Réseau des bibliothèques de Suisse occidentale
Westschweizer Bibliotheksverbund
Rete delle biblioteche della Svizzera occidentale
Library Network of Western Switzerland

Objet	**Complément d'étude sur l'indexation matières dans RERO** (suite au rapport "Indexation matières dans RERO" publié le 01.04.2011)	**Date** 07.07.2011 **Version** finale
Description	**Complément d'étude sur l'indexation matières dans RERO selon la feuille de route approuvée par le Bureau du Conseil stratégique en date du 10 mars 2011.** **Ce rapport a été établi par le groupe de travail composé de :** - Benoît-J. Pédretti, (responsable du groupe) - Pierre Buntschu, BCU-Fribourg - Jeannette Frey, BCU-Lausanne - Miguel Moreira, RERO	
Statut du document	En séance du 24 juin 2011, le Conseil stratégique RERO a approuvé ce rapport et sa recommandation. Il a décidé la mise en œuvre de la solution proposée, selon le calendrier indiqué.	
Diffusion	Libre	

www.rero.ch info@rero.ch
tél: +41 27 721 85 85 fax: +41 27 721 85 86 av. de la Gare 45 CH – 1920 Martigny

Réseau des bibliothèques de Suisse occidentale
Westschweizer Bibliotheksverbund
Rete delle biblioteche della Svizzera occidentale
Library Network of Western Switzerland

Complément d'étude sur l'indexation matières

Introduction

Suite au rapport du groupe de travail sur l'indexation matière mis en place en 2010, intitulé « Indexation matières dans RERO » du 13 décembre 2010, le Conseil stratégique, dans sa séance du 18 février 2011, n'a pas souhaité se prononcer définitivement sur les questions posées, mais a admis et validé les principes suivants :

- adoption d'une simplification de l'indexation matières et poursuite du dossier ;
- définition d'une feuille de route pour identifier et prioriser les problématiques ;
- création d'un groupe de travail pour complément d'étude, visant à apporter l'information nécessaire à une prise de décision par le Conseil stratégique le 24 juin 2011 (cf. feuille de route en annexe).

Le présent rapport précise et met en perspective l'analyse et les propositions du rapport de 2010. Il expose des éléments d'informations supplémentaires :

- en définissant deux grandes orientations d'évolution possible complémentaires soit une indexation manuelle simplifiée, soit une indexation automatique libre ou contrôlée par un vocabulaire ;
- en décrivant les conditions de mise en œuvre à envisager, les avantages, comme les limitations ;
- en présentant une recommandation.

Indexation manuelle simplifiée

Le renoncement à la précoordination

La mise en place d'une indexation manuelle simplifiée par les professionnels passe par le renoncement à la précoordination. La précoordination actuelle dans RERO est partielle, facile à abandonner et utilisée par une minorité de spécialistes.

Y renoncer générerait, à la fois, un gain de temps et d'énergie non négligeable au sein des institutions, mais permettrait aussi par une juxtaposition simple de descripteurs, de répondre, sans difficulté, aux requêtes des usagers dans un OPAC, muni d'un système à facettes.

Cette simplification dans la gestion au quotidien du vocabulaire peut s'effectuer sous deux options :

- le maintien du vocabulaire RERO, tel qu'existant aujourd'hui, sans précoordination ;
- l'adoption d'un autre vocabulaire également sans précoordination. RAMEAU, vocabulaire contrôlé par la Bibliothèque nationale de France (BnF) et dont s'inspire déjà très largement le vocabulaire RERO est le plus facilement adaptable à la situation du réseau.

RAMEAU, nouveau vocabulaire pour le réseau

Conditions d'adoption de RAMEAU pour les noms communs

Des possibilités ouvertes

La BnF, ayant déjà mis en place des coopérations internationales de mise à disposition du vocabulaire RAMEAU (par exemple, Belgique, Maroc, Pologne et Roumanie) envisage avec enthousiasme la mise en place d'une convention avec RERO, visant à la libre utilisation par le réseau du vocabulaire RAMEAU sans coût aucun (cf. courrier Bnf en annexe).

L'abandon de la précoordination

RAMEAU est utilisé pour le moment en précoordination, sans perspective d'abandon à court terme. Cette précoordination plus complexe que dans le vocabulaire RERO serait, dès lors, bien sur, abandonnée. La BnF, si elle souhaite que le vocabulaire RAMEAU soit respecté, ne met aucun obstacle à son emploi par RERO sans précoordination. Les responsables de RAMEAU sont ouverts et intéressés par l'utilisation de RAMEAU en post coordination.

L'utilisation de RAMEAU en post-coordination ne va cependant pas de soi. Le vocabulaire RAMEAU intègre – et pas seulement dans les règles de constitution des chaînes (pré coordination) – de nombreuses règles portant sur la combinaison possible de certains descripteurs entre eux. RAMEAU parle d'emploi "approprié". Si une bonne partie du vocabulaire serait utilisable sans problème particulier (utilisation des vedettes simples), plusieurs cas pourraient poser des problèmes (cf. Particularismes d'indexation. Subdivisions), auxquels il faudra être attentif.

Récupération des données pré-existantes: exploitation et valorisation

Cette récupération est un pré-requis impératif à toute modification du système d'indexation, afin de s'assurer de la capitalisation et de l'utilisation de l'indexation passée.

