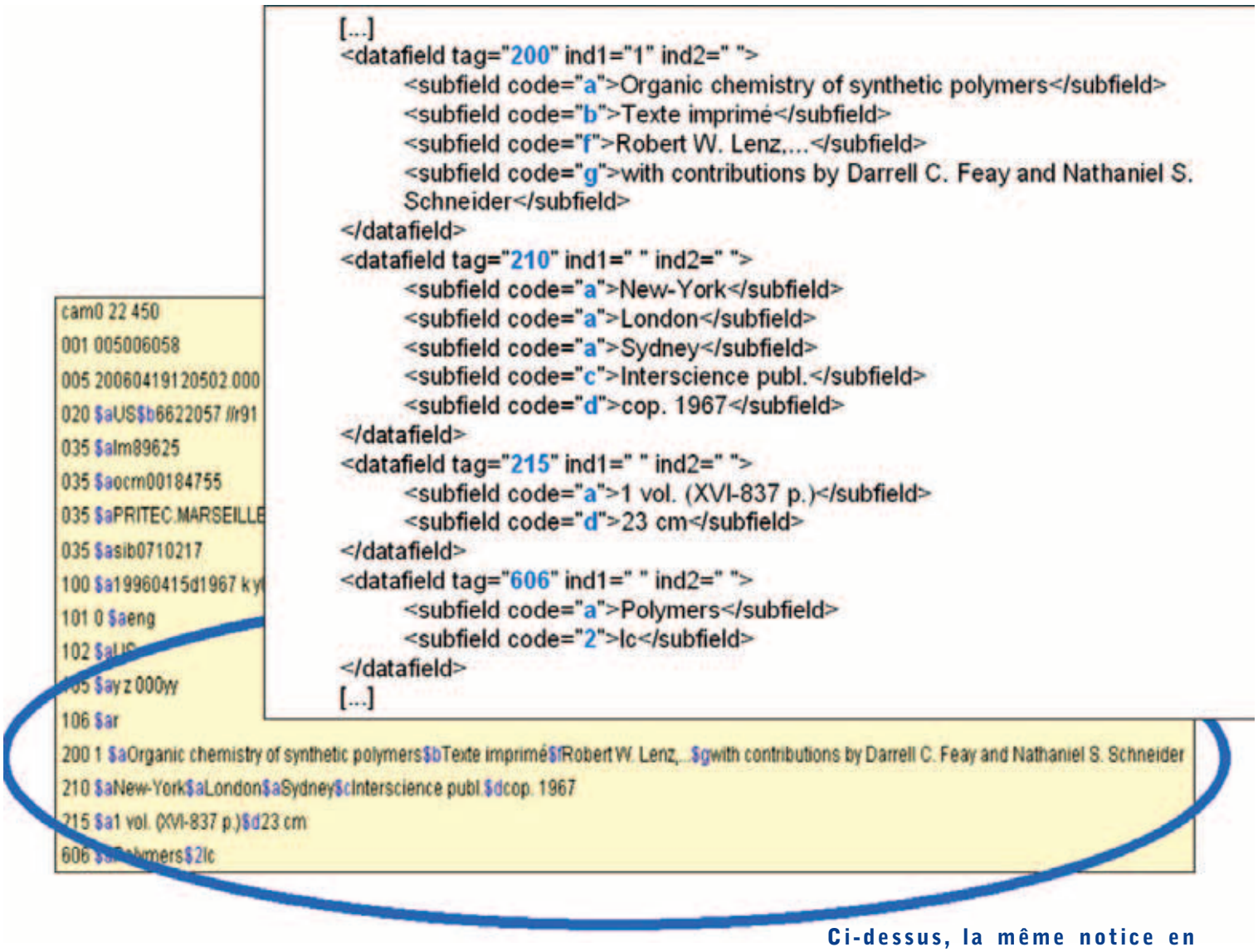
Google Scholar et le Sudoc
Transformations en douceur

Le moteur Google Scholar indexe plusieurs bases bibliographiques du domaine scientifique et permet de réaliser des recherches sur leurs notices.

L'échange de données avec Google s'effectue au format XML

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 devient le XML mais le vocabulaire MARC est conservé.



