Aligner ses données avec le catalogue BnF et le Sudoc : la genèse de Bibliostratus

Bibliostratus, logiciel d'alignement de données bibliographiques avec le catalogue de la BnF et le Sudoc, a été lancé en avril 2018 par le groupe Systèmes & Données, suite à l'identification d'un besoin de régénération de catalogue au sein du groupe. Voici l'histoire de sa naissance !

Au commencement

Au commencement, l'un des membres du groupe Systèmes & Données, travaillant au sein du réseau des médiathèques de Montpellier, souhaitait travailler au réalignement de ses données avec le catalogue de la BnF, notamment pour enrichir ses notices.
La BnF a accepté d'accompagner ce projet à condition que celui-ci soit documenté au fur et à mesure pour donner à voir une méthode de travail utile à d'autres. Même si le besoin exprimé par les médiathèques de Montpellier n'était pas prioritairement relié à la transformation du modèle de données (et à terme, des catalogues eux-mêmes), il est rapidement apparu que le projet pouvait constituer une pierre angulaire de la Transition bibliographique.
En effet, lorsque les agences bibliographiques nationales auront atteint une masse critique de FRBRisation de leurs catalogues respectifs, comment les bibliothèques pourront-elles dériver ces données de manière simple et fiable sans lien entre leurs notices et celles des agences ?
L'objectif initial de ce projet n'était donc pas de développer un logiciel, mais de définir un processus de travail, systématique et méthodique, à l'aune de l'exemple montpelliérain. C'est *a posteriori* qu'il est apparu faisable et plus pertinent de concevoir un logiciel fondé sur les différentes étapes identifiées (clés d'alignements, contrôles, solutions alternatives en cas d'absences de résultats, etc.).

Mai 2017 : premiers contacts

La nature de la collaboration a rapidement été définie : la BnF accompagnera les médiathèques de Montpellier « à titre d'exemple », ayant valeur d'expérimentation pour elle. Elle ne se lancera pas dans une nouvelle offre de services aux professionnels des bibliothèques, mais elle s'engagera à élaborer, avec l'établissement partenaire, une boîte à outils et ses modes d'emploi pour lui permettre de s'aligner avec BnF catalogue général. L'une des exigences était que toute bibliothèque souhaitant réaliser le même travail devra, elle aussi, pouvoir à terme exploiter facilement cette boîte à outils.

Montpellier a alors fourni plusieurs fichiers :

- Les monographies avec ISBN ;
- Les monographies sans ISBN ;
- Les périodiques avec ISSN ;
- Les périodiques sans ISSN.

Juin 2017 : Open Refine comme premier outil de travail

Le logiciel Open Refine a alors été identifié comme un outil idéal :

- Il est *open source*, gratuit, facile d'installation et d'utilisation ;
- Il permet de nettoyer des données grâce à un ensemble de fonctionnalités qui en font un outil très puissant ;
- Il permet d'interroger systématiquement, ligne à ligne, un web service comme le SRU de la BnF.

En effet, les données des catalogues pouvaient être très hétérogènes et nécessiter des opérations de nettoyage variées, difficiles à devenir génériques. Voici quelques cas emblématiques rencontrés lors de l'expérimentation avec Montpellier :

- Les zones ISBN contenaient généralement un ISBN. Mais parfois elles pouvaient en contenir deux ; ou contenir un ISSN ; ou la valeur « Br. » ou « 2- » ;
- Les titres pouvaient mélanger Titre et Auteurs ;
- Les dates pouvaient contenir « sans date » ou le nom de l'éditeur,ou les précisions « DL », « cop. », « Impr. », etc.

Pour chaque lot indiqué, on a donc défini une stratégie d'alignement, exploitant en priorité les identifiants internationaux (ISBN, ISSN) après avoir vérifié qu'il s'agissait bien d'ISBN/ISSN (présence de lettres, nombre de caractères).

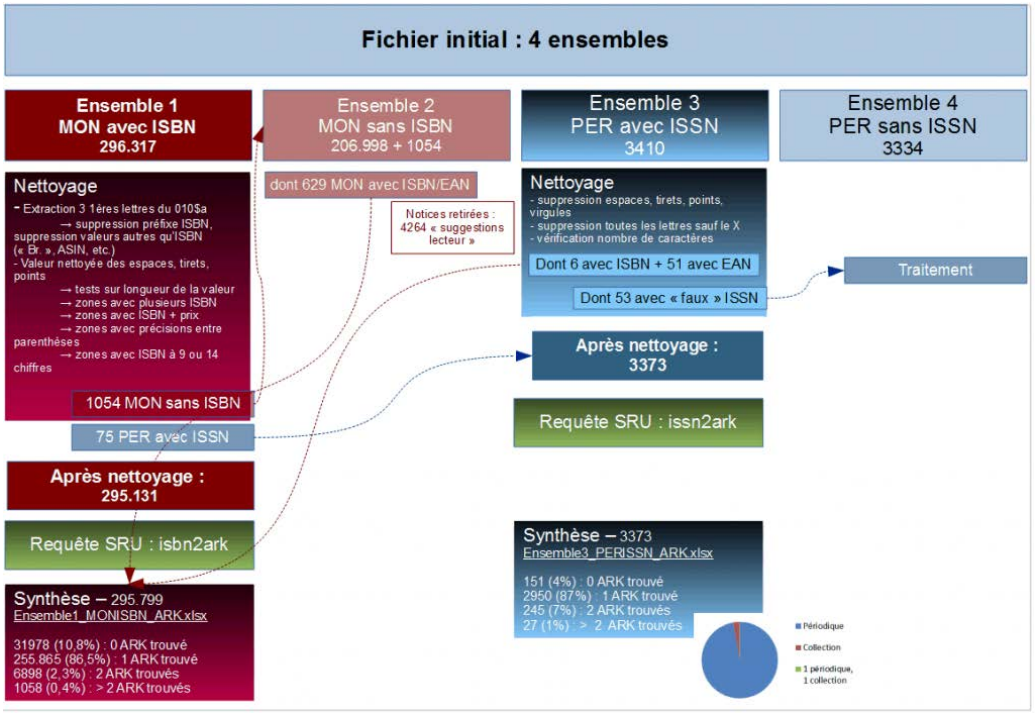

Tableau de suivi des différentes étapes d'alignement : volumétrie des notices traitées à chaque phase

