Les chantiers d'indexation rétrospective à la Bibliothèque nationale de France

Etienne Cavalié

L'indexation matière en transition

De la réforme de Rameau à l'automatisation

Chapitre 7. Les chantiers d'indexation rétrospective à la Bibliothèque nationale de France

Étienne Cavalié. Bibliothèque nationale de France.

Pour la session d'août 2018, l'appel à contributions de la *Subject Analysis and Access Section* de l'IFLA portait en grande partie sur les processus d'indexation automatisée à base de *machine learning*, et a donné l'occasion de mettre notamment en lumière les expériences allemande[1], norvégienne[2] et iranienne[3].

Les bibliothèques françaises n'ont pour le moment pas encore mis en place un tel mécanisme. Mais à défaut de s'appuyer sur des résumés ou du texte intégral, dont on parviendrait à extraire des concepts et attribuer ainsi automatiquement une indexation matière à chaque document, la Bibliothèque nationale de France a mené plusieurs chantiers d'indexation automatique sur des fonds déjà catalogués, dont le présent chapitre rend compte à travers trois exemples.

Profitons-en pour rappeler qu'un processus dit « automatique » inclut nécessairement des interventions manuelles à un ou plusieurs moments de son exécution. L'automatisation n'existe que par degrés, et ne peut jamais faire complètement abstraction de l'être humain, comme le rappelle par exemple le tout récent ouvrage d'Antonio Casili, *En attendant les robots*[4].

Dans la définition de sa politique d'indexation, la BnF précise les règles pour le traitement courant, mais ne s'engage pas sur une ré-indexation du rétrospectif : lors de chantiers de catalogage rétrospectif, ou de corrections de notices déjà existantes, la consigne n'est pas de les

---

[1] Ulrike Junger, « Automation first – the subject cataloguing policy of the Deutsche Nationalbibliothek », dans *Transform Libraries, Transform Societies*, présenté à IFLA WLIC 2018, Kuala Lumpur (Malaisie), 26 août 2018, [En ligne] <http://library.ifla.org/2213/1/115-junger-en.pdf> (Consulté le 21/02/2019).

[2] Svein Arne Brygfjeld, Freddy Wetjen et André Walsøe, « Machine learning for production of Dewey Decimal », dans *Transform Libraries, Transform Societies*, présenté à IFLA WLIC 2018, Kuala Lumpur (Malaisie), 26/08/2018, [En ligne] <http://library.ifla.org/2216/1/115-brygfjeld-en.pdf> (Consulté le 21/02/2019).

[3] Mahboubeh Ghorbani et Fattaneh Torkashvand, « Lessons learned from Automatic Indexing Projects regarding to Persian Language Specifications », dans *Transform Libraries, Transform Societies*, présenté à IFLA WLIC 2018, Kuala Lumpur (Malaisie), 26 août 2018, [En ligne] <http://library.ifla.org/2215/1/115-ghorbani-en.pdf> (Consulté le 21/02/2019).

[4] Antonio Casilli, *En attendant les robots*, Paris, Seuil, 2019.

indexer. Néanmoins la porte est laissée ouverte pour ces fonds déjà décrits, selon la possibilité d'exploiter des informations déjà présentes[5].

L'indexation rétrospective peut donc faire l'objet de chantiers ponctuels, s'il y a un « angle d'attaque » : soit un projet thématique ou intellectuel (par exemple dans la perspective de constituer un corpus qui doit être numérisé) ; soit une faisabilité technique, la possibilité d'utiliser une information déjà présente, pour la restructurer afin de générer une indexation identique à celle produite en catalogage courant (en l'occurrence, indexation Rameau et Dewey).

Commençons par évaluer la volumétrie des notices concernées, c'est-à-dire la part des notices du catalogue de la BnF qui n'ont pour le moment pas d'indexation matière. Une requête assez simple dans data.bnf.fr[6] nous permet de connaître la part *d'ouvrages indexés*, par année. Et une requête à peine différente permet d'obtenir le nombre *total* de notices par année (avec ou sans indexation). Le rapport entre les deux fournit le pourcentage des notices disposant d'une indexation matière, sur l'ensemble des notices, par année. Même si data.bnf.fr ne contient que 68 % du catalogue (décembre 2018 : 9,6 millions sur 14 millions), cela permet malgré tout de se rendre compte des proportions. Précisons aussi que la requête ne peut se faire que sur les notices pourvues d'une date de publication.

```
PREFIX frbr-rda: <http://rdvocab.info/uri/schema/FRBRentitiesRDA/>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX bnf-onto: <http://data.bnf.fr/ontology/bnf-onto/>
select ?datePub (count(distinct ?manif) as ?nbManifs) where {
?manif a frbr-rda:Manifestation; dcterms:subject ?sujet; bnf-onto:firstYear
?datePub.
FILTER (?datePub > 1800)
}
ORDER BY ?datePub
```

[5] « Ce document prend en compte les enjeux du catalogage courant et non des conversions rétrospectives, hautement dépendantes des données sources. » *Politique d'indexation de la Bibliothèque nationale de France*, juin 2018, p. 1, [En ligne] <http://www.bnf.fr/documents/politique_indexation_2018.pdf> (consulté le 12/01/2018).

[6] Interface de requête pour la base data.bnf.fr utilisant le langage SPARQL : <https://data.bnf.fr/sparql>.