Deux possibilités s'offrent au réseau:

- une juxtaposition : elle consiste à mettre dans une zone particulière de la notice, différente de celle de RAMEAU, les termes utilisés jusqu'à présent, avec un index commun, qui viendra moissonner ces deux zones. Cet index ne sera pas totalement cohérent, impliquera des répétitions mais sera utile pour l'utilisateur ; cette solution a les caractéristiques suivantes :
  - sans perte : récupération à l'identique de l'indexation existante, associé à RAMEAU, vocabulaire plus riche,
  - facile : un traitement de masse automatisé des zones correspondantes peut être mis en place.

  Cette solution est donc envisageable.
- une fusion avec conversion : elle consiste, d'abord, à fusionner de façon automatique les termes communs pour la mise en place d'un index commun RAMEAU + RERO dans l'interface de recherche. Les éléments qui n'auront pu être fusionnés seront soit juxtaposés (cf. supra), soit traités manuellement, de façon ponctuelle, non exhaustive et en privilégiant les termes spécifiques ; cette solution a les caractéristiques suivantes :
  - sans perte dans les deux cas : récupération des données fusionnées ou juxtaposées,
  - un traitement manuel, lourd et consommateur de ressources avec des règles complexes à mettre en place, mais permettant d'arriver à un index plus propre, dans le second cas.

  Cette solution est donc envisageable, mais plus consommatrice de temps.

Suivi et enrichissement du vocabulaire

La gestion des noms communs sera prise en charge par la BnF, qui a mis en place un fichier national des propositions RAMEAU (FNPR), maintenu par 8 collaborateurs avec :

- remontées de demandes argumentées et référencées (d'au moins une référence bibliographique), par les partenaire ;
- examen par l'équipe RAMEAU (1500 demandes par an et 8% de refus, souvent pour des problèmes formels liés à la demande) ;
- traitement sous 15 jours[1].

Les liens du thesaurus (association, généralisation, spécialisation), ainsi que l'équivalence LCSH, sont établis à chaque création de notice par l'équipe RAMEAU, ce qui constitue un atout sensible pour RERO.

Modèle de participation

L'adoption de RAMEAU réside dans le transfert d'une partie de la gestion du vocabulaire des équipes matières RERO vers l'équipe RAMEAU.

Les demandes de termes seront faites auprès de la BnF suivant une modélisation à définir (par exemple, la coordination locale désignera un ou plusieurs correspondants Matières pour chaque institution). L'ensemble de l'édifice de la gestion des noms communs tel qu'il existe actuellement sera donc considérablement réduit - moins de 10 personnes, correspondants matières auprès de la BnF (contre 44 actuellement) et seuls désignés pour lui faire des propositions. Une coordination de ce petit groupe pourrait être assurée par la centrale RERO pour uniformisation de la qualité et aide à la documentation des propositions.

La solution d'un correspondant par établissement est inenvisageable pour la BnF. En sus d'une grande difficulté à gérer sans pré filtrage, elle créerait des disparités visibles dans la qualité des propositions, qui nuiraient autant à l'efficacité qu'à l'image du réseau.

Mises à jour

Des modifications, mises à jour, nettoyages de l'index sont réalisés en continu par l'équipe RAMEAU: 18'000 descripteurs ont été revus depuis trois ans. Elles sont annoncées dans le "Journal RAMEAU des créations et des modifications" sur le site Internet de la BNF (version PDF avec fichiers d'autorités mis à jour, munis d'identifiants).[2]

Ces fichiers peuvent être récupérés, comme produits, à une périodicité à définir, permettant une mise à jour des autorités du réseau. La suppression/fusion de notices RAMEAU ou la séparation de séquences créant de nouvelles vedettes dans le vocabulaire RERO nécessitera un traitement automatique ou manuel des notices bibliographiques liées, doublées d'un signalement et d'un contrôle manuel de sécurité (particulièrement pour les vedettes construites). A la différence d'Amazon, qui reçoit les fichiers d'autorités RAMEAU contre paiement, les mises à jour seront à disposition de RERO gratuitement.

Prise en compte des besoins locaux RERO

La BnF assure le réseau de la sauvegarde et de la prise en compte du vocabulaire régional spécifique, helvétismes de deux catégories :

- terme désignant un particularisme local (ex.: un type de musique spécifique à une région) ; le terme sera intégré sans difficulté dans RAMEAU ;
- terme existant déjà dans RAMEAU, mais désignant une réalité différente (ex.: «élections cantonales » en France et en Suisse). Il sera alors adopté un descripteur composé, type "élections cantonales (Suisse - canton)".

Pour tous les besoins locaux, il sera aussi possible de rajouter un terme rejeté, permettant un renvoi opportun. En revanche, les termes associés/spécifiques/génériques seront gérés uniquement par l'équipe RAMEAU sur demande.

La BnF entrera en matière pour une opération de chargements par lots des descripteurs identifiés par RERO comme manquants actuellement dans RAMEAU (processus déjà expérimenté avec des bibliothèques belges).

[1] On notera qu'actuellement l'examen des descripteurs dans RERO prend environ une année, mais permet une utilisation en attente de validation. Cette possibilité de créer des notices provisoires dans l'attente de la validation RAMEAU sera maintenue.

[2] Ne sont pas annoncées les modifications de termes rejetés, mais uniquement les vedettes.

tél: +41 27 721 85 85 fax: +41 27 721 85 86 av. de la Gare 45 CH – 1920 Martigny

Particularismes d'indexation

- Vedettes construites: RAMEAU contient de nombreuses vedettes "construites" pour permettre la gestion des liens dans le thésaurus. Le choix va se poser de les utiliser en vedette construite ou la combinaison équivalente des descripteurs indépendants. Pour simplifier l'indexation, il serait intéressant d'écarter ces constructions ou de les corriger au moment de l'importation des données de RAMEAU. Certains descripteurs construits (termes composés ou vedettes de regroupement) devraient être entrés sous l'un des termes en vedette et l'autre en terme rejeté associé.

