Les moteurs de recherche

Petit précis de mécanique à l'usage des bibliothèques numériques

L'irruption des grands acteurs de la Toile, et notamment du moteur de recherche Google, dans le champ de la numérisation de masse, est aujourd'hui en train de bouleverser à la fois la perception que l'on a des bibliothèques numériques, et les outils et services sur lesquels elles s'appuient. Le moteur de recherche étant devenu le principal point d'accès à la collection en ligne, il tend souvent à se confondre avec elle, il en devient l'unique interface, le seul artefact visible, un peu comme si une bibliothèque se limitait soudain à sa salle de lecture.

Emmanuelle Bermès

Bibliothèque nationale de France
emmanuelle.bermes@bnf.fr

Pourtant, le moteur de recherche n'est que l'un des outils de la bibliothèque numérique ; s'il a un rôle particulier c'est qu'il est l'instrument du service rendu à l'utilisateur final, et à ce titre, il détermine ou est déterminé par presque toutes les autres activités qui constituent les différentes facettes de la bibliothèque numérique.

À la Bibliothèque nationale de France, le projet Europeana[1] a été l'occasion d'explorer plus en profondeur les tenants et les aboutissants de la mise en place d'un moteur de recherche comme outil d'accès à une collection de documents patrimoniaux numérisés. De cette expérience, la principale leçon à retenir est que, si l'idéal d'une bibliothèque numérique sur la Toile se cristallise dans la simplicité apparente de son moteur de recherche, ce dernier est un outil complexe, dont il faut commencer par comprendre et maîtriser le fonctionnement. Un défi qui n'est pas seulement technique, car un certain nombre de questions se posent au bibliothécaire, mettant en jeu ses compétences spécifiques de gestionnaire de la collection numérique : quelles sont les caractéristiques et les qualités des données à indexer ? Quels sont les outils à notre disposition pour procurer la meilleure pertinence ? Enfin, comment rendre le service simple et intelligible à un public distant souvent mal connu ?

L'essence des moteurs : les données

Par nature, les moteurs de recherche sont des outils adaptés au traitement de masses importantes de données textuelles, caractérisées par leur diversité et leur hétérogénéité. Entre les notices de bibliothèque et les milliards d'octets de texte qui naissent chaque jour sur la Toile dans le plus complet désordre, il semble y avoir un monde ; pourtant, l'enjeu du moteur de recherche est bien d'optimiser la manipulation simultanée de ces

1. Les informations sur le projet sont disponibles en ligne : www.bnf.fr/pages/europeana/europeana.htm
Le prototype Europeana a été rendu accessible au public à l'occasion du Salon du livre 2007 : www.europeana.eu

Archiviste-paléographe de formation, titulaire du DCB, ***Emmanuelle Bermès*** *est responsable fonctionnel de la bibliothèque numérique de la BnF. Elle a écrit « Un protocole pour l'échange de métadonnées : l'OAI » pour la Journée d'information Afnor/CG46 « Des métadonnées pour bien utiliser les ressources électroniques » et, dans le nº 40, déc. 2006 d'*International Preservation News, *« Des identifiants pérennes pour les ressources numériques : l'expérience de la BnF » et est l'auteur de Figoblog.*

objets de nature si variée. Confrontés à la croissance exponentielle de la Toile aussi bien en quantité qu'en diversité, les moteurs de recherche ont dû constamment adapter leurs mécanismes techniques, leurs algorithmes, à l'évolution du contexte environnant. Dans le domaine des bibliothèques numériques, la situation est plus stable : le matériau brut, les données que l'on fournit au moteur de recherche, est bien identifié et globalement beaucoup plus ordonné que ce que l'on trouve sur la Toile.

Des gisements de données structurées

Si on pose pour définition de la bibliothèque numérique le fait d'être organisée en suivant une logique raisonnée (la politique documentaire), et de se composer d'objets décrits de façon structurée (par des notices, également appelées métadonnées) même s'ils sont constitués eux-mêmes de plein texte, on peut permettre au moteur de recherche de se reposer sur au moins deux niveaux fiables de structuration de l'information.