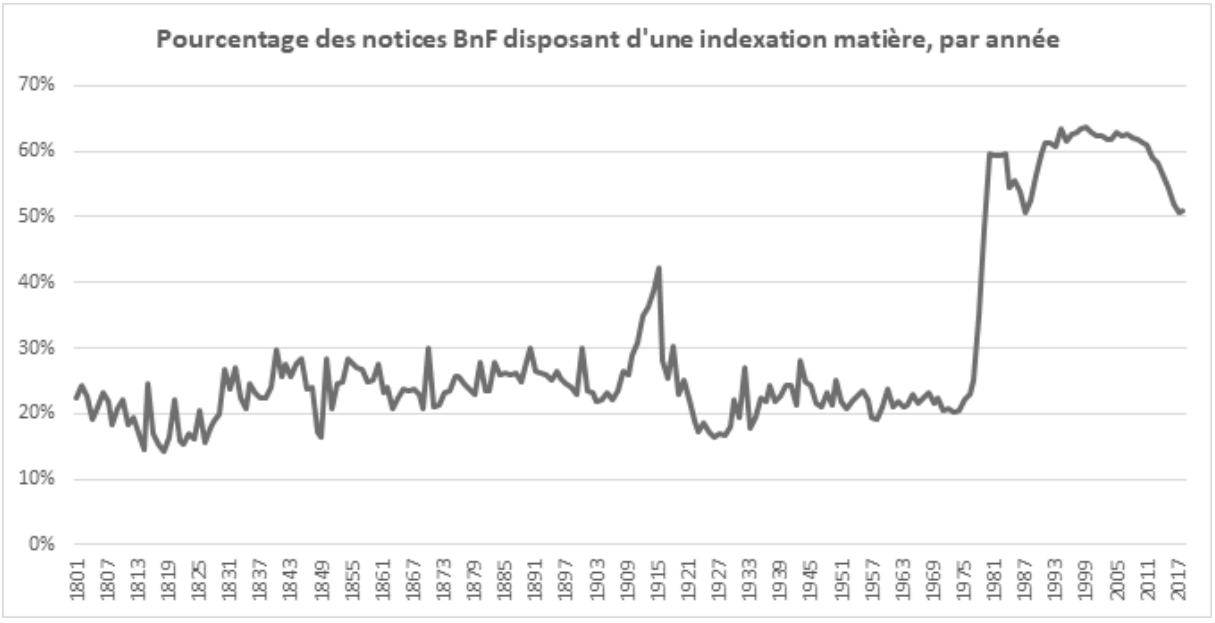

Figure 1. Pourcentage des notices BnF disposant d'une indexation matière, par année

Rameau n'existe que depuis 1980. Les 20 % (en moyenne) de notices antérieures déjà indexées l'ont été rétrospectivement : on se rend compte par exemple sur le graphe ci-dessus que les années 1914-1918 ont fait l'objet d'une indexation rétrospective plus fréquente, pour des raisons intellectuelles (intérêt pour la période) et non de faisabilité technique.

Par ailleurs, on voit bien d'une part le point de rupture (1980), mais aussi le plafond à atteindre : la politique d'indexation de la BnF correspond, en traitement courant, à 60 % des collections environ (et pas à 100 %, tous les documents n'ayant pour le moment pas vocation à être indexés). Donc pour aller vers une indexation homogène, qui permette d'avoir une meilleure vision du contenu des collections, il faut faire en sorte que le plus grand nombre de notices soit indexé selon ce référentiel Rameau, au moins jusqu'à une hauteur de 60 % de l'ensemble.

Le principe des chantiers décrits ci-dessous est de s'appuyer sur une « indexation » (ou ce qui peut en tenir lieu) déjà présente, ou sur une information structurée ou semi-structurée, que l'on puisse transformer en un ou plusieurs concepts Rameau.

Cas 1 : des cotes Clément aux vedettes Rameau (le « T catalogué »)

Du XVII^{e} siècle aux années 1990, les ouvrages entrant dans les collections de la BnF ont été classés selon la cotation Clément, du nom de ce garde de la bibliothèque de Louis XIV. Sans faire un complet historique de cette cotation, on peut en dire tout de même qu'elle s'appuie sur des lettrages désignant des grands domaines disciplinaires[7].

Chaque cote a ensuite eu son histoire, et certaines ont fait l'objet de subdivisions – selon le même mécanisme que la CDU ou la Dewey – alors que pour d'autres lettrages les livres y sont

[7] Voir la liste des lettrages dans l'article Wikipédia Cotation Clément : <https://fr.wikipedia.org/wiki/Cotation_Cl%C3%A9ment>.

numérotés simplement par ordre d'arrivée. Il est essentiel alors de distinguer deux catégories de lettrages : les lettres « cataloguées » et « non cataloguées ». Les lettres dites « cataloguées » sont celles qui ont la particularité de répartir les collections en chapitres (avec un identifiant alphabétique), eux-mêmes divisés en sous-chapitres (identifiant numérique).

Il existe huit lettrages « catalogués » au sein du cadre de classement élaboré par Nicolas Clément : L, N, O, O2, O3, P, P2 et T. Cela donne par exemple pour le « T catalogué » :

- T : Sciences médicales
    - Tf : Médecine légale
        - Tf1 : Considérations sur la médecine légale

Le chantier d'indexation du T catalogué a concerné les collections anciennes en sciences médicales et médecine vétérinaire, ce qui correspond à 130 000 notices. Le T catalogué n'a plus été enrichi depuis 1925 : on travaille donc sur des fonds qui ne sont plus cotés de cette manière. Le département des Sciences et Techniques, dont une large partie des collections est cotée au lettrage « T catalogué », a entrepris d'exploiter cette information pour générer une indexation Rameau et un indice Dewey. L'équipe a entrepris de générer une table de concordance, sous la forme d'un tableau à trois colonnes :

- Une tranche de cote (et sa définition)
- L'indexation Rameau correspondante, sous la forme d'une indexation construite, allant de 1 à 3 concepts
- L'indice Dewey correspondant

À la suite de cette table de conversion de cotes en sujets Rameau, toutes les notices concernées se sont vues enrichies des mots clés correspondants à leur cote (indexation Rameau et indice Dewey).