  A contrario, il est très important de conserver ces constructions, car elles servent de pivot pour les jeux de relations établies sans RAMEAU (termes associés, termes rejetés). Cette question devra être soigneusement étudiée.
- Subdivisions: un choix devra être fait entre les subdivisions à utiliser comme descripteur principal ou pas. Sauf pour les vedettes de forme, les subdivisions pures seront écartées au profit des descripteurs à part entière déjà existants dans RAMEAU.

  Il sera apporté une attention particulière aux points suivants :
  - Les subdivisions différentes du descripteur de base ;
  - les subdivisions difficilement utilisables de manière indépendante ;
  - les domaines d'application.
- Ouvrages généraux : Indépendamment de l'utilisation du vocabulaire RAMEAU ou RERO, l'utilisation en post-coordination pose aussi le problème de la recherche des documents généraux sur un sujet. Dans un système post-coordonné, cette notion est contenue "par défaut" dans les chaînes générales, avec un seul descripteur ou une combinaison de descripteurs. Une solution partielle pourrait être apportée par la création d'une forme "ouvrage général", identique à "Gesamtdarstellung" dans la SWD.
- Production concrète/parties de l'œuvre: leur gestion reste possible dans RAMEAU associés à des noms de personne.
- Codes de regroupement par domaines : ces codes utilisés dans RAMEAU et inspirés de la Dewey sont utiles pour regrouper les termes d'un même domaine. Leur utilisation aménagée pour RERO présente un vrai intérêt.

Le choix d'une approche pour les noms propres

Gestion des noms propres RAMEAU

Elle se décompose en deux lots d'autorités : les autorités utilisées par le seul établissement BnF, dites Autorités BnF (ATC) et les autorités utilisées par le réseau, dites Autorités RAMEAU.

Autorités BnF (ATC)

Maintenues et utilisées par la seule BnF, elles sont également utilisées dans l'indexation et juxtaposées aux descripteurs des Autorités RAMEAU. Ces autorités ne sont pas à disposition d'un partenaire extérieur.

Autorités RAMEAU

Ce vocabulaire à disposition du réseau comportent des :

- noms de peuples ;
- marques ;
- noms de personnes : dieux, déesses ou personnages mythologiques ;
- noms géographiques (+ des noms géographiques liés à des dates ou des tranches chronologiques (non gérés seuls)
- noms de personnes et titres avec des subdivisions spécifiques (ex : "Marie-Antoinette -- procès" ; "Bible -- concordances")

Au total en sus des 90'000 noms communs, les autorités RAMEAU comptent 9'345 noms de personnes et 54'612 noms géographiques. Le suivi est identique à la gestion des noms communs avec la possibilité de proposer de nouveaux noms propres s'ils sont accompagnés d'une subdivision.

Gestion des noms propres souhaitable dans RERO

La gestion des noms propres dans RERO est légère et conduite par les coordinations locales. Elle ne nécessite pas de grands changements, sinon l'abandon de la précoordination et la valeur ajoutée qu'apporte les autorités RAMEAU noms propres.

Il peut donc être envisagé l'association sous la forme d'autorités noms propres des :

- autorités ATC RERO (noms de personnes) ;
- autorités matières RERO ;
- autorités RAMEAU noms propres en totalité (les noms géographiques étant l'apport le plus appréciable).

Cette possible fusion, dont les conditions restent à définir, nécessitera :

- des travaux préalables d'examen des subdivisions ;
- une transformation des noms gérés actuellement dans RERO ;
- l'utilisation, si besoin, de formes alternatives, plus immédiatement accessibles dans un index virtuel, générés pour les interfaces de recherche à facettes.
- La prévision d'une synchronisation, lors de l'importation des notices RAMEAU.

Au sortir, RERO bénéficiera d'un réservoir d'autorités noms propres, dans lequel pourront être puisés et satisfaits les besoins auteurs/collectivités et toutes matières, mais aussi où seront créées les nouvelles autorités noms propres.

Conclusion

L'indexation manuelle simplifiée avec les autorités noms communs (RAMEAU) et les autorités noms propres ATC est une solution clairement envisageable :

- avec un modèle de fonctionnement et de participation négociée pouvant convenir à la structure et aux habitudes du réseau ;
- avec un vocabulaire bien pris en charge et maintenu ;
- avec un respect des besoins locaux ;
- ce changement n'a pas d'impact sur les autres vocabulaires employés dans le réseau (Jurivoc, MeSH, LCSH), non concernés.

Elle doit être envisagée avec un accompagnement bien pensé :

- une phase de communication encadrée ;
- une phase d'introduction, permettant une mise en place et une adaptation en souplesse au sein du réseau ;
- une phase opérationnelle incluant le suivi des différents outils et dispositifs mis en place, la prise en considération d'un travail de préparation important mais assimilable par le réseau et la gestion de l'indexation dans le fichier bibliographique en lien avec les modifications de RAMEAU.

Cette solution offre des avantages certains, comme autant des points très positifs pour l'amélioration de la qualité de service aux usagers de RERO :

- elle induit un gain immédiat incontestable qui transfère la gestion ordinaire, le suivi et la mise à jour du vocabulaire sur l'équipe RAMEAU ;
- elle permet une structure thesaurus, qui supporte le multilinguisme ;
- elle offre un lien vers le web de données (web sémantique).