Ci-dessus, la même notice en UNIMARC ... en MARC XML

Le partenariat de l'ABES avec cet outil permet d'améliorer la visibilité du catalogue Sudoc en proposant un nouveau point d'accès à celui-ci.
Ce projet a débuté en novembre 2006 pour aboutir à une mise en production officielle en mars 2007.
Il a mobilisé une grande variété des compétences de l'ABES depuis l'étude technique jusqu'aux développements, puis aux tests d'exportations de données vers Google Scholar.

Le format XML (langage de balisage extensible) est un format 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 vocabulaire.

On retrouve les mêmes informations décrivant les notices, mais représentées différemment.

Le traitement des notices

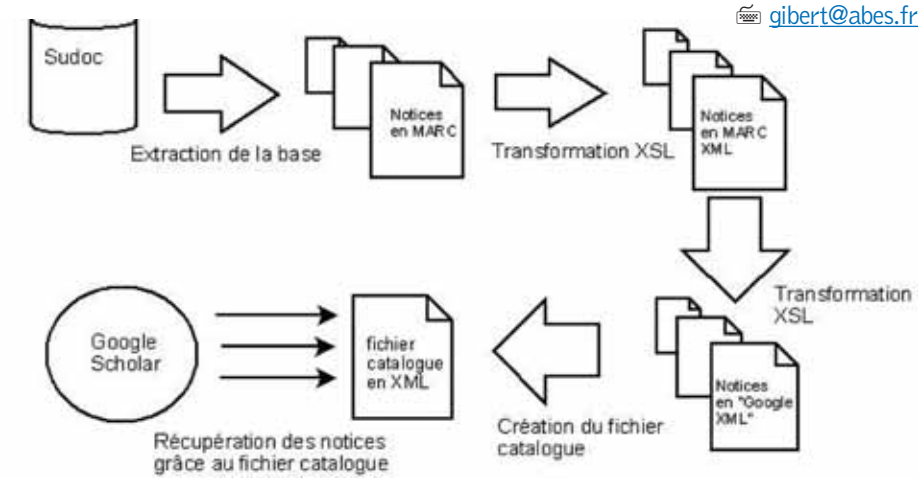
L'accès aux notices du Sudoc par le moteur de recherche de Google Scholar requiert une chaîne de traitement des notices à indexer.

.../...

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. 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.
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.

Julien Gibert
gibert@abes.fr



Chaîne de traitement des notices

Il était une fois un prototype...

Sudoc 1.99... version pré-alpha

En février 2006, à titre d'expérimentation, nous avons tenté d'imaginer une autre interface pour le Sudoc, en repartant des données brutes (de l'UNIMARC en XML) et en construisant dynamiquement autour de la notice un affichage enrichi. L'objectif n'était pas d'aboutir à une proposition clés en main, mais d'illustrer les possibilités offertes par les nouvelles technologies (données en XML, interfaces AJAX, XSLT) et par la présence sur le web de compléments d'information pertinents pour un catalogue bibliographique.
En voici une liste non exhaustive :

- couverture des livres ;
- suggestion d'autres éditions de la même œuvre (via le web service xISBN d'OCLC qui, depuis, est devenu un service commercial) ;
- affichage des éditions numérisées en ligne, incontournable avec la massification des programmes de numérisation ;
- rebond vers des recensions ou des critiques (sur « Google review », Persée, sur des blogs...) ;
- rebond vers des travaux qui citent le document affiché (via Google Scholar ou Google Books par exemple, mais demain via un portail des thèses en ligne ?) ;
- pour les périodiques, affichage du sommaire de la dernière livraison, via le fil RSS de la revue (ce qui, à partir de 2009, sera facilité par un service comme ticTOCs (http://www.tictocs.ac.uk/) ;
- présentation des localisations sous forme de carte.

Le prototype n'avait rien d'original ni d'abouti, mais il montrait bien l'étendue des possibilités qui s'offrent à nous. On voit que les principaux obstacles à la réinvention de nos catalogues ne sont plus techniques.

Christophe Bonnefond et Yann Nicolas
bonnefond@abes.fr
nicolas@abes.fr