

L'enrichissement automatique de l'indexation dans le réseau Renouvaud

Michael Hertig

Introduction

L'indexation matière consiste à décrire le contenu d'une ressource documentaire à l'aide de mots-clés, ou termes. Dans un catalogue de bibliothèque, elle sert à améliorer la recherche d'information. Traditionnellement dans les bibliothèques, l'indexation matière se base sur des vocabulaires contrôlés qui déterminent les termes que l'on peut utiliser pour décrire une ressource.

A l'heure actuelle, l'indexation matière contrôlée est remise en question. L'indexation elle-même et l'entretien d'un vocabulaire contrôlé sont des pratiques chronophages et coûteuses. Si une organisation bibliographique décide malgré tout de maintenir un programme d'indexation, elle doit alors trouver de nouvelles pistes pour surmonter les difficultés qui lui sont liées. L'enrichissement automatique de l'indexation est une solution possible. Cette pratique consiste à ajouter de manière automatique des informations à une ressource documentaire concernant son contenu. Depuis quelques années, des projets d'enrichissement automatique des données sont mis en place, notamment en France, en Allemagne ou encore en Suisse.

Cette contribution vise à décrire l'enrichissement automatique dans ses différentes variantes, et à évaluer son applicabilité aux données du réseau vaudois de bibliothèque Renouvaud. Nous allons tout d'abord présenter la notion d'indexation matière et les problèmes rencontrés quant à sa mise en œuvre. En particulier, nous allons mettre en avant la surabondance des ressources à traiter à l'heure actuelle, la diversité de leurs sources et l'hétérogénéité des données d'indexation que cela implique. L'enrichissement automatique de l'indexation sera ainsi introduit dans ce contexte comme solution à ces problèmes. Nous montrerons qu'en fonction des données à traiter, la méthode d'enrichissement appropriée peut varier. Enfin, nous examinerons les données d'indexation de Renouvaud pour évaluer quel type d'enrichissement automatique serait approprié.

Indexation matière et vocabulaires contrôlés

En bibliothéconomie, le terme d'indexation est souvent utilisé comme abréviation de « indexation matière », autrement dit indexation selon le sujet. Il s'agit d'un aspect du

traitement intellectuel du document, qui cherche à donner une représentation du contenu, ce dont traite le document.¹ Selon la définition de Salaün et Arsenault :

« L'indexation est l'opération qui permet de décrire et de caractériser le contenu thématique d'un document à l'aide de représentations verbales. » (Salaün & Arsenault, 2010, p. 81)

L'indexation matière améliore le signalement thématique des ressources et permet la recherche documentaire selon le sujet. En effet, un index est avant tout un outil de recherche d'information (Cleveland & Cleveland, 2013, p. 11). En produisant la représentation d'une information, l'indexation l'accompagne de sa *localisation* dans un ensemble. Ainsi le terme de l'index, en plus d'exprimer l'information à laquelle il renvoie, permet d'accéder à cette information dans l'ensemble d'origine. L'indexation matière donne ainsi un point d'accès *thématique* aux ressources documentaires.

Il existe deux méthodes pour ajouter des mots-clés à une ressource. La première laisse la liberté à l'indexeur de choisir les termes qu'il souhaite utiliser. On parle alors « d'indexation libre ». La seconde au contraire le contraint à puiser les mots-clés dans un répertoire de termes prédéfinis. On parle alors « d'indexation contrôlée », et ce répertoire est appelé un « vocabulaire contrôlé » (Bawden, 2012, p. 106).

Un vocabulaire contrôlé consiste en une liste de termes dont la forme est normalisée et la signification précisée. Par exemple, le terme contrôlé « Souris (informatique) » a une forme principale qui permet de distinguer l'accessoire informatique du petit rongeur. Ces termes sont définis par des autorités, qui déterminent leur forme principale ainsi que des variantes, établissent des renvois vers d'autres termes et citent des sources. Les vocabulaires contrôlés peuvent avoir un degré d'organisation plus ou moins élevé. Il peut s'agir d'une simple liste de termes, avec quelques renvois ou alors d'un thésaurus, où les termes sont organisés thématiquement et hiérarchiquement (Bawden, 2012, p. 120-121; Bertram, 2005, p. 130-131). Les vocabulaires contrôlés les plus courants dans le domaine des bibliothèques sont LCSH, GND, MeSH, ou encore RAMEAU.² Tous ces vocabulaires constituent des référentiels qui vont structurer les données bibliographiques.

La création de vocabulaires contrôlés répond à des problèmes posés par l'indexation libre. L'usage libre des termes dans un système de recherche d'information

1 En allemand, on parle de la *Sacherschliessung* ou *Inhalterschliessung*, opposée à la formale *Erschliessung*, le traitement documentaire formel qui détermine des caractéristiques formelles des ressources telles que l'intitulé, le créateur ou encore la date de création.

