, ou bien de termes d'indexation simples ou composés que l'on pourrait comparer à des descripteurs dans un thésaurus. Mais puisque l'indexation est totalement libre, les termes d'indexation peuvent prendre des formes hybrides ou ne ressembler à aucun modèle existant (p. ex. un ensemble de mots divisé par des

signes de ponctuation et agrémenté de dates). De plus, puisque il n'y a pas de vérification de ces termes, ils peuvent parfois (souvent) inclure des erreurs (frappe, orthographe, typographie, etc.). Un membre du personnel archiviste de l'établissement 2 a extrait la liste des termes présents dans le champ d'indexation libre. Ce champ est disponible à la fois dans la notice descriptive en anglais et en français. Ce fichier énumère toutes les chaînes de caractères ayant été insérées dans ce champ, soit 85 163 termes. La première tâche a donc été de trier ces termes en fonction de la probabilité des termes d'appartenir au français (traitement automatique de la langue), à partir de fichiers de vocabulaire anglais ou français. Nous nous sommes alors reportée à des outils de traitement automatique de la langue (voir p. ex. Bougoin 2015, 18sq); nous avons utilisé principalement *Indexo*, logiciel d'indexation qui permet notamment d'extraire et de compter les formes et termes différents et de lemmatiser les termes extraits, développé par Da Sylva (2009a). Nous avons identifié automatiquement 56 532 termes anglais (66%), 9 856 termes français (12%), 7 598 termes indifférenciés, une forme pouvant appartenir à l'une ou l'autre langue (9%) et 6 281 termes à la langue non reconnaissable, une forme reconnue comme ne pouvant appartenir ni au français ni à l'anglais, mais à d'autres langues (7%). Le restant consiste en des suites de caractères invalides soit 4869 lignes dans la base de données (6%). Nous avons utilisé le fichier comprenant des termes clairement identifiés comme appartenant au français (9 856 termes français).

Établissement 3

1. La mission

L'établissement 3 « a pour mandat d'acquérir les documents institutionnels et privés qui constituent les archives (…), d'en assurer le traitement, la conservation ainsi que la communication aux chercheurs et aux citoyens. » (*site Internet de l'établissement 3*, 2016) Si le territoire couvert par ce service d'archives est moins étendu que les deux autres établissements, sa mission – et, ce qui nous intéresse, la diversité thématique – est toute aussi large. Pour mener à bien sa mission, il a développé un site *Internet* et un portail de diffusion qu'il décrit en ces mots :

> Établissement 3 est un site entièrement consacré aux archives historiques de NOM. (…) Ce site constitue l'entrée principale de notre plate-forme. Il comprend plus de 400

articles élaborés en phase avec l'actualité, qui constituent autant de portes d'entrées thématiques vers nos archives. (*site Internet de l'établissement 3*)

Figure 8. Portail Web – Établissement 3

Cette figure rend compte de l'interconnexion entre les « outils spécialisés de diffusion : catalogue en ligne, médias sociaux, données ouvertes, expositions virtuelles et autres » (*site Internet de l'établissement 3*), développés par l'établissement 3.

2. La description

L'établissement 3 suit la norme de description des *RDDA* dans un logiciel qui adopte une mise en forme selon la norme internationale *ISAD(G)*, nous y revenons ci-après. Le portail Web de l'établissement 3 inclut, comme nous l'avons vu ci-dessus, le catalogue. Le catalogue en ligne « est l'instrument de recherche de base pour toute consultation dans nos archives » (site Internet de l'établissement 3, 2017).

> [Il] répertorie nos fonds et nos collections d'archives, tout en offrant un accès direct à nos documents numérisés. Toutes les descriptions disponibles de nos documents d'archives s'y retrouvent. Certains outils de recherche complémentaires (cartes-index, microfilms, etc.) doivent par ailleurs être consultés sur place, à notre salle de consultation. (site Internet de l'établissement 3, 2017).

Le catalogue repose sur le logiciel libre ICA-AtoM depuis 2013. ICA-AtoM est une application Web de description archivistique qui repose sur les normes du Conseil international des archives. *AtoM* est un acronyme signifiant: *« Access to Memory »*, Accès à la Mémoire. Le catalogue en ligne fait partie d'un « projet en développement », « toutes les descriptions disponibles de nos documents d'archives sont toutefois dans le catalogue » (site Internet de l'établissement 3, 2017).

Ainsi, le catalogue est basé non pas uniquement sur la norme canadienne des *RDDA* mais aussi sur ISAD(G). Les différences entre les deux normes sont minimes; par exemple, la note associée au titre se trouve sous le titre dans *ISAD(G)* et non dans une section des notes regroupées en fin de notice dans les *RDDA*. Une notice visualisable dans le catalogue indique de manière récurrente les mêmes sections, indiquées dans le Tableau III en romain alors qu'en italique sont la description ou l'exemplification que nous en faisons. La section intitulée « Document numérique – métadonnées » n'apparaît que pour les documents numériques. La notice descriptive se présente dans une page dont les colonnes de droite et de gauche servent respectivement à situer l'unité dans l'arborescence et accéder à des fonctionnalités variées telles que « télécharger l'instrument de recherche », « Parcourir les documents numériques », « exporter » la notice en « XML Dublin Core 1.1 » ou « XML EAD 2002 ». Certaines notices descriptives comportent une description abrégée. Il s'agit généralement de la description faite lors de l'acquisition et donc temporaire[45]. Lorsque le fonds n'est pas encore totalement traité, cette ébauche offre déjà au chercheur un aperçu de ce qu'il pourrait trouver dans le contenant qu'il pourrait demander à consulter.

[45] Pour une description de la différence entre une notice descriptive temporaire ou après traitement, voir Champagne, 2017.

Tableau III. Éléments d'une notice descriptive de l'établissement 3

niveau de description et cote (p. ex. Fonds J999) et titre et date(s) (police de grosse taille)	
chemin d'accès à l'unité archivistique rendant compte du positionnement dans l'arborescence (p. ex. Fonds NOM. – 1840-2002 >> Construction et aménagement des infr… . – 1880-1960 >> Permis de raccordement d'égout. – 189…)	
aperçu des documents numérisés, mention du nombre d'images disponibles (p. ex. Résultats 1 à 25 sur 477) et bouton « Afficher tout »	
Titre propre	*officiel ou composé*
Niveau de description	*indique le niveau hiérarchique de l'unité archivistique dans le reste de l'unité, p. ex. dossier, pièce*
Dépôt	*l'identification de l'établissement 3 avec une mention particulière*
Cote	*sous sa forme internationale, p. ex. CA J999 JJ999-J-9-99-J9999*
Date(s)	*date(s) de production (création) de l'unité archivistique*
Description matérielle	*la collation (p. ex. 1 gravure : entoilée ; 36 x 48 cm.)*
Nom du producteur/créateur	*NOM (dates de vie)*
Historique de la conservation	*mention des divers détenteurs de l'unité archivistique (généralement de premier niveau, fonds ou collection)*
Portée et contenu	*sujet et types des documents*
Localisation des originaux	
Source immédiate d'acquisition	*mention du passage de la garde du producteur/créateur au service*
Disponibilité d'autres formats	*mention de l'existence de copies sous d'autres formats (p. ex. Les dossiers sont en partie disponibles en version numérique.)*
État de conservation	*remarque sur l'état dans lequel est l'unité archivistique (p. ex. Le document est sale, jauni, taché et déchiré par endroits.)*
Classement	*note sur l'organisation physique de l'unité archivistique*
Instrument de recherche	*interne ou externe, identification (p. ex. Planche contact : J-999) ou lien URL sur le site Internet de l'établissement vers un pdf (p. ex. fonds-service-des-affaires-institutionnelles-99-2015.pdf)*
Éléments associés	*autres parties du fonds ou d'autres fonds ou encore référence vers un autre service d'archives ou une référence bibliographique*
Document numérique – métadonnées	
Nom du fichier	*J999-99_99J999_j999_jj.pdf*
Taille du fichier	*p. ex. 606.9 KiB*[46]

Ces notices d'acquisition sont considérées à part entière comme des notices descriptives dans notre recherche puisqu'elles incluent une description thématique du contenu et parfois des termes d'indexation.

3. L'indexation

L'établissement 3 ne procède pas à l'indexation thématique de ses archives ni de leurs notices descriptives. Certains types d'archives, à certaines périodes, ont été indexés par sujet, mais ce n'est pas le cas de la majorité des fonds. Pour les dossiers décisionnels, les dossiers

[46] *KiB* : un kibioctet représente 1024 octets (soit 8192 bits).

thématiques et les dossiers de photographies, il existe des index onomastiques ou thématiques, qui ont été rédigés à différentes périodes de la vie du service d'archives et relèvent de certaines habitudes de travail propres à l'établissement lors de ces périodes. Se présentant par exemple sur de petites cartes, les instruments de recherche ont été numérisés pour permettre une recherche plus rapide par ordinateur. Cependant, la lecture de l'écriture manuscrite n'est pas toujours évidente. L'établissement 3 procure un accès thématique aux usagers grâce à des billets sur son blogue rédigés autour d'une thématique et jouant ainsi le rôle des guides thématiques bien connus des services d'archives. Également, les archivistes de l'établissement 3 procèdent à l'indexation de certains documents déposés dans *Flickr*.

Présence sur Flickr

Les trois services d'archives dont émanent nos données diffusent des images de documents d'archives sur la plateforme *Flickr* : principalement documents iconographiques, mais aussi textuels, cartes et plans. Leur profil étant similaire, nous en avons rassemblé la description dans cette section. *Flickr* est un service de gestion et de partage de photographies en ligne. Sur le site de *Flickr*, à partir du profil de l'utilisateur, nous avons pu relever quelques informations sur les trois établissements source; ces informations sont à jour du 6 avril 2017. Les trois établissements ont des comptes *Flickr* « PRO ». Ce sont des comptes à l'abonnement payant qui offrent des services dont l'affichage du « badge pro » est visible sur le profil (Flickr 2017), ce qui assoit la crédibilité du détenteur de ce compte. Un membre de *Flickr* peut classer ses images en classeurs contenant des albums; une certaine classification peut être opérée et faciliter ainsi la recherche du visiteur. La recherche par mot clé porte sur le nom du membre, des classeurs, des albums, des images ou leur description.

L'établissement 1 est membre de *Flickr* depuis juillet 2015; il n'a aucun classeur, mais 21 albums rassemblant entre quatre et 165 photographies, totalisant 774 photographies, elle a 83 abonnés et trois abonnements. L'établissement 2 est membre de *Flickr* depuis juillet 2008, il a 15 classeurs de trois à 60 albums rassemblant entre deux et 133 photographies ou vidéos, totalisant 5564 documents; elle a 1300 abonnés et 117 abonnements. L'établissement 3 est membre depuis juin 2009, elle a onze classeurs d'un à 38 albums rassemblant entre huit et 122 photographies, totalisant 5273 photographies; elle a 1000 abonnés et 23 abonnements.

Nous avons décrit dans cette section les milieux où nous avons effectué notre collecte. À partir de cette description, nous remarquons que leur mission respective est similaire, bien qu'elle porte sur des territoires d'étendue et d'administrations différentes. Nous avons aussi détaillé les éléments présents dans une notice ; ceci révèle certaines différences entre les établissements dans la manière de décrire et d'indexer les documents d'archives. Les établissements ne présentent pas tout à fait le même profil dans *Flickr* : ils sont tous les trois membres PRO, mais depuis plus ou moins de temps et offrent ainsi un nombre de documents plus ou moins grand. Malgré quelques divergences, les établissements montrent des similitudes qui permettent de regrouper leurs données en un seul corpus. La prochaine section décrit la procédure de constitution du corpus.

4.1.2. Procédure de constitution du corpus

Cette section comprend la justification de l'analyse par corpus dans notre thèse, la description des sources et des critères d'inclusion dans notre corpus, ainsi que l'exposé des procédures de collecte.

4.1.2.1. Justification de l'analyse par corpus

Notre recherche sur le VATAP se concentre sur l'aspect sémantique du vocabulaire. Polguère (2008, 60) présente trois méthodes pour décrire le sens des mots : l'introspection, l'enquête linguistique et le corpus linguistique. L'introspection est une méthode courante en linguistique, elle apporte des données inatteignables autrement. Elle est parfois critiquée car trop subjective et sans réel moyen de vérification des résultats; la méthode ne satisfait pas les critères de qualité habituellement utilisés en sciences sociales, ce n'est donc pas notre premier choix. La deuxième méthode, l'enquête linguistique, repose sur la perception des locuteurs d'une langue et leur vision du monde. Elle est également subjective, mais le nombre de participants peut assurer une certaine représentativité et validité scientifique. Elle est particulièrement indiquée pour rendre compte d'une norme linguistique, ce qui n'est pas notre objectif ici. La troisième méthode a l'avantage d'être stable dans le temps, moins subjective et plus paramétrable, il s'agit de l'étude sur corpus linguistique. C'est cette dernière que nous avons retenue pour cette recherche.

4.1.2.2. Sources et types de données

Notre corpus est constitué à partir des documents auxquels les archivistes des services d'archives participants ont accepté de nous donner accès et de documents librement accessibles en ligne. Nous avions prévu qu'il soit formé de termes émanant des quatre sources suivantes :

1) les courriels d'usagers à la référence;
2) les notices descriptives;
3) les index contenant les termes d'indexation professionnelle;
4) les nuages d'étiquettes, les termes d'indexation collaborative.

Le Tableau IV présente les quatre sources de données envisagées pour notre corpus sous l'angle de l'étendue linguistique (mots isolés et phrases) et sous l'angle de l'auteur du discours (usager ou archiviste).

Tableau IV. VATAP : 4 sources

Auteur du discours / Étendue linguistique	**Vocabulaire employé dans les instruments de recherche**	**Vocabulaire des usagers**
Phrases	Notices descriptives : Termes de description	Courriels d'usagers à la référence : Termes de recherche
Mots isolés	Index : Termes d'indexation professionnelle	Nuages d'étiquettes : Étiquettes ou termes d'indexation collaborative

Les termes issus de ces sources appartiennent à trois catégories : les termes de recherche, les termes de description, les termes d'indexation (indexation professionnelle et indexation collaborative ou étiquettes). Cet ensemble de termes constitue le vocabulaire pour l'accès thématique aux documents d'archives patrimoniales (VATAP).

Afin d'étudier le vocabulaire que les archivistes ont employé dans les instruments de recherche, nous avons collecté les notices descriptives des fonds ou parties de fonds jusqu'à la pièce recommandés par les archivistes de référence ainsi que les termes d'indexation thématique qui leur étaient rattachés. Le vocabulaire employé dans les instruments de recherche émane de l'archiviste ou des archivistes qui ont rédigé ces instruments.

Afin d'étudier le vocabulaire des usagers, nous avons collecté les questions de recherche que les usagers ont librement formulées dans un courriel envoyé à l'archiviste de référence. Nous avons aussi exploré les étiquettes, qui sont des termes d'indexation librement générés par des usagers; elles sont issues de l'indexation collaborative. Après avoir examiné les premières données collectées pour évaluer si le projet était réalisable, il nous est apparu que le courriel de l'usager était en réalité un échange de courriels entre un usager et un archiviste de référence menant un entretien de référence en différé. Nous avons pris en compte les expressions linguistiques employées par l'archiviste de référence dans sa réponse en tant qu'indice de la médiation sémantique qu'il opère entre l'usager et les documents d'archives (voir Guitard 2018); nous considérons que ces données textuelles appartiennent au vocabulaire des archivistes tout en étant des termes de recherche. Nous revenons sur ce point dans l'analyse des résultats.

Les étiquettes apparaissent dans des fiches descriptives. Les fiches descriptives correspondent à la description, généralement écourtée, des images de documents d'archives diffusés dans *Flickr* et auxquelles sont rattachées les étiquettes. Pour susciter un premier repérage et la collaboration des usagers à l'indexation, les archivistes attribuent eux-mêmes quelques étiquettes. Puisqu'il n'y a pas d'inscription sur la plate-forme *Flickr* qui permette d'identifier l'usager, il ne nous est pas possible de savoir si les étiquettes ont été ajoutées par l'archiviste ou un usager, alors nous aurions pu les considérer toutes comme appartenant au vocabulaire des usagers. Cependant, comme nous l'expliquons plus loin (voir section 4.1.2.4), au final, nous n'avons pas pu intégrer d'étiquettes à notre corpus.

4.1.2.3. Critères d'inclusion

Nous limitons notre corpus en fonction des critères suivants que nous expliquons ci-après :

a. La langue (pour les questions de recherche, les notices descriptives, les termes d'indexation et les étiquettes);

b. La nature grammaticale (pour les questions de référence et les notices descriptives);

c. Le temps (pour les questions de référence);

d. La lisibilité (pour les questions de recherche, les notices descriptives, les termes d'indexation);

e. L'origine humaine des termes.

a. La langue

Pour pouvoir procéder à une analyse linguistique du sens des mots en français, il est important sinon crucial que les personnes qui ont produit ces segments de langue maîtrisent bien la langue française. Ces segments ne devraient pas avoir été traduits, au risque de briser l'élan naturel de la langue française (Rey 1972, 8). Ne pouvant nous assurer de la compétence des locuteurs, nous nous sommes basée sur la bonne formation des termes en langue française pour vérifier ce critère, à partir de notre propre compétence de locutrice native du français. En outre, au cours du prétraitement (voir section 4.2.5) et de l'analyse (voir notamment section 5.1.1), nous avons eu recours à des ouvrages de référence : dictionnaires et ressources terminologiques (voir section 4.2.5.3).

Au sein de la francophonie, nous circonscrivons notre collecte aux établissements situés dans la région du français laurentien afin de conserver une certaine unité dans l'emploi des mots au sein de la même langue (variation topolectale).

b. La nature grammaticale

Les questions de référence formulant une recherche thématique doivent inclure des noms communs qui expriment le sujet et non seulement des noms propres de personnes, de lieux ou d'événements et des dates. Les notices descriptives emploient généralement des noms communs pour exprimer le sujet des documents, d'après les formulations recommandées dans les *Règles pour la description des documents d'archives* (*RDDA)* (p. ex., *le fonds traite de* + nom commun). Mais si des notices ne respectent pas ce principe, alors elles sont écartées. Le mot principal (appelé en linguistique *la tête* ou *le noyau*) des termes d'indexation et des étiquettes appartient traditionnellement à la catégorie grammaticale du nom. Le nom, aussi appelé *substantif* (dénommé et défini ainsi par Aristote, voir p. ex. Bécherel 1994), désigne généralement la substance de la matière, c'est-à-dire le sujet. En outre, les termes constitués uniquement de chiffres ou de nombres, y compris les dates, bien qu'ils puissent porter un

sujet, sont exclus; p. ex. *1967* (année de l'exposition universelle tenue à Montréal en 1967 sur le thème « Terre des hommes »).

c. Le temps

Les questions de référence que nous avons collectées sont datées de janvier 2015 à avril 2016 pour deux établissements et début 2015 à novembre 2015 pour l'établissement 3. Ceci a permis de limiter la variation temporelle dans le corpus. Une période de moins de cinq ans nous semble suffisante pour rendre compte du vocabulaire actuellement en usage. En outre, nous voulons étudier l'écart sémantique présent à l'heure actuelle, pour des usagers contemporains qui consultent des instruments de recherche existants et actuellement utilisés. Ainsi, la date de création des instruments de recherche n'a pas d'importance dans notre recherche; Oliver, Jamieson et Daniel soulignent que les archivistes utilisent divers outils associés à diverses technologies et que ces outils évoluent dans un continuum relatif au processus de création des documents (2017, 4, n. t.). Finalement, la limite de temps permet de restreindre le volume des courriels d'usagers à la référence à une taille raisonnable pour procéder à une analyse humaine. Par exemple, il ne nous semble pas raisonnable d'envisager de traiter humainement davantage que les 10 000 questions de référence par courriel annuelles de l'établissement 1. Les étiquettes seront nécessairement récentes, la technologie du Web 2.0 étant encore jeune (depuis 2004-2005 selon Le Deuff 2006 et Voss 2007) et l'adoption de cette technologie par les services d'archives encore plus[47]. En effet, les services d'archives participants gèrent des albums *Flickr* depuis peu (voir section 4.1.1.2). Les notices descriptives et les termes d'indexation professionnelle pouvaient éventuellement être anciens, mais ils étaient inclus dans des instruments de recherche qui étaient au moment de la collecte à la disposition de l'usager; c'est pourquoi nous n'avons pas émis de limite de temps pour ces éléments.

[47] Par exemple, Rorissa 2010 et Angel 2012 font partie des premières études incluant des documents d'archives étiquetés, Skinner 2014 fait un état de l'art sur la question dans les institutions patrimoniales.

d. La lisibilité

Pour des raisons de lisibilité, nous excluons les éléments manuscrits de notre corpus afin d'éviter toute erreur de déchiffrage. Certains instruments de recherche détenus par les trois établissements participants sont manuscrits. Le fait de privilégier les questions d'usagers par courriel[48] rend le critère de lisibilité de l'écriture manuscrite non pertinent pour cette source de données. Mais dans certains services d'archives les usagers remplissent manuellement, sur place, un formulaire d'entrée mentionnant l'objet de leur recherche, parfois sous forme de question. Les étiquettes ne sont pas touchées par ce critère puisqu'elles sont nécessairement tapées à la saisie.

e. L'origine humaine des termes

Si les questions d'usagers et les réponses d'archivistes – *i. e.* les courriels entre les usagers et la référence – sont naturellement créées par des êtres humains, cela est moins évident pour les termes d'indexation qui peuvent être automatiquement extraits de diverses sources textuelles (p. ex. de la fiche descriptive pour les étiquettes, avec le recours à des dictionnaires ou non comme filtres, voir indexation automatique dans section 3.4.3.2 Archivistes et langage documentaire).

Tableau V. Pertinence des limites en fonction des sources du corpus

Source / Limite	Courriels entre les usagers et la référence	Notices descriptives	Termes d'indexation	Fiches descriptives et étiquettes
Langue	1	1	1	1
Nature grammaticale	1	1	1	1
Temps	1	0	0	0
Lisibilité	0	1	1	0
Origine humaine	0	0	1	1

Le Tableau V indique si le critère présenté dans la colonne de gauche est pertinent (1) ou non (0) pour chacune des sources de notre corpus. Ces critères d'inclusion ont été jumelés à nos procédures de collecte pour constituer notre corpus.

[48] Martin indiquait déjà en 2001 que de 1995 à 1999 le courriel était devenu le mode de communication préféré des usagers à la référence.

4.1.2.4. Procédure de collecte

En fonction des critères indiqués dans la section précédente, nous exposons ici la manière dont nous avons collecté les données de notre corpus, matériau de base de notre recherche. Notre collecte s'est effectuée en plusieurs temps. Nous avons envisagé trois procédures de collecte (voir Annexe 2 – Procédures de collecte envisagées), mais n'en avons retenu en définitive qu'une seule que nous exposons ici. Ensuite, nous expliquons ce qui nous a amenée à avoir un second corpus et présentons sa méthode de constitution.

Procédure effectuée

Les courriels d'usagers collectés dans les boîtes de gestion de courriels des trois services d'archives participants nous ont permis de collecter deux autres types de données : les notices descriptives et les termes d'indexation. En effet, les réponses des archivistes et parfois les usagers eux-mêmes indiquaient des cotes archivistiques. Les notices descriptives ont été récupérées en ligne, à partir du catalogue public des trois établissements participants. Les notices indiquaient les termes d'indexation qui leur étaient rattachés. Ainsi, nous avons sélectionné les notices descriptives accessibles en ligne auxquelles menaient des courriels entre les usagers et la référence et parmi celles-ci les notices qui ont été indexées. Cette dernière partie de la collecte en deux étapes nous aura été utile pour constituer un second corpus, comme nous le verrons plus loin. Finalement, les cotes nous auraient permis de chercher dans *Flickr* les éventuelles étiquettes et fiches descriptives correspondant à l'unité archivistique mentionnée par l'archiviste.

Ainsi, théoriquement (voir Figure 9), nous devions pouvoir collecter des fiches descriptives et des étiquettes à partir des termes d'indexation présents dans les notices, mais notre corpus n'a pas présenté ce cas de figure, puisqu'aucun des termes des fiches descriptives ni aucune étiquette trouvés ne correspondaient à nos critères de recherche (voir section 4.1.2.3). En effet, les étiquettes collectées étaient des dates (p. ex. *1967*), des noms propres (p. ex. *Pères-du-Saint-Esprit, Terre des hommes*), des expressions en anglais (p. ex. *Mothers of Invention*) ou bien générées par la machine et ainsi indiquées en grisé dans la fiche descriptive (p. ex. *noir et blanc, monochrome, foule, personnes, auditorium*).

Afin que les trois établissements participant à notre étude soient représentés dans nos données et afin d'élargir notre corpus que nous jugions trop restreint bien qu'obtenu par une sélection rigoureuse, nous avons décidé de constituer un second corpus qui contiendrait des questions d'usagers, des réponses d'archivistes et des notices descriptives, mais pas de termes d'indexation professionnelle.

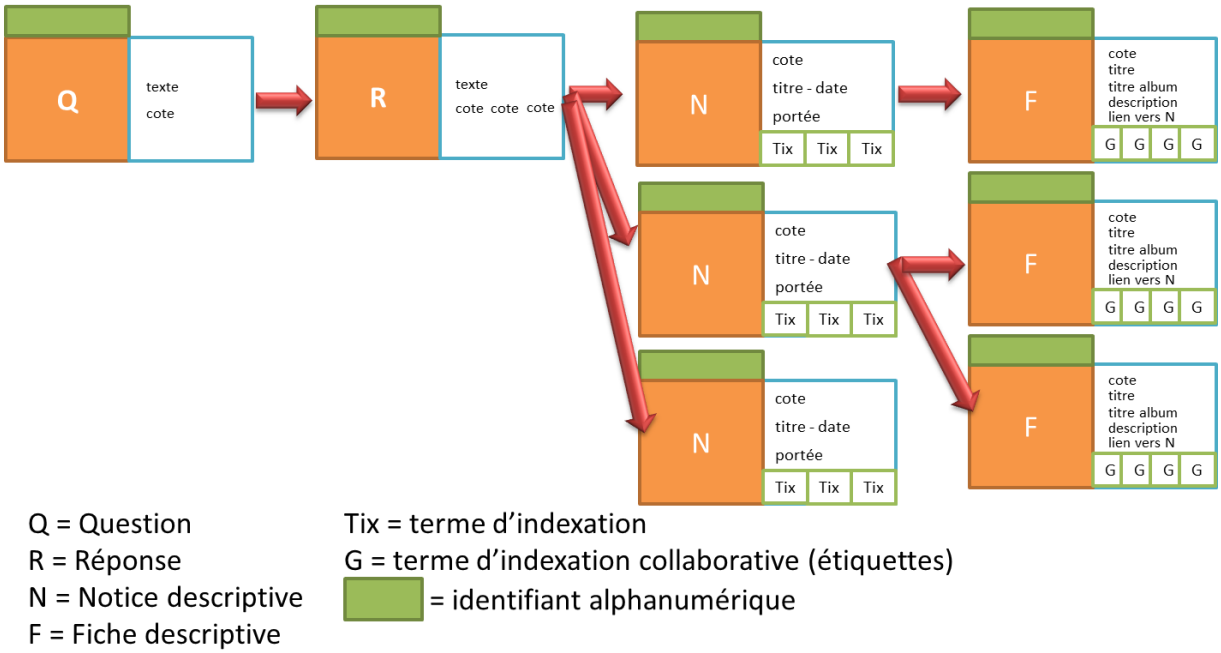

Figure 9. Procédure de collecte du VATAP

Pour ce faire, nous avons inclus les questions d'usagers provenant de deux établissements ayant une réponse positive (voir section 4.1.3.1) et des cotes archivistiques repérables dans le catalogue en ligne, soit six questions de l'établissement 3 et 14 de l'établissement 2. Nous n'avons pas appliqué d'autre critère de sélection. Comme l'indique la Figure 9, les questions d'usagers sont à la source de notre collecte. Le Tableau VI indique le nombre de questions d'usagers selon le service d'archives dont elles émanent, pour le corpus 1 (QRNTix) et le corpus 2 (QRN).

Tableau VI. Nombre de questions d'usagers selon l'établissement dans les 2 corpus

	Établissement 1	Établissement 2	Établissement 3	Total
Corpus 1	8	2	0	10
Corpus 2	0	14	6	20
Total	8	16	6	30

Le corpus 1 et le corpus 2 forment notre corpus. Sa taille semble modeste, mais comme nous le verrons dans la section suivante, le nombre d'expressions linguistiques porteuses de sujet (thémanymes) est satisfaisant pour une analyse entièrement humaine et manuelle.

4.1.3. Description des données

Dans cette section, nous décrivons les données examinées[49] et collectées, en distinguant les difficultés rencontrées lors de la collecte selon le type de données.

Nos données ont différentes caractéristiques (quantité et état) selon le milieu dont elles proviennent (trois établissements), leur source (quatre sources : Q, R, N, Tix) et leur type (deux types : textes et termes). Les quatre sources de données ont évolué lors de la collecte, comme précisé dans la section 4.1.2.4.

Nous présentons les courriels examinés puis retenus après avoir été triés selon nos critères de sélection énoncés plus haut (voir section 4.1.2.3). Les caractéristiques du corpus 1 et du corpus 2 sont similaires; nous décrivons d'abord le corpus 1, puis le corpus 2.

4.1.3.1. Les questions d'usagers (Q) et les réponses d'archivistes (R)

À la base de notre collecte, les questions d'usagers (Q) sont essentielles à notre recherche. Elles expriment le besoin des usagers dans leurs mots. Les questions d'usagers proviennent de plusieurs services d'archives, aux missions semblables sur des territoires différents (voir section 4.1.1.2). Nous avons fait une sélection des questions envoyées par courriel à partir des données qui étaient mises à notre disposition. Nous présentons ici les motifs de rejet des courriels.

Certains courriels de questions d'usagers n'étaient pas complets. Il manquait parfois les pièces jointes ou bien celles-ci étaient dans des liens *Wetransfer* – un outil d'envoi de fichiers volumineux – qui ne sont valides que sept jours. Dans les données conservées par le service

[49] Bien que notre corpus ne contienne pas d'étiquettes, nous avons souhaité maintenir la description de l'état des étiquettes que nous avons examinées en grand nombre pour le bénéfice des lecteurs intéressés par ce sujet.

d’archives, il manquait parfois la question d’usager à l’origine de l’échange; dans ce cas-là, l’échange n’était pas conservé.

Certaines questions d’usager sont en double, lorsqu’il y a l’oubli de joindre une pièce ou quand deux équipes d’archivistes travaillent conjointement sur la même question d’usager. Les doublons ont été éliminés.

Nous avons procédé à la sélection des questions d’usagers à teneur thématique dont le sujet était exprimé par des noms communs. Nous avons formalisé ce processus de prise de décision sous la forme d’un arbre décisionnel (voir Figure 10 *Arbre décisionnel pour la sélection des questions d’usager à teneur thématique*).

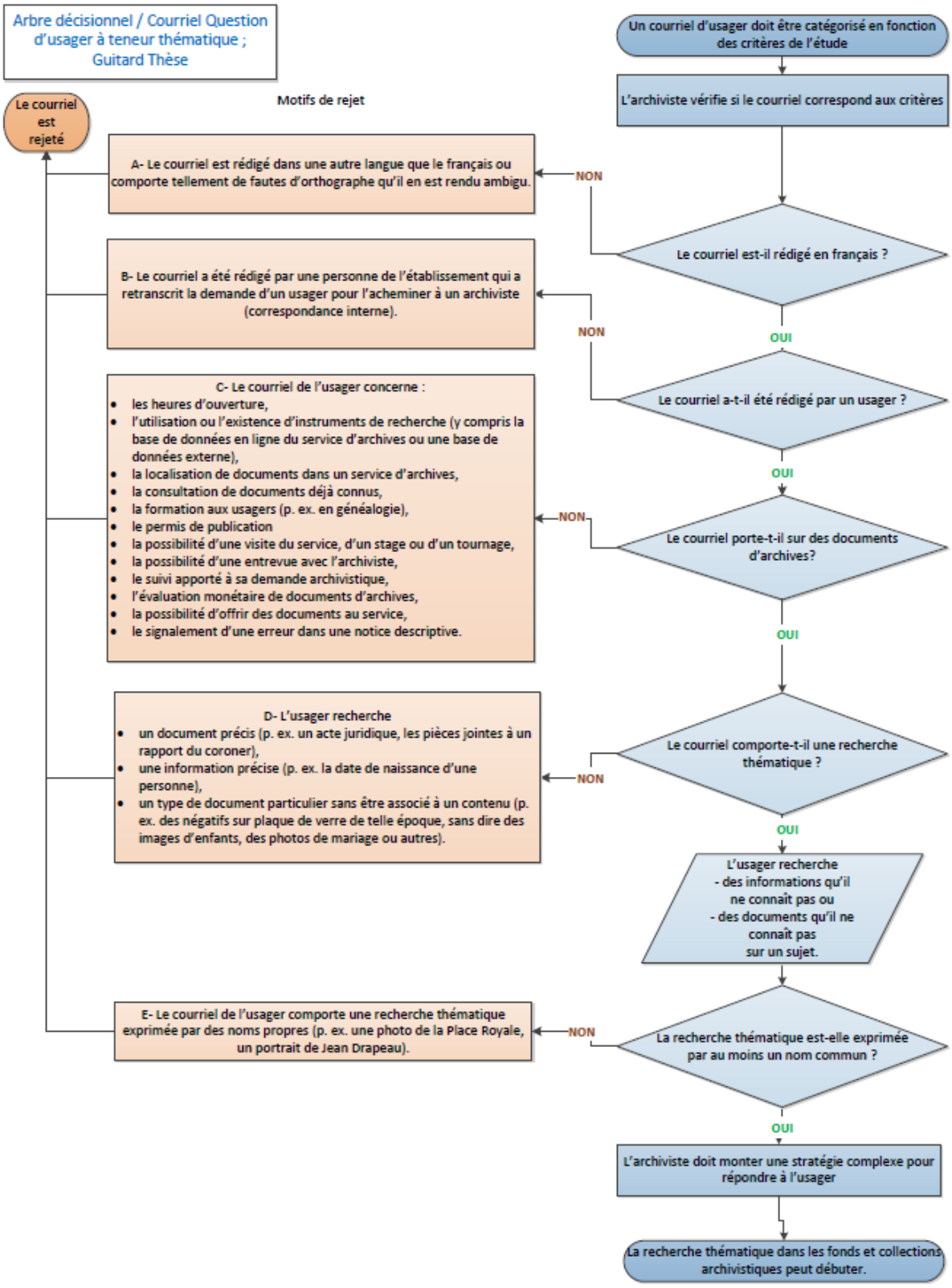

Figure 10. Arbre décisionnel pour la sélection des questions d'usager à teneur thématique

Pour les courriels restants, parfois (28 cas), nous n'avons pas eu accès à la réponse de l'archiviste ou encore le contact avec l'usager a été fait par téléphone et seule une note était inscrite en regard de la question dans le fichier de gestion des questions de référence. Parfois (31 cas), la réponse ne contenait pas de cote archivistique. Dans ces cas, les courriels (questions d'usagers et réponses d'archivistes) ont été exclus. Par exemple, les réponses des archivistes de l'établissement 2 contiennent des liens url vers des notices descriptives, parfois des numéros d'identification de notice, parfois des mots clés ou des numéros d'instruments de recherche, parfois le titre d'un document (difficile à retrouver par la suite), parfois une cote archivistique.

Ainsi, pour être retenue dans notre corpus, chaque question à teneur thématique devait inclure une réponse ayant au moins une cote archivistique afin de représenter les éléments de chaîne communicationnelle des usagers aux documents d'archives (voir section 3.2.1). Mais la réponse d'archiviste n'inclut pas nécessairement ni uniquement des cotes de fonds à consulter ni des liens vers des résultats de recherche. L'archiviste indique d'abord des conseils de recherche en expliquant sa démarche pour former l'usager à la recherche; il exprime son raisonnement, indique les mots clé utilisés et dans quels fonds, il indique aussi ses résultats négatifs pour éviter que l'usager ne refasse la recherche inutilement. La réponse contient aussi des liens vers des instruments de recherche, des références de sources publiées, des liens vers des collections numériques ou des bases de données hébergées par d'autres organismes. Ces éléments sont conformes en tout point à la description de l'interaction de référence décrite par Duff et Yakel (2017, 199-204).

Pour chaque établissement, nous indiquons les quantités de données qui ont été accessibles et qui correspondaient à nos critères après lecture et évaluation. Nous avons appliqué notre arbre de codage plusieurs fois sur les mêmes ensembles de questions; nous avons calculé le pourcentage de questions correspondant à nos critères après le dernier tri, par rapport au nombre initial de questions examinées.

Établissement 1

Nous avons eu accès aux courriels de référence de diverses manières pour cet établissement : soit nous avons eu directement accès aux boîtes de courriels et avons procédé à

nos recherches dans les courriels reçus et envoyés, soit nous avons consulté des impressions des courriels traités. Nous avons ainsi consulté 2281 courriels, nous en avons anonymisé 449, nous en avons retenu 54 (1,9%) par un dernier tri effectué à partir de l'arbre décisionnel, c'est-à-dire la dernière évaluation du courriel examiné pour savoir si la question d'usager répond à nos critères. En effet, nous avons mené plusieurs tris sur le même corpus à divers intervalles de temps (en laissant passer entre deux et trois mois) afin de procéder à une analyse plus rigoureuse et pour être certaine que l'ensemble des données ait été traité de la même manière.

Cet établissement distingue la référence pour les documents textuels et pour les autres types de documents. Le responsable de la référence du second type a mis en place un tableau de suivi à partir de 2013, plus systématiquement à partir de l'année financière 2014-2015 et de manière tout à fait régulière pour 2015-2016. Ce tableau attribue un numéro interne à la demande d'usager, le nom de l'usager, la date de réception de la demande, le nombre de jours de traitement, l'archiviste auquel la demande est attribuée, le caractère urgent ou non, la technologie par la laquelle la demande a été reçue et, ce qui nous intéresse tout particulièrement, la nature de la demande, à savoir s'il s'agit d'une demande de licence, de reproduction, de mention ou d'une recherche. Une même demande peut être de plusieurs natures. Ainsi, par exemple, en janvier 2015, 57 demandes ont été reçues dont 24 nécessitaient une recherche. Précisons que ce ne sont pas toutes les demandes qui nécessitent un travail de recherche de l'archiviste qui sont des questions thématiques, mais elles y sont assurément incluses. Nous n'avons examiné que 373 courriels « recherche » plutôt que les 907 courriels reçus pour la référence de ce second type.

Établissement 2

Nous avons reçu un fichier incluant toutes les demandes de référence nécessitant la réponse d'un archiviste. Nous avons anonymisé les 518 courriels, sans les trier. Il est resté 59 courriels (11%) après le dernier tri.

Établissement 3

Aux côtés de l'archiviste, nous avons parcouru de nombreux courriels, dans deux boîtes de courriels d'archivistes de référence. Nous n'avons pas le nombre exact de courriels examinés (le tri a duré 1h30 env.), mais avec l'archiviste, nous estimons le nombre à environ

300. Nous avons retenu 29 courriels. Nous les avons rendus anonymes puis retriés à l'université. Il est resté douze courriels (4%) après le dernier tri.

Sur les 3729 courriels examinés, nous atteignons donc un total de 125 questions d'usagers répondant à nos critères (3,4%) (voir Tableau VII). Nous leur avons attribué un identifiant alphanumérique dont le nombre s'incrémente et qui est suivi de la lettre *q* pour les questions : a001q à a125q.

Tableau VII. Répartition des courriels examinés et retenus par établissement

Établissements / Courriels (Q-R)	**Examinés**	**Retenus (%)**
Établissement 1	2911	54 (1,9%)
Établissement 2	518	59 (11%)
Établissement 3	300	12 (4%)
Total	**3729**	**125 (3,4%)**

Parmi les 125 courriels d'usagers, 98 présentent une réponse d'archiviste à laquelle nous avons eu accès; les réponses héritent du même code que la question à la dernière lettre près qui est remplacée par un *r*, p. ex. a001r. Nous ajoutons une lettre (a, b, c) à la fin du code s'il y a plusieurs échanges ou si l'archiviste répond à deux questions posées dans deux courriels envoyés par le même usager sur le même sujet, ce qui arrive rarement. Il va de soi que la réponse de l'archiviste répond à la question de l'usager et que le sujet de la question est le sujet de la réponse.

4.1.3.2. Les notices descriptives (N)

Lorsque la réponse de l'archiviste incluait des cotes archivistiques, nous avons collecté les notices descriptives associées; ceci, dans le but d'avoir une filiation sémantique dans le vocabulaire employé pour l'accès thématique aux archives patrimoniales (VATAP). La filiation sémantique (voir Figure 11) est un terme que nous employons pour désigner le repérage des mêmes éléments de sens d'une source à l'autre dans le sens de la chaîne communicationnelle des usagers aux documents d'archives : de la question d'un usager (Q), à la réponse d'un archiviste de référence (R), à la notice descriptive (N) jusqu'au terme d'indexation (Tix).

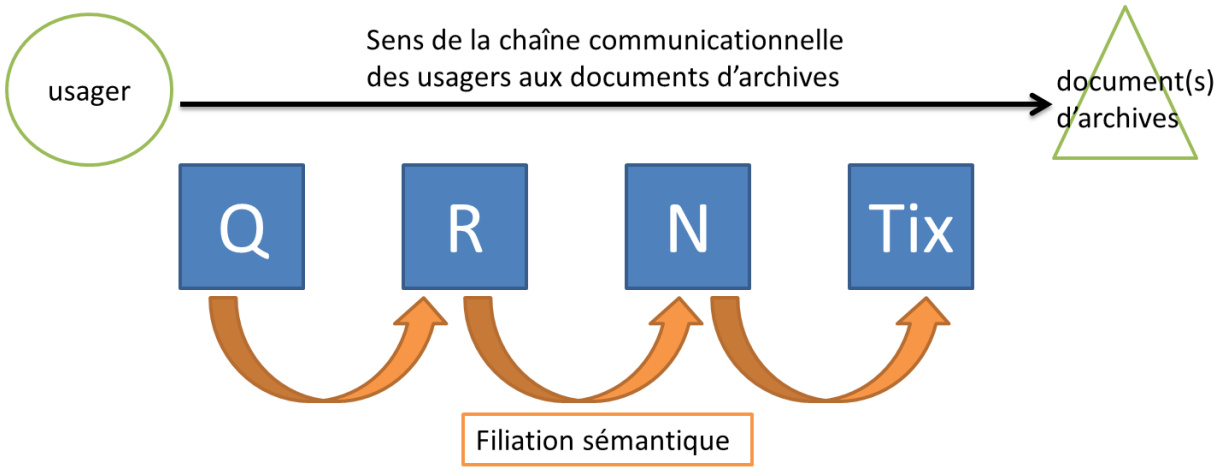

Figure 11. Filiation sémantique dans une séquence (QRNTix)

La filiation sémantique permet la filiation thématique qui est le repérage des expressions linguistiques porteuses de sujet (thémanymes) d'une source à l'autre dans le sens de la chaîne communicationnelle. Nous présentons un exemple de la filiation sémantique. Voici les trois portions de texte des sources concernées par une question d'usager (Q, R, N) :

> Question de l'usager (Q) : l'important héritage patrimonial de la Brasserie Molson de la rue Notre-Dame Brasserie Molson de la rue Notre-Dame photos d'archives des lieux, à différentes époques
>
> Réponse de l'archiviste (R) : les photos que nous possédons où l'on voit la brasserie Molson : 1936 : 1942 : 1966 : Début des années 1980 :
>
> Notice descriptive du document d'archives (N) : *Titre de la pièce.* Brasserie Molson : Rue Notre-Dame. 1 avril 1936. *Portée et contenu.* Photographie de la brasserie Molson (1590, rue Notre-Dame Est) On y aperçoit la rue Notre-Dame vers l'est et la structure du pont Jacques-Cartier

Nous avons repéré dans ces trois portions de texte des thémanymes qui se rapportent les uns aux autres, par exemple, les variantes *Brasserie Molson* et *Brasserie Molson* (Q), *brasserie Molson* (R) et *Brasserie Molson* et *brasserie Molson* (N). Des expressions linguistiques porteuses de sujet, très similaires hormis la casse, assurent la filiation sémantique entre les sources et permettent aux deux interlocuteurs – l'usager et l'archiviste – de continuer à parler de la même chose. D'autres éléments participent du sujet, par exemple, ici, les dates : *à différentes époques* (Q), *1936 :*, *1942 :*, *1966 :*, *Début des années 1980* (R) et *1 avril 1936* (N). Dans certains cas, la filiation sémantique est facile à établir formellement (p. ex. reprise

de termes exacts); dans d'autres cas, elle nécessite une interprétation (p. ex. « à différentes époques » / « 1936 »). Nous verrons par la suite des cas plus complexes.

Dans notre recherche, nous avons repéré la cote archivistique soit littéralement dans la réponse d'archiviste, soit dans la liste des résultats de recherche dont le lien de résultats de recherche avait été copié par l'archiviste dans sa réponse à l'usager. Nous n'avons pas exploré les résultats de recherche si le nombre de résultats dépassait la centaine, ce qui est arrivé à quelques reprises. Les notices devaient en outre présenter au moins un terme d'indexation répondant à nos critères pour être retenues. Dans une notice descriptive, nous trouvons la cote, la description physique et la description intellectuelle des documents d'archives. Cette dernière est représentée par le titre, la section *portée* de la notice descriptive et les termes d'indexation qui lui sont rattachés. L'accès thématique repose sur la description intellectuelle des documents, c'est pourquoi nous nous attachons à décrire en détail les notices descriptives. Au niveau de la pièce, l'instrument de recherche indique parfois le titre propre de la pièce qui peut se présenter dans une autre langue que le français (en anglais par exemple); à ce moment-là, le titre n'est pas conservé dans nos données puisque la langue n'est pas le français (voir section 4.1.2.3a). La cote rend unique l'unité archivistique dans un service d'archives. Elle permet de situer son niveau hiérarchique dans un fonds ou d'indiquer sa place dans une collection. Elle fait partie des éléments essentiels de la description selon la norme *ISAD(G)* (2000, 9, section I.12). Dans notre recherche, elle est le fil d'Ariane entre les réponses d'archivistes et les notices descriptives, l'élément concret qui permet d'assurer une correspondance entre le sujet des questions d'usagers tel que compris par l'archiviste et les ressources archivistiques correspondantes identifiées par la cote et repérées par l'archiviste.

Comme vu précédemment (voir section 3.4.2), il existe une norme canadienne sur la description archivistique, les *Règles pour la description des documents d'archives* (les *RDDA*), et la *Norme générale et internationale sur la description archivistique*, *ISAD(G)* développée par le Conseil international des archives. Mais chaque établissement adapte la norme choisie pour répondre aux particularités de ses documents et de son public. Chaque établissement compose également en fonction des choix technologiques qui ont été faits auparavant et des ressources allouées à la description par le passé et de nos jours. Ainsi, les notices descriptives des trois services d'archives participants sont différentes mais recèlent les

éléments thématiques nécessaires aux usagers, aux archivistes et à notre étude : le titre et la portée.

Nous avons recueilli 361 cotes archivistiques directement dans les réponses d'archivistes aux questions d'usagers ou dans les liens de résultats qu'ils avaient copiés dans leur réponse. Pour chaque notice descriptive, nous relevons le cas échéant la présence d'une description abrégée sur *Flickr*, que nous appelons une fiche descriptive. Ainsi, nous avons associé à ces cotes le titre et la date de l'unité archivistique, sa portée et contenu en français, les éventuels termes d'indexation en français et, le cas échant, un lien vers la fiche descriptive dans *Flickr* (dans un album ou non). Nous obtenons 237 notices descriptives qui contiennent une portée et contenu, 155 notices qui contiennent des termes d'indexation, six notices dont l'objet se trouve dans un album *Flickr*. Parmi les réponses, certaines cotes étaient répétées. Nous avons trouvé trois doublons et un triplet. Les notices acceptables selon nos critères sont au nombre de 155. Les termes d'indexation subissent à leur tour une épuration et sont sélectionnés en fonction de nos critères, essentiellement la langue et le fait que le nom principal ou tête soit un nom commun; nous obtenons ainsi 56 notices auxquelles sont associés des termes d'indexation qui satisfont nos critères (voir Tableau VIII).

Tableau VIII. Nombre de notices descriptives

Type de données	Nombre	Pourcentage
Cotes présentes dans les réponses	361	100%
Notices descriptives en français, avec portée	237	65,65%
Notices ou fiches descriptives qui contiennent des termes d'indexation	155	42,94%
Notices qui contiennent des termes d'indexation selon nos critères	56	15,51%

Par ailleurs, dans le corpus 1, les réponses d'archivistes présentent d'une à dix cotes archivistiques qui respectent nos critères (c'est-à-dire dont les notices descriptives sont en français, avec une portée et contenant des termes d'indexation) Quant au corpus 2, il comprend des réponses d'archivistes qui comptent d'une à huit cotes qui respectent nos critères.

4.1.3.3. Les termes d'indexation (Tix)

Nous nous sommes assurée que chaque notice sélectionnée contienne minimalement un terme d'indexation dont la tête était un nom commun et avons éliminé les termes d'indexation à la syntaxe complexe (vedettes-matière, termes suivis de qualificatifs, énumérations). Pour les termes d'indexation qui semblent être un agglomérat de deux termes d'indexation, nous séparons les deux parties et en faisons deux termes (voir Tableau IX).

Tableau IX. Termes d'indexation (dédoublement)

Terme d'indexation collecté	Terme d'indexation dans notre recherche
Carpathe-russes - Tchécoslovaquie, 1930 Société de nations, 1919[sic]	*Carpathe-russes - Tchécoslovaquie, 1930*
	Société de nations, 1919
John Leard - Testament, 1836 Grande-Bretagne. Imperial War Cabinet, 1919	*John Leard - Testament, 1836*
	Grande-Bretagne. Imperial War Cabinet, 1919

Une notice descriptive est ainsi passée de 42 termes d'indexation dans la notice à 67 expressions linguistiques considérées comme des termes d'indexation dans notre recherche.

Certains termes présentaient un amalgame de français et d'anglais en leur sein (p. ex. *Hôpitaux militaires – England*, *Jews – France*). Nous les avons exclus. Certains termes d'indexation en français présentaient des fautes d'orthographe (p. ex. *Tariff douanier - Canada, 1878-1892, Parti porgressiste-conservateur - Canada*), nous les avons exclus aussi.

Nous avons exclu également les termes d'indexation qui comportent des parenthèses – appelés qualificatifs. Le contenu de la parenthèse peut remplir plusieurs rôles, essentiellement la désambigüisation en cas de polysémie (p. ex. *Parlement (édifices)*) ou la clarification de l'expression linguistique utilisée comme terme d'indexation (p. ex. *Noronic (bateau à vapeur)*). L'étude de ces compléments ne fait pas partie de nos objectifs; les termes avec compléments (n=19) ont été exclus.

Plus précisément dans notre corpus, les notices descriptives sélectionnées indiquent généralement un ou plus plusieurs termes d'indexation; les 56 notices descriptives

sélectionnées présentent d'un à 127 termes d'indexation par notice pour totaliser 1264 termes d'indexation. Parmi ceux-ci, 968 sont en français. Parmi ces termes d'indexation, certains revêtent la forme de vedettes-matière[50] ou sont des termes suivis d'un qualificateur entre parenthèses (n=510). Ces termes ont une syntaxe complexe difficile à appréhender sous forme d'unité (p. ex. *Clergé - Costume, [17-], [18-]* ou *Habitations – Démolition* ou encore *R-100 (Dirigeable)*). Nous avons aussi identifié les termes d'indexation dont le mot principal – i. e. la tête – est un nom commun (n=571) afin d'exclure les noms propres pris en charge par les règles d'indexation onomastique (p. ex. *Loppinot, Joseph-Chrysostome, né 1688*). Les termes d'indexation en français qui n'ont pas une syntaxe complexe et qui contiennent un nom commun sont alors au nombre de 458 (p. ex. *Anciens de l'université d'Ottawa*, *Pré-arrangements funéraires* ou *Flottage*) (voir Tableau X).

Tableau X. Termes d'indexation (sélection)

Type de terme d'indexation	Nombre
collectés	1264
en français	968
dont le mot principal est un nom commun	458

Ces 458 termes d'indexation sont attribués à des notices descriptives dont la cote est citée dans la réponse d'un archiviste à une question d'usager. Ce nombre étant trop important pour une analyse humaine à effectuer dans un délai raisonnable, nous avons examiné le niveau de description auquel ces termes étaient rattachés. De ces 458 termes d'indexation, 417 termes appartiennent à des notices descriptives du niveau du fonds ou de la collection, cinq termes à des notices de la série, cinq termes pour la sous-série, six termes pour la sous-sous-série, neuf termes pour le dossier et 16 termes pour la pièce (voir Figure 12).

[50] Nous rappelons la définition de ce terme. « Vedette-matière : Terme d'indexation constitué d'une tête de vedette utilisée seule ou d'une tête de vedette complétée par une ou plusieurs subdivisions, et dont la fonction est de représenter de façon complète et précise le sujet dont traite un document ». (Hudon 2013, 281)

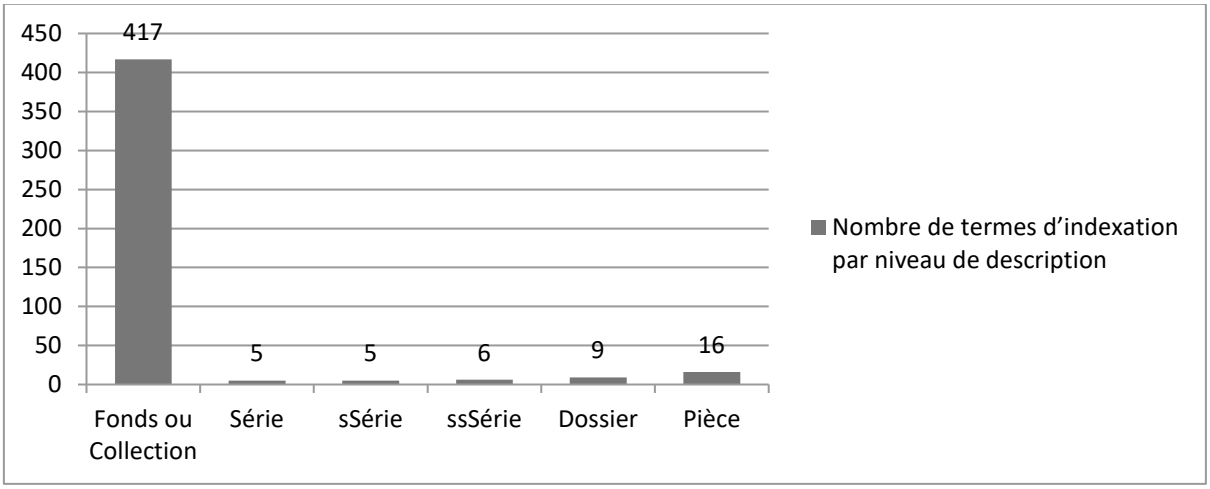

Figure 12. Termes d'indexation et niveau de description

De manière tout à fait attendue, le niveau du fonds était surreprésenté dans notre corpus, niveau où les notices comptent *a priori* le plus grand nombre de termes d'indexation et sont souvent les plus complètes. Nous avons d'emblée sélectionné les 41 termes appartenant à des notices des niveaux inférieurs (de la série à la pièce), mais puisque sept de ces termes ne répondaient pas aux critères de sélection (voir section 4.1.2.3), nous avons ajusté notre sélection.

Nous appelons un ensemble de séquences qui ont la même question et la même réponse une polyséquence. Une polyséquence contient une unité de sujet puisqu'elle repose sur le sujet de la question d'usager à l'origine de la polyséquence (voir Figure 13).

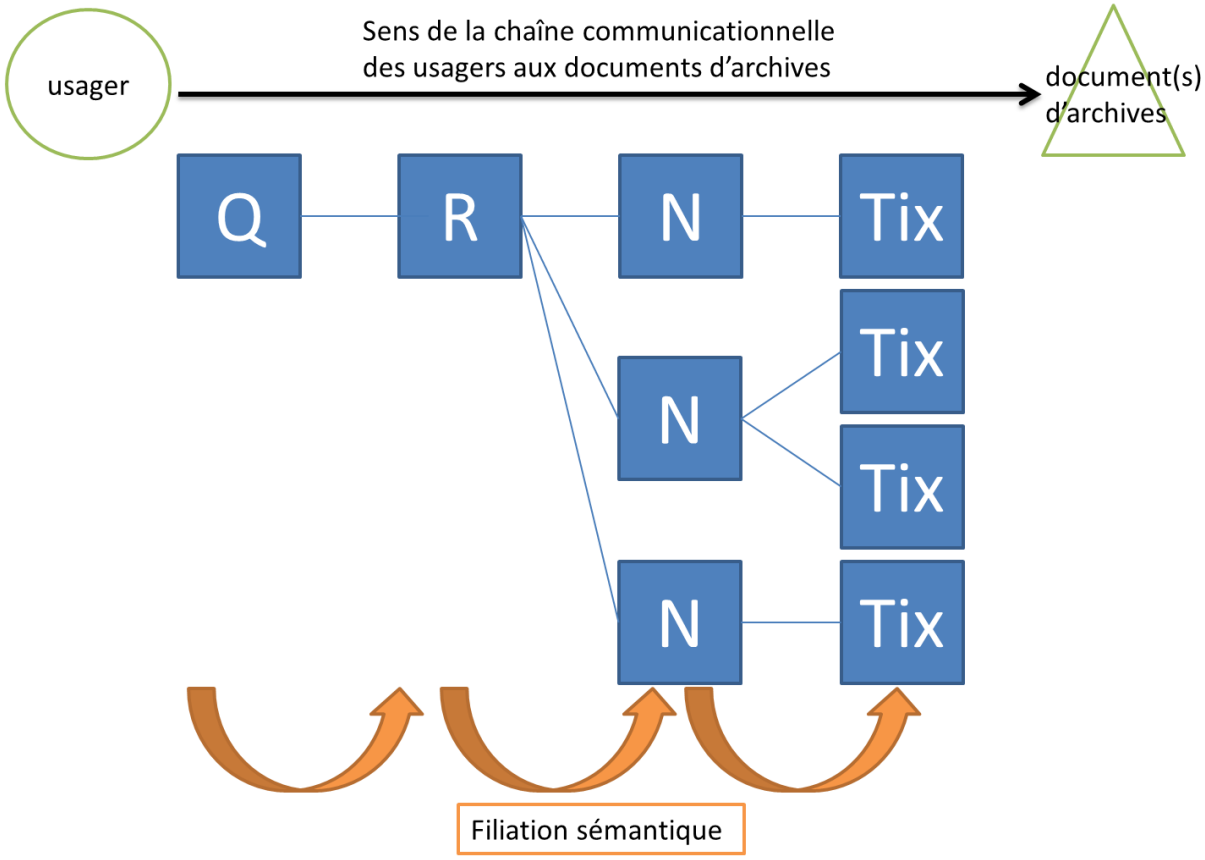

Figure 13. Filiation sémantique dans une polyséquence (QRNTixNTixTixNTix)

Les 34 termes d'indexation retenus étaient contenus dans des notices descriptives, elles-mêmes associées à dix réponses d'archivistes. Ces réponses incluaient d'autres cotes renvoyant à d'autres notices descriptives du niveau du fonds ou de la collection. Les termes associés à ces autres cotes et notices (n=63) ont été ajoutés à notre corpus pour étendre la sélection et avoir un portrait plus juste du traitement d'une question d'usager. Par exemple, *Foires* et *Parcs d'attractions* (tix35 et tix36) sont reliés à la notice d090 mentionnée dans la réponse 014r (voir Figure 14).

