

Alignements des données : s'habituer à finir plus... sans finir¹

Comment, dans le contexte de FRBRisation progressive, commencer le chantier de l'alignement des données, qui ne peut plus être repoussé ? Le groupe Systèmes & données du programme Transition bibliographique a créé et expérimenté, avec le Département des métadonnées de la BnF, un logiciel libre qui tire parti des bases et API de la BnF et du Sudoc.

Le programme Transition bibliographique a été lancé en novembre 2015, un an après la publication du communiqué officialisant la position française par rapport à RDA. Ce sont donc plus de deux ans qui se sont écoulés, et même quatre, si l'on considère les prémices de l'adaptation du code RDA à l'analyse catalographique française. Ce programme ambitieux est un défi pour les acteurs des systèmes d'information documentaire et des données bibliographiques. Ils pourraient soit être tentés d'adopter une attitude d'attentisme (tant que le code RDA-FR n'est pas complet et de nouveaux formats d'échange entérinés) ; soit, pour ceux qui sont allés plus vite dans la FRBRisation de leurs catalogues, se confronter à l'écueil de l'ensilage et de la perte des liens avec les métadonnées des agences bibliographiques. Or, tout avance en parallèle – évolutions de l'Unimarc, projet InterMarc-NG, publication de RDA-

FR, FRBRisations itératives ou/et incrémentales des agences – et l'écosystème doit s'adapter à un contexte nouveau.

UNE PÉRIODE DE COHABITATION

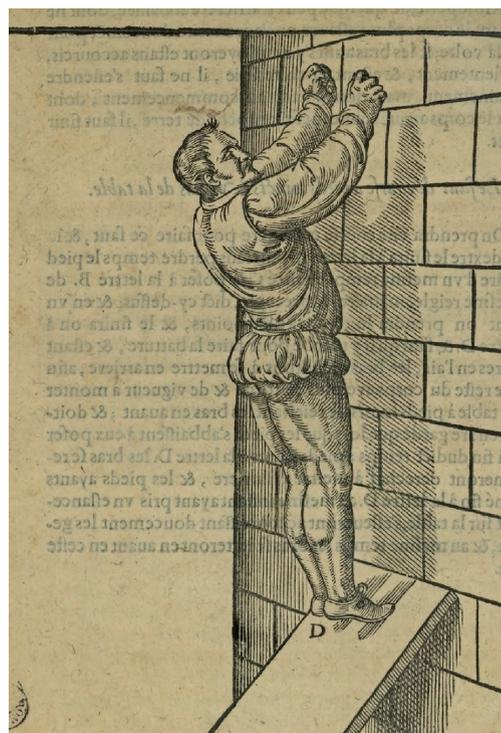
Le choix français étant celui de la transition plutôt que de la rupture, les éléments du code RDA-FR sont publiés chapitre après chapitre plutôt qu'en une unique livraison. Et, bien qu'une étape importante ait été franchie en août 2017 avec l'annulation de la norme Afnor NF Z 44-061 (remplacée par le Chapitre 9 – Identification des personnes de RDA-FR), une phase de cohabitation délicate s'est désormais ouverte.

De plus, une fois des éléments du code RDA-FR publiés, leur mise en application ne peut être que progressive. En effet, trois conditions, au moins, doivent impérativement être remplies : leur implémentation au sein de l'outil de production ; la formation des catalogueurs ou « rédacteurs de métadonnées » ; et, pour accueillir les nouvelles zones, l'évolution des formats d'échange – Unimarc dans le contexte français, mais rappelons que la BnF propose également un InterMarc de diffusion. Dans ce contexte mouvant, la décision de l'Abes et de la BnF d'adopter un calendrier commun pour l'application des mises à jour de RDA-FR avec une synchronisation prévue à partir de janvier 2019 a produit un « choc de simplification » pour les ré-utilisateurs, c'est-à-dire les éditeurs et administrateurs de SIGB et de métadonnées.

LES LIENS QUI LIBÈRENT NOS DONNÉES

Pour autant, la situation n'est pas totalement stabilisée, comme l'a montré, en 2017, l'imbroglio autour de la proposition, émise en 2016 par le Comité français Unimarc, d'une zone 214 – pour les éléments de l'ex-zone de l'adresse – refusée par le Permanent Unimarc Committee². Cet état de fait a contraint les agences et leurs ré-utilisateurs à définir une zone à usage national et provisoire, la 219, sans réelle visibilité sur sa pérennité – poussant légitimement certains acteurs à surseoir à son implémentation jusqu'à la prochaine décision du PUC en mars 2018...

Le succès des récentes journées professionnelles



Source : Gallica - BnF

➔ Trois dialogues de l'exercice de sauter et voltiger en l'air... par le Sr Archange Tuccaro, 1599, p. 146v.

[1] Dans une lettre de 1954 à son marchand d'art, le peintre Nicolas de Staël écrit, à propos de son travail : « Il faut s'habituer à finir plus, sans finir. Ce n'est pas facile. »

[2] rda.abes.fr/2017/04/01/souvenir-de-lisbonne

autour de la FRBRisation démontre l'intérêt des éditeurs et des administrateurs de SID/SIGB ainsi que celui des fournisseurs de métadonnées. La journée d'études de 2016 du groupe Systèmes & Données avait d'ailleurs permis de prendre connaissance des voies innovantes empruntées par quelques-uns de ces acteurs (sans ambition d'exhaustivité ni de labellisation)³.

