

Compte rendu de l'atelier Nettoyer vos données

L'objectif de l'atelier était de proposer une méthode pas à pas pour analyser son catalogue et le corriger. La présentation de chaque étape de cette méthode a largement été ouverte aux questions des participants, permettant une réflexion sur la façon d'appréhender les problèmes concrets que l'on peut rencontrer dans son catalogue.

Avant de nettoyer ses données, il convient d'abord de définir ce qu'est une donnée propre, ce qui dépend en réalité beaucoup du contexte dans lequel une bibliothèque se positionne et de ses objectifs. Trois strates ont néanmoins pu être définies : le dédoublonnage des données, l'homogénéité de la base, et le respect des normes. L'accent a été mis sur le fait de ne pas nécessairement chercher à avoir les notices parfaites, mais d'avoir un niveau de qualité des données adapté à son catalogue, ses usages et surtout aux moyens disponibles pour le maintenir.

Cette définition établie, des méthodes pour analyser son catalogue ont été partagées : l'archéologie du catalogue, l'échantillonnage de données, le requêtage via les outils de son SIGB et enfin la possibilité d'utiliser des outils tels que MarcEdit pour analyser ses données après export.

Cette analyse est une première étape nécessaire à la définition des corrections à effectuer, qu'il faut préciser en répondant à un certain nombre de questions :

-Quel est l'objectif de la correction ? Rendre plus visibles les notices au catalogue ? Homogénéiser les données ? Les mettre à jour pour respecter les normes ? Préparer une migration dans un catalogue collectif ?

-En pondérant l'importance des corrections à leur difficulté, quels chantiers sont prioritaires ?

-Quel est le périmètre précis des données à corriger ? En cas d'import ou de modification par lot : quels champs souhaite-t-on protéger et conserver ? Lesquels peuvent être écrasés ou perdus ?

Des méthodes de corrections par lot existent permettant soit d'intervenir sur un ensemble de notices de sa base pour en corriger certains champs, soit d'importer depuis un réservoir extérieur des notices d'une meilleure qualité, en acceptant de remplacer ses propres notices ou simplement en n'important que certains champs choisis. Trois outils ont été proposés permettant d'effectuer ces opérations sans être dépendant de son SIGB :

-Bibliostratus, un outil développé dans le cadre de la Transition bibliographique permettant de s'aligner avec le catalogue de la BnF et le Sudoc pour importer leurs données.

-MarcEdit qui est un outil permettant d'analyser ses données, et de faire des modifications par lot en dehors de son SI, donc sans risquer de faire d'erreur et d'écraser des données.

-OpenRefine, présenté plus succinctement, qui affiche les données sous forme de tableur et permet ainsi de les analyser et de les corriger.

Dans les échanges qui ont suivi, ces logiciels ont pu être présentés comme une solution possible aux problématiques spécifiques des bibliothèques d'archives. En l'absence d'un SIGB permettant une description correcte des documents, Bibliostratus permet de préparer une potentielle migration en récupérant des données complètes. Dans le cas d'un catalogue partagé avec une Bibliothèque départementale, MarcEdit permet de modifier et de corriger ses données sans risquer d'écraser celles des autres bibliothèques présentes dans la base.

Pour certaines corrections nécessitant un contrôle humain, il a été proposé l'organisation de chantiers de correction, ou de définir une liste de corrections simples, dites « au fil de l'eau » à appliquer dès lors que l'on intervient sur une notice.

Enfin, l'atelier s'est conclu sur une présentation de l'offre de données et de services de la BnF, et le partage d'une fiche pratique récapitulant les éléments présentés lors de l'atelier.