Relevons au passage que l'« indexation » initiale exploitée se trouvait au niveau des exemplaires, et que cette information a été exploitée pour enrichir les notices bibliographiques.

La nature du traitement a consisté à réaliser un alignement entre deux référentiels : les cotes du T catalogué et des concepts Rameau. Ainsi, le début de cote Tf1 (Considérations sur la médecine légale) a été aligné avec le concept Rameau « Médecine légale » (identifiant: ark:/12148/cb11932336s).

Au total, 617 lettrages différents se sont vus attribuer une équivalence avec une indexation Rameau (et Dewey quand c'était possible), ce qui a permis l'indexation des 130 000 notices ayant un exemplaire coté sous ce lettrage.

Cas 2 : des LCSH à Rameau pour 100 000 e-books

Le second corpus présenté est un cas différent : entre 2011 et 2017, la BnF a acquis à l'unité 3000 e-books, et environ 100 000 titres par bouquets. Plus exactement, elle a acquis des accès pérennes aux documents, sur les plates-formes des diffuseurs. Et pour chaque notice, on disposait d'un ISBN 13 comme métadonnée initiale.

La politique de catalogage de la BnF consiste à ne faire entrer dans son catalogue[8] que les ressources présentes de manière pérenne dans les collections. Les abonnements électroniques (y compris les abonnements à des plates-formes qui contiennent des e-books) n'y sont donc pas signalés. Mais dans le cas de ces 103 000 documents mentionnés, la pérennité de l'accès rendait le signalement légitime. Ce qui implique d'obtenir des notices en Intermarc[9] pour chacune d'entre elles.

WorldCat a été identifié comme source première pour générer les notices : la dérivation en masse de plusieurs milliers de titres pour récupérer la description de chaque document (titre, éditeur, collation) à partir de son ISBN a été réalisée en reprenant le mécanisme déjà utilisé au quotidien par les catalogueurs lorsqu'ils récupèrent les notices WorldCat pour cataloguer un ouvrage étranger.

Mais lors d'une dérivation à l'unité, le catalogueur complète la notice obtenue par conversion de la notice Marc21 vers Intermarc. Il crée notamment des liens vers les notices d'autorité Auteur et Sujet.

Dans le cas d'un corpus de 103 000 documents, il était exclu, pour des raisons de moyens humains notamment, de faire reprendre à la main chacune des notices : il fallait donc trouver une manière de les lier, autant que possible, aux notices déjà présentes dans la base. Cela signifiait donc identifier les notices d'auteurs et les notices de vedettes matières, à partir des informations déjà présentes dans les notices bibliographiques dérivées.

Le mécanisme de dérivation « simple » a été enrichi à l'issue du processus qui a généré les notices en Intermarc. Pour chaque ISBN, on a réalisé successivement trois recherches :

- Recherche par ISBN dans WorldCat, afin de récupérer les métadonnées descriptives (titre, nom d'auteur, date de publication, éditeur) ainsi que l'indexation matière, généralement exprimée avec les LCSH (*Library of Congress Subject Headings*).
- Recherche par Titre-Auteur dans le catalogue de la BnF (à partir du titre et de l'auteur fournis par WorldCat à l'étape précédente). Dans ce cas, on ne cherche pas directement l'ISBN puisqu'on sait que l'ouvrage n'a encore jamais été catalogué. En revanche, on peut retrouver une autre édition, imprimée, du même livre, et donc identifier son auteur et son sujet dans l'autre notice.
- Recherche par ISBN dans le Sudoc[10] : si un catalogueur du réseau Sudoc a déjà dérivé et enrichi la notice bibliographique, en ajoutant des liens vers IdRef[11], on peut exploiter cette information dans le cas où la notice IdRef mentionne un identifiant BnF (c'est le cas systématiquement pour les vedettes-matières Rameau, moins fréquemment pour les notices de personnes et de collectivités).

---

[8] Le catalogue de la BnF est consultable sur <https://catalogue.bnf.fr/>.

[9] Si elle diffuse ses notices en Unimarc, format d'échange largement partagé dans le monde des bibliothèques, la BnF catalogue nativement dans un format qui lui est spécifique, Intermarc, qui lui permet notamment de mieux décrire les collections particulières de ses départements spécialisées.

[10] Le Sudoc est le catalogue collectif de l'enseignement supérieur, administré par l'Abes (Agence bibliographique de l'enseignement supérieur).

[11] IdRef est la base de notices d'autorité de l'Abes. Elle est consultable à l'adresse <https://idref.fr/>.

Chacune de ces bases est bien sûr interrogée par web service[12].

Passons sur le traitement des liens aux auteurs, qui n'est pas le thème de cet ouvrage.

Pour l'indexation Sujet[13], on a déterminé un ordre de préférence entre les différentes sources interrogées :

- si on a trouvé au moins une notice BnF (autre édition, imprimée, du même livre), on récupère son indexation Sujet et on la réapplique à la notice de l'e-book
- à défaut, si on a trouvé une notice dans le Sudoc (avec indexation Rameau), on récupère l'indexation Rameau du Sudoc (en convertissant les liens aux notices Rameau IdRef en liens aux notices Rameau BnF)
- à défaut (cas le plus fréquent), on s'appuie sur l'indexation LCSH souvent déjà présente en essayant de la convertir en indexation Rameau

En effet les gestionnaires des notices Rameau s'efforcent autant que possible, lorsqu'ils créent ou mettent à jour une notice Rameau, d'indiquer le libellé correspondant dans le référentiel LCSH : 90 000 notices Rameau contiennent un équivalent LCSH. La conversion de l'indexation LCSH trouvée dans les notices, en indexation Rameau, s'appuyait donc sur ces équivalences Rameau-LCSH présentes dans les notices Rameau elles-mêmes. Une fois la conversion réalisée, concept à concept, le programme vérifiait si la chaîne construite était autorisée (s'assurant par exemple que le premier terme pouvait effectivement être utilisé en tête de vedette : cette autorisation est présente sous la forme d'une valeur codée dans la notice Rameau).

