https://www.transition-bibliographique.fr/2018-02-14-frbrisation-catalogues-traitements-algorithmiques-abes-bnf/

Chantiers de FRBRisation par traitements algorithmiques de l'Abes et de la BnF

Cet article est le compte-rendu de la session parallèle n°3 « Qu'attendre des agences pour la FRBRisation des catalogues ? Chantiers de FRBRisation par traitements algorithmiques de l'Abes et de la BnF » de la 2e journée professionnelle du groupe Systèmes & Données, Métadonnées en bibliothèques : attention, travaux !, qui s'est tenue le mardi 14 novembre 2017 à la BnF.
Le support, les notes explicatives et les réponses aux questions posées lors de la session, ont été réalisés par les membres du groupe Systèmes & Données : Wilfried ARRONDEL, Étienne CAVALIÉ, Marianne CLATIN (coordinatrice du groupe), Philippe LE PAPE, Françoise LERESCHE, Nuria PASTOR MARTINEZ, Véronique PLAUT ; avec la participation d'Olivier ROUSSEAUX.
Si malgré toute l'attention apportée à l'écriture de ce billet, certaines informations étaient erronées, vous pouvez nous le signaler en nous contactant via ce formulaire.

Les agences bibliographiques mènent parallèlement des travaux de FRBRisation. Cet atelier présentait les chantiers en cours avec les outils utilisés de part et d'autre, leurs calendriers, leurs méthodologies, ainsi que le travail permanent de collaboration et de rapprochement entre les deux bases. Il précisait aussi la manière dont les établissements des deux réseaux pourront en bénéficier.

En complément de cette présentation et lorsque cela s'avérait nécessaire pour la compréhension, les membres du groupe Systèmes & Données ont tenté de synthétiser les compléments d'information à certaines diapositives ou les éléments de réponse aux questions posées, dans les notes ci-dessous :

Diapo 4 :

Le contexte de 2013 a évolué depuis et l'échéance de cette fin de vie programmée n'est plus d'actualité en 2017.
Le temps passé par les quelques bibliothécaires de l'Abes ayant contribué aux spécifications et tests de validation des résultats avant mise en production est difficile à évaluer. La mise en oeuvre technique des algorithmes ayant nécessité un changement de version du coeur du système informatique hébergeant le catalogue Sudoc, le calendrier initial en a été assez sérieusement modifié. Cette diapositive ne mentionne pas le travail technique, notamment lors de la mise en production des algorithmes ayant eu lieu mi-octobre, et qui a fortement sollicité l'équipe informatique en charge du Sudoc.
Les "quelques (bouts de) de bibliothécaires" évoqués sont là en écho de la présentation BnF qui mentionne "une équipe d'experts Biblio et Autorités".

Diapo 6 :

Par mode "vitrine", il faut comprendre un fonctionnement dans lequel les données générées par les algorithmes (notices de regroupement ou liens bibliographiques vers ces dernières) sont rendues visibles aux catalogueurs du réseau Sudoc dans l'interface professionnelle de catalogage mais ne sont pas modifiables (aucune création / modification / suppression n'est possible).
– Pour les agrégats (ressources contenant plusieurs expressions, d'oeuvres généralement différentes), la structure actuelle des données en format Unimarc rend très complexe leur exploitation. L'association des différentes oeuvres à leurs auteurs respectifs a été jugée insuffisamment pertinente lors des tests effectués.
– Les ressources continues ont la caractéristique d'être le plus souvent dépourvues d'accès auteur. La clé de regroupement utilisée par les algorithmes étant fondée sur les titres + auteurs présent, elle en devient insuffisamment pertinente pour ce type de ressource.
– La limite portant sur l'absence "de notice de regroupement pour les notices bibliographiques seules à représenter une oeuvre" n'est pas une limite technique. Elle correspond seulement au choix de générer une notice de regroupement quand il y a … quelque chose à regrouper (plus d'une notice bibliographique). Elle pourrait être levée dans la perspective d'une création systématique de notice d'oeuvre pour toute les manifestations.