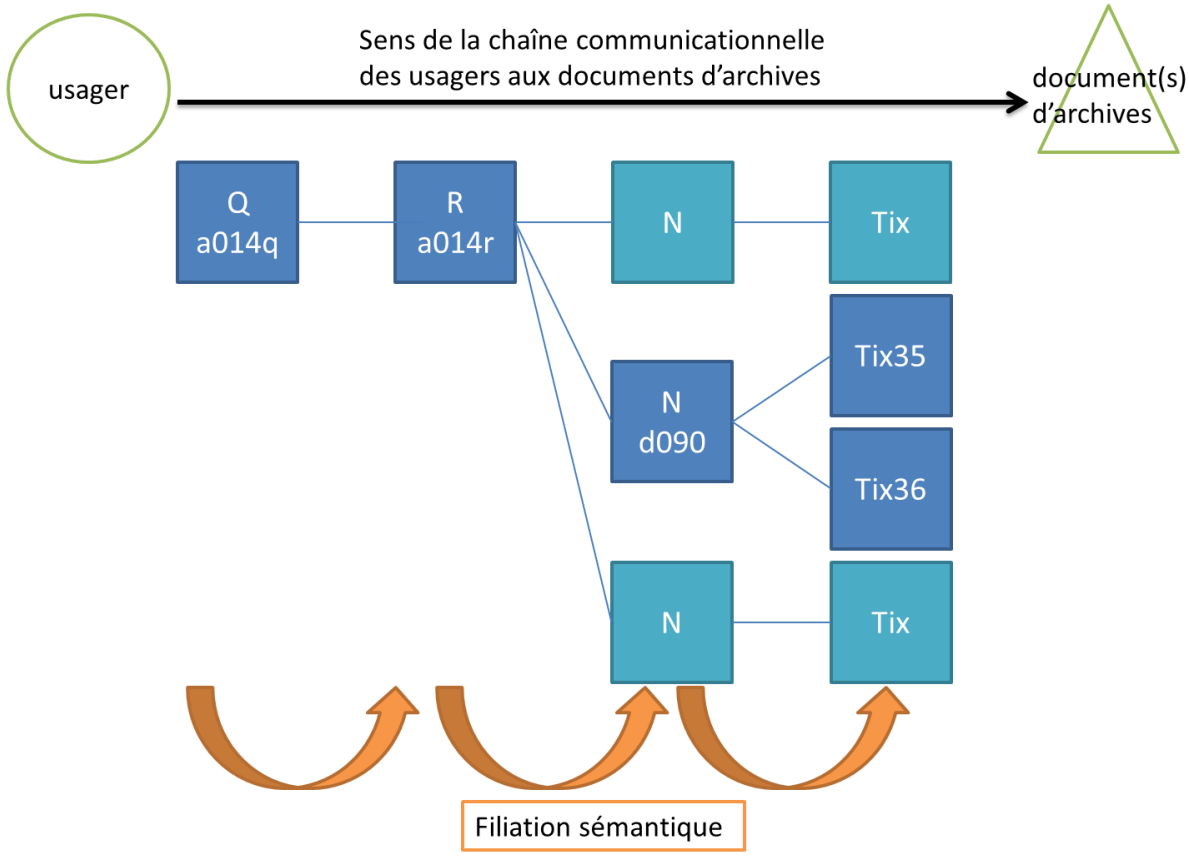

Figure 14. Filiation sémantique dans une polyséquence : un exemple

Nous avons alors retenu les termes d'indexation qui étaient contenus dans des notices dont les cotes étaient indiquées dans une même réponse. Reprenons l'exemple; la réponse a014r incluait : *Couvents* tix26, *Softball* tix27, *Viaducs* tix28, *Bâtiments publics* tix19, *Comités de citoyens* tix20, *Expropriation* tix 21, *Groupes sociaux* tix22, *Logement social* tix23, *Sociologie urbaine* tix24, *Urbanisme* tix25, *Manifestations* tix31, *Rassemblements* tix32, *Théâtre* tix33, *Administration municipale* tix03, *Anciens de l'université d'Ottawa* tix04, *Clubs de hockey* tix05, *Généalogie* tix06, *Livres comptables* tix07, *Nationalisme* tix08, *Base-ball* tix34. Ceux-ci étaient rattachés à sept notices descriptives différentes (d047, d049, d072, d074, d075, d082 et d089).

En tout, 97 termes d'indexation ont ainsi été associés à 26 notices descriptives parmi celles retenues. Nous avons vérifié que ces termes correspondaient bien à nos critères de sélection déjà identifiés et nous avons ajouté des critères de sélection supplémentaires propres aux termes d'indexation. Ces critères sont les suivants :

- la tête du terme d'indexation doit être un nom commun (voir section 4.1.2.3b). P. ex. les termes d'indexation suivants ont été retenus : « Anciens de l'université d'Ottawa » ou « Suffrage féminin »; les deux noms – anciens et suffrage – sont la tête des syntagmes et des noms communs.
- s'il y a agglutination de plusieurs mots ou expressions, ils sont scindés en deux et chaque segment constitue un terme d'indexation à part entière. Lors de la séparation des termes qui semblaient doubles, les noms propres sont écartés (voir section 4.1.2.3b). P. ex. dans le terme d'indexation « Canada Gouverneurs généraux Portraits », nous identifions « Gouverneurs généraux » et « portraits ». « Canada » est supprimé.
- les dates sont retirées des termes d'indexation (voir section 4.1.2.3b). P. ex. le terme d'indexation « Réciprocité, 1911 », nous avons retiré la date.
- seuls les termes en français sont retenus (voir section 4.1.2.3a). P. ex., les termes d'indexation suivants n'ont pas été retenus : « University of Toronto, 1921 », « The Memoirs of Sir Robert Borden, [ca.1938] ».
- les énumérations sont disloquées en plusieurs termes. P. ex., le terme d'indexation « Clergé Nomination, choix et élection » est décomposé en « clergé », « nomination », « choix » et « élection ».

Certains critères ont été appliqués conjointement. Par exemple, dans le terme d'indexation suivant « Borden (Famille) - Généalogie, [1778-1937] », nous conservons « Généalogie ». Dans le terme d'indexation « Éloquence politique - Canada, 1891-1917 », nous n'avons retenu que « Éloquence politique ». Après l'application de ces critères sur les termes d'indexation, nous obtenons 104 termes d'indexation sur les 458 termes d'indexation répondant à nos critères. Mais certains de ces termes sont répétés. Nous obtenons alors 96 termes d'indexation différents.

4.1.3.4. Étiquettes

Notre procédure de collecte incluait la possibilité de collecter des étiquettes. C'est pourquoi nous dressons une courte description de l'état des termes que sont les étiquettes. Les

trois établissements dont émanent nos données diffusent des images de documents d'archives sur la plateforme *Flickr*, un « service de gestion et de partage de photographies en ligne » (Flickr 2017). Les services d'archives participant à notre recherche gèrent ainsi des albums et attribuent à leurs photographies des étiquettes. Certaines étiquettes sont générées automatiquement[51] par *Flickr* et apparaissent en grisé dans le nuage d'étiquettes. Ces dernières ne sont pas retenues dans notre recherche puisqu'elles ne sont pas d'origine humaine.

Selon l'étude de terrain de Hoyle-Rojas, *Flickr* est un « site d'hébergement et de partage en ligne de photographies ». Flickr est

> [u]tilisé par des millions d'amateurs et de professionnels de l'image à travers le monde, il est le plus populaire des réseaux sociaux fondé sur la photo et offrant différents dispositifs de catégorisation et de production de métadonnées autour de l'image comme le tagging [*étiquetage*], les favoris, les albums et les groupes de discussion. (…) *Flickr* a la particularité d'avoir été choisi par de nombreuses institutions patrimoniales comme vitrine pour leurs collections d'images et comme dispositif de médiation pour proposer à leurs publics des activités d'indexation collaborative et de crowdsourcing [*externalisation ouverte*]. (Hoyle-Rojas 2016, 115)

Nous retenons de la perception de Hoyle-Rojas que *Flickr* est très utilisé, donc connu tant des archivistes que du public et choisi par de nombreux établissements patrimoniaux, ce qui est le cas des trois participants à notre recherche.

Nous avons ouvert un compte *Flickr* (abonnement gratuit) et nous nous sommes abonnée aux pages des trois services d'archives concernés par la recherche. Nous avons ainsi pu récupérer les listes des étiquettes de chacun des établissements qui n'étaient pas accessibles sans être abonnée/connectée. Chacune de nos sources a un compte *Flickr* « PRO » (voir section 4.1.1.2). Au 2017-03-02, nous comptabilisions 160 étiquettes pour l'établissement 1, 17 998 étiquettes pour l'établissement 2 et 825 étiquettes pour l'établissement 3. L'établissement 2 crée des étiquettes pour chacun de ses mots et dans les deux langues officielles du Canada, ainsi, le total de ses étiquettes n'est pas comparable aux deux autres établissements. Considérant la taille de l'établissement, il est étonnant de constater que

[51] Nous ne savons pas comment *Flickr* génère ces étiquettes grisées.

l'établissement 3 ait bien davantage d'étiquettes que l'établissement 1. Ceci s'explique principalement par la taille de la collection sur la plateforme. L'album *Flickr* de l'établissement 3 a été créé plusieurs années avant les deux autres établissements. Les étiquettes des trois établissements présentent, en plus des noms communs, parfois uniquement des chiffres ou des nombres (dont des dates), des « mots vides » (ayant un sens grammatical p. ex. *du, de, la* et non des « mots pleins » ayant un sens lexical, p. ex. *fourrure, hockey*) ou encore de nombreux noms propres.

4.1.3.5. Récapitulatif : données accessibles et corpus

Fortin, Côté et Filion (2006) définissent les notions de « population » et « population cible » de la manière suivante. La population est « un ensemble d'éléments (individus, spécimens, dossiers) qui ont des caractéristiques communes » (p. 249). « La population cible est l'ensemble des personnes qui satisfont aux critères de sélection définis d'avance et qui permettent de faire des généralisations. La population accessible est la portion de population cible qu'on peut atteindre » (p. 250). La population de la recherche est composée des termes du vocabulaire pour l'accès thématique aux archives patrimoniales (VATAP). Les données correspondant à la population ciblée sont issues des courriels de référence, des notices descriptives et des termes d'indexation. Les données correspondant à la population accessible sont composées des termes du VATAP des trois établissements participants (voir section 4.1.1.1) et des critères identifiés dans les limites du corpus (voir section 4.1.2.3). Finalement, des critères subséquents, définis lors du contact avec les données (voir sections ci-dessus 4.1.3.1 à 4.1.3.4) ont été utiles pour établir des corpus de taille raisonnable pour un traitement entièrement humain.

Tableau XI. Corpus général : données accessibles et données retenues

	Données accessibles	Corpus 1	Corpus 2	Corpus général	Pourcentage des données accessibles
Questions	125	10	20	30	24%
Réponses	98	10	20	30	30,6%
Notices descriptives	155	26	58	84	54,2%
Termes d'indexation	458	96	0	96	21%

Le Tableau XI montre la taille de la population accessible de nos données et celle des échantillons finaux. La population accessible a été restreinte par des critères de convenance,

justifiés plus haut (voir section 4.1.2.3), afin d'avoir un échantillon non généralisable, mais représentatif du VATAP. Les données du corpus 1 sont composées de 8% des questions, 10% des réponses, 17% des notices et de 21% des termes d'indexation des données accessibles. Les données du corpus 2 sont composées de 16% des questions, 20% des réponses et 37% des notices des données accessibles. Avec les corpus 1 et 2 réunis, nous traitons 24% des questions, 30,6% des réponses, 54,2% des notices et 21% des termes d'indexation des données accessibles.

Conclusion

La section 4.1. rend compte de la collecte des données, à savoir la sélection et la description des milieux, la constitution des corpus et la description des données elles-mêmes. Il est regrettable que notre corpus qui regroupait théoriquement quatre types de données (questions d'usagers, notices descriptives, termes d'indexation et étiquettes) ait dû se modifier. En effet, nous avons perdu lors du processus de sélection les étiquettes – seconde source du vocabulaire des usagers. Cependant, nous avons gagné la réponse des archivistes aux courriels envoyés à la référence comme autre source du VATAP ; elle constitue l'illustration de la médiation sémantique de l'archiviste avec l'usager. Les étiquettes auraient sans nul doute apporté un éclairage intéressant sur le vocabulaire des usagers sous forme de termes et non de texte. Cette partie de la recherche n'est, nous l'espérons, que reportée momentanément (voir section 7).

4.2. Prétraitement

Notre corpus constitué, la première tâche à accomplir pour mener notre étude est d'identifier les expressions linguistiques porteuses de sujet (thémanyme). Ces thémanymes se répartissent en termes de recherche, termes de description et termes d'indexation et constituent le vocabulaire pour l'accès thématique des archives patrimoniales (VATAP).

Avant tout, nous exposons ici quelques données brutes[52] afin que les différents traitements présentés dans cette section s'enracinent dans un exemple concret. Voici :

- un exemple de question d'usager dans un courriel (voir Figure 15);
- un exemple de réponse d'archiviste de référence à ce courriel (voir Figure 16);
- un exemple de notice descriptive dont la cote figure dans la réponse de l'archiviste (voir Figure 17);
- un exemple (fictif) de termes d'indexation (voir Figure 18).

De : M. Untel [untel@blabla.com]
A : Archie Viste [archie.viste@servicedarchives.ca]
Date : 2015-07-28 19:41
Objet : Molson

Bonjour,

Mon nom est M. Untel, je suis journaliste au journal Blabla. Je prépare un dossier sur l'important héritage patrimonial de la Brasserie Molson de la rue Notre-Dame. Je suis à la recherche de photos d'archives des lieux, à différentes époques. Je me demandais si vous n'auriez pas de telles photos et s'il nous serait possible de les utiliser?

Merci de votre aide,
M. Untel
Journaliste
Blabla
untel@blabla.com
438 765-4321

Figure 15. Exemple de question d'usager par courriel

[52] Les informations personnelles ont été modifiées pour assurer la confidentialité des personnes.

De : Archie Viste [archie.viste@servicedarchives.ca]
Date : mercredi 29 juillet 2015 14:09
À : M. Untel [untel@blabla.com]
Objet : RE: Molson

Bonjour M. Untel,

Voici les photos que nous possédons où l'on voit la brasserie Molson :
1936 :
1942 :
1966 :
Début des années 1980
Si vous désirez utiliser ces photos, vous devrez indiquer la mention de crédits suivante : NOM

Si vous souhaitez des versions en plus haute résolution, il nous fera plaisir de vous les faire parvenir via XXX.

Cordialement
Archie Viste
Service de référence
438 123-4567 (tél.)
514 123-4567 (téléc.)
Courriel : archie.viste@servicedarchives.ca
Site web : http://www.servicedarchives.ca/

[6 pièces jointes]

Figure 16. Exemple de réponse d'archiviste par courriel

Le courriel de réponse de l'archiviste est accompagné de fichiers joints. Voici la liste des noms de fichiers joints suivis des cotes) :

- JJ94-Z111.jpg JJ J999 JJ094-Y-1-17-D0110
- J9999-2_1555E(1971-)-001.jpg JJ J999 JJ166-1-2-D010
- J9999-2_1555E(1971-)-002.jpg JJ J999 JJ166-1-2-D010
- J9999-2_1555E(1971-)-003.jpg JJ J999 JJ166-1-2-D010
- J9999-A-002.jpg JJ J999 JJ166-1-2-D030
- JJ94-B019-001.jpg JJ J999 JJ94-B19-001

Parmi ces cotes de documents recommandés par l'archiviste, une seule correspond à une notice pourvue d'une portée et contenu (voir Figure 17).

Brasserie Molson: Rue Notre-Dame . - 1er avril 1936
Photographie de la brasserie Molson (1590, rue Notre-Dame Est) On y aperçoit la rue Notre-Dame vers l'est et la structure du pont Jacques-Cartier. Numéro original de la pièce: Z-111.

Figure 17. Exemple de notice descriptive recommandée par l'archiviste

Quand ils sont présents, les termes d'indexation sont rattachés à une notice descriptive. Les documents présentés dans cet exemple ne présentant pas de terme d'indexation, nous

avons puisé dans les termes d'indexation associés à d'autres documents pour constituer cet exemple (voir Figure 18) et compléter la séquence question-réponse-notice descriptive-termes d'indexation (désormais QRNTix).

Brasserie Molson Centres villes Molson Companies Ltd Montréal (Québec : Ville) Pont Jacques-Cartier (Montréal, Québec) Vues aériennes

Figure 18. Exemple (fictif) de termes d'indexation

De ces documents, nous avons tiré des expressions linguistiques porteuses de sujet (thémanymes), suite à une chaîne de prétraitement comprenant l'anonymisation (réalisée au préalable), l'épuration (c'est-à-dire le retrait des formules périphériques à l'expression du sujet de la recherche), la segmentation en groupes de mots avec un nom commun à leur tête et la caractérisation sémantique de ces groupes nominaux (selon l'analyse en facettes ou par champ sémantique, décrite ci-dessous). La liste suivante présente le résultat final de cette chaîne de prétraitement pour l'exemple en question; elle inclut les répétitions des usagers et archivistes (voir Figure 19).

Question : héritage patrimonial Brasserie Molson rue Notre-Dame Brasserie Molson rue Notre-Dame lieux Réponse : brasserie Molson	Notice descriptive : Brasserie Molson Rue Notre-Dame brasserie Molson 1590, rue Notre-Dame Est rue Notre-Dame vers l'est la structure du pont Jacques-Cartier Termes d'indexation : Brasserie Molson Centres villes Montréal Pont Jacques-Cartier Vues aériennes

Figure 19. Exemple d'expressions linguistiques porteuses de sujet (thémanymes)

Dans la Figure 19, nous voyons que certains termes se rapportent au même sujet, par exemple, il est question de la brasserie Molson dans les quatre sources. Ainsi, nous voyons l'expression du même sujet qui se déroule, à la manière d'une même bobine, de la question d'usager aux termes d'indexation.

Les termes de recherche, les termes de description et les termes d'indexation sont trois types de termes qui présentent des caractéristiques que nous abordons dans cette étude selon une approche linguistique (section 4.2.1). Après avoir décrit ces caractéristiques, nous décrivons les quatre étapes du prétraitement : anonymisation (section 4.2.2), l'épuration (section 4.2.3), la segmentation (section 4.2.4) et la caractérisation sémantique (section 4.2.5).

4.2.1. Approche linguistique des données

Les quatre sources de notre corpus (question, réponse, notice descriptive et terme d'indexation) résultent de trois opérations : rechercher, décrire et indexer. Les termes de recherche, de description et d'indexation nécessitent des méthodes de prétraitement différentes selon qu'ils se présentent dans des phrases (termes de description et termes de recherche) ou non (termes d'indexation).

4.2.1.1. Termes de recherche

Les termes de recherche sont présents dans les échanges de courriels entre un usager qui pose une question et un archiviste de référence qui lui répond.

Dans sa thèse sur les émotions dans les questions de recherche posées par courriel à la référence, Veal (2011) a décrit les tests inter- et intra-codeurs; elle a indiqué les instructions de codage pour un échantillon et les a raffinées après le test inter-codeur pour le corpus complet. Nous nous sommes inspirée de cette méthodologie pour élaborer la démarche de notre propre recherche.

Kim (2014, 316) présente trois types d'ambigüités (voir section 3.2.2). Nous n'avons pas repris ces éléments dans notre analyse pour deux raisons. Ces éléments requièrent des connaissances linguistiques poussées qui ne coïncident pas avec notre approche plus proche du monde professionnel des archivistes. Ensuite, pour analyser l'écart sémantique, notre méthodologie ne requérait pas une analyse aussi profonde du type de questions.

Plusieurs classifications des courriels d'usagers à la référence ont été proposées telles que la double catégorisation de Mas pour 1) le type de question et 2) le type de contenu des questions (Mas 2013-2014, 89-91), élaborée à partir de celles de Duff et Johnson (2001, 48-51) et Gagnon-Arguin (1998).

Nous avons examiné les courriels d'usagers à la référence pour distinguer ceux qui constituent un accès thématique dont le sujet est exprimé par des noms communs (p. ex. *je cherche des informations sur le droit de vote des femmes*) de ceux qui formulent un autre type de demande, par exemple une demande de document connu (p. ex. *je cherche le bulletin de vote complété par madame Mae O'Connor, première femme à s'être portée candidate au Québec*). Afin d'assurer une constance dans l'application des divers critères de sélection des questions d'usagers pour notre recherche, nous avons créé un arbre décisionnel à partir de la double catégorisation de Mas (2013-2014) (voir section 4.1.3.1).

Dans les études sur la référence archivistique, ce sont plus souvent les questions d'usagers au service de référence qui sont au cœur de l'étude (Duff et Johnson 2001, Yakel 2000, Duff et Fox 2006). Nous considérons la référence en tant que situation de communication (Duff et Yakel 2017, Guitard 2014, voir aussi 3.2.1). Ce fait nous invite à examiner non seulement les questions des usagers, mais aussi les réponses des archivistes.

4.2.1.2. Termes de description

Nous entendons par *termes de description* les expressions linguistiques porteuses de sujet (les thémanymes) à la fois dans la notice descriptive (dans la portée et contenu de la zone de description des documents d'archives, selon les *Règles pour la description des documents d'archives* (*RDDA*) du Bureau canadien des archivistes (BCA 2008, cf. section 331) et dans le titre des documents d'archives. Ces termes véhiculent le sujet contenu dans les documents d'archives. Jusqu'à présent, nous n'avons pas trouvé d'étude sur les termes employés dans les notices descriptives archivistiques et les types de termes qu'elles contiennent. Peut-être à cause de la normalisation de la structure d'une notice et de la liberté laissée à l'archiviste lors de la rédaction de la description, les chercheurs n'ont pas encore ressenti le besoin d'étudier les termes de description. Nous avons repris la catégorisation du contenu des courriels d'usagers à la référence qui, elle, est développée et attestée dans la littérature afin de l'adopter dans notre analyse du contenu des descriptions. Utiliser la même méthode pour analyser le contenu des deux sources de notre corpus (notices descriptives et courriels échangés entre usagers et archivistes de référence) permet de s'assurer d'avoir des résultats comparables.

4.2.1.3. Termes d'indexation

Les termes d'indexation se présentent sous la forme des mots isolés, c'est-à-dire qu'ils ne sont pas présents dans des phrases ni un texte. L'étude des mots isolés est particulière (Dubois 1983); les mots isolés manquent de contexte qui permettrait de rapidement lever les ambigüités et identifier le référent, la chose à laquelle renvoie le mot. « Les signes linguistiques permettent au locuteur de parler de la réalité qui l'entoure. Ils ont, en effet, la propriété de pouvoir renvoyer aux objets du monde, extérieurs à la langue ; ces objets sont les **référents** *[mise en forme des auteures]*. Les référents sont des entités matérielles ou conceptuelles (êtres, objets, lieux, processus, propriétés, événements, etc.). Ils relèvent de l'univers extralinguistique réel ou fictif (par exemple, la sirène). » (Lehman et Martin-Berthet 2008, 10) Envisageons la situation de communication d'un usager qui consulte un index de documents d'archives. Pour l'usager, les termes d'indexation ont pour seul contexte les autres termes d'index s'ils lui sont présentés sous forme d'index visible ou de nuage d'étiquettes (de Keyser 2012). Cependant certains index ne sont pas visibles des usagers et ne servent qu'au moteur de recherche pour présenter des résultats (de Keyser 2012).

Pour les raisons évoquées précédemment (voir section 4.2.1.2), nous avons repris la même caractérisation sémantique pour identifier le sujet des termes d'indexation que celle utilisée pour les termes de recherche et les termes de description.

4.2.2. Anonymisation

Certaines données de notre recherche ont une nature sensible; elles contiennent des renseignements personnels, selon la définition des lois fédérales et québécoises (voir section 4.4.). Nous présentons ci-dessous le code d'anonymisation auquel nous avons eu recours. L'anonymisation touche les courriels, questions d'usagers et réponses d'archivistes; elle concerne plusieurs éléments dans un courriel. Nous présentons notre code d'anonymisation dans le Tableau XII Code d'anonymisation.

Tableau XII. Code d'anonymisation

CODE	Description de l'élément à remplacer
NOM	les noms propres de personnes, sauf les personnes célèbres, de notoriété publique, p. ex. P. E. Trudeau, les noms de lieu s'ils sont trop précis ou permettent de rattacher l'usager à un lieu, les noms d'entreprises, de services
USGR	le nom de l'usager et son courriel
ARVST	le nom de l'archiviste de l'institution et son courriel
PRFL	le nom ou le courriel d'un professionnel de l'institution ou d'une autre institution, souvent les répartiteurs de courriels
@ (*a commercial*)	les adresses courriel qui ne sont pas identifiées autrement (USGR ou ARVST ou PRFL)
ADRESSE	les adresses qui s'échelonnent généralement sur plusieurs lignes
TEL	numéro de téléphone, cellulaire ou télécopie
OUTIL	le nom de certains outils tels que le catalogue
SITE	les sites internet hors URL qui renvoient vers une description de l'établissement
Établissement 1, 2 ou 3	le nom des établissements en question

Dans le but de réduire au maximum le risque de bris de la confidentialité relative aux informations incluses dans les courriels (voir section 4.4), nous avons appliqué le code d'anonymisation au fur et à mesure que nous stockions les données. Tous les courriels (questions et réponses) ont subi l'anonymisation. Une seule fois un usager était double (deux adresses courriels étaient associées à la même demande et les deux personnes signaient le courriel); à ce moment-là, nous avons répété USGR à deux reprises. Mais si une personne écrit pour une autre (ce qui est souvent le cas dans les études de notaires), la personne qui faisait la demande était inscrite USGR et le nom du notaire pour lequel elle travaillait était anonymisé en NOM, ainsi que le nom de l'étude.

Ici, nous voyons la limite de l'anonymisation des courriels. Dans nos données, afin d'accéder aux résultats identifiés dans la réponse de l'archiviste, nous ne supprimons pas le lien URL pointant vers une description sur le site Internet de l'établissement étudié, il reste donc identifiable (p. ex., un lien URL vers Pistard dans les données issues de BAnQ).

Pour des raisons évidentes de confidentialité, nous donnons un exemple d'anonymisation avec des noms fictifs (voir Tableau XIII).

Tableau XIII. Exemple d'anonymisation d'un courriel

courriel d'usager reçu	**courriel d'usager anonymisé**
De : M. Untel [untel@blabla.com] A : Archie Viste [archie.viste@servicedarchives.ca] Date : 2015-07-28 19:41 Objet : Molson Bonjour, Mon nom est M. Untel, je suis journaliste au journal Blabla. Je prépare un dossier sur l'important héritage patrimonial de la Brasserie Molson de la rue Notre-Dame. Je suis à la recherche de photos d'archives des lieux, à différentes époques. Je me demandais si vous n'auriez pas de telles photos et s'il nous serait possible de les utiliser? Merci de votre aide, M. Untel Journaliste Blabla untel@blabla.com 438 765-4321	De : USGR [USGR] A : ARVST [ARVST] Date : 2015-07-28 19:41 Objet : Molson Bonjour, Mon nom est USGR, je suis journaliste au journal NOM. Je prépare un dossier sur l'important héritage patrimonial de la Brasserie Molson de la rue Notre-Dame. Je suis à la recherche de photos d'archives des lieux, à différentes époques. Je me demandais si vous n'auriez pas de telles photos et s'il nous serait possible de les utiliser? Merci de votre aide, USGR Journaliste NOM USGR TEL

L'anonymisation a touché les questions d'usagers et les réponses d'archivistes envoyées par courriel.

4.2.3. Épuration

Après l'anonymisation, le texte des questions des usagers, des réponses des archivistes et des notices descriptives est épuré des formules qui ne sont pas pertinentes pour notre recherche.

Dans les courriels (questions des usagers et réponses des archivistes), nous avons retiré les formules de politesse (p. ex. *bonjour, cordialement, merci*), les portions de communication à l'interne entre collègues qui reçoivent la question de l'usager et le réacheminent (p. ex. *je n'ai pas réussi à trouver des informations pour ce client, peux-tu m'aider svp ?*), certains éléments propres au mode de communication, ici le courriel (p. ex. la mention *Objet, À, De*, la signature qui inclut les coordonnées), les excuses (p. ex. pour le retard dans la réponse), les indications liées à la reproduction (mention de la source, demande de droits), aux divers frais

possibles, aux heures d'ouverture et autres informations relatives à une démarche concrète. Ces éléments ne se rapportent pas au sujet de recherche en tant que tel. Cependant, certains de ces éléments sont conservés si la manière dont l'auteur du courriel a rédigé impose de les conserver pour comprendre l'objet de la demande.

Dans les notices descriptives, nous avions envisagé dans un premier temps de retirer tous les éléments de « contenu » au sens des *RDDA – i. e.* des éléments relatifs à la catégorie de document, au genre de document, à la classification, p. ex. *documents sonores*, *photographies, lettre, dossier* – pour ne retenir que les éléments de « portée » qui décrivent le sujet des documents. Toutefois, il nous est apparu en examinant la section portée et contenu des notices que les éléments de « contenu » étaient souvent imbriqués dans l'expression du sujet si bien qu'il était difficile de les retirer sans constater une perte de sens relatif au sujet des documents. La plupart du temps, l'identification de l'élément de « contenu » est clairement séparable de l'élément de « portée ». Voici des exemples dans lesquels nous séparons par une double barre oblique les deux éléments de contenu :

- *illustrer avec des photos d'archives // la vie de NOM* (a017q),
- *des photos en lien avec // la première fois où les femmes ont pu voter au Québec* (a033q),
- *nous recherchons une vidéo // du dirigeable R-100 survolant la ville de Montréal en 1930* (a026q).

Mais d'autres fois, les données incluent des formulations telles que *des photos d'époque (aucune époque précisée encore) du centre-ville de Hull* où il est difficile de séparer le sujet qui inclut ici une dimension temporelle, grâce à l'adjectif épithète du type de document. C'est pour cette raison que l'épuration est minime pour les notices descriptives. Cependant, les éléments de « contenu » et de « forme » ont été codés différemment lors de la phase de caractérisation sémantique (voir section 4.2.5).

Prenons un exemple. La question de l'usager anonymisée se présente ainsi : *Je prépare un dossier sur l'important héritage patrimonial de la Brasserie Molson de la rue Notre-Dame. Je suis à la recherche de photos d'archives des lieux, à différentes époques*. La question porte sur la notion d'héritage patrimonial, un sujet selon notre analyse. La réponse de l'archiviste de

référence se présente ainsi : *Voici les photos que nous possédons où l'on voit la brasserie Molson : [hyperlien] 1936 : [hyperlien] 1942 : [hyperlien] 1966 : [hyperlien] Début des années 1980 : [hyperlien]*. La réponse de l'archiviste présente ici des hyperliens vers des notices descriptives de leur catalogue ou directement vers des pièces numérisées (le plus souvent des photographies). Le document VM94-Z111.jpg dont l'hyperlien est inscrit dans le courriel a pour cote archivistique CA M001 VM094-Y-1-17-D0110 ; la notice qui le décrit est disponible dans leur catalogue en ligne. Cette notice a pour titre « Brasserie Molson : Rue Notre-Dame.- 1er avril 1936 ». La collation ou description matérielle qui est habituellement associée au titre dans la tradition archivistique ne se présente pas dans le titre dans ce catalogue basé sur la norme ISAD(G)[53]. Dans la zone *Description* ou *Portée et contenu*, nous pouvons lire des informations de contenu et l'ancienne cote : « Photographie de la brasserie Molson (1590, rue Notre-Dame Est) On y aperçoit la rue Notre-Dame vers l'est et la structure du pont Jacques-Cartier. Numéro original de la pièce: Z-111. » Le texte de ces trois sources est épuré pour le réduire aux éléments relatifs au sujet de la recherche (voir Tableau XIV).

Tableau XIV. Exemple d'épuration des données

Source	Données épurées
Question de l'usager (Q)	l'important héritage patrimonial de la Brasserie Molson de la rue Notre-Dame photos d'archives des lieux, à différentes époques
Réponse de l'archiviste de référence (R)	les photos où l'on voit la brasserie Molson : 1936 : 1942 : 1966 : Début des années 1980 :
Titre et notice descriptive (N)	Brasserie Molson : Rue Notre-Dame.- 1er avril 1936 Photographie de la brasserie Molson (1590, rue Notre-Dame Est) On y aperçoit la rue Notre-Dame vers l'est et la structure du pont Jacques-Cartier

Le texte présente encore des verbes conjugués qui participent parfois indirectement à l'expression du sujet; rappelons que selon la tradition d'indexation nous ne nous intéressons qu'aux expressions nominales. Ainsi, dans l'exemple que nous avons pris, le verbe

[53] La description des catalogues utilisés par les services d'archives participants à notre recherche est incluse dans la section 4.1.1.2 sur la description des services d'archives participants.

apercevoir[54] implique que l'objet de l'image n'est pas centré sur les éléments indiqués en complément direct à ce verbe, mais fait seulement partie des éléments visibles sur la photographie. D'autres formules linguistiques introduisant le sujet ont été répertoriées dans la recherche (voir Tableau LXXVI).

Contrairement à ce que nous pensions initialement, nous avons dû épurer les termes d'indexation de la notice, au même titre que les textes des questions, des réponses et des notices. En effet, certains termes d'indexation ne sont pas systématiquement des syntagmes nominaux bien formés en français (p. ex. *Hôpitaux militaires – England,* voir section 4.1.3.3).

4.2.4. Segmentation

La segmentation consiste à découper le texte étudié en segments qui correspondent à des expressions linguistiques porteuses de sujet (les thémanymes). Par exemple, voici une question d'usager anonymisée et épurée ainsi :

> Photos demandées. L'hebdomadaire NOM est à préparer un dossier spécial sur le 125e anniversaire du décès du Curé Labelle et nous aurions besoin de photos de bonnes dimensions, des photos du Curé Labelle et/ou de la construction du chemin de fer dans les Laurentides, des premiers trains à y circuler. (a034q)

Le texte a été découpé en segments de la manière suivante (la méthode utilisée est détaillée ci-dessous) :

> Photos demandées ; Photos ; L'hebdomadaire NOM est à préparer un dossier spécial sur le 125e anniversaire du décès du Curé Labelle ; photos de bonnes dimensions ; des photos ; du Curé Labelle ; curé ; la construction du chemin de fer dans les Laurentides ; construction ; chemin de fer ; dans les Laurentides ; des premiers trains à y circuler ; premiers ; trains ; y (= dans les Laurentides) ; à circuler. (a034q)

[54] « apercevoir : I. v.tr. A. Saisir par la vue, en un instant, une personne ou une chose, en dépit de certains obstacles, en particulier l'éloignement, le rétrécissement du champ de vision, le manque de luminosité. » (*Trésor de la langue française* 2017)

Certains éléments sont répétés afin de correspondre à une seule facette[55]. La segmentation de textes en segments plus courts est une étape classique en analyse de texte en vue de la recherche d'information (Jacquemin et Zweigenbaum 2000). Le principe de segmentation repose sur l'analyse grammaticale (4.2.4.1) et l'identification de formules introductives (4.2.4.2).

4.2.4.1. Analyse grammaticale

Lors de l'analyse grammaticale, le linguiste identifie la nature (nom ou substantif, article, adjectif, pronom, verbe, adverbe, préposition, conjonction, interjection) et la fonction (une dizaine de fonctions recensées : sujet, compléments, attribut, apposition, épithète, épithète détachée, apostrophe) (Wilmet 2016). L'analyse grammaticale est appliquée sur des phrases. Alors nous segmentons d'abord en phrases – *i. e*. ensemble de mots entre deux points finaux – et en propositions (groupes de mots centrés autour d'un verbe conjugué), puis en groupes de mots plus petits (groupe nominal, verbal, prépositionnel[56], etc.) établis à partir des fonctions (sujet grammatical, verbe, compléments de toutes sortes). Ainsi, reprenant l'exemple précédent (a034q), nous passons de :

> L'hebdomadaire NOM est à préparer un dossier spécial sur le 125e anniversaire du décès du Curé Labelle et nous aurions besoin de photos de bonnes dimensions, des photos du Curé Labelle et/ou de la construction du chemin de fer dans les Laurentides, des premiers trains à y circuler

à des segments plus courts :

> L'hebdomadaire NOM est à préparer un dossier spécial sur le 125e anniversaire du décès du Curé Labelle et
> nous aurions besoin de photos de bonnes dimensions,
> des photos du Curé Labelle et/ou
> de la construction du chemin de fer dans les Laurentides, des premiers trains à y circuler

[55] Nous employons *facette* au sens d'un terme générique renvoyant à une dimension d'un sujet (voir aussi section 3.3.3).

[56] Ainsi, un groupe nominal est un groupe de mots régi par un nom; p. ex. le groupe nominal *le petit chat noir* est régi par le nom *chat*.

et encore plus courts :

L'hebdomadaire NOM est à préparer un dossier spécial sur le 125e anniversaire du
décès du Curé Labelle et
nous aurions besoin de
photos
de bonnes dimensions,
des photos
du Curé Labelle et/ou
de la construction
du chemin de fer
dans les Laurentides,
des premiers trains à y circuler

Et ainsi de suite pour atteindre des unités véhiculant une seule unité de sens, une seule facette. La segmentation nous oblige parfois à répéter un segment avec et sans complément quand son sens change. Par exemple, à la dernière ligne de l'exemple précédent, le segment

des premiers trains à y circuler

sera segmenté selon les étapes successives suivantes :

1/ des premiers trains à y circuler

2/ des trains + premiers à y circuler

3/ des trains + premiers + circuler + y (= dans les Laurentides).

4.2.4.2. Identification de formules introductives du sujet

Notre recherche porte sur l'accès thématique. Nous nous intéressons non pas à toutes les expressions linguistiques des données que nous avons collectées, mais seulement à celles qui expriment le sujet d'une recherche ou d'un document, les thémanymes. Nous avons segmenté l'ensemble de nos données pour obtenir des thémanymes. Le jugement humain identifie quelles sont les expressions linguistiques porteuses de sujet et les discrimine des autres segments de texte relativement facilement. Mais dans la perspective d'une formalisation et d'une éventuelle automatisation, nous exprimons dans cette section des critères de jugement habituellement implicites.

Les thémanymes se trouvent de manière attendue dans :

- les questions des usagers (parfois l'objet, toujours le corps du message)
- les réponses des archivistes de référence (parfois l'objet, toujours le corps du message)

- les notices descriptives des documents d'archives (souvent les titres, généralement la portée (et contenu)),
- les termes d'indexation.

Dans les courriels, il y a des indices formels qui précèdent les thémanymes tels que l'emploi de la préposition « sur » (p. ex. *je fais des recherches sur, voici des documents qui portent sur*), la préposition « de » avec certains verbes comme *parler de, traiter de* (p. ex. *je cherche des documents qui parlent de*) ou dans un groupe nominal (p. ex. *je recherche des photos de tel événement*), ou la locution « à propos de » (p. ex. *je cherche des documents à propos de*). Dans les notices descriptives, on trouve des formulations telles que *les documents portent sur, la série illustre, le fonds témoigne de*. Le style rédactionnel des notices est parfois direct, surtout au niveau de la pièce : *photographie de.* Parfois, aucun élément formel ne permet d'identifier les thémanymes. Il s'agit alors de repérer les thémanymes par la fonction grammaticale; p. ex., complément d'objet direct d'un verbe de perception visuelle : *voici les photos que nous possédons où l'on voit*…, *on y aperçoit…* Une autre manière d'envisager cette situation linguistique qui permet à un être humain de repérer les thémanymes sans élément formel distinct serait de considérer la nature des actants d'un prédicat selon la sémantique des cadres (Fillmore 1976); nous n'élaborerons pas davantage ici sur ce point.

Pour les termes d'indexation, leur seule présence dans un index et le rôle sémiotique que joue l'index leur confère la valeur de sujet (Guitard 2014). Ils sont constitués d'un thémanyme ou de plusieurs combinés ensemble, ce que l'on constate lors de la segmentation (voir des exemples dans la section 4.1.3.3).

L'analyse grammaticale et l'identification des formules introductives du sujet sont les deux moyens selon lesquels nous avons repéré les thémanymes et les avons isolés du reste des éléments présents dans les sources. La linguistique fournit d'autres éléments permettant d'identifier formellement un sujet. Par exemple, la reprise d'une expression linguistique par des pronoms, surtout en début de phrase, indique que cette expression linguistique est le thème dont il est dit quelque chose dans le texte étudié (Aussenac-Gilles et Séguéla 2000), mais ceci constituerait une étude en soi. Ces marqueurs linguistiques formels pourraient participer à

l'identification automatique de l'expression du sujet dans des données similaires à celles de notre corpus (courriels ou résumés).

La segmentation automatique a été envisagée, mais la diversité des formes de textes à segmenter et le temps de se former à ces outils ont été rédhibitoires. La segmentation a donc été réalisée manuellement, par la chercheuse, à partir de ses connaissances linguistiques en tant que locutrice native du français et en tant que spécialiste de la langue[57], à partir de la méthodologie de l'introspection, classique en linguistique (Talmy 2007).

Après cette première segmentation, nous pouvions identifier parfois, voire souvent, plusieurs catégories sémantiques dans un groupe de mots. Par exemple, les termes d'indexation suivants sont complexes : *Présidentes d'association, Fondation de paroisses, Administration militaire*. Voici d'autres exemples, dans les réponses d'archivistes cette fois : *Commission royale d'enquête sur l'incendie du Laurier Palace et sur certaines autres matières*, ou *ministre des Communautés culturelles et de l'Immigration*.

4.2.5. Caractérisation sémantique

Le but de notre recherche est d'étudier l'écart sémantique présupposé entre le vocabulaire des usagers qui recherchent des documents d'archives patrimoniaux et celui employé par les archivistes dans les instruments de recherche pour le vocabulaire employé pour l'accès thématique aux archives patrimoniales (VATAP). Pour atteindre ce but, nous avons procédé à une dernière étape de prétraitement. Nous n'avons pas comparé toutes les expressions linguistiques porteuses de sujet (thémanymes) d'une source de nos données à l'autre. D'abord, certains sujets n'étaient présents que dans un des deux vocabulaires, soit

[57] Nous détenons une maîtrise recherche en linguistique (nos recherches portaient sur la sémantique parémique), une maîtrise professionnelle en lexicographie, terminographie et traitement automatique de corpus et nous avons collaboré pendant plusieurs années à divers projets lexicographiques (groupe de recherche Franqus : dictionnaire *Usito* 2011, équipe de la maison des dictionnaires Le Robert : *Dictionnaire des combinaisons de mots* 2007) et métalexicographiques (assistanat de recherche sur le traitement des proverbes dans un dictionnaire bilingue français-anglais du XVIII^{e} siècle 2008-2009).

celui de l'usager, soit, plus fréquemment, celui employé dans les instruments de recherche. Ainsi une notice descriptive pouvait contenir de nombreux sujets (p. ex. *Brasserie Molson*, *Pont Jacques Cartier*) dont seulement certains étaient contenus dans la question de l'usager ou la réponse de l'archiviste (p. ex. *Brasserie Molson*). Ensuite, afin d'atteindre notre but de comparer le vocabulaire des usagers et celui employé dans les instruments de recherche, nous avons comparé les thémanymes des termes de recherche (c'est-à-dire dans les questions d'usagers et les réponses d'archivistes) aux thémanymes des termes de description et des termes d'indexation. Ainsi, nous n'avons pas comparé les thémanymes contenus dans les notices descriptives et ceux contenus dans les termes d'indexation puisqu'ils appartiennent tous deux au vocabulaire employé dans les instruments de recherche. Finalement, nous ne voulions pas comparer des thémanymes ne se rapportant pas les uns aux autres (p. ex. *1936* et *Brasserie Molson*). Ainsi, nous voulions comparer des éléments proches sur le plan du sens (p. ex. *1936* et *à différentes époques*), dont des éléments de sens similaires pouvaient être identifiés des uns aux autres, nous verrons lesquels dans cette section.

Nous avons identifié les thémanymes se rapportant à un même sujet à partir d'une caractérisation sémantique. Nous appelons la caractérisation sémantique une forme d'étiquetage sémantique[58] des thémanymes. Elle s'applique à toutes les sources de nos données.

Dans cette section, nous présentons un exemple de filiation des identifiants d'une source à une autre. Puis nous exposons l'analyse en facettes appliquée dans cette recherche en tant que première caractérisation sémantique, suivi de l'analyse par champs sémantiques en tant que seconde caractérisation sémantique.

4.2.5.1. Filiation dans les sources de données

Nous exposons ici un exemple de filiation des identifiants alphanumériques dans une séquence d'analyse. Il s'agit d'un mécanisme de vérification au cours de notre analyse pour

[58] Pour une définition linguistique d'étiquetage sémantique, voir p. ex. Cartier 2009.

assurer la qualité de notre recherche. Cependant, la filiation des identifiants repose sur la filiation sémantique sous-jacente, c'est-à-dire le repérage des éléments de même sens d'une source à l'autre dans le sens de la chaîne communicationnelle : de la question, à la réponse, à la notice descriptive jusqu'au terme d'indexation. La filiation sémantique permet la filiation thématique qui est le repérage des expressions linguistiques porteuses de sujet (thémanymes) d'une source à l'autre dans le sens de la chaîne communicationnelle (voir section 3.2.1).

Ainsi, la question d'usager a014q a pour réponse d'archiviste a014r; celle-ci contient 44 cotes dont l'une correspond à la notice d047 qui contient 14 termes d'indexation dont six contiennent un nom commun : tix03 à tix08. La notice d048 dont la cote est citée elle aussi dans a014r contient 29 termes d'indexation dont neuf contiennent un nom commun : tix09 à tix17. En plus de ces deux notices, la séquence contient dix autres notices retenues dans notre recherche (voir Figure 20).

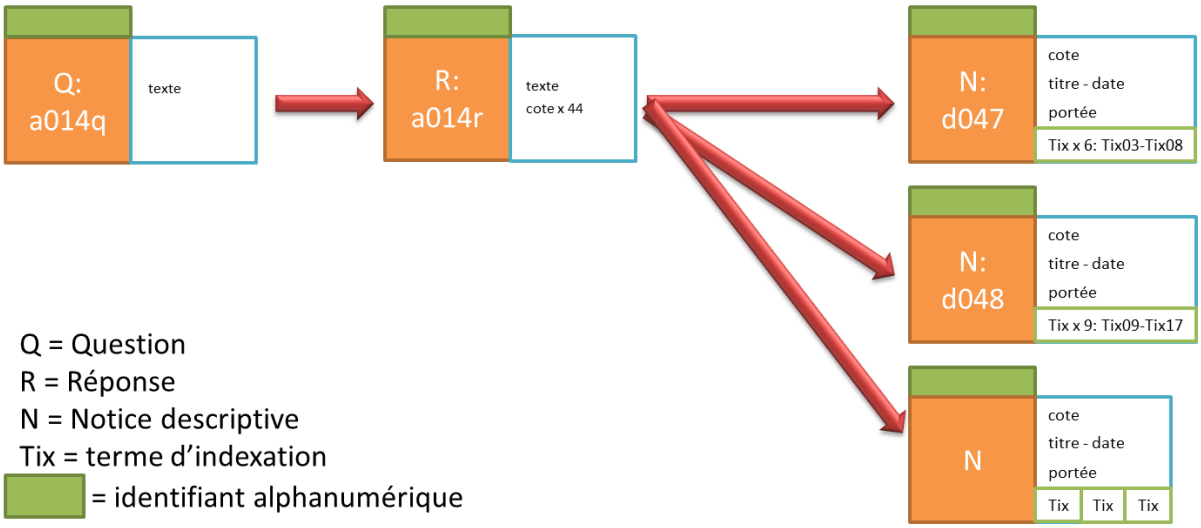

Figure 20. Filiation des identifiants d'une source à l'autre

Parmi ces quatre bassins de termes (les quatre sources, soit Q, R, N et Tix), nous examinons quels thémanymes correspondent à nos critères. Voici un extrait d'une notice descriptive épurée :

> Fonds Romulus Beauparlant . - [Vers 1885-vers 1970] Le fonds permet une étude de l'engagement social d'un homme dans sa paroisse et dans sa ville, durant la première moitié du XXe siècle. Il comprend des documents produits par M. Beauparlant durant

ses études, sa carrière et ses activités sociales. On y retrouve entre autres, sa correspondance personnelle, ainsi que les textes de discours qu'il a prononcés. (d047)

Voici un extrait du fichier de prétraitement de cette notice descriptive. Toutes les facettes – les dimensions d'un sujet – ne sont pas représentées ici.

Tableau XV.Notice descriptive : exemple de la segmentation et de l'analyse en facettes

séquence	code	nom commun	facette	thémanyme
1	d047	1	60	Fonds Romulus Beauparlant
2	d047	0	10	Romulus Beauparlant
3	d047	0	51	[Vers 1885-vers 1970]
4	d047	1	60	Le fonds
5	d047	1	30	l'engagement social
6	d047	0	20	social
7	d047	1	10	un homme
8	d047	1	40	sa paroisse
9	d047	1	40	sa ville
10	d047	1	51	durant la première moitié du XXe siècle
11	d047	1	60	des documents
12	d047	0	30	produits par
13	d047	0	10	M. Beauparlant
14	d047	1	52	ses études
15	d047	1	52	sa carrière
16	d047	1	30	ses activités sociales
17	d047	0	20	sociales
18	d047	1	60	sa correspondance personnelle
19	d047	0	20	personnelle
20	d047	1	60	les textes
21	d047	1	60	discours
22	d047	0	30	qu'il a prononcés

Les facettes qui ne sont pas exprimées par un nom commun (généralement par un adjectif ou un nom propre) sont retirées (0 dans la colonne « nom commun »). Les facettes non thématiques (60 le type de document, 71 le contexte de la recherche, 72 les stratégies de recherche, 80 les formules introductives du sujet) ne sont pas prises en compte (voir section suivante 4.2.5.2). Le numéro de séquence sert seulement à retrouver l'ordre originel du texte de la source.

Reprenons cet exemple avec la séquence complète (QRNTix). Pour la question a014q, nous comptons deux thémanymes issus de la question qui ne contient pas de sujet exprimé par

l'usager lui-même, mais plutôt une liste de cotes archivistiques préparée par l'usager. Les deux thémanymes sont : *du centre-ville* et *d'époque (aucune époque précisée encore)*. Il n'y aucun thémanyme dans la réponse, en effet dans sa réponse l'archiviste propose à l'usager, après lui avoir expliqué comment faire une recherche archivistique avec les outils de son établissement, de préciser sa demande : *Je vous laisse donc faire le tri dans les descriptions sélectionnées et nous soumettre à nouveau votre demande*. La question et la réponse contiennent toutefois 44 cotes archivistiques. Parmi elles, douze notices descriptives ont des termes d'indexation contenant des noms communs (notices d047, d048, d049, d072, d074, d075, d081, d082, d087, d088, d089 et d090). La notice descriptive d047 contient 28 thémanymes. À cette notice sont rattachés 14 termes d'indexation :

Action catholique canadienne
Administration municipale
Anciens de l'université d'Ottawa
Beauparlant, Romulus, 1894-1971
Clubs de hockey
Collège Notre-Dame de Hull
Commission de police du Québec
Fanfares (Musique)
Généalogie
Hull (Québec : Ville)
Juniorat du Sacré-Cœur
Livres comptables
Nationalisme
Société d'histoire régionale du Nord de l'Outaouais

Parmi ceux-ci, six répondent à nos critères :

Administration municipale
Anciens de l'université d'Ottawa
Clubs de hockey
Généalogie
Livres comptables
Nationalisme

et après la segmentation, nous avons huit thémanymes issus des termes d'indexation rattachés à d047 :

Administration municipale
Anciens de l'université d'Ottawa
université
Clubs de hockey

hockey
Généalogie
Livres comptables
Nationalisme

dont un reçoit une facette qui n'est pas thématique : livres comptables (facette 60 : type de document). La notice descriptive d047 descriptive contient alors sept thémanymes. De la même manière, les 12 notices sélectionnées contiennent 171 thémanymes et les termes d'indexation 39 thémanymes. Pour cette séquence (QRNTix), nous obtenons un total de 212 thémanymes à comparer en fonction de leur facette.

4.2.5.2. Analyse en facettes

Pour caractériser sémantiquement les thémanymes, nous avons d'abord attribué de grandes classes générales utilisées pour classer des sujets, telles que le lieu et le temps. Nous nous sommes rapportée à la méthode de l'analyse en facettes développée par Ranganathan (PMEST pour *personnality, matter, energy, space, time*), bien connue en sciences de l'information (voir section 3.3.3 Notion de facettes). Une facette est un terme générique dénotant une dimension d'un sujet. Notre code reprend les facettes de Ranganathan (1957) et inclut donc les éléments thématiques suivants :

- Personnalité (entité);
- Matière (fonction ou attribut);
- Énergie (action ou activité);
- eSpace (lieu);
- Temps (date ou période et événement).

Outre le sujet, nous nous intéressons à des éléments non thématiques présents dans nos données :

- la mise en contexte de la recherche (parfois liée dans le texte à la formulation du sujet de recherche dans un courriel),

- la stratégie de recherche (tentée par l'usager avant sa question ou développée par l'archiviste dans sa réponse à l'usager),

- la forme documentaire que prend le sujet évoqué (le type de documents tel que décrit dans les *RDDA* (p. ex. *document iconographique, plan*), le genre de documents (p. ex. *discours, lettre*), le type d'information (p. ex. *document, information, renseignements*), la quantité de cette forme documentaire (p. ex. *trois pages, 28 documents, 17 dossiers*), éventuellement une cote archivistique complète ou incomplète),

- l'expression du sujet, c'est-à-dire les mots utilisés pour introduire un sujet (voir section 4.2.4.1).

Tableau XVI. Code de l'analyse en facettes

CATÉGORIE **code**	**Description et exemples**
01_PERSONNALITÉ 10_entité	Personne physique ou morale et les regroupements de personnes (p. ex. *association, comité, union*), animal, végétal, chose concrète ou abstraite (p. ex. *établissement, édifice, chemin de fer, hôtel*), discipline scientifique (p. ex. *histoire*, *histoire de l'art (une seule entité)*) P. ex. *Bourrassa, Henri Bourrassa, Monsieur Bourrassa, M. Bourrassa, la famille Bourrassa*, *les alouettes*, *les ponts*, *le pont Jacques Cartier* (une seule entité), *la mode, un prix*.
02_MATIÈRE 20_fonction ou attribut	Fonction d'une entité (p. ex. titre d'un emploi; y compris les relations de famille), attribut en tous genres (p. ex. *potentiel, populaire, administratif ou d'administration, toujours en activité; en bois, en français*)
03_ÉNERGIE 30_action ou activité	Action subie ou effectuée, généralement par une entité; activité. Y compris les sports, les arts, les activités (p. ex. *cuisine, couture*), les maladies. P. ex. *une émission de radio* (au sens de produit fini = entité) *diffusée sur* (action) *ondes courtes* (entité). *La remise* (action) *de prix* (entité).
04_ESPACE 40_lieu	Évocation d'un lieu, associé à un nom propre ou pas (p. ex. *le jardin, la place Bonaventure, la rue Laurier*)
05_TEMPS 51_date ou période	Évocation du temps par une date, un seul point dans le temps (p. ex. *1915*) ou par une période, une durée ou un intervalle (p. ex. *1903-1915*, *à partir de 1940*, *le 19e s.*) Y compris la mention de temps relative à une personne ou à une action : *une nouvelle* (temps) *voiture* (entité); *une ancienne* (temps) *secrétaire* (fonction)
05_TEMPS 52_événement	Évocation du temps par un événement nommé (p. ex. la 2e Guerre mondiale) ou relatif (p. ex. *son enfance*, par contre, *ses études* sera codé par (action) car il y a une activité clairement indiquée) P. ex. *fête, événement, accident, séance, spectacle, attractions foraines*
06_DOCUMENT 60_type, support ou cote	Mention du type de document voulu (p. ex. photographie, vidéo ou image en mouvement, textuel; catégories des *RDDA*) ou du genre de document (discours, reportage, lettre, projet, sommaire, etc.). Y compris les caractéristiques matérielles et l'étendue (p. ex. 3 photos de 10x10cm), la mention du type de support (p. ex. numérique, analogique ou papier). Ou simplement des expressions comme *informations* ou *renseignements* ou encore *documents, copies, dossiers, dossiers,* etc.
07_RECHERCHE 71_contexte de la recherche	Évocation du contexte de la recherche, du but de la recherche, du cadre ou du projet dans lequel prend place la recherche de l'usager et autres informations contextuelles sur le sujet de recherche. P. ex. *Dans le cadre de mon doctorat sur l'histoire de l'art* (contexte de recherche), je recherche des *informations* (document) *sur* (linguistique) les *ingrédients* (entité) des *peintures* (entité) *au 18e s* (temps).