2 Ces vocabulaires sont gérés par des institutions prescriptrices qui les maintiennent. LCSH (Library of Congress Subject Headings) est le vocabulaire entretenu par la Library of Congress (LoC) ; GND (Gemeinsame Normdatei) est géré par la Deutsche Nationalbibliothek (DNB) ; MeSH (Medical Subject Headings) par la National Library of Medicine (USA) ; RAMEAU (Répertoire d'autorité-matière encyclopédique et alphabétique unifié) par la Bibliothèque nationale de France (BnF).

conduit en effet à la multiplication des entrées d'un index faisant référence à la même notion et donc à une efficacité moindre de la recherche d'information.³

L'indexation matière contrôlée est traditionnellement une activité manuelle (on parle souvent de « traitement intellectuel » par opposition au « traitement automatique »). C'est aussi une activité chronophage qui requiert de grands moyens en ressources humaines. Elle requiert des compétences avancées, comme la maîtrise d'un lexique spécialisé, l'analyse de contenus et la synthèse de celle-ci en termes contrôlés. Pour ces raisons, l'indexation matière est remise en question.⁴ L'indexation telle qu'elle est pratiquée traditionnellement ne répond plus aux défis actuels, en particulier en terme du nombre de ressources à traiter.

Surabondance et hétérogénéité

L'information croît de manière exponentielle. Le monde de l'édition produit toujours plus de ressources, sans parler des ressources créées sur le web. En particulier, la digitalisation de l'édition, surtout scientifique, a permis l'augmentation importante du nombre de ressources disponibles (notamment avec la publication des notices d'articles, et l'acquisition par les bibliothèques de livres numériques par « paquets »).

Avec la multiplication des ressources documentaires, les bibliothèques n'arrivent plus à gérer le traitement intellectuel. Avant l'avènement des nouvelles technologies d'échange d'information (internet et les protocoles d'échange), une bibliothèque pouvait prétendre traiter de manière exhaustive toutes les ressources qu'elle acquérait et mettait à la disposition de ses usagers. Aujourd'hui, cela n'est plus possible. Le nombre de ressources provenant des grands éditeurs dépassent les forces de travail des bibliothèques.⁵ De plus, les technologies actuelles permettent d'échanger facilement des données bibliographiques et donc de l'indexation matière. L'indexation systématique de chaque ressource par une seule bibliothèque perd de son sens, quand on sait que l'information a déjà été produite ailleurs et qu'elle pourrait être réutilisée.

La conséquence de cette surabondance est que les bibliothèques se retrouvent avec des ressources indexées de façon hétérogène. Certaines ressources sont indexées

3 Pour plus de détails sur les problèmes posés par l'indexation libre et comment les vocabulaires contrôlés viennent les résoudre, cf. (Gross, Taylor, & Joudrey, 2015; Kempf, 2013).

4 Les vocabulaires contrôlés eux-mêmes sont critiqués à l'heure actuelle. Il s'agit d'une pratique coûteuse, en concurrence avec des solutions de traitement automatique de l'information. A la fin des années 2010, un grand débat sur l'utilité des vocabulaires contrôlés a eu lieu. Malgré tout, les grandes institutions dans le monde des bibliothèques persistent à maintenir ce genre de référentiels. Cf. (Gross et al., 2015).

5 Le nombre de ressources proposées par les grands éditeurs et agrégateurs de contenus s'élève à plusieurs dizaines, voire centaines, de milliers. Par exemple, les cinq plus gros fournisseurs de ressources numériques du réseau Renouvaud (Elsevier ScienceDirect, JSTOR, Springerlink, Wiley Online Library et Taylor & Francis Online) proposent à eux seuls près de 27 millions de documents (Hertig, 2018, p. 48-51).

de façon méticuleuse, d'autres selon un vocabulaire libre, d'autre pas du tout. Ces indexations dépendent d'une multiplicité de référentiels, dont un grand nombre, principalement pour les ressources numériques, ne sont pas accessibles librement.⁶ Les bibliothèques sont contraintes de mettre à disposition des ressources qu'elles n'ont pas pu traiter selon les normes et les standards qu'elles préconisent.

L'hétérogénéité des vocabulaires entraîne des problèmes majeurs tant pour l'utilisateur que pour le professionnel de l'information documentaire. Du point de vue de l'utilisateur, la conséquence est l'absence de point d'accès unique pour interroger l'ensemble des ressources à partir d'une recherche de sujets matières. Un utilisateur faisant une recherche thématique ne peut pas interroger l'ensemble des ressources en une seule requête. Se familiariser avec un vocabulaire d'indexation perd alors de son intérêt puisque ce vocabulaire ne couvre pas l'intégralité des ressources. Ce problème est d'autant plus aigu avec le multilinguisme des vocabulaires d'indexation. Une recherche en français doit être répétée en anglais et possiblement dans d'autres langues encore.

Du point de vue de l'indexeur, la situation est également insatisfaisante. En effet, actuellement le moyen le plus direct de donner un point d'accès thématique unique est d'indexer l'ensemble des ressources dans un seul vocabulaire. Or la surabondance des ressources documentaires rend cet usage impraticable. En outre, l'indexation systématique dans un seul vocabulaire entraîne que l'indexation est souvent redondante puisqu'il faut indexer une ressource comportant déjà une indexation dans un autre vocabulaire. Une indexation LCSH d'un document doit être suppléée par une indexation RAMEAU alors que ce sont souvent les mêmes concepts que l'on fait ressortir, dans des langues différentes. L'indexation dans ces conditions est une activité chronophage qui présente peu de plus-value.