Si les stratégies diffèrent fortement face au calendrier au long cours de la transition bibliographique, des questions récurrentes sont posées : « Où en est la FRBRisation des données des agences bibliographiques ? », « Pour une bibliothèque ou un éditeur en avance de phase sur la FRBRisation des agences, comment garantir l'alignement futur des données ? Et quelles méthodes et outils privilégier pour la récupération des premières FRBRisations des agences ?⁴ » Questions fondamentales car, rappelons-le, l'enjeu n'est pas uniquement l'adoption d'un nouveau code de catalogage ni de nouvelles interfaces de consultation, mais l'exposition et l'insertion des entités des catalogues des bibliothèques dans le *linked open data*. Il faut donc veiller à ne pas recréer (pour des raisons techniques, juridiques ou autres), à l'occasion de leur FRBRisation, de nouveaux silos sans liens avec le Web de données. C'est pourquoi l'un des axes de travail et l'une des recommandations fortes du groupe Systèmes & Données est, aujourd'hui, l'alignement des propres données de son établissement avec celles des agences. Ce chantier – qui suppose d'abord de

connaître ses propres données ! – s'adresse prioritairement à la lecture publique, avant d'espérer s'élargir vers le réseau Sudoc en fonction des retours et besoins des bibliothèques de l'ESR.

MAKE OUR DATA GREAT AGAIN ?

Avec l'appui technique et scientifique du Département des métadonnées de la BnF, des expérimentations ont été menées par les médiathèques de Montpellier avec l'objectif de régénérer tout ou partie des zones à aligner et particulièrement l'identifiant ARK. En effet, les URI sont désormais les clefs indispensables pour ces alignements, quelles que soient les pratiques internes de catalogage. Cette première expérience a produit une méthode et des outils permettant d'analyser ses données, de les nettoyer et, le cas échéant et en fonction des résultats, de les aligner avec le catalogue BnF⁵. À la lumière de ses enseignements, un **logiciel libre** a été développé⁶ (voir l'encadré ci-dessous). Ces travaux (outils, résultats, documentation, etc.) produits par le groupe Systèmes & Données sont et seront à la disposition de tous les acteurs (éditeurs, fournisseurs, administrateurs etc.). Il appartient donc à cet écosystème de s'en emparer, voire d'y contribuer, afin de mener à bien l'immense chantier collectif et ramifié de reprise et d'évolution des données.

RENAUD AÏOUTZ

Pilote du groupe Systèmes & Données
renaud.aioutz@puy-de-dome.fr

[3] http://www.bnf.fr/fr/professionnels/anx_journees_pro_2016/a.jp_161109_inovations_metadonnees.html

[4] Retrouvez les questions – et les réponses ! – dans un compte rendu de la Journée professionnelle S&D tinyurl.com/J-SD-2018-quest

[5] Pour plus de détails, voir ce document en ligne sur bnf.fr/documents/jsyd2017_franceschi-cavalie.pdf

[6] Présentation en avant-première à un panel de bibliothèques pilotes le 19 mars 2018. Le logiciel et la documentation sont accessibles à cette adresse : github.com/Transition-bibliographique/alignements-donnees-bnf

● ● ● LE LOGICIEL DÉVELOPPÉ

L'outil développé par le groupe Système et données se compose, pour l'instant, de trois briques :

1) à partir d'un fichier Unimarc ISO 2709 en entrée (correspondant à l'export d'un catalogue de bibliothèque, potentiellement entier), le logiciel constitue autant de fichiers tabulés qu'il y a de types de documents ;

2) à partir de chaque fichier tabulé de notices bibliographiques, il mène une opération d'alignement, c'est-à-dire de mise en correspondance stricte, avec les notices de la BnF ;

3) à terme, le logiciel disposera d'une liste d'identifiants ARK BnF, afin d'extraire les notices bibliographiques, et, si besoin, les notices d'autorité liées.

À chaque fois, le programme effectue des tests d'alignements : d'abord sur

les identifiants internes BnF (contrôle sur le titre ou sur l'auteur) ; si le test est négatif, l'on utilisera les identifiants internationaux (contrôle sur le titre ou sur l'auteur) ; enfin, en cas de test toujours négatif, le programme se tournera vers les métadonnées en texte libre...

Conçu principalement à l'intention des bibliothèques municipales qui souhaitent récupérer les données de la BnF pour mettre à jour leur catalogue, ce logiciel prévoit donc des traitements *ad hoc*. Mais rien n'empêche de l'utiliser pour d'autres projets, y compris dans le contexte du réseau Sudoc. Par exemple, l'une de ses fonctionnalités permet de mettre en exergue des interactions fécondes entre la base de la BnF et celle du Sudoc. Il s'agit du cas particulier de l'utilisation – déli-

cate – de l'ISBN comme clef de recherche et en particulier du problème des documents absents à la BnF et de certaines réimpressions. Le programme cherche d'abord l'ISBN dans le catalogue. BnF. S'il ne le trouve pas, il convertit l'ISBN₁₃ en ISBN₁₀ (si l'ISBN commence par 978) ou l'inverse ; s'il ne le trouve toujours pas, il va interroger le Sudoc sur l'ISBN initial comme sur l'ISBN converti.

Si cet ISBN se trouve dans le Sudoc, c'est parce que :

- le document est signalé dans le Sudoc mais absent des collections de la BnF,
- un ISBN de réimpression différent de celui déposé à la BnF au titre du Dépôt légal est signalé. Dans ce cas, il est fort probable que la notice Sudoc comporte le FRBNF correspondant, et l'algorithme remontera ainsi à l'ARK de la notice BnF.