CATÉGORIE code	Description et exemples
07_RECHERCHE 72_stratégie de recherche	Avec succès ou échec, stratégie de recherche indiquée par l'usager ou l'archiviste. Y compris des éléments de réponse de l'archiviste tels que les liens URL vers des ressources ou des références bibliographiques ou l'invitation à aller voir un autre service d'archives plus susceptible d'avoir des documents intéressants. P. ex. Je cherche des photos (type de doc) de Mme Y (entité). J'ai regardé dans le fonds sur M. X, mais je n'ai pas trouvé de photos de sa femme (stratégie de recherche).
08_LINGUISTIQUE 80_expression du sujet	Expression linguistique introduisant un sujet de recherche ou le sujet de documents. P. ex.: *sur, portant sur, à propos de, relatif à, en lien avec, concerner*.

Le Tableau XVI (pour une version complète du tableau, voir Annexe 4 – Protocole de test du codage) est une extraction du code utilisé dans le logiciel QDA Miner (suite Provalis) pour coder les segments de texte issus des quatre sources de notre corpus. Dans QDA Miner, on ne peut pas créer de codes en dehors de catégories et les catégories ne sont pas utilisables pour coder des segments de texte; ainsi, notre codage présente un code, parfois deux, dans une catégorie. Nous avons repris la dénomination des facettes de Ranganathan (1957) pour les cinq premières catégories. Nous avons numéroté les catégories afin que le triage alphanumérique du code les maintienne l'ordre des facettes PMEST. Les cinq premières catégories sont thématiques : personnalité, matière, énergie, espace et temps.

La catégorie du temps se subdivise en deux pour identifier d'une part l'évocation claire du temps et l'évocation d'événements qui ont une base temporelle, mais agrègent d'autres éléments (typiquement personnalité ou action). La sixième catégorie porte sur l'évocation des documents, quelle qu'elle soit : support, type, genre ou cote archivistique. La forme documentaire est souvent imbriquée et participe à l'expression même du sujet. Il pourrait être intéressant d'examiner ultérieurement, dans un plus large corpus, si un même sujet est exprimé de la même façon pour des documents textuels et iconographiques. La septième catégorie porte sur des éléments présents dans les courriels (question et réponse) relatifs à la recherche : le contexte de la recherche et les stratégies de recherche. D'abord, le contexte de la recherche a constitué un bassin de mots permettant de désambigüiser certains mots polysémiques de l'expression du sujet; ils ont permis aussi parfois d'éclairer le sujet lui-même, quand il y a imbrication du sujet de recherche avec la mise en contexte. Prenons en exemple la question d'usager suivante :

> Je fais présentement une recherche historique de l'ilot de l'actuelle Place des Festivals et j'aimerais trouver de l'information sur les bâtiments qui se trouvaient anciennement sur

> le site afin de retracer leurs années de construction et leurs usages. Il me faut de la documentation (cartes, photographies, autres) qui date d'après le grand feu (1852) jusqu'à la construction de la Place de [*sic*] Arts. Pouvez-vous me dire si vous avez des archives pertinentes à ce sujet et si ce serait possible de les consulter ?

L'expression « à ce sujet » renvoie aux segments de texte précédents. La filiation du sujet d'une source à une autre rappelle les travaux sur les chaînes de référence (Schnedecker et Landragin (2014). On appelle *chaîne de référence* une suite d'expressions d'un texte qui se rapportent les unes aux autres et renvoient à la même réalité; elles entretiennent une relation d'identité référentielle (Corblin 1995, 123, cité par Schnedecker et Landragin 2014, 3[59]). Marandin (1988) décrit l'analyse linguistique humaine de la notion de thème dans le récit – la thématisation. Quant à Longo et Todiraşcu (2010, 2014), ils se sont intéressés à la détection automatique des thèmes selon le genre textuel des documents en exploitant la catégorie de marqueurs de cohérence que sont les chaînes de référence; le système qu'ils ont développé recourt à des méthodes linguistiques et statistiques, communes en traitement automatique de la langue (TAL).

Dans le cas particulier présenté ci-dessus, à cause du renvoi d'un segment de phrase à un autre, les éléments de la question d'usager ont été codés doublement : d'abord avec le contexte de la recherche puis avec les éléments des cinq premières catégories à cause de l'expression « à ce sujet ». Ensuite, nous avons identifié les stratégies de recherche entreprises par l'usager ou détaillées par l'archiviste identifiées en tant que telles (par le code 72) ; lors de la discussion, nous verrons que ces éléments nous ont permis de comparer les courriels de notre corpus avec ceux décrits dans la littérature relative à l'étude des courriels de référence afin de vérifier si notre corpus était typique ou non. Ces éléments pourraient être décortiqués et codés plus finement afin de servir à la formation à la référence, par exemple. La huitième et dernière catégorie est linguistique : elle nous permet de recenser les formules introductives du sujet du VATAP et sert au repérage du sujet. Elle était facultative pour les codeurs (voir section 4.4). La raison d'être de cette catégorie et de ce code fait l'objet de la section 4.2.4.1.

---

[59] La référence donnée par les auteurs est la suivante : Corblin, F. (1995). *Les formes de reprise dans le discours. Anaphores et chaînes de référence.* Rennes : Presses Universitaires de Rennes.

Pour déterminer quelles expressions linguistiques expriment les sujets de recherche d'un usager, les sujets de la réponse de l'archiviste de référence, les sujets dont rend compte la notice descriptive (plus exactement, les sections du titre et de la portée) ainsi que les sujets représentés par des termes d'indexation, nous avons exercé notre jugement, encadrée par les normes internationales telles que *ISO 999*:1996 et *ISO 5963* :1985 du Conseil international des archives (ICA), à la manière d'un indexeur.

L'exemple suivant présente l'attribution de facettes à des thémanymes issus de la question, de la réponse, de la notice descriptive ou des termes d'indexation. Rappelons que le code indique la source : *-q* et *-r* à la fin du code indiquent respectivement une question ou une réponse; *d-* ou *tix-* au début du code indiquent respectivement une notice descriptive ou un terme d'indexation. Suivent la facette attribuée au thémanyme et le thémanyme lui-même :

a025q	10	compagnie de La Baie d'Hudson
a025q	10	Pelleteries
a025q	10	Fourrure brute
a025q	30	Chasse
a025q	30	couture
a025q	40	Pêcheries
a025q	40	Salon de couture
a025r	40	pêcheries
a025r	40	Salon de couture
d099	30	teinte
d099	10	fourrures
d099	10	Fourrure
d099	30	Chasse
d099	40	Pêcheries
tix075	10	Fourrures

Si la caractérisation sémantique issue de l'analyse en facettes n'a pas suffi à regrouper les thémanymes, elle nous a parfois servi par la suite à distinguer les acceptions[60] dans les articles de dictionnaires lors de la recherche du champ sémantique.

[60] Acception : Sens particulier d'un mot, admis et reconnu par l'usage. ➞Signification (→ 2. Original, cit. 1). Acception propre ou figurée. Mot à plusieurs acceptions, polysémique. (*Grand Robert* 2017)

4.2.5.3. Identification du champ sémantique

La catégorisation sémantique à partir de l'analyse en facettes n'a pas été suffisante pour nos besoins. En effet, le but de la caractérisation sémantique est de rapprocher seulement les thémanymes comparables pour en analyser le lien sémantique. Or, après la caractérisation sémantique à partir de l'analyse en facettes, nous trouvons regroupées au sein d'une même facette des thémanymes tels que *couture* et *chasse* (dans la facette Activités), *salon de couture* et *pêcheries* (dans la facette Lieux), *fourrure brute*, *pelleterie* ou *(la) mode* (dans la facette Entités). Nous avons essayé une autre caractérisation sémantique pour nous permettre de comparer des thémanymes comparables. Nous avons regroupé les thémanymes par champ sémantique. Le champ sémantique (Polguère 2016, 228 et 2013) est, tel qu'indiqué section 3.3.3, mais cette fois-ci en nos mots, un ensemble de termes dont le dénominateur commun est un élément de sens, servant généralement à dénommer le champ sémantique en question. Par exemple, *immigrants, diverses traditions, francophonie, la recherche interculturelle, Fête nationale du Québec, réfugiés* et *personnes âgées italo-canadiennes* appartiennent au champ sémantique de l'immigration. Le regroupement des thémanymes dans un champ sémantique s'est fait en 3 temps :

(1) le point de départ a été l'introspection (ou l'intuition, conformément à la méthodologie usuelle en linguistique – Talmy, 2007);

(2) nous avons également exploré les définitions de plusieurs dictionnaires généraux (le *Grand Robert*, le *Trésor de la langue française informatisé* (*TLFi*) et *Usito*) et de banques terminologiques (le *Grand Dictionnaire* de l'Office québécois de la langue française et *Termium Plus* du Bureau de la traduction du Canada);

(3) un dernier outil a été utilisé pour valider notre intuition, soit le « dictionnaire des champs lexicaux » d'*Antidote*[61].

[61] Nous avons écrit au Service à la clientèle d'*Antidote* pour connaître leur distinction entre champ sémantique et champ lexical. Ils nous ont répondu ceci : « Le terme de « Champ sémantique » aurait certes pu s'appliquer; d'ailleurs, l'édition anglaise d'*Antidote* appelle ce dictionnaire celui des « Semantic Fields ». Mais le corps

Nous présentons ici un extrait du guide d'utilisation du logiciel afin d'indiquer de quoi est constitué cet outil en particulier :

> Le dictionnaire des champs lexicaux présente l'ensemble des mots sémantiquement apparentés au mot en vedette, regroupés par sens et classés par force relative. Il illustre le voisinage sémantique des mots et permet de naviguer à travers leur écheveau conceptuel. Au total, 69 000 entrées, dont 10 000 noms propres, sont mises en relation au moyen de 2,8 millions de liens. (…)
>
> La liste des mots du champ lexical
>
> Les mots appartenant au champ lexical de la vedette sont présentés en liste verticale et regroupés par catégorie. La nature des liens lexicaux entre ces mots est variée : un champ lexical peut être composé à la fois de cooccurrents, de mots de même famille, de synonymes, d'hyponymes, de méronymes, etc. (…)
>
> Les grandes sections sémantiques
>
> Les entrées sont parfois divisées en grandes sections sémantiques afin de faciliter le repérage. Ainsi, le nom *feuille* comporte quatre sections, chacune formant un champ sémantique : « partie d'une plante », « partie d'une substance », « support » et « document ». (*Posologie,* Druide informatique 2017, 75-76)

Bien des fois, lorsque les dictionnaires n'explicitaient pas un lien, les noms communs principaux des thémanymes à comparer se trouvaient dans le champ sémantique l'un de l'autre. Notons que Polguère (2013) rappelle la distinction entre champ sémantique (classe de lexies, *i. e*. une forme ayant une seule acception) et champ lexical (classe de vocables, *i. e*. une forme pouvant être polysémique).

AMERTUME
A. Saveur amère
Il est des laits filants, d'une viscosité accusée, et des laits amers, d'une amertume de bière.
B.
1. Sentiment (ou caractère propre du sentiment) mêlé de découragement et de rancœur, éprouvé à la suite d'un échec, d'une désillusion. *En ce moment, il y a mélange dans mon âme, mélange d'amertume et de douceur, confusion de miel et de fiel, pêle-mêle étrange.*
2. Caractère mordant, agressif (du langage, du comportement d'une pers.) où se

enseignant québécois, à qui ce dictionnaire était destiné en priorité, connaissait et utilisait la notion de « Champ lexical », que nous avons retenue. » (communication personnelle, 29-08-2017). Nous préférons recourir à la notion de champ sémantique dans ce travail.

reconnaît de la rancœur, du ressentiment. *Tout ce qu'il y avait en moi de légèreté, de vanité, de puérilité, de sécheresse, d'ironie ou d'amertume d'esprit pendant ces mauvaises années de mon adolescence disparaissait tellement que je ne me reconnaissais plus moi-même.*

Figure 21. Vocable et lexies : l'exemple d'amertume dans le TLFi

Une entrée de dictionnaire est un vocable : elle rassemble tous les sens d'une forme. Ainsi, le vocable *amertume* tel que décrit par le *TLFi* (voir Figure 21) rassemble les lexies *amertume*-saveur amère, *amertume*-sentiment et *amertume*-caractère mordant.

De notre côté, nous avons utilisé le champ sémantique – *i. e*. le rassemblement de thémanymes ayant un élément de sens commun : nous ne nous intéressons pas aux autres acceptions d'une forme que celles qui sont actualisées dans notre corpus. Ici, l'encodage des facettes et le retour aux données textuelles ont plusieurs fois été utiles pour se concentrer sur une seule acception de termes polysémiques (p. ex. *couture* l'activité et *couture* le résultat).

Dans notre étude, l'écart sémantique a été calculé à l'intérieur d'un champ sémantique. Les thémanymes ont été rapprochés lors de la caractérisation sémantique par le degré maximal d'éloignement de notre échelle d'écart sémantique, à savoir le champ sémantique (voir section 4.3.1). Cet élément qui a émergé spontanément lors de l'analyse nous apparaît à présent déterminant et incontournable pour la cohérence de notre recherche.

La caractérisation sémantique a consisté en l'attribution d'une facette puis d'un champ sémantique à chaque thémanyme dans le corpus 1 et seulement en l'attribution d'un champ sémantique pour le corpus 2. Cette différence tient au fait que l'analyse en facettes ne permettait pas une caractérisation sémantique discriminant suffisamment les thémanymes. Différents classements d'un même échantillon de nos données permettent de rendre visible le bénéfice de cette seconde caractérisation sémantique : par source, par facette et par champ sémantique. Les tableaux suivants (Tableau XVII, Tableau XVIII, Tableau XIX) présentent ces classements pour une même séquence (QRNTix) relative à l'histoire de la mode.

Tableau XVII. Classement des thémanymes par source

source	code	facette	thémanyme	champ sémantique
Q	a025q	10	compagnie de La Baie d'Hudson	compagnie
Q	a025q	10	compagnie de Baie d'Hudson	compagnie
Q	a025q	10	Pelleteries	animaux
Q	a025q	10	Fourrure brute	animaux

source	code	facette	thémanyme	champ sémantique
Q	a025q	30	Chasse	animaux
Q	a025q	30	couture	couture
Q	a025q	40	Pêcheries	animaux
Q	a025q	40	Salon de couture	couture
R	a025r	40	pêcheries	animaux
R	a025r	40	Salon de couture	couture
N	d099	30	teinte	animaux
N	d099	10	fourrures	animaux
N	d099	10	Fourrure	animaux
N	d099	30	Chasse	animaux
N	d099	40	Pêcheries	animaux
Tix	tix075	10	Fourrures	animaux

Le classement par code permet d'identifier le nombre de thémanymes par source dans une séquence (Q : 8 thémanymes, R : 2 thémanymes, N : 5 thémanymes, Tix : 1 thémanymes). Nous reparlerons de la quantification dans la description de la méthode d'analyse des quantités (voir section 4.3.3). Rappelons que notre code d'analyse en facettes comprend cinq facettes thématiques et trois non-thématiques. Les facettes thématiques présentes dans cette séquence sont 10-Personnalité, 30-Activité et 40-Lieu. Nous présentons les mêmes données en regroupant par facette dans le Tableau XVIII.

Tableau XVIII. Classement par facette

facette	source	code	thémanyme	champ sémantique
10	Q	a025q	compagnie de La Baie d'Hudson	compagnie
10	Q	a025q	compagnie de Baie d'Hudson	compagnie
10	Q	a025q	Pelleteries	animaux
10	Q	a025q	Fourrure brute	animaux
10	N	d099	fourrures	animaux
10	N	d099	Fourrure	animaux
10	Tix	tix075	Fourrures	animaux
30	Q	a025q	Chasse	animaux
30	Q	a025q	couture	couture
30	N	d099	teinte	animaux
30	N	d099	Chasse	animaux
40	Q	a025q	Pêcheries	animaux
40	Q	a025q	Salon de couture	couture
40	R	a025r	pêcheries	animaux
40	R	a025r	Salon de couture	couture

facette	source	code	thémanyme	champ sémantique
40	N	d099	Pêcheries	animaux

Par exemple, après le classement par facette, on aurait dû comparer dans cette séquence les thémanymes d'activité (30) *Chasse* (deux fois : une fois dans la question et une dans la réponse), *couture* et *teinte* ensemble, thémanymes issus de deux types de données (question et notice). N'arrivant pas à déterminer un élément de sens commun, cette comparaison nous a semblé insuffisante. Les champs sémantiques sont au nombre de trois dans cette séquence : compagnie, animaux et couture. Nous présentons les mêmes données par champ sémantique dans le Tableau XIX.

Tableau XIX. Classement par champ sémantique

champ sémantique	source	code	facette	thémanyme
compagnie	Q	a025q	10	compagnie de La Baie d'Hudson
compagnie	Q	a025q	10	compagnie de Baie d'Hudson
animaux	Q	a025q	10	Pelleteries
animaux	Q	a025q	10	Fourrure brute
animaux	N	d099	10	fourrures
animaux	N	d099	10	Fourrure
animaux	Tix	tix075	10	Fourrures
animaux	Q	a025q	30	Chasse
animaux	N	d099	30	teinte
animaux	N	d099	30	Chasse
animaux	Q	a025q	40	Pêcheries
animaux	R	a025r	40	pêcheries
animaux	N	d099	40	Pêcheries
couture	Q	a025q	30	couture
couture	Q	a025q	40	Salon de couture
couture	R	a025r	40	Salon de couture

Le classement par champ sémantique regroupe des thémanymes de plusieurs sources et de plusieurs facettes. Ces thémanymes sont plus proches thématiquement que dans les autres classements puisqu'elles partagent un élément de sens. Le champ sémantique joue donc un double rôle. Il est l'unité minimale de sens qui est à la base des rapprochements de thémanymes qui, eux-mêmes, serviront à l'identification de l'écart sémantique. Ce dénominateur commun – de nature sémantique – permet d'établir une filiation sémantique entre les thémanymes lorsque l'on passe d'une source à l'autre. Cette filiation sémantique se

fait naturellement par un être humain sur la base d'éléments implicites, internes aux mots, que nous avons tenté de mettre au jour par le prétraitement.

Les champs sémantiques sont identifiés dans la question d'usager et la réponse d'archiviste puis cherchés dans les notices descriptives et les termes d'indexation attachés à chacune de ces notices. Nous n'examinons pas le lien sémantique entre le vocabulaire de la notice et celui des termes d'indexation, deux types d'instruments de recherche établis par l'archiviste. Cette analyse ne rendrait pas compte de l'écart sémantique entre le vocabulaire des usagers et celui des archivistes, mais plutôt de la complémentarité des instruments de recherche, ce qui ne fait pas partie de nos objectifs de recherche.

Nous pensions initialement effectuer d'abord la segmentation puis la caractérisation sémantique. Il est apparu que les deux opérations étaient intimement liées dans notre esprit. Après plusieurs essais non concluants de dissociation des opérations, nous avons décidé de les mener de manière concomitante et récurrente (au moins deux fois à trois semaines d'intervalle).

Conclusion

À partir d'une approche linguistique des données, le prétraitement a consisté en une série d'étapes ayant pour but de comparer des thémanymes issus de sources différentes, mais ayant en commun un élément de sens. Pour ce faire, nous avons procédé à l'anonymisation des courriels, à l'épuration de toutes les sources, à leur segmentation et caractérisation sémantique selon l'analyse en facettes et/ou l'identification du champ sémantique. Ces opérations de prétraitement ont préparé les données à l'analyse. Nous avons complété ces étapes manuellement. Pour les accomplir avec soin et nous assurer de la qualité de notre recherche au fur et à mesure, ces étapes ont nécessité beaucoup de temps. Une automatisation partielle de la méthode constituerait un préalable à la récurrence de ce type de recherche.

Nous avons rapproché les thémanymes appartenant à un même champ sémantique. Cela nous a permis une comparaison ajustée des expressions linguistiques issues des quatre sources de données de notre corpus (questions des usagers, réponses des archivistes, notices descriptives, termes d'indexation). Cette comparaison a été le point de départ de notre analyse dont la méthode est exposée dans la section suivante.

4.3. Méthode d'analyse

Après la collecte et le prétraitement, nous présentons dans cette section la méthode d'analyse des données. Elle s'est faite en quatre étapes, répondant à chacun de nos objectifs dont découlent nos questions de recherche. Chaque étape est détaillée ci-dessous :

1. Valider empiriquement l'existence de l'écart sémantique présupposé en relevant les reprises et les non-reprises entre les différentes sources du corpus (OR1, QR1);
2. nommer les relations sémantiques (OR2, QR2);
3. calculer la fréquence de l'écart sémantique (OR3, QR3), et des relations sémantiques (OR3, QR4);

4.3.1. Comparaison des expressions linguistiques (QR1)

Pour vérifier l'écart sémantique présupposé en comparant les thémanymes entre les différentes sources du corpus, nous avons comparé des éléments qui se rapportent les uns aux autres. La comparaison est le sixième des dix grands principes de l'analyse qualitative selon Tesch (1990, 95-97). Elle nous a permis de faire émerger les relations sémantiques entre les expressions linguistiques porteuses de sujet (thémanymes) de sources différentes.

Nous avons procédé à la comparaison des thémanymes pour vérifier empiriquement l'existence d'un écart sémantique présupposé, sur la base d'une variation de forme et de sens entre les thémanymes. Ces derniers ont été comparés deux à deux (Q-R, Q-N, Q-Tix, R-N et R-Tix), manuellement, à l'intérieur d'une séquence d'analyse. Une séquence d'analyse est constituée des sources présentes dans le corpus : QRNTix pour le corpus 1 et QRN pour le corpus 2.

Nous avons établi une échelle des variations à laquelle nous attribuons des valeurs symboliques qui nous ont permis de quantifier l'écart sémantique entre deux expressions linguistiques de deux sources différentes.

Échelle d'écart sémantique

L'échelle d'écart sémantique nous a permis d'évaluer la variation de forme et de sens entre les expressions linguistiques porteuses du sujet (thémanymes). Elle a été élaborée à partir

des grilles examinées dans la littérature (voir section 3.5.3). Nous avons appliqué cette grille dans un autre projet de recherche mené conjointement avec notre directrice de recherche; ce projet porte sur les index de fin de livre qui s'apparente grandement aux index d'archives (voir Da Sylva, soumis). Au fur et à mesure de l'analyse, nous l'avons affinée en fonction des données. L'échelle se présente sous la forme d'un tableau (voir Tableau XX). La première colonne indique la valeur de l'écart associé à la relation. La deuxième et la troisième colonne expriment une typologie des relations sémantiques : type et éventuellement sous-type. Un exemple ou parfois un commentaire permettent d'illustrer ou d'expliciter la relation. Dans la perspective de présenter notre travail à des archivistes non linguistes, nous avons essayé de vulgariser la relation, à la manière de Buckland (1999) (voir section 3.5.3).

Plus l'écart est grand entre les formes, plus la valeur est importante. Mais un échelon au début de l'échelle ne représente pas le même éloignement sémantique, la même distance qu'un échelon de la fin de l'échelle; les échelons de l'échelle d'écart sémantique ne sont pas équidistants. La comparaison de deux thémanymes identiques résulte en l'écart nul (écart 0) : nous avons appelé cette relation, la relation d'identité.

D'abord précisons que si l'expression n'est présente que dans une source, alors nous avons affaire à une forme isolée ou un isolat; puisque ce thémanyme ne peut pas entretenir de relation avec un autre thémanyme, il n'est pas comptabilisé dans l'écart. Il peut s'agir d'un élément de la question qui est redondant, qui n'a pas de correspondance dans les collections du service d'archives où la question a été posée, d'un élément de la réponse de l'archiviste qui tente de compléter ou de préciser le sujet de la question, ou encore d'un élément de la notice qui ne répond pas à la question si le document couvre plusieurs sujets.

Ensuite, sans reprendre tous les éléments de l'échelle, nous soulignons quelques éléments. La variation orthographique (écart 2) nécessite normalement minimum deux locuteurs pour être réalisée. On ne s'attend pas à ce qu'une même personne emploie deux graphies d'un même mot dans le même document. Le champ sémantique (écart 13) – en plus d'être la méthode de caractérisation sémantique des thémanymes commune aux deux corpus – est le niveau maximal de notre échelle de comparaison des thémanymes, utilisée pour identifier l'écart sémantique entre deux thémanymes lors de leur comparaison.

Tableau XX. Échelle d'écart sémantique

Écart	Type	Sous-type	Exemple ou commentaire	Vulgarisation
0	Identité		Deux expressions linguistiques identiques	• même forme, même sens
1	Variation orthographique		• clé, clef	• presque même forme (orthographe), même sens
2	Variation de longueur	Sigle et expansion	• TAL / traitement automatique de la langue	• presque même forme (forme courte), même sens
		Abréviations	• moto / motocyclette	
3	Variation graphique	Guillemets	• indexation orientée usager / indexation orientée « usager »	• presque même forme (signe typographique), même sens
		Autre signe typographique	• vedettes matière / vedettes-matière	
4	Variation de casse		• brasserie Molson / Brasserie Molson	• presque même forme, (presque) même sens
5	Variation flexionnelle	Sur la tête	• prison / prisons	• presque même forme, (presque) même sens
		Sur le complément	• protection des enfants / protection de l'enfant • satisfaction de l'usager / satisfaction des usagers	
6	Variation dérivationnelle	Mots de la même famille	• utilisation / utiliser; financement / financer	• lien de forme (dérivation/racine), presque même sens
			• troupes canadiennes / Canadiens	
		Même tête	• flux du sang / flux sanguin	
			• protection des enfants / protection de l'enfance	
7	Variation syntaxique	Possessif	• besoins de l'utilisateur / ses besoins	• lien de forme (syntaxe), presque même sens
		Déterminant	• évaluation du besoin / évaluation d'un besoin	
		Omission ou ajout d'un déterminant	• recherche de l'information / recherche d'information	
		Insertion / suppression adjectivale / adverbiale / prépositionnelle	• soutien aux utilisateurs / soutien adéquat aux utilisateurs ; canal de Lachine / canal Lachine	
		Redondance	• la Ville de Montréal / Montréal	
8	Ellipse d'un complément		• support de documents audiovisuels / le support	• forme différente ou partielle, presque même sens ou sens plus spécifique
9	Paraphrase	Glose / définition	• Zouave / La garde paroissiale s'inspire des corps de gardes du Vatican formés au XV^{e} siècle pour servir et protéger le pape	• forme différente, même sens

Écart	Type	Sous-type	Exemple ou commentaire	Vulgarisation
10	Synonymie ou équivalence de la tête	Synonymie (ou Antonymie)	• vélo / bicyclette • usager / utilisateur • vainqueur / vaincu	• forme différente, même sens (ou sens opposé)
11	Hiérarchie	Méronymie	• guidon / vélo ou vélo / guidon; 1600 rue Notre-Dame Est / rue Notre Dame	• forme différente ou partielle, sens plus spécifique
		Hyponymie	• chat / animal	• forme différente, sens plus spécifique
		Hyperonymie	• animal / chat	• forme différente, sens moins spécifique
		Co-hyponymie	• jument / étalon (co-hyponyme de cheval)	• forme différente, plusieurs éléments de sens communs (même hyperonyme)
12	Relation associative	cause/effet action/agent action/produit action/objet action/lieu agent/lieu science/objet objet/propriété objet/application objet/matériau, etc.	• récolte / culture (action / objet) • enseignement / apprentissage (concepts complémentaires)	• forme différente, un élément de sens commun précis
13	Champ sémantique		• technologie / logiciel • détérioration du support / les frictions causent la modification du signal	• forme différente, un élément de sens commun vague
14	Relation référentielle	Correspondance référentielle	• étoile du matin / étoile du soir (Vénus)	• forme différente, sens différent, même référent

Notons quelques précisions. À l'écart 5, la tête est le nom commun principal d'un thémanyme. En linguistique, on emploie *tête* ou *noyau* pour le mot principal d'un syntagme ou groupe de mots. À l'écart 9, la périphrase est une définition. Mais cette définition présente dans une notice descriptive est erronée. En effet, le mot *zouave* apparaît vers 1830 et c'est en 1871 que les zouaves pontificaux ont été créés en corps de gardes au Vatican (*Grand Robert* 2017 et *TLFi* 2017). À l'écart 12, les couples de relations associatives et notre sélection parmi celles-ci ont été identifiés à la section 3.5.3 Les relations sémantiques.

Lors de l'analyse, nous avons ressenti le besoin d'introduire la relation référentielle (écart 14). En effet, en comparant des thémanymes de deux sources (p. ex., question d'usager

et notice descriptive) – par exemple *centre-ville* et *île* qui désignent le même endroit dans le contexte d'apparition de ces mots – il nous est apparu que le seul lien qui pouvait expliquer la recommandation par l'archiviste de la notice reposait sur la correspondance référentielle. La relation référentielle (écart 14) a été identifiée lors de l'analyse. Toute personne qui lirait les documents (ici, la question d'usager et la notice descriptive) serait en mesure d'identifier qu'il existe un lien (référentiel), mais le lien n'apparaît pas dans un dictionnaire ni dans le champ lexical d'*Antidote*. Elle dépasse le cadre de recherche de cette thèse étant donné que sa nature n'est pas sémantique, mais plutôt encyclopédique et contextuelle.

Les relations sémantiques « peuvent être vues comme des liens orientés entre les concepts ou entités qui participent de la relation » (Khoo et Na 2006, 159, n. t.) Elles s'expriment dans un sens et, pour celles qui relient des unités qui ne sont pas similaires (contrairement à la synonymie et l'antonymie), changent de nom selon leur orientation : hyperonymie *versus* hyponymie, méronymie *versus* holonymie, relation de cause à effet, etc. Ainsi, dans notre étude, la relation s'opère dans le sens de la chaîne documentaire archivistique (de l'usager au document) : question, réponse, notice descriptive, terme d'indexation. L'orientation de la relation est préservée dans notre échelle.

Nous avons élaboré l'échelle d'écart sémantique à la fois pour notre recherche doctorale au moment de la mise en place de la méthodologie ainsi qu'avec notre directrice de recherche, pour une recherche sur les index de fin de livre (Guitard et Da Sylva 2014; recherche en cours). Parmi toutes ces listes de relations associatives, nous avons retenu les suivantes parce qu'elles appartenaient à plusieurs listes vues dans la littérature ou bien parce qu'elles étaient présentes dans les premiers cas traités dans les deux recherches auxquelles nous participions et semblaient donc avoir une certaine fréquence d'occurrence. Les relations associatives que nous avons retenues sont les suivantes :

- Cause/effet
- Action/agent
- Action/produit
- Action/objet
- Science ou discipline ou entité/objet
- Objet/propriété
- Objet/application
- Objet/matériau

Encore une fois, les relations sont orientées. Ainsi celles-ci ne sont pas toujours réciproques, contrairement aux relations syntagmatiques ou hiérarchiques. Notre liste est assez restreinte dans le but de rendre compte avec exactitude du lien sémantique plutôt que d'avoir une classe aux limites vagues et permissives. Ainsi, les liens tels que discipline / lieu (santé / le milieu de la santé) ou discipline / acteur (santé / le défunt Comité de la santé mentale du Québec (CSMQ)) n'ont pas été inclus dans la classe des relations associatives, mais dans le champ lexical. Également, les liens sémantiques couvrant plusieurs relations, ou plusieurs « nœuds » d'un hypothétique thésaurus (santé / les infections ; précision qui explique les nœuds : antonyme/hyper (maladie) + hypo), ont été inclus dans le champ lexical.

La comparaison a constitué l'étape de base sur laquelle le reste de l'analyse a porté. Une fois les paires de thémanymes comparées, nous avons nommé les relations sémantiques à partir de la grille (section 4.3.2). Ensuite, la quantité de relations d'identité en nombre et en pourcentage a été calculée (section 4.3.3). Nous avons ainsi pu déterminer s'il y avait un écart sémantique entre le vocabulaire des usagers et celui employé dans les instruments de recherche (voir section 5.2) et de quelles nature et importance il était (5.3).

4.3.2. Nommage des relations sémantiques (QR2)

Nous avons nommé les relations sémantiques qu'entretiennent les thémanymes entre deux sources de données de notre corpus se rapportant au vocabulaire des usagers ou à celui employé dans les instruments de recherche. Nous nous sommes reportée à Aussenac-Gilles et Seguela (2000), Mortureux (2008), Lehmann et Martin-Berthet (2013) et Polguère (2016) notamment (voir section 3.5.1.2.) pour l'identification des liens sémantiques entre les thémanymes étudiés. Le nommage de la relation sémantique sert à indiquer le type ou la signification de la relation (Khoo et Na 2006, 159). Après avoir identifié le lien sémantique entre deux thémanymes, nous nous sommes reportée à l'échelle d'écart sémantique pour le caractériser.

Le nommage de relations sémantiques n'est pas une tâche facile. Dancette souligne « les difficultés à observer une cohérence totale dans l'attribution des étiquettes de relations et la part d'arbitraire qui en découle » (2011, 289). « L'attribution d'une étiquette de RS [relation sémantique] à un couple de termes est une opération délicate. » (2011, 295) Pour assurer la

qualité de sa recherche, Dancette a recours à des classes de relations sémantiques bien définies au préalable et tâche de faire preuve de systématicité tout au long du processus de nommage de la relation sémantique entre deux termes. De notre côté, nous nous sommes reportée à l'échelle d'écart sémantique pour nommer les relations de manière constante.

Nous avons identifié les relations sémantiques entre les thémanymes de notre corpus, en fonction de l'auteur du discours (usagers / archivistes). La distinction relative à l'étendue linguistique initiale (voir Tableau IV, section 4.1.2.2), c'est-à-dire entre les mots isolés et les phrases n'a plus lieu d'être à partir du moment où les thémanymes ont été établis. De plus, nous avons étudié séparément les données émanant de chaque établissement afin de pouvoir envisager les phénomènes observés dans chacun d'eux.

Suit la description du calcul de la fréquence des résultats obtenus qualitativement sur laquelle s'effectue la sélection des relations sémantiques. Le calcul de la fréquence est nécessaire pour discriminer les relations les plus courantes des plus rares. Nous nous sommes intéressée aux relations les plus fréquentes dans le corpus lors de l'interprétation des résultats du VATAP.

4.3.3. Calcul de la fréquence (QR3, QR4)

En nous basant sur les résultats de l'analyse qualitative de nos données, nous souhaitions d'abord connaître la fréquence de la correspondance exacte des thémanymes entre le vocabulaire des usagers et celui employé dans les instruments de recherche (la relation d'identité) et par là même connaître la fréquence des cas d'écart sémantique, quelle que soit la nature de la relation sémantique qui constitue l'écart (Q3). Ensuite, nous souhaitions connaître la fréquence de chacune des relations sémantiques qui caractérisent un écart sémantique entre les thémanymes entre le vocabulaire des usagers et celui employé dans les instruments de recherche (Q4). Identifier les patrons récurrents à partir du codage fait partie du processus d'analyse qualitative (Zhang et Wildemuth 2009). La fréquence déterminée à cette phase de la recherche participe de la description de l'écart sémantique.

Nous avions envisagé la lexicométrie comme méthode pour quantifier nos résultats qualitatifs. « La lexicométrie est un ensemble de méthodes d'analyse appliqué aux corpus textuels » Fiala (1994, 113). Tournier, dans l'article intitulé « D'où viennent les fréquences de

vocabulaire ? La lexicométrie et ses modèles », se pose une question directement reliée à notre recherche : « Étant donné un inventaire de fréquences dans un corpus et ses partitions, peut-on en inférer des interprétations « en langue » ? » (Tournier 1980, 189). Des chercheurs ont publié plus récemment sur l'usage de la lexicométrie dans les sciences sociales et humaines. Glady et Leimdorfer (2015) font une synthèse des présentations de deux journées d'études sur les usages de la lexicométrie et l'interprétation sociologique. Leblanc (2015) propose une réflexion sur les outils d'analyse des données et compare la lexicométrie et la statistique textuelle pour proposer une démarche expérimentale. Finalement, les auteurs phares en la matière, Labbé et Labbé rappellent que « la lexicométrie est l'alliance des sciences du langage, des statistiques et de l'informatique. Elle permet de traiter de vastes ensembles de textes (corpus), d'établir leur vocabulaire, de classer les vocables en fonction de leur fréquence, de leur répartition, de leurs catégories grammaticales. » (2013, 1) La lexicométrie ne se limite pas au calcul de la fréquence des mots sur quelques exemples contrairement à ce que nous souhaitons faire dans cette étude. Ainsi, nous constatons qu'elle est trop puissante pour les résultats de nature quantitative que nous cherchions à obtenir. Nous avons donc calculé la fréquence d'apparition des relations sémantiques au sein des séquences (QRNTix ou QRN), polyséquences (QRNTixNTixNTix ou QRNNN), corpus 1 (dix questions d'usagers), corpus 2 (20 questions d'usagers), corpus général (30 questions d'usagers) ainsi qu'au sein de chaque établissement ayant participé à notre recherche. Ceci nous a permis de rendre compte de l'ampleur de l'apparition des relations sémantiques identifiées dans l'échelle d'écart sémantique.

4.4. Qualité de la recherche

Dans cette section, nous présentons les dispositions prises pour assurer la qualité de notre recherche. Puisque notre approche est qualitative (voir section 4), nos critères de qualité sont ceux d'une recherche qualitative (Pickard 2013). D'abord, nous exposons les critères de qualité d'une recherche qualitative (section 4.4.1). Ensuite, pour renforcer la qualité et notamment la fiabilité de notre recherche, nous avons exposé notre méthodologie à plusieurs reprises; la section 4.4.1 retrace ces éléments. La section 4.4.2 indique les procédures de vérification mises en place pour pallier les erreurs humaines dans notre traitement. Nous

rappelons la nature sensible des données (section 4.4.3) et les mesures prises pour assurer le respect des lois et de l'éthique scientifique (section 4.4.4) dans les deux dernières sections : ces mesures participent à l'appareil méthodologique et à la qualité de la recherche.

4.4.1. Critères de qualité

Plusieurs critères permettent à des chercheurs d'évaluer la qualité d'une recherche qualitative telle que la nôtre. Pickard (2013, 20-22) présente quatre critères couramment adoptés de nos jours en recherche qualitative : la crédibilité, la transférabilité, la fiabilité et la confirmabilité. Nous les décrivons ci-dessous en identifiant quels moyens nous avons mis en œuvre dans notre recherche pour les remplir.

La crédibilité est rattachée à la valeur de vérité de la recherche, elle est démontrable selon Pickard (2013, 21) par la triangulation des données, de leurs sources et des chercheurs ainsi que par la multiplicité des techniques de collecte. Nos données proviennent de trois établissements sélectionnés et d'usagers inconnus. Les questions d'usagers nous ont été transmises par les établissements. Les notices descriptives ont été récupérées par la chercheuse directement dans le catalogue public en ligne de chacun des établissements. Nous avons collecté des données en ligne (les notices et les termes d'indexation). Les données ont été créées par des personnes de divers milieux, à diverses époques. Nous sommes la chercheuse principale, mais nous avons été supervisée lors de cette recherche par un comité de recherche doctoral formé de notre directrice de recherche, de notre co-directrice et d'un professeur de l'École de bibliothéconomie et des sciences de l'information.

En outre, la crédibilité de notre recherche est renforcée par la présentation de la méthodologie ou des résultats préliminaires auprès des pairs. Nous avons présenté les concepts linguistiques de notre recherche aux journées de linguistique, colloque étudiant à Québec en 2013, puis, quelques mois plus tard, à Res per nomen, colloque sur la linguistique et la philosophie, à Reims (France). Cette année-là, le colloque portait sur la notion de référence en linguistique et était organisé en hommage aux travaux de Georges Kleiber, dont les écrits ont nourri notre pensée. En outre, nous avons présenté les critères de sélection de notre corpus (méthode d'échantillonnage) et la méthode de caractérisation sémantique (analyse en facettes) au colloque VocUM, colloque des étudiants des cycles supérieurs de l'Université de Montréal

dédié à la langue et à la communication, organisé par les étudiants du département de linguistique et de traduction, à l'automne 2016. La présentation de notre recherche à divers moments devant un public davantage linguiste qu'archiviste nous a permis d'explorer plus en profondeur cet aspect de notre recherche. La manière dont s'exerce la référence dans l'index de documents d'archives et l'introduction aux rouages sémantiques et sémiotiques de l'index ont été présentés et publiés dans les actes du colloque Res per nomen IV (Guitard 2014).

La méthode d'analyse des termes a été éprouvée au préalable sur un corpus de termes d'indexation d'index de fin d'ouvrage de la collection *Gestion de l'information* des Presses de l'Université du Québec (voir section 4.3.1). Cette recherche menée conjointement avec notre directrice de recherche a permis de préciser certains paramètres utilisés dans cette recherche. La propriété linguistique qu'est l'opposition dénomination/désignation mise en jeu dans les index – dont l'étude était initialement incluse dans cette recherche, mais qui a finalement été reportée à des recherches futures – a été présentée au congrès de l'Acfas 2014.

Nous avons présenté des résultats préliminaires relatifs aux relations sémantiques entre les termes de questions d'usagers, de réponses d'archiviste et de notices descriptives au congrès de l'Association des archivistes du Québec (AAQ) en juin 2016. Cette communication a donné naissance à une publication sur le thème de l'archiviste, médiateur sémantique (Guitard 2018).

La transférabilité est rattachée à l'applicabilité de la recherche. Afin de pouvoir évaluer si les résultats de notre recherche seront transférables à d'autres milieux, nous décrivons en détail les milieux où nous avons effectué notre collecte ainsi que les données elles-mêmes. Le vocabulaire est caractérisé par un lieu, une époque et des personnes particulières; cependant, nous espérons avoir pu dégager une certaine transférabilité de nos résultats sémantiques, éventuellement pour d'autres services d'archives du Québec où des locuteurs s'expriment en français de nos jours. Bien que nous ayons diversifié les milieux de collecte, nous ne savons pas si les données collectées sont représentatives de données disponibles dans des centres et services d'archives plus modestes, à une mission moins large, aux fonds et collections moins nombreux et peut-être moins diversifiés. La spécificité du vocabulaire employé par les usagers est probablement plus importante dans un milieu déjà associé à une thématique en particulier (p. ex. un service d'archives dans un musée de l'architecture).

La fiabilité est rattachée à la cohérence de la recherche et à la constance des chercheurs. Il s'agit de démontrer que les méthodes et les techniques employées par la chercheuse sont adéquates au projet et justifiées. Notre démarche de recherche est explicitée dans ce texte. Nous avons tenu un journal de bord. Les tests inter- et intra-codeurs sont reconnus en sciences sociales (Zhang et Wildemuth 2009) pour assurer une constance dans la recherche, un recoupement entre des jugements individuels et permettre de trouver un accord sur le principe de l'intersubjectivité partagée. En particulier, le test inter-codeur doit être appliqué lors de la préparation du code afin de le rendre plus robuste avant le codage de toutes les données (Zhang et Wildemuth 2009, 310-313). Nous avons recouru à un test intra-codeur (la chercheuse a codé les données à plusieurs reprises espacées dans le temps pour évaluer si elle codait d'une manière similaire à chaque fois) pour la sélection des questions d'usagers, l'identification des facettes et l'établissement des relations sémantiques. Nous avons eu recours également à un test inter-codeur pour l'identification des facettes. Nos deux directrices ont utilisé le code du protocole du test de codage (voir Annexe 4). Le test portait à la fois sur la segmentation et la caractérisation sémantique par l'analyse en facettes, si bien que le taux d'accord était faible (environ 30%) entre la chercheuse et l'une ou l'autre des deux codeuses. Celles-ci et la chercheuse se sont rencontrées pour discuter des différences de perception du code. Nous avons affiné nos catégories, ajouté des exemples et segmenté le texte au préalable avant de resoumettre le test trois semaines plus tard sur un échantillon partiellement différent aux deux mêmes codeuses pour atteindre un taux suffisant (75%).

La confirmabilité est rattachée à la neutralité de la recherche. Bien que nous apposions inévitablement un biais aux données, nous avons tenté de montrer dans la section méthodologie comment les données ont été analysées (en particulier dans la section 4.1.3). Le test intra-codeur nous a aussi permis d'évaluer si l'analyse que nous tirions des données correspondait bel et bien au contenu des données. Également, chacune des étapes de prétraitement, soit la transformation des textes en des expressions linguistiques plus courtes et en thémanymes, ont clairement été identifiées et décrites (voir section 4.2). Finalement, nous avons utilisé des sources autorisées comme référence linguistique : des dictionnaires et des ressources terminologiques publiées nous ont aidée à déterminer le sens des thémanymes et à identifier les facettes (voir section 4.2.5).

4.4.2. Regard critique sur la recherche

Cette section permet de poser un regard critique sur la recherche effectuée afin d'estimer la confiance que l'on peut porter à sa qualité.

Les données ont été collectées de manière similaire pour les trois établissements, cependant, il est à noter que l'établissement 3 ne disposait d'une boîte courriel dédiée à la référence que depuis peu ; nous avons récupéré seulement 6 mois de données au lieu de seize mois dans les deux autres établissements. Malgré un nombre restreint de courriels à la référence, nous avons obtenu des résultats intéressants pour notre étude exploratoire du vocabulaire pour l'accès thématique des archives patrimoniales (VATAP). L'un de nos objectifs était de quantifier l'écart sémantique existant dans le VATAP; nos résultats consistent en une première approche de cet écart afin d'avoir un ordre d'idée de l'ampleur du phénomène.

Le nombre de questions traitées est beaucoup moins important que nous l'aurions voulu au départ. Cependant, les critères de sélection émis au fur et à mesure de la recherche ont permis de ne pas avoir à réaliser d'échantillonnage final sur une base aléatoire : les données traitées répondent à des critères motivés. Par ce corpus, nous ne visons pas la représentativité de l'ensemble des centres et services d'archives du Québec. Cependant, il est intéressant de constater que les trois établissements appartiennent chacun à un des trois paliers gouvernementaux : fédéral, provincial et municipal. Cette diversité s'associe bien au dessein d'une recherche exploratoire.

La grille d'analyse en facettes a été vérifiée par des tests inter- et intra-codeurs. L'arbre décisionnel pour la sélection des questions d'usagers à teneur thématique et l'analyse en champs sémantiques ont fait l'objet de tests intra-codeurs. L'arbre décisionnel pour la sélection des questions d'usagers à teneur thématique a formellement été établi après la collecte dans le troisième établissement. Cependant, notre première sélection des questions dans les établissements a toujours été de prime abord plus large afin justement de nous laisser le temps d'apprivoiser les données, comme en témoigne la description de notre démarche de sélection et notamment des tris successifs (voir section 4.1.3.1).

L'arbre décisionnel pour la sélection des questions d'usagers à teneur thématique a permis de catégoriser les questions à la référence de plusieurs types : recherche d'une information précise (p. ex. la date de naissance d'une personne), d'un document précis (p. ex. un acte juridique, les pièces jointes à un rapport du coroner), d'une information inconnue (recherche thématique) par des noms propres ou des noms communs. Nous avons ajouté des critères pour nous assurer d'avoir des questions de même acabit : la langue (le français non traduit), l'auteur du message (nous avons refusé les courriels d'une personne de l'établissement retranscrivant la demande d'un usager qui avait téléphoné), la portée archivistique (et non des informations sur les heures d'ouverture ou toute autre information sur l'établissement ou l'obtention d'un service). Adaptable, l'arbre décisionnel pourrait servir à d'autres cheminements logiques, à d'autres sélections de questions de référence pour d'autres usages. En effet, l'application d'un code de catégories tel que celui de Mas (2013-2014) ne convenait pas pleinement à notre recherche qui tentait de ne sélectionner qu'une seule sorte de questions. Nous n'avions pas besoin de distinguer les autres catégories que celles de la teneur thématique.

Comme nous l'avons mentionné (voir section 4.2.5.2), l'analyse en facettes, bien qu'elle ait permis une meilleure compréhension du sens des thémanymes n'a pas directement servi à la filiation sémantique. Il a plutôt fallu procéder par analyse de champs sémantiques, méthode qui, elle, peut être appliquée directement. Ainsi, elle a été appliquée seule au corpus 2. Le dédoublement de la méthodologie rend compte de l'exploration méthodologique qui a été exercée tout au long de cette recherche.

Dans l'analyse linguistique, le nommage des relations sémantiques a été opéré par la chercheuse. Cette opération, s'apparentant à du codage, aurait pu être consolidée par un test intercodeur. Nous avons manqué de temps pour le réaliser. Loiseau (2012) met en garde contre la recherche à tout prix de la fréquence des mots ou des relations entre des mots; la fréquence est à pondérer en fonction du contexte linguistique dans lequel les mots ou expressions linguistiques évoluent. Nous ajoutons au contexte linguistique, le contexte archivistique tel qu'entendu par Yakel et Torres (2003) dans le modèle d'intelligence archivistique (voir section 3.2.1) peut jouer un rôle.

Toutefois, nous pensons – et nous reprenons la formulation de Lamoureux (2000, 238) – que nous pouvons avoir une confiance raisonnable dans la crédibilité (validité interne) de notre recherche et ainsi, dans les résultats présentés à chacune des questions de recherche associées à un objectif de recherche.

4.4.3. Nature sensible des données collectées

Notre corpus inclut le contenu de courriels envoyés par des usagers à des centres ou services d'archives. Même si nous ne nous intéressions qu'au contenu des courriels, ce fut à nous dans la plupart des cas de lire tous les courriels échangés, sélectionner dans l'ensemble de ces courriels envoyés aux services de référence des établissements ceux correspondant à notre recherche et d'en copier le contenu dans un tableau, par la suite anonymisé et vérifié par l'archiviste responsable de l'établissement. Ainsi, nous avons dû consulter les courriels entiers d'usagers, ayant accès à des renseignements personnels d'individus sans leur consentement, mais à des fins de recherche.

Notre collecte s'est effectuée sur un territoire couvert par une double législation émanant de la fédération du Canada et de la province du Québec.

Les lois concernées sont pour le Canada la *Loi sur la protection des renseignements personnels* (L.R.C. ch. P-21) et pour le Québec *la Loi sur la protection des renseignements personnels dans le secteur privé* (L.R.Q. ch. P-39.1) et la *Loi sur l'accès aux documents des organismes publics et sur la protection des renseignements personnels* (L.R.Q ch. A-2.1). L'article 8 section j de la LRC P-21, l'article 18, section 8 de la LRQ P-39.1 et les articles 59, section 5 et 125 de la LRQ A-2.1 portent sur la possibilité de communiquer des renseignements personnels d'individus sans leur consentement pour fins d'études et de recherche ou de statistiques, sous certaines conditions qui sont décrites dans ces textes (voir Annexe 5 – Extraits de textes de loi).

Selon ces lois, les centres et services d'archives, qu'ils soient publics ou privés, n'ont pas le droit de communiquer les renseignements personnels que sont les adresses courriel des usagers ainsi que le contenu de certains courriels comportant des informations permettant de les identifier. Ils ne peuvent les communiquer qu'à certaines conditions et doivent s'assurer de minimiser les risques de bris de la confidentialité. Afin de se conformer à cette loi, les

archivistes nous ont demandé quels étaient les moyens selon lesquels nous pouvions assurer la protection ou la non-divulgation des informations personnelles auxquelles nous aurions éventuellement accès. Nous nous sommes livrée à des enquêtes de sécurité ou autres vérifications du sérieux de notre démarche et de notre identité exigées par les centres et services d'archives pour avoir accès aux courriels d'usagers, même en simple lecture. Nous avons par ailleurs demandé un certificat d'éthique au Bureau Recherche Développement Valorisation (BRDV) de l'Université de Montréal (voir section suivante 4.4.4). Dans la mesure du possible, et comme notre recherche ne porte pas sur les informations personnelles des courriels en tant que telles (principalement le nom et l'adresse courriel de l'usager, éventuellement ses coordonnées), nous avons au fur et à mesure de la collecte supprimé ces informations selon un code d'anonymisation que nous avons élaboré (voir Tableau XII). Seul le contenu thématique des échanges de courriel entre l'usager et l'archiviste a été conservé.

Le nom des établissements contactés a été conservé et diffusé dans la présentation générale des données, après acceptation de leur part. Ceci nous a permis de rendre compte du vocabulaire pour l'accès thématique aux archives patrimoniales (VATAP) dans chacun des établissements. Ce portrait est à la mesure des données collectées, restreint pour les établissements qui avaient peu de données disponibles correspondant à notre recherche, et plus étoffé pour les établissements qui en avaient davantage.

4.4.4. Certificat d'éthique

L'Université de Montréal enjoint les chercheurs à déposer une demande de certificat d'éthique avant de débuter la recherche proprement dite. Notre démarche s'est assurée de minimiser le bris de la confidentialité et de la protection de la vie privée auxquels sont soumis les services d'archives. Les données recueillies ont été anonymisées au plus tôt, soit dans les centres et services d'archives soit ultérieurement, mais après une entente de sécurité avec l'établissement. Nous avons obtenu le certificat d'éthique numéro CERAS-2016-17-169-D.

La qualité de la recherche a été principalement assurée à partir de la multiplicité des sources de données, de tests inter- et intra-codeurs, d'une description détaillée des milieux de collecte et des données ainsi que de l'explicitation claire de notre démarche de recherche.

Conclusion du chapitre

Notre recherche est exploratoire par la taille de ses corpus et descriptive par l'analyse linguistique menée en profondeur sur les phénomènes linguistiques identifiés. Notre méthodologie consiste en une étude linguistique des termes de recherche issus des questions dans les courriels d'usagers à la référence, des termes de description contenus dans les notices descriptives et des termes d'indexation éventuellement associés aux notices. Le corpus de ces termes émane de trois centres et services d'archives de grande envergure francophones ou bilingues. Les données ont été anonymisées, puis épurées des éléments non pertinents à notre recherche. La segmentation et la caractérisation sémantique ont mis au jour les expressions linguistiques porteuses de sujet, les thémanymes. Les thémanymes issus du vocabulaire des usagers et de celui employé dans les instruments de recherche ont été comparés par paire. Nous avons nommé la relation sémantique à partir de l'échelle d'écart sémantique que nous avons développée afin de qualifier l'écart sémantique entre les deux vocabulaires. Les relations sémantiques sont comptées afin de quantifier l'écart sémantique entre les deux vocabulaires. Finalement, nous avons abordé la qualité de la recherche et en avons identifié les faiblesses pour en nuancer la discussion.

5. Résultats

Notre recherche porte sur le vocabulaire employé pour l'accès thématique aux documents d'archives patrimoniaux (VATAP). Elle a pour but d'étudier l'écart sémantique présupposé entre le vocabulaire des usagers qui recherchent des documents d'archives patrimoniaux et celui employé par les archivistes dans les instruments de recherche pour l'accès thématique à des documents d'archives patrimoniaux (VATAP). Nous avons vérifié empiriquement l'existence de l'écart sémantique présupposé entre ces deux vocabulaires, l'avons qualifié selon une approche linguistique et l'avons quantifié.

Ce chapitre décrit les résultats de l'analyse portée sur les données avec la méthodologie présentée au chapitre précédent. Il comprend d'une part l'analyse descriptive des données (section 5.1) et d'autre part les résultats (sections 5.2, 5.3, 5.4 et 5.5). Ces éléments permettent de répondre aux trois objectifs de la recherche (voir section 2.2) : valider l'écart sémantique entre le vocabulaire des usagers et celui des instruments de recherche, le qualifier linguistiquement et le quantifier.

Le VATAP est constitué d'expressions linguistiques porteuses de sujet, les thémanymes. Rappelons d'abord que les liens sémantiques entre les thémanymes du vocabulaire des usagers qui recherchent des documents d'archives patrimoniaux par sujet et les thémanymes du vocabulaire employé par les archivistes dans les instruments de recherche sont implicites mais existent réellement. Ils ont été rendus manifestes par la réponse de l'archiviste – par la correspondance qu'il a établie entre la question et les notices retenues. L'analyse a consisté à documenter le plus explicitement possible ces liens, grâce à des outils et techniques linguistiques. Ainsi, nous avons tenté de rendre saillants les liens sémantiques déjà établis. Tout locuteur de la langue française serait en mesure de rétablir ces liens, du moins, du point de vue linguistique. Cette analyse de nature linguistique ne peut néanmoins se faire hors contexte et d'un point de vue purement théorique (linguistique); elle nécessite une

compréhension de la référence assurée par un service ou centre d'archives[62], du traitement opéré sur des fonds d'archives, ainsi que du contexte précis de la question.

5.1. Portrait synthèse des données

Dans cette section, nous rendons compte des données issues de l'analyse menée telle que décrite dans la section 4.3 Méthode d'analyse. Bien que nos données soient linguistiques, le prétraitement et en particulier l'application de notre échelle d'écart sémantique leur ont conféré une nature quantitative. Puisque les données de la recherche sont des valeurs numériques, nous en traçons d'abord un portrait synthèse (Lamoureux 2000, 207), c'est-à-dire une description qualitative. Le portrait synthèse concerne le résultat de l'application de la grille ainsi que les cas particuliers qui se sont présentés à nous. Puis, nous donnons les chiffres - essentiellement le mode, la moyenne et la médiane - relatifs à la présence des relations sémantiques dans les données et la fréquence de leur occurrence.

5.1.1. Résultat de l'application de l'échelle d'écart sémantique

Dans cette section, nous traitons de la manière dont nous avons appliqué notre échelle d'écart sémantique, en soulignant en particulier les cas complexes ou plus difficiles.

Nous avons appliqué aux données l'échelle d'écart sémantique que nous avons développée pour cette recherche (voir section 4.3.1). Cette échelle attribue une valeur à chaque variation, la valeur étant de plus en plus élevée au fur et à mesure que la variation est importante. Cependant, tel que mentionné ci-dessus, les échelons de l'échelle d'écart sémantique ne sont pas équidistants. La variation entre deux thémanymes peut être de nature formelle, avec très peu d'incidence sur le plan sémantique ou bien la variation peut affecter à

[62] Compréhension telle qu'entendue selon le modèle de la référence archivistique « ARK » (Duff, Yakel et Tibbo 2013, voir section 3.2.1)

la fois grandement la forme et le sens entre deux thémanymes. Nous présentons ci-après des exemples[63] illustrant chaque relation de notre échelle.