La situation actuelle présente ainsi une multiplicité désorganisée de vocabulaires contrôlés, à laquelle les acteurs concernés ne peuvent faire face s'ils se limitent à leurs pratiques traditionnelles. Le manque de moyens pour l'indexation intellectuelle des ressources se fait ressentir un peu partout. En France, l'Agence Bibliographique de l'Enseignement Supérieur (ABES) fait le constat, dans un rapport de 2013, de l'hétérogénéité grandissante des sources de données à agréger et de la surabondance des ressources (ABES, 2013, p. 4). Le manque de temps pour l'indexation intellectuelle est mis en avant comme élément déclencheur du projet « Digitale Assistent » à la Zentralbibliothek de Zurich (ZB), qui met en place une solution d'indexation semi-

6 En effet, un éditeur ou un fournisseur peut utiliser son propre référentiel et n'est souvent pas transparent sur le vocabulaire ou thésaurus utilisé. Par exemple, Taylor & Francis propose une navigation à partir des mots-clés affichés dans les notices, ce qui laisse penser qu'ils utilisent un vocabulaire contrôlé. Mais on ne trouve aucune information sur un tel vocabulaire dans les pages web de l'éditeur (cf. e.g. la ressource https://www.tandfonline.com/doi/abs/10.1300/J104v43n03_03 - consultée le 26.06.2019). On observe la même chose chez JSTOR (Hertig, 2018, p. 48-51).

automatique (Malits & Schäuble, 2014, p. 132-133).⁷ Le rapport stratégique 2017-2020 de la DNB annonce également la mise en place de processus automatisés et basés sur l'importation de données externes (Deutsche Nationalbibliothek, 2017, p. 9).

Quelles solutions sont envisageables pour résoudre ces problèmes de surabondance des ressources et d'hétérogénéité des référentiels ? Une solution radicale serait d'affirmer que l'indexation matière est superflue et d'abandonner tout bonnement cette pratique.⁸ Une autre solution est l'indexation à la source, qui part du principe que le traitement intellectuel doit se faire par son auteur, au moment de sa création. La ressource est donc décrite une seule fois et l'indexation peut être réutilisée ailleurs. On peut encore aller plus loin et admettre que les usagers eux-mêmes prennent en charge le traitement intellectuel des ressources (*crowdsourcing*). Le catalogue est ouvert aux utilisateurs qui peuvent ajouter leurs propres mots clés (ou *tags*) aux notices. Une autre solution serait l'indexation automatique. A la place d'un humain, c'est une machine qui analyse le contenu d'une ressource et en fait ressortir les thèmes principaux.⁹

Ces solutions rencontrent cependant des difficultés avec les vocabulaires contrôlés. Soit elles ne les exploitent tout simplement pas, soit, dans le cas de l'indexation automatique, ils représentent un défi technique non négligeable.¹⁰ Nous allons explorer une voie alternative : il s'agit de l'enrichissement automatique de l'indexation matière.

L'enrichissement automatique des notices bibliographiques

Christoph donne une définition de l'enrichissement automatique spécifique aux ressources documentaires des bibliothèques :

“Mit Kataloganreicherung (englisch catalog enrichment) werden Einträge eines Bibliothekskatalogs um weiterführende Informationen ergänzt, die über die reguläre Formal- und Sacherschließung hinausgehen.» (Christoph, 2013, p. 140)

L'enrichissement consiste principalement dans l'importation de données à partir de réservoirs externes. Cette importation peut prendre deux formes principales. Soit on

7 Le Digitale Assistent est le précurseur de FRED, que nous présentons dans la section suivante.

8 En 2012, RERO a sérieusement envisagé d'abandonner l'indexation matière, mais a finalement décidé de la maintenir.

9 Depuis septembre 2017, la DNB utilise des processus d'indexation automatique de certaines séries de leurs fonds. Le contenu des ressources documentaires est analysé et indexé (au sens informatique) par une machine. Les données bibliographiques sont ensuite enrichies par l'attribution automatique d'indices de classification DDC et de descripteurs GND correspondant au contenu analysé. Cf. <http://www.dnb.de/DE/Erwerbung/Inhaltserschliessung/aenderungInhaltserschliessungSeptember2017.html> (consulté le 26.06.2019)

10 L'efficacité de l'indexation automatique pratiquée par la DNB est remise en question (Wiesenmüller, 2017).

ajoute simplement des liens vers d'autres ressources, soit on importe de nouvelles données « en dur » dans le système local. Dans le premier cas, l'institution qui importe les données ne fait que renvoyer à des données externes, tandis que dans le dernier cas les données sont ajoutées à l'index du système. L'institution importatrice en devient ainsi dépositaire (Christoph, 2013, p. 142).

Les données importées peuvent être plus ou moins bien structurées. Dans le cas d'ajout de contenus supplémentaires, les données seront peu structurées s'il s'agit de mots clés libres, ou de résumés, etc. Elles peuvent au contraire être bien structurées, dans le cas par exemple de l'importation de vedettes matières conformes à un vocabulaire d'indexation. La structuration sera encore meilleure avec l'ajout des identifiants des autorités correspondant aux vedettes. Dans le cas de l'ajout de liens, s'ils s'agit de simples URL statiques, l'information sera peu structurée, alors que dans le cas d'URI conformes aux modèles du web sémantique (en premier lieu RDF), l'information sera structurée et pourra être réutilisable par le système local (Christoph, 2013, p. 143-145).

