Cederilic : constitution d'un livret d'un index numérique

Jean Charlet, Touria Aït El Mekki, Didier Bourigault, Adeline Nazarenko, Régine Teulier, Baruk Toledano

CEDERILIC : constitution d'un livre et d'un index numériques*

Jean Charlet[1], Touria Aït El Mekki[2], Didier Bourigault[3], Adeline Nazarenko[2], Régine Teulier[4], Baruk Toledano[5]

[1]STIM/DSI/AP–HP & INSERM ERM 202, Université Paris VI
jc@biomath.jussieu.fr

[2]LIPN, CNRS-Université Paris-Nord
{taem,nazarenko}@lipn.univ-paris13.fr

[3]ERSS, CNRS-Université Toulouse Le Mirail
[4]CRG
[4]LIP6, Université Paris VI

Résumé :

Nous décrivons une expérience en grandeur réelle de constitution d'un index thématique pour un ouvrage scientifique. Cet ouvrage est constitué d'une sélection de vingt-et-un articles de trois éditions des journées Ingénierie des connaissances (1999-2001). Ce corpus a été traité par l'analyseur SYNTEX puis par le système INDDOC, logiciel dédié à la constitution d'index. Ce travail a été réalisé dans un contexte entièrement numérique, c'est-à-dire à partir de fichiers numériques et pour constituer la collection des articles de l'ouvrage en un ensemble de fichiers HTML au sein duquel l'utilisateur navigue via un navigateur. Nous présentons les principaux problèmes rencontrés et les solutions adoptées.

MOTS-CLES : livre numérique, indexation, ingénierie des connaissances, acquisition de connaissances à partir de textes, structuration de terminologie, condensation de l'information, XML, DTD DocBook.

Abstract:

We describe a real experiment in order to build a thematic index of a scientific book. This book is a compilation of 21 articles from the French Knowledge

* CEDERILIC pour « Cédérom pour indexer le livre IC » est un projet soutenu par France-Télécom. En dehors de la forte activité de recherche suscitée par le projet, le soutien a principalement consisté dans le financement du stage de DESS de Baruk Toledano qui a réalisé les programmes de transformation et d'enrichissement des fichiers.

Engineering conferences (1999-2001). The corpus has been analysed by SYNTEX then by INDDOC, software dedicated to index formation. This work has been realized in a full digital context, with digital HTML articles and HTML index. The user uses a browser for exploring the articles through the index. We describe the work, the main problems and the chosen solutions.

KEYWORDS : digital book, indexation, knowledge engineering, knowledge acquisition from texts, terminology structuration, XML, DTD DocBook.

1. Introduction

Nous décrivons une expérience en grandeur réelle de constitution d'un index thématique pour un ouvrage scientifique. Cet ouvrage [TEUL 04] est constitué d'une sélection de vingt-et-un articles de trois éditions des journées Ingénierie des connaissances (1999-2001). Ce travail a été réalisé dans un contexte entièrement numérique, c'est-à-dire à partir de fichiers numériques et pour constituer la collection des articles de l'ouvrage en un ensemble de fichiers HTML que l'utilisateur peut consulter via un navigateur. Ce travail tire parti des expériences acquises par les auteurs :

- La constitution d'un index pour le livre sur l'Ingénierie des connaissances [CHAR 00], regroupant 35 articles des années 1995-1998 effectuée par D. Bourigault et J. Charlet [BOUR 99]. Plutôt que de faire appel aux auteurs des articles, ce travail innovait en exploitant les résultats fournis par un outil de traitement automatique des langues, l'analyseur syntaxique de corpus LEXTER (prédécesseur de SYNTEX [BOUR 00]) à partir de l'analyse automatique du corpus électronique constitué des trente-cinq articles sélectionnés. Si le repérage des candidats termes (Cf. § 3) nous a rapidement paru n'être qu'un point de départ, l'expérience nous a permis de repérer un certain nombre de difficultés et d'y apporter des solutions [BOUR 99].
- Le développement d'un système de constitution d'index par T. Aït El Mekki et A. Nazarenko, INDDOC [AITE 02]. Tirant parti des enseignements du travail précédent, l'équipe du LIPN a proposé une réflexion et une nouvelle architecture de constitution d'index. Cette architecture considère un index comme une ressource, constituée à partir d'un corpus, que des outils permettent d'ébaucher (index ébauche), que l'utilisateur complète (index source) et qui peut être visualisé.

Ainsi, que ce soit la première expérience, vis-à-vis de la complexité de l'index construit, ou la seconde, vis-à-vis de la complexité des fonctions attendues, tout concourait au développement d'un index numérique permettant de naviguer dans une collection d'articles numériques. Nous avons donc décidé de monter un projet qui visait à associer à un livre « papier », un cédérom proposant les articles indexés et permettant d'y accéder via l'index. Le projet comporte cinq étapes principales (*Cf.* figure 1) :

1. la transformation des articles du format d'origine (RTF) dans un format XML,
2. l'enrichissement de ce format selon plusieurs contraintes (visualisation,indexation),

3. le traitement du corpus ainsi constitué par SYNTEX pour obtenir les candidats termes nécessaires à l'étape suivante,
4. la constitution de l'index grâce à INDDOC,
5. la réalisation finale des fichiers à visualiser et de l'interface de navigation.

Nous décrivons, section 2, les tenants et aboutissants de notre approche, section 3, la constitution des ressources XML, section 4, le repérage des candidats termes par SYNTEX, section 5, le mode de constitution de l'index par INDDOC, et, section 6, la constitution de la ressource HTML sur laquelle navigue l'utilisateur. La section 7 décrit l'expérimentation d'un point de vue qualitatif et quantitatif. Enfin, nous essayons de tirer des conclusions et de proposer des perspectives à ce travail dans la section 8.