Le lien sémantique entre deux thémanymes n'était pas toujours facile ni simple à étiqueter en fonction de l'échelle d'écart sémantique. C'est d'ailleurs pour cette raison que nous parlons de *liens sémantiques* pour les liens identifiés dans le corpus; nous réservons *relations sémantiques* aux liens sémantiques clairement établis par la littérature ou notre échelle. Dans le corpus, le lien sémantique était parfois complexe : plusieurs valeurs de l'échelle d'écart sémantique pouvaient être pertinentes conjointement. Par exemple[64], entre « le ministère d'état de la science et la technologie » et « le ministère d'État, Sciences et Technologie », nous remarquons une variation graphique (écart 3 : sans virgule/avec virgule), une variation de casse (écart 4 : tout en minuscule/majuscule à trois mots du thémanyme) et une variation syntaxique avec l'insertion de déterminants devant deux des noms composant le thémanyme 1 (écart 7 : 2 déterminants/absence de déterminant). Nous avons attribué à ce type de variations la valeur la plus élevée dans notre échelle d'écart sémantique, indiquant la relation sémantique la plus éloignée. Dans le corpus, nous avons fréquemment observé que plusieurs relations sémantiques apparaissaient conjointement et les exemples que nous reprenons dans ce texte l'illustrent parfois. Essayant de rendre compte des données le plus fidèlement possible, nous reproduisons dans les exemples et les tableaux la graphie exacte observée.

---

[63] Les divers exemples présentés ici sont récapitulés dans l'Annexe 6 – Quelques exemples des relations sémantiques de l'échelle.

[64] Dans notre analyse, nous séparons les précisions relatives à la description de l'écart entre les thémanymes par une barre oblique (précisions sur le thémanyme 1/précisions sur le thémanyme 2). En outre, la colonne précision contient des informations de nature linguistique si bien que N désigne la catégorie grammaticale du nom (N pour nom, NC pour nom commun, etc.) et non la notice descriptive (symbole utilisé dans QRNTix).

0. Relation d'identité

La relation d'identité, telle que comprise dans notre recherche et indiquée dans notre échelle d'écart sémantique, consiste en une similitude en tous points : forme et sens.

Tableau XXI. Illustration de la relation d'identité (écart 0)

Écart	Thémanyme 1	Thémanyme 2
0	anatomie	anatomie
0	Prix des communautés culturelles	Prix des communautés culturelles

La relation d'identité s'entretient entre deux thémanymes absolument identiques, qu'ils soient constitués d'un seul mot ou de plusieurs (voir Tableau XXI).

1. Variation orthographique

Notre corpus général ne présente que très peu de cas de variation orthographique. La variation peut la plupart du temps être attribuée à une autre valeur de l'échelle d'écart sémantique.

Tableau XXII. Illustration de la variation orthographique (écart 1)

Écart	Thémanyme 1	Thémanyme 2	Précision
1	les Chinois et Canadiens d'origine chinoise	les Chinois et Canadiens d'origine chinoisent	faute d'orthographe

Ainsi, le Tableau XXII indique une variation orthographique inattendue mais tout à fait probable.

2. Variation de longueur

La variation de longueur consiste en le repérage d'abréviations, y compris les sigles et acronymes (court) et leurs expansions (long). Comme le présente le Tableau XXIII, la forme courte peut se présenter tant dans le thémanyme 1 (Q et R) que dans le thémanyme 2 (R, N, Tix).

Tableau XXIII. Illustration de la variation de longueur (écart 2)

Écart	Thémanyme 1	Thémanyme 2	Précision
2	le 22e	22ième bataillon	court/long
2	le marché Ste-Anne	le marché Sainte-Anne	court/long
2	le CN	la Compagnie des chemins de fer nationaux du Canada	court/long
2	les Jeux Olympiques	les JO	long/court

Écart	Thémanyme 1	Thémanyme 2	Précision
2	le jardin zoologique du Parc Lafontaine (vers 1905-1955)	du zoo du parc La Fontaine	long : adjectif sans majuscule /court : N[65]; avec majuscule

Certaines abréviations sont connues et passées dans la langue courante, comme le sigle JO. Pour d'autres, la connaissance encyclopédique propre au territoire où est employée l'expression est nécessaire, comme le sigle CN pour *Canadian National*, la compagnie fédérale de chemin de fer. Dans ce cas-ci justement, nous pourrions nous demander si le thémanyme en regard de CN n'est pas une expansion du sigle mais plutôt une définition ou une paraphrase (écart 9). L'abréviation écrite *Ste* pour *sainte* est très courante et n'entraîne pas de changement de sens. Lors de notre analyse, certaines abréviations nécessitaient une certaine connaissance du contexte historique relatif au sujet pour pouvoir être interprétées Ici, la comparaison des thémanymes a permis de lever certaines ambigüités. Ainsi, le 22^{e}, dans le contexte de la question de l'usager, faisait référence au 22^{e} bataillon (question sur la participation financière du capitaine A.V. Roy à la constitution de ce bataillon).

3. Variation graphique

La variation graphique inclut l'ajout ou la suppression de signes typographiques tels que les traits d'union, les virgules et plus particulièrement l'ajout ou la suppression de guillemets. Nous avons trouvé très peu de variation graphique liée à l'ajout ou la suppression de guillemets. Cet élément avait été inclus dans la grille suite à l'apparition de cas dans la recherche parallèle à celle-ci, menée par Guitard et Da Sylva (2014; recherche en cours, voir section 4.4) sur les index de fin de livre.

Tableau XXIV. Illustration de la variation orthographique (écart 3)

Écart	Thémanyme 1	Thémanyme 2	Précision
3	la ville de Tracadie au Nouveau-Brunswick	Ville de Tracadie, Nouveau Brunswick	préposition/virgule

[65] Dans la colonne précision, ce sont des informations de nature grammaticale qui sont données. Ainsi, N signifie *nom* et non *notice descriptive.* Les abréviations grammaticales courantes telles que N pour nom sont indiquées dans la table des abréviations (p. 8).

Écart	Thémanyme 1	Thémanyme 2	Précision
3	la Grosse-Île	au Grosse Isle	trait d'union (3), orthographe diachronique (1)

Dans la seconde ligne du Tableau XXIV, nous avons repéré une éventuelle variation de genre tout en conservant le même sens. Cette variation n'avait pas été prévue dans notre grille. Dans une étude diachronique[66] de la langue, il s'agit effectivement d'une variation qu'il faudrait inclure, éventuellement entre 0 et 1 puisque la forme est la même, le sens aussi, mais seul le genre diffère; cette variation n'est repérable que dans un seul segment de texte dans notre corpus. Ici, cependant, si l'article diffère (la/le), l'adjectif reste au féminin (grosse), probablement par influence de l'oral. Cet exemple identifie également une variation orthographique (Île/Isle), liée elle aussi au passage du temps.

4. Variation de casse

La variation de casse concerne la présence ou l'absence de majuscules/minuscules (maj/min) à certains mots du thémanyme.

Tableau XXV. Illustration de la variation de casse (écart 4)

Écart	Thémanyme 1	Thémanyme 2	Précision
4	la Brasserie Molson	la brasserie Molson	maj/min
4	Pêcheries	pêcheries	maj/min
4	santé	Santé	min/maj
4	Cercles de Fermières	des cercles de fermières	maj maj/min min
4	cet inspecteur d'anatomie	l'Inspecteur d'anatomie	min min/maj min
4	la ville de Tracadie, Nouveau-Brunswick	Ville de Tracadie, Nouveau Brunswick	min maj/maj maj

Le Tableau XXV indique que la variation de casse peut affecter un ou plusieurs mots des thémanymes comparés.

Nous pensions initialement que la variation de casse pourrait jouer un rôle dans la distinction de certains sens, par exemple entre *la ville* – au sens d'une agglomération urbaine, par opposition à la campagne par exemple – et *la Ville* – au sens de la personne morale.

[66] Diachronie : « Évolution des faits linguistiques dans le temps (opposé à synchronie) » (*Grand Robert* 2017).

Pourtant, nous nous sommes rendu compte que la variation n'était pas significative dans notre corpus. En effet, les usagers emploient la majuscule ou non au sein du même courriel pour référer à la même entité. Les termes d'indexation débutent tous par une majuscule. Si le mot apparaît dans le titre d'une unité archivistique, il prend également une majuscule, même si dans le corps de la notice, le mot s'écrit avec une minuscule. Ainsi, cette variation n'a pas revêtu l'importance qu'on avait anticipée. La faible représentation d'une valeur de l'échelle d'écart sémantique dans le corpus général est une des raisons qui nous mènera à regrouper certaines relations sémantiques pour avoir une meilleure vue d'ensemble des résultats (voir section 5.4).

5. Variation flexionnelle

La flexion est la modification d'un mot par une marque grammaticale porteuse d'information grammaticale, typiquement le nombre (singulier/pluriel, soit sg/pl) et le genre (masculin/féminin).

Tableau XXVI. Illustration de la variation flexionnelle (écart 5)

Écart	Thémanyme 1	Thémanyme 2	Précision
5	fourrure	Fourrures	sg/pl + min/maj (4)
5	la date et le lieu de l'assermentation des ministres et des premiers ministres depuis la confédération	des dates et des lieux d'assermentation des ministres et du premier ministre depuis la Confédération	sg/pl (de la tête et d'un des compléments) + min/maj (4)

La variation flexionnelle, comme la variation de casse, peut toucher un ou plusieurs mots des thémanymes comparés (voir Tableau XXVI). Dans les deux exemples présentés ici, la variation flexionnelle est couplée à une variation de casse, c'est fortuit. Le premier cas est cependant typique d'un mot présent dans un texte (un courriel) pour le premier thémanyme et d'un terme d'indexation, isolé, débutant par une majuscule, pour le second thémanyme.

6. Variation dérivationnelle

La dérivation concerne la création de mots à partir d'autres mots de la même famille, avec une racine commune. Ainsi, le nom (*danse*) et le verbe (*danser*) sont reliés formellement et étymologiquement. Le plus souvent, la dérivation s'accompagne d'un changement de catégorie grammaticale (principalement nom, verbe, adjectif, adverbe). Lors du passage d'une catégorie grammaticale à une autre, un affixe (préfixe ou suffixe) peut se greffer à la racine, en

plus des éléments grammaticaux propres à la catégorie (p. ex. une terminaison verbale de l'infinitif immigr*er*, une terminaison d'un nom immigr*ation*). La dérivation s'applique sur le nom principal d'un thémanyme – appelé en linguistique la tête ou le noyau – ou sur le complément de ce nom principal. On trouve dans le Tableau XXVII des thémanymes ayant un sens proche et dont la variation formelle repose sur la dérivation de la tête d'un des deux thémanymes.

Tableau XXVII. Illustration de la variation dérivationnelle (écart 6)

Écart	Thémanyme 1	Thémanyme 2	Précision
6	l'Immigration	immigrants	mots de la même famille
6	Radio	Émissions radiophoniques	N/adjectif
6	Ce navire est parti de La Rochelle le 13 mai	départ	mots de la même famille : partir/départ
6	ont pu voter	des votes	verbe/N déverbal[67]
6	les troupes canadiennes	les Canadiens	N adjectif/adjectif nominalisé

Partir et *départ* (avec le préfixe *dé-*), *voter* et *vote* sont des cas typiques de la variation dérivationnelle. La nominalisation d'un adjectif dans le contexte de la guerre est assez courante également. Le mot *radio* est polysémique, mais il était employé ici bel et bien dans le sens d'émissions radiophoniques :

– Contexte du thémanyme 1 : Titre de la notice présent dans la réponse de l'archiviste : *Radio*. Nous ajoutons une partie de la portée et contenu de la notice descriptive pour rendre compte du sens du mot : « Trois invitées de l'émission "Votre opinion Mesdames" sous la direction de madame Pierre Casgrain (Thérèse Casgrain), réunies autour d'une table dans les studios de la station C.B.C. (Radio-Canada) à Montréal. » (d121).

– Contexte du thémanyme 2 : Terme d'indexation *Émissions radiophoniques*, rattaché à la notice descriptive en question.

[67] Un nom déverbal est un nom construit à partir d'un verbe ; un adjectif nominalisé est un adjectif employé comme un nom.

Si le mot radio dans son contexte avait mobilisé une autre de ses acceptions (s'il s'était agi par exemple de *radio* en tant qu'appareil), cette paire thémanymique aurait été codée différemment, en relation associative domaine/produit (écart 12) par exemple.

L'exemple de la première ligne du Tableau XXVII pourrait être considérée comme une relation associative (12) : action / agent. Mais la parenté de famille a préséance. En effet, nous nous plaçons toujours dans le contexte de la référence archivistique et donc de la recherche d'information au cours de laquelle on peut recourir à la troncature. Quand la racine étymologique est similaire d'un mot à l'autre, la troncature permet d'ouvrir la recherche à l'ensemble des mots d'une même famille. En outre, la description vulgarisée des relations associatives dans l'échelle d'écart sémantique indique une forme différente alors que celle de la dérivation indique un lien de forme.

7. Variation syntaxique

La variation syntaxique rend compte d'une différence de forme au niveau syntaxique (soit l'ordre et la combinaison des mots) : ajout, suppression ou modification d'un déterminant, ajout ou suppression adjectivale, adverbiale. À la deuxième ligne du Tableau XXVIII, l'exemple est simplement une différence entre un texte (dans un courriel) et un titre (d'une notice descriptive).

À la première ligne du même tableau, l'absence de déterminant (ou *article zéro,* Benetti 2008) après la préposition DE au pluriel est un fait linguistique largement étudié et connu. Il s'agit de l'indétermination du complément de nom. Nous pensions que ce fait, corroboré par la présence ou l'absence de majuscules, serait à mettre en relation avec la dénomination officielle de l'association. Le site web porte sur les « Cercles de Fermières » (Cercles de Fermières 2017) : l'expression ne contient pas de déterminant pour le complément et deux majuscules, tel que noté dans le thémanyme 1.

Tableau XXVIII. Illustration de la variation syntaxique (écart 7)

Écart	Thémanyme 1	Thémanyme 2	Précision
7	Cercles de Fermières	Cercles des fermières	absence de déterminant au pl/présence de déterminant au pl et maj maj/maj min 4
7	le Fonds du Service des parcs, jardins et espaces verts	Fonds Service des parcs, jardins et espaces verts	N déterminant N/N N

Écart	Thémanyme 1	Thémanyme 2	Précision
7	Montréal	la Ville de Montréal	redondance court/long

À l'échelon d'écart 7, nous incluons aussi la redondance, phénomène de tautologie au sein même d'un thémanyme. L'exemple donné dans le Tableau XXVIII est le plus fréquent de notre corpus : *Montréal/Ville de Montréal*. À partir de l'étude du contexte d'emploi des expressions, nous avons déterminé qu'il ne s'agissait pas de synonymes parce que la variation formelle entre les deux expressions est trop ténue et uniquement formelle (la synonymie inclut une petite variation de sens). Il ne s'agit pas non plus d'une paraphrase puisque le thémanyme 1 est repris dans le thémanyme 2 et qu'ils ne sont pas totalement distincts. On aurait pu considérer le thémanyme 1 comme une ellipse, soit une forme raccourcie et incomplète *Ville de Montréal* (voir section suivante), du thémanyme 2, mais cela ne semble pas le cas. La lecture du thémanyme 1 et son interprétation ne donnent pas le sentiment qu'il s'agisse d'une forme raccourcie et incomplète. Nous nous attendions à trouver la dénomination officielle dans les instruments de recherche élaborés par le service d'archives de cette entité même et cela a été le cas.

8. Ellipse d'un élément

La valeur 8 de l'échelle d'écart sémantique est attribuée à des phénomènes d'ellipse tels que *l'exposition* pour *l'Exposition Universelle et Internationale de Paris de 1900*. L'ellipse est plutôt courante en discours ; lorsque le contexte est clair pour tous les interlocuteurs d'une conversation, il n'est pas nécessaire de reprendre l'expression au complet. La reprise s'associe bien souvent par ellipse des compléments de la tête. Remarquons que le Tableau XXIX présente des thémanymes qui ne sont pas de nature nominale, nous y reviendrons en discussion.

Tableau XXIX. Illustration de l'ellipse d'un élément (écart 8)

Écart	Thémanyme 1	Thémanyme 2	Précision
8	Adolphe V. Roy a été envoyé étudier les mines du Yukon (autre version: de l'Alaska) en 1894 et 1895	qu'il fut envoyé là-bas par le gouvernement du Québec	Long/ court (reprise par un pronom) avec ajout d'un complément d'agent
8	l'Exposition Universelle et Internationale de Paris de 1900	l'exposition	Long (maj maj maj)/court (min)
8	des zoos de Montréal	le Jardin zoologique	abréviation (2) au pl (5) avec complément/ long sans complément

À la première ligne du Tableau XXIX, ce thémanyme a été retenu en tant que tel à cause de la présence du thémanyme auquel il est comparé, celui-ci étant un groupe nominal. La reprise s'effectue par le recours à des pronoms : *il, là-bas*; il s'agit de l'anaphore dont nous ne parlerons pas ici[68]. Ce même cas pourrait être envisagé comme une variation syntaxique (écart 7) mais la similarité de construction est trop importante pour cela.

9. Paraphrase

La paraphrase consiste en la reformulation généralement synthétique d'une expression généralement longue. Nous avons trouvé au début de notre exploration des index d'archives un terme d'indexation qui renvoyait à une longue expression dans une notice descriptive qui n'était rien d'autre que sa définition (voir Guitard 2014[69]). Partant de ce constat, nous avons inclus cette possibilité dans l'échelle d'écart sémantique. Cependant, dans le corpus de cette recherche, le phénomène est peu fréquent. Comme pour la valeur 8 de l'échelle, on observe ici le phénomène de la reprise. Mais cette fois-ci, il ne s'effectue pas par le recours à un pronom mais par le recours à d'autres mots qui recouvrent le sens de l'expression employée précédemment (voir première ligne du Tableau XXX).

Tableau XXX. Illustration de la paraphrase (écart 9)

Écart	Thémanyme 1	Thémanyme 2	Précision
9	Adolphe V. Roy a été envoyé étudier les mines du Yukon (autre version: de l'Alaska) en 1894 et 1895	ce mandat	explicitation/dénomination qui recouvre la phrase
9	une transaction ayant eu cours au milieu des années 40 (1946 je pense) entre le Canada et l'Argentine	l'exportation de castors en Patagonie en 1946	nom général avec noms de pays et date approximative ou incertaine/nom précis avec complément, nom de région et date précise

[68] Anaphore : Ling. Reprise d'un segment de discours antécédent par un mot qui y renvoie (ex. : de l'argent, j'*en* ai) (*Grand Robert* 2017). Pour s'informer sur l'anaphore, consulter notamment Kleiber (1994).

[69] Dans l'index figurait le terme d'indexation *Zouave* avec un localisateur qui renvoyait à une notice descriptive dans laquelle on ne pouvait lire qu'un seul segment qui pouvait correspondre sémantiquement au terme d'indexation. Le segment était le suivant : « La garde paroissiale s'inspire des corps de gardes du Vatican formés au XVe siècle pour servir et protéger le pape. (…) » (Archives nationales du Québec 1991).

Par contre, nous avons pu appliquer cette valeur à la correspondance entre deux thémanymes qui avaient le même sens : nous aurions pu catégoriser l'exemple présenté à la seconde ligne du Tableau XXX en tant que phénomène de hiérarchie de sens (ici, hyperonyme/hyponyme, *transaction/exportation*). Dans le contexte de notre corpus, l'exportation dont il est question est une transaction. Mais il ne s'agit pas à proprement parler d'une relation d'hyperonymie/hyponymie entre les deux thémanymes dans le sens où, hors contexte, une personne pourrait difficilement établir cette relation entre ces deux mots. Ce qui nous a poussée à refuser cette catégorisation est le fait que les éléments contextuels sont répétés et synthétisés entre le thémanyme 1 et le thémanyme 2.

10. Équivalence

L'équivalence inclut la synonymie entre deux thémanymes (deux mots aux sens similaires), leur antonymie (deux mots aux sens contraires) et l'équivalence entre deux noms principaux de thémanymes (deux expressions dont la tête, le noyau est identique ou synonyme). En effet, à la manière des thésaurus (Clarke 2001, Hudon 2009, voir section 3.5.3), nous incluons les antonymes dans cette catégorie. Ceci s'explique par la différence entre un contexte linguistique et un contexte de recherche d'information. Du point de vue de la recherche d'information, deux antonymes peuvent être inclus dans des ressources traitant du même sujet et ainsi apparaître simultanément dans des résultats de recherche. En outre, à la suite de Kleiber (2009), nous ne restreignons pas la synonymie à une définition d'identité sémantique totale, y compris par les connotations associées au mot, ni d'interchangeabilité dans tous les contextes. La synonymie est entendue ici de manière large, au sens d'équivalence sémantique globale, dans les contextes présentés par chacune de nos questions, dans le contexte de la recherche d'information. Il s'agit de « synonymie documentaire » (Hudon 2013, 110). Si la relation d'équivalence dans notre recherche ressemble en partie à la relation d'équivalence telle qu'entendue dans les langages documentaires, elle s'en distingue aussi. En effet, il est à noter qu'elle ne reprend pas les éléments relevant de phénomènes linguistiques identifiés dans des écarts plus faibles, contrairement à ce qu'elle peut inclure traditionnellement. Par exemple, la variation orthographique (écart 1), habituellement prise en charge par la relation de synonymie dans les thésaurus est ici distinguée et constitue un échelon en soi de l'échelle d'écart sémantique.

Tableau XXXI. Illustration de la relation d'équivalence (écart 10)

Écart	Thémanyme 1	Thémanyme 2	Précision
10	envoyer 25 couples de castors en Patagonie - Terre de feu	l'exportation de castors en Patagonie en 1946	dérivation (6) verbe/N synonymes
10	Colonies	les différentes régions ayant fait partie de la colonie de la Nouvelle-France	Synonymes
10	la formation du 22e	Constitution (du 22e)	Synonymes
10	la formation du 22e	Création (du 22e)	Synonymes
10	le gouvernement provincial	le gouvernement du Québec	provincial/du Québec

Dans le Tableau XXXI, la première ligne présente un cas de synonymie couplée à une dérivation verbe/nom : envoyer – exportation. Seule la tête du thémanyme est concernée par la synonymie-dérivation, les compléments n'influençant pas sur le sens de la tête. Le cas de la deuxième ligne du Tableau XXXI aurait pu être perçu comme de la paraphrase, mais la reprise du mot *colonie* au sein du thémanyme 2 rend caduque cette hypothèse. Nous aurions aussi pu envisager l'ellipse avec variation de nombre (écarts 8 et 5), mais ce n'est pas la même acception de *colonie* qui est mobilisée. Dans le thémanyme 1, *colonies* – au pluriel – désigne l'« ensemble des territoires colonisés » (*Grand Robert* 2017). Le thémanyme 2, quant à lui, mobilise le sens suivant : « Établissement fondé par une nation appartenant à un groupe dominant dans un pays étranger à ce groupe, (…) et qui est placé sous la dépendance et la souveraineté du pays occupant dans l'intérêt de ce dernier » (*Grand Robert* 2017)[70]. Aux lignes suivantes, les mots *formation, constitution, création* se rapportent au 22e bataillon; dans le contexte de recherche d'information, ils sont sémantiquement équivalents. Chacun de ces mots entretient une relation de synonymie l'un avec l'autre, au niveau de l'une de ses acceptions; ainsi, dans le dictionnaire de synonymes d'*Antidote*, les uns sont présents dans les entrées des autres. Nous pouvons alors conclure qu'ils sont réellement synonymes les uns des autres, dans notre corpus. Le cas suivant est différent : les expressions *gouvernement provincial / gouvernement du Québec* ne sont équivalentes que dans le contexte de cette province. L'ancrage du discours dans une situation communicationnelle inclut un lieu, un temps et des personnes. La relation de synonymie, comme l'ensemble des relations de notre

[70] Comme cette interprétation est rejetée, nous n'intégrons pas la polysémie dans notre analyse.

échelle d'écart sémantique, est analysée dans le contexte de la référence archivistique, à partir de l'ensemble des occurrences qui forment notre corpus. Ce lien sémantique observé aurait pu être placé à un autre échelon (correspondance référentielle) pour rendre compte de l'absence de correspondance sémantique au niveau linguistique pur. Cependant, la fréquence de l'expression et la perspective de la recherche d'information nous ont poussée à l'intégrer ici, en tant que relation d'équivalence de la tête.

11. Hiérarchie

La relation de hiérarchie regroupe un ensemble de relations sémantiques basées sur l'appartenance à des catégories qui en incluent d'autres, plus restreintes. Ces relations sont classiques dans un thésaurus (Hudon 2009). Nous trouvons la relation d'hyperonymie/hyponymie (fleur/rose), d'hyponymie/hyperonymie (rose/fleur). Ainsi, le Tableau XXXII indique à sa première ligne *pelleteries/fourrure. Pelleterie* se définit ainsi : « (SOUVENT AU PLUR.) Peau qu'on transforme en fourrure; la fourrure elle-même. *La traite des pelleteries.* » (Dictionnaire *Usito* 2017) La fourrure est un type de peau et toutes les peaux ne sont pas des fourrures, *pelleterie*s appartient donc à une classe plus large que *fourrure* et le sens de *peau* (ici *pelleterie*) est inclus dans celui de *fourrure.* En plus de la hiérarchie, nous voyons une variation en nombre. Le degré d'éloignement entre un hyperonyme et un hyponyme est variable. En effet, *rose* peut être l'hyponyme du plus distant *végétal* ou du plus proche *fleur*. Notre corpus présente le cas de *installations/Hôpitaux* (voir Tableau XXXII) où la relation est lâche, indirecte, la distance hiérarchique entre deux expressions est grande. Du point de vue d'un thésaurus, la relation générique/spécifique « saute » plusieurs nœuds de l'arborescence. Dans le dictionnaire *Usito*, l'incluant – c'est-à-dire le premier mot de la définition – d'*hôpital* est *établissement*, l'incluant d'*établissement* est *ensemble des installations*. Mais on sent encore la hiérarchie entre les mots, c'est pour cela qu'il ne s'agit pas d'un champ lexical.

Tableau XXXII. Illustration de la hiérarchie (écart 11)

Écart	Thémanyme 1	Thémanyme 2	Précision
11	Pelleteries	Fourrure	hyper/hypo
11	Cercles de Fermières	l'Union catholique des femmes rurales (UCFR)	co-instances de la classe des associations féminines
11	femmes	Personnes	hypo/hyper

Écart	Thémanyme 1	Thémanyme 2	Précision
11	les zoos	des parcs	hypo/hyper
11	installations (« ens. des objets, bâtiments » Petit Robert)	Hôpitaux	hyper très large/hypo
11	réparations	Construction	co-hypo de travaux
11	suffrage	Suffrage féminin	hyper/hypo
11	les Chinois et Canadiens d'origine chinoise	Groupes ethnoculturels	hypo/hyper
11	la Ville	Service des parcs	méronymie tout/partie
11	la frontière avec les USA	la frontière alaskienne	tout/partie
11	le Premier ministre	sa carrière politique	méronymie partie/tout 1er min est un des degrés d'une carrière politique
11	le regroupement Le Cercle des fermières du Québec	Cercles de fermières du Québec	instance/classe
11	le Premier ministre	la carrière de premier ministre de Brian Mulroney	Classe maj min (4)/redondance (7) min min instance

À la quatrième ligne du Tableau XXXII, nous voyons la relation hiérarchique entre *zoos* et *parcs*. Zoo est l'abréviation de jardin zoologique, hyponyme de parc, du point de vue de la gestion municipale. En effet, la Ville de Montréal a une Division des Parcs qui gérait tant le jardin zoologique, le parc Belmont, le parc la Fontaine que le jardin des Merveilles, entre autres.

La co-hyponymie c'est-à-dire la relation entre deux mots qui entretiennent une relation d'hyponymie avec le même hyperonyme (par exemple, pour l'hyperonyme *fleur*, *rose* et *tulipe* sont des co-hyponymes) s'illustre aussi dans notre corpus. Ainsi, *réparations* et *Construction* sont tous deux hyponymes de *travaux* et à ce titre ont été inclus dans la relation de hiérarchie.

Nous avons inclus dans la relation de hiérarchie la méronymie ou relation partie-tout (ou tout-partie), tout comme on l'observe dans les normes sur la relation hiérarchique dans les thésaurus (*ISO 25 964* :2011[71]). Ainsi, nous voyons à la neuvième ligne du Tableau XXXII que la méronymie s'opère entre *la Ville* et le *Service des parcs* de la Ville, l'une de ses unités constitutives.

La relation classe/instance est habituellement traitée parmi les relations hiérarchiques.

[71] Pour une invitation à distinguer soigneusement les deux types de relations hiérarchiques, voir notamment Hudon 2013, 58-59 et 131.

12. Relation associative

La relation associative est en réalité une classe de relations de plusieurs natures, d'ailleurs l'expression est le plus souvent employée au pluriel. À partir de la littérature (voir section 3.5.3), nous avons établi une liste de relations associatives étiquetées en tant que telles dans notre échelle d'écart sémantique (voir section 4.3.1). Nous les rappelons ici :

- Cause/effet
- Action/agent
- Action/produit
- Action/objet
- Action/lieu
- Agent/lieu
- Science ou discipline/objet de la science ou discipline
- Objet/propriété
- Objet/application
- Objet/matériau

La liste des relations associatives présentes dans notre corpus est plus fine, la voici (précision de la relation associative et le nombre d'occurrences dans notre corpus, de la plus à la moins fréquente) :

activité/lieu	11	événement/lieu	2
objet/activité	10	fonction/lieu	2
action/agent	8	lieu/administrateur du lieu	2
activité/objet ou produit	8	lieu/état	2
lieu/acteur	8	objet/acteur	2
lieu/activité	8	public/événement destiné à un public	2
objet/action	8		
lieu/fonction	7	acteur/action	1
entité/fonction	6	art/performance	1
agent/lieu	5	discipline ou domaine/événement	1
objet de la science/science	5		
discipline/objet de la discipline	4	discipline/activité	1
entité/activités de l'entité	4	entité/activités dans lieu	1
matière/objet	4	événement/activité	1
activité/acteur	3	fonction/attribut	1
agent/action	3	lieu/événement	1
discipline/acteur	3	objet/lieu	1
lieu/entité	3	science/objet de la science	1
action/lieu	2		
action/objet	2		

Contrairement à l'analyse en facettes, nous distinguons ici *action* d'*activité*, une activité étant grosso modo une action régulière, répétée. Nous illustrons quelques-unes de ces 36 relations (orientées) dans le Tableau XXXIII.

Tableau XXXIII. Illustration de la relation associative (écart 12)

Écart	Thémanyme 1	Thémanyme 2	Précision
12	Immigration	réfugiés	action/acteur
12	ministère	sous-ministres	lieu/acteur
12	navires	la Marine	objet/discipline
12	fourrure	Chasse	objet/activité
12	l'équipe de William Ogilvie ("Dominion Surveyor')	les arpenteurs canadiens envoyés au Yukon pour la question des frontières	instance/classe des arpenteurs

Cet échelon de la grille d'écart sémantique, nous le voyons dans le Tableau XXXIII, couvre de nombreux phénomènes. Pourtant, ce n'est pas la relation la plus fréquente dans notre corpus.

13. Relation qui relève d'un champ sémantique

L'inclusion dans un champ sémantique révèle un lien sémantique thématique entre des thémanymes sans que ce lien soit explicitement dénommé. Contrairement aux relations précédentes qui sont étiquetées, celle-ci est plus floue. L'ensemble des expressions linguistiques et leurs liens forment un champ sémantique (voir section 3.5.3). Cet échelon regroupe les liens que la chercheuse était en mesure d'étiqueter par des éléments en précision. Ces liens étaient absents de notre liste préalable. Ce sont des liens sémantiques plus lâches, plus ténus, mais que l'archiviste et parfois l'usager ont établis entre des thémanymes de la question et de la notice descriptive ou des termes d'indexation et que la chercheuse a été capable de retracer dans des ressources lexicographiques – notamment dans l'onglet « champ lexical » d'*Antidote* – ou par introspection linguistique.

Tableau XXXIV. Illustration du champ sémantique (écart 13)

Écart	Thémanyme 1	Thémanyme 2	Précision
13	le gouvernement du Canada	politique générale	entité/domaine (politique)
13	colonies	découvertes géographiques	lieu/action (histoire et géographie)
13	ont pu voter	le suffrage	action/droit (droits civiques)

Écart	Thémanyme 1	Thémanyme 2	Précision
13	Pêcheries	la teinte [de fourrure]	lieu et activité dans un même domaine (chasse et pêche)
13	retour en politique en 1989	maire	événement (action-temps)/fonction dans un même domaine (politique)
13	le gouvernement fédéral ou provincial	des communications officielles entre le gouvernement fédéral et les gouvernements provinciaux	entité vague/action d'entités précises (politique)
13	les irlandais	la Société Saint-Patrice de Montréal	peuple/association ethnique (culture)
13	Les Canadiens de Montréal	les jeux	équipe de sport / hockey sport et jeu (culture et sports)
13	événements sportifs plus "officiels"	les jeux	évt/activité (culture et sports)
13	événements sportifs plus "officiels"	les activités	évt/activité (culture et sports)
13	des abords canal Lachine	l'ensemble des activités industrielles, ferroviaires et portuaires, les bâtiments existants et les campagnes environnantes	lieu : abords du canal/ activité et lieu : campagnes environnantes (géographie)
13	le typhus	l'hôpital de la marine et des émigrants	maladie/lieu pour soigner (santé)
13	des Irlandais	l'hôpital de la marine et des émigrants	instance peuple qui a émigré au Canada/lieu de soin pour émigrants (santé)

Nous avons essayé dans la mesure du possible d'identifier le lien sémantique flou et ténu entre des thémanymes, dans nos mots, pas toujours de manière synthétique. Nous avons formulé dans la colonne « Précision », à la pièce, l'identification de ce lien sémantique verbalisé (Kister, Jacquey et Gaiffe 2011). Nous avons tenté, dans la mesure du possible également, d'identifier la notion centrale du champ sémantique et l'avons indiquée entre parenthèses dans la colonne *Précision* du Tableau XXXIV. Même si, à la suite des archivistes et des usagers, nous sentions qu'il y avait un lien sémantique entre deux thémanymes, il n'était pas toujours évident de poser des mots sur une compréhension qui paraissait évidente ou intuitive plus qu'explicitement sémantique.

Le champ sémantique est l'échelon sémantiquement le plus éloigné de notre échelle d'écart sémantique, telle qu'elle a été conçue et testée dans une recherche parallèle (voir section 4.4). Pourtant, nous avons ressenti le besoin de créer un quatorzième échelon à notre échelle d'écart sémantique, un échelon qui regroupe des liens, qui ne sont pas réellement sémantiques et qui donc sortent du cadre de notre analyse et de l'objet de notre thèse.

14. Une classe supplémentaire : correspondance référentielle

La correspondance référentielle est un lien d'ordre référentiel, c'est-à-dire que le lien entre deux thémanymes ne s'effectue pas sur le plan du sens, mais sur celui du *monde*, comme l'on dit en linguistique, le plan de la réalité des choses. Comme nous ne nous attendions pas à trouver cette relation dans notre corpus et qu'elle dépasse légèrement le cadre de notre recherche, nous indiquons ici quelques éléments bibliographiques. « Le référent d'un énoncé linguistique est un élément du monde que cet énoncé permet de désigner dans un contexte donné de parole (c'est-à-dire d'utilisation de la langue). Le sens appartient à la langue alors que le référent n'existe que dans la parole : ce n'est que lorsque l'on considère une instance particulière d'utilisation ou de manifestation d'un énoncé qu'on peut identifier un référent donné. » (Polguère 2016, 154) « [L]e langage en tant que système de signes est tourné vers le dehors, vers ce qu'on appelle ou ce qu'on croit être la réalité, ou encore le monde, précisément parce qu'un signe n'est signe que s'il représente quelque chose d'autre que lui-même. » (Kleiber 1997, 16) « [L]es expressions linguistiques, si elles réfèrent, réfèrent à des éléments « existants », réels ou fictifs, c'est-à-dire conçus comme existant en dehors du langage : cette existence leur est garantie par cette modélisation intersubjective stable à apparence d'objectivité qui caractérise notre appréhension du monde » (Kleiber 1997, 17).

Tableau XXXV. Illustration de la correspondance référentielle (écart 14) ex.1

Écart	Thémanyme 1	Thémanyme 2	Précision
14	notariat	directeurs de compagnies	les actes notariés rendent compte de certaines fonctions/certains postes
		gouverneurs particuliers	
		secrétaire d'État	
		gouverneurs généraux	
		ministre	

Le Tableau XXXV illustre la correspondance référentielle entre la réalité du notariat et celle des titres de certaines fonctions que l'on peut trouver dans les documents d'archives : le lien sémantique entre les thémanymes n'est pas sémantique mais référentiel, encyclopédique, plus précisément, ancré dans un savoir contextuel relié à la connaissance des fonds et à la connaissance des types d'informations qui se trouvent dans certains types de documents d'archives.

Tableau XXXVI. Illustration de la correspondance référentielle (écart 14) ex. 2

Écart	Thémanyme 1	Thémanyme 2	Précision
14	qui ont marqué la province	présidente	le créateur du fonds est une personne influente
		Présidentes d'association	

Le créateur du fonds suggéré par l'archiviste de référence pour illustrer les personnes influentes de la province est une personne influente. L'acquisition de fonds d'archives privées est le produit d'une évaluation préalable sur l'importance ou le rôle du créateur du fonds, en lien avec la politique d'acquisition du service d'archives acquéreur du fonds d'archives. L'influence de certaines personnes dans la société est particulièrement difficile à établir si elle est envisagée à partir de la classe qui regroupe ces personnes (classe identifiée normalement par un nom commun, ici par une proposition relative) et non à partir de ses instances (identifiées généralement par des noms propres). Cette connaissance repose sur la culture générale de l'archiviste, en plus de sa connaissance des fonds que conserve l'établissement pour lequel il travaille. Le Tableau XXXVI illustre ce cas avec une correspondance référentielle entre un fait et des fonctions. Ce lien n'est pas inscrit purement dans le sens des thémanymes, mais bien plus dans la connaissance historique et générale de la société : les présidentes (notamment d'associations) ont pu jouer un rôle sur le plan provincial. Il s'agit d'une connaissance encyclopédique.

Tableau XXXVII. Illustration de la correspondance référentielle (écart 14) ex. 3

Écart	Thémanyme 1	Thémanyme 2	Précision
14	la rivière Saint-Pierre de Montréal	le fleuve Saint-Laurent	2 cours d'eau à Montréal
		la petite rivière (Saint-Antoine)	
14	du poste de traite à Lachine	canal Lachine	proximité dans la réalité physique
14	du vieux Montréal	depuis le fleuve Saint-Laurent (...) le secteur situé entre le pont Victoria, Pointe-Saint-Charles, l'île Sainte-Hélène et le port au sud, le secteur des docks et du canal Lachine (aux environs de la rue des Seigneurs) à l'ouest, le parc du Mont-Royal au nord et les frontières est de la ville	lien référentiel : un lieu et la description de son territoire
14	du centre-ville	l'île	ce centre-ville appartient à une ville qui est sur une île
14	l'Immigration	les membres du Conseil des communautés culturelles	activité/fonction liée à des entités reliées à cette activité
14	autres grands moments historiques concernant les femmes qui ont marqué la province	vie	correspondance référentielle
		Conférences	

Écart	Thémanyme 1	Thémanyme 2	Précision
14	la mode hippie	Exposition artisanale	correspondance référentielle
14	les activités secrètes allemandes au Canada pendant la Première Guerre mondiale	les camps d'internement au Canada pendant les deux guerres mondiales	correspondance référentielle

La première ligne du Tableau XXXVII rapproche des thémanymes qui réfèrent à des cours d'eau de Montréal, sans qu'il y ait de lien sémantique à proprement parler entre les deux. Nous pourrions éventuellement parler de deux co-instances d'une même classe, les cours d'eau à Montréal.

La deuxième ligne du Tableau XXXVII illustre bien un lien dans la réalité, les deux lieux se situent dans la même zone géographique. Il n'y a pas de lien sémantique entre un poste de traite et un canal : le lien ne se situe pas sur le plan sémantique. Il y a peut-être une étude à faire sur le lien entre les lieux nommés par le même nom propre; elle se situerait en onomastique. Nous mentionnons ici que nous avons été obligée d'inclure quelques noms propres très courants pour ne pas trop réduire le nombre de relations sémantiques observées; nous avons conservé par exemple Montréal (pour faire le lien avec *la ville* et *municipal*) ainsi que Québec (pour faire le lien avec *provincial* et *québécois*).

À la troisième ligne du Tableau XXXVII, le lien entre le thémanyme « vieux Montréal » et le long thémanyme qui lui est comparé rend compte d'un lien similaire à la paraphrase (9), mais sur le plan référentiel. Ainsi, il ne s'agit pas ici d'une définition, mais plutôt d'une description référentielle de l'espace associé à un lieu précis représenté par une dénomination[72].

Les deux dernières lignes du Tableau XXXVII illustrent le cas d'un lien qui s'appuie sur des connaissances encyclopédiques : les hippies aimaient particulièrement l'artisanat, les camps d'internement ont été établis dans le contexte de la surveillance politique et militaire au cours des deux guerres mondiales. Aussi, à la période déterminée par la mode hippie, des images d'un événement public couru tel qu'une exposition artisanale sont censées rendre compte de l'habillement des personnes de cette époque. Les dossiers sur les camps

[72] La dénomination est un terme linguistique défini dans la revue de la littérature (voir section 3.5.1.6).

d'internement au Canada pourraient inclure d'autres affaires de surveillance politique et militaire dans le contexte évoqué, dont des activités secrètes allemandes, ce qui intéresse l'usager. Nous voyons ici le raisonnement de l'archiviste pour répondre aux questions des usagers qui s'appuie à nouveau sur des connaissances des fonds et des connaissances historiques, nous y reviendrons en discussion.

5.1.2. Cas particuliers de l'application de l'échelle d'écart sémantique

Dans la perspective de rendre compte de l'étendue des phénomènes auxquels nous avons été confrontée lors de l'analyse, nous relevons ici quelques-uns des cas difficiles ou extrêmes. Pour chacun de ces cas, nous introduisons le phénomène à l'étude en présentant d'abord un tableau identifiant le cas; puis un paragraphe suit pour discuter de la difficulté rencontrée et de la position que nous avons prise par rapport à celle-ci. Lorsque pertinent, nous indiquons si la difficulté présentée est survenue à plusieurs reprises, mais ces cas restent tout de même généralement limités.

Liens complexes

Tableau XXXVIII. Cas particulier : liens complexes

Écart	Thémanyme 1	Thémanyme 2	Précision
13	du centre-ville	la ville de Hull	partie/instance d'un tout
13	des premiers trains à y circuler	chemin de fer	entité action/ entité
14	Montréal	conseiller municipal	instance de lieu qui est une municipalité /fonction municipale

Nous l'avons déjà mentionné au début de la section 5.1, les liens sémantiques dans notre corpus étaient bien souvent complexes : plusieurs échelons de notre échelle d'écart sémantique étaient identifiables pour caractériser un lien entre deux thémanymes. Dans le flot de liens, nous avons eu, en particulier, de la difficulté à identifier la relation instance/classe couplée à la méronymie, telle qu'indique la première ligne du Tableau XXXVIII. Un autre type de cas difficile à étiqueter était celui où des éléments de même niveau se présentaient; à la seconde ligne du tableau, les deux entités sont *trains* et *chemin de fer*. Nous avons dans ces cas procédé par élimination, en essayant d'appliquer d'autres relations telles que la hiérarchie ou l'une des relations associatives (ici, p. ex. objet/domaine), des plus basses aux plus élevées dans l'échelle d'écart sémantique, à partir de la consultation d'ouvrages lexicographiques. En

outre, nous sommes généralement retournée à la source, i. e. le texte dont émanait le thémanyme : Q, R, N ou Tix. Le thémanyme dans le texte complet dont il est issu fournissait un contexte phrastique plus large qui nous a permis d'identifier le sens des mots polysémiques en jeu. La dernière ligne du Tableau XXXVIII illustre un lien complexe sur le plan référentiel : une instance de lieu et une fonction possible - il y en a bien d'autres - associée à ce lieu.

Thémanymes complexes

Tableau XXXIX. Cas particulier : thémanymes complexes

Écart	Thémanyme 1	Thémanyme 2	Précision
8	Adolphe V. Roy a été envoyé étudier les mines du Yukon (autre version: de l'Alaska) en 1894 et 1895	son envoi au Yukon	événement précis avec verbe/ événement imprécis avec N (ellipse de 2 compléments + variation dérivationnelle 6 verbe/N)
11	dans les parcs	la piste du parc Kent	Classe/ instance + méronymie

Certains thémanymes sont longs et complexes. Le lien qu'ils entretiennent avec un autre thémanyme est ainsi, de manière tout à fait attendue, complexe lui aussi. Dans le Tableau XXXIX, nous voyons que le premier thémanyme est non pas un groupe nominal (dont la tête est un nom), mais plutôt une proposition (dont la tête est un verbe conjugué qui régit sujet et compléments). Ce thémanyme a été retenu en tant que tel à cause de la présence du thémanyme auquel il est comparé, celui-ci un groupe nominal. La présence de thémanymes qui n'ont pas de nom commun à leur tête s'explique par le fait que, dans le corpus 2, les textes n'ont pas été segmentés et analysés en facettes au préalable (contrairement au corpus 1). La comparaison s'est alors effectuée à la vue des textes; les éléments ayant un rapport les uns avec les autres entre deux sources (ici question de l'usager et réponse de l'archiviste) étaient délimités à ce moment-là (voir section 4.1.2.4).

Le lien sémantique peut englober plusieurs nœuds d'une représentation thésaurale classique. La seconde ligne du Tableau XXXIX rend compte de ce phénomène. Formellement, le mot *parc* ne subit qu'une variation de nombre, pourtant, la variation réelle porte sur la signification du thémanyme entier. On passe ainsi d'une classe – les parcs – à une instance – le parc Kent – puis encore à une partie du parc – la piste du parc de Kent. Nous constatons alors que l'étude des mots isolés, décontextualisés et non des thémanymes entiers – même si ceux-ci

accusent déjà une certaine perte de contexte – ne rendrait pas compte de la compréhension humaine des termes du VATAP.

Interférence syntaxique

Tableau XL. Cas particulier : interférence syntaxique

Écart	Thémanyme 1	Thémanyme 2	Précision
11	Inspecteur d'anatomie	ce poste	hypo/hyper

Les thémanymes ont été extraits de textes, parfois longs, dans lesquels s'appliquaient les règles d'écriture du français rédigé. Comme dans tout texte rédigé, nous avons constaté le recours à des pronoms (p. ex. *il*) et des déictiques[73] (p. ex. *celle-ci*) : c'est le phénomène de l'anaphore. Dans le Tableau XL, la présence d'un adjectif démonstratif (*ce* poste*)* a perturbé notre identification du lien. Nous l'avons d'abord supprimé, mais alors le mot perdait son contexte et notamment l'identification de l'homonyme par le genre (*le/la poste*), nous l'avons donc réintégré. Un locuteur qui fait une première mention d'une chose utilise, pour les mentions suivantes, des pronoms, des déictiques, des moyens linguistiques pour éviter de se répéter. Le jeu de l'antécédent avec la reprise par un hyperonyme est somme toute assez classique. Mais nous avons dû nous y reprendre à plusieurs fois pour l'identifier clairement une fois les thémanymes hors contexte. Dans ce cas-ci, nous aurions pu décider de rattacher le lien sémantique à l'échelon 12 : instance/classe. La différence entre la relation hiérarchique et la relation associative est théoriquement claire (voir section 3.5.3) : une instance est un hyperonyme du plus bas niveau qui soit dans une hiérarchie étant donné qu'il désigne une entité unique. Cette différence est moins évidente dans la pratique. Une instance est unique, or, ce n'est pas le cas des inspecteurs d'anatomie, malgré la majuscule.

[73] Déictique : Log., ling. Qui sert à montrer, à désigner un objet singulier. *« Ceci » est un mot déictique* (*Grand Robert* 2017).

Distribution de la tête

Tableau XLI. Cas particulier : distribution de la tête

Écart	Thémanyme 1	Thémanyme 2	Précision
5	le gouvernement fédéral ou provincial	le gouvernement fédéral et les gouvernements provinciaux	distribution de la tête avec sg/pl

Il existe une nuance entre les deux thémanymes présentés dans le Tableau XLI. Mais nous avons rapproché syntaxiquement ces deux thémanymes par la répétition du mot principal - appelé en linguistique la tête ou le noyau - du groupe nominal (le gouvernement) et la distribution des compléments opérée par la conjonction de coordination *ou* entre les deux adjectifs. La présence du pluriel dans le thémanyme 2 indique qu'il ne s'agit pas du palier fédéral ou provincial au choix, mais bien de l'ensemble des deux paliers de gouvernement canadien dont il est question, du Canada et des provinces (il n'est pas fait mention des territoires dans notre corpus).

Co-hyponymie incertaine

Tableau XLII. Cas particulier : co-hyponymie incertaine

Écart	Thémanyme 1	Thémanyme 2	Précision
11	communautés culturelles	francophonie	co-hyponymes ? hyper : groupe de personnes

L'identification du lien sémantique n'allait pas toujours de soi, même si les deux thémanymes étaient des termes connus, ayant une définition dans des dictionnaires. Ainsi le Tableau XLII présente un cas où nous avons attribué une relation de hiérarchie à la comparaison de deux thémanymes. Francophonie est définie comme suit : « Ensemble des populations francophones, communauté que forment les États qui emploient le français (France, Belgique, Suisse, Canada, Louisiane, Afrique, Madagascar, Liban, Antilles, etc.). » (*Antidote* 2017) *Communautés culturelles* est une expression documentée dans *Antidote* à l'article *communauté*, avec une épithète (un adjectif qui se présente directement relié au nom) : *membres des communautés culturelles, issu des communautés culturelles, représentants des communautés culturelles*. Nous pensons que ce couple entretient une relation de co-hyponymie. Mais alors il devrait nous être possible d'identifier l'hyperonyme commun, éventuellement *groupe de personnes* ou *groupe social*. Or, après la consultation de ressources lexicographiques, nous n'avons pas été en mesure de le faire avec certitude. Il

semblerait que les deux thémanymes ne soient pas sur le même paradigme, la francophonie rassemblant des communautés culturelles d'un certain type – francophones. Pourtant, on ne définirait pas la francophonie comme étant un regroupement de communautés culturelles francophones. Le lien ne semble pas être non plus une relation associative présentant deux co-instances d'une classe puisque les deux thémanymes désignent des classes et non des entités uniques par des noms propres. Nous aurions pu placer ce lien dans un champ sémantique (pouvant être appelé *culture*). Cependant, le lien entre les deux thémanymes nous semblait plus proche, plus resserré que celui d'un champ sémantique. C'est pourquoi nous avons conservé la relation de hiérarchie pour ce couple.

Redondance ou correspondance référentielle

Tableau XLIII. Cas particulier : redondance ou correspondance référentielle

Écart	Thémanyme 1	Thémanyme 2	Précision
7	Montréal	la Ville de Montréal	redondance court/long
14	Montréal	la Ville	correspondance référentielle ou 7 (redondance) avec une ellipse 8 (de la redondance justement)

La relecture de notre application de la grille à divers intervalles nous a permis de rapprocher un cas particulier pour lequel nous avons hésité entre les échelons 7 et 14. À la première ligne du Tableau XLIII, la comparaison des thémanymes permet d'identifier un lien de redondance, classée dans notre échelle en tant que variation syntaxique (7). À la seconde ligne du tableau, nous voyons le couple de thémanymes Montréal/la Ville. Nous nous sommes alors demandé s'il s'agissait d'une redondance avec une ellipse (=> 7 la Ville de Montréal, avec omission du complément 8 => la Ville) ou d'une correspondance référentielle (14). L'étude du contexte de la séquence nous a amenée à pencher pour la correspondance référentielle : nous trouvons d'une part *les parcs de la Ville de Montréal* et d'autre part *les olympiades de Montréal*, dans les notices descriptives de la même série. Il nous semble que les deux thémanymes identifiés dans le tableau sont complets et équivalents implicitement dans le contexte montréalais.

Classe et instance

Tableau XLIV. Cas particulier : classe et instance

Écart	Thémanyme 1	Thémanyme 2	Précision
12 ou 14	Libération Nord-Pas-de-Calais	Boulogne est tombée aux mains des Canadiens	classe d'événement/instance d'événement opposé
9 ou 14	les troupes canadiennes libérant des villes autour de Boulogne sur mer	Boulogne est tombée aux mains des Canadiens	antonymie ou correspondance référentielle
12 ou 14	une ville de sports	la Ville de Montréal	Classe/instance ou correspondance référentielle avec partage de la tête du syntagme (redondance, écart 7); min min / maj maj
12	Irlandais	Sir William Hales Hingston	Classe/instance

Le lien sémantique présenté à la première ligne du Tableau XLIV a posé quelque difficulté à la chercheuse. Il est à nouveau question de la relation hiérarchique classe/instance, mais cette fois-ci en lien avec des événements, qui plus est des événements vus de deux points de vue opposés (*libération/tomber aux mains*). Les relations sémantiques sont généralement présentées dans les ouvrages de lexicologie avec des noms concrets. Bien sûr, cela est nécessaire pour bien faire comprendre la réalité qui se cache derrière la relation sémantique. Cependant, notre corpus présente de nombreux cas où les thémanymes sont des noms abstraits, beaucoup plus difficiles à cerner par la pensée (Flaux et Van de Velde 2000). Encore une fois, il s'est agi alors de nous y reprendre à plusieurs fois pour arriver à placer le lien sémantique entre un couple de thémanymes à l'échelon de l'échelle d'écart sémantique que nous estimions le plus juste. La deuxième ligne rend compte de la même action désignée par les thémanymes, vue de deux points de vue (*libérer/tomber aux mains*) avec inversion des actants[74] (*troupes canadiennes et Boulogne sur mer/ Boulogne et les Canadiens*).

La troisième ligne du Tableau XLIV illustre un autre type de cas particulier, étant donné que les thémanymes partagent la tête (*ville*), avec une différence de casse (écart 4). Ce lien rend compte d'une classe de lieu et d'une instance de lieu précis; cependant il est difficile d'identifier le lien comme appartenant au plan purement sémantique et non référentiel.

[74] Les actants sont les rôles associés à une action, quelle que soit la manière dont ils agissent, subissent ou contribuent à l'action.

L'identification du lien est perturbée par le partage du mot principal du groupe nominal, *ville*, qui entre dans la redondance (écart 7) dans le thémanyme 2.

Les relations de classe/instance ne présentent pas systématiquement un lien évident *a priori*. Par exemple, ne va pas de soi le lien entre les Irlandais (thémanyme de la question et de la réponse) et Sir William Hales Hingston (titre de la notice descriptive), dont voici la description : « Médecin, chirurgien, professeur, auteur et homme politique, maire de Montréal de 1875 à 1877, né le 29 juin 1829 dans le canton de Hinchinbrook, Bas-Canada, fils du lieutenant-colonel Samuel James Hingston et d'Eleanor McGrath; le 16 septembre 1875, il épousa à Toronto Margaret Josephine Macdonald, fille de Donald Alexander Macdonald (Sandfield), lieutenant-gouverneur de l'Ontario, et ils eurent cinq enfants; décédé le 19 février 1907, à Montréal, et inhumé le 21 suivant au cimetière de Notre-Dame-des-Neiges » (notice c2d017). Il s'agit bien d'un lien de classe – les Irlandais, comme peuple ou communauté – à instance – un individu, une personne d'origine irlandaise. Pourtant, rien dans la notice descriptive ne laisse envisager l'origine de cette personne puisqu'elle est née au Québec. Les patronymes peuvent sous-entendre une origine particulière, mais il est nécessaire d'avoir des connaissances en onomastique.

Évocation

Tableau XLV. Cas particulier : évocation

Écart	Thémanyme 1	Thémanyme 2	Précision
13	néant (sème de l'époque révolue véhiculé par le mot « archives » répété + « ne date pas d'hier »)	au cours des années 1940 à 1960	(époque révolue)

Le Tableau XLV présente le cas extrême d'un lien sémantique très difficile à étiqueter. À la lecture des textes (question de l'usager et réponse de l'archiviste), il nous a semblé qu'un lien manquait dans la restitution des liens sémantiques qui unissaient les deux sources lorsque nous faisions la caractérisation sémantique. Ce n'est qu'en examinant les notices descriptives reliées qu'il nous est apparu qu'un élément implicite dans la question était présent explicitement dans la réponse et les notices. L'archiviste proposait à l'usager des notices de certaines époques sans que l'usager n'ait identifié explicitement de dates dans sa question. Par contre, l'usager demandait des « archives » – en répétant le mot – et parlait du sujet qui l'intéressait en ajoutant la mention « qui ne date pas d'hier ». Nous avons conclu que ces

expressions véhiculaient de manière implicite la notion de l'époque révolue et que l'archiviste, à défaut de date explicite, proposait à l'usager des documents d'une époque qui devait lui semblait suffisamment révolue – ici les années 1940-1960. Ce cas d'interprétation poussée est dû à la simple évocation d'un concept. La présence d'un même élément de sens (« sème », en linguistique) dans les deux sources dont sont issus les thémanymes nous a incitée à inclure leur lien au niveau de l'échelon du champ sémantique. Ce cas est unique dans notre corpus.

Nous avons présenté l'application de l'échelle d'écart sémantique ainsi que quelques cas particuliers présents dans notre corpus. Ces deux premières sections du portrait synthèse de nos données sont complétées par la prochaine section qui porte sur les caractéristiques numériques de thémanymes et de relations sémantiques dans notre corpus.

5.1.3. Caractéristiques numériques des données

Cette section rend compte du nombre et de la proportion des thémanymes et des relations sémantiques entre les thémanymes, au sein de notre corpus.