Enfin, l'enrichissement des données peut se faire sur la base de correspondances (matching) à deux niveaux différents. En effet, les correspondances peuvent être établies à un premier niveau entre les ressources elles-mêmes (par exemple entre les notices du catalogue local et celles du catalogue externe). Elles peuvent également être tirées au niveau supérieur, entre les référentiels eux-mêmes, autrement dit entre les termes des vocabulaires contrôlés. On parle alors d'alignement des vocabulaires.

Deux programmes d'enrichissement automatique permettent d'illustrer ces différentes distinctions. Il s'agit de l'application FRED (pour FREmDaten Anreicherung), de la Bibliothèque de l'Université de Zürich (Bucher et al., 2018),¹¹ et du programme Europeana Semantic Enrichment, de la bibliothèque virtuelle Europeana.¹²

Concernant la nature des données importées, l'enrichissement avec FRED consiste à ajouter des vedettes matières aux notices bibliographiques pour améliorer la recherche d'information. Il s'agit donc de données importées en dur dans les ressources. Au contraire, dans le cas d'Europeana, l'enrichissement se fait par le biais d'URI. Grâce aux standards du web sémantique, l'interface utilisateur d'Europeana peut réutiliser les données présentes dans des réservoirs externes.

Du côté de la structuration, les deux projets s'intéressent à des données structurées, mais pas de la même manière. Les données enrichies dans Europeana sont structurées selon le modèle RDF (*linked data*). L'enrichissement se fait à partir de réservoirs qui proposent leurs données en RDF (Geonames, GEMET, DBpedia, ULAN (Getty), Semium Time). On peut ainsi enrichir l'information sur une personne,

11 De nombreuses informations sur FRED m'ont été transmises directement par l'équipe FRED de la ZB lors d'un entretien qui a eu lieu le 7 juin 2018. Je les remercie chaleureusement pour leur aide.

12 <https://pro.europeana.eu/page/europeana-semantic-enrichment>

un lieu, un concept, etc. en exploitant des données présentes dans ces réservoirs (dates de naissance, traductions en d'autres langues, etc.). Il s'agit de jeux de données non-bibliographiques, qui proviennent d'acteurs extérieurs aux bibliothèques.¹³ Pour FRED, les données importées sont des données bibliographiques structurées selon le format MARC21. Mais le degré de structuration n'est pas le même. Les vedettes matières importées dans FRED ne sont pas conformes aux standards *linked data*, bien que cela soit envisageable.¹⁴

En ce qui concerne le niveau de correspondance, pour FRED, les correspondances sont établies entre des notices bibliographiques. Le travail essentiel de FRED consiste à comparer les vedettes matières de la notice locale avec celles de la notice externe et à importer les champs qui ne sont pas déjà présents dans la notice locale. Les vedettes matières sont comparées sur la base des chaînes de caractères et non pas des identifiants. Il n'y a pas d'alignement entre les termes de différents vocabulaires. Dans le cas d'Europeana, la correspondance est établie entre les termes des référentiels, qui sont donc alignés. Les entités des différents référentiels sont donc directement liées, ce qui permet ensuite de récupérer des informations disponibles dans un référentiel spécifique alors qu'elles ne sont pas présentes localement.

Un aspect particulier de l'enrichissement des données bibliographiques est l'usage du format MARC. La plupart des grands catalogues de bibliothèque utilisent actuellement des variantes du MARC (MARC21, UNIMARC, INTERMARC, etc.). C'est un format bien maîtrisé des bibliothécaires, puisqu'il est utilisé depuis les années 1960, mais il n'est plus adapté aux technologies actuelles. Toutefois, il permet un échange de données assez efficace entre catalogues de bibliothèques. Tant que l'on reste confiné aux réservoirs de bibliothèques, le MARC est approprié.

Cela se complique dès que l'on veut exploiter des données provenant d'autres domaines, comme les éditeurs ou les institutions culturelles et plus généralement les ressources du web. La majorité des ressources disponibles pour les outils de découverte modernes proviennent d'acteurs qui ne produisent pas de données nativement au format MARC. Les plateformes de partage des publications scientifiques (ResearchGate, Academia, etc.), les réservoirs open access, les éditeurs commerciaux, les fournisseurs (EBSCO, Electre, etc.), ou encore les organisations culturelles sont autant d'acteurs incontournables pour les données bibliographiques qui n'ont pas la contrainte du MARC. Le format MARC représente donc un obstacle à l'échange de

13 L'enrichissement pourrait potentiellement se faire depuis des référentiels de bibliothèques. Cela a été envisagé au cours de la phase projet (Simon et al., 2014 Appendix 3) mais n'a pas encore été mis en production.

14 La tendance actuelle est d'ajouter des identifiants aux données bibliographiques afin de permettre à terme leurs conversion dans des modèles de données *linked data*. Du reste, plusieurs organisations bibliographiques publient leurs données en RDF (DNB, BnF, Library of Congress, entre autres). Mais pour l'heure aucune ne produit des données nativement RDF.

données, car il nécessite une conversion des données coûteuse et souvent approximative.