D'une part, le fait de disposer de collections organisées est un atout essentiel car, en organisant la collection, on garantit une forme d'homogénéité des données, même si on reste à un niveau très global. Ainsi, une collection de livres numérisés en mathématiques, une sélection de revues, un choix de livres qui sont tombés dans le domaine public, une collection de photographies anciennes, sont autant d'ensembles qui sont suffisamment circonscrits pour garantir des modèles communs : la langue, le vocabulaire, la structure, la quantité et la qualité des informations disponibles seront plus ou moins constants.

D'autre part, les notices bibliographiques sont caractérisées par la richesse de leur structure : fruits d'une analyse fine réalisée par les bibliothécaires à travers le catalogage, elles représentent l'or noir de la recherche d'information par leur fiabilité et le niveau de détail adopté dans la qualification des différents éléments.

> Ce sont justement les recherches les plus communes, les plus évidentes, qui tendront à poser problème, d'où l'importance de travailler sur le classement de pertinence des listes de résultats

Pour prendre un exemple, si « Victor Hugo » dans le plein texte n'est qu'une chaîne de caractères comme les autres, dans une notice bibliographique, on dispose de son nom tel qu'il est inscrit sur le document, de son nom normalisé par des conventions (les notices d'autorités), d'autres informations comme le rôle qu'il a joué dans l'élaboration de l'œuvre (était-il ou non auteur principal ?), de ses dates… Toutes ces informations étant correctement qualifiées, prêtes à être utilisées par le moteur de recherche. L'utilisation de ces données structurées sera, nous le verrons, déterminante pour garantir un service de qualité.

Enfin, certains formats d'encodage des textes, comme la TEI (Text Encoding Initiative), permettent de structurer de façon logique les contenus mêmes des ouvrages et de fournir au moteur des informations précises telles que le repérage des références bibliographiques ou des notes de bas de page, des citations, mais aussi des noms de personnes, de lieux, etc.[2] Cependant, dans le cadre de la numérisation de masse, les coûts ne permettent généralement pas d'atteindre ce niveau de structuration, et le « plein texte » obtenu par saisie ou par reconnaissance optique de caractères (OCR) n'est que peu ou pas du tout structuré. Ainsi, le format Alto (Analyzed Layout and Text Object)[3], choisi par la Bibliothèque nationale de France pour l'OCR de ses ouvrages numérisés, sert principalement à encoder des informations de structure physique. Pour chaque mot, il enregistre sa position dans la page numérisée en mode image, la taille des caractères et la police. Ces informations ne sont que de peu d'intérêt pour le moteur de recherche.

Le plein texte ou la pollution par le bruit

Dans une bibliothèque numérique contenant des livres numérisés en mode texte, on est donc confronté à deux types de matériaux : les données structurées (tirées des notices bibliographiques) et le plein texte (peu structuré ou uniquement de manière physique), qui vont tous deux être fournis au moteur de recherche en vue de l'indexation et de l'interrogation par les utilisateurs. Dès lors, se pose la question de la gestion de l'effet d'échelle : si on prend en

2. On peut voir l'application de ces fonctions sur ce type de documents structurés dans Telma, Centre de ressources numériques pour le traitement électronique des manuscrits et des archives, rubrique « corpus » : www.cn-telma.fr
3. Pour plus de détails sur l'utilisation de ce format à la BnF, on peut consulter le document *Conversion en mode texte des images de la presse périodique de la BnF* (mai 2006), en ligne : http://bibnum/numerisation/index.html

compte le plein texte, compte tenu de la quantité de mots à disposition, le bruit généré par n'importe quelle requête, mais surtout par celles qui portent sur les mots les plus courants, est forcément important. Ce sont justement les recherches les plus communes, les plus évidentes, qui tendront à poser problème, d'où l'importance de travailler sur le classement de pertinence des listes de résultats.

