Acquisitions Istex : *traiter et enrichir les données*

L'enrichissement et la valorisation des données fournies par les éditeurs est une des nombreuses tâches gérées par le département Adéle et l'équipe du hub de métadonnées de l'Abes. Méthodes et enjeux.

Dans le cadre du projet Istex (Initiative d'excellence de l'information scientifique et technique) initié par le ministère de l'Enseignement supérieur et de la Recherche, l'Abes, *via* son département achat de documentation électronique (Adéle), passe des marchés avec des éditeurs scientifiques afin d'acquérir des données pour la communauté scientifique française.

DONNEZ, DONNEZ LES MÉTADONNÉES

Par « données », il faut bien sûr entendre à la fois les contenus – bases de données, collections de revues numériques, e-books, etc. – et les métadonnées relatives à ces mêmes contenus. Effectivement, l'acquisition de métadonnées est un élément particulièrement discuté. Et pas seulement leur acquisition d'ailleurs. Car ces métadonnées, il faut pouvoir les manipuler, les traiter, les enrichir, les redistribuer. Au-delà de l'acquisition, ce sont donc surtout les droits d'utilisation et d'exploitation de ces données qui sont négociés. Le contrat de licence, signé par l'éditeur lors de la passation du marché, définit d'ailleurs précisément ces règles d'utilisation. Il permet ainsi de « *fournir les métadonnées à l'ensemble des bénéficiaires afin qu'ils puissent les intégrer dans leur catalogue local ou leur outil de découverte* ». Le contrat prévoit également la modification du format des métadonnées et leur enrichissement. Concernant leur format, la licence n'impose pas de standard particulier. S'il est préférable que les métadonnées soient livrées en XML, l'éditeur garde néanmoins la possibilité de fournir des données en format Marc. En revanche, le contrat insiste sur la nécessité d'inclure « *l'intégralité des informations bibliographiques disponibles* » (l'expression apparaît plusieurs fois dans le document, c'est dire !). Conclusion : le format importe peu, finalement, tant que l'éditeur nous procure l'ensemble des données qu'il détient sur ses propres contenus.

LE RÔLE DU HUB DE MÉTADONNÉES

• *Avant la signature du marché*

En collaboration avec Adéle, le hub de métadonnées joue un rôle important dans les acquisitions Istex, en intervenant dès la phase de négociations. En effet, le hub est impliqué dans la constitution des listes de titres à acquérir qui sont proposées aux éditeurs, en réponse à leurs premières offres de contenus. Bien souvent effectivement, les premières listes de titres transmises par les éditeurs sont erronées ; bibliographiquement parlant, s'entend. Pour les listes de périodiques, par exemple, dont nous connaissons bien la complexité – une revue, au contraire d'un e-book, a en effet une vie bien mouvementée. Le chemin parcouru, depuis sa naissance, n'est jamais simple : fusion avec un titre cousin, scissions en un, deux, trois titres, voire plus, etc. – l'éditeur ne fournit généralement que le dernier titre paru d'un périodique, alors que les dates transmises correspondent bien souvent à un ensemble de titres. Le travail du hub consiste donc à reconstituer l'historique des revues proposées. Pour ce faire, un outil a été développé : il s'agit de *Métarevues*, un programme bien pratique, basé sur les données Sudoc, qui permet de lister dans l'ordre chronologique, tous les titres liés à la revue proposée, à partir d'un ISSN donné. Cet outil est d'ailleurs utilisé par l'application *Périscope* de l'Abes – en plus d'afficher les états de collection des périodiques signalés dans le Sudoc, le service permet désormais de visualiser les historiques complets liés à ces revues[1].

Bien souvent, les premières listes de titres transmises par les éditeurs sont erronées ; bibliographiquement parlant, s'entend.

• *Après la signature du marché*

Après la signature du marché, lorsque les listes de contenus ont été fixées et que l'ensemble des données a été livré, le hub procède au travail de signalement des ressources, et pas seulement dans le Sudoc, même si cela reste l'une des priorités. Des données de qualité, standardisées, sont ainsi transmises aux établissements bénéficiaires et aux éditeurs de bases de connaissance, afin d'assurer le signalement de ces ressources dans d'autres environnements que le Sudoc.