Dans le cadre d'un projet d'enrichissement automatique d'indexation contrôlée, le choix des formats de données à toute son importance. Si une bibliothèque se limite aux données disponibles en MARC, elle se limite également à exploiter uniquement les réservoirs d'autres bibliothèques, comme c'est le cas pour FRED. Cela règle bien un certain nombre de questions. La conversion des données d'indexation est facilitée. La bibliothèque peut exploiter les référentiels répandus dans les bibliothèques. Elle peut par exemple bénéficier de l'indexation RAMEAU effectuée dans d'autres réseaux de bibliothèques. Le revers de la médaille est alors que le nombre et la diversité des réservoirs exploités sont considérablement réduits.

Le programme « Hub de métadonnées » de L'Agence bibliographique pour l'enseignement supérieur (ABES), en France, illustre bien la difficulté de sortir du paradigme MARC pour les données bibliographiques. L'objectif du "Hub" est précisément de récupérer des données de sources et de formats multiples et de les rendre disponibles pour les bibliothèques. En particulier, sont prises en charge les notices de livres électroniques des éditeurs Springer et Oxford University Press. Dans ce contexte, l'objectif est surtout de compléter des données lacunaires fournies par l'éditeur grâce à de l'enrichissement à partir de réservoirs externes (ABES, 2013, p. 15). Concernant les données d'indexation, les notices de livres numériques ont été identifiées avec des notices disponibles dans le Sudoc et à la BnF, ce qui a permis de récupérer les indexations correspondantes (ABES, 2013, p. 24). Même si les données originales ne sont pas en MARC, l'indexation récupérée provient bel et bien de réservoirs MARC.

En quoi l'enrichissement automatique constitue une réponse aux problèmes de surabondance et d'hétérogénéité des données d'indexation ? Concernant la surabondance, l'outil FRED constitue un véritable cas d'enrichissement de notices qui n'ont pas été traitées au préalable. En allant chercher automatiquement de l'information ailleurs, la charge de travail des indexeurs se voit diminuer. Au contraire, un projet d'enrichissement basé sur des *linked data* comme Europeana n'est a priori pas d'une grande utilité. En effet, comme les correspondances sont établies sur la base des référentiels, il faut que l'information soit déjà présente dans la ressource locale. Si l'information n'est pas présente, l'enrichissement n'est pas possible puisqu'il n'y a pas de terme permettant le lien vers un autre référentiel.

Pour ce qui est de l'hétérogénéité des référentiels, c'est un peu l'inverse. FRED ne peut pas vraiment y parvenir, puisqu'il ne fait qu'importer des données supplémentaires sans mettre les référentiels en relation. Pour assurer un accès thématique homogène au catalogue, il faudrait que chaque ressource ait reçu une indexation dans chaque référentiel souhaité. Mais cette stratégie ne fonctionnerait pas forcément, car

d'une culture à l'autre les ressources peuvent grandement varier. Toutes les publications en françaises n'ont pas été indexées en allemand et vice versa. Enrichir des ressources en français avec de l'indexation en allemand ne permettrait donc pas de couvrir l'ensemble des publications en français. Un tel système d'enrichissement ne répond donc pas vraiment au problème d'hétérogénéité. Concernant Europeana, au contraire, l'enrichissement par alignement de référentiels est plus approprié. Si les référentiels sont alignés, une ressource indexée en RAMEAU est virtuellement aussi indexée dans les autres référentiels alignés avec RAMEAU.

Une solution d'importation automatique d'indexation devrait chercher à exploiter les deux aspects. D'une part récupérer des données produites dans d'autres réservoirs, d'autre part exploiter l'équivalence entre référentiels pour mieux mettre les données importées en relation. Quoiqu'il en soit, une solution d'enrichissement automatique devrait s'émanciper du format MARC pour pouvoir exploiter pleinement les réservoirs non bibliographiques.

Il faut à présent examiner comment l'enrichissement automatique de l'indexation pourrait être implémenté dans les données du réseau Renouvaud.

Enrichir les données de Renouvaud ?

Le réseau Renouvaud regroupe près de 113 bibliothèques du Canton de Vaud.¹⁵ Il est né le 22 août 2016 de la décision du Canton de Vaud de quitter RERO, le Réseau des bibliothèques de Suisse occidentale. L'ensemble des bibliothèques du réseau utilise le même SIGB, le logiciel Alma de la société Ex Libris. Le système de gestion est utilisé en couple avec un outil de découverte de la même compagnie, Primo.