Nous rappelons d'abord brièvement quelles sont nos données. Le corpus 1 contient des données de quatre sources (Corpus 1 : QRNTix), le corpus 2 ne contient pas de termes d'indexation (Corpus 2 : QRN). Nous comparons deux vocabulaires. D'une part, nous examinons le vocabulaire des usagers représenté par les expressions linguistiques porteuses de sujet (les thémanymes) issues de la question posée à la référence des trois services d'archives qui sont les milieux de notre collecte (Q). Et, d'autre part, nous examinons le vocabulaire des archivistes représenté par les thémanymes dans les notices descriptives (N) et les termes d'indexation (Tix) pour le corpus 1 et seulement les notices pour le corpus 2. Le vocabulaire des archivistes est également représenté par les réponses des archivistes aux questions d'usagers à la référence (R) (voir Tableau XLVI). Et ce sont également des termes de recherche puisqu'ils rendent compte de l'activité de recherche et incluent bien souvent les stratégies de recherche essayées par l'archiviste, qu'elles soient fructueuses ou non (voir section 4.2.1.1).

Tableau XLVI. Types de termes et vocabulaires (corpus 1 et 2)

	Vocabulaire des usagers		Vocabulaire des archivistes	
	Corpus 1	**Corpus 2**	**Corpus 1**	**Corpus 2**
Termes de recherche	Q	Q	R	R
Termes de description			N	N
Termes d'indexation			Tix	
Termes dans les instruments de recherche			Instr (N+Tix)	Instr (N)

Dans le corpus général (corpus 1 et corpus 2), bien souvent les archivistes reprennent les mots de la question afin d'y répondre. Nous avons alors effectué deux analyses distinctes : par source (QRNTix) et par types de termes (termes de recherche *versus* de description et d'indexation). La première analyse porte sur les relations présentes entre certaines sources : Q-R, Q-N, Q-Tix, R-N et R-Tix. Puis, pour répondre à nos questions de recherche, nous avons mesuré l'écart sémantique entre le vocabulaire des usagers et celui des instruments de recherche qui lui sont habituellement mis à disposition, par exemple en ligne, sans l'intervention systématique d'un archiviste de référence. Les instruments de recherche dont nous parlons sont les notices descriptives et les termes d'indexation. Cette seconde analyse est d'un intérêt certain pour les services d'archives concernés parce qu'elle indique, par exemple, si les attentes de l'archiviste quant à l'autonomie des usagers ayant à disposition les instruments de recherche seraient réalistes ou non.

Deux types de termes ont des affinités l'un avec l'autre. Ainsi, il est naturel de regrouper les termes de description et d'indexation puisqu'ils font tous deux partie des instruments de recherche que les archivistes établissent pour permettre aux usagers (usagers externes, usagers internes, archivistes eux-mêmes) d'être autonomes dans leur recherche. Les résultats de N et Tix sous l'appellation *Termes des instruments de recherche* sont alors groupés dans certaines analyses pour évoquer les termes des instruments de recherche (voir Tableau XLVI).

5.1.3.1. Nombre de thémanymes

Le nombre d'expressions linguistiques porteuses de sujet (thémanymes) est très variable d'une séquence à l'autre. Nous appelons une « séquence » l'ensemble formé par une

question, une réponse, une notice descriptive et éventuellement des termes d'indexation rattachés à cette notice (QRNTix dans le corpus 1 ou QRN dans le corpus 2).

Le corpus 1 totalise 1410 thémanymes. Dans le corpus 1, le nombre de thémanymes varie de 7 à 392 selon la séquence (2 à 28 % du nombre de thémanymes du corpus 1, n=1410). Dans le corpus 1, la moyenne du nombre de thémanymes par séquence (QRNTix) se situe à 54 alors que la médiane se situe à 31 thémanymes. Une polyséquence est un ensemble de séquences qui ont la même question et la même réponse. Par exemple, a055q inclut une réponse, mais quatre notices descriptives avec leurs termes d'indexation chacune : QRNTix NTix NTix NTix. La moyenne du nombre de thémanymes par polyséquence est 141 alors que la médiane se situe à 69 (voir Tableau XLVII).

Tableau XLVII. Total, minimum, maximum, moyenne et médiane des thémanymes par séquence et par polyséquence (corpus 1)

Corpus 1	Nombre	Pourcentage
Nombre total thémanymes / corpus 1	1410	100%
Minimum thémanymes / séquence	7	0,5%
Maximum thémanymes (mode) / séquence	392	27,8%
Moyenne / séquence	54	-
Médiane / séquence	31	-
Minimum thémanymes / polyséquence	24	1,7%
Maximum thémanymes (mode) / polyséquence	392	27,8%
Moyenne / polyséquence	141	-
Médiane / polyséquence	69	-

La différence entre la moyenne et la médiane est particulièrement élevée. En effet, une notice descriptive de notre corpus comporte 369 thémanymes dans une séquence qui en contient 392 (voir Tableau XLVI), ce qui fausse la perception de l'ensemble. Nous avons alors calculé la moyenne et la médiane pour chacune des sources (question Q, réponse R, notice descriptive N et terme d'indexation Tix). Le Tableau XLVI présente ces résultats pour le corpus 1. Nous indiquons le nombre minimal rencontré dans les données en plus du nombre maximal (mode).

Tableau XLVIII. Mode, moyenne et médiane de thémanymes par source (corpus 1)

Calcul / source	Q	R	N	Tix
Minimum < Maximum	2 < 12	1 < 43	3 < 369	1 < 15
Moyenne	6	8	35	5
Médiane	7	4	13	3

Nous voyons dans le Tableau XLVIII que le nombre de thémanymes dans les notices est plus élevé que celui des autres sources et que la médiane donne une meilleure idée de la distribution que la moyenne puisque les nombres de thémanymes dans les sources sont hétérogènes.

Dans le corpus 2, nous n'avons pas isolé les thémanymes en tant que tels lors de notre travail d'établissement des liens sémantiques; nous avons directement attribué des champs sémantiques à des segments de texte (voir section 4.2.5.3).

En outre, rappelons que le lien sémantique a été posé dans un sens précis, celui de la séquence (QRNTix). Nous avons attribué une valeur de l'échelle d'écart sémantique aux liens sémantiques entre les thémanymes des diverses sources toujours de la même manière. Ainsi, les thémanymes 1 reprennent des thémanymes présents dans la question et la réponse alors que les thémanymes 2 reprennent des thémanymes présents dans la réponse et la notice descriptive et éventuellement les termes d'indexation. Les thémanymes 1 sont toujours des termes de recherche, alors que les thémanymes 2 peuvent appartenir tant aux termes de recherche, de description que d'indexation. Cependant, dans nos outils, les sources (Q, R, N ou Tix) sont toujours rattachées aux thémanymes dont ils sont extraits. Puisque nous analysons la filiation sémantique dans une séquence, les thémanymes repérés dans la réponse sont présents à deux reprises dans nos calculs. Ils sont considérés à la fois dans thémanymes 1 avec les thémanymes de la question et à la fois dans thémanymes 2 avec les thémanymes de la notice descriptive et des termes d'indexation. Un même thémanyme peut être mentionné plusieurs fois dans notre corpus s'il entretient une relation sémantique avec des thémanymes différents ou de plusieurs sources. Par exemple, dans le corpus 1, le même thémanyme 1 d'une question peut être en relation avec un thémanyme 2 de la réponse, deux thémanymes 2 de la notice descriptive et un thémanyme 2 de termes d'indexation.

5.1.3.2. Nombre de liens sémantiques

Au cœur de notre analyse se trouvent les liens sémantiques entre les thémanymes des quatre sources du corpus général (QRNTix). Le nombre de liens sémantiques entre les sources correspond au nombre de paires de thémanymes analysées. Les deux corpus contiennent tous les deux environ 600 liens sémantiques (voir Tableau XLIX).

Tableau XLIX. Nombre de liens sémantiques entre les sources (corpus général)

Sources comparées	Corpus 1		Corpus 2		Corpus général	
	Nombre	Pourcentage	Nombre	Pourcentage	Nombre	Pourcentage
Q R	142	23,7%	116	19,0%	258	21,3%
Q N	127	21,2%	349	57,1%	476	39,4%
Q Tix	28	4,7%	-	-	28	2,3%
R N	258	43,0%	146	23,9%	404	33,4%
R Tix	44	7,4%	-	-	44	3,6%
Total	**599**	100%	**611**	100%	**1210**	100%

En effet, le nombre de liens entre les deux corpus est du même ordre (599 et 611), malgré une différence du nombre de questions (10 pour le corpus 1 et 20 pour le corpus 2). Cette divergence s'explique par la manière dont les liens sémantiques ont été codés : de manière systématique dans le corpus 1, de manière plus intuitive dans le corpus 2 (voir section 4.2.5). La technique appliquée pour constituer le corpus 2 nous apparaît complémentaire de la première parce qu'elle reproduit la démarche intuitive d'un archiviste qui répond à une question d'usager en fonction de ses connaissances, y compris linguistiques et encyclopédiques. Remarquons d'ores et déjà que le nombre de liens sémantiques entre la question et la notice descriptive est deux fois plus important dans le corpus 2 (57,1%) que dans le corpus 1 (21,2%) et que le nombre de liens sémantiques entre la réponse et la notice descriptive est plus élevé dans le corpus 1 (43%) que dans le corpus 2 (23,9%).

Tableau L. Nombre de liens sémantiques entre les sources (corpus général) Q Instr

Sources comparées	Corpus 1		Corpus 2		Corpus général	
	Nombre	Pourcentage	Nombre	Pourcentage	Nombre	Pourcentage
Q R	142	23,7%	116	19,0%	258	21,3%
Q Instr	155	25,9%	349	57,1%	504	41,7%
R Instr	302	50,4%	146	23,9%	448	37
Total	**599**	100%	**611**	100%	**1210**	100%

Le Tableau L rend compte des mêmes calculs mais du point de vue de la dichotomie vocabulaire des usagers *versus* vocabulaire employé dans les instruments de recherche. Les liens se répartissent selon les polyséquences et les établissements de la manière suivante (voir Tableau LI).

Tableau LI. Répartition du nombre de liens par établissement (corpus général)

ID polyséquence	nb liens / polyséquence	établissement	nb liens / établ.	nb polyséquences /établ.
a014q	16	Établ 1	480	8
a017q	72			
a025q	88			
a033q	46			
a034q	12			
a037q	5			
a050q	74			
a055q	167			
a072q	80	Établ 2	119	2
a078q	39			
corpus 1	**599**	**total corpus 1**	**599**	**10**
a001q	13	Établ 3	295	6
a002qa	20			
a003q	106			
a004qa	47			
a006q	34			
a012qa	75			
a070q	13	Établ 2	316	14
a074q	15			
a075q	66			
a076q	17			
a079q	19			
a082q	35			
a085q	10			
a094q	36			
a105q	35			
a108q	28			
a109q	12			
a111q	12			
a116q	6			
a117q	12			
corpus2	**611**	**total corpus 2**	**611**	**20**
corpus général	**-**	**total corpus général**	**1210**	**30**

Les 30 questions d'usager sont à l'origine de 30 polyséquences, dont 8 sont issues de l'établissement 1, 16 sont issues de l'établissement 2 et six sont issues de l'établissement 3.

5.2. La relation d'identité

La relation d'identité a été comptabilisée en tant que telle lorsque deux thémanymes étaient absolument identiques, du point de vue de la forme et du sens, tel qu'indiqué dans notre échelle d'écart sémantique (voir section 4.3.1). La présente section indique les résultats relatifs à l'objectif de recherche 1 : *Vérifier empiriquement l'existence de l'écart sémantique présupposé*. Elle est organisée en fonction des deux questions de recherche émises pour atteindre cet objectif, à savoir la question de recherche 1 : *Quels thémanymes sont différents entre les termes de recherche, de description, d'indexation ?* (voir section 5.2.1) et la question de recherche 3 : *Quelle est la fréquence d'identité entre les thémanymes des usagers (dans les courriels d'usagers à la référence) et ceux des instruments de recherche (dans la notice descriptive ou l'index) ?* (voir section 5.2.2).

5.2.1. Validation de l'écart sémantique présupposé (QR1)

À partir de l'application de l'échelle d'écart sémantique, nous constatons que le corpus général n'inclut pas 100% de relations d'identité (voir section 5.2.2.1). Nous pouvons donc affirmer qu'il existe un écart sémantique entre le vocabulaire des usagers et celui employé par les archivistes.

5.2.2. Fréquence de la relation d'identité (QR3)

L'échelle d'écart sémantique a été appliquée à l'ensemble de nos données. Mais nous sommes en mesure d'identifier les proportions d'apparition de la relation d'identité dans le corpus général (voir section 5.2.2.1), ainsi qu'en fonction des établissements participants dont émanent les thémanymes comparés (voir section 5.2.2.2).

5.2.2.1. La relation d'identité dans le corpus

La présence de la relation d'identité (écart 0 dans l'échelle d'écart sémantique) a d'abord été calculée au sein de chaque corpus (corpus 1 et 2), puis dans le corpus général. La relation d'identité représente 12,1% des liens identifiés dans notre corpus général (voir Tableau LII).

Tableau LII. Relation d'identité (corpus général)

Relations sémantiques entre des thémanymes	**Corpus 1** Nombre (%)	**Corpus 2** Nombre (%)	**Corpus général** Nombre (%)
Relation d'identité (0)	72 (12%)	74 (12,1%)	146 (12,1%)
Autres liens	527 (88%)	537 (87,9%)	1064 (87,9%)
Total	599 (100%)	611 (100%)	1210 (100%)

La relation d'identité n'est pas présente uniformément entre toutes les sources (voir Tableau LIII). Dans le corpus 1 et dans le corpus 2, elle représente environ 5% des liens entre la question et la réponse; c'est entre ces deux sources qu'elle est la plus forte. Entre toutes les autres sources cumulées, la proportion de la relation d'identité oscille entre 0,3% et 3,7% pour représenter cumulativement 7% des relations d'identité dans les deux corpus.

Tableau LIII. La relation d'identité par sources comparées (corpus général)

Sources comparées	**Corpus 1 (n=599)**	**Corpus 2 (n=611)**	**Corpus général (n=1210)**
	Nombre (%)	Nombre (%)	Nombre (%)
Q R	32 (5,3%)	30 (4,9%)	62 (5,1%)
Q N	12 (2,0%)	21 (3,4%)	33 (2,7%)
R N	22 (3,7%)	23 (3,8%)	45 (3,7%)
Q Tix	2 (0,3%)	-	2 (0,3%)
R Tix	4 (0,7%)	-	4 (0,7%)
Total	72 (12%)	74 (12,1%)	146 (12,1%)

Il est à souligner que dans le corpus 2, les questions et les notices entretiennent 3,4% de relation d'identité alors qu'elles n'entretiennent que 2% de cette relation dans le corpus 1.

Le but de notre recherche est de connaître quelles sont la nature et l'ampleur de l'écart sémantique présupposé entre le vocabulaire des usagers qui recherchent des documents d'archives patrimoniaux et le vocabulaire employé par les archivistes dans les instruments de recherche pour l'accès thématique à des documents d'archives patrimoniaux. Ainsi, parmi les calculs suivants, nous avons regroupé les résultats relatifs à la notice descriptive et aux termes d'indexation. Nous cherchons la relation d'identité entre les thémanymes de la question ou de la réponse et ceux des instruments de recherche – Q (N + Tix), appelé Q Instr ou R (N + Tix), appelé R Instr (voir Tableau LIV).

Tableau LIV. La relation d'identité par sources comparées (corpus général) Q Instr

Sources comparées	Corpus 1 (n=599)	Corpus 2 (n=611)	Corpus général (n=1210)
	Nombre (%)	Nombre (%)	Nombre (%)
Q R	32 (5,3%)	30 (4,9%)	62 (5,1%)
Q Instr	14 (2,3%)	21 (3,4%)	35 (2,9%)
R Instr	26 (4,4%)	23 (3,8%)	49 (4,1%)
Total	72 (12%)	74 (12,1%)	146 (12,1%)

Nous trouvons 2,9% de correspondance à l'identique entre les thémanymes appartenant au vocabulaire des usagers et ceux présents dans les instruments de recherche pour l'ensemble du corpus. Ces chiffres prendront davantage de sens en discussion.

5.2.2.2. La relation d'identité dans les établissements participants

La relation d'identité a ensuite été calculée pour chacun des établissements participants (voir Tableau LV). Le calcul a été opéré pour les deux corpus et le corpus global afin de rassembler les données issues de l'établissement 2 dont les données sont présentes dans le corpus 1 et dans le corpus 2.

Tableau LV. Relation d'identité par établissement (corpus général)

Établissements	Corpus 1		Corpus 2		Corpus général	
	Nombre	Pourcentage % relation d'identité / autres relations	Nombre	Pourcentage % relation d'identité / autres relations	Nombre	Pourcentage % relation d'identité / autres relations
Établ. 1	65	90,3 / 10,9	-	-	65	44,5 / 5,4
Établ. 2	7	9,7 / 1,2	41	55,4 / 6,7	48	32,9 / 4,0
Établ. 3	-	-	33	44,6 / 5,4	33	22,6 / 2,7
Total	72	100 / 12	74	100 / 12,1	146	100 / 12,1

Le pourcentage a été calculé d'abord pour la relation d'identité, puis par rapport à l'ensemble des liens sémantiques de chaque corpus. Ainsi, dans les données issues de l'établissement 1, nous avons identifié 65 liens sémantiques étiquetés en tant que relation d'identité (écart 0). Ces 65 liens représentent 90,3% des liens sémantiques étiquetés en tant que relation d'identité dans le corpus 1 (n=72) et 10,9% des liens sémantiques identifiés dans le corpus 1.

La répartition inégale de la relation d'identité entre les établissements participants laisse penser que les habitudes de réponse de l'archiviste de référence divergent d'un

établissement à l'autre. En effet, dans l'établissement 2, les réponses de l'archiviste débutaient toutes par la reprise quasi mot pour mot des termes employés par l'usager, alors que dans les réponses des autres établissements, la reprise du sujet de recherche de l'usager se faisait souvent avec une variation dans le vocabulaire. Cependant, l'apparition de la relation d'identité pourrait également être liée à d'autres facteurs, tels que la complexité de la question d'usager, par exemple.

La relation d'identité a été traitée séparément pour faire ressortir l'existence d'un écart sémantique démontré empiriquement (cf. objectifs de recherche 1 et 3, voir section 2.2). La prochaine section traite de l'ensemble des liens sémantiques, identifiés grâce à l'échelle d'écart sémantique, entre les thémanymes du vocabulaire des usagers et celui employé dans les instruments de recherche de notre corpus.

5.3. Les relations sémantiques qui caractérisent un écart sémantique

La présence d'une relation sémantique autre que l'identité (écart 0 dans l'échelle d'écart sémantique) indique qu'il existe un écart entre les deux thémanymes. Les liens sémantiques ont été comptabilisés entre deux thémanymes qui différaient du point de vue de la forme ou du sens, selon la gradation établie dans l'échelle d'écart sémantique (voir section 4.3.1). La présente section indique les résultats relatifs aux objectifs de recherche 2 *Qualifier l'écart sémantique entre le vocabulaire des usagers qui recherchent des documents d'archives patrimoniaux par sujet et le vocabulaire présent dans les instruments de recherche pour l'accès à des documents d'archives patrimoniaux* et 3 *Quantifier cet écart sémantique*. Elle est organisée en fonction des deux questions de recherche émises pour atteindre ces objectifs, à savoir respectivement la question de recherche 2 : *Quelles sont les relations sémantiques qui existent entre les thémanymes appartenant au vocabulaire des usagers et ceux appartenant à celui des instruments de recherche ?* (voir section 5.3.1) et la question de recherche 4 : *Quelle est la fréquence de chacune des relations sémantiques qui caractérisent un écart sémantique entre les thémanymes appartenant au vocabulaire des usagers et ceux appartenant à celui des instruments de recherche ?* (voir section 5.3.2).

5.3.1. Les relations qui caractérisent l'écart sémantique (QR2)

La comparaison de deux thémanymes de deux sources différentes a permis d'identifier les liens sémantiques qu'ils entretenaient, à partir de l'échelle d'écart sémantique et de son application à nos données (voir section 5.1.1 et 5.1.2). Après comptabilisation (voir section 5.3.2), nous constatons que les relations de variation orthographique (1), de variation graphique (3), d'ellipse (8) et de paraphrase (9) sont absentes du corpus 1. Par contre, le corpus 2 contient l'ensemble des valeurs de l'échelle d'écart sémantique, même si parfois le lien n'est présent qu'une seule fois.

5.3.2. La fréquence des relations qui caractérisent l'écart sémantique (QR4)

Cette section identifie la fréquence des liens présents dans les deux corpus, selon les sources dont émanent les thémanymes. Nous rappelons les résultats relatifs à la relation d'identité et ajoutons la relation de correspondance référentielle (écart 14) afin de faciliter la comparaison ultérieure et la discussion ultérieure.

Les deux tableaux suivants identifient le nombre de liens sémantiques présents dans le corpus 1, en fonction des sources dont sont tirés les thémanymes (voir Tableau LVI et Tableau LVII).

Tableau LVI. Liens sémantiques présents dans le corpus 1 (nombre)

Relation /Source	écart 0	écart 2	écart 4	écart 5	écart 6	écart 7	écart 10	écart 11	écart 12	écart 13	écart 14	Total général (nombre)
Q R	32	1	9	11	2	0	3	21	28	33	2	142
Q N	12	0	0	3	1	0	2	21	29	52	7	127
R N	22	0	8	6	1	1	2	33	58	114	13	258
Q Tix	2	0	1	6	1	0	0	5	5	5	3	28
R Tix	4	0	4	2	2	0	0	10	6	16	0	44
Total général (nombre)	72	1	22	28	7	1	7	90	126	220	25	599

Tableau LVII. Liens sémantiques présents dans le corpus 1 (%)

Relation /Source	écart 0	écart 2	écart 4	écart 5	écart 6	écart 7	écart 10	écart 11	écart 12	écart 13	écart 14	Total général (%)
Q R	5,3	0,2	1,5	1,8	0,3	0,0	0,5	3,5	4,7	5,5	0,3	23,7
Q N	2,0	0,0	0,0	0,5	0,2	0,0	0,3	3,5	4,8	8,7	1,2	21,2
R N	3,7	0,0	1,3	1,0	0,2	0,2	0,3	5,5	9,7	19,0	2,2	43,1
Q Tix	0,3	0,0	0,2	1,0	0,2	0,0	0,0	0,8	0,8	0,8	0,5	4,7
R Tix	0,7	0,0	0,7	0,3	0,3	0,0	0,0	1,7	1,0	2,7	0,0	7,3
Total général (%)	12,0	0,2	3,7	4,7	1,2	0,2	1,2	15,0	21,0	36,7	4,2	100,0

Les deux tableaux suivants identifient les liens sémantiques présents dans le corpus 2, en fonction des sources dont sont tirés les thémanymes (Tableau LVIII et 0). Ici aussi, nous rappelons les résultats relatifs à la relation d'identité et ajoutons la relation de correspondance référentielle (écart 14) afin de faciliter la comparaison ultérieure.

Tableau LVIII. Liens sémantiques présents dans le corpus 2 (nombre)

Relation /Source	éc. 0	éc. 1	éc. 2	éc. 3	éc. 4	éc. 5	éc. 6	éc. 7	éc. 8	éc. 9	éc. 10	éc. 11	éc. 12	éc.13	éc. 14	Total général (nombre)
Q R	30	1	4	1	6	5	4	9	8	2	5	29	33	31	15	183
Q N	21	0	0	1	10	0	3	13	4	1	0	54	43	113	19	282
R N	23	0	1	0	8	2	2	3	1	3	0	20	33	49	1	146
Total général (nombre)	74	1	5	2	24	7	9	25	13	6	5	103	109	193	35	611

Tableau LIX. Liens sémantiques présents dans le corpus 2 (%)

Relation /Source	éc. 0	éc. 1	éc. 2	éc. 3	éc. 4	éc. 5	éc. 6	éc. 7	éc. 8	éc. 9	éc. 10	éc.11	éc.12	éc. 13	éc. 14	Total général (%)
Q R	4,9	0,2	0,7	0,2	1,0	0,8	0,7	1,5	1,3	0,3	0,8	4,7	5,4	5,1	2,5	30,0
Q N	3,4	0,0	0,0	0,2	1,6	0,0	0,5	2,1	0,7	0,2	0,0	8,8	7,0	18,5	3,1	46,2
R N	3,8	0,0	0,2	0,0	1,3	0,3	0,3	0,5	0,2	0,5	0,0	3,3	5,4	8,0	0,2	23,9
Total général (%)	12,1	0,2	0,8	0,3	3,9	1,1	1,5	4,1	2,1	1,0	0,8	16,9	17,8	31,6	5,7	100,0

Les deux tableaux suivants identifient les liens sémantiques présents dans le corpus (corpus 1 et 2), en fonction des sources dont sont tirés les thémanymes (Tableau LX et

Tableau LXI). Certaines paires de thémanymes ont été comparées sans qu'aucun lien sémantique ne soit identifié; ceci prend la forme de zéros dans les tableaux.

Tableau LX. Liens sémantiques présents dans le corpus général (nombre)

Relation /Source	éc. 0	éc. 1	éc. 2	éc. 3	éc. 4	éc. 5	éc. 6	éc. 7	éc. 8	éc. 9	éc. 10	éc. 11	éc. 12	éc. 13	éc. 14	Total général
Q R	62	1	5	1	15	16	6	9	8	2	8	50	61	64	17	325
Q N	33	0	0	1	10	3	4	13	4	1	2	75	72	165	26	409
R N	45	0	1	0	16	8	3	4	1	3	2	53	91	163	14	404
Q Tix	2	0	0	0	1	6	1	0	0	0	0	5	5	5	3	28
R Tix	4	0	0	0	4	2	2	0	0	0	0	10	6	16	0	44
Total	146	1	6	2	46	35	16	26	13	6	12	193	235	413	60	1210

Tableau LXI. Liens sémantiques présents dans le corpus général (%)

Relation /Source	éc. 0	éc. 1	éc. 2	éc. 3	éc. 4	éc. 5	éc. 6	éc. 7	éc. 8	éc. 9	éc. 10	éc. 11	éc. 12	éc. 13	éc. 14	Total général (%)
Q R	5,1	0,1	0,4	0,1	1,2	1,3	0,5	0,7	0,7	0,2	0,7	4,1	5,0	5,3	1,4	26,9
Q N	2,7	0,0	0,0	0,1	0,8	0,2	0,3	1,1	0,3	0,1	0,2	6,2	6,0	13,6	2,1	33,8
R N	3,7	0,0	0,1	0,0	1,3	0,7	0,2	0,3	0,1	0,2	0,2	4,4	7,5	13,5	1,2	33,4
Q Tix	0,2	0,0	0,0	0,0	0,1	0,5	0,1	0,0	0,0	0,0	0,0	0,4	0,4	0,4	0,2	2,3
R Tix	0,3	0,0	0,0	0,0	0,3	0,2	0,2	0,0	0,0	0,0	0,0	0,8	0,5	1,3	0,0	3,6
Total (%)	12,1	0,1	0,5	0,2	3,8	2,9	1,3	2,1	1,1	0,5	1,0	16,0	19,4	34,1	5,0	100,0

Les deux tableaux suivants identifient les liens sémantiques présents dans le corpus général (corpus 1 et 2), en fonction des sources dont sont tirés les thémanymes, mais cette fois-ci en faisant ressortir l'écart entre le vocabulaire des usagers et celui employé dans les instruments de recherche (Tableau LXII et Tableau LXIII).

Tableau LXII. Liens sémantiques présents dans le corpus général Q Instr (nombre)

Relation /Source	éc. 0	éc. 1	éc. 2	éc. 3	éc. 4	éc. 5	éc. 6	éc. 7	éc. 8	éc. 9	éc. 10	éc. 11	éc. 12	éc. 13	éc. 14	Total général
Q R	62	1	5	1	15	16	6	9	8	2	8	50	61	64	17	325
Q Instr	35	0	0	1	11	9	5	13	4	1	2	80	77	170	29	437
R Instr	49	0	1	0	20	10	5	4	1	3	2	63	97	179	14	448
Total	146	1	6	2	46	35	16	26	13	6	12	193	235	413	60	1210

Tableau LXIII. Liens sémantiques présents dans le corpus général Q Instr (%)

Relation /Source	éc. 0	éc. 1	éc. 2	éc. 3	éc. 4	éc. 5	éc. 6	éc. 7	éc. 8	éc. 9	éc. 10	éc. 11	éc. 12	éc. 13	éc. 14	Total général (%)
Q R	5,1	0,1	0,4	0,1	1,2	1,3	0,5	0,7	0,7	0,2	0,7	4,1	5,0	5,3	1,4	26,9
Q Instr	2,9	0,0	0,0	0,1	0,9	0,7	0,4	1,1	0,3	0,1	0,2	6,6	6,4	14,0	2,4	36,1
R Instr	4,0	0,0	0,1	0,0	1,7	0,8	0,4	0,3	0,1	0,2	0,2	5,2	8,0	14,8	1,2	37,0
Total (%)	12,1	0,1	0,5	0,2	3,8	2,9	1,3	2,1	1,1	0,5	1,0	16,0	19,4	34,1	5,0	100,0

À nouveau, ces chiffres prendront davantage de sens en discussion.

5.4. Regroupements de relations sémantiques

Afin de dégager une plus grande variété de remarques des résultats relatifs aux liens sémantiques et à leur fréquence, nous avons regroupé les échelons de l'échelle d'écart sémantique selon divers modes. L'échelle d'écart sémantique présente 13 à 14 degrés d'écart – des valeurs – qui permettent d'identifier les phénomènes de forme et de sens en jeu dans la relation sémantique entre deux expressions linguistiques porteuses du sujet (thémanymes), en fonction de la source (QR, QN, QTix, QInstr [N+Tix], RN, R Tix, RInstr [N+Tix]). Ainsi, les regroupements ont été effectués sur la base d'une part de la nature linguistique (forme et sens) des relations sémantiques de l'échelle d'écart sémantique et d'autre part de leur signification potentielle pour dans un contexte de référence archivistique.

5.4.1. Regroupement multiple (écarts 0-5, 6-9, 10-12, 13-14)

À partir de l'observation de l'échelle d'écart sémantique (voir section 4.3.1, Tableau XX), nous pouvons identifier des écarts qui pourraient être regroupés. Ainsi, les écarts 1 à 5 ont *a priori* un faible impact sur la compréhension : variation orthographique, variation de longueur (sigles et acronymes), variation graphique (p. ex. l'insertion de traits d'union), variation de casse et variation flexionnelle (p. ex. passage du singulier au pluriel). Notre échelle inclut une colonne de vulgarisation des relations sémantiques. Dans tous ces cas, nous mentionnons qu'il s'agit de la même forme ou de presque la même forme, tant la variation affecte peu le sens. La suite de l'échelle contient les variations dérivationnelles, syntaxiques, l'ellipse d'un complément et la paraphrase (écarts 6 à 9). Les écarts suivants correspondent

aux relations thésaurales (écarts 10-12). Les deux derniers écarts (écarts 13 et 14) rendent compte d'un lien existant entre les deux thémanymes de sources différentes mais parfois ténu.

Tableau LXIV. Regroupement multiple (0-5, 6-9, 10-12, 13-14) (corpus 1)

Sources	écarts 0-5	écarts 6-9	écarts 10-12	écarts 13-14	Total
Q R	53 (37,3%)	2 (1,4%)	52 (36,6%)	35 (24,7%)	142 (100%)
Q N	15 (11,8%)	1 (0,8%)	52 (40,9%)	59 (46,5%)	127 (100%)
R N	36 (14%)	2 (0,8%)	93 (36%)	127 (49,2%)	258 (100%)
Q Tix	9 (32,1%)	1 (3,6%)	10 (37,7%)	8 (28,6%)	28 (100%)
R Tix	10 (22,7%)	2 (4,5%)	16 (36,4%)	16 (36,4%)	44 (100%)
Total	123 (20,5%)	8 (1,3%)	223 (37,3%)	245 (40,9%)	599 (100%)

Tableau LXV. Regroupement multiple (0-5, 6-9, 10-12, 13-14) (corpus 2)

Sources	écarts 0-5	écarts 6-9	écarts 10-12	écarts 13-14	Total
Q R	47 (25,7%)	23 (12,6%)	67 (36,6%)	46 (25,1%)	183 (100%)
Q N	32 (11,3%)	21 (7,5%)	97 (34,4%)	132 (46,8%)	282 (100%)
R N	34 (23,3%)	9 (6,2%)	53 (36,3%)	50 (34,2%)	146 (100%)
Total	113 (18,5%)	53 (8,7%)	217 (35,5%)	228 (37,3%)	611 (100%)

Tableau LXVI. Regroupement multiple (0-5, 6-9, 10-12, 13-14) (corpus général)

Sources	écarts 0-5	écarts 6-9	écarts 10-12	écarts 13-14	Total
Q R	100 (30,8%)	25 (7,7%)	119 (36,6%)	81 (24,9%)	325 (100%)
Q N	47 (11,5%)	22 (5,4%)	149 (36,4%)	191 (46,7%)	409 (100%)
R N	70 (17,3%)	11 (2,7%)	146 (36,2%)	177 (43,8%)	404 (100%)
Q Tix	9 (32,1%)	1 (3,6%)	10 (37,7%)	8 (28,6%)	28 (100%)
R Tix	10 (22,7%)	2 (4,5%)	16 (36,4%)	16 (36,4%)	44 (100%)
Total	236 (19,5%)	61 (5%)	440 (36,4%)	473 (39,1%)	1210 (100%)

Tableau LXVII. Regroupement multiple (0-5, 6-9, 10-12, 13-14) (corpus général) Q Instr

Sources	écarts 0-5	écarts 6-9	écarts 10-12	écarts 13-14	Total
Q R	100 (30,8%)	25 (7,7%)	119 (36,6%)	81 (24,9%)	325 (100%)
Q Instr	56 (12,8%)	23 (5,3%)	159 (36,4%)	199 (39,1%)	437 (100%)
R Instr	80 (17,8%)	13 (2,9%)	162 (63,2%)	193 (43,1%)	448 (100%)
Total	236 (19,5%)	61 (5%)	440 (36,4%)	473 (39,1%)	1210 (100%)

Les Tableau LXIV, Tableau LXV, Tableau LXVI et Tableau LXVII se lisent de gauche à droite pour les premières lignes et non de haut en bas. Nous avons créé le Tableau LXVII pour voir plus directement les nombres et pourcentages entre les questions ou les réponses et les instruments de recherche.

5.4.2. Regroupement d'identité à équivalence (écarts 0-10 et 11-14)

La colonne de vulgarisation de notre échelle d'écart sémantique permet d'identifier une césure entre les écarts qui existent entre deux formes de même sens et les écarts qui rendent compte d'une différence de sens. Cette césure se produit lorsque l'on passe de la relation d'équivalence (écart 10) à la relation de hiérarchie (écart 11). Un premier regroupement consiste alors à regrouper les valeurs 0 à 10 d'un côté et 11 à 14 de l'autre. Les écarts 11 à 13 rendent compte en particulier de relations paradigmatiques, non syntagmatiques, auxquels sont ajoutées les relations référentielles.

Tableau LXVIII. Écart de forme et écart de sens (corpus 1)

Sources	écarts 0-10	écarts 11-14	Total
Q R	58 (40,8%)	84 (59,2%)	142 (100%)
Q N	18 (14,2%)	109 (85,8%)	127 (100%)
R N	40 (15,5%)	218 (84,5%)	258 (100%)
Q Tix	10 (35,7%)	18 (64,3%)	28 (100%)
R Tix	12 (27,3%)	32 (72,7%)	44 (100%)
Total	138 (23%)	461 (77%)	599 (100%)

Tableau LXIX. Écart de forme et écart de sens (corpus 2)

Sources	écarts 0-10	écarts 11-14	Total
Q R	75 (41%)	108 (59%)	183 (100%)
Q N	53 (18,8%)	229 (81,2%)	282 (100%)
R N	43 (29,5%)	103 (70,5%)	146 (100%)
Total	171 (28%)	440 (72%)	611 (100%)

Tableau LXX. Écart de forme et écart de sens (corpus général)

Sources	écart 0-10	écart 11-14	Total
Q R	133 (40,9%)	192 (59,1%)	325 (100%)
Q N	71 (17,4%)	338 (82,6%)	409 (100%)
R N	83 (20,5%)	321 (79,5%)	404 (100%)
Q Tix	10 (35,7%)	18 (64,3%)	28 (100%)
R Tix	12 (27,3%)	32 (72,7%)	44 (100%)
Total	309 (25,5%)	901 (74,5%)	1210 (100%)

Tableau LXXI. Écart de forme et écart de sens (corpus général) Q Instr

Sources	écart 0-10	écart 11-14	Total
Q R	133 (40,9%)	192 (59,1%)	325 (100%)
Q Instr	81 (18,5%)	356 (81,5%)	437 (100%)
R Instr	95 (21,2%)	353 (78,8%)	448 (100%)
Total	309 (25,5%)	901 (74,5%)	1210 (100%)

Le Tableau LXVIII, le Tableau LXIX, le Tableau LXX et le Tableau LXXI se lisent de gauche à droite pour les premières lignes et non de haut en bas. Nous pouvons lire les nombres et pourcentages de chaque paire de thémanymes selon la source. En outre, nous voyons que l'écart de la relation d'identité, les variations formelles jusqu'à l'équivalence représente 23% des liens sémantiques dans le corpus 1, 28% dans le corpus 2 et 25,5% dans le corpus général.

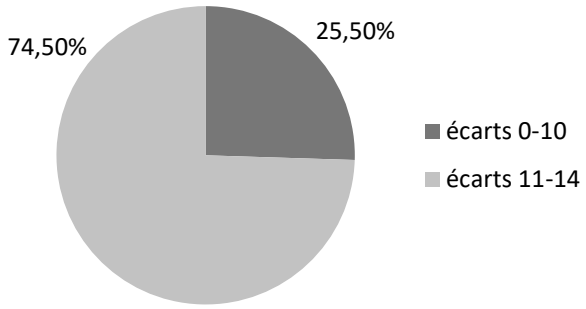

Figure 22. Regroupement des relations d'identité à équivalence (écarts 0-10)

Sur la Figure 22, nous constatons une proportion d'un quart pour les relations d'identité à équivalence et de trois quarts pour les relations qui modifient profondément le

sens et non seulement la forme (relations hiérarchiques, associatives, champ sémantique et relations référentielles).

5.4.3. Regroupement de mise en valeur des forts écarts (0-9, 10 à 14)

L'observation des résultats pour chacune des relations de l'échelle d'écart sémantique nous pousse à regrouper les petits écarts (écarts 0-9) dont les liens ont une faible fréquence, pour faire ressortir la fréquence de chacun des grands écarts (écarts 10, 11, 12, 13 et 14).

Tableau LXXII. Regroupement sur les forts écarts (0-9, 10 à 14) (corpus 1)

Sources	écarts 0-9	écart 10	écart 11	écart 12	écart 13	écart 14	Total
Q R	55 (39,7%)	3 (2,1%)	21 (14,8%)	28 (19,7%)	33 (23,2%)	2 (1,4%)	142 (100%)
Q N	16 (12,6%)	2 (1,6%)	21 (16,5%)	29 (22,8%)	52 (44,2%)	7 (5%)	127 (100%)
R N	38 (14,7%)	2 (0,8%)	33 (12,8%)	58 (22,5%)	114 (44,2%)	13 (5%)	258 (100%)
Q Tix	10(35,7%)	0	5 (17,9%)	5 (17,9%)	5 (17,9%)	3 (10,6%)	28 (100%)
R Tix	12(27,3%)	0	10 (22,7%)	6 (13,6%)	16 (36,4%)	0	44 (100%)

Tableau LXXIII. Regroupement sur les forts écarts (0-9, 10 à 14) (corpus 2)

Sources	écarts 0-9	écart 10	écart 11	écart 12	écart 13	écart 14	Total
Q R	70 (38,3%)	5 (2,7%)	29 (15,8%)	33 (18%)	31 (16,9%)	15 (8,2%)	183 (100%)
Q N	53 (18,8%)	0	54 (19,1%)	43 (15,2%)	113 (40,1%)	19 (6,7%)	282 (100%)
R N	43 (29,5%)	0	20 (13,7%)	33 (22,6%)	49 (33,6%)	1 (0,7%)	146 (100%)
Total	166 (27,2%)	5 (0,8%)	103 (16,9%)	109 (17,8%)	193 (31,6%)	35 (5,7%)	611 (100%)

Tableau LXXIV. Regroupement sur les forts écarts (0-9, 10 à 14) (corpus général)

Source	écarts 0-9	écart 10	écart 11	écart 12	écart 13	écart 14	Total
Q R	125 (38,4%)	8 (2,5%)	50 (15,4%)	61 (18,8%)	64 (19,7%)	17 (5,2%)	325 (100%)
Q N	69 (16,9%)	2 (0,5%)	75 (18,3%)	72 (17,6%)	165 (40,3%)	26 (6,4%)	409 (100%)
R N	81 (20,1%)	2 (0,5%)	53 (13,1%)	91 (22,5%)	163 (40,3%)	14 (3,5%)	404 (100%)
Q Tix	10 (35,7%)	0	5 (17,9%)	5 (17,9%)	5 (17,9%)	3 (10,6%)	28 (100%)
R Tix	12 (27,3%)	0	10 (22,7%)	6 (13,6%)	16 (36,4%)	0	44 (100%)
Total	297 (24,5%)	12 (1%)	193 (16%)	235 (19,4%)	413 (34,1%)	60 (5%)	1210 (100%)

Tableau LXXV. Regroupement sur les forts écarts (0-9, 10 à 14) (corpus général) Q Instr

Source	écarts 0-9	écart 10	écart 11	écart 12	écart 13	écart 14	Total
Q R	125 (38,4%)	8 (2,5%)	50 (15,4%)	61 (18,8%)	64 (19,7%)	17 (5,2%)	325 (100%)
Q Instr	79 (18,1%)	2 (0,5%)	80 (18,3%)	77 (17,6%)	170 (38,9%)	29 (6,6%)	437 (100%)
R Instr	93 (20,8%)	2 (0,4%)	63 (14,1%)	97 (21,7%)	179 (39,9%)	14 (3,1%)	448 (100%)

Les Tableau LXXII, Tableau LXXIII, Tableau LXXIV et Tableau LXXV présentent les résultats relatifs à la mise en valeur des forts écarts.

5.5. Autre résultat : les formules introductives du sujet

Notre étude a permis de dégager d'autres résultats, parallèles à nos objectifs, mais venant enrichir la connaissance et la reconnaissance du VATAP. Nous avons identifié des formules introductives du sujet dans le contexte de la référence archivistique (voir section 4.2.4.1) grâce à l'analyse en facettes, lors de la caractérisation sémantique (voir section 4.2.5). Le Tableau LXXVI présente les principales formules identifiées, la source où elles ont été repérées ainsi que la forme lemmatisée. La forme lemmatisée est une forme considérée neutre, non marquée du point de vue grammatical. Nous avons regroupé les formules en fonction d'abord de leur fréquence et ensuite de la sémantique de leurs constructions. Une étude des actants sémantiques (tels qu'entendus par la Sémantique des cadres, Fillmore 1976) couplée à l'analyse en facettes déjà effectuée pourrait permettre d'approfondir le sujet.

Tableau LXXVI. Formules introductives du sujet

Source	Formules introductives	Forme lemmatisée
Q, N	je recherche une information/un document sur	Sur
Q, N	portant sur / sur lesquels portent / portent sur / portant sur le contenu de / portant notamment sur / porte sur	Porter sur
Q	je suis à la recherche d'archives de/sur	De/sur
R, N	concernent / concernant / qui concernent la / ne concerne pas exclusivement / concernant principalement / des informations concernant	Concerner
N	qui se rapportent à / se rapportent aussi à / se rapportant aux / rapportent les problématiques se rapportant à	Se rapporter à
Q	Auriez-vous des (documents) à ce sujet ou des pistes à nous fournir pour notre recherche ?	À ce sujet (par anaphore)
Q	Pourriez-vous me faire parvenir des documents (il m'est malheureusement impossible de me déplacer à votre institution) portant sur les aspects suivants	Porter sur les aspects suivants
Q	Plus spécifiquement, nous souhaitons effectuer des recherches en lien avec	En lien avec (qqch)
N	couvrant divers aspects liés à	Couvrir des aspects liés à

Source	Formules introductives	Forme lemmatisée
N	touchent des évènements liés au	Toucher des évènements liés à
N	liée à/ liés au	Liés à
N	traitent, entre autres sujets, de	Traiter de
N	plus particulièrement en ce qui a trait à	Avoir trait à
N	ils documentent / documente / sont également documentés avec / est aussi bien documenté dans	Documenter
N	nous présentent	Présenter qqch.
N	Témoigne de	Témoigner de
N	Ils ajoutent à notre connaissance de	Ajouter à notre connaissance de (qqch)
N	touchent à	Toucher à
N	représentant principalement	Représenter
N	À travers ces documents, on peut retracer	Retracer
N	relatifs à la / relatives au	Relatif à
N	racontant	Raconter
N	On y voit	Voir
N	À sa gauche, nous identifions	Identifier
N	présentent	Présenter
N	illustrent	Illustrer
N	Nous assistons	Assister
N	À gauche on aperçoit / Nous apercevons de gauche à droite / Nous apercevons	Apercevoir
N	montre	Montrer

Les formules introductives qui se démarquent sont sans nul doute : la préposition *sur* et la locution verbale *porter sur*, le verbe *concerner* surtout au participe présent, la préposition *de* (p. ex. les archives de la Seconde Guerre mondiale), *documenter qqch*., la locution adjectivale *lié à* précédée ou non de noms tels que *aspects* ou *événements* ou la locution prépositive *en lien avec*. Les derniers verbes du tableau sont issus de notices descriptives relatives à des documents iconographiques et rappellent le fait de voir.

Notre corpus est cependant actuellement trop restreint pour vérifier si ces formules sont suffisamment récurrentes pour être intégrées dans un patron de reconnaissance automatique du sujet exprimé dans des notices descriptives ou dans des questions d'usager posées par courriel. En effet, nous pensons que le genre de document (courriel de demande à la référence, notice descriptive) pourrait influencer le choix de la formule introductive, mais ceci reste une pure supposition dans le cadre de cette recherche.

Conclusion du chapitre

La présente section a rapporté la nature des liens sémantiques trouvés dans notre corpus : la relation d'identité ainsi que les relations sémantiques. Notre analyse rend compte par des calculs (nombre total, mode, moyenne et médiane) de la fréquence des liens

sémantiques en jeu dans notre corpus. Par là même, ils permettent d'affirmer la présence d'un fossé sémantique entre le vocabulaire des usagers et celui des archivistes et de le quantifier. Les proportions de liens des différents types de relations catégorisées à partir de l'échelle d'écart sémantique que nous avons bâtie nous permettent en sus de qualifier ce fossé sémantique et d'identifier les relations sémantiques les plus présentes dans notre corpus : le champ lexical, les relations thésaurales et la relation d'identité. En outre, l'identification des sources dont émanent les thémanymes nous permet de caractériser avec plus de justesse les deux vocabulaires appartenant au VATAP. Finalement, d'autres résultats tels que l'identification de formules introductives du sujet dans un contexte de référence archivistique sont venus enrichir la recherche. Dans la prochaine section nous interprétons ces résultats.

6. Discussion

Nous rappelons que le but de notre recherche est d'étudier l'écart sémantique présupposé entre le vocabulaire des usagers qui recherchent des documents d'archives patrimoniaux et celui employé par les archivistes dans les instruments de recherche pour l'accès thématique à des documents d'archives patrimoniaux. Pour ce faire, nous avons établi trois objectifs, à savoir vérifier l'existence de cet écart présupposé et s'il existe réellement, de le qualifier par l'étude des relations sémantiques en jeu entre les expressions porteuses de sujet (thémanymes) et de le quantifier par le relevé de leur fréquence. Ainsi, dans cette section, nous discutons dans un premier temps des caractéristiques générales de la recherche. Puis nous revenons sur la relation d'identité et les divers regroupements de relations sémantiques. Ces résultats qui portent sur le corpus général sont comparés avec les résultats de chacun des établissements. Cette comparaison nous permet de relativiser nos résultats : certaines tendances seraient-elles liées à la nature des regroupements et donc des relations sémantiques ? Ou bien seraient-elles liées à des habitudes issues de la culture organisationnelle de chaque établissement ? Ou encore seraient-elles liées à la nature même, intrinsèque de chaque séquence – Ensemble formé par une question, une réponse, une notice descriptive et éventuellement des termes d'indexation rattachés à cette notice (QRNTix) – ou chaque polyséquence – Ensemble de séquences qui ont la même question et la même réponse (p. ex. QRNTixNTixNTix) ? Notre recherche ne cherchait pas à découvrir les raisons à l'origine des phénomènes observés, mais elle permet de soulever les questions, point de départ de recherches futures.

Dans cette section, nous discutons des résultats relatifs à la relation d'identité (section 6.1) et aux trois regroupements de relations sémantiques (section 6.2), établis au chapitre précédent (voir section 5.4 : regroupement multiple, regroupement d'identité à équivalence et regroupement de mise en valeur des forts écarts). Nous présentons ici une mise en perspective de la recherche (section 6.3) Nous terminons par l'apport de la recherche sur les plans théorique, méthodologique et professionnel (section 6.4).

6.1. La relation d'identité

Notre recherche a pour but de vérifier empiriquement l'existence d'un écart présupposé entre le vocabulaire des usagers et celui que les archivistes ont employé dans les instruments de recherche. Le premier objectif de notre recherche est le suivant : *Valider empiriquement l'existence d'un écart sémantique présupposé entre le vocabulaire des usagers qui recherchent des documents d'archives patrimoniaux par sujet et le vocabulaire employé par les archivistes dans les instruments de recherche pour l'accès à des documents d'archives patrimoniaux*. La relation d'identité a pour corollaire inverse la vérification empirique de l'écart sémantique. En effet, si deux expressions linguistiques porteuses de sujet (thémanymes) n'entretiennent pas une relation d'identité, alors elles sont différentes et il existe un écart – faible ou fort – entre elles. La relation d'identité est donc le premier élément à prendre en considération pour répondre à notre première question de recherche : *lors de la comparaison du vocabulaire des usagers et celui employé dans les instruments de recherche, quelle proportion de thémanymes entretiennent une relation d'identité ?* Nous pouvons reformuler cette question de la façon suivante : est-ce que le corpus général n'inclut que la relation d'identité (un taux de 100%) ? Si c'était le cas, alors il n'y aurait pas d'écart, même faible, entre le vocabulaire des usagers et celui employé dans les instruments de recherche. Notre analyse montre que 12,1% des liens sémantiques identifiés dans le corpus général sont de l'identité stricte (écart 0 seulement). Ainsi, 87,9% des thémanymes comparés ne sont ni formellement ni sémantiquement identiques (voir Tableau LII). L'écart sémantique présupposé entre le vocabulaire des usagers et le vocabulaire dans les instruments de recherche est prouvé empiriquement dans notre corpus. Même en tenant compte des différences formelles minimes (voir section 6.2), nous sommes encore loin d'atteindre un taux de 100%, ce qui renforce l'observation de la présence de l'écart sémantique étudié.

La Figure 23 (basée sur le Tableau LII Relation d'identité (corpus général)) montre la faible proportion de liens sémantiques associés à la relation d'identité (écart 0) par rapport aux autres liens sémantiques dans notre corpus.

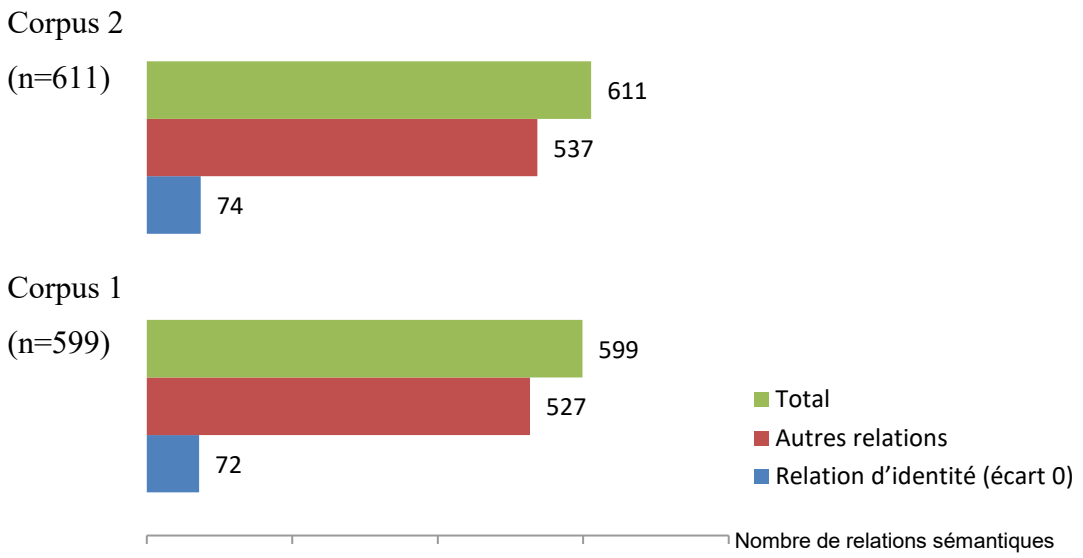

Figure 23. Nombre de liens sémantiques associés à la relation d'identité et aux autres relations (corpus général, n=1210)

La part de liens sémantiques est similaire dans les corpus 1et 2 qui sont de même taille. Par contre, les établissements présentent de légères différences quant à la présence de la relation d'identité (voir Figure 24). L'établissement 1 obtient le plus grand pourcentage. La répartition inégale de la relation d'identité entre les établissements participants laisse penser que les habitudes de réponse des archivistes de référence divergent d'un établissement à l'autre. Pourtant, les réponses des archivistes de l'établissement 2 débutaient toutes par la reprise quasi mot pour mot des formulations employées par l'usager, alors que dans les réponses des deux autres établissements, la reprise du sujet de recherche de l'usager se faisait souvent avec une variation (de faible écart). Ce résultat nous étonne donc particulièrement. Mais il est à noter que la reprise des éléments de la question de l'usager ne forme qu'une toute petite partie de la réponse de l'archiviste.

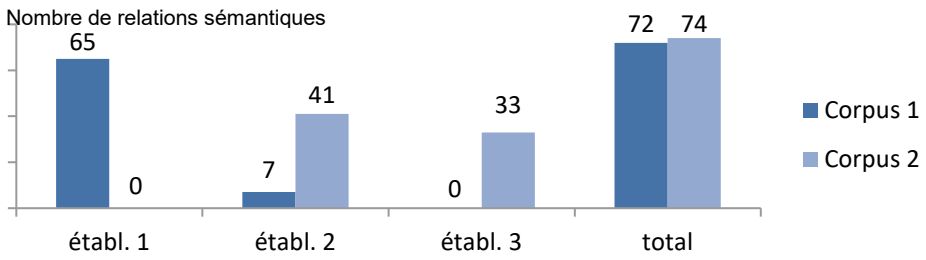

Figure 24. Relation d'identité dans les trois établissements (corpus général, n=1210)

La présence de la relation d'identité pourrait également être liée à d'autres facteurs, tels que la complexité de la question d'usager. En examinant les sources dont les paires de thémanymes sont issues, nous approchons mieux la disparité de la présence de la relation d'identité selon les établissements. La source dont sont extraits les thémanymes semble jouer un rôle dans la présence de la relation d'identité (voir Figure 25).

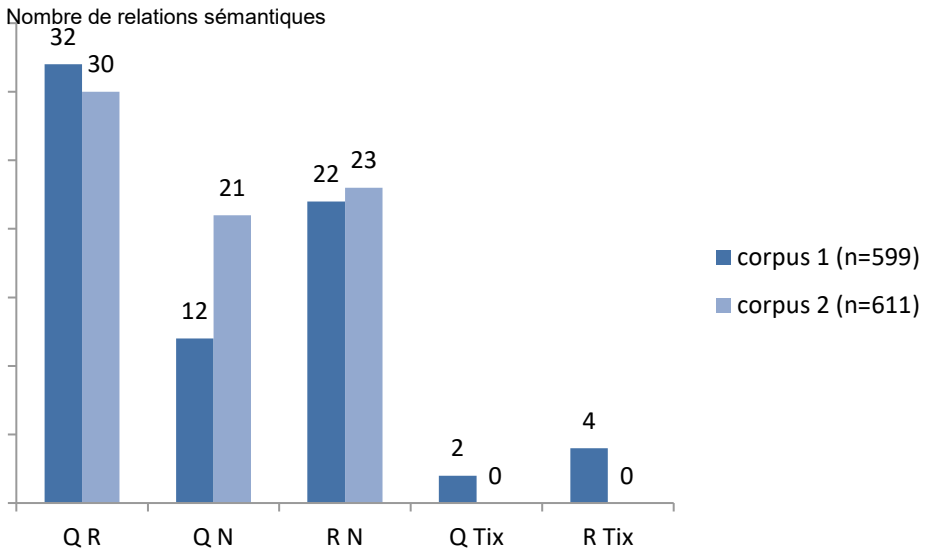

Figure 25. Relation d'identité selon les paires de thémanymes dans les sources (corpus général, n=1210)

Nous voyons que le nombre de liens sémantiques associés à la relation d'identité dans les corpus 1 et 2 est le plus élevé entre les thémanymes de la question de l'usager et de la réponse de l'archiviste (Q R), ensuite entre la réponse et la notice descriptive (R N). La Figure 26 montre ces mêmes calculs de manière à faire ressortir la relation d'identité entre le vocabulaire des usagers et celui employé dans les instruments de recherche, avec (Q R + R Instr) ou sans (Q Instr) l'intervention d'un archiviste de référence.