Indexation automatique

L'indexation automatique, sujet d'actualité s'il en est, fait l'objet d'initiatives nombreuses en Suisse comme à l'étranger. Elles répondent aux besoins signalés par de nombreuses institutions désireuses de trouver une solution simple, pratique et économe pour assurer les tâches quotidiennes du traitement documentaire.

Conditions de mise en œuvre

La mise en place d'une indexation automatique présuppose, comme une indexation manuelle simplifiée, le choix d'une solution d'avenir qui permette :

- le transfert et la récupération complète des données actuelles ;
- une souplesse, une adaptabilité du système et une rapidité du processus ;
- la définition d'une chaîne de traitement idoine par l'acquisition aisée et rapide de ressources, permettant une indexation automatique et l'extraction de mots matières :
  - par numérisation de tables de matière et/ou de résumés, puis océrisation (nécessitant l'acquisition de scanners sur les sites, sans transfert d'ouvrages, ni mutualisation trop difficiles à mettre en place),
  - par acquisition de tables de matière et/ou de résumés électroniques, avec ou sans index ;
- la définition des modalités de contrôle :
  - les erreurs potentielles de la machine étant de nature différente de celles de l'intervention humaine et l'indexation de certains termes (ex. termes génériques), pouvant donner des résultats moins probants, une attention particulière doit être portée aux modalités de contrôle, afin de ne pas péjorer le gain de temps escompté.
  - parmi les solutions possibles, un contrôle systématique et complet par l'indexeur ou le catalogueur est exclu, aucun contrôle serait dommageable, un simple pointage par échantillonnage serait raisonnable.

Il est essentiel, en outre, de veiller à :

- une coordination avec la Discovery solution (OPAC) qui sera choisie fin 2011-début 2012 et sa propre gestion de l'indexation ;
- un accompagnement du changement pour le réseau ;
- la mise en place d'une procédure de sélection encadrée avec un cahier des charges et une évaluation des solutions qui seront proposées.

Avantages et limitations

Avantages

- une économie immédiate et réelle de moyens par l'intégration des opérations liées à l'indexation automatique dans la gestion des tâches courantes ;
- l'enrichissement de la notice par intégration automatique des mots matières et surtout par la mise en place d'un lien notice-table des matières/résumés ;
- la facilitation de la recherche pour les professionnels et le public ;
- la mise à disposition par RERO de la solution choisie pour l'ensemble du réseau avec une uniformisation des pratiques.

Limitations et questions en attente

- La transformation d'une activité traditionnelle intellectuellement stimulante en une activité de contrôle plus ordinaire, mais nécessitera, au final, moins de temps, de compétences et de formation ;
- L'existence d'expérimentations, suite à de sérieuses études, permettant de démontrer l'intérêt effectif de l'indexation automatique, sans développement pour le moment de solutions immédiatement adaptables à RERO ;
- Eléments à prendre en compte et qui n'ont pu encore être évalués :
  - le traitement du multilinguisme : l'indexation automatiquement de documents dans plusieurs langues (français, allemand, anglais, etc…), appuyée sur un seul vocabulaire francophone est-il possible, dès lors que des liens linguistiques préalables auront été établis ou faut-il plutôt employer un vocabulaire différent pour chaque langue dont les liens seront établis ensuite ?
  - le traitement des documents non textuels (image, son, etc…) : l'indexation automatique des bases images et des notices de supports sonores fait l'objet de

nombreuses expérimentations ou projets aboutis. Toutefois, pour le moment, aucun projet ambitieux d'indexation automatique tous supports n'est identifié.

- les termes des vocabulaires contrôlés demeurent parfois éloignés du langage naturel des utilisateurs. Les habitudes de recherches évoluent et les renvois/termes rejetés prévus dans les vocabulaires contrôlés sont-ils suffisants ?
- l'ordinateur n'étant pas force de proposition de nouveaux termes, comment faire évoluer et enrichir le vocabulaire pour les besoins locaux s'il n'y a aucune indexation manuelle ?

Success stories / initiatives

Avec vocabulaire contrôlé

Petrus, DNB, Allemagne

http://www.d-nb.de/wir/projekte/petrus.htm

http://files.d-nb.de/pdf/petrus/petrus_dialog_2011_1.pdf

En 2009, la Bibliothèque nationale allemande a commencé à tester des procédures d'indexation automatique. L'objectif de ce projet, nommé « PETRUS », est d'établir un lien entre le traitement conventionnel de documents et les procédures de traitement automatique. Ce projet part du constat que, dans l'accomplissement du mandat légal d'indexer toutes les publications, il n'y a pas d'autre alternative que de faire confiance aux logiciels. La croissance rapide du nombre de titres imprimés, ainsi que le nombre énorme de ressources Web dans le cadre du domaine internet « .de » exige des flux de travail non intellectuel et une accélération des mécanismes habituels de catalogage. Dans la mesure du possible, des procédures assistées par ordinateur devraient aider les indexeurs et contribuer ainsi à économiser des ressources humaines pour ce qui est vraiment nécessaire dans l'amélioration de la recherche en ligne.

Les recherches dans le cadre du projet portent essentiellement sur quatre scénarios d'application :

- détection automatique de documents parallèles ou similaires, et échange de métadonnées entre eux ;
- génération automatique de nouvelles données d'autorité personne par l'importation de notices et la mise en relation des noms de personnes et des titres ;
- classification automatique des publications en ligne avec la classification décimale de Dewey ;
- attribution automatique de mots-clés basés sur le vocabulaire contrôlé SWD.