Le vocabulaire d'indexation principal du réseau est RAMEAU. Quelques bibliothèques spécialisées utilisent d'autres vocabulaires : les bibliothèques de médecine et santé indexent selon le vocabulaire MeSH, tandis que les bibliothèques de droit indexent selon le vocabulaire Jurivoc, thesaurus suisse spécialisé pour le droit. Le scénario idéal pour une recherche par sujets voudrait donc que toutes les ressources des collections générales soient indexées selon RAMEAU, tandis que les collections spécialisées en médecine et en droit soient indexées intégralement dans leur vocabulaire respectif. Un bref état des lieux des données d'indexation du réseau montre que l'on est encore loin de cet idéal.¹⁶

Cet état des lieux tient compte du type de ressource, du réservoir de données et des divers vocabulaires d'indexation utilisés. Au niveau du type de document, l'indexation ne se fait pas de manière homogène. Du côté des ressources imprimées, les monographies sont bien indexées systématiquement, mais ce n'est pas le cas d'autres

15 Rapport annuel Renouvaud 2018 (Renouvaud. Réseau vaudois des bibliothèques, 2019, p. 8)

16 Un état des lieux plus complet a été effectué dans mon mémoire de master (Hertig, 2018, p. 41-59).

formats, comme les articles, les travaux universitaires ou encore les documents manuscrits. En ce qui concerne les ressources en ligne, seuls les travaux universitaires sont indexés, et encore sans vocabulaire contrôlé.

Concernant l'origine des ressources mises à disposition des usagers, on compte cinq réservoirs différents : (1) le catalogue Renouvaud (ressources cataloguées dans Alma) ; (2) Primo Central : la base de connaissance d'Ex Libris, comportant avant tout des ressources électroniques ;¹⁷ (3) Serval : le dépôt institutionnel de l'Université de Lausanne ;¹⁸ (4) la Biodiversity Heritage Library : bibliothèque virtuelle liée à la biodiversité ;¹⁹ (5) Rero doc (Vaud) : le serveur institutionnel des hautes écoles du Canton de Vaud. Les données patrimoniales de la BCUL ont également été prises en compte. Elles sont en cours de migration vers une nouvelle plateforme dénommée Patrinum.²⁰

Concernant les vocabulaires d'indexation, chaque réservoir a ses propres caractéristiques et l'on observe souvent une multiplicité de référentiels par réservoir. C'est particulièrement le cas du catalogue Renouvaud, qui réunit des dizaines de vocabulaires d'indexation différents. L'indexation de certaines ressources est parfois tout simplement absente. C'est le cas en particuliers des ressources de PCI. Les réservoirs agrégés par la base de connaissance d'Ex Libris ne proposent en effet pas de traitement intellectuel systématique. Il se peut aussi que PCI n'ait pas accès à l'indexation des ressources faite sur les plateformes des fournisseurs. Et même quand c'est le cas, chaque fournisseur va proposer son propre référentiel d'indexation, qui est le plus souvent en anglais, ce qui multiplie les étapes de recherche d'information pour l'utilisateur. D'autre part, comme mentionné précédemment, les fournisseurs de contenus ne sont pas transparents quant aux référentiels utilisés. Ils ne déclarent pas les vocabulaires utilisés et ne les mettent pas à disposition.

Dans le cas des dépôts institutionnels Serval et Rero doc (désormais inclus dans Patrinum), les producteurs ne sont pas des professionnels de l'information documentaire, mais les auteurs des travaux universitaires eux-mêmes. On ne peut donc pas attendre de leur part une indexation contrôlée. On perd donc les bénéfices de l'usage de vocabulaires contrôlés pour les usagers, notamment pour la recherche documentaire.

On retrouve ainsi le constat d'hétérogénéité des référentiels dans les données Renouvaud, à cause de la multiplicité des sources et de l'import de notices. Les

17 https://knowledge.exlibrisgroup.com/Primo/Content_Corner/Product_Documentation/PC_Index_Configuration_Guide/010Overview/010What_is_Primo_Central_Index%3F

18 <http://wp.unil.ch/infoserval/>

19 <https://about.biodiversitylibrary.org/>

20 <https://www.patrinum.ch/> Depuis l'été 2018, les ressources Rero doc (Vaud) ont été migrées sur Patrinum. Depuis mai 2019, les ressources de Patrinum sont moissonnées par Primo. Les ressources Rero doc sont donc mises à disposition via Patrinum, qui est ainsi devenu le cinquième réservoir du réseau.

indexeurs du réseau sont également confrontés au problème de la surabondance des ressources, principalement dans le catalogue Renouveau et dans PCI. À l'heure actuelle, l'indexation matière dans le réseau est concentrée sur les ressources imprimées et les indexeurs travaillent déjà au maximum de leur productivité, pour traiter un seul type de document (monographies imprimées). Il est donc impensable d'élargir l'indexation à d'autres types de document du catalogue, en particuliers les livres numériques. On n'ose même pas évoquer le reste des ressources numériques disponibles dans PCI.²¹

Il ressort que l'écrasante majorité des ressources ne passe pas par un traitement intellectuel selon les normes choisies dans le réseau. En fonction de la problématique développée dans le chapitre 0, on voit que les données Renouveau sont directement touchées par les problèmes de surabondance des ressources et d'hétérogénéité des référentiels. Elles constituent donc un bon terreau pour mettre en place un programme d'enrichissement automatique de l'indexation. Nous allons maintenant esquisser comment les dispositifs d'enrichissements automatiques décrits dans la section précédente pourraient être appliqués aux données Renouveau.

Il est important de distinguer entre les données produites par les institutions locales et celles produites ailleurs. Lorsque des données présentes dans les ressources locales sont également disponibles dans des réservoirs externes, une solution d'enrichissement de type FRED, par importation de vedettes matières, est envisageable. C'est le cas du catalogue Renouveau, par exemple. C'est aussi le cas d'une bibliothèque virtuelle comme la BHL, où les données sont en accès libre.