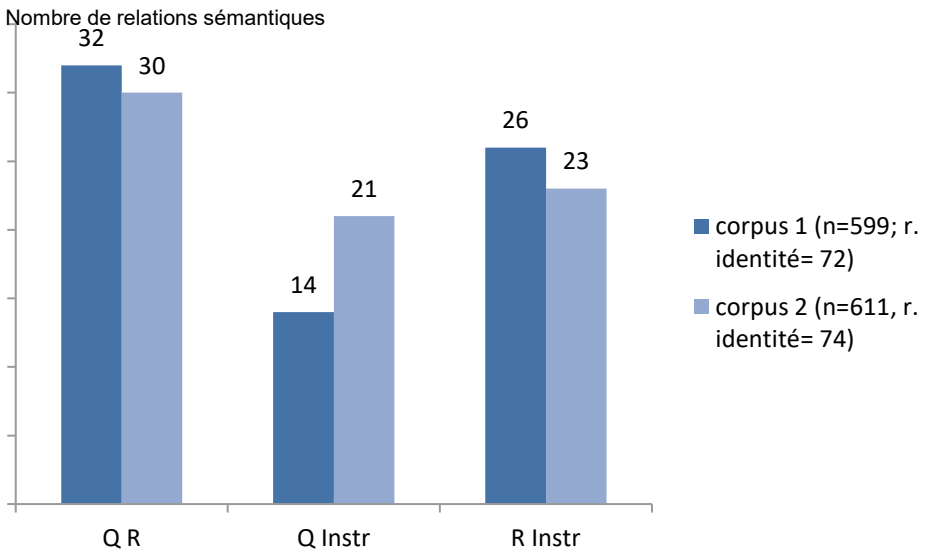

Figure 26. Proportion de la relation d'identité selon les paires de thémanymes de diverses sources (corpus général, n=1210) Q Instr

La plus faible présence de la relation d'identité ou – dit de la manière inverse – la présence du plus grand écart réside entre les thémanymes issus de la question et ceux issus des instruments de recherche. L'écart sémantique le plus important dans nos données prend bel et bien place entre le vocabulaire des usagers et celui dans les instruments de recherche.

Dans les deux dernières figures, nous voyons une disparité importante entre le corpus 1 et le corpus 2, à propos de la relation d'identité lors de la comparaison entre la question et la notice descriptive (Q N); nous avons alors poussé l'analyse plus loin (voir Figure 27). Les disparités sont effectivement assez prononcées entre les établissements. Ainsi, les établissements 1 et 2 montrent un grand nombre de relations d'identité entre la question de l'usager et la réponse de l'archiviste; cette représentation est la plus forte parmi toutes les sources. L'établissement 3 montre une faible présence de la relation d'identité entre la question d'usager et la réponse de l'archiviste, mais par contre obtient la plus forte présence de la relation d'identité entre la question d'usager et la notice descriptive.

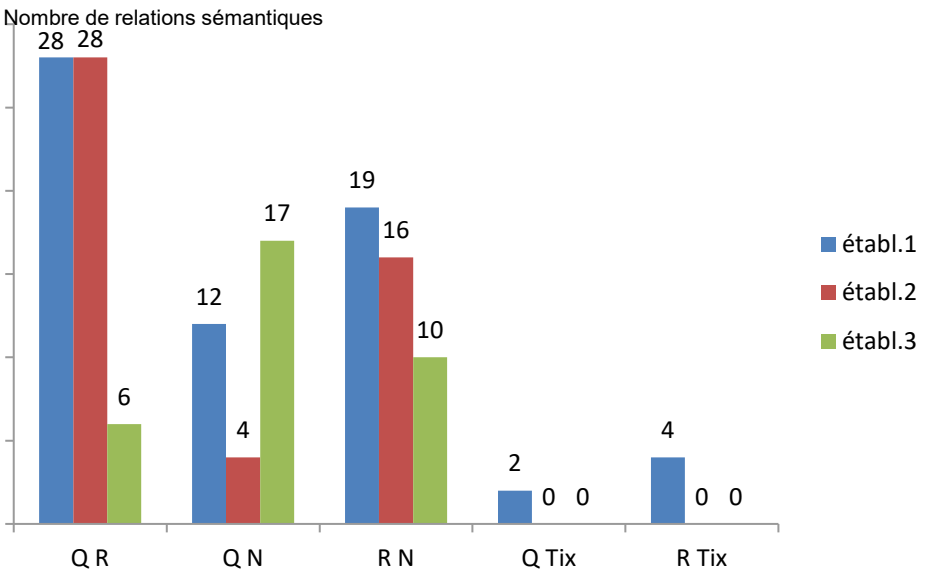

Figure 27. Proportion de la relation d'identité selon les paires de thémanymes dans les sources (corpus général, n=1210)

Il est à noter que, malgré tout, des termes d'indexation (pour l'établissement 1 seulement) et des termes de notices descriptives sont identiques à ceux de la question. Puisque les catalogues des trois établissements consultés présentent des notices descriptives et des termes d'indexation (pour deux des trois établissements) qui sont librement accessibles en ligne, nous nous demandons pourquoi l'usager s'est tourné vers un archiviste plutôt que faire lui-même sa recherche à l'aide des instruments de recherche en ligne. Nous nous demandons si sa recherche préalable a été infructueuse et alors quels mots il a inclus dans sa requête lorsqu'il a interrogé la base de données en ligne. Peut-être les résultats pertinents à sa recherche étaient-ils noyés dans des résultats qui lui convenaient moins (bruit à la recherche) ? Ou encore, l'usager a-t-il préféré passer directement par les services d'un archiviste ? Ces éléments n'entraient pas dans notre recherche et pourraient faire l'objet d'une recherche future.

La relation d'identité entre deux thémanymes peut « correspondre » à la recherche en texte intégral par un usager dans les instruments de recherche en ligne, c'est-à-dire une recherche caractère par caractère dans des ensembles de caractères qui forment le texte. Ainsi, un usager obtiendrait un résultat de 2,9% de correspondance (voir Tableau LIV sur La relation d'identité par sources comparées (corpus général) Q Instr) s'il avait à sa disposition les notices

descriptives et éventuellement les termes d'indexation. Sans l'intervention de l'archiviste de référence, l'usager fait donc face à un fossé sémantique de 97,1%, entre son vocabulaire et celui employé dans les instruments de recherche classiques. Les outils informatiques étant plus ou moins sensibles à la casse ou incorporant des dictionnaires d'équivalences pour les sigles et acronymes ou les synonymes par exemple, nous verrons que ce chiffre peut être cependant nuancé par l'étude plus en détail des relations sémantiques que nous avons codées en fonction de notre échelle d'écart sémantique.

6.2. Les relations qui caractérisent un écart sémantique

Dans un premier temps, nous rappelons que la distribution des relations sémantiques identifiées à partir de l'échelle d'écart sémantique que nous avons développée (voir section 4.3.1, Échelle d'écart sémantique) est particulièrement inégale. Dans la Figure 28 (basée sur le Tableau LXII), le vocabulaire des usagers est comparé avec celui employé dans les instruments de recherche (Q Instr) – sans médiation de l'archiviste de référence – et chacun de ces deux vocabulaires avec le vocabulaire de l'archiviste de référence (Q R et R Instr). Elle offre un panorama des résultats et fait ressortir les relations sémantiques les plus présentes dans le corpus.

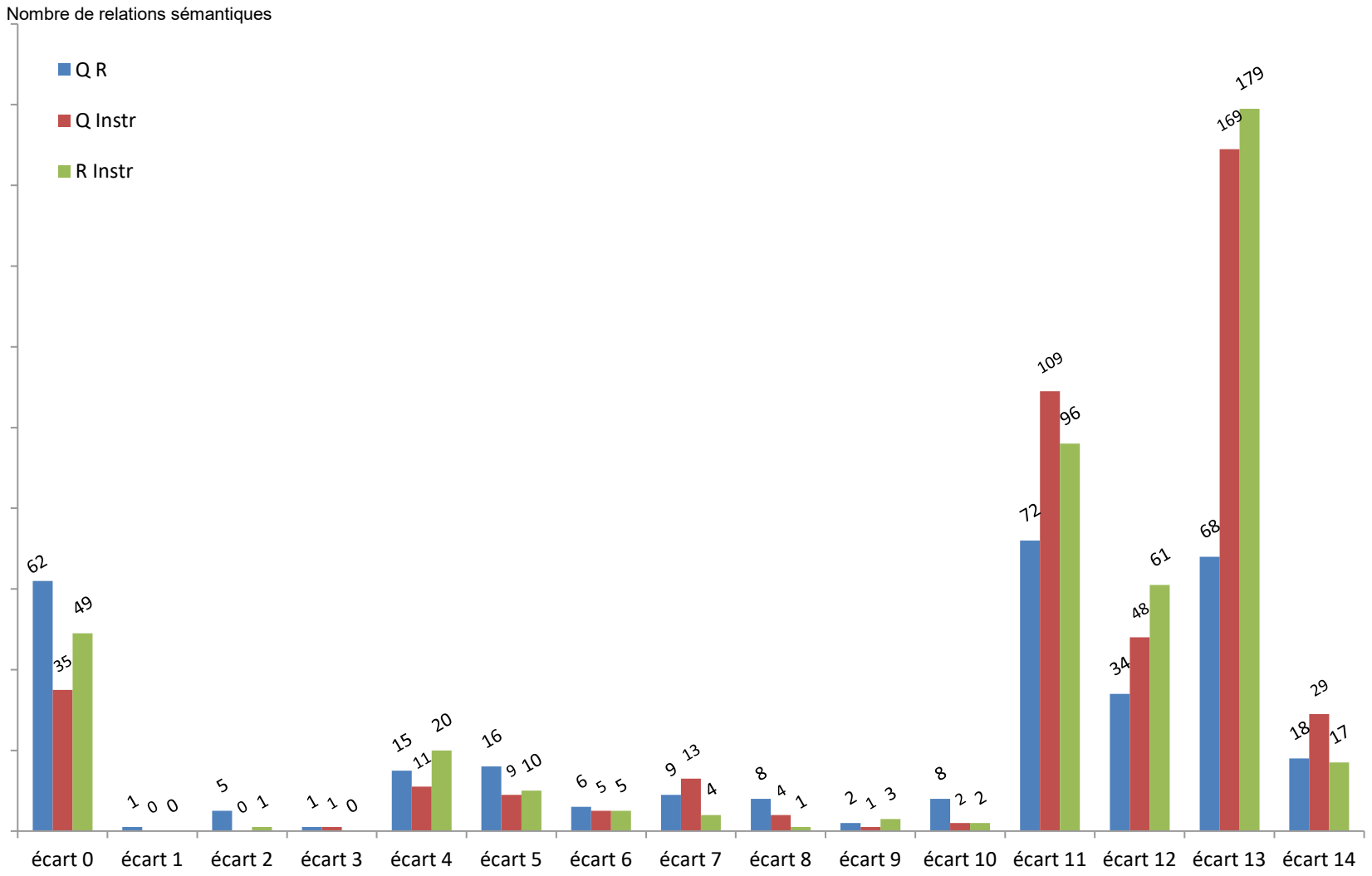

Figure 28. Panorama des relations sémantiques selon les sources comparées (corpus général, n=1210)

Cette figure montre ainsi que les liens sémantiques principaux identifiés au cours de notre analyse sont d'une part la relation d'identité (écart 0) - traitée dans la section précédente - et d'autre part les relations à fort écart, *i. e*. des relations hiérarchiques (écart 11), associatives (écart 12), relatives à un champ sémantique (écart 13) ou encore des relations référentielles (écart 14).

L'échelle d'écart sémantique a permis de catégoriser assez finement la relation en jeu entre deux thémanymes. La Figure 28 montre l'étendue de cette catégorisation et fait ressortir l'absence ou la faible représentation de certains échelons de l'échelle. À partir de l'observation des résultats, il nous est ainsi apparu intéressant d'établir des regroupements effectués selon divers modes, présentés au chapitre 5 (voir section 5.4). Une analyse de ces regroupements fait l'objet des trois prochaines sections : le regroupement multiple (section 6.2.1), le regroupement d'identité à équivalence (section 6.2.2) et le regroupement de mise en valeur des forts écarts (section 6.2.3).

6.2.1. Regroupement multiple : relation d'identité élargie (écarts 0-5) et relations thésaurales (écarts 10-12)

Les regroupements d'échelons nous permettent d'envisager les données sous divers angles. Ainsi, le regroupement multiple rassemble d'abord les liens sémantiques de faible écart de forme (écarts 1 à 5) et la relation d'identité (écart 0), tel que le montre la Figure 29 (basée sur le Tableau LXVII sur le Regroupement multiple (0-5, 6-9, 10-12, 13-14) (corpus général) Q Instr). Les résultats des liens sémantiques d'écart faible ont été regroupés à ceux de la relation d'identité parce qu'ils semblent avoir un impact faible sur la compréhension du thémanyme et la recherche d'information; ils forment ce que nous appelons la relation d'identité élargie.

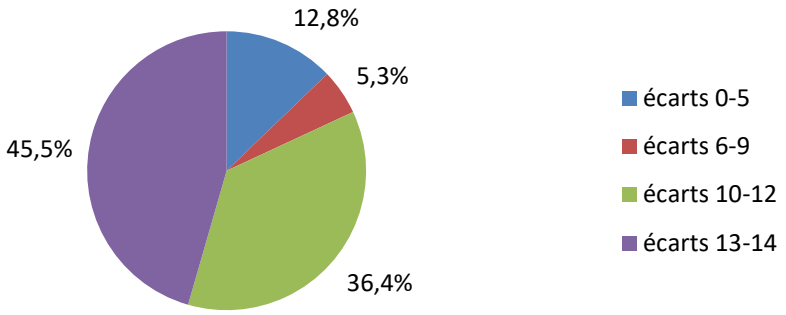

Figure 29. Répartition de la relation d'identité élargie (écarts 0-5) dans Q Instr

Si nous considérons la relation d'identité élargie (écarts 0 à 5), l'écart entre le vocabulaire des usagers et le vocabulaire dans les instruments de recherche (écarts 6 à 14) reste important puisqu'il couvre 87,2% des relations en jeu dans le corpus général.

Le regroupement multiple permet également d'observer la part des liens pris en charge habituellement par un thésaurus (écarts 10, 11, 12). Les relations thésaurales sont connues des professionnels de l'information et leur traitement est encadré par des normes (notamment ISO 5963 :1985 *Documentation – Méthodes pour l'analyse des documents, la détermination de leur contenu et la sélection des termes d'indexation*, NF Z 47-102 :1993 *Principes généraux pour l'indexation des documents*) et manuels (p. ex. Hudon (2009)). Elles présentent des défis pour l'indexation des archives mais des centres ou services d'archives les utilisent (voir sections 3.3.2 et 3.4.3.2). À partir de la Figure 29, nous observons que les liens qui ne sont ni des liens de la relation d'identité élargie (écarts 0-5 : 12,8%), ni des relations thésaurales (écarts 10-12 : 36,4%) constituent la moitié des relations sémantiques de notre corpus général (écarts 6-9 et écarts 13-14 : 50,8%). Ainsi, la moitié des relations sémantiques entre le vocabulaire de l'usager et celui employé dans les instruments de recherche est constituée de relations indéfinies, qui ne sont habituellement pas identifiées dans les outils documentaires. L'échelle d'écart sémantique a permis cependant de les identifier et de les catégoriser.

6.2.2. Regroupement d'identité à équivalence (écarts 0-10)

Nous avons vu dans la Figure 29 que les liens sémantiques relatifs aux écarts 6 à 9 représentent 5,3% des liens sémantiques entre le vocabulaire des questions des usagers et le vocabulaire des instruments de recherche (Q Instr seulement). Les liens sémantiques relatifs aux écarts 6 à 9 représentent 5% des relations du corpus général (voir Tableau LXVII). Ces liens sémantiques impliquent un changement de forme mais très peu de variation sémantique, si bien que nous les avons regroupés avec le premier regroupement relatif à l'identité. La relation thésaurale de synonymie (écart 10) a également été ajoutée à ce regroupement pour rendre compte de cette notion d'équivalence entre les deux thémanymes. La part de cette relation d'identité à équivalence est représentée par la Figure 30 (basée sur le Tableau LXXI Écart de forme et écart de sens (corpus général) Q Instr).

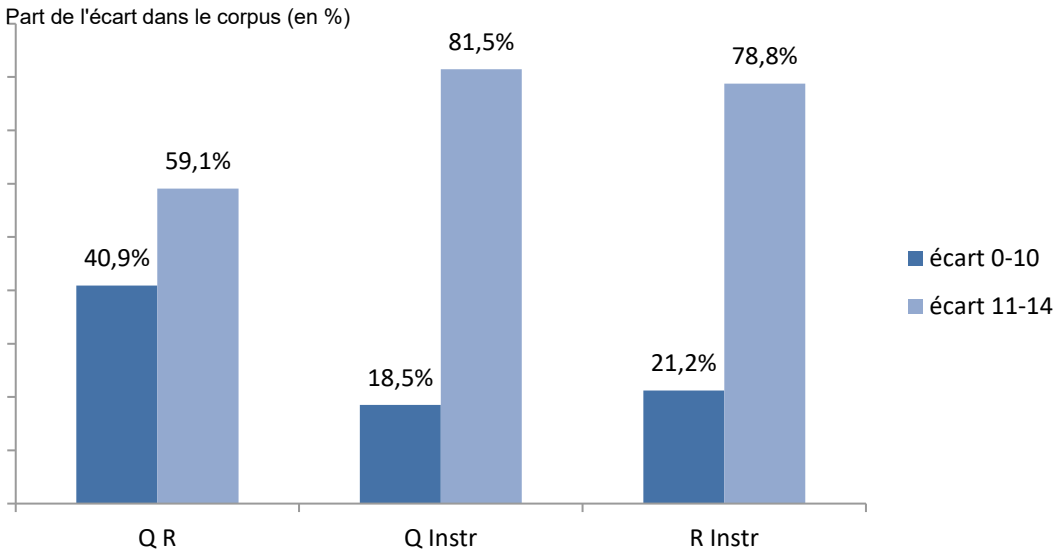

Figure 30. Répartition de la relation d'identité-équivalence (écarts 0-10) entre les sources (corpus général)

L'écart sémantique qui réside entre le vocabulaire des questions des usagers et le vocabulaire employé dans les instruments de recherche, c'est-à-dire celui qui contient le plus de relations de fort écart (écarts 11 à 14), couvre les trois quarts des liens sémantiques identifiés (voir Figure 22 Regroupement des relations d'identité à équivalence (écarts 0-10)). Remarquons que le vocabulaire des réponses des archivistes et le vocabulaire employé dans

les instruments de recherche entretiennent un écart similaire. Les instruments de recherche entretiennent donc un écart sémantique particulièrement élevé avec non seulement les questions des usagers mais aussi les réponses des archivistes de référence. Ce point soulève des interrogations. En effet, nous nous demandons dès lors comment le travail de médiation sémantique est opéré par l'archiviste de référence et sur quels éléments, non linguistiques, il repose. Les éléments identifiés par les archivistes eux-mêmes et inclus dans le modèle ARK (Duff, Yakel et Tibbo 2013, voir section 3.2.1) tels que la connaissance des fonds et des connaissances générales, pourraient entrer dans la réponse. Ce questionnement pourrait là encore faire l'objet de recherches ultérieures.

La proportion des relations d'identité à équivalence entre les thémanymes invite à souligner le bénéfice d'une recherche en texte intégral dans les notices et les termes d'indexation. Les petites variations formelles induisant de très légères, voire imperceptibles nuances de sens pourraient n'être prises en compte dans la recherche en texte intégral que si l'usager le demande. Cela favoriserait le taux de rappel sur le taux de précision lors d'une recherche. Mais l'usager des archives a besoin d'aide pour se retrouver parmi la « myriade des outils » dans chaque établissement (Oliver, Jamieson et Daniel 2017, 3). En outre, les instruments de recherche ne sont pas complets ni complètement en ligne, ce qui empêche les usagers d'effectuer une recherche en texte intégral. « The lack of a good and comprehensive finding aid system greatly hampers reference archivists and at times makes it impossible for them to do their jobs. » (Duff et Johnson 2002, 494)

6.2.3. Regroupement de mise en valeur des forts écarts (écarts 10 à 14)

Le troisième regroupement met en valeur les relations sémantiques les plus éloignées dans la grille d'écart sémantique (écarts 10 à 14). La Figure 31 (basée sur le Tableau LXXV qui porte sur le Regroupement sur les forts écarts (0-9, 10 à 14) (corpus général) Q Instr) montre la forte présence de la relation qui relève du champ sémantique (écart 13) par rapport aux relations d'identité très élargie (écarts 0-9), les relations thésaurales (écarts 10, 11 et 12) ou référentielles (écart 14).

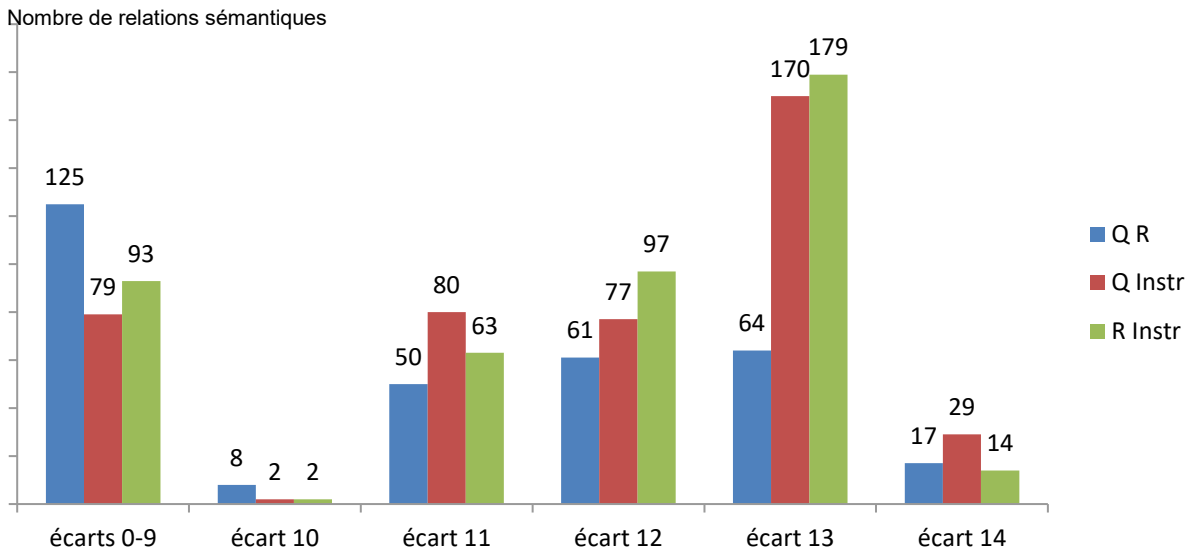

Figure 31. Répartition des relations à fort écart selon les sources (corpus général, n=1210) Q Instr

Les relations hiérarchiques (écart 11) et associatives (écart 12), prises en charge par les thésaurus ou ressources terminologiques utilisées de manière classique, sont bien représentées dans le corpus. Elles représentent respectivement 16% et 19,4% des liens sémantiques identifiés dans le corpus général (voir Tableau LXXV Regroupement sur les forts écarts (0-9, 10 à 14) (corpus général) Q Instr). Environ un quart des relations en jeu lors de la référence pourraient alors bénéficier d'une prise en charge thésaurale. Cependant, il est à noter que si la relation est de nature thésaurale, les thémanymes, eux, ne sont pas des termes biunivoques, caractéristique des termes normalement de mise dans un thésaurus (Hudon 2009). En outre, nous avons établi les liens sémantiques entre deux thémanymes dans le sens de la chaîne communicationnelle de la référence archivistique; or, dans un thésaurus, les relations sémantiques sont réciproques. Notre remarque est dès lors à prendre avec circonspection.

Un champ sémantique est, dans notre recherche, un regroupement de thémanymes ayant en commun dans leur définition un élément de sens. La relation qui relève d'un champ sémantique est l'assurance qu'un élément de sens est repris entre deux thémanymes. À elle seule, la relation qui relève du champ sémantique (écart 13) couvre 38,9% des relations en jeu entre le vocabulaire des usagers et celui employé dans les instruments de recherche (Q Instr) et 39,9% des relations entre le vocabulaire de l'archiviste de référence et celui dans les instruments de recherche (R Instr). Avec 34,1% de tous les liens sémantiques en jeu dans le

corpus général, elle est la relation sémantique la plus présente dans notre corpus. Présente surtout entre les termes de recherche (thémanymes de Q et de R) et les termes de description et d'indexation (thémanymes de N et Tix, *i. e.* Instr), nous pensons que sa forte présence est due au fait que les sujets formulés par les usagers et repris par les archivistes ne sont que rarement précisément traités dans des documents d'archives et leurs représentations. Nous hasardons une hypothèse : la probabilité est faible que des documents d'archives – uniques et constitués en tant que traces organiques des fonctions et activités d'une personne morale ou physique ou d'une famille – portent exactement sur les éléments d'une question d'usager de nature thématique exprimée par des noms communs. Le fait que les documents d'archives n'aient pas été créés de manière intentionnelle pour porter une information thématique pourrait également peut-être jouer sur la formulation du sujet par l'archiviste ou l'usager – notamment quant au choix lexical – ou sur la granularité de l'information[75].

La relation qui relève du champ sémantique est absente des outils classiques de formalisation d'un langage documentaire, tels que les thésaurus. Elle est à ce jour peu reconnue. Nous n'avons trouvé pour l'instant qu'une seule ressource informatique en français qui propose les mots du même champ sémantique que le mot recherché. Il s'agit du logiciel *Antidote*, que nous avons utilisé dans notre recherche comme outil de référence pour cette relation.

Basée sur le Tableau L, la Figure 32 représente le nombre de liens sémantiques entre les questions des usagers et les réponses des archivistes (Q R), les questions des usagers et les instruments de recherche (Q Instr) et les réponses des archivistes et les instruments de recherche (R Instr). Nous observons que les liens sémantiques entre les questions et les réponses reflètent à la fois de faibles écarts (écarts 0-5) et de forts écarts (écarts 10-14). Les

[75] En effet, le créateur des documents d'archives –notamment s'il s'agit d'une personne morale ou d'une famille de plusieurs générations – n'a pas conscience au moment de la création du tout que forme(ra) l'ensemble des documents conservés (après évaluation) pour constituer le fonds d'archives patrimoniales accessible à l'usager.

résultats révèlent aussi que les questions comme les réponses entretiennent des relations d'écarts élevés (10 à 14) avec les instruments de recherche.

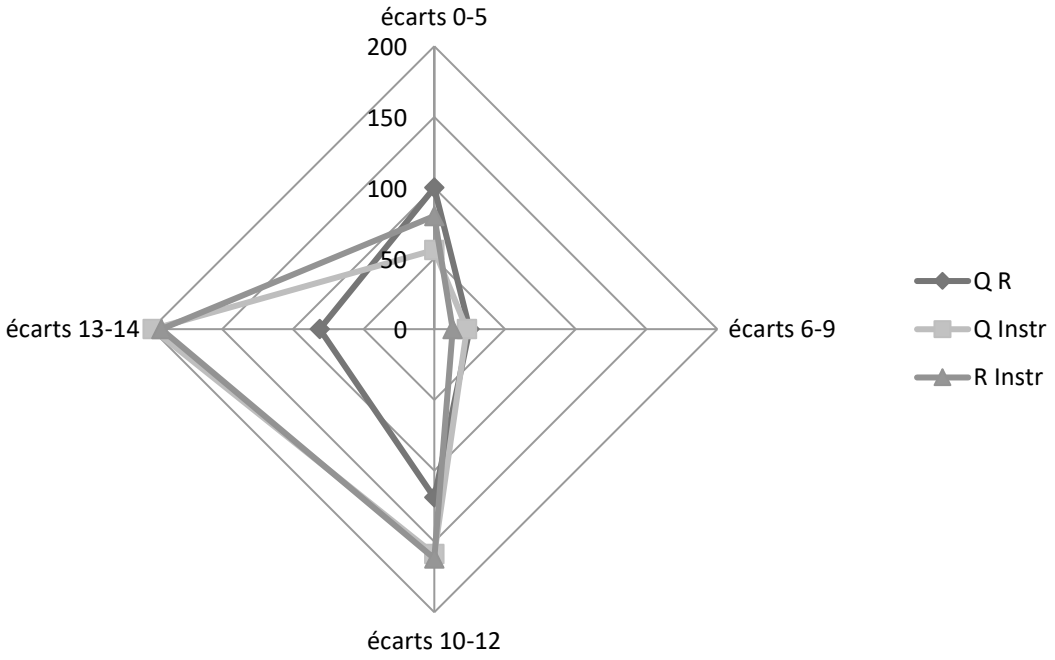

Figure 32. Nombre de relations entre Q R, R Instr et Q Instr (corpus général, n=1210)

Le grand nombre de relations à forts écarts (écarts 13 et 14) ainsi que les liens sémantiques complexes (liens sémantiques qui sont constitués par plusieurs échelons de l'échelle d'écart sémantique) dans le corpus général pourrait fournir un élément d'explication (ou une piste à creuser) au faible succès dans les archives des vocabulaires formalisés et autres outils lexicaux hiérarchisés tels que les thésaurus (voir section 3.4.3.2).

Les relations relatives au champ sémantique (écart 13) et les relations référentielles (écart 14) ne sont généralement pas prises en compte par les ressources terminologiques ou thésaurales habituelles. Pourtant elles représentent environ 40% des liens sémantiques identifiés dans nos données (41% dans le corpus 1 et 38% dans le corpus 2) et ne sont donc pas prises en compte par les moteurs de recherche basés sur les relations thésaurales.

C'est dans le rapport Q Instr que se situe le point le plus profond du fossé entre le vocabulaire des usagers et le vocabulaire employé dans les instruments de recherche. Seul (pour l'instant) l'archiviste peut franchir ce fossé et seul lui sait apparier les termes de la question qu'il reformule avec des notices. Bien qu'il explique ses stratégies dans sa réponse

par courriel, il ne donne pas tout le cheminement de son raisonnement lexical et conceptuel. En témoignent les résultats également élevés entre le vocabulaire de l'archiviste de référence et celui employé dans les instruments de recherche. Nous constatons que la « mécanique » de la référence n'est pas encore mise au jour et explicite. Une analyse recourant au protocole verbal serait donc intéressante pour ce faire. Duff et Fox (2006) se sont intéressées au processus de la référence archivistique du point de vue de l'archiviste. Elles concluent que « [t]he data (…) supports Yakel and Ruth's suggestion that finding aids play a vital role as reference tools ». (Duff et Fox 2006, 149) D'après le schéma de la chaîne communicationnelle que nous avons élaboré (voir section 3.2.1), nous soulignons à notre tour le rôle crucial des instruments de recherche dans la recherche de documents d'archives. Leur mise en ligne – même s'ils ne sont que temporaires ou préliminaires – permettrait un premier accès aux usagers qui ne sont pas novices et ont déjà un certain entendement archivistique (*archival intelligence*, *ARK* : Duff, Yakel, Tibbo 2013). Dans l'étude de Duff et Fox, « The participants in the study suggested that onsite reference includes educating users about how to use finding aids and offsite reference predominantly involves finding answers to research questions ». (2006, 150) Notre étude s'est intéressée uniquement aux questions d'usagers posées par courriel, dont les réponses contenaient des cotes archivistiques et non une instruction de contacter en personne ou par téléphone le service d'archives pour des recherches sur place. Nos données montrent que certains archivistes développent leur réponse de manière à apprendre à l'usager comment procéder à une recherche dans leur base de données. Mais quand les instruments de recherche ne sont pas en ligne, quand ils sont internes, sur papier, les archivistes ne peuvent pas les suggérer à distance aux usagers ni proposer à ces derniers de venir et de les former à leur utilisation sur place. Oliver, Jamieson et Daniel (2017, 1-2) soulignent le peu d'études sur la référence du point de vue de l'archiviste et la pertinence de l'approfondissement de ce champ d'études pour rendre compte de l'évolution des manières de faire dues au contexte technologique. Les retombées de l'explicitation du processus de référence archivistique seraient nombreuses; nous pensons d'abord à la formation des usagers comme l'ont rappelé Duff et Fox (2006) ainsi que Yakel et Torres (2003), la formation des nouveaux archivistes comme l'ont rappelé Oliver, Jamieson et Daniel (2017) et le contrôle de la qualité en vue d'améliorer l'efficacité du processus de la référence.

L'écart 14 contient des variations davantage référentielles que sémantiques; ces variations s'ancrent dans la connaissance et la compréhension du contexte de communication, tel que présenté dans le schéma de Jakobson (voir section 3.2.1). Par exemple, dans une notice descriptive de notre corpus (notice d145), le fossé thématique entre Q R et N Tix était flagrant : il n'existait aucun lien identifié d'après notre grille d'analyse entre les termes de recherche et les termes des instruments de recherche. La question et la réponse traitaient du poste d'inspecteur d'anatomie et de morgues. Mais la notice descriptive et les termes d'indexation traitent quant à eux de paroisses et de constructions ou réparations d'édifices. Pourtant, il ne s'agit pas d'une erreur de l'archiviste d'avoir suggéré cette unité archivistique à l'usager. Dans ces données, le lien n'est pas explicité par l'archiviste, mais il est *a priori* compris par un humain (par opposition à une machine). Il semble que le lien implicite s'effectue entre *morgue* et *paroisse.* La paroisse gérait alors ce qui entourait les corps, les sépultures et donc les registres les recensant pourraient contenir quelque information sur ce sujet. Le lien n'est pas sémantique, mais référentiel : les connaissances encyclopédiques sont nécessaires pour l'établir et peut-être aussi pour le comprendre. Cependant, la thématique des termes de recherche étant tellement éloignée de celle des termes de description et d'indexation, nous n'avons pu cibler les thémanymes à rapprocher à partir des outils d'analyse définis dans cette recherche. Le décalage du plan sémantique au référentiel pourrait expliquer pourquoi le recours à des ressources de nature purement sémantique n'apporte qu'une modeste contribution à l'accès aux ressources archivistiques. Ici encore, le lien qui a été établi par l'archiviste relève de la connaissance des collections et d'un savoir général, compétences identifiées par le modèle *ARK* (Duff, Yakel et Tibbo 2013).

La relation référentielle couvrant 5% des liens sémantiques identifiés dans le corpus général, elle y est la cinquième relation la plus représentée : 1. Le champ sémantique (34,1%), 2. La relation associative (19,4%), 3. La relation hiérarchique (16,0%), 4. La relation d'identité (12,1%), 5. La relation référentielle (5,0%). Nous présentons ci-dessous les échelons de l'échelle d'écart sémantique en fonction de leur fréquence dans nos données (voir Tableau LXXVII).

Tableau LXXVII. Représentation des écarts de l'échelle d'écart sémantique (corpus général, n=1210)

rang	écart	type	valeur (%)
1	13	Champ sémantique	34,1
2	12	Relations associatives	19,4
3	11	Hiérarchie	16
4	0	Relation d'identité	12,1
5	14	Relation référentielle	5
6	4	Variation de casse	3,8
7	5	Variation flexionnelle	2,9
8	7	Variation syntaxique	2,1
9	6	Variation dérivationnelle	1,3
10	8	Ellipse d'un complément	1,1
11	10	Équivalence	1
12	2	Variation de longueur	0,5
13	9	Paraphrase	0,5
14	3	Variation graphique	0,2
15	1	Variation orthographique	0,1

La représentation de ces chiffres sous forme de figure rend plus évidente encore la part des cinq premières relations (voir Figure 33).

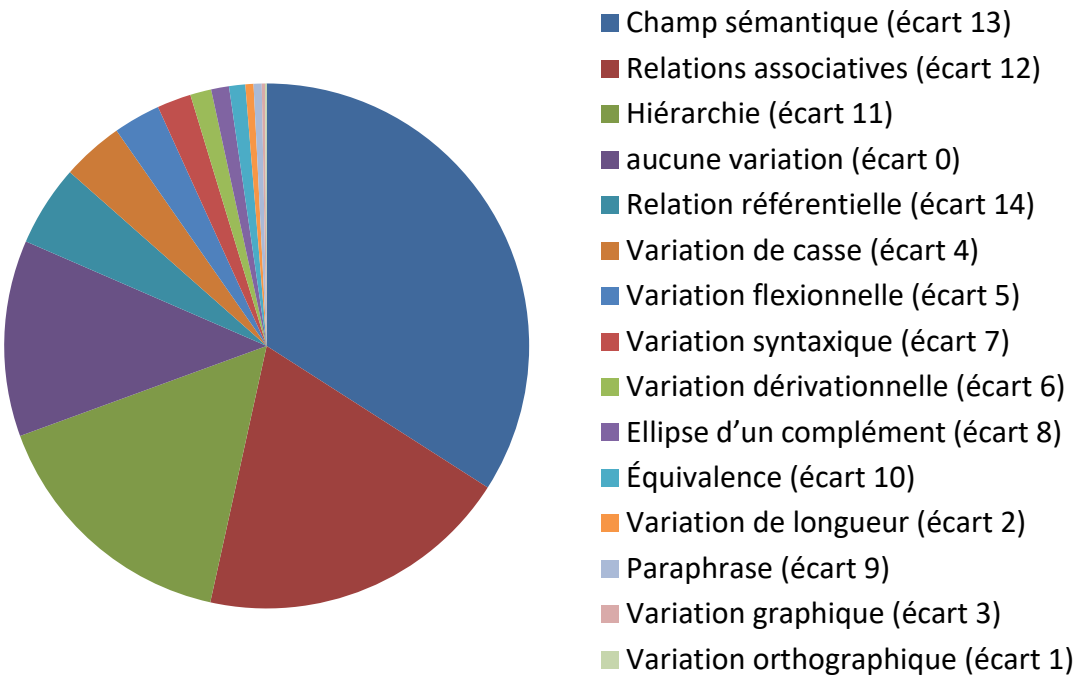

Figure 33. Part des relations sémantiques (corpus général, n=1210)

Selon nous, la relation référentielle mériterait d'être prise en compte et distinguée des relations purement sémantiques afin d'améliorer les outils de recherche. Tout comme la relation qui relève du champ sémantique, la relation référentielle fait appel à l'interprétation humaine. De même, les relations thésaurales, à défaut d'outil déjà constitué et utilisable dans les archives, sont attribuées par un être humain. Nos données permettent donc de souligner le rôle essentiel de l'archiviste de référence dans le contexte archivistique actuel et de justifier les fonds alloués à la référence : la rémunération du personnel de référence, sa formation, les moyens technologiques pour assurer la référence quel que soit le mode de communication de l'usager, pour ne citer que quelques-uns des éléments du modèle *ARK* de Duff, Yakel et Tibbo (2013).

De nos jours, on appelle beaucoup de choses *sémantiques*, notamment en informatique. À cause de ces nouvelles acceptions, on perd parfois de vue le sens du mot *sémantique* en linguistique qui est « relatif au sens », relatif au *signifié* (d'après Saussure 1960). Les sciences de l'information nous semblent particulièrement sujettes à la méprise parce qu'elles s'intéressent à la fois aux systèmes informatiques et aux systèmes linguistiques dans le traitement et la diffusion de l'information, consignée ou non. Dans notre étude, des phénomènes certes linguistiques, mais pas nécessairement sémantiques, ont été mis au jour. Ces relations de correspondance référentielle, loin d'être minoritaires, représentent un pourcentage non négligeable, soit 5% des relations identifiées. Ce pourcentage est plus élevé que les relations d'écart faible à modéré : écarts 1 à 10 pris chacun individuellement (écart 1 à 0,1% et écart 4 à 3,8%). C'est pourquoi nous invitons les concepteurs de langages documentaires et toute personne intéressée par les langages de description ou d'indexation et les relations sémantiques à distinguer le plan sémantique du référentiel.

Certaines relations sémantiques telles que la hiérarchie ou la méronymie sont valides en dehors d'un usage dans un discours particulier; elles sont donc valides du point de vue des archivistes qui préparent les instruments de recherche ou font de la référence, comme de celui des usagers, quels qu'ils soient et quels que soient leurs besoins. Elles sont valides tout au long de la chaîne communicationnelle des documents d'archives aux usagers (voir section 3.2.1), à partir de la conception des instruments de recherche. Par exemple, entre *santé* et *santé mentale* ou *santé publique*, le lien d'hyperonymie/hyponymie est *a priori* stable, quelle que soit la situation de communication. De même, *réparations* et *agrandissement* en parlant de bâtiments

sont, de manière *a priori* stable, deux co-hyponymes de l'hyperonyme de *travaux*. Finalement, « le gouvernement fédéral ou provincial » et « le premier ministre du Québec » entretiennent une relation d'association précisée *entité/fonction* : à partir de l'un des deux thémanymes et de la relation précise, nous pourrions suggérer à un usager l'autre thémanyme. L'identification précise des relations sémantiques pourrait ainsi servir l'aide à la recherche. D'autres relations semblent rattachées à un point de vue ou du moins à un contexte particulier d'emploi entre deux termes (ici, deux thémanymes) et ne sont valides que dans un tel contexte. Par exemple, « des premiers ministres » et « sa carrière politique » entretiennent une relation de méronymie (partie/tout); le premier thémanyme est un des degrés d'une carrière politique. Cependant, à l'inverse, il serait difficile de retrouver à partir du second thémanyme et de la relation précise qu'elle serait la partie du tout qu'est la carrière; les degrés d'une carrière sont très variables d'une profession à l'autre. Pour être située, la relation a besoin d'un contexte. Si nous la situons dans le champ lexical de la politique, alors, déjà, il semblerait que le premier thémanyme vienne plus facilement à l'esprit. Cependant, ceci est une hypothèse, il faudrait tester ces relations auprès d'usagers pour vérifier si elles peuvent entrer dans la suggestion de raffinement d'une recherche.

Le fait de connaître le contexte pour pouvoir identifier la relation a déjà été repéré par Svenonius (1997) et Clarke (2001) à propos des relations hiérarchiques (voir section 3.5.3). Nous avons pu constater dans notre recherche l'existence de ces relations stables, valides en tout temps ainsi que l'existence de relations contextuelles, identifiables sur le plan référentiel. Sont-elles particulièrement présentes dans un contexte de référence archivistique ? La nature même des documents d'archives jouerait-elle alors un rôle dans la nature des sujets exprimés et ainsi dans les relations sémantiques entretenues entre les divers éléments de la chaîne communicationnelle de l'usager vers les documents d'archives ?

Nous avons étudié les relations sémantiques entre les quatre sources (Q, R, N, Tix) mais non à l'intérieur de celles-ci. L'étude des relations sémantiques au sein de la réponse de l'archiviste permettrait de comprendre le raisonnement qu'il opère pour passer du vocabulaire de recherche au vocabulaire de description et d'indexation. Ainsi, nous avons pu déjà repérer certains mécanismes. L'un de ces mécanismes est la « 'traduction' de requêtes formulées sur

des sujets en des stratégies de recherche basées sur le contexte (ou la provenance) » (McCausland, 2011, 312 et Pugh 2005).

Bien que les questions retenues dans notre recherche aient nécessité l'explicitation du sujet par des noms communs, il est arrivé que l'archiviste de référence ait recours à des noms propres dans sa recherche décrite dans sa réponse. Par exemple, l'archiviste indique qu'il doit effectuer le repérage des noms des ministres et premiers ministres et donne à l'usager plusieurs fonds à consulter : Duncan Cameron, Mackenzie King, Brian Mulroney, Jeanne Sauvé, Flora MacDonald. L'archiviste présume ici que l'usager connaît le nom de ces personnes et leur fonction. Dans la même réponse, l'archiviste inclut la collection de Ted Grant. Le lien entre le photographe officiel – la fonction de cette personne est inscrite dans la notice descriptive de la collection – et la recherche de photographies d'assermentations de ministres (sujet de la question d'usager) est implicite, mais peut-être n'est-ce pas obscur pour l'usager renseigné sur le sujet.

6.3. Mise en perspective de la recherche

Cette section porte sur une certaine mise en perspective de notre recherche, par rapport à la reconnaissance d'une question d'usager (section 6.3.1), à la notion de question par sujet en archivistique (section 6.3.2), des considérations générales sur les thémanymes (section 6.3.3) et sur les relations sémantiques (section 6.3.4), à des remarques méthodologiques (section 6.3.5) et finalement des remarques générales sur les résultats (section 6.3.6).

6.3.1. Reconnaissance d'une question d'usager à teneur thématique

Il est difficile de définir clairement les critères qui permettent de reconnaître une question d'usager à teneur thématique exprimée par des noms communs. La notion de question d'usager à teneur thématique n'est pas évidente pour les archivistes, peut-être en raison d'un manque de définition claire dans la théorie et d'une tradition onomastique ancrée dans la pratique. Pourtant ce type de question existe. Nous avons développé un arbre décisionnel pour systématiser la discrimination de ce type de questions par rapport aux autres. Les recherches pour répondre aux questions d'usagers à teneur thématique demandent beaucoup de temps aux archivistes. Les réponses à ces demandes sont généralement longues et

fouillées, explorant diverses stratégies de recherche. Le courriel de réponse ressemble souvent à un compte rendu des recherches menées par l'archiviste; cela assure la crédibilité de sa réponse et en même temps forme l'usager à la recherche. Ce dernier point rejoint les compétences de pédagogie d'un archiviste de référence selon le modèle ARK (Duff, Yakel et Tibbo 2013). Certains liens tels que les relations référentielles sont établis par l'archiviste et restent inaccessibles aux usagers car ils ne connaissent pas les fonds d'archives conservés par tel établissement. L'étude des relations sémantiques (voir section 6.2) montre que l'archiviste procède parfois intuitivement à partir de ses connaissances encyclopédiques propres et de sa connaissance des fonds et collections sous sa garde. Ainsi, il ne peut pas être demandé à un usager de répondre sans assistance à ce genre de questions. L'autonomie des usagers que les archivistes escomptent comporte des limites dues notamment à la nature des questions (Oliver, Jamieson, Daniel 2017).

6.3.2. Proportion des questions thématiques par rapport aux autres types de questions d'usagers

Nous cernons ici le nombre (et pourcentage) de questions thématiques identifiées par rapport au nombre de questions examinées. Il s'agit de savoir s'il réside des différences entre les établissements. Notre recherche présente un nombre total de questions par sujet particulièrement petit. Tel qu'indiqué aux Tableau VI et Tableau VII, les 125 questions thématiques dont le sujet est exprimé par des noms communs représentent 3,4% des 3729 courriels examinés, ce qui constitue une proportion relativement faible. La proportion est plus faible que les proportions présentées dans l'étude de Gagnon-Arguin (1998), soit 40% environ. Cependant dans cette dernière étude, les questions par sujet recouvraient également les noms propres. Et comme le spécifie bien l'auteure, ce nombre portait sur un seul service d'archives. La disparité entre les établissements participant à notre étude est grande, allant de presque 2% à 12%, le taux entre les établissements est donc extrêmement variable. En outre, la sélection des questions thématiques est plus stricte de notre côté en raison de plusieurs autres critères tels que l'expression du sujet par un nom commun et l'absence de documents précis identifiés. La notion de « sujet » en archivistique est un concept aux contours flous et peut faire l'objet de

multiples interprétations, comme nous l'avons déjà indiqué en revue de littérature (voir section 3.3.1).

6.3.3. Considérations sur la quantité et la nature des thémanymes

Nous voyons dans le Tableau XLVIII sur les Mode, moyenne et médiane de thémanymes par source (corpus 1) que le nombre de thémanymes dans les notices est plus élevé que celui des autres sources, allant jusqu'à 369 thémanymes dans une notice. Ceci n'est pas étonnant, étant donné qu'une notice peut porter sur une unité archivistique de haut niveau et décrire ainsi un grand nombre de sujets variés. Par ailleurs, le nombre médian donne une meilleure idée de la proportion de thémanymes dans notre corpus et probablement dans des notices descriptives en général. Notre corpus comportait des séquences particulièrement complexes et développées; ainsi, la moyenne (n =35) n'apportait pas une idée juste du nombre de thémanymes, contrairement à la médiane. Mais nous sommes frappée par le petit nombre médian de thémanymes dans les questions et les réponses, respectivement sept et quatre. Le nombre médian de thémanymes dans les notices descriptives est porté à 13 mais il est variable et dépend notamment du niveau de description dans la hiérarchie. Ainsi, au niveau du fonds, il y en aura davantage qu'au niveau des séries puis, le nombre sera un peu plus important au niveau de la pièce. Le nombre de thémanymes issus des termes d'indexation est de trois; ce score était attendu pour des termes d'indexation par sujet. En effet, c'est le propre de l'indexation de représenter le sujet par une ou plusieurs formes synthétiques. Nous constatons que le sujet est véhiculé d'une source à l'autre par seulement quelques thémanymes et qu'il y est concentré. Par contre, les thémanymes peuvent contenir un grand nombre de mots et sont des expressions linguistiques aux contours non prédéfinis.

Les thémanymes que nous avons comparés ne sont pas toujours, ou plutôt sont rarement, des *termes*, au sens terminologique, c'est-à-dire une expression linguistique entretenant une relation biunivoque avec un concept dans un domaine donné. Deux éléments se dégagent de cette définition, la biunivocité et l'appartenance à un domaine. La biunivocité est la « propriété d'un langage documentaire au sein duquel chaque terme représente un seul concept, toujours le même, et chaque concept est représenté par un seul et même terme » (Hudon 2013, 264). Deuxièmement, les termes – en terminologie – appartiennent à un

domaine. « Les termes sont des unités lexicales dont le sens est envisagé par rapport à un domaine de spécialité, c'est-à-dire un domaine de la connaissance humaine, souvent associé à une activité socio-professionnelle » (L'Homme 2004, 22). Les thémanymes analysés dans notre recherche appartiennent le plus souvent à la langue générale. Ceci s'explique par la mission des services d'archives : l'usager pose une question générale à un établissement qui a pour mission de collecter, sauvegarder et diffuser le patrimoine, la mémoire collective d'une communauté de personnes sur une large étendue de territoire (voir section 4.1.1.2). La variation de langue – essentiellement par rapport à l'absence de domaine d'ancrage du vocabulaire – pourrait être une des raisons qui empêchent un langage documentaire – relié à un découpage de la réalité selon un seul point de vue – de répondre de manière satisfaisante aux besoins en recherche d'information archivistique.

La comparaison des thémanymes est une première étape dans l'étude du vocabulaire pour l'accès thématique aux archives patrimoniales (VATAP) (voir Tableau IV). Outre les relations sémantiques qui ont fait l'objet de la présente étude, les propriétés linguistiques des thémanymes seraient à étudier dans le cadre d'une autre recherche. La notion même de thémanyme, développée pour les besoins de cette recherche, mériterait un approfondissement quant à ses assises théoriques (distance avec un terme tel qu'entendu en terminologie, en lexicologie, en étude du discours) et sa compréhension pratique dans divers milieux. Ces autres milieux pourraient être d'autres services d'archives, d'autres milieux où la référence touche un public large, notamment les milieux culturels tels que l'ont fait Angel (2012) pour les étiquettes dans des archives, musées et bibliothèques ou Sattar Chaudry et Jiun (2005) pour les termes d'un thésaurus commun à des archives et à un musée de Singapour.

6.3.4. Répartition des relations sémantiques selon les établissements

D'après le Tableau L, il y a 504 liens sémantiques entre le vocabulaire des usagers et celui dans les instruments de recherche (Q Instr), ce qui représente 41,7% des liens sémantiques dans le corpus général (n=1210). Étant donné qu'il y a un débalancement entre le corpus 1 (155, 25,9%, n=599) et le corpus 2 (349, 57,7%, n=611), nous avons voulu mettre ce chiffre en perspective selon les établissements (voir la Figure 34 et la Figure 35 basées sur Tableau LI sur la Répartition du nombre de liens par établissement (corpus général)).

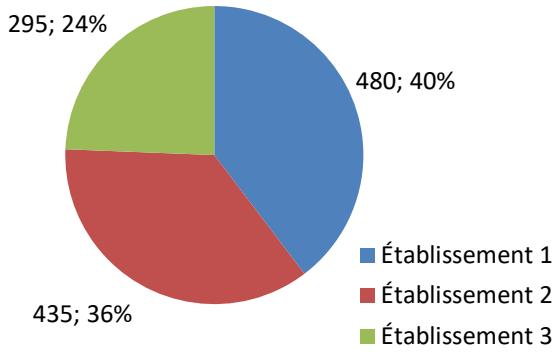

Figure 34. Liens sémantiques (nombre et %) par établissement (corpus général, n=1210)

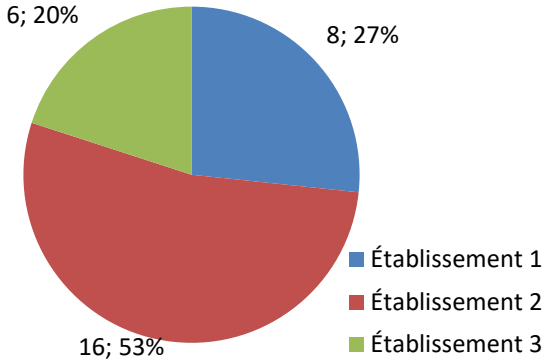

Figure 35. Polyséquences (nombre et %) par établissement (corpus général, n=1210)

Rappelons d'abord qu'une séquence est un ensemble formé par une question, une réponse, une notice descriptive et éventuellement des termes d'indexation rattachés à cette notice (p. ex. QRNTix) et qu'une polyséquence est un ensemble de séquences qui ont la même question et la même réponse de référence (p. ex. QRNTixNTixNTix). Alors que le nombre de liens sémantiques est similaire pour les établissements 1 et 2, le nombre de polyséquences de l'établissement 1 est plus petit de moitié que celui de l'établissement 2. Ainsi, les questions d'usagers à la source des polyséquences étudiées ont généré davantage de liens pour l'établissement 1. Ce résultat est-il lié aux établissements et à leurs habitudes de travail ? L'est-il plutôt lié à la nature des questions reçues et collectées dans le cadre de notre recherche ? Notre étude exploratoire ne nous permet pas de nous prononcer avec confiance sur ce fait. Cependant, le Tableau LI sur la Répartition du nombre de liens par établissement (corpus général) montre une grande variation dans le nombre des liens par polyséquence, quel que soit l'établissement : de 5 à 167 pour l'établissement 1, de 6 à 80 pour l'établissement 2 et de 13 à 106 pour l'établissement 3. Ces résultats semblent indiquer qu'il n'existerait pas une tendance par établissement mais plutôt par question d'usager. Ceci resterait à vérifier par des recherches précises sur le rapport entre la nature des questions d'usagers et le processus de référence du point de vue de l'archiviste.

Parmi tous les liens entre thémanymes, nous nous intéressons particulièrement au rapport entre les thémanymes issus des questions des usagers et ceux présents dans les

instruments de recherche : le rapport Q Instr (Q N et Q Tix dans le corpus 1 et Q N dans le corpus 2). Il compare les ensembles de thémanymes les plus éloignés dans la chaîne communicationnelle des usagers aux documents (voir Figure 3, section 3.2.1). Il est ainsi compréhensible que les instruments de recherche mis en ligne sans aucun outil pour reformuler le vocabulaire de recherche soient insuffisants pour les usagers.

6.3.4.1. La réciprocité des relations sémantiques

La réciprocité d'une relation sémantique est nécessaire dans une structure lexicale, ou plutôt terminologique, telle que les thésaurus. Cette réciprocité permet d'engager un deuxième sens de lecture de la chaîne communicationnelle archivistique : du document (ou de sa représentation) à l'usager. Ce sens de lecture est inverse à celui présenté à la Figure 3, qui va de l'usager au document, tel qu'étudié dans cette recherche.

Les relations sémantiques non réciproques induisent un sens de lecture et offrent une moins forte prédictibilité du lien sémantique dans l'autre sens de lecture. Par exemple, un centre-ville contient nécessairement des rues, le centre-ville est donc le tout dont une partie serait une rue. Pourtant, il existe des rues ailleurs qu'au centre-ville. Ainsi, la relation de méronymie (tout-partie) posée dans un contexte particulier n'est pas toujours réciproque. Les liens référentiels, qui font appel à une connaissance non seulement encyclopédique comme dans le cas des relations de méronymie par exemple, mais aussi à une connaissance contextualisée du réel (*ego, hic, nunc* : liée à une personne, un lieu et un moment) sont les plus nombreux. Certes un lien existe entre les deux thémanymes comparés qui entretiennent une relation référentielle. Mais ce lien est tellement ancré dans le réel, contingent d'un contexte – le contexte des documents d'archives et de la situation de communication de la référence archivistique – que seule une bonne connaissance des documents d'archives conservés par son institution permet à l'archiviste de mettre les thémanymes des instruments de recherche en relation avec ceux de la demande de l'usager. Cette connaissance fait partie des savoirs à acquérir pour un archiviste de référence selon Duff, Yakel et Tibbo (cf. Modèle *ARK*, 2013).

Lors de l'établissement de l'échelle d'écart sémantique, nous avions dans un premier temps séparé la relation classe/instance des relations hiérarchiques pour les inclure dans les relations associatives. En effet, il nous semblait – et il nous semble toujours – que la relation

classe/instance – comme d'autres relations associatives – repose sur des connaissances encyclopédiques et non seulement linguistiques. Puisque cette relation est établie entre une classe et un nom propre (voir notamment Hudon 2013, 59), les connaissances encyclopédiques sont mobilisées pour établir ce lien et attribuer au nom propre des traits qui servent de critères d'inclusion à la classe dénommée par un nom commun. Mais la littérature établissait que cette relation était traditionnellement incluse dans les relations hiérarchiques et nous n'avons pas souhaité nous en démarquer.

6.3.5. Difficultés méthodologiques

Dans cette section, nous rapportons quelques difficultés méthodologiques qui portent sur notre recherche en général.

Nous avions d'abord envisagé l'analyse en facettes comme méthode de caractérisation sémantique. Cependant, après l'avoir appliquée sur le corpus 1, nous avons constaté qu'elle n'était pas suffisante pour nos besoins, pour identifier des thémanymes à comparer entre deux sources. Ainsi, nous n'avons pas procédé à l'analyse en facettes pour le second corpus. Alors, nous avons étendu la notion de thémanyme dans le corpus 2 de « expression linguistique porteuse de sujet » à « portion de texte constituant une unité de sens porteuse de sujet ». En fin de compte, les thémanymes peuvent s'étendre du mot à – dans de rares cas – la proposition comportant par exemple un verbe conjugué avec des compléments et un sujet : « Adolphe V. Roy a été envoyé étudier les mines du Yukon (autre version: de l'Alaska) en 1894 et 1895 » ou « qu'il fut envoyé là-bas par le gouvernement du Québec » (Tableau XXIX Illustration de l'ellipse d'un élément (écart 8)).

