Le Sudoc dans Google Scholar

Raymond Bérard
berard@abes.fr

Julien Gibert
gibert@abes.fr

Agence bibliographique de l'enseignement supérieur

Conservateur général de bibliothèque, **Raymond Bérard** *dirige l'Abes. Il a auparavant dirigé la bibliothèque municipale et interuniversitaire de Clermont-Ferrand avant d'être directeur des études à l'École nationale supérieure des sciences de l'information et des bibliothèques puis directeur du Centre technique du livre de l'Enseignement supérieur. Il est également président de la CG46 de l'Afnor (Documentation). Il contribue régulièrement au* BBF.

Après l'obtention d'un Dest III Système d'information au Cnam, **Julien Gibert** *a été développeur dans une société de création de sites web puis au CRDP de Montpellier. Assistant ingénieur à l'Abes, il est développeur d'applications et de scripts de traitement des notices du Sudoc.*

Les premiers contacts de l'Agence bibliographique de l'enseignement supérieur (Abes) avec Google pour l'indexation du Système universitaire de documentation (Sudoc) par Google Scholar remontent au début de l'année 2006. Ils sont nés du constat que font toutes les bibliothèques : rares sont les étudiants et chercheurs commençant leurs recherches par le catalogue de leur bibliothèque. D'après une étude d'OCLC*, ils sont 89 % à lui préférer un moteur de recherche commercial, au premier rang desquels Google. À l'ère d'« Amazoogle », selon l'expression imagée de Lorcan Dempsey, directeur de la recherche à OCLC, il faut aller chercher le public là où il est, c'est-à-dire sur les moteurs de recherche et ne pas se contenter de l'attendre sur nos sites, même si la consultation de la version web du Sudoc continue de progresser régulièrement (24 millions de recherches annuelles).

> « À l'ère d'‘Amazoogle', il faut aller chercher le public là où il est »

Comme Google avait un projet déjà bien avancé d'indexation des catalogues collectifs nationaux (Library Link), l'Abes s'est très rapidement penchée sur la possibilité de s'associer à cette opération. Elle n'était pas pionnière en la matière puisque, outre OCLC (Open WorldCat), les catalogues collectifs de douze pays avaient alors déjà rejoint le programme « Library Link » : Suède, Suisse (Réro), Hongrie, Israël, Islande, Portugal, Australie, Chine, République tchèque, Danemark, Taïwan et Slovénie.

Avant de s'engager avec Google, l'Abes s'est informée auprès de ses partenaires suisses du réseau Réro (Réseau des bibliothèques de Suisse occidentale). Qu'il se soit agi de faisabilité technique, de protection des données, de nature des relations avec Google, d'incidence de la charge supplémentaire sur les serveurs, nos voisins nous ont persuadés de la pertinence du projet.

Une licence préservant les intérêts de l'Abes et de ses partenaires

Soucieuse de préserver les intérêts de l'Abes, des bibliothèques du réseau Sudoc et de ses fournisseurs de données, l'Agence a soumis l'accord de licence proposé par Google à l'expertise juridique d'un cabinet spécialisé : celui-ci a confirmé que le contrat d'accès de Google aux données bibliographiques provenant du Sudoc n'était cessible à aucun tiers et ne transférait pas un droit de reproduction sauf à des fins internes et de sauvegarde. Seules seraient transférées les données bibliographiques du catalogue public (les fichiers d'autorité ne sont pas concernés). La conclusion était claire : « *Ce contrat* [a] *les caractéristiques que*

* *College Students' Perceptions of Libraries and Information Resources*, OCLC, 2006.

l'on rencontre habituellement dans les contrats de licence de données et ne comporte pas de stipulation anormale. »

Les autres craintes émises ont été rapidement dissipées :

- la licence proposée n'était pas exclusive. Le Sudoc pourrait ainsi être indexé par d'autres moteurs de recherche proposant un service similaire. Des contacts ont depuis été pris, notamment avec MSN, sans résultat jusqu'à présent ;
- aucuns frais ne seraient facturés par Google. Les développements nécessaires chez chacun des partenaires resteraient à leur charge ;
- l'Abes pouvait facilement se désengager en cas de difficultés, le contrat pouvant être dénoncé à tout moment avec un préavis de 60 jours.