Souvent, la pertinence est un subtil équilibre qui dépend aussi très largement de ce que l'utilisateur recherche : s'il saisit « Victor Hugo », souhaite-t-il en priorité obtenir des ouvrages du romancier, ou des biographies le concernant ? Comment être capable de répondre aussi bien au lecteur qui cherche le roman *Notre-Dame de Paris*, qu'à celui qui cherche des textes ou des photographies évoquant la cathédrale parisienne ? L'une des clés de la pertinence réside dans la granularité des résultats de recherche. Souvent, les usagers ne sont pas en quête d'une occurrence dans le plein texte, mais ils essayent plutôt de constituer ou d'accéder directement à un corpus cohérent autour d'un auteur, d'une notion, d'un thème. Le plein texte apparaît dans ces conditions comme une solution de repli pour décrire le contenu de documents qui ne pourraient pas bénéficier de métadonnées suffisamment structurées.

À cet égard, l'exemple de la presse ancienne numérisée est significatif : l'utilisateur peut s'intéresser aussi bien au titre d'un journal, et donc aux différents numéros disponibles, qu'à un seul numéro pour une date précise ou parce qu'il contient une expression ou un nom. Or la recherche plein texte a pour conséquence d'écraser cette granularité du document. Si on ne propose que le niveau le plus élevé (le titre), on ne représente que par un nombre extrêmement limité d'informations structurées une réalité plus massive (des milliers de fascicules pour un quotidien). Au contraire, si on se focalise sur le plein texte, comme le fait par exemple le projet Chronicling America[4], on risque de survaloriser les fascicules de presse par rapport à d'autres ressources et de déconstruire le document et la collection, au détriment des recherches les plus simples.

Dans ce contexte, il est important de bien maîtriser à la fois la connaissance de la structure des collections, et celle des besoins utilisateurs, pour être en mesure de se doter d'outils qui sauront manipuler les données et les enrichir, en vue d'optimiser la pertinence des résultats de recherche.

Mécanique des textes : peser et classer les données

Si on projette dans le domaine des moteurs de recherche l'expertise classique de sélection qui est celle des bibliothécaires, il apparaît clairement que ces derniers ont un rôle à jouer pour déterminer comment le moteur peut calculer la pertinence des documents en fonction d'une requête. Pour cela, il leur faudra développer une expertise dans le domaine du *text-mining*, qui regroupe l'ensemble des technologies de traitement et d'analyse mises en œuvre par les moteurs de recherche pour tirer du sens à partir des masses de données textuelles qui leur sont fournies[5].

Ces traitements mettent en jeu un type d'outil particulier : le moteur d'indexation ou indexeur, qui peut être un module du moteur de recherche global, ou un outil spécifique. Pendant la phase d'indexation, celui-ci extrait des données les chaînes de caractères qui seront utilisées pour la recherche et constitue un index inversé qui permettra de répondre plus rapidement aux requêtes. Il procède ensuite à des analyses qui vont préparer le calcul de la pertinence en associant des valeurs aux mots indexés.

Des rouages bien huilés : la pondération et l'analyse statistique

Dans le cas d'une bibliothèque numérique, on peut utiliser les données structurées des notices bibliographiques pour valoriser certains champs comme l'auteur ou le titre. Des expériences menées par la National Library of Australia pour le projet Libraries Australia[6], ou par le Melvyl Recommender Project à la California Digital Library[7], ont montré que les notices de catalogue se prêtaient particulièrement bien à cette exploitation des données structurées, que ce soit dans le contexte d'un catalogue traditionnel, ou dans celui d'une bibliothèque numérique. Lors de l'indexation, le moteur de recherche pondère les différents champs : il attribue une valeur plus ou moins élevée à un mot en fonction du champ dans lequel il le trouve. C'est ce qui va permettre à la qualité de ne pas être noyée par la quantité : ainsi, sur la requête « Victor Hugo », la pondération forte du champ auteur permet de faire remonter en haut de liste des ouvrages de Victor Hugo, sans qu'ils soient éclipsés par les récits biographiques dans lesquels cette chaîne de caractères apparaît à de bien plus nombreuses reprises.

Toutefois, on utilise également les informations quantitatives pour pondérer la « valeur » d'un mot. Cette

4. www.loc.gov/chroniclingamerica
Ce projet piloté par la Library of Congress vise à mettre en ligne les titres de presse américaine numérisés en partenariat dans le cadre du NDNP (National Digital Newspaper Program). Les documents sont numérisés en mode image et OCR. Le moteur de recherche permet uniquement de « chercher des pages » contenant certains mots.