Pour étudier la faisabilité des différents scénarios, la DNB teste l'emploi de plusieurs solutions logicielles du commerce.

Le dernier scénario mentionné (indexation automatique) est celui qui le plus proche de notre préoccupation actuelle. Il constitue même, parmi les différents projets de ce genre menés dans différents pays, celui qui correspond le plus aux objectifs visés par RERO.

IDS, Suisse

http://files.d-nb.de/pdf/petrus/computerunterstuetzte_sacherschliessung_schweiz.pdf

La KDH (Konferenz Deutschschweizer Hochschulbibliotheken), partant du constat que le coût en ressources humaines dédiées à l'activité d'indexation matières manuelle est très élevé, a adopté en 2009 l'objectif stratégique de revoir entièrement le processus général d'indexation matières dans les bibliothèques du réseau IDS.

Parmi les mesures principales préconisées figure l'adoption de méthodes d'indexation matières automatique avec une réduction drastique du travail intellectuel, limité à un ensemble très spécifique de cas.

Des expériences menées en 2010 par la ZBZ (Zentralbibliothek Zürich) avec l'entreprise Eurospider, dans le cadre d'un projet pilote visant un benchmarking comparatif manuel et automatique, ont permis de conclure que le taux de précision des approches automatiques peut atteindre environ 80%, ce qui correspond aux résultats obtenus avec l'indexation manuelle (les erreurs ou inconsistances produites par les deux approches sont, en revanche, de types différents, ce qui correspond à des conclusions similaires retrouvées dans la

littérature spécialisée). Autre conclusion : certains termes de caractère général sont plus difficiles à traiter avec des approches automatiques.

Le projet se poursuit en 2011 avec l'introduction graduelle des méthodes automatiques dans un sous-ensemble de domaines scientifiques (droit, économie, sciences sociales) et l'élargissement des tests à des documents dans d'autres langues que l'allemand (français, anglais).

Avec ou sans vocabulaire contrôlé

JISC, Royaume Uni

Projet Merlin (Metadata Enrichment for Repositories in a London Institutional Network)

Financé par le JISC, avec pour participants :

- UCL (University College London)
- University of London Computing Centre
- University of Nottingham

L'objectif de ce projet est d'évaluer les possibilités d'utilisation d'outils existants de text mining, associés à des thésaurus, pour extraire des mots-sujets descriptifs à partir du texte intégral des documents déposés dans un serveur institutionnel (LASSO), et ainsi améliorer les possibilités d'accès et de navigation dans les contenus de ce serveur.

http://www.ucl.ac.uk/ls/merlin/about.shtml

Plus généralement, le JISC finance un groupe de projets qui visent, entre autres, l'expérimentation et l'évaluation de différentes techniques de génération automatique de métadonnées et leur impact dans l'amélioration des procédures de recherche documentaire.

http://www.jisc.ac.uk/whatwedo/programmes/inf11/resdis.aspx

BnF, France

http://bbf.enssib.fr/consulter/bbf-2008-06-0020-004

La Bibliothèque nationale de France conduit depuis 2006 un projet d'indexation automatique des sites web du dépôt légal : une structure d'archivage du web sur petabox (baies de stockage allant jusqu'à 120 téraoctets) avec un processus d'indexation automatique complet permet une recherche par URL (logiciel Wayback Machine). Associé à cela, une expérimentation sur l'indexation automatique des contenus de ces sites est en cours :

« L'indexation plein texte reste un objectif prioritaire pour permettre la recherche par mot. Elle s'avère en effet l'outil le plus intuitif et le plus attendu des utilisateurs, qui veulent naturellement rechercher dans les archives comme ils le font sur le web. Le logiciel NutchWAX est encore expérimental, et moins de 5 % des collections de la BnF ont pu être indexés en plein texte. Compte tenu des performances et des coûts d'indexation actuels, on envisage de réserver, dans un premier temps, ce mode d'indexation aux collections issues d'une sélection humaine, afin de les valoriser.

D'autres opérations de signalement sont en cours d'expérimentation. Pour celles-ci, on le verra, l'intervention du bibliothécaire peut être sollicitée. Cette contribution est cependant davantage éditoriale (valorisation de corpus, thèmes, actualités ou parcours) que descriptive. Elle s'inscrit dans l'évolution plus globale de la bibliothèque éditrice, c'est-à-dire agissant en tant qu'architecte d'interfaces et de services plutôt que comme opérateur direct du traitement des données.

À terme, au-delà de la consultation des données, la BnF souhaite ainsi s'impliquer dans la recherche-développement d'outils d'exploitation et d'analyse de gros volumes (fouille de données, cartographies dynamiques du web…). La structure de la collection constituée permet en effet de retrouver non seulement des sites web, mais aussi des informations précises sur leur contexte de publication, les relations entre sites, ainsi que l'évolution de ces relations dans le temps et la géographie du web (ce qu'on appelle le link mining). C'est un secteur particulièrement prometteur, notamment pour la nouvelle sociologie du web, où des développements importants sont attendus. »

Technologies disponibles

Produits et entreprises impliqués dans les projets germanophones (DNB, IDS) :

- Averbis Extraction Platform (Averbis GmbH) http://www.averbis.de/
- TopicFinder + iFinder (Intrafind Software GmbH) http://www.intrafind.de/

- iSquare SmartSearch (iSquare GmbH) http://www.isquare.de/
- RapidMiner (Rapid-I GmbH) http://rapid-i.com/content/view/181/190
- Eurospider http://www.eurospider.com/