On voit donc dans ce chantier sur les e-books trois stratégies à l'œuvre, mises en place tour à tour, pour essayer de générer une indexation automatique :

- La première s'appuie sur la notion d'œuvre et le modèle IFLA LRM[14] : si une autre édition (manifestation) existe et dispose d'une indexation matière, cette indexation vaut pour une autre édition de la même œuvre.
- La seconde va puiser dans un autre réservoir : si l'indexation Rameau est déjà faite ailleurs, pourquoi ne pas la récupérer ?
- La troisième exploite les mécanismes d'alignements : puisqu'une indexation a déjà été faite, mais dans un autre vocabulaire, ne peut-on la transposer selon notre propre référentiel ?

---

[12] Le web service est pour un programme informatique ce que l'interface de consultation est à l'internaute : une manière adaptée d'interroger une base de données. L'interface de consultation est conçue pour l'être humain, avec une mise en page, des couleurs, des images, des sauts de ligne. Ces éléments de présentation n'ont pas d'intérêt pour un programme qui devrait interroger un catalogue et en extraire (pour les retraiter) des informations : le web service lui propose un mécanisme de requêtage avec un format d'affichage différent, non pas en HTML (affichage visuel mais à la sémantique pauvre), mais le plus souvent en XML ou en JSON, avec conservation de la signification de chaque information.

[13] Étienne Cavalié, « Des milliers de ebooks (et de liens !) dans le catalogue de la BnF », *Bibliothèques [*reloaded*]*, 9 juillet 2018, [En ligne] <https://bibliotheques.wordpress.com/2018/07/09/des-milliers-de-ebooks-et-de-liens-dans-le-catalogue-de-la-bnf/> (Consulté le 16/01/2019).

[14] Sur le modèle IFLA LRM (Library Reference Model), cf. *supra* le chapitre de Françoise Leresche.

Cas 3 : exploiter l'indexation libre

Certaines collections à la BnF disposent d'une indexation libre, issue des différentes strates de catalogage : si la zone Intermarc 619 n'est aujourd'hui contrôlée par aucun référentiel, elle est issue, pour les lettrages LN27 et LK7 (et d'autres) de versions antérieures du catalogue.

La lettre L (histoire de France) est bien un lettrage « catalogué », mais sans subdivisions aussi fines que le lettrage T décrit plus haut :

- LN : Biographies françaises
    - LN27 : Biographies individuelles, par ordre alphabétique de nom de personne
- LK : Histoire et géographie locales, France continentale
    - LK7 : Histoire des villes et localités diverses

Mais au sein des milliers de notices ayant une cote commençant par LK7, il est impossible d'utiliser la seule cote pour générer une indexation (et une indexation globale « Villes -- France -- Histoire » pour toutes les notices cotées LN27 n'aurait aucun sens). De même, LN27 regroupe des ouvrages à caractères biographiques sur des personnes de l'histoire de France. Mais il n'y a pas de subdivision plus précise (il n'existe pas de cote spécifique regroupant les ouvrages sur Louis XIV, ou sur Du Guesclin).

En revanche, le lettrage LN27 (130 200 documents) contient souvent en indexation libre le nom de la personne qui est sujet de la biographie : c'est-à-dire que les prénom et nom de la personne sont indiqués dans cette zone 619, mais sans lien à la notice d'autorité. Or dans le cas des autobiographies, les mêmes prénom et nom sont à la fois auteur et sujet de l'ouvrage : la différence entre ces deux zones, c'est que dans la première le lien à la notice d'autorité existe. Le travail de reprise semi-automatisé, réalisé en 2016-2017, a donc consisté à valider le doublement systématique du lien à la notice d'autorité en tant qu'auteur comme lien en tant que sujet. Quand c'était possible (avec une information déjà présente et exploitable), on a aussi ajouté une mention de genre/forme en fin de la zone d'indexation sujet.

Enfin, si la notice d'auteur contenait une précision de domaine d'activité (sous la forme d'un indice Dewey à 3 chiffres : valeur 620 pour « Ingénierie et activités connexes », 800 pour « Littérature » (dans le cas d'un écrivain), un indice Dewey a aussi été ajouté aux notices bibliographiques liées : si l'auteur est une personne du monde des lettres, son autobiographie va porter sur « le monde des lettres » également. Si on trouve « 620 » dans la notice de l'auteur, la notice bibliographique pour son autobiographie se verra ajouter l'indice Dewey « 620.009 2 » (« Ingénierie et activités connexes – Biographie »).

4 000 notices bibliographiques enrichies d'une indexation Rameau grâce à ce chantier, au nom de personne, dont 1 000 avec une précision de genre (correspondance, autobiographie, etc.).

La cote LK7 (106 000 documents) regroupe les documents sur l'histoire de lieux en France. Là encore, la cote Clément ne nous donne pas suffisamment d'informations pour la convertir en indexation Rameau.

En revanche une partie des notices dispose d'une indexation libre qui mentionne notamment les noms de lieux sur lesquels portent les documents.

On trouve en zone 619 de ce corpus 16 000 valeurs différentes (12 500 quand on les a un peu nettoyées).