Au contraire, dans le cas des données Patrinum ou du dépôt institutionnel Serval, les données sont produites par les bibliothécaires du réseau ou par les chercheurs locaux. Il s'agit de données originales pour lesquelles le réseau est en première ligne quant au traitement documentaire et ne peut pas compter sur des indexations à importer. Une solution d'enrichissement automatique de type FRED ne semble donc pas être une option puisqu'elle implique de reprendre des données disponibles ailleurs. Un enrichissement sémantique, par alignement de référentiels, comme celui mis en place dans Europeana, serait plus approprié.

Le cas de PCI est particulier. En effet, l'outil Primo Central d'Ex Libris agrège des ressources identiques à partir de différents fournisseurs afin de les mettre à disposition de l'utilisateur via l'outil de découverte Primo.²² Dans ce cadre, les indexations présentes sur une plateforme sont récupérées dans la notice Primo. Le système fait

21 Pour se donner une idée de l'ampleur de la tâche, prenons l'exemple des ressources des 5 plus gros fournisseurs de ressources numériques du réseau Renouveau (près de 27 millions de documents, cf. note de bas de page 5). À temps plein, un indexeur confirmé dans le réseau indexe en moyenne 3000 documents par année. Réindexer ces 27 millions de documents dans un délai raisonnable, par exemple 3 ans, mobiliserait donc une équipe de 3000 indexeurs !

22 https://knowledge.exlibrisgroup.com/Primo_Central/Product_Documentation/PC_Index_Configuration_Guide/010Overview/020How_Does_Primo_Central_Index_Work%3F (consulté le 24.07.2018)

donc déjà de l'enrichissement par correspondance entre ressources. Cependant, la bibliothèque cliente n'a pas de contrôle là-dessus. Qui plus est, les correspondances sont faites en dehors du format MARC. Du fait que les fournisseurs ne sont pas transparents sur les référentiels utilisés, il n'est pas possible de les aligner avec des vocabulaires d'indexation propres aux bibliothèques.²³

Par contraste, le catalogue Renouvaud présente des caractéristiques propices à l'enrichissement automatique. Le format des données est normalisé (MARC21) et les données des autres catalogues de bibliothèques sont en accès libre. Dans mon travail de master, j'ai tenté d'exploiter le potentiel d'Alma pour mettre au point un processus d'importation automatique de vedettes matières, à l'instar de FRED (Hertig, 2018, p. 61-72).

Il est ressorti des investigations dans Alma que le SIGB n'est pas approprié pour mettre en place une solution d'enrichissement automatique de l'indexation. Le problème principal est qu'Alma ne peut pas établir de correspondance entre les vedettes, mais seulement entre les notices et les étiquettes des champs MARC. Plus généralement, l'architecture du système ne s'adapte pas à une solution d'enrichissement automatique. Ses possibilités techniques sont trop limitées pour s'adapter aux spécifications des divers réservoirs à moissonner. Il ne permet pas de mettre en place une solution de contrôle qualité de l'enrichissement, indispensable pour un tel dispositif. La conclusion générale de l'examen est qu'un système externe est mieux approprié pour enrichir les ressources du réseau Renouvaud.²⁴

Enfin, il faut encore relever qu'un système proche de FRED se limite à un cas limité d'enrichissement. Les réservoirs sont les catalogues de bibliothèques et les données sont au format MARC. Il ne permet pas d'exploiter des réservoirs en dehors des bibliothèques et des formats de données plus hétérogènes. Un tel système ne répond donc qu'à un aspect des problèmes rencontrés avec l'indexation, celui de la surabondance des ressources, mais des ressources disponibles au format MARC et présentes dans le catalogue de la bibliothèque. Une politique d'enrichissement de l'indexation devrait viser plus large et envisager des réservoirs et des formats de données plus diversifiés.

23 On pourrait imaginer un enrichissement supplémentaire des ressources en exploitant des données produites localement, sur les serveurs institutionnels par exemple. Si une ressource présente sur la plateforme d'un éditeur est également présente dans le dépôt institutionnel de l'Unil, on pourrait envisager de récupérer l'indexation de la notice Serval pour l'importer dans la notice Primo. Mais dans la pratique, l'on risque fort de se retrouver avec des redondances, car l'auteur de la publication a sans doute fourni les mêmes métadonnées à Serval et à son éditeur. De plus, il ne s'agit pas d'indexation contrôlée.

24 Des tests supplémentaires pourraient malgré tout être effectués afin d'évaluer l'intérêt d'un moissonnage ponctuel sur un seul réservoir pour faire de l'enrichissement rétrospectif. On pourrait par exemple faire tourner Alma sur le catalogue de la DNB pour récupérer les données d'indexation GND.

Conclusion

L'enrichissement automatique de l'indexation est une solution aux problèmes de surabondance des ressources et d'hétérogénéité des référentiels. Deux méthodes principales ont été distinguées dans cette contribution (sans être les seules possibles).