Diapo 8 :

Limites vis à vis du modèle :
– Niveau de l'expression non traité.
– Pas de mécanisme d'héritage : les accès au concepteur de l'oeuvre, de même que les accès sujet ou le résumé, par exemple, restent présents dans les notices bibliographiques et sont seulement dupliqués dans les notices de regroupement.

Diapo 9 :

Un premier bilan qualitatif est prévu pour la mi-2018.
Les suites données à cette expérimentation (et notamment la mise en oeuvre de tout ou partie des "chantiers à prévoir") dépendront pour partie de ce bilan.

Cette expérimentation se déroule dans l'environnement de production du Sudoc. Elle n'en demeure pas moins à l'heure actuelle une expérimentation. Pour cette raison, dans cette phase, aucune des données générées ne sort du Sudoc :
– ni en direction des SIGB des établissements membres du réseau Sudoc ;
– ni dans l'interface publique du catalogue Sudoc (mais il est bien entendu que la destination finale d'une telle opération est de rendre service à l'usager) ;
– ni dans l'interface de consultation des autorités IdRef.

Diapo 10 :

Plan identique.
Points de départ différents -> méthodes et outils différents -> à la fin, résultats BnF et résultats Abes à comparer et enrichir réciproquement.

Diapo 11 :

Par rapport à situation de l'Abes, la BnF dispose de 3 caractéristiques majeures qui expliquent que la BnF ait pris des directions différentes :
1. Un trésor de guerre de plusieurs dizaines de milliers de notices d'autorité Titre, qui peut servir à constituer une "base d'oeuvres".
2. Une équipe dédiée à la reprise des données du catalogue, dont l'activité peut se concentrer au moins en partie sur des chantiers liés à la Transition bibliographique.
3. Un projet d'expérimentation (en interface, en modélisation des données et en algorithmes de retraitement automatique) sur le modèle FRBR. En 2016, certains de ces algorithmes ont été extraits de data.bnf.fr pour en faire une plate-forme de traitement de données à part entière, permettant à la BnF de faire tourner ces traitements non plus seulement sur tout le catalogue, comme le fait data.bnf.fr, mais sur des corpus sélectionnés : c'est le logiciel RobotDonnées.

Diapo 13 :

Chantier particulier sur les agrégats (ouvrages regroupant plusieurs oeuvres, par exemple le tome I des Rougon-Macquart d'Emile Zola dans la collection Bouquins chez Laffont) : identification de notices d'autorité Titre déjà existantes, pour les documents agrégatifs identifiant la liste des oeuvres qu'ils contiennent d'une manière suffisamment structurée.

Diapo 14 :

Choix du corpus (auteurs français du XXe siècle) parce que :
1. Dépôt légal censément exhaustif -> la BnF possède l'édition la plus ancienne et peut donc déterminer le titre d'origine et la date de création (première publication) de l'oeuvre.
2. Une première expérimentation sur la constitution d'oeuvres par regroupement de notices bibliographiques avait été menée en 2016 dans le cadre du projet ReLire (oeuvres indisponibles du XXe siècle).

Diapo 15 :

Principes de RobotDonnées : à partir d'une liste d'auteurs (c'est-à-dire d'une liste de leurs identifiants : "NNA" pour "Numéro de Notice d'Autorité"), on réalise une succession d'étapes, en faisant tourner une succession d'algorithmes.
Ici, pour identifier des oeuvres à partir de leurs auteurs :
1. [bleu] WScatalog extrait et nettoie les formes de titres.

2. [violet] Minhashing regroupe les formes de titre identiques ou presque (ou minshift, un autre algorithme de regroupement).
3. [vert] Dedupe regroupe des formes de titre qui n'ont rien en commun mais désignent bien la même oeuvre (exemple : titre traduit / titre original).

Diapo 16 :

Après avoir regroupé des manifestations en oeuvres (cf. diapos précédentes), un dernier algorithme (Merge) permet de constater que 2 auteurs homonymes ont publié des oeuvres aux titres identiques :
s'il s'agit d'un doublon, il faut les fusionner ;
s'il s'agit d'homonymes avérés, il faut réattribuer toutes les notices bibliographiques portant le même titre à un seul et même auteur.