En s'appuyant sur l'expérience acquises avec le développement du logiciel Bibliostratus[15], conçu pour faciliter les alignements avec les notices bibliographiques de la BnF en interrogeant systématiquement le web service du catalogue[16], un programme a été développé pour rechercher systématiquement les chaînes de caractères présentes dans la zone d'indexation libre afin d'identifier les termes Rameau correspondants. Des règles complémentaires sont mises en place pour favoriser les noms de lieux : l'entrée « Rouen » va ainsi plutôt être alignée avec la ville (nom géographique Rameau) dont le point d'accès est « Rouen (Seine-Maritime) », et non avec la notice de la collectivité, dont le point d'accès est strictement « Rouen » (donc, *a priori*, plus conforme à la valeur en entrée puisque rigoureusement identique).

L'homogénéité relative du corpus (on sait qu'on y trouve pour l'essentiel des noms de lieux français) rend les résultats plus fiables. Mais la difficulté réside principalement dans le fait qu'il s'agit de valeurs saisies librement, non adossées à un référentiel. Il y a un référentiel implicite, mais sans contrôle informatique : la liste des communes, ou autres structures géographiques administratives, en France. Les catalogueurs qui ont indexés ces ouvrages ne se sont pas sentis libres d'ajouter tous types de termes, ou de les écrire de n'importe quelle manière. Mais cela n'empêche pas des variantes orthographiques (dans les accents, les tirets, les abréviations). Et par ailleurs on y trouvera aussi malgré tout des noms de personnes ou d'organisations…

Le traitement consiste donc à prendre chaque chaîne de caractère saisie librement pour la faire correspondre à une valeur dans un référentiel fermé. Un tel traitement n'est pas encore envisagé pour une indexation matière de noms communs, qui pose de nouvelles questions[17], mais pose un premier jalon vers ce genre d'expérience. Il permet tout de même de définir un processus de traitement et des méthodes de contrôles.

De l'alignement entre référentiels, à l'extraction d'entités nommées

L'alignement des référentiels, un enjeu essentiel dans le cadre du web de données

Le point commun de ces chantiers réside dans l'exploitation d'une information antérieure, très structurée, et dans la réalisation d'un travail de traduction d'un référentiel à l'autre : c'est ce qu'on désigne par le terme d'alignement (ou *mapping*), qui consiste à déclarer des équivalences de concepts ou entités entre deux référentiels : entre Tf1 (Considérations sur la médecine légale, selon Clément) et ark:/12148/cb11932336s (Médecine légale, selon Rameau). Pour décrire une ressource en adoptant un nouveau référentiel, il faut donc une information pré-existante quelque part selon un référentiel précédent, soit dans la notice elle-même sous une autre forme, soit

---

[15] Pour en savoir plus sur Bibliostratus, consulter la page <https://www.transition-bibliographique.fr/systemes-et-donnees/bibliostratus-alignement-donnees-catalogues/> sur le site de la Transition bibliographique.

[16] Le catalogue de la BnF est interrogeable par web service à l'adresse <http://catalogue.bnf.fr/api>.

[17] L'exemple de « Rouen » donné plus haut montre bien qu'il ne suffit pas d'avoir une égalité stricte des termes pour conclure à l'équivalence des concepts : dans la notice, « Rouen » désigne le lieu ; alors que la seule notice d'autorité dans le référentiel BnF dont le point d'accès est exactement « Rouen » ne désigne pas le lieu, mais la municipalité.

même dans une autre base de données : en effet dans le cas des e-books dont on a récupéré l'indexation Sudoc, il y a eu en effet un double mapping, d'abord au moyen de l'ISBN entre la notice dérivée de WorldCat et la notice Sudoc, puis au moyen du PPN (identifiant Sudoc/IdRef) entre la notice Rameau IdRef et la notice Rameau BnF.

Dans le projet de diffuser nos données bibliographiques dans le web de données, il y a un énorme enjeu concernant l'alignement entre référentiels, une mission essentielle notamment pour les agences nationales. En effet, dans le cadre du signalement bibliographique universel[18], chaque agence nationale en charge du dépôt légal signale, pour le monde, la production éditoriale de son pays. Et, selon sa politique d'indexation, elle assure une indexation matière selon son propre référentiel linguistique (LCSH aux États-Unis, Rameau en France, GND[19] en Allemagne, etc.). Donc les bibliothèques qui n'utiliseraient pas le même vocabulaire pour réaliser l'indexation matière dans leur catalogue, en particulière pour des raisons linguistiques, ne pourraient pas bénéficier de cette indexation matière – et tout est à refaire. Si des alignements existent entre référentiels (soit de pair à pair, soit en passant par un référentiel pivot), l'indexation matière LCSH peut être convertie en indexation Rameau : c'est précisément ce qui a été réalisé pour le corpus des e-books, grâce au travail d'alignement présent dans les notices Rameau. Dans le partage des données des bibliothèques, on pense généralement d'abord à la diffusion des données bibliographiques : le partage des données d'autorité, des référentiels, est bien souvent une richesse dont la diffusion trouve plus facilement des bénéficiaires.

Toutefois la conversion automatisée n'est possible que si les règles d'utilisation pour chaque concept peuvent être reconnues par la machine : si le concept LCSH en tête de vedette correspond à un concept Rameau qui lui, ne peut être utilisé qu'en subdivision Nom commun, cette information « À n'utiliser qu'en subdivision » est bien codée dans la notice Rameau[20]. En revanche l'information qui précise que tel terme « s'emploie uniquement en subdivision aux guerres et révolutions »[21] est une règle syntaxique qui n'existe que sous forme textuelle : il est difficile de concevoir un programme informatique qui saura interpréter cette information dans les zones de notes (avec tous les risques d'erreurs de saisie que cela implique), et s'assurer en outre que le concept qui précède est bien une guerre ou une révolution. C'est pourquoi le travail d'alignement entre référentiels, que data.bnf.fr a spécifiquement pour mission de valoriser, a une cohérence forte avec la réforme Rameau en cours, dont un des objectifs est de rendre la chaîne d'indexation beaucoup plus prévisible, et de généraliser les règles syntaxiques dans la construction de la chaîne d'indexation.