5. Voir l'introduction proposée par Christian Fauré sur son blog Hypomnemata : supports de mémoire : www.christian-faure.net/2007/05/30/introduction-au-text-mining (consulté le 20 août 2007).

6. « Relevance ranking of results from MARC-based catalogues : from guidelines to implementation exploiting structured metadata » par Alison Dellit et Tony Boston, Bibliothèque nationale d'Australie, février 2007 : www.nla.gov.au/nla/staffpaper/2006/documents/Boston_Dellit-relevance-ranking.pdf

7. Consulter le site du projet : www.cdlib.org/inside/projects/melvyl_recommender

analyse, appelée analyse statistique, est commune à tous les moteurs de recherche aujourd'hui et, comme la pondération, repose sur des principes bien connus que le bibliothécaire doit maîtriser pour être en mesure d'ajuster les paramètres du moteur. De façon basique, lorsqu'un mot apparaît de nombreuses fois dans un document, c'est un indice de pertinence élevé pour ce document. Mais ce rôle de la masse est pondéré par des analyses statistiques plus fines. Ainsi, un mot rare dans un texte aura plus de valeur qu'un mot qui revient très souvent, un mot trouvé dans un champ très court aura plus de valeur que le même mot dans un champ très long.

D'autres types de données que celles qui décrivent au sens strict les documents pourraient être prises en compte dans l'analyse statistique, par exemple des données de popularité, à l'image du fameux « *pagerank* » qui évalue la notoriété d'une page web en fonction du nombre de liens qui la citent. On pourrait ainsi utiliser des données d'usage, comme les statistiques de consultation, pour pondérer les résultats du moteur : cela équivaudrait à une forme de recommandation. Toutefois, ce type de comportement naturellement induit par la vocation commerciale des moteurs de recherche de la Toile, dans une logique qui n'est pas celle des bibliothèques en principe, nous ramène au traditionnel débat entre l'offre et la demande, bien connu dans le domaine de la politique documentaire. La recommandation ne doit pas conduire à fragiliser l'équilibre qui existe entre les besoins conscients d'un utilisateur, et la possibilité pour lui de découvrir des ressources qu'il n'aurait pas pensé à chercher.

L'utilisation des folksonomies, ou tags, est également aujourd'hui considérée comme une piste possible : si les utilisateurs peuvent librement attribuer des mots-clés aux documents, l'existence de ces mots-clés est un indice de popularité en même temps que cela fournit un vocabulaire d'indexation particulier, plus proche en théorie des usages courants que les vocabulaires bibliothéconomiques, et donc susceptible de fournir un service complémentaire. Toutefois, cette théorie demande encore à être expérimentée en bibliothèque[8].

Plus de puissance avec les analyses linguistiques et sémantiques

Les opérations d'analyse statistique et de pondération sont communes à tous les outils d'indexation, ce qui n'est pas le cas des analyses linguistiques et sémantiques. Celles-ci relèvent d'outils spécifiques utilisés pour ajouter de la structuration aux textes peu ou pas structurés.

L'analyse linguistique relève des outils de traitement automatique des langues (TAL). Dans cette phase, l'outil repère les mots, leur emplacement dans les phrases, et les structures grammaticales qui lui permettent d'identifier les mots qui ont le plus de valeur et que l'on appelle généralement les « entités nommées » (noms de personnes, noms de lieux, concepts). Il peut également procéder à une lemmatisation, c'est-à-dire qu'il recherche les différentes formes dérivées d'une même racine (par exemple, les conjugaisons d'un verbe).

L'analyse sémantique consiste à confronter le résultat de l'analyse linguistique avec des significations connues par des référentiels, parfois englobés sous le nom de bases de connaissances. Il s'agit de modélisations de concepts sous une forme proche de nos vocabulaires contrôlés et thésaurus ; on emploie également le terme d'ontologie. Elles ont pour rôle de rajouter au texte initial les informations trouvées dans le référentiel : une mention de Victor Hugo dans une biographie passera alors du statut de simple chaîne de caractère à celui de personne identifiée, au même titre que l'auteur signalé dans les métadonnées structurées.

