Metadata mining : *fouiller les données des catalogues ?*

Les grands catalogues de bibliothèques auraient-ils vocation à devenir des sources de *data mining* ? Pourquoi et comment ? Plaidoyer appuyé sur quelques cas d'usage rencontrés à la Bibliothèque nationale de France (BNF).

1

Les débats en cours sur le *text* et *data mining* (TDM), qu'ils soient opérationnels (comment fouiller, pour quelle recherche, quels services et avec quels outils ?) ou juridiques (de quel droit, à quel coût ?) embrassent potentiellement une grande variété de données aux périmètres en réalité très différents. On imagine ou on pratique déjà la fouille de données dans les données brutes de la recherche, dans les publications couvertes par les grandes bases de connaissance éditoriales et commerciales de l'information scientifique et technique, ou encore dans les bibliothèques numériques issues de la numérisation patrimoniale.

Pour le chercheur, il ne s'agit plus seulement de trouver et de lire des documents, mais aussi d'interpréter, d'analyser, de confronter et de « faire parler » par le calcul de grandes masses de données, et ce, dans des domaines disciplinaires qui recouvrent aussi bien les sciences dures que les sciences humaines et sociales, notamment dans le cadre du développement des « humanités numériques ».

LES BIBLIOTHÈQUES, NOUVEAUX LABORATOIRES DE FOUILLE

Pour les bibliothèques, un enjeu stratégique est de trouver leur juste place et leur valeur ajoutée propre dans l'invention de ces nouvelles pratiques et dans les dispositifs au sein desquels elles vont s'élaborer. Leur connaissance intime de la constitution et de la structuration des collections peut les conduire à devenir des fabricants et des fournisseurs de corpus de données à fouiller. Leur maîtrise du signalement et leur pratique du traitement informatique des catalogues et des données bibliographiques peuvent aussi les aider à structurer ces corpus et ces bases en les qualifiant, en les organisant. D'ores et déjà placées, par leurs missions, au plus proche des attentes et des pratiques de recherche des usagers, elles pourraient alors concevoir, gérer et animer par elles-mêmes les services et l'outillage de traitement des données.

Dans la bibliothèque du XXIe siècle, il faut donc imaginer (sur place et/ou en ligne) un laboratoire où un chercheur pourra venir fouiller de grands corpus de données et où le bibliothécaire sera en mesure de l'aider à effectuer et valoriser ces opérations en mettant à sa disposition des outils, des méthodes, des référentiels. Si les types de données à exploiter, par leur masse et leur nature, peuvent, pour certaines, représenter un changement radical de paradigme (on ne parle plus de livres, ni même de bases d'articles scientifiques) pour les professionnels, les fonctions et services à inventer s'inscrivent en réalité dans la continuité de leurs missions et de leurs pratiques.

DES MÉTADONNÉES À REVISITER : CAS D'USAGE

Cet horizon peut paraître lointain et futuriste, et pourtant, très près de nous, quelques cas d'usage m'ont récemment amené à penser que les bibliothèques, ou pour le moins les agences en charge de la maintenance des grands catalogues nationaux comme l'Abes et la BNF, pourraient d'ores et déjà proposer des services de *data mining* à partir de données que nous connaissons particulièrement bien car nous les avons créées : les notices bibliographiques et d'autorité, bref, les métadonnées de nos catalogues. Voici ces cas d'usage.

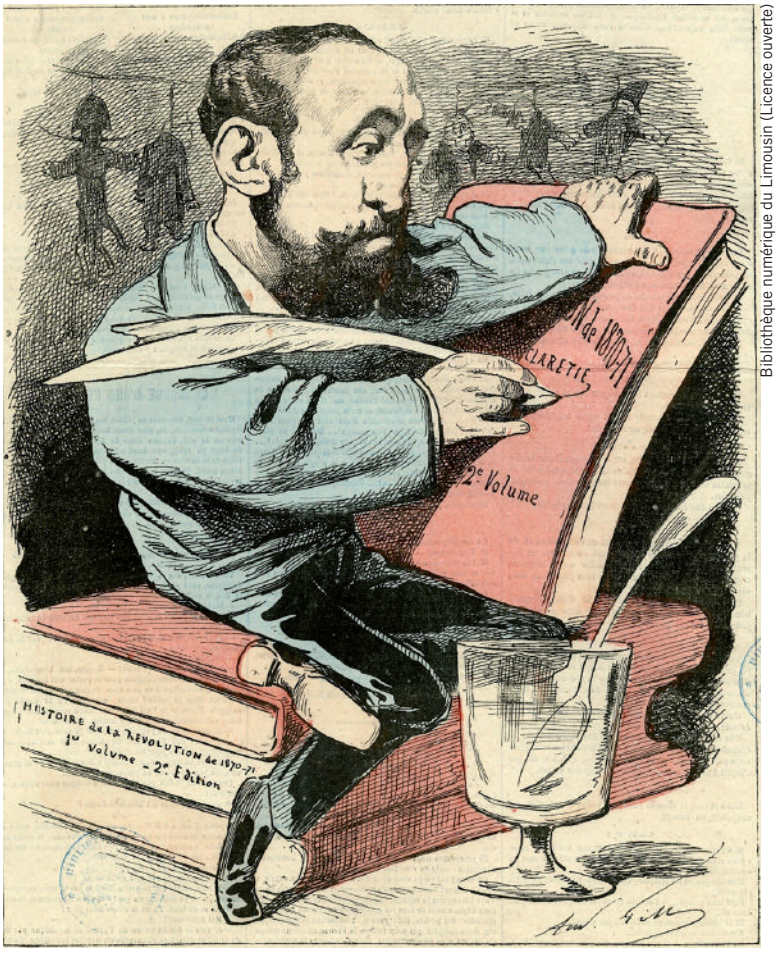
Un maître de conférence en sociologie dans un institut d'études politiques a récemment contacté la BNF car il conduisait une recherche sur la production de livres par les journalistes : pourquoi écrivent-ils ? Comment évolue cette production ? Afin de donner une dimension quantitative et statistique à son étude, il a eu recours au catalogue général de la BNF, dans lequel il a repéré près de 30 000 notices d'autorités correspondant à des journalistes. Pour ce faire, il a cherché à exploiter par lui-même de longues séries de notices d'autorité, mais s'est heurté à différents obstacles dont il s'est ouvert à ses correspondants de la bibliothèque.