---

[18] IFLA, « Bibliographic control », 5 décembre 2017, [En ligne] <https://www.ifla.org/best-practice-for-national-bibliographic-agencies-in-a-digital-age/node/8911> (Consulté le 05/01/2019).

[19] *Gemeinsame Normdatei*, référentiel sujet géré par la Deutsche Nationalbibliothek.

[20] Zone 106 du format Unimarc Autorités. Cf. la page de documentation de l'Abes < http://documentation.abes.fr/sudoc/formats/unma/zones/106.htm> (Consulté le 09/03/2019).

[21] C'est le cas des notices Rameau « Participation + adj. de nationalité », comme « Participation russe ».

L'alignement : pour une recontextualisation des notices et une homogénéisation des catalogues

L'indexation matière n'a réellement de sens et n'est pleinement exploitable que si elle est présente de manière homogène dans tout le catalogue. Ce qui est vrai pour une utilisation normale du catalogue l'est plus encore dans une perspective de recherche[22] : proposer une recherche sujet, sans avoir le moyen de compenser le fait qu'une moitié seulement des données est effectivement indexée (voire beaucoup moins !), est trompeur. Ces documents passent au travers des recherches en question et font défaut à l'utilisateur. L'indexation rétrospective vise à restituer une certaine cohérence à un catalogue dont la constitution s'est étirée dans le temps.

On voit avec l'exemple du T catalogué que c'est précisément l'historique de ce catalogue qui pose question : les documents du T catalogué sont indexés (par leur cote), ceux du LN27 et du LK7 également. Mais cette indexation pré-existante n'est plus exploitable aujourd'hui, et le travail consiste non pas à inventer une nouvelle indexation, mais à actualiser l'indexation précédente. Ce qui avait du sens en 1920 (le lettrage Tf1, par exemple), n'en a plus aujourd'hui, dans un nouveau *contexte*.

C'est bien un problème de contexte qui est posé : toutes ces notices sont déjà indexées (en lettrage Clément, en indexation libre, en LCSH), mais dans d'autres contextes (catalogue BnF 1920, catalogue BnF 1960, catalogue américain) qui n'ont pas de sens pour le catalogue de la BnF aujourd'hui.

Cette question du contexte des données est un des points majeurs du présent ouvrage : à l'heure où l'on promeut le partage des données et la réutilisation des métadonnées, comment parvenir à recontextualiser systématiquement ce qu'on récupère ? La BnF n'échappe pas à la question : elle aussi doit s'efforcer de créer de la cohérence dans son catalogue, en fonction des objectifs de sa politique d'indexation (Rameau et Dewey, et quelques autres vocabulaires pour des types de documents très spécifiques comme les images fixes ou les cartes et plans) et de l'origine des notices existantes.

Sémantiser le catalogue

Pour peu qu'on dispose du dictionnaire des cotes (mais celui des cotes Clément n'est pas aussi bien diffusé que la base des notices Rameau), il est possible d'accéder à la signification du code « Tf1 » d'une manière non ambiguë. Ce n'est pas forcément le cas pour le corpus des documents de la cote LK7 (histoire de France - lieux) où l'indexation matière peut parfois être : « Saint-Martin », ou « St-Martin ». Cela dit, même « Paris » n'est pas un lieu complètement univoque : certes, au sein de la cote LK7 qui est une sous-collection de la cote L qui traite de l'histoire de France, Paris est forcément la capitale française. Mais cette chaîne de caractères ne saurait constituer un identifiant universel dans un contexte de diffusion des données : la base de données géographiques Geonames recense près de 16 000 lieux contenant le mot « Paris ».

Dans le traitement du corpus LK7, le travail consiste à convertir une chaîne de caractères, compréhensible dans un certain contexte, en une ressource identifiable de manière univoque et désignée par son URI. Cela revient donc à sémantiser le catalogue, à partir de termes présents

[22] Voir par exemple les explorations du catalogue de la BnF par Pierre-Carl Langlais sur son carnet <https://scoms.hypotheses.org/>, ou par Frédéric Glorieux sur <https://resultats.hypotheses.org>.

dans les notices, et identifiés comme correspondant à des concepts présents dans le référentiel cible. Par « sémantiser », on entend associer à un terme (une chaîne de caractères) une définition qui lui enlève toute ambiguïté quant à ses conditions d'utilisation : c'est le rôle de la notice d'autorité.

On voit ce processus de sémantisation à l'œuvre dans les nouvelles règles de catalogage publiées au fur et à mesure dans la publication de RDA-FR, transposition française du code de catalogage RDA[23]. L'objectif est de renseigner sous une forme codée toute information qui ne soit pas pure *transcription* (la date, le nombre de pages, ou encore le numéro de volume ou d'édition) d'une manière beaucoup plus systématique qu'aujourd'hui. Dans le futur format de catalogage en cours d'élaboration à la BnF, pour son catalogue interne (« Intermarc nouvelle génération », ou Intermarc NG), l'ensemble des valeurs codées seront adossées à des notices d'autorité selon le même mécanisme que Rameau actuellement, et non plus seulement à des listes de valeurs contrôlées.