Soumis au conseil d'administration de l'Abes, le projet d'accord avec Google a donné lieu à un débat nourri. La polémique autour des projets de numérisation massive de fonds de bibliothèques par Google était alors à son comble et a pu être source d'amalgame entre des projets de nature radicalement différente. Après avoir recueilli l'accord formel de ses fournisseurs de données, au premier rang desquels la Bibliothèque nationale de France, l'Abes a signé l'accord de licence le 31 octobre 2006.

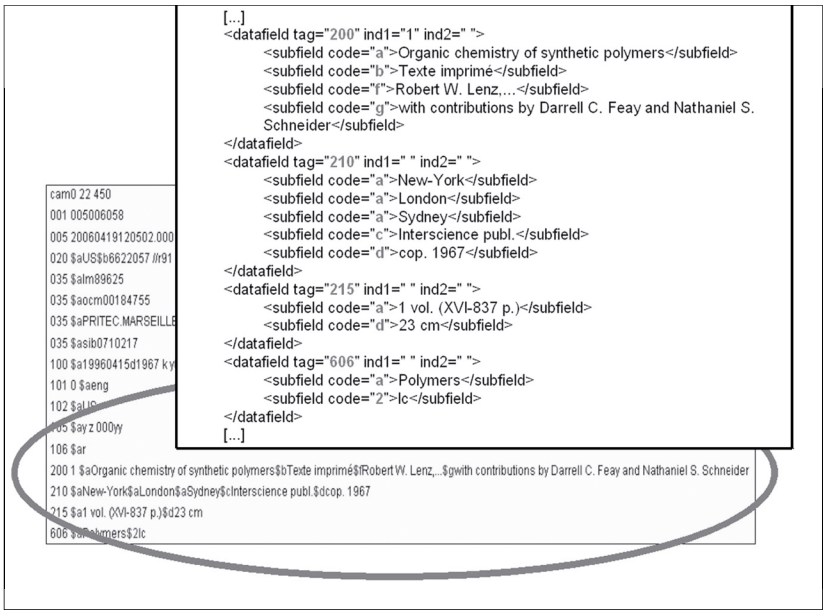

Figure 1
La même notice en Unimarc et en Marc XML

Le mode de consultation est simple : à partir de la saisie d'un terme de recherche, Google Scholar affiche dans les résultats de la recherche les liens vers les notices du Sudoc et les localisations dans les bibliothèques. La visibilité des ressources documentaires des bibliothèques du réseau Sudoc en est grandement améliorée. Un bémol toutefois : si le Sudoc est systématiquement affiché par défaut pour les interrogations effectuées sur le territoire français, ce n'est pas le cas pour les utilisateurs interrogeant Google Scholar à l'étranger qui doivent cocher le Sudoc dans la liste des catalogues proposés.

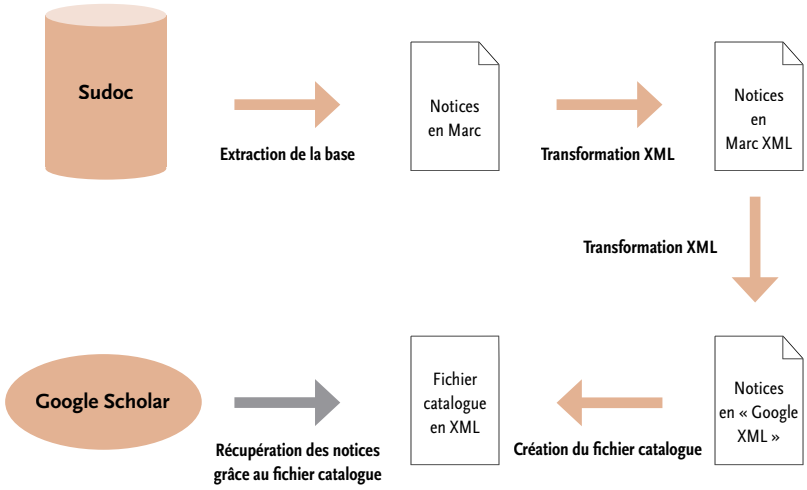

Figure 2
Chaîne de traitement des notices

Mise en œuvre technique

Les équipes de l'Abes se sont mises au travail avec leurs homologues de Mountain View immédiatement après la signature de la licence et ont abouti à une mise en production publique en avril 2007. Le projet a mobilisé une grande variété de compétences au sein de l'Abes, depuis l'étude technique jusqu'aux développements, puis aux tests d'exportations de données vers Google Scholar.