À la même époque, une doctorante allemande dont le sujet de thèse est « *La circulation du savoir africain. Présence et réception de la littérature académique africaine en Allemagne et en France* » nous a également sollicités. Elle souhaitait relever et classer par nationalité actuelle l'ensemble des auteurs nés en Afrique, dans le domaine des sciences humaines, publiés en France depuis les années 60, dans n'importe quelle langue. Mais les fonctionnalités de recherche standard du catalogue de la BNF ne suffisent pas pour obtenir des données fiables et en masse, et pour tendre à l'exhaustivité :

il faut lancer des requêtes sur un ensemble de critères croisant les notices bibliographiques et les données d'autorités, requêtes pour lesquelles notre catalogue n'est absolument pas conçu. Dans un premier temps, nous avons essayé de l'assister en utilisant les outils professionnels habituels pour l'exploitation de nos métadonnées par d'autres bibliothèques ou des fournisseurs de systèmes de gestion de bibliothèque. Jusqu'à ce que l'on arrive au constat que ces outils n'étaient ni économiquement ni techniquement adaptés au besoin de cette doctorante.

ADAPTER LES OUTILS SIGNALÉTIQUES AUX NOUVEAUX BESOINS

Ces deux histoires nous ont ouvert les yeux sur une réalité : la demande de *data mining* est déjà à la porte de nos catalogues et nous devons prendre en compte les questions bien légitimes de ces nouveaux usagers dans l'évolution de nos dispositifs de signalement : pourquoi est-il si difficile d'extraire des notices en Marc des informations dans des formats simples d'exploitation bureautique par le commun des mortels ? Pourquoi, comme l'un de nos cobayes nous l'a fait remarquer, y-a-t-il « deux catalogues, l'un pour les autorités, l'autre pour les documents » ? Comment cerner dans le catalogue la population des journalistes alors qu'il n'y a pas de référentiel « emploi » dans les notices d'autorité personne ? Ces questions nous ont plus globalement interrogés sur la nécessité de rendre plus intelligible et plus explicite la structure du catalogue (et pas seulement des informations qu'il contient), de proposer des fonctionnalités d'extraction (sous forme de fichiers tabulés simples) et de retraitement massif de ses données adaptés à ce type d'usage, en donnant par exemple la possibilité aux chercheurs d'exploiter des catégorisations et indexations conçues initialement pour la gestion du catalogue et de la recherche documentaire, mais qui pourraient s'avérer extrêmement utiles pour ce type de traitement.
L'évolution engagée des catalogues vers les standards du web sémantique permettra sans nul doute d'aborder ces fonctionnalités de TDM au moyen d'outils plus adaptés que ceux des systèmes actuels, mais gardons d'ores et déjà cette idée : le catalogue de demain ne sera plus seulement un outil de recherche et de localisation d'information ; il faut aussi l'imaginer comme une vaste base de données à part entière que les chercheurs auront envie de faire parler.



Bibliothèque numérique du Limousin (Licence ouverte)

Jules Clarétie par André Gill, détail de la Une de *L'Eclipse*, n° 245, 6 juillet 1873. De nombreux journalistes, comme ici Jules Clarétie, sont également écrivains. Les métadonnées des catalogues des bibliothèques ne permettent pas d'effectuer un croisement judicieux entre ces 2 statuts à défaut de « référentiel emploi » dans les notices d'autorité personne.

GILDAS ILLIEN
Directeur du département de l'Information bibliographique et numérique, Bibliothèque nationale de France
gildas.illien@bnf.fr

LORSQUE LES MÉGADONNÉES CHASSENT LE BIG DATA

Le *Journal officiel* du 22 août 2014 a publié un avis de la Commission générale de terminologie et de néologie de l'informatique et des composants électroniques (CSTIC) qui propose d'utiliser le mot « mégadonnées », en lieu et place de l'expression anglaise « *big data* ».
Dès lors qu'il faudra parler du *big data*, c'est-à-dire ces « *données structurées ou non dont le très grand volume requiert des outils d'analyse adaptés* », les différentes administrations et autres établissements publics français devront adopter le terme « mégadonnées » proposé par les membres de la commission. L'expression « données massives » reste également admise.
Un peu plus tôt dans le mois (5 août), la même commission a préconisé de remplacer « *crowdfunding* », terme qui désigne le travail collaboratif d'internautes, par « production participative ».

www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000029388087&dateTexte=&categorieLien=id