L'information ainsi inscrite dans les notices doit alors être compréhensible et exploitable à la fois pour l'être humain et pour la machine. L'indexation rétrospective automatisée sur des notices bibliographiques vise à convertir du texte (la valeur déjà présente dans la notice) en un autre texte qui est, lui, codé.

Les différents chantiers décrits ci-dessus offrent une typologie de cas qui glissent progressivement, mais sans l'atteindre encore vraiment, vers l'extraction d'entités nommées. Le chantier de reconnaissance des noms de lieux pour les collections du lettrage LK7 est ce qui s'y rapproche le plus : on sait que dans cette zone d'indexation locale va se trouver un nom de lieu, et qu'il s'agit du sujet du document. En cherchant quelle notice Rameau a comme libellé ce même nom de lieu, on réalise une forme d'alignement entre une valeur textuelle et une entité dans un référentiel.

Une entité « nommée » est une entité mentionnée par son nom, et non par son identifiant, dans un texte. L'extraction d'entités nommées consiste précisément en cela : analyser un texte (une phrase, un paragraphe, un document) et y repérer des termes qui correspondent à des concepts dans un référentiel choisi.

Il ne s'agit pas seulement de faire en sorte que le mot trouvé dans le texte corresponde au libellé d'un concept. Par exemple, si le terme « séisme » est présent, même dans l'hypothèse où il est à prendre au sens propre (et non pour qualifier un événement politique, par exemple), il peut aussi bien renvoyer à la sismologie (ark:/12148/cb11933257q), à l'étude ou des illustrations de séismes (ark:/12148/cb11933194n) ou à un cas spécifique (comme le séisme de Lisbonne de 1755, ark:/12148/cb146521983). Chaque terme doit donc être croisé avec ceux qui l'entourent pour favoriser l'une de ces hypothèses. C'est pourquoi cette technologie est essentiellement adossée à du *machine learning* : on donne au programme les moyens d'engranger d'une mémoire, à partir de laquelle elle va faire des choix chaque fois plus pertinents, grâce à la validation (ou à l'invalidation) des choix précédents. À la Deutsche Nationalbibliothek, le circuit du dépôt légal prévoit que les déposant fournissent dans un certain format le résumé et la table des matières associés au document déposé. Partant du principe que les termes présents

[23] À propos de RDA (*Resource Description and Access*), voir *supra* le chapitre de Françoise Leresche.

dans ces zones (résumé et table des matières) identifient les sujets dont traitent le livre, un programme les analyse pour y reconnaître le libellé (ou *point d'accès*) de concepts présents dans les Gemeinsame Normdatei (GND).

Le chantier de reprise du LK7 à la BnF n'utilise pas de mécanisme de *machine learning* pour valider les choix d'alignement entre un nom de lieu et son concept Rameau : la validité du résultat est assurée par la nature de la source elle-même, vu ce qu'on en connaît à l'avance (il s'agit d'un nom de lieu situé en France). Une autre différence cruciale avec l'extraction d'entité nommée réside dans le fait que l'indexation rétrospective s'appuie sur une indexation pré-existante, qui a été réalisée manuellement. Néanmoins on voit bien qu'il y a un glissement possible, par degrés successifs, qui nous rapprochent de l'expérience allemande.

Retour sur l'indexation en texte intégral

Ce que nous apprend également le mécanisme d'extraction d'entités nommées, c'est le lien entre l'indexation matière et l'indexation « plein texte ».

À l'origine, l'indexation consiste à élaborer des index, c'est à dire à lister les occurrences de mots et leur emplacement pour un texte (ou un corpus de textes). L'indexation sujet consiste à identifier, parmi tous les termes réellement présents dans un texte, quels sont les concepts effectivement mentionnés dans le texte, un même concept pouvant être successivement désigné par plusieurs termes distincts : un concept est alors un *cluster* de termes effectivement présents dans un texte (ce qui ne veut pas dire que ce concept est le sujet du document).

La dernière étape consistant à identifier parmi les concepts présents dans le texte, quels sont ceux qui sont réellement sujets du texte.

C'est précisément le cheminement qu'a parcouru le moteur de recherche Google lorsqu'il a constitué et exploité son *Knowledge Graph* à partir de 2012[24]. À l'origine moteur d'indexation en texte intégral, la société Google a développé à la fin des années 2000 une technologie pour identifier les « termes rejetés » d'un même concept, et constituer en entités ces grappes de termes, afin de mettre en avant ces entités dans les listes de résultats.

Cette évolution dans le fonctionnement du moteur de recherche web, tout autant que les différents chantiers de la BnF évoqués plus haut, témoignent de la continuité technique et intellectuelle entre l'indexation plein texte et l'indexation matière, en particulier dès lors qu'on a affaire à des ressources textuelles numériques, et de la difficulté de pouvoir clairement marquer une frontière entre les deux.

Indexation et algorithmique

Plusieurs raisons convergentes amènent les bibliothèques à explorer la piste de l'indexation automatique. En voici trois, qui sont déjà suffisantes.

D'abord, les ressources publiées sont chaque année plus nombreuses, parce que beaucoup moins coûteuses à produire pour leurs auteurs. C'est évidemment le cas pour les documents numériques (qu'il s'agisse de texte, de son ou de vidéo), mais cela concerne aussi les documents

[24] Amit Singhal, « Introducing the Knowledge Graph: things, not strings », *Official Google Blog*, 16 mai 2012, [En ligne] <https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html> (Consulté le 16/01/2019).

sur support physique : l'Observatoire du Dépôt légal indique[25] une augmentation de 12 % des livres imprimés entre 2012 et 2017, et pour la même période une augmentation de 33 % du nombre d'ouvrages auto-édités.

Ensuite, les ressources humaines disponibles pour en assurer le traitement documentaire ne vont pas s'accroissant dans les mêmes proportions.