Ces solutions d'analyse linguistique et sémantique ont aujourd'hui fait leurs preuves sur des corpus homogènes du point de vue du contenu (par exemple des textes relevant du domaine médical ou juridique, des dépêches d'actualité…). En effet, elles relèvent d'une démarche qui ne peut être que partielle : les outils linguistiques sont spécifiques à une langue, les outils sémantiques à un domaine de la connaissance ou un métier. S'agissant de corpus généralistes, multilingues et très étalés dans le temps, comme le sont les collections des bibliothèques numériques, l'enjeu est de taille.

> Cette vue d'ensemble demandée par les internautes n'est pas seulement un supplément d'âme, mais un besoin cognitif lié à la nécessité d'appréhender son environnement en tant qu'espace, même lorsque celui-ci est virtuel

Pourtant, la mutualisation de l'effort dans ce sens apparaît comme un véritable espoir, avec la mise en œuvre des technologies et des standards du « web sémantique ». Les bibliothèques disposent déjà de données structurées et de référentiels qui les décrivent : imaginons la valeur ajoutée que pourrait apporter

8. Voir Jonathan Furner, « L'indexation des ressources des bibliothèques par les usagers : vers un modèle d'évaluation ». Traduit de l'anglais par Françoise Bourdon : www.ifla.org/IV/ifla73/papers/157-Furner-trans-fr.pdf

à un texte le fait de relier aux termes de Rameau les occurrences des entités nommées. Ces référentiels, une fois convertis dans les standards d'encodage des ontologies tels que RDF[9] ou SKOS[10], et exposés sur la Toile, présenteraient un double avantage : d'une part, leur usage étant largement répandu en bibliothèque, cet effort bénéficierait à l'ensemble de la communauté ; d'autre part, ces outils de référence généralistes pourraient faire autorité comme base de connaissance pour d'autres acteurs de la Toile qui attendent cette expertise de la part des bibliothèques. C'est d'ailleurs le sens du projet VIAF[11], partenariat entre OCLC, la Library of Congress et la Deutsche Nationalbibliothek, maintenant rejointes par la BnF, pour rapprocher et mettre en ligne les notices d'autorité de ces différentes bibliothèques.

Moteurs et carrosserie : les interfaces de recherche

Les techniques de *text-mining* que nous venons d'évoquer ont pour vocation d'enrichir les données en vue d'améliorer le service fourni aux lecteurs internautes de la bibliothèque numérique : le moteur de recherche proprement dit. Si les usages ont consacré sur internet le rôle de la « recherche simple » avec un seul champ de recherche, le moteur de la bibliothèque numérique, avec ses données structurées et ses contenus enrichis, dispose de suffisamment de puissance et de complexité pour proposer des interfaces de recherche avancée qui rappellent, par certains aspects, nos catalogues.

De fait, s'agissant d'une collection de documents traditionnels (livres, revues, etc.), bien que celle-ci soit numérisée, on peut constater de la part des internautes un certain nombre d'attentes assez traditionnelles qui se sont clairement exprimées dans les études d'usage réalisées autour d'Europeana[12]. De nombreux lecteurs ont exprimé le besoin de chercher spécifiquement un titre ou un auteur et, de façon corollaire, celui de lister la collection. Depuis l'origine de Gallica, les internautes n'ont cessé de réclamer la liste de tous les auteurs, de tous les titres présents dans la bibliothèque numérique, alors même qu'elle serait trop longue pour pouvoir être véritablement parcourue ; cette demande a été répétée pour Europeana.