L'enrichissement par correspondance entre ressources permet de récupérer de l'information produite ailleurs. En revanche, elle ne permet pas d'agréger les données, car elle ne présuppose pas de lier les termes des référentiels entre eux. Dans le pire des cas, ce type d'enrichissement peut même déboucher sur l'augmentation de l'hétérogénéité des référentiels.

L'enrichissement par alignement de référentiels est approprié pour les ressources dont il n'existe qu'une seule description, pour lesquelles il n'est pas possible d'importer des données supplémentaires qui seraient disponibles dans les descriptions d'autres réservoirs. Dans le cas présent, l'enrichissement repose sur le fait que certaines métadonnées possèdent des équivalents dans d'autres référentiels et ces métadonnées peuvent ainsi être liées à ces données équivalentes.

Concernant l'alignement des référentiels, il y a un gros travail à faire sur les données. Les référentiels utilisés localement doivent être alignés avec d'autres vocabulaires, en particuliers RAMEAU, LCSH et GND. Tant que ce travail d'alignement n'aura pas été effectué, une solution d'enrichissement sémantique ne sera pas envisageable. Cet état de fait concerne les institutions bibliographiques en général. L'alignement entre référentiel n'a pas encore été effectué de manière complète. Le développement par des acteurs privés de référentiels propres est un obstacle aux projets d'alignements.

L'autre grande entreprise d'enrichissement serait d'ouvrir les données bibliographiques à des référentiels extérieurs aux bibliothèques, à l'instar du Semantic Enrichment d'Europeana ou du Hub de métadonnées de l'ABES. Le catalogue Renouvaud pourrait de la même manière être enrichi avec des données comme VIAF ou DBpedia pour les personnes, Geonames pour les lieux, ISIL pour les collectivités, etc. Toutefois, ce genre d'enrichissement nécessite la conversion des données en *linked data* pour que le liage vers ces ressources soit possible.

L'analyse des données de Renouvaud a montré qu'il est important de considérer chaque réservoir et chaque fournisseur comme un cas unique, en tout cas dans un premier temps. On ne peut pas mettre en place des procédures généralisées pour l'ensemble des réservoirs. Il faudrait mener une analyse sur les différents jeux de données afin de proposer des solutions d'enrichissement appropriées à chaque cas.

Enfin, il ne faut pas oublier que l'enrichissement automatique de l'indexation n'est qu'une partie de la réponse aux problèmes soulevés plus haut. Cette solution doit être complétée par d'autres, comme l'indexation à la source et l'indexation automatique.

Bibliographie

- ABES. (2013). *Etude sur la faisabilité et le positionnement d'un hub de méta-données ABES*. ABES.
- Bawden, D. (2012). *An introduction to information science*. London: London : Facet Publ.
- Bertram, J. (2005). *Einführung in die inhaltliche Erschliessung: Grundlagen - Methoden - Instrumente*. Würzburg: Ergon Verlag.
- Christoph, P. (2013). Datenanreicherung auf LOD-Basis. In *(Open) Linked Data in Bibliotheken*. <https://doi.org/10.1515/9783110278736.139>
- Cleveland, A. D., & Cleveland, D. B. (2013). *Introduction to Indexing and Abstracting, 4th Edition* (4 edition). Santa Barbara, California: Libraries Unlimited.
- Deutsche Nationalbibliothek. (2017). *Strategische Prioritäten 2017-2020*.
- Gross, T., Taylor, A. G., & Joudrey, D. N. (2015). Still a Lot to Lose: The Role of Controlled Vocabulary in Keyword Searching. *Cataloging & Classification Quarterly*, 53(1), 1-39. <https://doi.org/10.1080/01639374.2014.917447>
- Hertig, M. (2018). *L'enrichissement automatique de l'indexation dans le réseau Renouvaud*. Universität Bern, Bern.
- Kempf, A. O. (2013). Automatische Inhaltserschließung in der Fachinformation / Automatic indexing of domain-specific information / L'indexation automatique dans l'information spécialisée. *Information - Wissenschaft & Praxis*, 64(2-3), 96–106. <https://doi.org/10.1515/iwp-2013-0011>
- Malits, A., & Schäuble, P. (2014). Der Digitale Assistent: Halbautomatisches Verfahren der Sacherschließung in der Zentralbibliothek Zürich. *ABI Technik*, 34(3-4), 132–143. <https://doi.org/10.1515/abitech-2014-0024>
- Renouvaud. Réseau vaudois des bibliothèques. (2019). *Renouvaud. Rapport annuel 2018*. Consulté à l'adresse https://www.bcu-lausanne.ch/wp-content/uploads/2019/05/20190513_RA2018-RENOUVAUD_final_web.pdf
- Salaün, J.-M., & Arsenault, C. (2010). *Introduction aux sciences de l'information*. Paris: La Découverte.
- Simon, A., Suero, D. V., Hyvönen, E., Guggenheim, E., Svensson, L. G., Freire, N., ... Petras, V. (2014). *EuropeanaTech Task Force on a Multilingual and Semantic Enrichment Strategy: final report* (p. 44). Europeana.
- Wiesenmüller, H. (2017, août 2). Das neue Sacherschließungskonzept der DNB in der FAZ. Consulté 16 juillet 2018, à l'adresse Basiswissen RDA website: <http://www.basiswissen-rda.de/neues-sacherschliessungskonzept-faz/>