En redistribuant ces données vers d'autres outils de recherche documentaire, le hub permet le développement de l'accessibilité et de la visibilité des contenus – et par là même des auteurs et des édi-

[1] Pour plus d'informations, consulter le billet sur le blog technique de l'Abes, Punktokomo : **http://punktokomo.abes.fr/2014/06/26/metarevues-un-outil-dedie-au-traitement-des-periodiques**

teurs de ces contenus – acquis en licence nationale. C'est également un moyen de sensibiliser les éditeurs au travail bibliographique.
Après l'étape de signalement, le hub de métadonnées travaille aux enrichissements de ces données. D'autres informations viendront s'y ajouter. Ces informations complémentaires peuvent provenir de différentes sources. Il s'agira, par exemple, d'ajouter des liens à des autorités comme IdRef, Viaf, Rameau, ou encore d'insérer des codes de langues…

L'HOMME ET LA MACHINE

Si l'ensemble de ces traitements est grandement facilité par l'exploitation d'outils et de programmes informatiques, l'analyse humaine reste indispensable pour détecter certaines anomalies et penser les corrections nécessaires. En outre, chaque type de ressources comprend son lot de difficultés. Le traitement des données de périodiques, par exemple, impose plutôt un travail d'expertise en amont, lors de la constitution des listes contractuelles, tandis que l'effort majeur livré sur le traitement des données d'e-books se concentre un peu plus sur l'aval de la chaîne, lors de l'étape de signalement. Le hub doit générer lui-même des notices Marc à partir des métadonnées transmises par les fournisseurs, au contraire des périodiques. En effet, grâce aux notices du centre ISSN déjà présentes dans le Sudoc, le travail de signalement des revues ne consiste généralement qu'à ajouter des données d'exemplaires et compléter certaines données bibliographiques, comme celles de la zone 207 (informations données sur les dates et les numéros de volumes).
Au-delà des difficultés liées à la nature de la ressource traitée, il faut également bien noter que chaque nouveau corpus fraîchement débarqué peut également (se) présenter à chaque fois (avec) ses propres complexités, (avec) ses propres cas particuliers. À ce véritable travail sur mesure s'ajoute une autre complication : puisque plusieurs négociations sont menées de front, l'équipe du hub de métadonnées – du reste relativement réduite – doit faire face à plusieurs corpus simultanément. Dans ce contexte, il est difficile de dégager des priorités, ce qui rend l'organisation du travail assez complexe. Car en établissant des priorités, on sacrifie forcément certaines tâches : faut-il privilégier le travail en amont de la signature ou le travail en aval ? Faut-il traiter le maximum de corpus le plus tôt possible ou viser un travail de qualité pour chacun d'entre eux, quitte à en retarder le signalement ?

Jean-Raphaël Guillaumin / Flickr (CC BY-SA 2.0)

Du pollen au miel : une orgie de données pour les abeilles du hub !

UNE COOPÉRATION AVEC L'INIST RENFORCÉE

En guise de conclusion, il convient de noter que la collaboration Abes-Inist doit s'intensifier dès la rentrée de septembre 2014, laquelle permettra d'optimiser le travail sur le traitement des données. L'Inist est effectivement l'un des acteurs majeurs du projet Istex. Il travaille notamment à la validation des données livrées et à la réalisation de la plateforme qui accueillera l'ensemble des contenus et des données acquis en licence nationale et dont la mise en œuvre est prévue pour fin 2015.

MARION GRAND-DÉMERY
Abes
grand-demery@abes.fr

LE HUB, QU'ES AQUÒ ?

Pour rappel, le hub de métadonnées est un projet porté par l'Abes qui consiste à mettre en place des méthodes et des outils d'aide au traitement des métadonnées issues des éditeurs, et notamment de celles provenant des achats Istex. Ces procédures et ces outils sont sans cesse améliorés et développés au fur et à mesure de leur utilisation au sein même du hub, puisque ses équipes continuent effectivement de travailler sur le traitement des corpus dans un même temps. C'est d'ailleurs l'une des particularités du hub, qui se fait donc, au-delà du projet, véritable service en activité.