Autres entreprises et projets :

- AUTINDEX http://www.agi-imc.de
- iVia LCSH Metadata Assignment (University of California Riverside Libraries) http://ivia.ucr.edu/projects/Metadata/LCSH.shtml

Autres logiciels :

- LINGO http://www.lex-lingo.de
- EXTRAKT http://www.textec.de/
- Semtinel/K.A.I.E.C. http://www.kaiec.org/
- NutchWAX http://archive-access.sourceforge.net/projects/nutch/
- Wayback machine http://archive-access.sourceforge.net/projects/wayback/
- IDX, Produit de Microsoft, intégré dans les projets de MILOS I+II et KASCADE

Indexation automatique : libre ou contrôlée

Deux options sont envisageables pour l'indexation automatique :

- une indexation automatique libre sans appui sur un vocabulaire ;
- une indexation automatique contrôlée appuyée sur un vocabulaire.

Chacune de ces deux options peut éventuellement cohabiter avec une indexation manuelle simplifiée.

Indexation automatique libre

Dans le cas d'une indexation automatique libre, les résultats viendraient se cumuler avec l'indexation pré-existante récupérée de RERO. Si cette solution venait en cohabitation avec une indexation manuelle simplifiée, elle offrirait plusieurs désavantages :

- la possibilité offerte au réseau soit d'une indexation manuelle simplifiée, soit d'une indexation automatique libre, sans cohérence. Une telle solution irait à l'encontre des pratiques d'harmonisation souhaitables pour le réseau ;
- des résultats associés mais disparates.

Indexation automatique appuyée

L'option d'une indexation automatique appuyée sur un vocabulaire contrôlé aurait pour avantages :

- le passage à un vocabulaire cohérent comme RAMEAU ;
- un enrichissement du catalogue par un taux d'indexation supérieur de meilleure qualité ;
- une meilleure aide à la recherche au service du public ;
- cette solution peut aussi venir en cohabitation avec une indexation manuelle simplifiée, qui sera encadrée comme un moyen de traiter les reliquats non pris en charge par l'indexation automatique, ou comme un moyen de contrôle a posteriori. En aucun cas, cette indexation manuelle simplifiée ne devra être une alternative, permettant de continuer, à loisir, une indexation traditionnelle, avec un vocabulaire à choix (RERO ou RAMEAU). Dans ce cas, des directives très précises devront être mises en place ;
- le maintien d'une cohérence au travers d'un vocabulaire commun, suivi et contrôlé a minima sans précoordination ;
- la mise à disposition de l'utilisateur d'une interface de recherche basée sur des données rapidement extraites et cohérentes.

Schémas

Les schémas suivants permettent de mieux appréhender les processus impliqués.

Schéma 1 : traitement des données actuelles lors de la transition

Ce schéma illustre le traitement des notices d'autorités et bibliographiques, avec une adaptation automatique à affiner dans ses modalités[3] et un passage à RAMEAU en douceur avec une ouverture sur une indexation automatique appuyée pour les acquisitions à venir.

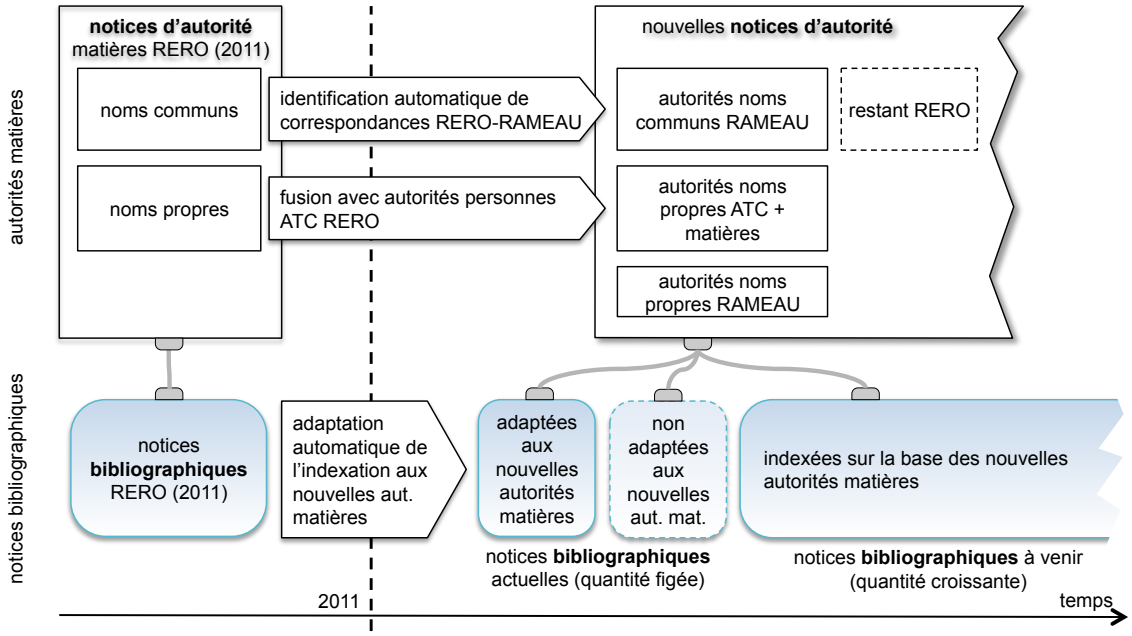
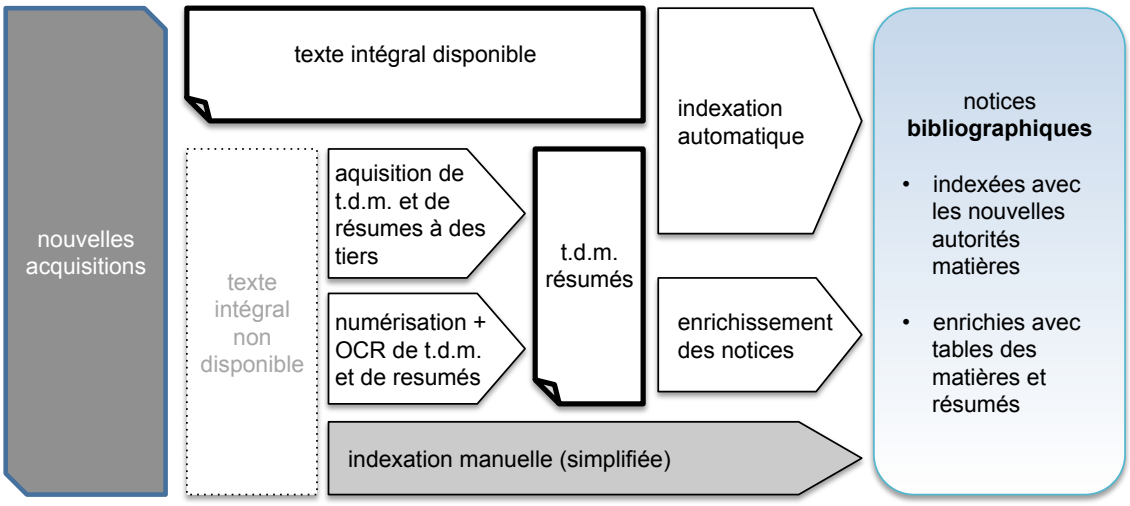