Diapo 17 :

Autre exemple d'algorithme : PRR crée un lien entre notices bibliographiques et notices d'oeuvres qui ont le même auteur et portent le même titre.

Diapo 18 :

Exemple de processus de traitement d'un corpus avec RobotDonnées : de l'extraction de formes de titre à la création d'oeuvres en vue de leur chargement dans data.bnf.fr.

Diapo 19 :

D'ici le printemps 2018, les oeuvres des auteurs français du XXe siècle auront été traitées (du moins pour les auteurs présents dans data.bnf.fr). Puis, courant 2018, on élargira ce processus aux imprimés du XXIe siècle et à d'autres supports.
Parallèlement, des chantiers de correction et d'amélioration des notices dans le catalogue sont et seront menés avec la perspective que cela permettra aux algorithmes de mieux les prendre en compte.
Sans abandonner le traitement unitaire de notice/document, on accorde une part croissante au traitement de masse sur des éléments d'information (des données, donc).

Diapo 20 :

Après une phase de fonctionnement indépendant l'un vis-à-vis de l'autre des deux processus de FRBRisation, en raison des contextes et environnements informatiques propres à chacune des institutions, devrait venir une phase de recherche de convergence (étude) et de résultats partagés, sachant qu'est déjà mené un travail de réalignement des données actuelles, notamment autour des identifiants.

Diapo 21 :

Expérimentation avec la BM de Montpellier pour voir comment faciliter la récupération par les bibliothèques de lecture publique du travail de FRBRisation, en ré-associant chacune des notices locales avec une notice BnF (identifiant de l'ARK BnF pour chaque notice bibliographique ou d'autorité d'une bibliothèque).

Questions à la BnF :

Question 1 : Quelle priorité est donnée à la résolution des problèmes de doublons pour les opérations de FRBRisation ?
La BnF va FRBRiser avec des doublons. Elle mène par ailleurs des chantiers de dédoublonnage, mais il est impossible d'attendre qu'il n'y ait plus de doublons dans son catalogue pour FRBRiser les notices.

Question 2 : Avez-vous récupéré des données du catalogue BnF Archives et manuscrits (BAM) ? Et pouvez-vous nous les réinjecter ?
BnF Catalogue général et BnF Archives et manuscrits partagent le même fichier d'autorité (renvoyant donc aux mêmes identifiants), donc les pages de data.bnf.fr renvoient vers les instruments de recherche de BAM qui sont liés à la notice d'autorité concernée par la page.

Question 2 bis : Exploitez-vous les "œuvres" de BAM ? Est-il envisagé de faire évoluer l'utilisation de l'EAD pour y rajouter des éléments d'information issus du modèle FRBR-LRM (références aux notices d'œuvres à l'intérieur des instruments de recherche notamment) ?
Il existe déjà dans BAM des liens vers des notices d'oeuvres (attribut authfilenumber de unititle/title). Exemple : avec le manuscrit Smith-Lesouëf 62 Le Roman de la rose, http://archivesetmanuscrits.bnf.fr/ark:/12148/cc9017p, ce qui permet dans BAM d'avoir un lien vers la notice d'autorité dans le catalogue général ; et dans la page data.bnf.fr Le Roman de la rose (http://data.bnf.fr/ark:/12148/cb166125510) de déclarer le manuscrit comme une manifestation du roman dans les triplets RDF et dans la rubrique « Documents d'archives et manuscrits » de la page web.

C'est une piste à creuser.

Questions à l'ABES :

Question 1 : Que voient les catalogueurs du Sudoc des résultats de l'opération de FRBRisation du 23 octobre 2017 ?
Ils voient, dans l'interface professionnelle de catalogage uniquement :
– un type d'autorité supplémentaire : notice de regroupement ;
– dans les notices bibliographiques regroupées par les algorithmes : une nouvelle zone de lien vers les notices de regroupement générées.