L'échange de données avec Google s'effectue au format XML qui, en permettant de structurer l'information, favorise l'échange de données via internet. Les documents peuvent être associés à un vocabulaire spécifique, défini dans une DTD ou dans un schéma : on dit que le document XML est valide s'il respecte ce voca-

bulaire. L'utilisation d'un vocabulaire commun facilite donc l'échange de documents.

Le format Marc, qui est le format bibliographique utilisé dans le Sudoc, peut être converti en XML : le format est donc XML, mais le vocabulaire Marc est conservé. On retrouve les mêmes informations décrivant les notices, mais présentées différemment (figure 1).

L'accès aux notices du Sudoc par le moteur de recherche de Google Scholar requiert une chaîne de traitement des notices à indexer.

Dans un premier temps, un script extrait les notices en format Marc de la base Sudoc. Une extraction complète de la base a lieu une fois par semestre (figure 2). Cette tâche, d'une durée d'une vingtaine d'heures, est planifiée la nuit pour ne pas surcharger les serveurs. Parallèlement, tous les mois, a lieu une extraction des nouvelles notices uniquement.

Tableau 1
Consultations avril 2007 à janvier 2008 (dix mois)

Avril 2007	2 250
Mai 2007	6 609
Juin 2007	6 852
Juillet 2007	4 702
Août 2007	6 416
Septembre 2007	11 397
Octobre 2007	14 394
Novembre 2007	15 745
Décembre 2007	23 866
Janvier 2008	32 465
Total	124 696

Tableau 2
Les 15 domaines ayant effectué le plus de requêtes (janvier 2008)

wanadoo.fr	35,01 %
proxad.net	29,37 %
inconnu	10,46 %
gaoland.net	10,22 %
club-internet.fr	2,42 %
belgacom.be	2,33 %
tiscali.fr	2,02 %
bluewin.ch	1,64 %
tele2.fr	1,41 %
noos.fr	1,04 %
jussieu.fr	0,93 %
numericable.fr	0,81 %
fti.net	0,79 %
completel.net	0,79 %
telenet.be	0,76 %

> "L'indexation du Sudoc par Google est un élément de la stratégie globale de l'Abes"

Les fichiers Marc résultants sont transformés en fichiers Marc au format XML puis, à la volée, en un format XML spécifique à Google Scholar. Ces transformations sont réalisées grâce à la technologie XSL qui permet de manipuler des arborescences XML.

Un script se déclenche alors pour parcourir les fichiers XML et fabriquer un fichier catalogue utilisé par Google pour repérer les notices.

Enfin, une date est convenue à laquelle les équipes techniques de Google Scholar utilisent le fichier catalogue pour récupérer les fichiers XML sur leurs serveurs et mettre ainsi les notices à disposition de leur moteur de recherche.

Premier bilan après le démarrage

Ainsi qu'en témoigne le tableau 1, le démarrage a été très progressif, le véritable décollage intervenant à fin du dernier trimestre 2007. Comparés aux 24 millions de consultations du Sudoc Web, les chiffres de consultation restent pour le moment marginaux mais, si la progression se confirme, ils pourraient représenter à terme jusqu'à 1 ou 2 points de la consultation totale du Sudoc. Des perspectives honorables, mais qui démentent le mirage entretenu parfois autour de Google d'une explosion des consultations.

La prédominance des recherches en provenance du territoire national est imputable au mode d'affichage du Sudoc dans la page de préférences de Google Scholar (par défaut en France, volontaire à l'étranger).

La principale réserve porte sur l'indexation non exhaustive du Sudoc par Google Scholar. Le mode d'indexation de Google, sur lequel l'Abes n'a pas d'informations, exclut en effet une proportion non négligeable de documents. Des discussions sont en cours avec Google pour améliorer cette indexation.

L'indexation du Sudoc par Google est un élément de la stratégie globale de l'Abes pour assurer une plus grande visibilité aux ressources des bibliothèques universitaires françaises sur le web. Cette stratégie porte à la fois sur l'amélioration du Sudoc Web (sa customisation par les bibliothèques est prévue en 2008) et sur l'ouverture des données du Sudoc sur le web. L'accord avec Google constitue un élément de cette stratégie, sans doute emblématique en raison de la position dominante du moteur de recherche sur le marché et de sa surexposition médiatique, mais l'objectif de l'Agence est d'être encore plus présente dans une démarche proactive avec un large éventail de partenaires. ●

Janvier 2008