Dans notre recherche, nous avons expérimenté à quel point les noms communs et les noms propres étaient porteurs de sujet et qu'il était dommage de perpétuer cette dichotomie dans l'étude des thémanymes. En effet, le lexique forme un tout et les noms propres entretiennent des relations sémantiques diverses avec les noms communs : le lexique est un continuum. Nous les trouvons particulièrement dans la relation classique d'hyperonymie/hyponymie entre un nom commun et un nom propre de même paradigme (Kleiber 1996) en lien avec la relation de classe/instance, p. ex. : *séducteur/don juan*. Dans notre corpus, le thémanyme « Rideau Hall » est un lieu associé thématiquement et

sémantiquement au cabinet ministériel. Nous pensons dorénavant qu'une étude sur le sujet des documents d'archives devrait inclure les noms propres et s'appuyer sur leur pouvoir de connotation dans l'évocation du sujet.

Il est difficile d'attribuer certaines relations au type hiérarchique ou au type associatif. Il existe là une zone grise (Clarke 2001, 46). Clarke remarque même que « [l]a présence d'une relation associative dépend plus du ressenti de l'éditeur du thésaurus sur ce qui sera utile aux usagers que d'une analyse sémantique précise » (2001, 46).

L'échelle d'écart sémantique a été bâtie selon un point de vue linguistique où la syntaxe joue un rôle assez important. Pourtant la variation syntaxique – déjà haute dans l'échelle (7^{e} échelon sur 13) – semble ne pas avoir de conséquence dans l'application de l'échelle dans notre corpus. Les échelons sont basés sur des différences linguistiques entre deux expressions linguistiques comparées sur le plan de la forme puis du sens. Ils ne sont pas équidistants sémantiquement et cela peut affecter l'interprétation des résultats. Les chiffres ne sont pas directement significatifs, c'est seulement leur fréquence que nous avons étudiées. La grille pourrait être améliorée sur ce point, en fonction de la signification des échelons.

6.3.6. Remarques générales sur les résultats

Les remarques qui suivent portent non pas sur notre recherche en général, la relation d'identité ou les relations qui caractérisent un écart sémantique, mais sur des éléments périphériques à notre recherche, à savoir l'aspect diachronique du vocabulaire et polysémie (section 6.3.6.1), les formules introductives du sujet (section 6.3.6.2) et les stratégies de recherche (section 6.3.6.3).

6.3.6.1. Variation de sens d'un mot : aspect diachronique du vocabulaire et polysémie

Contrairement à ce que nous pensions, l'aspect diachronique du vocabulaire si souvent souligné dans la littérature et rappelé en introduction (voir section 1 Contexte et problématique) n'a pas été un grand obstacle. Nous avons repéré un terme d'indexation incluant un sens « Vieux » selon le *Grand Robert* (2017) : *anatomie* dans l'ancien intitulé de poste d'*inspecteur d'anatomie* dont l'équivalent actuel pourrait être *médecin légiste*. Ceci peut s'expliquer en partie en raison de la nature de notre corpus. Nous avons étudié les instruments

de recherche et non les documents d'archives eux-mêmes. La langue est vivante et continue d'évoluer inlassablement. Ainsi, quand un fonds porte sur de longues périodes, sur plusieurs générations, il serait attendu de trouver des variations dans la langue utilisée. Ce cas ne s'est pas présenté car nous avons étudié le VATAP dans le cadre de notices descriptives et termes d'indexation sélectionnés par l'archiviste en réponse à la question d'un usager; la discrimination dans la variation de sens a été opérée par l'archiviste de référence. Cela est vrai également pour la polysémie.

Deux formes identiques peuvent ne pas avoir le même sens, il s'agit de la polysémie; celle-ci est extrêmement fréquente dans la langue. Pour en prendre conscience, il suffit de comparer le nombre de mots et le nombre d'acceptions indiqué dans la préface du *Petit Larousse* (2018) : plus de 63 000 mots, 125 000 sens. Dans le Tableau XXI Illustration de la relation d'identité (écart 0), dans la section 5.1.1 Résultat de l'application de l'échelle d'écart sémantique, le terme *anatomie* est employé dans un sens marqué : « *Vx* Dissection » (*Grand Robert* 2017). Bien que cette acception ne soit pas la plus courante actuellement, l'usager et l'archiviste ont employé ce mot dans le même sens. Ainsi, notre corpus ne présente pas de cas de polysémie entre deux thémanymes. On aurait pu l'identifier entre *la Ville* et *Montréal*, mais l'analyse en facettes des expressions dans leur contexte a pu distinguer leurs acceptions : *la Ville* (10, l'institution) et *Montréal* (40, le lieu). Cette propriété sémantique d'un mot avait pourtant été envisagée puisqu'elle constitue un problème majeur dans la recherche d'information (Chu 2010; Lancaster 1986, 2003). La polysémie est parfois évoquée par l'homographie c'est-à-dire le fait que deux mots différents aient la même forme (Chu 2010, 59; Hudon 2013, 121), par exemple « son » (enveloppe extérieure du blé; étymologie : forme latinisée *seonnum,* 1243, *Grand Robert* 2017) et « son » (phénomène acoustique; étymologie anc. franç. *suen,* v. 1120 ; lat. *sonum, Grand Robert* 2017). L'établissement d'un langage documentaire permet justement de contrôler cet aspect de la langue (Arsenault 2006). La polysémie et l'homonymie ont été rapidement évacuées de l'échelle d'écart sémantique. En effet, il était peu probable qu'il y ait une telle ambigüité car le mot – du point de vue de la forme – était employé dans un discours qui offrait un contexte désambiguïseur et non utilisé en terme de recherche isolé. Nous avons rencontré dans deux thémanymes le mot *colonie(s)*, au sens des terres éloignées (lieu) et au sens de l'établissement d'une nation (activité). Cette distinction d'acception nous a permis de

catégoriser au mieux le lien sémantique entre les thémanymes (voir les explications du Tableau XXXI). De plus, les résultats de la recherche dans les bases de données ont été filtrés, sélectionnés par l'archiviste de référence qui avait compris précisément quelle acception l'usager employait. Ainsi, la polysémie n'a pas été incluse dans l'échelle d'écart sémantique et est absente des résultats.

6.3.6.2. Normalisation des formules introductives du sujet dans les notices descriptives

Nous remarquons dans le Tableau LXXVI sur les Formules introductives du sujet (voir section 5.5) que les formules introductives du sujet sont plus variées dans les notices descriptives. L'harmonisation des formules faciliterait le repérage thématique et peut-être même l'indexation automatique à partir des notices descriptives. Dans un rapport pour l'OCLC, Schaffner rappelle que dès 1993, Spindler et Pearce-Moses avaient proposé d'adapter les méthodes de description pour améliorer le repérage ainsi que Light (2008) a suggéré des stratégies pour adapter la description de manière à améliorer les possibilités de repérage. De son côté, elle préconise une adaptation quitte à « tordre les règles de description » (*bending rules for description,* 2015, 91, n. t.). Ainsi, à la suite de ces auteurs, nous pensons que la zone de la portée et contenu (spécifiquement la portée) devrait faire l'objet d'une normalisation, précisément en ce qui a trait au formules introductives du sujet (p. ex. cette unité archivistique *porte sur/traite de*…).

6.3.6.3. Explicitation des stratégies de recherche

Dans le codage des données, nous avons écarté de l'expression du sujet les éléments relatifs aux stratégies de recherche (voir section 3.5.3). Voici un exemple de réponse d'archiviste répondant à une question sur « l'exportation de castors en Patagonie en 1946 » :

> L'événement semble bien avoir eu lieu, comme le mentionnent les sites Internet cités plus bas, mais une recherche dans les fonds du ministère de l'Agriculture (COTE), du ministère des Affaires extérieures (COTE) ou du Programme des Affaires du Nord (COTE) n'a pas porté fruit. Voici les liens vers les pages web portant sur le sujet : SITE, SITE. Le seul dossier potentiellement intéressant que j'ai identifié dans notre collection documentaire porte sur l'envoi de rennes en Patagonie en 1947. Voici le lien vers la description du dossier en question : SITE. Si vous consultez l'onglet « Conditions d'accès », vous verrez que la documentation est disponible sur bobine de microfilm (COTE). Dans le cadre du Projet NOM, cette bobine a été numérisée et est disponible en ligne. Voici le lien vers la bobine : SITE. Le dossier 18323 débute à l'image 1346. À la

> lecture des documents dans le dossier précédemment mentionné, il semble clair qu'il a existé un programme d'envoi d'animaux avec l'Argentine dans les années 1940. L'ambassadeur du Canada en Argentine semble avoir été étroitement impliqué. J'ai donc poursuivi les recherches avec cette nouvelle information. M. NOM était l'ambassadeur du Canada en Argentine durant une bonne partie des années 1940 et 1950. Nous n'avons malheureusement pas le fonds d'archives de M. NOM (que je n'ai d'ailleurs pas été capable de retrouver dans aucune autre institution archivistique), mais sa trace est tout de même bien présente dans plusieurs dossiers à ARV. Voici le lien vers les descriptions de ces documents : SITE. Les documents identifiés avec la cote archivistique COTE proviennent du fonds du premier ministre NOM. Les descriptions ne sont pas très éloquentes, mais il se pourrait que l'envoi de castors fût mentionné dans l'une de ces lettres. Il est possible que la documentation portant sur votre sujet de recherche n'ait pas été préservée. Une recherche en profondeur serait nécessaire pour totalement élucider ce mystère. (a105r)

Comme mentionné précédemment, nous voyons que les stratégies de recherche sont explicitées dans le courriel de réponse d'un archiviste de référence, qu'elles soient fructueuses ou non. Par exemple, les premières recherches indiquées par l'archiviste n'ont pas donné de résultats positifs mais permettent à l'usager de comprendre qu'il ne sert à rien de continuer à chercher de ce côté-là. Par la suite, l'archiviste explicite son raisonnement et passe du sujet exprimé par des noms communs à des éléments du contexte du fait recherché exprimés par des noms propres, notamment de provenance. Cet élément serait à étudier plus en profondeur pour en dégager des éléments pertinents. Les stratégies de recherche décrites par les archivistes font partie de la formation aux usagers; les courriels de la référence participent à la littératie informationnelle pour les sources primaires (Yakel 2004, 63). L'explicitation du raisonnement serait utile pour la formation aux usagers, pour la formation aux futurs archivistes de référence et pour une meilleure compréhension du processus de la référence archivistique.

6.4. Apports de la recherche

Cette section rend compte des apports de la recherche sur les plans théorique, méthodologique et professionnel. Nous indiquons quelles sont les notions mobilisées dans deux disciplines, quels sont les outils développés, les méthodes employées et nous rappelons comment notre recherche appuie le rôle essentiel de l'archiviste de référence dans l'accès intellectuel aux archives.

6.4.1. Apports théoriques

Le premier apport théorique concerne le but même de notre recherche : la validation empirique d'un écart qui n'était jusqu'alors que présupposé. Le fossé sémantique entre le vocabulaire des usagers et celui employé dans les instruments de recherche est évoqué à de nombreuses reprises dans la littérature (voir section 3.2.3). Nous avons repéré des études sur la comparaison des termes des questions d'usagers avec des étiquettes (p. ex. dans le contexte muséal, par Angel 2012) ou avec un vocabulaire contrôlé (p. ex. Smiraglia 1990, voir section 3.4.3.2) dans la perspective de rendre compte du fossé sémantique. Mais sa vérification empirique, sa nature et sa valeur n'ont pas encore, à notre connaissance, été réalisées au sein d'une séquence (QRNTix) révélant la situation de communication de la référence archivistique. De plus, nous avons qualifié finement par une variété de relations sémantiques, certaines habituellement utilisées en sciences de l'information et d'autres moins connues des sciences de l'information. Finalement, nous avons apporté une première quantification de l'écart, pour laquelle d'autres études, recourant à des tests inter-codeurs, seraient nécessaires afin de solidifier nos résultats.

La notion de sujet en archivistique a été approchée d'un point de vue pratique; une question d'usager est thématique quand elle concerne des informations ou des documents qu'il ne connaît pas sur un sujet. Dans notre recherche, nous nous sommes intéressée seulement à l'expression du sujet par des noms communs. Nous en dégageons une définition formulée positivement dans notre recherche, mais non généralisable à la notion de sujet formulé avec des noms communs en archivistique.

La filiation sémantique opérée par l'attribution de champs sémantiques aux thémanymes permet de rassembler les thémanymes qui se rapportent les uns aux autres dans une séquence. L'identification de la filiation sémantique au sein d'une séquence ou d'une polyséquence est également un apport de cette recherche. Cet élément s'appuie sur des méthodes linguistiques, les chaînes de référence c'est-à-dire les éléments qui se rapportent les uns aux autres dans des textes (Schnedecker et Landragin 2014). La recherche sur ce sujet est en cours sur des textes scientifiques. Même si la notion de champ sémantique est ancienne, son application nécessite encore des développements. Cependant, les ouvrages et recherches

sur la linguistique textuelle telle que décrite par Adam (1999, 2008) pourraient consolider la démarche appliquée dans des études comme la nôtre.

Nous avons recouru à la notion linguistique de champ sémantique pour déterminer le sujet de documents représentant des sujets de documents d'archives patrimoniales. La linguistique est une discipline contributive des sciences de l'information (Salaün et Arsenault 2009). Notre recherche a rappelé l'importance de la langue et notamment de la sémantique dans une situation de communication, telle que la référence archivistique.

6.4.2. Apports méthodologiques

Dans cette recherche, nous avons développé pour nos besoins deux outils : un arbre décisionnel pour la sélection de questions d'usagers et une grille d'analyse pour l'identification des relations sémantiques entre les thémanymes étudiés.

Un arbre décisionnel pour la sélection des questions d'usagers à teneur thématique a été élaboré pour notre recherche, mais nous croyons qu'il est adaptable à d'autres milieux ou recherches sur les courriels. Selon la réponse que l'on choisit dans l'arbre décisionnel (oui/non), il pourrait servir à l'identification des questions d'usagers qui recherchent des documents précis, des informations connues ou qui ont formulé leur sujet avec uniquement des noms propres. Ainsi identifiés, ces types de questions pourraient être traités selon des procédures propres à chacun d'eux. Par exemple, les sujets exprimés par des noms propres pourraient au préalable être recherchés dans des bases de données onomastiques, avant que le travail de l'archiviste débute.

L'échelle d'écart sémantique est une grille d'analyse de la variation de forme entre deux expressions linguistiques et des relations sémantiques qui l'accompagne. Cette grille formée sur des critères linguistiques pourrait être utile dans des recherches de comparaison de termes en archivistique, en sciences de l'information et dans divers domaines.

L'identification des formules introductives des sujets a rapidement saturé dans notre recherche. La liste de ces formules peut d'ores et déjà permettre un repérage automatique non exhaustif des sujets présents dans une question d'usager ou dans une notice descriptive. Cette liste n'est pas un outil en soi. Mais sous forme de graphe, ces formules pourraient participer à

l'étude des structures d'introduction du sujet avec une plus grande rigueur dans l'identification des mécanismes d'introduction du sujet en français.

Mais c'est également toute la démarche méthodologique que nous avons créée qui pourrait être reprise, si elle était en partie automatisée. Les étapes d'anonymisation, d'épuration, de segmentation et de caractérisation sémantique nous ont permis d'obtenir des unités propices à l'analyse sémantique que nous avions choisie. En particulier, nous nous sommes servie d'une catégorisation thématique typique des sciences de l'information (l'analyse en facettes) pour identifier les acceptions des mots polysémiques, mobilisées dans le contexte étudié.

6.4.3. Apports professionnels

L'analyse linguistique a mis en lumière l'apport d'un agent cognitif tel que l'archiviste de référence par rapport à une analyse automatique des expressions linguistiques et de leurs liens sémantiques. Cette formalisation des relations sémantiques d'un travail humain (celui de l'archiviste) par un autre travail humain (notre analyse) a permis d'identifier les liens sémantiques récurrents dans les questions d'usager thématiques. Elle pourrait servir à l'automatisation d'une certaine partie du travail d'appariement de la question d'usager aux instruments de recherche ou encore à enrichir un outil de suggestion de termes lors de la recherche de l'usager.

La mise en valeur de l'accès thématique dans notre étude dépasse le cadre des fonds et collections d'un service d'archives. Nous pensons à la mutualisation de fonds de plusieurs services d'archives de moins grande envergure que les établissements de notre corpus – et donc aux ressources financières et humaines plus restreintes, mais aux collections non moins riches – qui recouvriraient des thématiques similaires, c'est-à-dire que les collections partageraient un même vocabulaire. L'accès thématique par l'indexation par sujet pourrait ajouter une couche d'information qui rendrait la recherche cohérente pour les usagers. Il serait dès lors possible de répondre aux usagers à partir de l'ensemble de ces fonds ainsi rassemblés thématiquement et indexés par sujet.

Comme le montre Mortureux (1993) dans son étude des reformulations (voir section 3.5.3), nous observons une variation dans les expressions linguistiques employées pour

évoquer un référent dans un même champ lexical. Cette variation n'est pas étonnante dans un contexte de recherche d'information : l'usager ne connaît pas nécessairement d'avance l'expression linguistique sous laquelle le sujet qu'il recherche a été mentionné dans les documents d'archives et dans leurs instruments de recherche. Pourtant, « certains aspects [des] reformulants [du terme] peuvent avoir l'avantage de focaliser des aspects du référent et du concept particulièrement adaptés à la situation d'énonciation. » (Mortureux 1993, par. 65) Ainsi, lorsque l'usager parle de certains concepts (p. ex. *bateau*), il identifie par son vocabulaire l'aspect sous lequel il mène sa recherche. Le travail de médiateur sémantique de l'archiviste de référence qui répond à une question d'usager est de permettre à cet usager de trouver des documents qu'il a sous sa garde et qui traitent du référent, mais pas nécessairement sous l'angle recherché (p. ex. *vaisseau*). Ceci offre alors à l'usager la possibilité d'interpréter les documents à sa manière (Guitard 2018).

Les ensembles de mots identifiés dans une situation de communication qu'est la référence archivistique (la question d'usager, la réponse de l'archiviste, la notice descriptive et éventuellement les termes d'indexation) fournissent un bassin de termes qui caractérisent thématiquement les documents qui traitent des sujets présents dans les questions d'usagers et pourraient servir à l'indexation des documents d'archives ou plutôt de leurs notices descriptives. Ces mots additionnels pourraient être suggérés à un indexeur, qu'il soit professionnel ou amateur, ou bien utilisés pour indexer automatiquement des documents. Nous proposons l'ajout de ces thémanymes comme termes d'indexation une fois qu'un usager a formulé une demande thématique à laquelle ces documents pouvaient répondre. Par ailleurs, l'extraction automatique de relations sémantiques ainsi identifiées pourrait permettre de suggérer des documents à l'usager à partir des mots employés. Cependant, une approche automatique serait confrontée à deux problèmes. D'une part, le nombre d'usagers et le nombre de notices descriptives faisant partie de notre corpus sont faibles et toute tentative de modélisation souffrirait de la pauvreté des données. D'autre part, la portée et contenu des notices descriptives et les courriels sont des textes hétéroclites, et donc difficiles à modéliser. Alors que l'environnement numérique invite la recherche en linguistique et en sciences de l'information à tenter d'automatiser certains processus tels que l'indexation, il semble que cette voie connaisse certaines limites et que le rôle de médiateur des professionnels du monde

documentaire doive, au contraire, être réaffirmé. Notre recherche sur le vocabulaire employé pour l'accès thématique aux archives patrimoniales (VATAP) appuie le rôle essentiel de l'archiviste de référence dans l'accès intellectuel aux archives. Ainsi, avec Coutaz, nous pensons que : « À la croisée du passé et de l'avenir, les archivistes ne seront pas remplacés par le clic informatique qui, s'il est un outil indispensable, ne sera jamais un substitut. » (2016, 21)

Conclusion du chapitre

Dans ce chapitre, nous avons répondu à l'ensemble des questions de recherche identifiées dans la section 2 But, objectifs et questions de recherche. Nous avons d'abord approfondi la description de la relation d'identité dans le contexte archivistique et la mise en perspective de ces résultats dans nos deux corpus, selon les sources et pour chacun des établissements. Nous avons aussi décrit les relations qui caractérisent un écart sémantique, en identifiant la part importante des relations à fort écart et des relations complexes qui rendent l'être humain pour l'instant indispensable. Ensuite, nous avons souligné la distinction entre le plan sémantique et le plan référentiel d'une relation et l'implication de cette distinction quant à son manque de réciprocité et de prédictibilité. Finalement, nous avons identifié les apports de la recherche sur les plans théorique, méthodologique et professionnel.

7. Conclusion de la thèse

Cette section résume notre recherche, présente des recommandations et se termine par des pistes de recherche futures.

7.1. Résumé de la recherche

Notre recherche a porté sur le vocabulaire employé pour l'accès thématique aux documents d'archives patrimoniaux (VATAP). Elle avait pour but d'étudier l'écart sémantique présupposé entre le vocabulaire des usagers qui recherchent des documents d'archives patrimoniaux et celui employé dans les instruments de recherche pour l'accès thématique à des documents d'archives patrimoniaux.

Les archivistes souhaitent améliorer les accès existants et offrir davantage d'accès thématiques pour mieux répondre aux besoins des usagers qui recourent principalement à ce type d'accès. L'accès thématique a longtemps été délaissé dans l'élaboration d'instruments de recherche archivistiques au profit de la provenance. Il en résulte que les usagers doivent recourir à un archiviste pour traduire leurs termes de recherche thématiques formulés avec des noms communs en noms propres de personnes (provenance) puis en titre de fonds et collections employés dans chacun des services d'archives qu'ils fréquentent. Les services d'archives ont peu recours aux langages documentaires et utilisent plutôt le vocabulaire libre; ils manquent en outre de lignes directrices pour encadrer leur travail relativement à ce sujet. Dans l'accès thématique aux documents et la recherche d'information en vocabulaire libre, la question du vocabulaire – les noms communs – est centrale. Mais elle est à ce jour peu étudiée. Les archivistes de référence, dans leurs réponses aux questions d'usagers, rapprochent le vocabulaire employé dans les instruments de recherche de celui des usagers. L'existence d'un écart sémantique identifié par certains auteurs n'avait pas été validée empiriquement. Une fois constaté, cet écart a été décrit. Pour ce faire, nous avons choisi l'approche linguistique qui est la plus à même de rendre compte de différences d'ordre sémantique entre des expressions linguistiques. La connaissance de l'accès par sujet aux archives patrimoniales est à développer. En adoptant un point de vue sémantique pour étudier les vocabulaires utilisés pour rechercher, décrire et indexer des documents d'archives patrimoniaux, nous souhaitions

contribuer à l'étude du sens dans les termes d'indexation thématique et plus largement dans le vocabulaire employé pour l'accès par sujet aux archives patrimoniales (VATAP).

Dans notre revue de la littérature, nous avons abordé à la fois des notions archivistiques et des notions linguistiques. Les documents d'archives recèlent en leur définition une nature singulière et hiérarchique au sein d'un fonds qui leur confère un statut unique. De cette spécificité découle un traitement qui requiert d'être lui aussi spécifique : il n'est pas possible de reprendre les lignes directrices documentaires telles quelles pour les appliquer aux archives. Les sujets portés par des fonds d'archives ne font pas exception. Les questions d'usagers à teneur thématique lors d'un entretien de référence différé tel que par courriel rendent compte de cette spécificité dans la complexité des liens qui relient les termes de recherche aux termes de description et d'indexation. L'archiviste de référence occupe une position de médiateur sémantique entre les usagers et les documents d'archives. Dans le processus de référence archivistique, il est un maillon inaliénable de la chaîne communicationnelle des usagers aux documents d'archives et les instruments de recherche jouent un rôle essentiel dans ce processus. L'archiviste s'appuie sur les instruments de recherche et, pour la question de l'accès thématique, il s'appuie particulièrement sur l'indexation thématique.

En ce qui concerne les notions linguistiques, nous avons défini ce qu'est le vocabulaire pour l'accès thématique aux archives patrimoniales en décrivant les termes de recherche, de description et d'indexation. La lexicologie traite des mots, tant du point de vue de leurs propriétés que des relations sémantiques qu'ils entretiennent entre eux. Nous avons décrit ces deux éléments – les propriétés et les relations sémantiques – dans la littérature linguistique et dans la littérature des sciences de l'information.

Nous avons développé une méthodologie complexe pour ramener l'ensemble des expressions linguistiques porteuses de sujet à un format similaire par un traitement incluant l'anonymisation, l'épuration, la segmentation et la caractérisation sémantique des données, par l'analyse en facettes et l'attribution d'un champ sémantique. De cette caractérisation sémantique, nous avons pu établir une filiation sémantique qui est aussi une filiation du sujet entre les diverses sources de notre corpus : question, réponse, notice descriptive et terme d'indexation. Cette filiation du sens et du sujet est basée sur la chaîne communicationnelle

entre l'usager et les documents d'archives, à laquelle contribue l'archiviste de référence par sa réponse et la sélection d'instruments de recherche pertinents. Pour nous aider à assurer une certaine régularité dans l'attribution des relations sémantiques, nous avons bâti une grille d'analyse de repérage et de qualification des relations sémantiques que nous avons appelée *échelle d'écart sémantique*. Elle s'échelonne de l'absence de relation par un degré zéro (écart 0) qui rend compte de l'identité des expressions linguistiques porteuses de sujet (thémanymes) comparées, à une relation sémantique assez lâche et difficile à cerner qu'est le champ sémantique (écart 13). Nous avons ajouté en bas de cette échelle la relation référentielle (écart 14), déjà identifiée dans la littérature, sans identification claire cependant.

Nos résultats ont montré que 12,1% des liens sémantiques identifiés dans le corpus général sont de l'identité stricte (écart 0 seulement). Cette proportion chute à 2,9% des liens sémantiques dans le corpus général si l'on ne considère que ceux identifiés entre le vocabulaire des usagers et celui dans les instruments de recherche. Si l'on considère seulement les faibles variations telles que celles prises en charge par les cinq premiers échelons de l'échelle d'écart sémantique (écart 1 à 5), variations qui sont prises en compte par de nombreux moteurs de recherche, alors nous pouvons regrouper plusieurs résultats autour de la notion de relation d'identité large (écarts 0 à 5). Les liens sémantiques entre le vocabulaire des usagers et celui dans les instruments de recherche atteignent alors 4,6% du corpus général (n=1210). Si nous poursuivons le raisonnement et tentons de regrouper l'ensemble des phénomènes liés à l'identité non seulement large (écarts 0 à 5) mais aussi ceux liés à des variations syntagmatiques plus importantes jusqu'à l'équivalence telle qu'elle est entendue en tant que relation thésaurale (écarts 6 à 10), alors nous créons un regroupement de résultats que nous appelons identité-équivalence (écarts 0 à 10). Le regroupement identité-équivalence représente 6,7% de toutes les relations du corpus général (n=1210). Ainsi, l'écart présupposé a été vérifié empiriquement et représente entre 97,1% du corpus général selon la considération d'une relation d'identité stricte, 95,4% selon celle d'une relation d'identité élargie et 83,3% selon celle d'une relation d'identité-équivalence.

Les retombées de notre recherche pourraient être nombreuses. D'un point de vue théorique, nous avons établi une caractérisation (à ce jour inédite) de l'écart entre le vocabulaire des usagers et celui employé dans les instruments de recherche. L'application

d'une méthodologie linguistique à un corpus archivistique est également novatrice et constitue une avancée dans la collaboration des deux disciplines. D'un point de vue méthodologique, nous avons développé une démarche d'analyse dans la présente étude qui pourra être reproduite par les services d'archives afin d'avoir un portrait linguistique de leurs usagers quand ils le jugeront nécessaire. Pour ce faire, nous avons développé deux outils : un arbre décisionnel pour la sélection des questions d'usagers à teneur thématique et une échelle d'écart sémantique pour nommer les relations sémantiques entre deux thémanymes. D'un point de vue professionnel, la formalisation des relations sémantiques pourrait servir à l'établissement de règles de création des termes d'indexation thématique en vocabulaire libre pour les archives patrimoniales et à l'automatisation de cette indexation. Une meilleure connaissance des tendances linguistiques des usagers d'un centre ou service d'archives pourrait permettre à l'établissement d'ajuster ses instruments de recherche issus de la description et de l'indexation. La modification de la notice descriptive à partir de critères linguistiques clairs pourrait faciliter l'indexation automatique. L'identification des relations sémantiques les plus fréquentes dans trois établissements pourrait permettre le développement d'un outil de raffinement de la recherche par la suggestion de termes issus de l'application de ces relations aux premiers termes saisis. D'un point de vue sociétal, notre recherche contribue à mieux comprendre et indirectement à améliorer l'accès intellectuel des documents d'archives et de l'information qu'ils contiennent à l'ensemble des usagers. Ceci favorise l'enrichissement de citoyens éclairés par la connaissance de leur patrimoine et de leur culture ainsi que le développement de leur savoir.

7.2. Recommandations à l'attention des archivistes

À partir de notre étude du vocabulaire pour l'accès thématique aux archives patrimoniales (VATAP) dans trois établissements ainsi que de notre lecture de la littérature sur le sujet, nous émettons des recommandations à l'attention des archivistes.

- Diversifier le vocabulaire dans les notices descriptives et indexer par sujet les notices en variant le vocabulaire de la portée permettraient d'offrir davantage de clés d'accès thématiques. Cela favoriserait la « rencontre entre les archives et les usagers » (Bédard et Morel 2013-2014).

- Formaliser la description au niveau de la description thématique. À la suite de Duff (2010), nous pensons que les instruments de recherche ne devraient pas seulement être améliorés mais repensés. Une fois dégagés de l'emprise de l'accès traditionnel par provenance, les archivistes pourraient imaginer des instruments de recherche thématiques réellement conçus pour servir les usagers.
- Mettre les instruments de recherche même temporaires, même inachevés, dès l'acquisition, en ligne. Cette recommandation s'inscrit dans le courant « More Products Less Process, MPLP » (« moins de procédures, plus de produits », n. t.) présentée par Green et Meissner (2005). Plus il y a des ressources archivistiques en ligne, plus les questions d'usagers sont « substantielles » (Pugh et Power 2015, cité par Duff et Yakel 2017, 201-202[76]). Cette nouvelle habitude pourrait réduire le temps pour donner accès aux fonds et collections.

7.3. Recherches futures

Les recherches futures potentielles sont nombreuses. Ayant dû couper à plusieurs reprises l'étendue de notre sujet, nous envisageons de multiples manières dont la recherche sur le vocabulaire employé pour l'accès thématique des archives patrimoniales pourrait continuer. Dans la section 4.1.3, nous évoquons le fait de ne pas avoir conservé les données dont la notice ne correspondait pas exactement au niveau de la cote indiquée par l'archiviste. L'étude du vocabulaire des notices descriptives des niveaux supérieurs pourrait indiquer si le vocabulaire employé à ce niveau–là est plus général ou non et de quelle manière il peut rencontrer ou non le vocabulaire employé par les usagers.

[76] La référence donnée par les auteures est la suivante : Pugh, J.et C. Power. (2015). « Swimming the Channels: An Analysis of Online Archival Reference Enquiries ». Dans Abascal J., S. Barbosa, M. Fetter, T. Gross, P. Palanque et M. Winckler (Dir.). *Human-Computer Interaction – INTERACT 2015*. Lecture Notes in Computer Science. vol. 9298. Springer, Cham.

De nombreuses étapes méthodologiques ont été réalisées humainement, sans le concours ou si peu de la machine. Des logiciels spécialisés existants ou à développer auraient facilité ou accéléré le traitement des données, ce qui serait souhaitable dans un milieu professionnel. Développer à partir de la méthodologie créée ici des applications informatiques en vue de l'automatisation de certaines tâches, pour rendre les usagers plus autonomes, permettrait de dégager du temps à l'archiviste afin qu'il dispose de davantage de temps pour répondre plus rapidement à un plus grand nombre d'usagers et avant cela pour traiter davantage de fonds d'archives.

Pour avoir un portrait plus complet des relations sémantiques en jeu dans toute la chaîne communicationnelle des documents à l'usager, il faudrait poursuivre l'analyse des relations à l'intérieur d'une même source (Q, R, N ou Tix) et ainsi avoir une meilleure idée des rouages exercés dans la thématisation, notamment dans la question d'usager. Pour ce faire, la linguistique textuelle (Adam 1999, 2008) pourrait apporter un éclairage intéressant à l'étude de la filiation du sujet à l'intérieur d'une source.

Nous avons collecté des fichiers de termes d'indexation provenant de deux établissements et les étiquettes pour trois établissements. Nous pourrions utiliser les notes associées aux termes d'indexation professionnelle (dans le fichier de l'établissement 1) et les pages de description dans *Flickr* présentant des termes d'indexation collaborative (les trois établissements) pour identifier les notices descriptives. Ceci nous permettrait d'observer d'une part dans quels cas un professionnel juge nécessaire d'introduire des termes d'indexation dans une notice descriptive constituant déjà un champ de mots où l'on peut rechercher en texte intégral. Et d'autre part, nous avons une association de termes provenant d'usagers et de professionnels pouvant être comparés. À partir de l'étude des étiquettes par rapport aux termes d'indexation issus de l'indexation professionnelle, nous pourrions peut-être proposer des stratégies pour inclure l'indexation collaborative dans l'élaboration ou l'enrichissement d'instruments de recherche actuels. Une meilleure connaissance des étiquettes liées à des archives patrimoniales pourrait notamment contribuer à évaluer leur qualité pour éventuellement les inclure dans les systèmes de repérage ou de suggestion à l'indexation professionnelle.

La lexicologie a pour domaine l'étude du lexique, c'est-à-dire non seulement des relations qui s'établissent entre les mots mais aussi des propriétés de ces mots. Nous avions envisagé d'explorer des propriétés linguistiques telles que la massivité ou la concrétude des noms. Les recherches en linguistique sur ce point sont en cours; elles n'ont pas encore formalisé clairement la reconnaissance de ces critères en français; une telle recherche sur corpus pourrait faire avancer la recherche linguistique appliquée. En outre, nous avons déjà identifié (voir section 3.5.1) les propriétés linguistiques qui seraient à étudier prioritairement en tant qu'elles sont citées dans les normes documentaires et dans les manuels de lexicologie. L'étude des propriétés linguistiques de termes de recherche, de description et d'indexation complèterait bien le portrait linguistique du VATAP. Est-ce que certaines relations sémantiques s'établissent préférablement entre des mots ou des thémanymes qui ont certaines propriétés linguistiques ? La question reste à explorer.

Les textes présentent diverses caractéristiques propres à leur genre, qui est défini par la forme, le contenu et la fonction du document (Alberts 2009, 2015). Dans notre corpus, nous avons traité des courriels et des notices descriptives. Leur rédaction subit nécessairement l'influence de leur genre : nous reconnaissons une notice descriptive et la distinguons clairement d'une question d'usager ou d'une réponse d'archiviste par courriel. Les deux revêtent une forme – dont la structure – qui leur est propre; il s'agit de rédaction adaptée au genre de document. Les éléments porteurs du sujet apparaissent dans ces deux genres de documents dans des sections propres à chacun d'eux : dans l'objet du courriel, dans des phrases impliquant des formules introductives du sujet ou bien dans le titre d'une unité archivistique et dans sa portée en recourant également à des formules introductives du sujet. Lors de l'épuration, nous avons retiré les éléments caractéristiques du genre de chaque source (pour le courriel : salutations, remerciements, identification du locuteur, p. ex.; pour la notice : contenu) pour les ramener à des segments de texte thématique similaires. Ces éléments, rejetés dans notre étude, pourraient servir à caractériser les genres de documents utilisés lors du processus de référence archivistique.

Si le processus de référence a déjà été étudié par d'autres chercheurs (notamment Duff et Fox 2006, voir section 3.2.1), le vocabulaire qui permet l'accès thématique aux archives patrimoniales n'a jamais été étudié à ce jour sous la forme globale que nous avons proposée

(voir Tableau IV VATAP : 4 sources). Notre étude est une toute première étape dans l'exploration du portrait global du vocabulaire pour l'accès thématique aux archives patrimoniales (VATAP) au Québec à l'heure actuelle. Lors de notre collecte, nous avons procéder à des choix afin de restreindre notre corpus et de mener notre recherche doctorale en un temps raisonnable. Des recherches futures pourraient permettre de procéder aux autres méthodes de collecte identifiées en 4.1.2.4 et d'élargir l'éventail des services participants (en variant les tailles et les milieux des services d'archives, leurs langues de travail et leur localisation afin de couvrir diverses méthodes de référence, de description et d'indexation). Pour approfondir l'étude du processus de référence en tant que tel du point de vue de l'archiviste, une recherche recourant à l'analyse de protocole verbal permettrait de rendre compte plus finement de la démarche complète de l'archiviste.

Nous espérons que notre travail pourra servir de base à d'autres travaux, qui voudraient explorer les pistes que nous avons relevées ici.

Bibliographie

La bibliographie présente quelques repères alphabétiques pour faciliter sa consultation.

A

Adam, J.-M. (1999). *Linguistique textuelle. Des genres de discours aux textes*. Paris : Éditions Nathan.

———. (2008). *Linguistique textuelle : introduction à l'analyse textuelle des discours.* Collection Lettres, Cursus. 2e éd. Paris : A. Colin.

AFNOR (Association française de normalisation). (1981). *Règles d'établissement des thésaurus monolingues*. Norme NF Z 47-100. Paris : Association française de normalisation.

———. (1993). *Information et documentation : principes généraux pour l'indexation des documents*. Norme *NF Z 47-102*. Paris : Association française de normalisation.

Alberts, I. (2009). *Exploitation des genres de textes pour assister les pratiques textuelles dans les environnements numériques de travail : le cas du courriel chez des cadres et des secrétaires dans une municipalité et une administration fédérale canadiennes* (Thèse de doctorat, Montréal : Université de Montréal).

———. (2015). « Caractérisation du genre des textes administratifs dans les environnements numériques de travail ». Dans Gagnon-Arguin, L., S. Mas et D. Maurel (dir.), *Les genres de documents dans les organisations : analyse théorique et pratique* Québec : Presses de l'Université du Québec :69-86.

Allen, G. (2000). *Intertextuality*. Londres et New York : Routledge.

Amar, M. (2000). *Les fondements théoriques de l'indexation : une approche linguistique*. Paris : ADBS.

Angel, C. M. (2012). *A Comparison of Descriptive Tagging Practices by Library, Archive and Museum Professionals using an Inter-Indexing Consistency Approach* (Thèse de doctorat, University of South Carolina).

ANSI/NISO (American National Standards Institute. National Information Standards Organization). (2005). *Guidelines for the construction, format and management of*

monolingual thesauri. Norme *ANSI/NISO Z 39.19*:2005. Baltimore (MD): National Information Standards Institute. Réexaminée en 2010.

Antidote (Antidote 9). (2017). Druide informatique. Correcteur grammatical, dictionnaires et guides linguistiques. Accès payant.

Archives nationales du Québec. (1991). *Archives écrites d'origine privée, conservées au Centre d'archives de Québec et de Chaudière-Appalaches*. Québec : Les publications du Québec.

Arsenault, C. (2006). « L'utilisation des langages documentaires pour la recherche d'information = The Use of Index Languages for Information Retrieval ». *Documentation et bibliothèques* 52(2): 139-148.

Assemblée nationale du Québec. (2018). *Bases de données. Bases de données produites par la bibliothèque*. Repéré à http://www.bibliotheque.assnat.qc.ca/fr/24-bases-de-donnees.

ASTED. (1998-2005). *Règles de catalogage anglo-américaines*. Traduction de *Anglo-American Cataloguing Rules*. Ottawa : Canadian Library Association, 1988. 2e éd. Révision : 1998. Modifications : 2001-2005.

Aussenac-Gilles, N. et P. Seguela. (2000). « Les relations sémantiques : du linguistique au formel ». *Cahiers de grammaire* (25): 175-198.

BAnQ (Bibliothèque et Archives nationales du Québec). (2012). *Guide d'indexation des archives*. Montréal : Bibliothèque et Archives nationales du Québec. [Document non publié].

———. (2018). *Pistard Archives*. Repéré à http://www.banq.qc.ca/aide/pistard/index.html

BCA (Bureau canadien des archivistes). Comité de planification sur les normes de description. (2008). *Règles pour la description des documents d'archives* (*RDDA*). 2e éd. Ottawa : Bureau canadien des archivistes (BCA).

BCA (Bureau canadien des archivistes). Groupe de travail sur l'indexation par sujet. (1992). *L'indexation par sujet en archivistique*. Ottawa : Bureau canadien des archivistes. Comité de planification sur les normes de description (BCA. CPND).

Bécherel, D. (1994). « L'opposition des deux parties du discours adjectif / substantif : Définitions et ajustements terminologiques ». *Meta* (39) : 625-635.

Bédard, S. et S. Morel. (2013-2014). « L'archiviste, outil de médiation entre les archives et l'usager ». *Archives* 45 (1) :47-56.

Bellec, P. (2016). « Actualité et enjeux de la Dewey : Classer, indexer, cartographier la collection ». *Bulletin des bibliothèques de France (BBF)*, 9: 18-31.

Benetti, L. (2008). *L'article zéro en français contemporain : aspects syntaxiques et sémantiques.* Bern ; New York : P. Lang.

Bertrand, A., J. M. Cellier et L. Giroux. (1994). « Expertises et stratégies dans une activité d'indexation de documents scientifiques ». *Le Travail Humain* 57(1): 25-51.

Bertrand-Gastaldy, S., L. Giroux, D. Lanteigne et C. David. (1994). « Les produits et processus cognitifs de l'indexation humaine ». *ICO-Québec* 6(1-2): 29-40.

Bibliothèque nationale de France. (2017). *Guide d'indexation RAMEAU*. 7e édition.

———. (2018). *BnF archives et manuscrits* Repéré à http://www.bnf.fr/fr/collections_et_services/catalogues_en_ligne/a.bnf_archives_et_manuscrits.html

Billaudele, A. (2015, 21 mai). *Table de concertation des archives religieuses de Montréal* [Billet de blogue]. Repéré à : https://regroupementarchivistesreligieux.wordpress.com/2015/05/21/table-de-concertation-des-archives-religieuses-de-montreal-tcarm/

Bougoin, A. (2015). *Indexation automatique par termes-clés en domaines de spécialité* (Thèse de doctorat, Université de Nantes Angers Le Mans).

Broudoux, E. (2013). « Quelles lectures du tagging? ». *Document numérique* 16(1): 55-71.

Broughton, V. (2003). « Facet analytical theory as a basis for a knowledge organization tool in a subject portal ». Dans María J. López-Huertas with the assistance of Francisco J. Muñoz-Fernández, *Challenges in knowledge representation and organization for the 21st century: integration of knowledge across boundaries:* proceedings of the the Seventh International ISKO Conference, 10-13 July 2002, Granada, Spain. Würzburg: Ergon Verlag, 8:135-142.

Broughton V. et Slavic A. (2004). *Facet analytical theory in managing knowledge structures for the humanities* (FATKS). Repéré à http://www.ucl.ac.uk/fatks/fat.htm

BSI (British Standards Institution). (2005). *Structured Vocabularies for Information Retrieval: Part 1: Definitions, Symbols and Abbreviations*. Norme *BS 8723-1*:2005. Annulée et remplacée par *ISO 25 964-1*:2011.

———. (2005). *Structured Vocabularies for Information Retrieval: Part 2: Thesauri*. Norme *BS 8723-2*:2005. Annulée et remplacée par *ISO 25 964-1*:2011.

Buckland, M. K. (1999). « Vocabulary as a Central Concept in Library and Information Science ». *The Third International Conference on conceptions of library and information science (CoLIS3)*, Dubrovnic, Croatie, Digital libraries: Interdisciplinary concepts, challenges, and opportunities. 3-12.

C

Cadieux, H. et N. Charbonneau. (2002). « Indexation aux Archives nationales du Québec ». *Archives* 33(3-4): 67-96.

Camart, C., F. Mairesse, C. Prévost-Thomas et P. Vessely. (2015). « Introduction ». Dans Camart, C., F. Mairesse, C. Prévost-Thomas et P. Vessely, *Les mondes de la médiation culturelle* vol. 2 *Médiations et cultures*. Collection Les cahiers de la médiation culturelle. Paris : L'Harmattan. 9-16.

Cardin, M. (1999). « Patrimoine archivistique religieux : enjeux et perspectives ». *Historical studies* 65: 53-66.

Cardinal, L., V. Chabot, J. Ducharme, G. Janson et G. Lapointe. (1984). *Les Instruments de recherche pour les archives.* La Pocatière, Québec : Documentor.

Cartier, E. (2009). « Étiquetage sémantique des textes : états des lieux, éléments de modélisation ». *L'Information Grammaticale* : 19-29.

Cartier, F. (2008). *Recueil de cours. ARV1055 Description des archives*. École de bibliothéconomie et des sciences de l'information. Université de Montréal. [Document non publié].

Cercle de Fermières. (2017). *Page d'accueil*. Repéré à https://cfq.qc.ca/a-propos/les-cercles-de-fermieres/

Chabin, M.-A., C. Scopsi, F. Dez, B. Müller, J.-J. Thomasson et L. Nardone (ill.). (2016). *Une journée dans la vie de Lili Eting (et son chat). Petit précis de diplomatique souriante*. Repéré à http://www.marieannechabin.fr/wp-content/uploads/2016/10/Une-journee-dans-la-vie-de-lili-eting-versionpdf.pdf

Champagne, M. (2017). « Le traitement définitif d'un fonds à la DGDA ». *Archives* 46(2) :61-80.

Chaumier, J. (1988). *Le traitement linguistique de l'information*. Paris : Entreprise moderne d'édition.

Cheminée P. (1990). *Quelques aspects de la désignation du musicien dans le discours de la critique musicale* (Mémoire de DEA, Université de Paris X[e]).

Chen, M., X. Liu, et J. Qin. (2008). *Semantic Relation Extraction from Socially-Generated Tags: A Methodology for Metadata Generation*. Proc. Int'l Conf. on Dublin Core and Metadata Applications.

Chichereau D., O. Contat, D. Dégez, A. Deniau, M. Lénart, C. Masse, D. Ménillet. (2007). « Les normes de conception, gestion et maintenance de thésaurus. Évolutions récentes et perspectives ». *Documentaliste-Sciences de l'Information* 44(1): 66-74. Repéré à http://www.adbs.fr/servlet/com.univ.collaboratif.utils.LectureFichiergw?ID_FICHE=376&OBJET=0017&ID_FICHIER=1380

Choi, Y. et S. Y. Syn. (2016). « Characteristics of tagging behavior in digitized humanities online collections ». *Journal of the Association for Information Science and Technology* 67(5): 1089-1104.

Chu, H. (2010). *Information Representation and Retrieval in the Digital Age.* Medford (NJ): Information Today.

Chung, E. K. et J. Yoon. (2009). « Categorical and Specificity Differences between User-Supplied Tags and Search Query Terms for Images. An Analysis of Flickr Tags and Web Image Search Queries ». *Information Research* 14(3): 3.

Clarke, S. G. D. (2001). « Thesaural Relationships ». Dans Bean, C. A. et R. Green (dir.), *Relationships in the Organization of Knowledge*. Springer Netherlands: Dordrecht.

Clarke, R. I. et J. H. Lee. (2015). « User Perceptions of Associative Thesaural Relationships: A Preliminary Study', *iConference 2015 Proceedings*.

Commission de toponymie du Québec. (2018). *Banque de noms de lieux du Québec*. Repéré à http://www.toponymie.gouv.qc.ca/ct/accueil.aspx

Corblin, F. (1995). *Les formes de reprise dans le discours. Anaphores et chaînes de référence*, Rennes : P. U. Rennes.

Coutaz, G. (2016). *Archives en Suisse : conserver la mémoire à l'ère numérique*. Lausanne : Presses polytechniques et universitaires romandes.

Couture, C. et M. Lajeunesse. (2014). *L'archivistique à l'ère du numérique : Les éléments fondamentaux de la discipline.* Québec : Presses de l'Université du Québec.

Creswell, J. W. (2009). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches.* Thousand Oaks (CA) : Sage Publications.

Da Sylva, L. (2009a) « Outil de butinage du contenu des documents de collections numériques ». *Actes du 12e Colloque International sur le Document Électronique* (CiDE). Université de Montréal, Montréal, 21-23 octobre 2009: 263-273.

———. (2009b). « Classes de vocabulaire et indexation automatique : le cas des index de livres ». *Premier Workshop international sur la Terminologie et la sémantique lexicale (TLS'09)*. Université de Montréal, Montréal : 67-76.

———. (2015). *SCI6133 Indexation et condensation – Notes de cours*. École de bibliothéconomie et des sciences de l'information. Université de Montréal.

———. soumis. « Analyse sémiotique de l'index de livre : étude de la construction complexe et unique d'un paratexte ». Destiné à la revue *Semiotica*.

Dancette, J. (2011). « L'intégration des relations sémantiques dans les dictionnaires spécialisés multilingues : du corpus ciblé à l'organisation des connaissances ». *Meta* 56: 284-300.

Daniels, M. et E. Yakel. (2010). « Seek and You May Find: Successful Search in Online Finding Aid Systems ». *The American Archivist* 73(2): 535-568.

de Keyser, P. (2012). *Indexing: from thesauri to the semantic web.* Oxford : Chandos.

Delmas, B. et F. Blouin. (1996). « Présentation : De la diplomatique médiévale à la diplomatique contemporaine (actes du colloque organisé par l'École nationale des chartes et la Bentley historical Library de L'université de Ann-Arbor [Michigan, États-Unis], Paris, 6-10 juillet 1992 et Ann-Arbor, 5-9 juillet 1993) ». *La Gazette des archives* 172: 9-11.

Deweze, A. (1981). *Réseaux sémantiques : essai de modélisation, application à l'indexation et à la recherche de l'information documentaire*. Lyon, Université Claude-Bernard.

Direction des archives de France. (1993). *La Pratique archivistique française*. Paris : Archives nationales.

Druide informatique. (2017). *Posologie : guide d'utilisation d'Andidote 9*.

Dubois, D. (1983). « Analyse de 22 catégories sémantiques du français : organisation catégorielle, lexique et représentation = An Analysis of 22 Semantic French Categories:

Categorial Organisation, Lexicon and Representation ». *Année Psychologique* 83(2): 465-489.

Duff, W. M. (2010). « Archival mediation ». Dans T. Eastwood et H. MacNeil (dir.), *Currents of Archival Thinking*. Santa Barbara, Californie, Libraries Unlimited: 115-136.

Duff, W. M. et A. Fox. 2006. « 'You're a guide rather than an expert': Archival reference from an archivist's point of view ». *Journal of the Society of Archivists* 27: 129-53.

Duff, W. M. et C. A. Johnson (2001). « A Virtual Expression of Need: An Analysis of E-mail Reference Questions ». *The American Archivist* 64(1): 43-60.

———. (2002). « Accidentally Found on Purpose: Information-Seeking Behavior of Historians in Archives ». *The Library Quarterly* 72(4): 472-496.

Duff, W. M., D. M. Johnson et J. M. Cherry (2013). « Archives' Use of Social Media in Canada ». *Archivaria* 75(Spring): 77-96.

Duff, W. M. et E. Yakel. (2017). « Archival Interaction ». Dans H. MacNeil et T. Eastwood (dir.), *Currents of Archival Thinking*: 193-223. Santa Barbara (CA) : Libraries Unlimited

Duff, W. M., E. Yakel, et H. Tibbo. (2013). « Archival Reference Knowledge ». *The American Archivist* 76(1): 68-94.

E

Ehrmann, M. (2008). *Les Entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation* (Thèse de doctorat, Université PARIS 7 - Denis Diderot).

Engerer, V. (2017). « Control and syntagmatization: Vocabulary requirements in information retrieval thesauri and natural language lexicons ». *Journal of the Association for Information Science and Technology* 68: 1480-1490.

Fasciolo, M. et M. Lammert. (2015). « Types de noms et critères définitoires ». *Travaux de linguistique*: 7-10.

Fellbaum, C. (dir.). 1998. *WordNet: An electronic lexical database*. Cambridge, MA : MIT Press.

Fenton, C. (2010). « Use of controlled vocabulary and thesauri in UK online finding aids ». *Journal of the Society of Archivists* 31(2): 187-205.

Fiala, P. (1994). « L'interprétation en lexicométrie. Une approche quantitative des données lexicales ». *Langue française* (103): 113-122.

Fidel, R. (1994). « User-Centered Indexing ». *Journal of the American Society for Information Science* 45(8): 572-576.

Fillmore, C. J. (1976). « Frame semantics and the nature of language ». *Annals of the New York Academy of Sciences* 280: 20-32.

Flaux, N. et D. Van de Velde (2000). *Les noms en français : esquisse de classement*. Gap, France; Paris : Ophrys.

Flickr. (2015). *Nouveau compte Flickr Pro – de meilleures stats, des avantages Adobe et bien plus encore !* [Billet de blogue]. Repéré à http://blog.flickr.net/fr/2015/07/23/nouveau-compte-flickr-pro-de-meilleures-stats-des-avantages-adobe-et-bien-plus-encore/

———. (2017). *À propos*. Repéré à https://www.flickr.com/about

Fortin, M. F. et J. Gagnon. (2016). *Fondements et étapes du processus de recherche: méthodes quantitatives et qualitatives.* 3e éd. Montréal : Chenelière éducation.

Fortin, M.-F., J. Côté et F. Filion. (2006). « Chapitre 14 : L'échantillonnage ». Dans *Fondements et étapes du processus de recherche*, Montréal : Chenelière éducation : 248-270.

Fortin, F. (2010). *Fondements et étapes du processus de recherche : méthodes quantitatives et qualitatives.* Montréal : Chenelière éducation.

Fortis, J.-M. (2010). « De l'hypothèse de Sapir-Whorf au prototype : sources et genèse de la théorie d'Eleanor Rosch ». *Corela. Cognition, représentation, langage* 8(2).

Frey, V. et M. Treleani. (2013). *Vers un nouvel archiviste numérique.* Paris ; Bry-sur-Marne : L'Harmattan ; Ina Éditions.

Gagnon-Arguin, L. (1998). « Les questions de recherche comme matériau d'études des usagers en vue du traitement des archives ». *Archivaria* (46): 86-102.

Gauthier, G. (2014). « Des données aux usages: les interfaces de médiation et de narration ». *Documentaliste* 51(2): 58-60.

Gilliland, A. J. (2014). *Conceptualizing Twenty-first-century Archives*. Chicago : Society of American Archivists.

Giuliano, F. (2012). « La référence en archives au XXIe siècle. L'impact du numérique sur le travail de référencier. État des lieux ». *Archives* 43(1): 3-19.

Glady, M. et F. Leimdorfer. (2015). « Usages de la lexicométrie et interprétation sociologique ». *Bulletin of Sociological Methodology / Bulletin de Méthodologie Sociologique* 127(1) : 5-25.

Golder, S. A. et B. A. Huberman (2006). « Usage patterns of collaborative tagging systems ». *Journal of Information Science* 32(2): 198–208.

Grand Dictionnaire. (2017). Office québécois de la langue française. Repéré à http://granddictionnaire.com/

Grand Robert (*Le Grand Robert de la langue française*). (2017). Dictionnaires Le Robert. Accès en ligne procuré par les bibliothèques de l'Université de Montréal.

Green, R. (2006). *Vocabulary Alignment via Basic Level Concepts*. OCLC / ALISE research grant report published electronically by OCLC Research. Repéré à https://www.oclc.org/content/dam/research/grants/reports/green/rg2005.pdf

Greene, M. A. et D. Meissner. (2005). « More Product, Less Process: Revamping Traditional Archival Processing ». *The American Archivist* 68 (2): 208-63.

Gross, G. (2015). « Traitement automatique de la polysémie ». *Studia Romanica Posnaniensia* 42: 15-33.

Guitard, L. A. (2013). « Indexation par sujet en archivistique et en bibliothéconomie : du pareil au même ? ». *Documentation et bibliothèques* 59(4) : 201-212.

———. (2014). « Une entrée d'index thématique : noms communs et références ». *Res per nomen IV, hommage à George Kleiber*. Reims, France : Res per nomen.

———. (2018). « Ça tombe sous le sens : la médiation sémantique de l'archiviste ». Dans Cardin, M. et A. Klein, *Médiation documentaire,* Avec la collaboration de la Chaire pour le développement de la recherche sur la culture d'expression française en Amérique du Nord (CEFAN). Québec : Presses de l'Université Laval, 119-139.

Guitard, L. A. et L. Da Sylva. (2014). *L'opposition dénomination / désignation dans l'index*, Acfas, 2014-05-15, Université Concordia, Montréal. Présentation par Lyne Da Sylva. Basée sur Guitard, L. A. et L. Da Sylva. (2014). L'opposition dénomination / désignation dans l'index. Rapport de recherche. Projet de recherche « Regards croisés sur le lexique : didactique, lexicologie, sciences de l'information et terminologie » Chercheuse principale : Marie-Claude L'Homme. Financement : FRQ-SC (soutien aux équipes de recherche). [document non publié].

H

Hennicke, S. (2011). « Leveraging EAD in a Semantic Web Environment to Enhance the Discovery Experience for the User in Digital Archives ». Dans S. Gradmann, F. Borri, C. Meghini et H. Schuldt (dir.), *Research and Advanced Technology for Digital Libraries* Berlin / Heidelberg, Springer : 511-514.

Héon, G. (1999). « Chapitre 6 : La classification ». *Les fonctions de l'archivistique contemporaine.* Québec, Presses de l'Université du Québec: 219-254.

Hjørland, B. (2007). « Chapter 8: Semantics and knowledge organization », *Annual Review of Information Science and Technology*, 41: 367-405.

———. (2008). « What is Knowledge Organization (KO)? », *Knowledge Organization. International Journal devoted to Concept Theory, Classification, Indexing and Knowledge Representation*, 35: 86-101.

Hoyle Rojas, S. (2016) *I. Mise en place d'une photothèque numérique au sein de la communauté d'agglomération du Grand Cahors. II. Le crowdsourcing à la croisée du web social et des archives : état des lieux, enjeux et perspectives.* Toulouse : Université Toulouse II - Jean Jaurès. Mémoire de master II.

Hudon, M. (1997-1998). « Indexation et langages documentaires dans les milieux archivistiques à l'ère des nouvelles technologies de l'information ». *Archives* 29(1): 75-98.

———. (2009). *Guide pratique pour l'élaboration d'un thésaurus documentaire.* Montréal : Éditions ASTED.

———. (2013). *Analyse et représentation documentaires : introduction à l'indexation, à la classification et à la condensation des documents*. Québec : Presses de l'Université du Québec.