Enfin, il y a une faisabilité technique, en particulier grâce à la politique d'ouverture et de partage des données, qui permet d'envisager d'utiliser des mécanismes d'alignement pour enrichir ses ressources.

La même ressource a été décrite ailleurs, dans un autre contexte, éventuellement dans un autre langage de description. Si on arrive à retrouver cette ressource « ailleurs » (en alignant sa propre notice bibliographique avec une autre notice bibliographique), on peut récupérer son indexation. Si cette indexation utilise le même langage documentaire (par exemple, Rameau), on la récupère, sinon on la transpose – à condition bien sûr de disposer d'un alignement entre les deux vocabulaires (ici, une conversion LCSH - Rameau).

En allant un cran technologique plus loin, on peut explorer l'extraction d'entités nommées, qui consistera à convertir non plus un terme contrôlé en un autre, mais une chaîne de caractères (par exemple un paragraphe correspondant au résumé d'un livre) en un ou plusieurs termes contrôlés. C'est l'extraction d'entités nommées. Un certain nombre de bibliothèques ont déjà commencé à explorer, voire à utiliser cette technologie, et il est vraisemblable que la Bibliothèque nationale de France s'y intéressera également. Mais il faut tout de même noter que l'exploitation d'alignements, qui manipule uniquement des métadonnées déjà structurées, est loin d'avoir été épuisée : des ressources non indexées en Rameau à la BnF, sont par ailleurs indexées dans le Sudoc (en Rameau), ou dans WorldCat (en LCSH), ou dans d'autres bases bibliographiques et dans d'autres vocabulaires, sans qu'il soit nécessaire de faire tourner un algorithme nécessairement plus complexe à base de *machine learning*.

Par ailleurs, le modèle IFLA LRM, qui porte l'indexation sujet au niveau de l'œuvre, ouvre d'autres perspectives : une même œuvre, dont certaines éditions ont paru avant 1980 (donc avec l'existence du vocabulaire Rameau) et d'autres après, pourra être indexée grâce aux éditions les plus récentes, sans qu'il paraisse alors pertinent, dans le cadre d'un catalogue LRMisé, de réindexer les *manifestations* anciennes.

Les processus d'alignement, le partage accru d'entités entre bibliothèques (en France, le projet de Fichier national d'entités va dans ce sens) l'adoption progressive du modèle LRM – tout concorde à enrichir en indexation structurée les fonds qui disposent aujourd'hui d'informations non exploitées parce que non conformes aux pratiques d'indexation actuelle.

Un ensemble de mécanismes nous offre encore de nombreuses pistes plutôt accessibles à explorer pour les années à venir, et la convergence de tous ces chantiers (Transition bibliographique, réforme Rameau, Fichier national d'entités[26], ouverture des données) favorise

[25] Bibliothèque nationale de France. Observatoire du Dépôt légal, *Observatoire du Dépôt légal : Données 2017*, 2018, [En ligne] <http://www.bnf.fr/documents/dl_observatoire_2017.pdf> (Consulté le 16/01/2019).

[26] Sur le projet de FNE, voir la rubrique <https://www.transition-bibliographique.fr/fne/fichier-national-entites/> sur le site de la Transition bibliographique.

à tous points de vue l'indexation Sujet rétrospective des ressources, dont on a pu voir avec l'exemple du *Knowledge Graph* de Google combien elle demeurait pertinente pour faire exister les ressources des bibliothèques, et offrir ainsi au monde une information riche.

Bibliographie

Bibliothèque nationale de France. Observatoire du Dépôt légal, *Observatoire du Dépôt légal : Données 2017*, 2018, [En ligne] <http://www.bnf.fr/documents/dl_observatoire_2017.pdf>.

Brygfjeld (Svein Arne), Wetjen (Freddy) et Walsøe (André), « Machine learning for production of Dewey Decimal », dans Transform Libraries, Transform Societies, présenté à IFLA WLIC 2018, Kuala Lumpur (Malaisie), 26 août 2018, [En ligne] <http://library.ifla.org/2216/1/115-brygfjeld-en.pdf>, 9 p.

Casilli (Antonio), *En attendant les robots*, Paris, Seuil, 2019 (La Couleur des idées).

Cavalié (Étienne), « Des milliers de ebooks (et de liens !) dans le catalogue de la BnF », *Bibliothèques [*reloaded*]*, 9 juillet 2018, [En ligne] <https://bibliotheques.wordpress.com/2018/07/09/des-milliers-de-ebooks-et-de-liens-dans-le-catalogue-de-la-bnf/>.

Ghorbani (Mahboubeh) et Torkashvand (Fattaneh), « Lessons learned from Automatic Indexing Projects regarding to Persian Language Specifications », dans *Transform Libraries, Transform Societies*, présenté à IFLA WLIC 2018, Kuala Lumpur (Malaisie), 26 août 2018, [En ligne] <http://library.ifla.org/2215/1/115-ghorbani-en.pdf>, 9 p.

IFLA, « Bibliographic control », 5 décembre 2017, [En ligne] <https://www.ifla.org/best-practice-for-national-bibliographic-agencies-in-a-digital-age/node/8911>.

Junger (Ulrike), « Automation first – the subject cataloguing policy of the Deutsche Nationalbibliothek », dans *Transform Libraries, Transform Societies*, présenté à IFLA WLIC 2018, Kuala Lumpur (Malaisie), 26 août 2018, [En ligne] <http://library.ifla.org/2213/1/115-junger-en.pdf>, 10 p.

Singhal (Amit), « Introducing the Knowledge Graph : things, not strings », *Official Google Blog*, 16 mai 2012, [En ligne] <https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html>.