Schéma 2 : chaîne de traitement des données des nouvelles acquisitions

Ce schéma illustre le traitement des nouvelles acquisitions par une indexation automatique appuyée sur RAMEAU, complétée par une indexation manuelle simplifiée si besoin et un enrichissement des notices par des données extraites ou acquises.



[3] Juxtaposition ou fusion devront s'assurer de valider à la fois le principe initial de simplification des pratiques, mais aussi celui de la plus grande automatisation possible du processus.

Recommandation

La recommandation du groupe de travail au Conseil stratégique est la suivante :

- La poursuite de ce dossier est indispensable dans la double perspective de simplification des pratiques des professionnels et de mise à disposition d'outils de recherche adaptés et de qualité pour le public.
- De nombreuses expérimentations d'indexation automatique avec ou sans vocabulaire à l'appui sont à l'œuvre et toutes les grandes institutions et réseaux ont compris la nécessité d'une solution pratique, rapide et efficace dans cette direction.
- La solution proposée devra faire l'objet d'un accompagnement avec des directives claires et une communication soignée.

Cette solution assurera à la fois une simplification des pratiques, une ouverture vers des solutions complémentaires (multilinguisme, web sémantique, etc…) et positionnera RERO directement et plus rapidement vers des solutions d'avenir plus ouvertes.

Le phasage suivant pourrait désormais être envisagé :

	Indexation manuelle simplifiée	**Indexation automatique appuyée**
Eté 2011	Mise en place d'un groupe de travail, chargé du suivi, appuyé de 2 sous-groupes spécialisés.	
2è semestre 2011	Conventionnement avec la BnF Simplification et transformation progressive de la structure de coordination.	Etude d'une chaîne de traitement documentaire automatisée Mise en place d'un cahier des charges et le lancement d'un appel d'offres pour une solution d'indexation automatique.
Courant 2012	Mise en correspondance RERO-RAMEAU Ediction de directives et de documentations Fusion des autorités noms propres.	Tests des solutions proposées.
1er juillet 2012[4]	**Abandon définitif de la pré-coordination** **Abandon du vocabulaire RERO au profit du vocabulaire RAMEAU**	
Fin 2012-début 2013	Conversion des données préexistantes.	Choix d'une solution coordonnée avec la nouvelle Discovery solution (OPAC), mise en place par RERO.

8 juin 2011

[4] En fonction de l'avancement des travaux, une option peut être envisagée d'un passage différé au 1er mars 2013.

Réseau des bibliothèques de Suisse occidentale
Westschweizer Bibliotheksverbund
Rete delle biblioteche della Svizzera occidentale
Library Network of Western Switzerland

Annexes

- Feuille de route pour un complément d'étude approuvée par le Bureau du CS du 10 mars 2011.

- Courrier de la Bibliothèque nationale de France (BnF), centre national RAMEAU du 7 juin 2011.

www.rero.ch info@rero.ch
tél: +41 27 721 85 85 fax: +41 27 721 85 86 av. de la Gare 45 CH – 1920 Martigny

Réseau des bibliothèques de Suisse occidentale
Westschweizer Bibliotheksverbund
Rete delle biblioteche della Svizzera occidentale
Library Network of Western Switzerland

Feuille de route
approuvée par le Bureau CS
en séance du 10.03.2011

Note au Bureau du Conseil stratégique

Sujet	**Politique de l'indexation matières dans RERO** **Feuille de route pour un complément d'étude**
Distribution	BCS
Auteurs	Direction RERO, M. Moreira
Date	08.03.2011

Le document de référence 110218.5.2 « Politique de l'indexation matières : Prise de décision du Conseil stratégique, à la demande du Bureau CS » invitait le CS à prendre des décisions sur des points essentiels dans le cadre de ce dossier. Dans sa séance du 18.02.2011 le CS n'a pas souhaité se prononcer définitivement sur les questions posées mais a validé les principes suivants :

- Adoption du principe d'une simplification de l'indexation matières et poursuite du dossier.
- Définition d'une feuille de route identifiant les différentes problématiques dans un ordre de priorité à développer.
- Création d'un groupe de travail restreint pour poursuivre l'étude, en préparant un rapport complémentaire à présenter lors de la séance du Conseil stratégique du 24.06.2011 pour une prise de décision.