Pourquoi un tel besoin apparaît-il dans la bibliothèque numérique, alors qu'il ne viendrait à l'idée de personne de demander à consulter la liste de tous les ouvrages conservés par la BnF ? Tout simplement à cause de l'un des effets pervers du moteur de recherche : en obligeant l'utilisateur à formuler une question a priori, il lui interdit d'avoir une visibilité globale sur la collection. Cette vue d'ensemble demandée par les internautes n'est pas seulement un supplément d'âme, mais un besoin cognitif lié à la nécessité d'appréhender son environnement en tant qu'espace, même lorsque celui-ci est virtuel. Elle est aussi la garantie de pouvoir trouver par hasard des documents que l'on ne cherchait pas forcément, un concept aujourd'hui largement valorisé sous le nom de sérendipité[13]. Or, comme nous l'avons déjà souligné, l'une des forces de la bibliothèque numérique est de disposer de la structure d'une collection, avec son organisation, et son homogénéité. Quels sont donc les moyens à notre disposition pour forcer le moteur de recherche à

Le moteur de recherche d'Europeana

En mars 2007, à l'occasion du Salon du livre, était présenté au public le site Europeana, contribution française à la bibliothèque numérique européenne. Cette première réalisation faisant suite à l'appel de Jean-Noël Jeanneney et à la lettre de mission reçue par la BnF est un prototype visant à servir de laboratoire pour l'expérimentation de nouvelles fonctionnalités de bibliothèque numérique, notamment la recherche plein texte. Ainsi, outre les métadonnées en Dublin Core de 12 000 ouvrages issus des collections de la Bibliothèque nationale de France, de la Bibliothèque Nationale Széchényi de Hongrie (Országos Széchényi Könyvtár) et de la Bibliothèque nationale du Portugal (Biblioteca Nacional de Portugal), Europeana donne accès en interrogation et consultation plein texte à quelque 7 000 documents de la BnF, sélectionnés dans Gallica pour leur intérêt documentaire et convertis en mode texte par OCR dans le cadre d'un marché spécifique.

Le moteur de recherche utilisé est un logiciel libre, Lucene ; le calcul de la pertinence, exprimée en pourcentage, repose sur l'algorithme par défaut inclus dans Lucene, avec un paramétrage de pondération spécifique portant sur les principaux champs de métadonnées.

Bien que seule la recherche simple soit proposée aux utilisateurs, la recherche par champs est également gérée par Lucene, ce qui lui permet de proposer des listes de documents par facettes et par thème, ainsi que l'affinage par provenance, date, langue et auteur. Pour gérer la granularité, le moteur d'Europeana dispose de deux index : l'un qui permet de rechercher des documents (recherche principale du site) et l'autre qui permet de rechercher des pages à l'intérieur de chaque document (module de recherche disponible sur la page de consultation du document).

Les retours utilisateurs collectés par le questionnaire en ligne sur le site ont montré que, si la pertinence était satisfaisante pour des requêtes simples, des améliorations étaient à prévoir sur la gestion des opérateurs booléens (ET plutôt que OU) et sur la lemmatisation qui donnait parfois des résultats surprenants. La recherche avancée a été réclamée avec insistance. Fortes de cette expérience, les équipes de la BnF ont continué à travailler sur Lucene et prévoient de l'intégrer, avec ces améliorations, dans la nouvelle version de Gallica qui devrait voir le jour à l'automne 2007.

E. B.

9. Voir l'interview de Yann Nicolas « Métadonnées : faut-il parier sur RDF ? » dans *Artist*, 30 mai 2007 : http://artist.inist.fr/article.php3?id_article=406

10. Simple Knowledge Organisation System est un standard du W3C qui permet d'encoder des thésaurus en utilisant RDF.

11. Virtual International Authority File : www.oclc.org/research/projects/viaf/default.htm

12. On peut consulter le rapport d'analyse de ces études en ligne : http://bibnum/usages/index.html

13. Olivier Ertzscheid et Gabriel Gallezot, « Chercher faux et trouver juste, sérendipité et recherche d'informations », Bucarest, CIFSIC, *Colloque bilatéral franco-roumain, CIFSIC Université de Bucarest, 28 juin–3 juillet 2003*, Bucarest : http://archivesic.ccsd.cnrs.fr/sic_00000689/fr

donner une vision spatiale de la collection numérique ?

La première méthode, expérimentée dans Europeana, est l'utilisation des données structurées pour construire un accès par facettes. Les facettes sont des critères de base qui permettent d'aborder la collection d'un certain point de vue : l'auteur, la période chronologique, le lieu, ou encore, comme dans la recherche thématique d'Europeana, les grands thèmes de la classification Dewey.