Ces données ne sont pour l'instant pas modifiables par les catalogueurs.

Elles ne sont pas non plus exportées vers les SIGB des établissements membres du réseau Sudoc, ni exploitées dans l'interface publique du Sudoc ou dans IdRef.

Question 2 : Les notices de regroupement sont elles stables ?
Elles sont stables dans la mesure où le catalogue l'est. Elles ne sont cependant pas figées. Les algorithmes de FRBRisation tournent quotidiennement en prenant en compte les évolutions apportées dans la journée aux données du catalogue et opèrent en conséquence des corrections/ajouts/suppressions, aussi bien dans le contenu des notices de regroupement que dans les liens depuis les notices bibliographiques vers les notices de regroupement. Ainsi, lorsqu'une notice bibliographique est modifiée, elle peut, selon les modifications qui lui ont été apportées,

contribuer différemment au contenu de la notice de regroupement à laquelle elle est liée, mais aussi, le cas échéant, se retrouver liée à une notice de regroupement différente.

Par ailleurs, si un regroupement de notices bibliographiques se réduit à une seule notice (par exemple si on dédoublonne une grappe de notices et qu'il n'en reste plus qu'une), alors la notice de regroupement précédemment générée pour la grappe disparaît (puisqu'elle n'existe que s'il y a plusieurs notices à regrouper).

Enfin, la stabilité de certaines notices de regroupement peut ne pas être signe de qualité, si le regroupement opéré par les algorithmes n'est pas pertinent.

Aujourd'hui, seule l'ABES est habilitée à modifier manuellement une notice de regroupement. En changeant le code de statut de la notice, cela a pour effet de la figer : l'algorithme ne touche alors plus cette notice qui cesse donc d'être mise à jour (enrichie) automatiquement par des apports ultérieurs en provenance de nouvelles notices bibliographiques entrant dans le catalogue ou de notices modifiées (ajout de variantes de titre, de points d'accès sujet, de résumé, etc.). Il en va de même pour les liens bibliographiques vers les notices de regroupement. Ils peuvent être validés manuellement et échappent alors à des d'éventuelles modifications automatiques ultérieures. Mais il est pour l'instant trop tôt pour entrer dans cette pratique qui priverait des bénéfices envisageables par des améliorations ultérieures des algorithmes.

Cette FRBRisation des données du Sudoc reste un chantier expérimental. Il faut maintenant passer par une phase de qualification de ce qui est « bon » dans ce processus automatique pour savoir ce qui mérite d'être stabilisé, partagé et exposé. Il existe aujourd'hui 1,5 million de notices de regroupement, mais combien méritent réellement de devenir des notices d'œuvres ?

Une exposition prématurée des résultats obtenus poserait à la fois la question de la qualité des contenus et celle de la pérennité d'un grand nombre des identifiants aujourd'hui attribués mais qui n'ont pas de pertinence. Ces notices de regroupement ne seront mises à disposition que lorsque des identifiants fiables (sans risque de disparition) pourront être garantis comme étant associés à des contenus suffisamment pertinents.

La BnF précise que l'on peut appliquer des processus de qualification par comparaison. Par exemple comparer si, pour les notices qui ont un ARK BnF, le regroupement effectué dans data.bnf.fr a été le même que celui effectué dans le Sudoc. Le qualificatif sera-t-il alors assez fiable pour valider côté ABES et côté BnF (d'où le travail d'alignement des deux base aujourd'hui) ?

Question 3 : L'ABES va-t-elle ouvrir des chantiers à réaliser par le réseau (ex. : modification des codes de fonction…) ?
L'expérimentation de FRBRisation est entrée depuis fin octobre 2017 dans une phase opérationnelle à l'échelle du catalogue Sudoc en production. L'ABES est pour le moment en mode "vitrine" (regarder sans toucher) sur ces nouvelles données, ce qui permet des premiers retours de catalogueurs, sur les listes de discussion, constatant par exemple pourquoi tel résultat est mauvais, quand il s'agit d'une erreur de catalogage (ex. : un mauvais code de fonction). Cela permettra ensuite d'identifier les limites qui relèvent de l'algorithme (et non des données).