Hudon, M., C. Arsenault, L. Da Sylva et D. Forest. (2009). « Le traitement du document ». Dans Salaün, J.-M. et C. Arsenault (dir.), *Introduction aux sciences de l'information*. Montréal : Presses de l'Université de Montréal : 53-100.

Hudon, M. et W. M. El Hadi. (2017). « Introduction. La classification à facettes revisitée. De la théorie à la pratique ». *Les Cahiers du numérique* 13: 9-24.

Hutchins, W. J. (1975). *Languages of Indexing and Classification: a Linguistic Study of Structures and Functions*. Stevenage, Royaume-Uni : P. Peregrinus.

Huyghe, R. (2015). « Les typologies nominales: présentation ». *Langue française* 185(1): 5-27.

ICA (Conseil international des archives). (2012). *Principes relatifs à l'accès aux archives*. Conseil international des archives = International Council on Archives.

———. (2016). *Research in Contexts. A conceptual Model for Archival Description (RiC-CM).* Exemplaire de consultation. Disponible à https://www.ica.org/site/default/files/RiC-CM-0.1.pdf

ICA (Conseil international des archives). Comité sur les normes de description. (2000). *ISAD(G) : norme générale et internationale de description archivistique* Norme *ISAD(G)*. 2e édition Ottawa : Conseil international des archives = International Council on Archives.

———. (2004). *Norme Internationale sur les notices d'autorité utilisées pour les Archives relatives aux collectivités, aux personnes ou aux familles.* Norme *ISAAR(CPF)* 2e édition. Ottawa : Conseil international des archives.

———. (2008). *Norme internationale pour la description des institutions de conservation des archives.* Norme *ISDIAH* Ottawa : Conseil international des archives.

———. (2011). *Norme de description des fonctions*. Norme *ISDF.* Ottawa : Conseil international des archives.

ISO (Organisation internationale de normalisation). (1985). *Documentation -- Méthodes pour l'analyse des documents, la détermination de leur contenu et la sélection des termes d'indexation.* Norme *ISO 5963*. Genève : Organisation internationale de la normalisation. Réexaminée en 2015.

———. (1985). *Documentation -- Principes directeurs pour l'établissement et le développement de thesaurus multilingues*. Norme *ISO 5964* :1985. Genève : Organisation internationale de la normalisation. Annulée en 2011 et remplacée par ISO 25 964-1:2011 et ISO 25 964-2:2013.

———. (1986). *Documentation -- Principes directeurs pour l'établissement et le développement de thésaurus monolingues*. Norme *ISO 2788* :1986. Genève :

Organisation internationale de la normalisation. Annulée en 2011 et remplacée par ISO 25 964-1:2011 et ISO 25 964-2:2013.

———. (1996). *Information et documentation -- Principes direct Information et documentation -- Principes directeurs pour l'élaboration, la structure et la présentation des index.* Norme *ISO 999*. Genève : Organisation internationale de la normalisation. Réexaminée en 2015.

———. (2011). *Information et documentation -- Thésaurus et interopérabilité avec d'autres vocabulaires -- Partie 1 : Thésaurus pour la recherche documentaire*. Genève : Organisation internationale de la normalisation. Réexaminée en 2017.

———. (2013). Inform*ation et documentation -- Thésaurus et interopérabilité avec d'autres vocabulaires -- Partie 2: Interopérabilité avec d'autres vocabulaires*. Genève : Organisation internationale de la normalisation.

Jacquemin, C. (2001). *Spotting and discovering terms through natural language processing*. Cambridge, Massachussetts : MIT Press.

Jacquemin, C. et P. Zweigenbaum. (2000). « Traitement automatique des langues pour l'accès au contenu des documents ». Dans J. Le Maître, J. Charlet et C. Garbay (dir.). *Le document en sciences du traitement de l'information*. Toulouse, France : Cepadues :71-109.

Jakobson, R. (1963). *Essais de linguistique générale*. Paris : Éditions de Minuit.

Jörgensen, C. (2007). « Image Access, the Semantic Gap, and Social Tagging As a Paradigm Shift ». Dans Joan Lussky (dir.), *Proceedings 18th Workshop of the American Society for Information Science and Technology Special Interest Group in Classification Research*. Repéré à http://dlist.sir.arizona.edu/2064/

K

Kakali, C. et C. Papatheodorou (2010). « Exploitation of Folksonomies in Subject Analysis ». *Library & Information Science Research* 32(3): 192-202.

Kim, Y.-w. (2014). « Typology of Ambiguity on Representation of Information Needs ». *Reference & User Services Quarterly* 53(4): 313-325.

Kipp, M. E. I. (2009). *Information organisation practices on the Web: Tagging and the social organisation of information*. (Thèse de doctorat, University of Western Ontario).

Kister, L., E. Jacquey et B. Gaiffe. (2011). « Du thesaurus à l'onto-terminologie: relations sémantiques vs relations ontologiques », *Corela. Cognition, représentation, langage* 9-1. Repéré à http://corela.revues.org/1962

Khoo, C. S. G. et J.-C. Na. (2006). « Semantic Relations in Information Science », *Annual Review of Information Science and Technology*, 40: 157-228.

Kleiber, G. (1981). *Problèmes de référence : descriptions définies et noms propres*. Metz; Paris : Centre d'analyse syntaxique Klincksieck.

———. (1984). « Dénomination et relations dénominatives ». *Langages* 19(76): 77-94.

———. (1990). *La Sémantique du prototype : catégories et sens lexical*. Paris : Presses universitaires de France.

———. (1994). « Contexte, interprétation et mémoire : approche standard vs approche cognitive ». *Langue française* (103): 9-22.

———. (1996). « Noms propres et noms communs : un problème de dénomination ». *Meta* 41(4): 567.

———. (1997). « Sens, référence et existence : que faire de l'extra-linguistique ? ». *Langages* 127: 9-37.

———. (2003). « Sur la Sémantique de la dénomination ». *Verbum* XXV(1): 97-106.

———. (2009). « La synonymie-« identité de sens » n'est pas un mythe ». *Pratiques. Linguistique, littérature, didactique*: 9-25.

———. (2012). « De la dénomination à la désignation : le paradoxe ontologico-dénominatif des odeurs ». *Langue française* (2): 45-58.

Kristeva, J. (1969). *Semeiotiké : recherches pour une sémanalyse.* Paris : Seuil.

Kyriaki-Manessi, D. et M. Dendrinos. (2014). « Developing Ontology for the University Archives: The Domain of Technological Education ». *Procedia - Social and Behavioral Sciences* 147: 349-359.

La Barre, K. (2010a). « Facet analysis ». *Annual Review of Information Science and Technology* 44: 243-84.

———. (2010b). « A Semantic (Faceted) Web? ». *Les Cahiers du numérique*, 6: 103-31.

Labbé, C. et D. Labbé. (2013) « Lexicométrie : quels outils pour les sciences humaines et sociales ? ». *Usages de la lexicométrie en sociologie*. Guyancourt, France. Repéré à https://hal.inria.fr/file/index/docid/834039/filename/LabbeLabbePrintemps.pdf

Lambert, J. (1999). « Chapitre 5 : L'accroissement (l'acquisition) ». *Les fonctions de l'archivistique contemporaine*. Québec, Presses de l'Université du Québec: 145-218.

Lamoureux, A. (2006). *Recherche et méthodologie en sciences humaines*. Montréal (Québec) : Beauchemin, Chenelière éducation.

Lancaster, F. W. (1986). *Vocabulary control for Information Retrieval*. Arlington, Virg : Information Resources Press.

———. (2003). *Indexing and Abstracting in Theory and Practice*. Champaign, Ill. : University of Illinois.

LeBlanc, C., F. Martineau et Y. Frénette (dir.). (2010). *Vues sur les français d'ici.* Québec : Presses de l'Université Laval.

Leblanc, J. M. (2015). « Proposition de protocole pour l'analyse des données textuelles : pour une démarche expérimentale en lexicométrie ». *Nouvelles perspectives en sciences sociales: Revue internationale de systémique complexe et d'études relationnelles 11*(1), 25-63.

Le Deuff, O. (2006). « Folksonomies. Les usagers indexent le Web ». *Bulletin des bibliothèques de France* (4).

Lee, H.-J. et D. Neal (2010). « A new model for semantic photograph description combining basic levels and user-assigned descriptors ». *Journal of information science* 36(5): 547-565.

Lehmann, A. et F. Martin-Berthet (2013). *Lexicologie : Sémantique, morphologie et lexicographie*. Paris : Armand Colin.

Lemay, Y. et A. Klein (2012). « Archives et émotions ». *Documentation et bibliothèques* 58(1): 5-16.

Lepage, T. et C. de Schaetzen (1997). « Évaluation d'un échantillon d'index ». *Meta* 42(2): 328-346.

Lévesque, M. (2001-2002). « L'indexation : luxe ou nécessité? » *Archives* 33(1): 17-45.

Lezcano, L., E. Garcia-Barriocanal et M.-A. Sicilia. (2012). « Bridging informal tagging and formal semantics via hybrid navigation ». *Journal of Information Science* 38(2): 140-155.

L'Homme, M.-C. (2004). *La terminologie : principes et techniques* Montréal : Presses de l'Université de Montréal.

Light, M. (2008). « The endangerment of trees », allocution présentée à EAD@10, 2008. Repéré à http://www.archivists.org/publications/proceedings/EAD@10/Light-EAD@10.pdf

Loi sur le patrimoine culturel. RLRQ, c. P-9.0002. Repéré à http://legisquebec.gouv.qc.ca/fr/ShowDoc/cs/P-9.002

Loi sur les archives. RLRQ, c. A-21.1. Repéré à http://legisquebec.gouv.qc.ca/fr/ShowDoc/cs/A-21.1

Loiseau, S. (2012). « Théories de la fréquence linguistique et interprétations des faits quantitatifs en sémantique ». *3e Congrès mondial de linguistique française (CMLF)*: 1861-1875.

Longo, L. et A. Todirascu. (2010). « Une étude de corpus pour la détection automatique de thèmes ». *Actes des 6es journées de linguistique de corpus*. 143-155.

———. (2014). « Vers une typologie des chaînes de référence dans des textes administratifs et juridiques ». *Langages* (3)195: 79-98. Repéré à https://www.cairn.info/revue-langages-2014-3-page-79.htm

Lu, C., J.-R. Park et X. Hu. (2010). « User tags versus expert-assigned subject terms: A comparison of LibraryThing tags and Library of Congress Subject Headings ». *Journal of Information Science* 36(6): 763–779.

Lytle, R. H. (1980a). « Intellectual Access to Archives I: Provenance and Content Indexing Methods of Subject Retrieval ». *The American Archivist* 43(1): 64–75.

———. (1980b). « Intellectual Access to Archives II: Report of an Experiment Comparing Provenance and Content Indexing Methods of Subject Retrieval ». *The American Archivist* 43(2): 191–207.

M

MacNeil, H. (2005). « Picking Our Text: Archival Description, Authenticity, and the Archivist as Editor ». *The American Archivist* 68(2): 264-278.

Marandin, J.-M. (1988). « À propos de la notion de thème de discours. Éléments d'analyse dans le récit ». *Langue française* (78): 67-87.

Martin, K. E. (2001). « Analysis of Remote Reference Correspondence at a Large Academic Manuscripts Collection ». *The American Archivist* 64: 17-42.

Marleau, Y., S. Mas et M. Zacklad. (2008). « Exploitation des facettes et des ontologies sémiotiques pour la gestion documentaire » Dans *Actes de la deuxième conférence Document numérique et société « Traitements et pratiques documentaires : vers un changement de paradigme? ».* Paris, France: ADBS. 91-110

Mas, S. (2011). *Classification des documents numériques dans les organismes : impact des pratiques classificatoires personnelles sur le repérage*. Québec : Presses de l'Université du Québec.

———. (2013-2014). « La notion de facettes et son application dans un contexte de recherche dans les fonds d'archives : analyse des questions de recherche et de l'expérience vécue par des usagers novices ». *Archives* 45(1): 85-102.

Maurel, D. et M. Champagne. (1999). « Chapitre 7 : la description et l'indexation ». Dans *Les Fonctions de l'archivistique contemporaine*. Sainte-Foy, Québec : Presses de l'Université du Québec : 255-371.

McCausland, S. (2011). « A Future Without Mediation? Online Access, Archivists, and the Future of Archival Research ». *Australian Academic & Research Libraries* 42: 309-19.

Mejri, S. (2000). « Figement et dénomination ». *Meta* 45: 609-21.

Mémoire du Québec, La. (2018) Communications Cournoyer Inc. Repéré à http://www.memoireduquebec.com/wiki/index.php?title=Accueil

Ménard, N. (1989). « Mesure des relations lexico-sémantiques dans des textes scientifiques: problèmes méthodologiques ». *Meta* 34: 468-78.

Ménard, E., S. Mas et I. Alberts. (2010). « Faceted classification for museum artefacts: A methodology to support web site development of large cultural organizations ». London : *Aslib Proceedings* 62: 523-32.

Miles, M. B. et A. M. Huberman. (1994). « Chapter 10: Making Good Sense: Drawing and Verifying Conclusions ». Dans *Qualitative Data Analysis: an Expanded Sourcebook*. Thousand Oaks, CA: Sage Publications.

Miles, M. B., A. M. Huberman et J. Saldaña. (2014). *Qualitative data analysis: A methods sourcebook.* Los Angeles : Sage.

Milstead, J. (2001). « Standards for Relationships between Subject Indexing Terms ». Dans Carol A. Bean and R. Green (dir.), *Relationships in the Organization of Knowledge* Dordrecht : Springer Netherlands.

Mortureux, M.-F. (1993). « Paradigmes désignationnels ». *Semen. Revue de sémio-linguistique des textes et discours* (8) :1-13.

———. 2008. *La lexicologie entre langue et discours*. Paris : Armand Colin.

Mortureux, M.-F. et G. Petit. 1989. « Fonctionnement du vocabulaire dans la vulgarisation et problèmes de lexique ». *DRLAV. Revue de Linguistique* 41-62.

Mounin, G. (1965). « Un champ sémantique : la dénomination des animaux domestiques ». *La Linguistique* 1: 31-54.

Munn, E. et D. Rioux. (1998). « La référence : une fonction archivistique à part entière ». *Archivaria* 45: 104-11.

Nahuet, R. (2009). « L'archivistique contemporaine à l'âge adulte : pertinence et actualité du respect des fonds ». *Archives* 41(1): 45-60.

Nougaret, C., B. Galland et Direction des archives de France. (1999). *Les instruments de recherche dans les archives*. Paris, Direction des archives de France : Documentation française.

Nouvel, D., M. Ehrmann et S. Rosset. (2016). « Named Entities, Referential Units ». Dans *Named Entities for Computational Linguistics* John Wiley & Sons, Inc.:11-46.

Office québécois de la langue française. (2013). Nétiquette. Dans *Vocabulaire d'Internet – Banque de terminologie du Québec*. Repéré à : http://www.oqlf.gouv.qc.ca/ressources/bibliotheque/dictionnaires/Internet/fiches/2071545.html

Oliver, A., A. Jamieson et A. Daniel. (2017). « Here, there and everywhere: an analysis of reference services in academic archives ». *Archives and manuscripts*: 1-19.

Oomen, J., L. Belice Baltussen, S. Limonard, A. van Ees, M. Brinkerink, L. Aroyo, J. Vervaart, K. Asaf et R. Gligorov. (2010). « Emerging practices in the cultural heritage domain-social tagging of audiovisual heritage ». Dans *Proceedings of Web Science 2010: Extending the Frontiers of Society On-Line.* The Web Science Trust.

P

Peters, I. (2009). *Folksonomies: Indexing and Retrieval in Web 2.0*. Berlin : De Gruyter / Saur.

Petit, G. (2001a). « Dénomination et lexique ». *Journal of French Language Studies* 11(1): 89-121.

———. (2001b). « Pour une conception lexicologique de la dénomination ». *Cahiers de praxématique* 36: 93-115.

———. (2009). *La dénomination : approches lexicologique et terminologique*. Louvain : Éditions Peeters.

———. (2012a). « Présentation : La dénomination ». *Langue française* (174) : 3-9.

———. (2012b). « Pour un réexamen de la notion de dénomination ». *Langue française* (174) : 27-46.

Petit Larousse, Le (Le Petit Larousse illustré). (2018) Éditions Larousse.

Pickard, A. (2013). *Research methods in information*. 2e éd. Chicago : Neal-Schuman.

Poitou, J. (2001). « Prototypes, saillance et typicalité ». *Nouvelles terminologies* (21): 17-26.

Polguère, A. (2008). *Lexicologie et sémantique lexicale*. Montréal : Presses de l'Université de Montréal.

———. (2013). « Les petits soucis ne poussent plus dans le champ lexical des sentiments ». Dans F. Baider et G. Cislaru (dir.), *Cartographie des émotions. Propositions linguistiques et sociolinguistiques*. Paris : Presses Sorbonne Nouvelle.

———. (2016). *Lexicologie et sémantique lexicale : notions fondamentales.* 3e éd. Montréal, Québec : Les Presses de l'Université de Montréal.

Pottier, B. (1963). *Recherches sur l'analyse sémantique en linguistique et en traduction mécanique: Linguistique appliquée et traduction automatique, série A, II*. Nancy : Université de Nancy.

Pugh, M. J. (2005). *Providing reference services for archives and manuscripts*. Chicago : Society of American Archivists.

———. (2009). « Archival Reference and Access ». Dans *Encyclopedia of Library and Information Sciences*, 3e éd. Boca Raton, FL : Taylor & Francis. 162-178.

Ranganathan S. R. (1957). « Library Classification, a discipline ». Conference on Classification for Information Retrieval. London, Royaume Uni : *Aslib* 3-14.

Rey, A. (1972). « Usages, jugements et prescriptions linguistiques ». *Langue française* 16(1): 4-28.

Rorissa, A. (2010). « A Comparative Study of Flickr Tags and Index Terms in a General Image Collection ». *Journal of the American Society for Information Science and Technology* 61(11): 2230-2242.

Rorissa, A. et H. Iyer (2008). « Theories of cognition and image categorization: What category labels reveal about basic level theory ». *Journal of the American Society for Information Science and Technology* 59(9): 1383-1392.

Rosch, E. (1975). « Cognitive Reference Points ». *Cognitive Psychology* 7(4): 532-547.

Rosch, E. et C. B. Mervis (1975). « Family Resemblances: Studies in the Internal Structure of Categories. » *Cognitive Psychology* 7(4): 573-605.

Ruth, J. (1988). « Educating the reference archivist ». *The American Archivist* 51: 266-76.

S

Salaba, A. (2005). *Term Selection Process in Subject Searching: End-user interactions with information retrieval systems and indexing languages* (Thèse de doctorat, University of Wisconsin-Madison).

Salaün, J.-M. et C. Arsenault (2009). *Introduction aux sciences de l'information*. Montréal : Presses de l'Université de Montréal.

Sattar Chaudhry, A. et T. P. Jiun. (2005). « Enhancing access to digital information resources on heritage: A case of development of a taxonomy at the Integrated Museum and Archives System in Singapore ». *Journal of Documentation* 61: 751-76

Saussure, F. de. (1960). *Cours de linguistique générale.* Paris : Payot.

Schaffner, J. (2015). *Making Archival and Special Collections More Accessible*. Dublin, Ohio : OCLC Research. Repéré à http://www.oclc.org/content/dam/research/publications/2015/oclcresearchmaking-special-collections-accessible-2015-a4.pdf

Schnedecker, C. et F. Landragin. (2014). « Les chaînes de référence : présentation ». *Langages* 195(3): 3-22.

Senécal, S. (1998). « La lecture et la description archivistique du document ». *Archives* 29(3 et 4): 49-56.

Skinner, J. (2014). « Metadata in Archival and Cultural Heritage Settings: A Review of the Literature ». *Journal of Library Metadata* 14(1): 52-68.

Smiraglia, R. P. (1990). « Subject Access to Archival Materials Using LCSH ». *Cataloging & Classification Quarterly* 11(3-4): 63-90.

Smith-Yoshimura, K. (2012). *Social Metadata for Libraries, Archives, and Museums: Executive Summary*. Dublin, Ohio (USA) : OCLC Research.

Soergel, D. (1974). *Indexing languages and thesauri: construction and maintenance*. Los Angeles, CA: Melville Publishing Company.

Stewart, B. (2013). *Pictures in words: indexing, folksonomy and representation of subject content in historic photographs* (Thèse de doctorat, Edith Cowan University, Perth, Western Australia).

Stock, W. G. (2010). « Concepts and Semantic Relations in Information Science ». *Journal of the American Society for Information Science and Technology* 61: 1951-69.

Svenonius, E. (1997). « Definitional approaches in the design of classification and thesauri and their implications for retrieval and for automatic classification ». Dans *Proc. Int. Study Conference on Classification Research* : 12-16.

Talmy, L. (2007). *Introspection as a Methodology in Linguistics*. (Conférence plénière). 10th International Cognitive Linguistics Conference (Krakow, July 2007).

Tesch, R. (1990). « Types of Qualitative Analysis ». In *Qualitative Research: Analysis Types and Software Tools.* Bristol : Falmer Press.

Termium Plus. (2017). Bureau de la traduction du Canada. Gouvernement du Canada. Repéré à http://www.btb.termiumplus.gc.ca/tpv2alpha/alpha-fra.html?lang=fra

Theimer, K. (2011). « What Is the Meaning of Archives 2.0? » *The American Archivist* 74(1): 58-68.

Tournier, M. (1980). « D'où viennent les fréquences de vocabulaire ? La lexicométrie et ses modèles ». *Mots* 1(1): 189-209.

Trésor de la langue française informatisé (TLFi). (2017). Centre national de ressources textuelles et lexicales (cnrtl). Repéré à www.atilf.atilf.fr. Accès libre.

Usito (Dictionnaire Usito). (2017). Les Éditions Delisme inc. Repéré à https://www.usito.com/ Accès par abonnement.

V

Veal, R. E. (2011). *A qualitative content analysis of e-mail reference questions* (Thèse de doctorat, Saint Mary's University of Minnesota). Repéré à http://search.proquest.com/docview/1025733889?accountid=12543

Vickery, B. C. (1960). *Faceted classification: A guide to construction and use of special schemes*. London, Royaume-Uni : Aslib.

———. (1971). « Structure and function in retrieval languages ». *Journal of Documentation* 27: 69-82.

Vilar, P. et A. Šauperl (2015). « Archives, Quo Vadis et Cum Quibus? : Archivists' self-perceptions and perceptions of users of contemporary archives ». *International Journal of Information Management* 35(5): 551-560.

Voss, J. (2007) « Tagging, Folksonomy & Co - Renaissance of Manual Indexing? » Rapport technique arXiv:cs/0701072. Repéré à http://arxiv.org/abs/cs/0701072

Waller, S. et C. Masse (1999). *L'analyse documentaire : une approche méthodologique*. Paris : ADBS Éditions.

Weinberg, B. H. (2009). « Indexing: History and Theory ». *Encyclopedia of Library and Information Sciences.* Boca Raton, FL : Taylor and Francis.

Wellisch, H. H. (2000). *Glossary of terminology in abstracting, classification, indexing, and thesaurus construction.* 2e édition. Medford, NJ : Information Today Inc.

Wilmet, M. (2016). « Postface ». *Le français aujourd'hui*, 192(1): 133-146.

Wolfram, D., H. A. Olson, et R. Bloom (2009). « Measuring consistency for multiple taggers using vector space modeling ». *Journal of the American Society for Information Science and Technology* 60(10): 1995-2003.

Yakel, E. (2000). « Thinking inside and outside the boxes: Archival reference services at the turn of the century ». *Archivaria* 49:140-160.

———. (2002). « Listening to users ». *Archival Issues* 26(2) : 111-27.

———. (2004). « Information literacy for primary sources: creating a new paradigm for archival researcher education ». *OCLC Systems & Services: International digital library perspectives* 20: 61-64.

Yakel, E. et D. A. Torres. (2003). « AI: Archival Intelligence and User Expertise', *The American Archivist* 66: 51-78.

Yeo, G. (2008). « Concepts of Record (2): Prototypes and boundary objects ». *The American Archivist* 71(1): 118-143.

Yu, H. et M. Young (2004). « The Impact of Web Search Engines on Subject Searching in OPAC ». *Information Technology and Libraries* 23(4): 168-180.

Zhang, J. (2011). *System evaluation of archival description and access* (Thèse de doctorat, University of Amsterdam).

Zhang, J. (2012). « Archival Representation in the Digital Age ». *Journal of Archival Organization* 10(1): 45-68.

Zhang, J. et D. Mauney (2013). « When Archival Description Meets Digital Object Metadata: A Typological Study of Digital Archival Representation ». *The American Archivist* 76(1): 174-195.

Zhang, Y. et B. M. Wildemuth. 2009. « Chapter 30: Qualitative Analysis of Content ». Dans B. M. Wildemuth (dir.), *Applications of Social Research Methods to Questions in Information and Library Science*. Westport, CN : Libraries Unlimited.

Annexes

Annexe 1 – Définitions

Dans le contexte de notre recherche, les définitions des concepts principaux sont les suivantes.

Archives patrimoniales ou *Documents d'archives patrimoniaux*

Documents d'archives conservés pour leur valeur patrimoniale (synonymes : archives définitives ou historiques). À rapprocher de la définition de *document patrimonial* de la *Loi sur le patrimoine culturel* : « selon le cas, un support sur lequel est portée une information intelligible sous forme de mots, de sons ou d'images, délimitée et structurée de façon tangible ou logique, ou cette information elle-même, qui présente un intérêt pour sa valeur artistique, emblématique, ethnologique, historique, scientifique ou technologique, notamment des archives » (*Loi sur le patrimoine culturel*, P-9.0002).

Écart sémantique

Éloignement relatif au sens entre deux expressions linguistiques. Par exemple, quand le terme qui figure dans un index n'est pas exactement identique aux expressions linguistiques dénotant le concept retenu pour l'indexation (exemples de différences : singulier/pluriel, synonymes, paraphrases ou définition, hyperonyme ou hyponyme), alors il y a nécessairement un écart. Nous présupposons que tout changement dans la forme induit un changement dans le sens de l'expression linguistique. Le seul changement que nous ne prenons pas en compte est la casse : la présence de majuscules ne nous semble *a priori* pas reliée à un écart sémantique.

Filiation sémantique

La filiation sémantique est le repérage des éléments de même sens d'une source à l'autre dans le sens de la chaîne communicationnelle : de la question, à la réponse, à la notice descriptive jusqu'au terme d'indexation. La filiation sémantique permet la filiation thématique qui est le repérage des expressions linguistiques porteuses de sujet (thémanymes) d'une source à l'autre dans le sens de la chaîne communicationnelle (cf 4132).

Polyséquence

Ensemble de séquences qui ont la même question et la même réponse de référence (p. ex. QRNTixNTixNTix).

Propriété linguistique

Propriété des mots de vocabulaire qui est relative à la forme ou au sens. Par exemple, *pomme* est un mot concret, cela renvoie à la propriété linguistique abstrait/concret.

Relation sémantique

Lien de sens entre deux expressions linguistiques. Par exemple, la synonymie (voiture et automobile; clerc et ecclésiastique), la polysémie (café la boisson et café le lieu et café les grains; paroisse le territoire et paroisse l'ensemble de personnes) sont deux relations sémantiques courantes. Les relations sémantiques donnent des noms à des écarts sémantiques récurrents, connus.

Séquence

Ensemble formé par une question, une réponse, une notice descriptive et éventuellement des termes d'indexation rattachés à cette notice (QRNTix dans le corpus 1 ou QRN dans le corpus 2).

Terme de description

Mot ou groupe de mots présent dans une notice descriptive (document secondaire), dénotant un concept ou un sujet et pouvant être indexé ou recherché.

Terme d'indexation

Mot ou groupe de mots représentant un concept ou un sujet présent dans un document d'archives (document primaire) ou une notice descriptive (document secondaire) et servant au repérage de celui-ci. Les termes d'indexation regroupent les termes d'indexation classiques, générés dans un service d'archives détenteur d'archives patrimoniales, et les étiquettes (*tags* en anglais), générées par des usagers.

Terme de recherche

Mot ou groupe de mots présent dans un échange de courriels entre un usager en recherche et un archiviste de référence, dénotant un concept ou un sujet qui peuvent correspondre à un terme d'indexation ou un terme de description.

Thémanyme

Expression linguistique porteuse de sujet. Les thémanymes issus du vocabulaire des usagers ou de celui employé dans les instruments de recherche sont comparés à partir de la filiation sémantique entre les sources.

Vocabulaire pour l'accès thématique aux archives patrimoniales (VATAP)

Ensemble des mots et expressions linguistiques employés pour qu'une recherche par sujet émanant d'un usager puisse aboutir à des documents d'archives. Il comprend le vocabulaire des usagers (courriels d'usagers à la référence et étiquettes) et le vocabulaire des instruments de recherche (notices descriptives et termes d'indexation) établi au préalable par les archivistes de l'institution détentrice des documents d'archives. Cet ensemble de mots couvre la description, l'indexation et la recherche.

Annexe 2 – Procédures envisagées

Au cours de notre recherche doctorale, nous avons envisagé trois procédures de collecte. Nous n'en avons retenu en définitive qu'une seule. Les deux autres procédures envisagées sont présentées ci-dessous.

Procédure envisagée numéro 1

Nous avions envisagé dans un premier temps d'étudier les étiquettes de *Flickr* des trois établissements participants et de remonter jusqu'à la notice descriptive dans leur catalogue respectif en ligne (voir Figure 36 Procédure de collecte envisagée numéro 1).

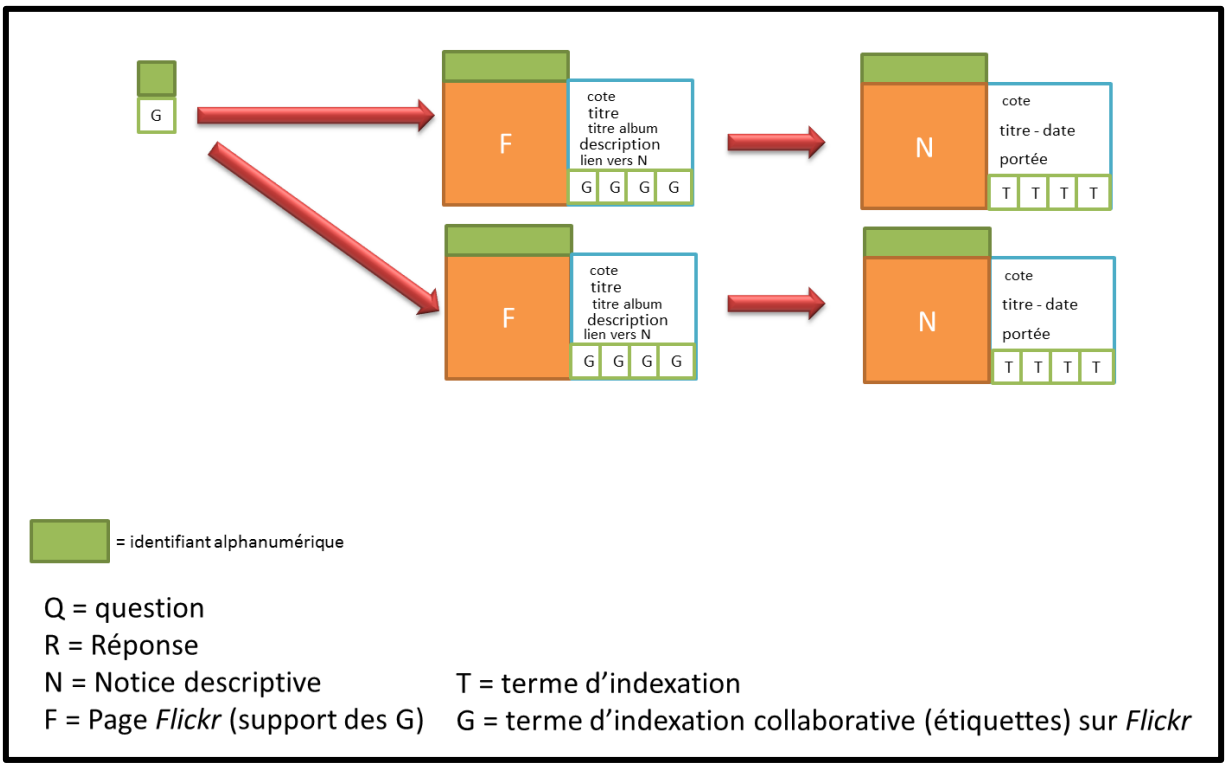

Figure 36. Procédure de collecte envisagée numéro 1

Par exemple, l'étiquette (G) *enfant* est présente dans les étiquettes des trois services d'archives. Rattachée à plusieurs fiches descriptives (F), elle nous aurait permis de remonter vers les notices descriptives (N) correspondantes, elles-mêmes éventuellement pourvues de termes d'indexation (T).

Nous avons établi un rapprochement des étiquettes communes aux trois établissements participants, mais n'avons pas mené l'analyse plus loin pour l'instant. Mais l'aspect technique de l'étude des étiquettes (mots concaténés, sans espace) nous a contrainte à reporter cette procédure de la collecte (voir Conclusion de la thèse).

Procédure envisagée numéro 2

Nous avions envisagé dans un deuxième temps de rechercher les termes d'indexation des listes reçues de deux établissements et de les chercher dans les catalogues en ligne et les albums *Flickr* des trois établissements participants pour collecter les notices descriptives et les fiches descriptives qui leur sont relatives (voir Figure 37). À partir des termes d'indexation auxquels sont parfois attribuée la cote d'une notice descriptive ou plusieurs, nous aurions pu

chercher cette cote dans les albums *Flickr* des participants pour trouver fiches descriptives et étiquettes.

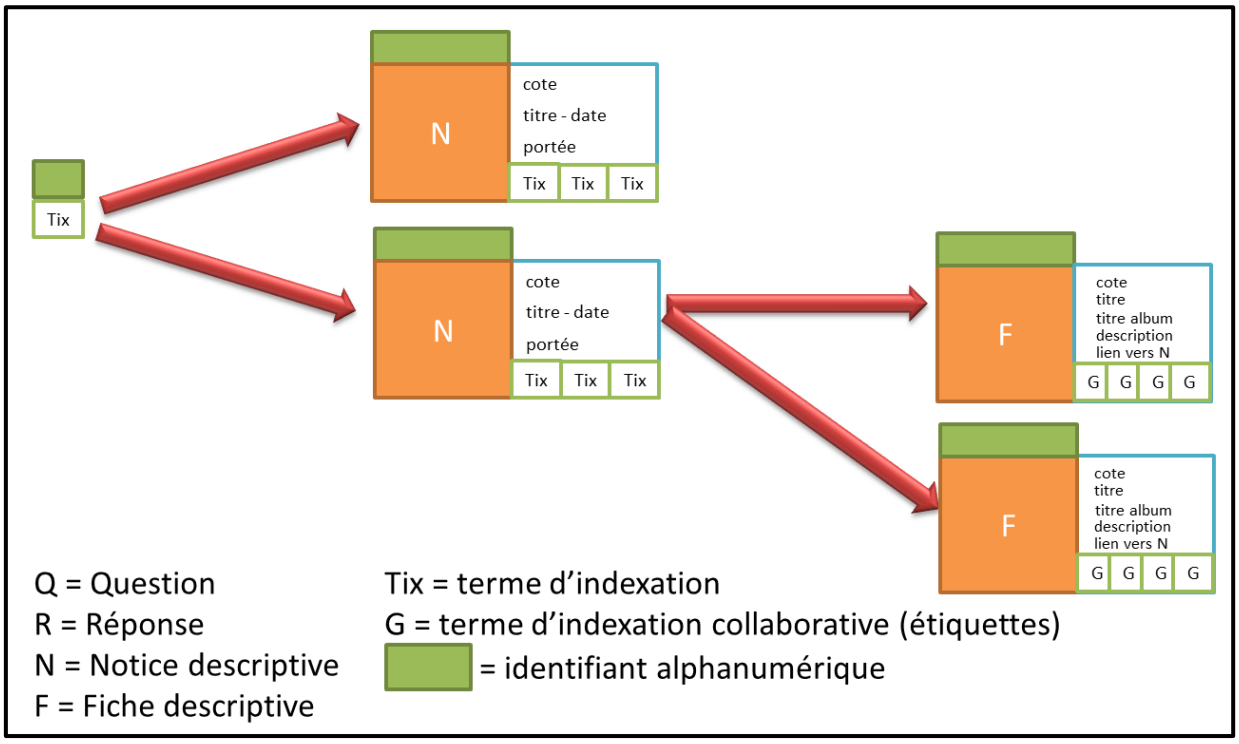

Figure 37. Procédure de collecte envisagée numéro 2

Bien que nous ayons eu à notre disposition les fichiers récapitulant l'ensemble des termes d'indexation utilisés dans deux établissements (voir 4.1.1.2), nous ne les avons pas utilisés dans cette recherche-ci. D'abord, ce moyen de collecte (à partir des termes d'indexation) nous permettait de n'obtenir que des données particulièrement bruitées qui auraient nécessité des prétraitements importants. Aussi, ce moyen de collecte nous permettait de n'obtenir que des notices descriptives et des termes d'indexation –i. e. deux des quatre groupes de données de notre corpus. Finalement, seulement deux des milieux pouvaient nous procurer ce type de données. Ainsi, nous avons décidé de ne pas le retenir. Ici encore, pour des raisons de faisabilité technique et de temps, nous avons dû reporter cette partie de la collecte de données à une éventuelle recherche ultérieure (voir Conclusion de la thèse).

Annexe 3 – Tableau méthodologique

Titre : Vocabulaire employé pour l'accès thématique aux documents d'archives patrimoniaux : étude linguistique exploratoire de termes de recherche, de description et d'indexation

But : Notre recherche porte sur le vocabulaire employé pour l'accès thématique aux documents d'archives patrimoniaux (VATAP). Elle a pour but d'étudier l'écart sémantique présupposé entre le vocabulaire des usagers qui recherchent des documents d'archives patrimoniaux et celui employé par les archivistes dans les instruments de recherche. Ces deux vocabulaires sont constitués d'expressions linguistiques porteuses de sujet; nous appelons ces expressions des thémanymes. Nous voulons vérifier empiriquement l'existence de l'écart sémantique présupposé entre les thémanymes de ces deux vocabulaires, qualifier cet écart selon une approche linguistique et le quantifier.

Question générale : quelles sont la nature et l'ampleur de l'écart sémantique présupposé entre le vocabulaire des usagers qui recherchent des documents d'archives patrimoniaux et le vocabulaire employé par les archivistes dans les instruments de recherche pour l'accès thématique à des documents d'archives patrimoniaux ?

Corpus 1 : QRNTix **Corpus 2 :** QRN

Objectif de recherche	Objectif spécifique	Question de recherche	Méthode d'analyse	Outil de collecte	Outil d'analyse	Résultat
OR 1. Valider empiriquement l'existence d'un écart sémantique présupposé entre le vocabulaire des usagers qui recherchent des documents d'archives patrimoniaux par sujet et le vocabulaire employé par les archivistes dans les instruments de recherche pour l'accès thématique à des documents d'archives patrimoniaux.	*néant*	QR1. Lors de la comparaison du vocabulaire des usagers et celui employé dans les instruments de recherche, des paires de thémanymes entretiennent-ils une relation d'identité ?	Analyse linguistique	Outils de collecte et de sélection des thémanymes : reconnaissance de la facette, identification du champ sémantique	Échelle d'écart sémantique : grille de comparaison de deux expressions du point de vue de la forme et du sens	L'identification des cas de répétition exacte et des cas d'écart

Objectif de recherche	Objectif spécifique	Question de recherche	Méthode d'analyse	Outil de collecte	Outil d'analyse	Résultat
OR 2. Qualifier l'écart sémantique entre le vocabulaire des usagers qui recherchent des documents d'archives patrimoniaux par sujet et le vocabulaire employé dans les instruments de recherche pour l'accès thématique à des documents d'archives patrimoniaux.	a. Caractériser les relations sémantiques entre les thémanymes des deux vocabulaires.	QR2. Quelles sont les relations sémantiques qui existent entre les thémanymes appartenant au vocabulaire des usagers et ceux appartenant à celui des instruments de recherche ?	Analyse linguistique	Échelle d'écart sémantique : outil de comparaison des expressions linguistiques entre sources	Échelle d'écart sémantique : grille d'analyse linguistique de deux expressions du point de vue de la forme et du sens	Un relevé des relations sémantiques entre les thémanymes utilisés dans la recherche et dans les la description et l'indexation (p. ex. la variation de casse, la paraphrase, l'hyperonymie)
OR 3. Quantifier l'écart sémantique entre le vocabulaire des usagers qui recherchent des documents d'archives patrimoniaux par sujet et le vocabulaire employé dans les instruments de recherche pour l'accès thématique à des documents d'archives patrimoniaux	a. Calculer la fréquence de correspondance et de non-correspondance entre les thémanymes des deux vocabulaires.	QR3. Quelle est la fréquence de la relation d'identité entre les thémanymes appartenant au vocabulaire des usagers et ceux appartenant aux instruments de recherche ?	Analyse de la fréquence	QDA Miner	QDA Miner	Des tableaux récapitulatifs du nombre d'occurrences de l'écart sémantique et sa fréquence exprimée en pourcentage par rapport au corpus 1, 2 ou général
OR 3. *Idem*	b. Calculer la fréquence de chaque type d'écart sémantique entre les thémanymes des deux vocabulaires	QR4. Quelle est la fréquence de chacune des relations sémantiques qui caractérisent un écart sémantique entre les thémanymes appartenant au vocabulaire des usagers et ceux appartenant à vocabulaire employé dans les instruments de recherche ?	Analyse de la fréquence	QDA Miner	QDA Miner	Des tableaux récapitulatifs du nombre d'occurrences des relations sémantiques et leur fréquence exprimée en pourcentage par rapport au corpus 1, 2 ou général

Annexe 4 – Protocole de test du codage par facettes

Durée

Les explications du test et les 2 exercices préparatoires prennent environ 15 minutes.
Le codage des cas du prétest prend environ 1 heure.

Consignes

- Ne coder que les segments **en français**
- Ne pas coder les déterminants
- Si le **segment est répété** à plusieurs reprise dans un même cas, le **coder à chaque fois**
- Vous pouvez coder **un même segment** deux fois **avec deux codes différents** ou coder une partie de segment différemment de l'expression complète (P. ex. *Ministère de l'Éducation du Québec* = entité + *Québec* = lieu)
- S'il y a coordination entre 2 segments de même code, coder séparément les 2 segments. P. ex. dans son lit (lieu) et dans le fauteuil (lieu).
- Ne pas coder les mots *USGR, ARVST, SERVICE ARCHIVES, Objet*, qui ont été laissés pour comprendre les données.
- Les codes « contexte de la recherche » et « stratégie de la recherche » englobent de longs éléments sur lesquels il n'est pas *a priori* utile d'apposer d'autres codes.
- Il faut analyser le sens du mot dans son contexte. p. ex. la couture comme activité est à coder dans "action", mais une couture, le résultat de l'action, le produit, est à coder dans "entité". La liste des termes d'indexation (donc des mots hors contexte) est difficile à coder : apposez autant de codes que vous le sentez nécessaire.
- Le code linguistique / expression du sujet est facultatif pour les intercodeurs, il n'est appliqué systématiquement que par la chercheuse elle-même.

Code

CATÉGORIE **code**	**Description et exemples**
01_PERSONNALITÉ entité	Personne physique ou morale et les regroupements de personnes (p. ex. *association, comité, union*), animal, végétal, chose concrète ou abstraite (p. ex. *établissement, édifice, chemin de fer, hôtel*), discipline scientifique (p. ex. *histoire*, *histoire de l'art (une seule entité)*) P. ex. *Bourrassa, henri Bourrassa, Monsieur Bourrassa, M. Bourrassa, la famille Bourrassa*, *les alouettes*, *les ponts*, *le pont Jacques Cartier* (une seule entité), *l'hippodrome Bleu Bonnets* (une seule entité versus *l'hippodrome de Montréal* (entité) + (lieu)), *l'hippodrome Blue Bonnets Inc* (une seule entité), *un peuplier*, *les boîtes à musique, un pont, la mode, un prix*. Y compris les noms des établissements d'enseignement comme personne morale, p. ex. *le collège a recruté trois peintres* (s'ils sont envisagés comme lieu, p. ex. *dans le collège*, alors ils seront codés avec (lieu))

CATÉGORIE **code**	**Description et exemples**
02_MATIÈRE fonction ou attribut	Fonction d'une entité (p. ex. titre d'un emploi; y compris les relations de famille), attribut en tous genres (p. ex. *potentiel, populaire, administratif ou d'administration, toujours en activité; en bois, en français*) Y compris le modèle d'un appareil p. ex. *un missile* (entité) *B52* (matière)), les relations de famille, un état. Le segment est parfois employé en remplacement de l'entité même : p. ex., *une femme* (entité) *présidente* (fonction) => *une présidente* (fonction); *des personnalités* (entité) *locales* (lieu).
03_ÉNERGIE action ou activité	Action subie ou effectuée, généralement par une entité; activité. Plusieurs catégories grammaticales acceptées : - verbe : *diffuser* - adjectif ou participe passé : *diffusé* - nom : *diffusion, activités, implantation, évaluation, courses, effet, conséquences, aide, service, étapes*, *présence, tenue* (action) *de discussions* (action). Y compris les sports, les arts, les activités (p. ex. *cuisine, couture*), les maladies P. ex. *une émission de radio* (au sens de produit fini = entité) *diffusée sur* (action) *ondes courtes* (entité). *La remise* (action) *de prix* (entité).
04_ESPACE lieu	Évocation d'un lieu, associé à un nom propre ou pas (p. ex. *le jardin, la place Bonaventure, la rue Laurier*)
05_TEMPS date ou période	Évocation du temps par une date, un seul point dans le temps (p. ex. *1915*) ou par une période, une durée ou un intervalle (p. ex. *1903-1915* ou *à partir de 1940*, *le 19e s.*) Y compris la mention de temps relative à une personne ou à une action : *une nouvelle* (temps) *voiture* (entité); *une ancienne* (temps) *secrétaire* (fonction); *champions* (fonction) *juniors* (temps)
05_TEMPS événement	Évocation du temps par un événement nommé (p. ex. la 2e Guerre mondiale) ou relatif (p. ex. *son enfance*, par contre, *ses études* sera codé par (action) car il y a une activité clairement indiquée) P. ex. *fête, événement, accident, séance, spectacle, attractions foraines* Un événement présuppose la conjonction d'au moins 2 des 4 éléments suivants : lieu, temps (date ou période), action et entité (personne ou groupe de personnes, etc.).
06_DOCUMENT type, support ou cote	Mention du type de document voulu (p. ex. photographie, vidéo ou image en mouvement, textuel; catégories des *RDDA*) ou du genre de document (discours, reportage, lettre, projet, sommaire, etc.). Y compris les caractéristiques matérielles et l'étendue (p. ex. 3 photos de 10x10cm). Y compris les caractéristiques matérielles et l'étendue. Y compris la mention du type de support (p. ex. numérique, analogique ou papier) Ou simplement d'expressions comme *informations* ou *renseignements* ou encore *documents, copies, dossiers, dossiers,* etc. Y compris chaque cote archivistique distincte, complète ou partielle : E6, S7, SS1, D1, P8189 ou P22 ou encore P12, D21 et P55, P6. S'il y a une suite de pièces ou de dossiers indiquée, n'apposer qu'une seule cote, p. ex. E6, S1, SS7, D1120 à 1125.

CATÉGORIE **code**	**Description et exemples**
07_RECHERCHE contexte de la recherche	Évocation du contexte de la recherche, du but de la recherche, du cadre ou du projet dans lequel prend place la recherche de l'usager et autres informations contextuelles sur le sujet de recherche. P. ex. *Dans le cadre de mon doctorat sur l'histoire de l'art* (contexte de recherche), je recherche des *informations* (document) *sur* (linguistique) les *ingrédients* (entité) des *peintures* (entité) *au 18e s* (temps).
07_RECHERCHE stratégie de recherche	Avec succès ou échec, stratégie de recherche indiquée par l'usager ou l'archiviste Y compris des éléments de réponse de l'archiviste tels que les liens URL vers des ressources ou des références bibliographiques ou l'invitation à aller voir un autre service d'archives plus susceptible d'avoir des documents intéressants. Y compris les réflexions relatives à la description ou au classement Y compris la mention d'instruments de recherche. Y compris les consignes pour accéder aux documents ou la question des droits (utilisation, reproduction). Y compris la demande de confirmation de l'usager que le service contient de tels documents. P. ex. Je cherche des photos (type de doc) de Mme Y (entité). J'ai regardé dans le fonds sur M. X, mais je n'ai pas trouvé de photos de sa femme (stratégie de recherche). Par contre, j'en ai trouvé dans la collection (type de document) des femmes célèbres (entité) (stratégie de recherche).
08_LINGUISTIQUE expression du sujet	Expression linguistique introduisant un sujet de recherche ou le sujet de documents. P. ex.: *sur, portant sur, à propos de, relatif à, en lien avec, concerner*. **! Facultatif !**

Merci de votre participation !

Annexe 5 – Extraits de textes de loi relatifs à notre recherche

La *Loi sur la protection des renseignements personnels dans le secteur privé* stipule :

« 1°l'usage projeté n'est pas frivole et que les fins recherchées ne peuvent être atteintes que si les renseignements sont communiqués sous une forme permettant d'identifier les personnes;

2°les renseignements seront utilisés d'une manière qui en assure le caractère confidentiel. » (LRQ P-39.1, art. 21)

Les renseignements personnels sont définis par les lois fédérale et provinciales :

Loi sur la protection des renseignements personnels :
« Les renseignements, quels que soient leur forme et leur support, concernant un individu identifiable, notamment :
a) les renseignements relatifs à sa race, à son origine nationale ou ethnique, à sa couleur, à sa religion, à son âge ou à sa situation de famille;
b) les renseignements relatifs à son éducation, à son dossier médical, à son casier judiciaire, à ses antécédents professionnels ou à des opérations financières auxquelles il a participé;
c) tout numéro ou symbole, ou toute autre indication identificatrice, qui lui est propre;
d) son adresse, ses empreintes digitales ou son groupe sanguin;
e) ses opinions ou ses idées personnelles, à l'exclusion de celles qui portent sur un autre individu ou sur une proposition de subvention, de récompense ou de prix à octroyer à un autre individu par une institution fédérale, ou subdivision de celle-ci visée par règlement;
f) toute correspondance de nature, implicitement ou explicitement, privée ou confidentielle envoyée par lui à une institution fédérale, ainsi que les réponses de l'institution dans la mesure où elles révèlent le contenu de la correspondance de l'expéditeur;
g) les idées ou opinions d'autrui sur lui;
h) les idées ou opinions d'un autre individu qui portent sur une proposition de subvention, de récompense ou de prix à lui octroyer par une institution, ou subdivision de celle-ci, visée à l'alinéa e), à l'exclusion du nom de cet autre individu si ce nom est mentionné avec les idées ou opinions;
i) son nom lorsque celui-ci est mentionné avec d'autres renseignements personnels le

concernant ou lorsque la seule divulgation du nom révélerait des renseignements à son sujet; » (LRC P-21, section Définitions)

Loi sur l'accès aux documents des organismes publics et sur la protection des renseignements personnels :

« Dans un document, sont personnels les renseignements qui concernent une personne physique et permettent de l'identifier. » (article 54) et « Le nom d'une personne physique n'est pas un renseignement personnel, sauf lorsqu'il est mentionné avec un autre renseignement la concernant ou lorsque sa seule mention révélerait un renseignement personnel concernant cette personne. » (article 56) et encore « Le fait qu'une signature apparaisse au bas d'un document n'a pas pour effet de rendre personnels les renseignements qui y apparaissent » (LRQ A-2.1, article 58)

Loi sur la protection des renseignements personnels dans le secteur privé :

« Est un renseignement personnel, tout renseignement qui concerne une personne physique et permet de l'identifier » (LRQ P-39.1, section I, article 2).

Références

Lois fédérales

Loi sur la protection des renseignements personnels. LRC. (1985). c. P-21. Disponible à http://laws-lois.justice.gc.ca/fra/lois/p-21/

Lois provinciales

Loi sur l'accès aux documents des organismes publics et sur la protection des renseignements personnels. RLRQ, c. A-2.1. Disponible à http://legisquebec.gouv.qc.ca/fr/ShowDoc/cs/A-2.1/

Loi sur la protection des renseignements personnels dans le secteur privé. RLRQ, c. P-39.1. Disponible à http://legisquebec.gouv.qc.ca/fr/showDoc/cs/P-39.1?&digest=

Annexe 6 – Quelques exemples des relations sémantiques de l'échelle d'écart sémantique

Le tableau suivant présente quelques exemples des relations sémantiques identifiées par l'échelle d'écart sémantique que nous avons développée.

Écart	Thémanyme 1	Thémanyme 2	Précision de la relation
0	anatomie	anatomie	
0	Prix des communautés culturelles	Prix des communautés culturelles	
1	les Chinois et Canadiens d'origine chinoise	les Chinois et Canadiens d'origine chinoisent	faute d'orthographe
2	le 22e	22ième bataillon	abréviation (pronom ?)
2	le marché Ste-Anne	le marché Sainte-Anne	abrév/long
2	le CN	la Compagnie des chemins de fer nationaux du Canada	court/long
2	les Jeux Olympiques de 76	les JO	long/court
3	la ville de Tracadie au Nouveau-Brunswick	Ville de Tracadie, Nouveau Brunswick	prép/virgule
3	la Grosse-Île	au Grosse Isle	NP: genre?, trait d'union (3), orthog diachronique (1)
4	la Brasserie Molson	la brasserie Molson	maj/min
4	Pêcheries	pêcheries	maj/min
4	santé	Santé	min/maj
4	Cercles de Fermières	des cercles de fermières	maj maj/min min
4	cet inspecteur d'anatomie	l'Inspecteur d'anatomie	min min/maj min
4	la ville de Tracadie, Nouveau-Brunswick	Ville de Tracadie, Nouveau Brunswick	min maj/maj maj
5	fourrure	Fourrures	sg/pl et min/maj (4)
5	la date et le lieu de l'assermentation des ministres et des premiers ministres depuis la confédération	des dates et des lieux d'assermentation des ministres et du premier ministre depuis la Confédération	sg/pl (de la tête et d'un des compléments)
6	l'Immigration	immigrants	mots de la même famille ; action/agent
6	Radio	Émissions radiophoniques	N/adj
6	Ce navire est parti de La Rochelle le 13 mai	départ	partir/départ
6	ont pu voter	les votes	verbe/N déverbal
6	les troupes canadiennes	les Canadiens	N adj/adj nominalisé
7	Cercles de Fermières	Cercles des fermières	var de dét et maj maj/maj min 4
7	Montréal	la Ville de Montréal	redondance court/long
7	le Fonds du Service des parcs, jardins et espaces verts	Fonds Service des parcs, jardins et espaces verts	N dét N/N N

Écart	Thémanyme 1	Thémanyme 2	Précision de la relation
8	Adolphe V. Roy a été envoyé étudier les mines du Yukon (autre version: de l'Alaska) en 1894 et 1895	qu'il fut envoyé là-bas par le gouvernement du Québec	ajout d'un complément d'agent et anaphore (nom complet/reprise par un pronom)
8	l'Exposition Universelle et Internationale de Paris de 1900	l'exposition	long/court
8	des zoos de Montréal	le Jardin zoologique	abrév (2) avec compl/long sasn compl
9	Adolphe V. Roy a été envoyé étudier les mines du Yukon (autre version: de l'Alaska) en 1894 et 1895	ce mandat	explicitation/dénomination qui recouvre la phrase
9	une transaction ayant eu cours au milieu des années 40 (1946 je pense) entre le Canada et l'Argentine	l'exportation de castors en Patagonie en 1946	Voir déf de dico
10	envoyer 25 couples de castors en Patagonie - Terre de feu	l'exportation de castors en Patagonie en 1946	dériv (6) V/N syn
10	colonies	les différentes régions ayant fait partie de la colonie de la Nouvelle-France	
10	la formation du 22e	constitution	
10	la formation du 22e	création	
10	le gouvernement provincial	le gouvernement du Québec	provincial/du Québec
11	Pelleteries	Fourrure	hyper/hypo (toutes les peaux n'ont pas de fourrure)
11	Cercles de Fermières	l'Union catholique des femmes rurales (UCFR)	co-instances de la classe des assos féminines
11	femmes	Personnes	hypo/hyper
11	les zoos	des parcs	hypo/hyper
11	installations (ens des objets, bâtiments Pti Rob)	Hôpitaux	hyper très large/hypo
11	réparations	Construction	co-hypo de travaux
11	suffrage	Suffrage féminin	hyper/hypo
11	les Chinois et Canadiens d'origine chinoise	Groupes ethnoculturels	hypo/hyper
11	la Ville	Service des parcs	méronymie tout/partie
12	le regroupement Le Cercle des fermières du Québec	Cercles de fermières du Québec	instance/classe
12	Immigration	réfugiés	action/acteur
12	ministère	sous-ministres	lieu/acteur
12	navires	la Marine	objet/discipline
12	fourrure	Chasse	objet/activité

Écart	Thémanyme 1	Thémanyme 2	Précision de la relation
12	l'équipe de William Ogilvie ("Dominion Surveyor')	les arpenteurs canadiens envoyés au Yukon pour la question des frontières	instance/classe arpenteur
12	la frontière avec les USA	la frontière alaskienne	tout/partie
12	le Premier ministre	la carrière de premier ministre de Brian Mulroney	classe/redondance (7) instance
12	le Premier ministre	sa carrière politique	méronymie partie/tout 1er min est un des degrés d'une carrière politique
13	le gouvernement du Canada		politique générale
13	colonies	découvertes géographiques	lieu/acteur
13	ont pu voter	le suffrage	action/droit
13	Pêcheries	la teinte [de fourrure]	lieu et activité dans un même domaine (chasse et pêche)
13	retour en politique en 1989	maire	action-temps/fonction dans un même domaine (politique)
13	le gouvernement fédéral ou provincial	des communications officielles entre le gouvernement fédéral et les gouvernements provinciaux	entité vague/action d'entités précises
13	les irlandais	la Société Saint-Patrice de Montréal	peuple/association ethnique