Cette approche peut être un point de départ, très apprécié de ceux des usagers qui trouvent dans cette classification le terme qui les intéresse, comme les généalogistes. Elle peut également être proposée dans un deuxième temps à l'utilisateur pour l'aider à affiner ou élargir sa requête. Ainsi, dans WorldCat [14], l'ensemble des champs de la notice est proposé sous forme de facettes pour feuilleter la collection, une fois qu'on a saisi une première recherche. Ce système permet de guider l'utilisateur dans la précision de sa requête, sans pour autant le forcer à passer par un formulaire de recherche avancée.

Si l'on étend ces principes aux bibliothèques numériques, il faut imaginer ce qu'apporteraient les fonctionnalités de *text-mining* à un tel mode de navigation : on pourrait créer entre les documents des passerelles hypertextuelles, proposer des rebonds sur des noms de personnages ou des références bibliographiques à l'intérieur d'un texte, et ainsi naviguer d'ouvrage en ouvrage comme sur autant d'étagères virtuelles.

Une autre piste pour l'exploration spatiale des collections serait dès lors de représenter cette navigation de façon visuelle : par des cartes ou des graphiques qui permettraient, grâce à un certain nombre de conventions simples, de donner une idée beaucoup plus concrète de la qualité et de la quantité de la collection numérique. Ce type de fonctionnalités est encore peu exploré à ce jour dans les bibliothèques ; on peut citer l'exemple d'AquaBrowser, un outil de la société Medialab qui permet de naviguer visuellement dans un Opac de bibliothèque.

Bilan et perspectives : la route à parcourir

L'expérience d'Europeana nous a montré qu'il restait bien des questions ouvertes avant que les moteurs de recherche puissent être considérés comme maîtrisés pour une bibliothèque numérique. La question du multilinguisme, dans le cas d'une bibliothèque européenne, reste entière : la plupart des fonctions décrites ci-dessus s'appliquent bien à un corpus homogène du point de vue de la langue, mais la multiplication des langues et en particulier la forte présence des langues anciennes, non prises en compte par les outils du marché, reste un défi. En ce qui concerne les technologies prometteuses du traitement linguistique et sémantique, des expérimentations seront menées au niveau européen, dans le cadre de projets pilotés en partenariat avec The European Library [15].

La multiplication des langues et en particulier la forte présence des langues anciennes, non prises en compte par les outils du marché, reste un défi

D'une façon plus générale, si certains moteurs de recherche procèdent de façon enchaînée à un traitement des données par étapes, de l'indexation à la recherche en passant par le traitement et l'analyse, on peut observer que la tendance actuelle est de décorréler ces différentes étapes en confiant chacune d'elle à un outil spécialisé ou à un module particulier. L'utilisation de moteurs *open source* comme Lucene facilite les expérimentations et le développement de modules de traitement spécifiques dans des champs encore réservés à l'innovation. Certaines solutions du marché se concentrent sur l'une des étapes seulement et s'utilisent en conjonction les unes avec les autres. D'autres produits gardent leur compétence sur l'ensemble de la chaîne d'indexation et d'analyse, mais leur paramétrage devient tellement précis et ouvert qu'il nécessite de très hautes compétences. Dans tous les cas, le moteur de recherche n'est pas, ou n'est plus, un logiciel « boîte noire » que l'on peut se contenter d'installer clé en main.

Dans ce contexte, il paraît illusoire d'imaginer que les bibliothécaires pourront se dispenser d'acquérir un minimum de connaissances et de savoir-faire concernant le fonctionnement et le paramétrage des moteurs de recherche. À terme, le moteur de recherche envisagé comme un élément central de la bibliothèque numérique nous amènera probablement à repositionner une partie de nos outils et de nos pratiques : le catalogue, comme un vivier de métadonnées structurées ; les actions des utilisateurs, comme une alternative à nos thésaurus ; nos classifications et nos vocabulaires contrôlés, comme une modélisation de la connaissance ; nos sites internet, comme un autre libre accès, dans lequel chacun viendrait réorganiser les rayonnages à sa guise.

Septembre 2007

14. Catalogue collectif international d'OCLC : http://worldcat.org

15. Voir http://theeuropeanlibrary.org