Cette étude complémentaire vise à apporter au CS les compléments d'information lui permettant de prendre les décisions.

Etude complémentaire

L'étude complémentaire doit couvrir les éléments spécifiés ci-après en encadré, dont la liste est énoncée en suivant la forme de présentation du document de référence 110218.5.2.

1. Décision du CS sur le principe du maintien ou de l'abandon d'une indexation matières dans RERO (Cf. proposition P12 du rapport)

   (Le CS adhère de manière générale au besoin d'une indexation matières dans RERO, sous une forme à repenser)

2. En cas de maintien d'une indexation matières, le CS est prié

2.a de confirmer le besoin d'une simplification de l'activité d'indexation dans RERO;

   (Ce principe est acquis par le CS)

2.b si la simplification est confirmée, de se prononcer sur le processus d'indexation, à savoir:

i) mise en place d'une indexation automatique (Cf. proposition P8, voire P9, à explorer)

Indexation automatique – étude des aspects suivants :

- technologies disponibles
- conditions de mise en oeuvre
- avantages et limitations
- success stories

www.rero.ch info@rero.ch
tél: +41 27 721 85 85 fax: +41 27 721 85 86 av. de la Gare 45 CH – 1920 Martigny

ii) mise en place d'une indexation manuelle simplifiée, par les professionnels, selon les recommandations du rapport (Cf. Proposition P7), avec l'une et/ou l'autre des modifications suivantes :

- le renoncement à la précoordination,
- l'abandon du vocabulaire RERO en faveur de RAMEAU pour les noms communs,
- le choix d'une approche pour les noms propres;

Indexation manuelle simplifiée – étude des aspects suivants :

- confirmation de la possibilité d'utiliser RAMEAU en postcoordination
- proposition d'un scénario pour la gestion des noms propres
- possibilités d'exploitation et de mise en valeur des données d'indexation RERO existantes
- avantages et limitations
- success stories

iii) ou la cohabitation des 2 techniques (automatique et manuelle)

Cohabitation des indexations automatique et manuelle – étude des aspects suivants :

- avantages et limitations
- gain d'une cohabitation (critères à prendre en compte : taux d'indexation, enrichissement du catalogue, ressources nécessaires en termes de moyens humains, temps de travail, infrastructure, ...)

3. Dans le cas d'une décision en faveur d'une indexation manuelle, le CS est invité à préciser si cette activité est libre ou obligatoire pour les partenaires RERO.

Groupe de travail

3-4 personnes :

- B. Pédretti, directeur adjoint RERO
- M. Moreira, chef de projet
- 1-2 spécialistes RERO

Délivrables

Le rapport de complément d'étude est à transmettre à la direction RERO pour le **8 juin 2011**, en vue de son traitement par le Conseil stratégique lors de sa séance du 24 juin.

Décision du Bureau du Conseil stratégique du 10.03.2011

- La feuille de route telle que présentée pour un complément d'étude relatif à la politique de l'indexation matières est approuvée.
- Le groupe de travail est formé de Mme Jeannette Frey (BCU-Lausanne), MM. Pierre Bunstchu (BCU-Fribourg), Miguel Moreira (RERO), Benoît Pédretti (RERO, responsable du groupe).

Bibliothèque nationale de France
Direction des Services et des Réseaux
Département de l'information bibliographique et numérique
Centre national RAMEAU

Paris, le 7 juin 2011

A l'attention de Madame Marylène Micheloud,
Directrice de RERO
avenue de la Gare, 45
1920 Martigny (VS)
Suisse

Madame la Directrice,

Permettez-moi de vous confirmer que la Bibliothèque nationale de France accueille très favorablement, et avec un vif intérêt, la possibilité que le réseau des bibliothèques de Suisse romane RERO utilise le vocabulaire Rameau (noms communs et noms propres) pour ses besoins documentaires.

La BnF se réjouit notamment que RERO, acteur et partenaire francophone de qualité, puisse devenir un contributeur régulier à l'enrichissement de Rameau par le biais du Fichier national des propositions Rameau (FNPR). En retour, RERO pourrait compter sur une mise à disposition du vocabulaire Rameau à titre gracieux ainsi que sur l'assurance d'une prise en compte bienveillante des besoins locaux et des helvétismes ; la direction du réseau pourrait également être invitée à contribuer aux projets scientifiques que la BnF développe, par exemple, en direction du web de données. Bien entendu, les modalités de ce partenariat devraient faire l'objet d'une convention signée entre la BnF et RERO.

La BnF se félicite, par avance, de la perspective d'une collaboration fructueuse et privilégiée qu'elle espère voir se concrétiser et qui serait confiée aux bons soins de M. Benoît-J. Pédretti, directeur adjoint de RERO, lien naturel entre nos deux établissements.

Je vous prie d'agréer, Madame la Directrice, l'expression de mes salutations les plus respectueuses et cordiales,

Michel Mingam
Centre national RAMEAU