Automne 2017 : premiers résultats et limites de l'outil et de la méthode

Les taux obtenus se sont avérés plutôt satisfaisants : 87% d'alignements uniques pour les notices avec ISBN ou ISSN.

Néanmoins un certain nombre de problèmes sont apparus à l'issue de ces expérimentations :

- La première version du document fournissant la méthode (avec mode d'emploi pour prise en main du logiciel) faisait 28 pages ;
- Le processus nécessitait plusieurs opérations successives en cas d'échec d'alignement. Donc, à chaque opération d'alignement, il fallait distinguer les notices pour lesquelles la requête a renvoyé des résultats de celles qui ont échoué, pour relancer une nouvelle opération d'alignement.

A l'analyse, la documentation de 28 pages décrivant les étapes systématiques une à une n'était rien d'autre qu'un algorithme – exprimé en langage naturel et non en langage informatique. Ce mode d'emploi a permis de constater qu'il était possible de standardiser les nettoyages et la restructuration des zones pour identifier et manipuler les ISBN, les titres, les auteurs et les dates.
C'est pourquoi en novembre 2017, la BnF a entamé le développement et les tests d'un petit programme d'alignement, s'affranchissant d'Open Refine. Il ouvrait d'ores et déjà des perspectives intéressantes, mais trop tardivement pour être évoqué comme une piste fiable lors de la journée professionnelle Systèmes & Données du 14 novembre 2017.

Avant Bibliostratus : une version alpha du logiciel

Malgré tout, les premiers tests ont permis de conclure qu'il était possible de fournir un traitement générique assez fiable :

- On pouvait tester si un « ISBN » (ou plutôt l'information présente dans la zone 010$a en Unimarc) est bien un ISBN (séquence de 10 ou 13 caractères, dont tous sauf le dernier sont des chiffres) ;
- On pouvait nettoyer les zones de dates ;
- On pouvait relancer facilement une nouvelle recherche si la précédente a donné 0 résultat, etc. ;
- Et on pouvait mettre en place des contrôles pour éviter les faux positifs (alignements proposés non pertinents).

Si les éléments d'informations n'étaient pas aux bons endroits, le programme renvoyait 0 résultat – à charge pour l'établissement de faire du nettoyage dans ses données : les tests d'alignement, rapidement effectués, permettaient de prendre la mesure du problème.

En somme, avec cette solution, la bibliothèque se concentrait sur l'analyse et l'exploitation des résultats obtenus, plutôt que sur le tutoriel pour les obtenir. Le réseau des médiathèques de Montpellier a d'ailleurs repris la main par la suite sur ses alignements pour enrichir son catalogue, en injectant d'abord les ARK à l'intérieur de ses propres notices, puis en important les notices BnF complètes.

Des traitements génériques des données grâce à Bibliostratus

En parallèle, le groupe Systèmes & Données s'est impliqué dans le projet et c'est ainsi que la première expérience a donné naissance au logiciel *Bibliostratus : Stratégie d'alignement d'URIs pour la Transition bibliographique*. Dans ce nouveau cadre, plusieurs établissements essentiellement de lecture publique mais également universitaires, ont participé aux tests et débogages (la phase de « recette » au sens informatique du terme), à ses évolutions, à sa mise à disposition et à sa documentation sur GitHub ou encore à la mise en place d'un forum sur Agorabib.
Mais au-delà, le groupe a souhaité mettre en place des démonstrations de Bibliostratus, afin d'exposer son fonctionnement aux gestionnaires de métadonnées et de le soumettre à la

critique. En effet, sans cette appropriation collective, le risque serait que Bibliostratus devienne une espèce de boîte noire.

Il ne s'agit pas de faire de Bibliostratus un « chapeau de prestidigitateur » : on y entre un catalogue, il en sort des alignements, *« on ne sait pas trop comment, mais ça marche »*. Il s'agit plutôt de s'appuyer sur l'expertise des bibliothécaires au sujet de leurs données et de transformer cette connaissance en algorithmes, au sein d'un logiciel libre et gratuit, accessible à tous et documenté.

C'est bien dans cette perspective qu'aura lieu le prochain atelier Bibliostratus du lundi 5 novembre 2018 organisé par le groupe Systèmes & Données à la BnF.

En savoir plus

- Page Bibliostratus : mettre en correspondance ses notices avec celles des catalogues des agences bibliographiques ;
- Vidéo de la présentation de la collaboration entre la BnF et les médiathèques de Montpellier, lors de la journée professionnelle Systèmes & Données du 14 novembre 2017 ;
- Téléchargement de la dernière version disponible de Bibliostratus sur Github ;
- Contact du groupe Systèmes & Données.

Précédemment publié sous l'ancienne URL https://www.transition-bibliographique.fr/2018-09-24-aligner-donnees-avec-catalogue-bnf-genese-bibliostratus/