Une évaluation qualitative des résultats obtenus (mi-2018) devra permettre de décider des suites à donner : améliorations à apporter aux algorithmes, modalités d'ouverture au catalogage courant des données de la FRBRisation (lesquelles sont à l'heure actuelle uniquement des données générées automatiquement), etc.

Parmi les chantiers envisagés (mais dont les périmètres restent à définir), il y a notamment ceux portant sur des corrections de données du catalogue lorsque des anomalies de catalogage courant ont un effet négatif sur les traitements automatiques de FRBRisation.

Question 4 : Les algorithmes créé par OCLC sont-ils propriété d'OCLC ?
Ce sont les algorithmes qu'OCLC a développés spécifiquement pour l'environnement technique CBS (le cœur technique du Sudoc) et sans doute en conséquence du code propriétaire. Dans le cadre de la prestation demandée à OCLC, l'ABES a cherché à en comprendre les principes et surtout les effets sur les données, mais n'est pas entrée dans des consile code. La réflexion a plutôt porté sur un paramétrage adapté aux données du catalogue Sudoc et à la pertinence du service rendu.

Question 5 : Pourquoi ne pas avoir utilisé l'expertise française de sociétés (ici Progilone) qui travaillent directement avec des professionnels impliqués dans RDA et LRM au niveau européen pour réaliser vos opérations de FRBRisation (puisque le résultat d'OCLC tel que présenté ne paraît pas très satisfaisant) ?
L'ABES a expliqué en introduction de sa présentation le contexte dans lequel l'expérimentation a pris naissance. Le choix a été fait de se saisir d'une opportunité (l'existence du processus de regroupement de données proposé par OCLC sur son système CBS hébergeant le Sudoc) permettant de générer un nouveau corpus de données, portable dans tout futur environnement technique pour le Sudoc en cas de migration de système, sans investissement financier et technique majeur. Les résultats obtenus relèvent d'une expérimentation à l'échelle d'un catalogue de 16 millions de notices bibliographiques. Leur pertinence reste à évaluer et l'ABES ne préjuge pas actuellement des suites qui seront données ni des conditions dans lesquelles elles le seront.

Question 6 : Comment les algorithmes d'OCLC distinguent-ils les Œuvres et les Expressions ? Dans le cas des cartes, distinguent-ils les cartes de France (la carte de France pouvant être considérée comme l'Œuvre) à différentes échelles (l'échelle étant un attribut de l'Expression) ? Comment font-ils les regroupements quand une Œuvre change de titre (par exemple, la carte topographique de la France au 1:50 000 qui devient la "Série Orange") ?
Rappel de modélisation : Dans l'analyse française, une œuvre cartographique ne se définit pas uniquement par le territoire cartographié et le créateur de la carte ; d'autres critères sont discriminants : la nature de la carte (carte topographique, géologique, routière, etc.) et l'échelle (qui peut être considérée comme un attribut d'Expression représentative selon la terminologie du modèle LRM). En conséquence, la carte topographique de la France au 1:50 000 établie par l'IGN est une œuvre différente de la carte topographique de la France au 1:25 000 également établie par l'IGN.

Comme il a été précisé dans la présentation, les algorithmes ne permettent pas de faire apparaître le niveau Expression.
Pour les cartes, comme pour les autres types de ressources, les regroupements de notices bibliographiques sont opérés par des rapprochements effectués sur la base de clés "titre-auteur" générées à partir du contenu des notices. Ceci peut aboutir parfois à des regroupements intempestifs, notamment lorsqu'aucun auteur n'est mentionné dans les notices, lorsque les titres sont très courts, ou au contraire lorsqu'ils sont longs mais qu'ils ne diffèrent que par des mots présents en fin de titre. Ce dernier point en particulier peut concerner les ressources cartographiques pour lesquelles la mention d'échelle apparaît dans le titre propre.

Cet écueil a néanmoins pu être partiellement évité en faisant jouer les titres de partie pour opérer les regroupements.