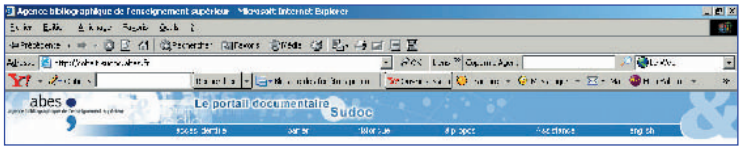
Bienvenue au club !

www.abes.fr



Si le portail du Sudoc est capable de donner accès à une grande diversité de ressources et de données, les universités et les autres partenaires potentiels se posent des questions sur les formats ou les procédures à respecter pour fédérer leurs ressources à travers cet outil. Voici quelques « bonnes pratiques ».

Le défi d'un portail comme celui du Sudoc est d'offrir sous une forme homogène toute une variété de données hétérogènes. Outre la présentation uniforme qu'elle propose, cette forme homogène permet de fusionner les résultats des requêtes et pas seulement de les juxtaposer. Cette fusion autorise des opérations souvent réservées aux recherches monobase telles que le tri ou le dédoublonnage. Mais pour arriver à cela, le portail doit gérer en coulisses l'hétérogénéité des données qu'il traite. Ces données sont hétérogènes par leur format (fichiers texte, tableaux, MARC, XML sous toutes les formes, formats propriétaires…), leur origine (catalogues, bibliographies, archives ouvertes, bases d'éditeurs ou de diffuseurs), leur statut (accès libre ou payant), leurs conditions techniques d'accès (récupération ou accès distant, par différents protocoles) et leur nature (métadonnées ou texte intégral). Techniquement, l'équipe est capable de gérer cette diversité, mais avec plus ou moins de facilité et de rapidité selon les cas. Sans surprise, plus les données d'origine seront transmises au portail sous une forme et selon des procédures standard, meilleur sera le service rendu, en termes de qualité des données, de fonctionnalités de recherche et de rapidité de mise en service.

La première question est celle du protocole : comment rendre les données accessibles ? Soit elles sont envoyées au portail puis chargées et indexées par ses bases internes, soit elles sont interrogées à distance par le moteur de recherche. Dans le premier cas, le portail peut régulièrement récupérer vos données sur un serveur FTP ou un serveur conforme au protocole OAI-PMH. Ce dernier est préférable. OAI-PMH s'impose comme la méthode la plus courante et la plus simple pour des échanges réguliers de métadonnées XML. Il implique de structurer vos métadonnées en XML, exprimées *a minima* en Dublin Core, mais il autorise aussi n'importe quel format XML pourvu qu'il soit documenté. Dans le cas d'une interrogation à distance, le portail se connectera bien sûr à votre serveur Z39.50, mais aussi via le protocole SRW/SRU, qui succède à Z39.50. SRW/SRU est plus simple (surtout sous sa forme SRU), mais en outre, il permet d'exploiter les données XML quelles qu'elles soient, et plus seulement des données MARC.

La seconde question est celle du format. On vient de le voir, les protocoles récents exigent des données en XML, ce qui n'est pas une restriction mais une ouverture, puisque le X de XML signifie « extensible » : vous pouvez définir votre propre format, si vos besoins sont particuliers et votre vocabulaire suffisamment documenté. Notez cependant que dans le cadre d'un portail multibase, les recherches ne peuvent se faire que sur les index communs aux différentes bases. Dès lors, les champs originaux de votre vocabulaire seront bien affichés dans les résultats, mais ne seront pas interrogeables en tant que tels. Pour les besoins ordinaires, autant suivre les standards (Dublin Core, MODS, MARCXML…). Dernier conseil sur le format : pour rendre efficaces les opérations de tris ou de dédoublonnage, il est conseillé de bien soigner et standardiser des zones comme la date ou la langue. Les conseils qui précèdent concernent avant tout les métadonnées, mais **la vocation du portail est aussi de permettre des recherches sur le texte intégral** (mais non de diffuser les documents primaires). Dans les cas encore rares où le texte intégral est structuré en XML, rien n'interdit d'utiliser les protocoles XML décrits plus haut pour transporter non seulement les métadonnées mais aussi le texte intégral. Le texte intégral devient interrogeable non comme un bloc, mais de manière structurée. Sinon, en l'absence de pratiques établies pour soumettre ensemble un document et ses métadonnées, on peut imaginer différentes solutions. Dans le cas le plus simple, les métadonnées sont intégrées au document, sous différentes formes : le portail ne reçoit que le document, mais en extrait les métadonnées. Autre configuration : on peut imaginer que le texte soit inséré dans les métadonnées, comme une longue note de texte libre qui servirait à l'indexation mais pas à l'affichage. Enfin, autre solution peu contraignante : il suffit que l'URL du document à indexer soit présente dans les métadonnées transférées (via OAI-PMH, par exemple). Les outils du portail exploitent alors cette URL pour aller indexer le document distant, sans le stocker. Cette solution implique que cette URL pointe bien vers le document à indexer, et non vers une page intermédiaire contenant le résumé ou d'autres informations. Dans tous les cas, l'idée est de ne gérer qu'un seul fichier, quitte à en faire deux usages.

Pour résumer, quand les standards – voire les normes – existent, le bon sens invite à les suivre de la manière la plus scrupuleuse. C'est le meilleur moyen de rejoindre le portail du Sudoc, mais aussi d'accéder à d'autres canaux de diffusion ou de profiter des outils qui gravitent autour des standards.

Bref, tenue correcte exigée, mais, en l'espèce, ce standing n'est qu'une question de standard.

Yann Nicolas
nicolas@abes.fr

Point technique

Le portail Sudoc s'appuie sur trois logiciels (Bookline, Masc et Sim) fournis par la société Archimed.
Pour le moment, ces logiciels fonctionnent sur un environnement Windows mais sont toutefois en passe de migrer vers l'architecture Dotnet qui autorisera l'utilisation de serveurs Windows et Unix.
Les composants de base utilisés dans la version actuelle sont Windows 2000, Sql Server 2000, DotNet et Index Server.
L'architecture utilisée par ces trois logiciels est constituée :
- de serveurs Web (Windows 2000) en load balancing pour la partie applicative (serveur Web et composants Métiers) ;
- de grappes de serveurs Sql : serveurs en réplication et répartition de charge (pour le stockage des métadonnées des notices et des données administratives) ;
- de grappes de serveurs Index : serveurs en répartition de charge pour la gestion des index (full text et autres). *C. B. bonnefond@abes.fr*