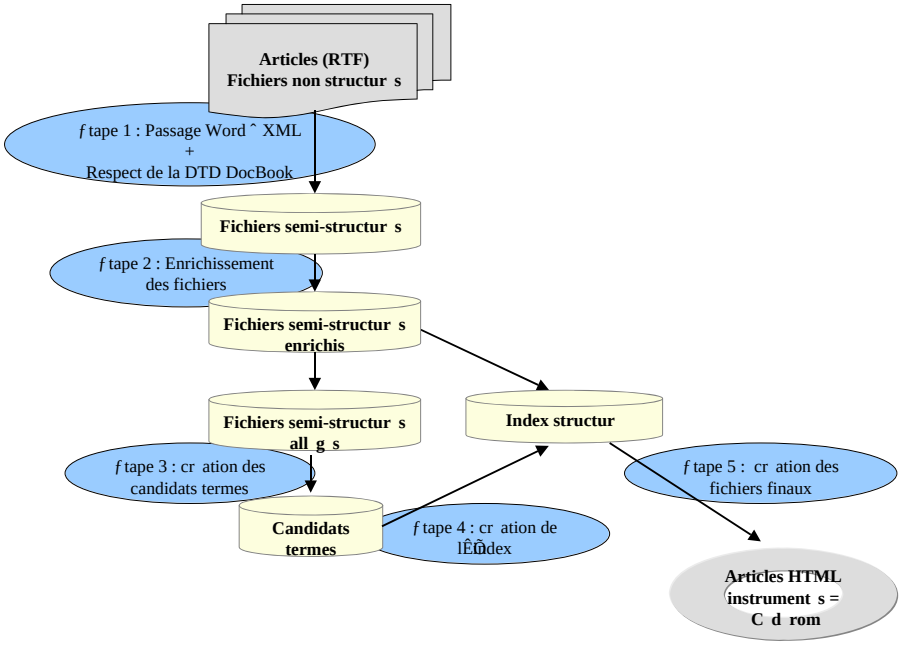

Fig. 1 – *Étapes principales de réalisation du cédérom.*

2. Approche

En nous appuyant sur la démarche proposée dans [AITE 02], nous concevons la construction d'un index comme un processus en deux étapes suivies d'une visualisation :

1. L'acquisition de l'index source qui est une base de connaissances qui contient le réseau des descripteurs accompagnés des renvois au texte.
2. La génération qui produit un ou plusieurs index dérivés (vues) à partir de l'index source en fonction des contraintes éditoriales. La génération exploite ainsi l'ordre de pertinence établi en 1 pour ne sélectionner qu'un certain nombre de descripteurs et de renvois par descripteur. Elle peut également ne conserver qu'un sous-ensemble de relations sémantiques.

Cette génération se fait sur la base de feuilles de styles produites par l'indexeur ou de modèles prédéfinis.

3. La visualisation finale est produite. Même si on peut représenter un index sous différentes formes dans un contexte numérique, dans le cadre de cette expérimentation, nous avons choisi de ne mettre en œuvre qu'un index textuel pour éviter, dans un premier temps, les problèmes de visualisation graphique.

On distingue deux grandes catégories de logiciels d'aide à la création d'index de fin de livres :

1. Les logiciels de gestion d'index que l'on trouve aujourd'hui dans les traitements de texte grand public et qui demandent à l'indexeur de saisir l'intégralité des entrées d'index et des renvois aux textes. Celui-ci le fait généralement à partir d'une lecture sur épreuve du document ou, plus rarement, à l'écran. Ces logiciels prennent donc en charge le tri alphabétique (qui est parfois complexe lorsqu'il est croisé avec les niveaux d'index) et la mise en forme du document « index » (format de sortie et feuille de style).
2. Les logiciels d'acquisition d'index qui proposent un premier jeu d'entrées d'index et de renvois au texte. Ces logiciels vont plus loin dans l'accompagnement de l'utilisateur. Pour ce faire, ils s'appuient sur la structure du document (*Cf.* HTML INDEXER[1], IXGEN[2]). Ils reposent parfois sur une analyse linguistique pour l'extraction de groupes nominaux (*Cf.* SYNTACTICA[3], INDEXING ONLINE[4]) mais la recherche d'occurrences pour le calcul des renvois se limite à une recherche de chaînes de caractères et ne tient pas compte de critères linguistiques.

En dehors des distinctions de casse, ces logiciels ne prennent pas en compte la variation: deux formes fléchies différentes rattachées à un même lemme sont proposées comme deux entrées différentes. Ils n'aident en rien l'indexeur à sélectionner ce qui doit figurer dans l'index. Enfin, ils limitent les entrées d'index aux seuls syntagmes nominaux. Dans ce travail, nous proposons une approche globale tenant compte de ces déficiences et nous proposons des évolutions.

3. De RTF au document semi-structuré instrumenté

Le travail consiste donc, dans un premier temps, à créer une chaîne permettant la transformation de documents RTF en document HTML avec un travail d'enrichissement sur un format intermédiaire, dit « pivot ». Deux choix s'imposaient d'eux-mêmes : *(a)* Puisque nous devions construire une ressource générique, suffisamment structurée, destinée à s'enrichir et à être analysée par les outils du projet, XML était le candidat idéal d'autant plus que de très nombreux programmes prêts à interpréter et transformer des fichiers XML existent ; *(b)* la DTD DocBook[5],

[1] http://www.html-indexer.com/
[2] http://www.fsatools.com/
[3] http://www.syntactica.com/login/login1.htm
[4] http://www.indexingonline.com/index.php
[5] http://www.oasis-open.org/docbook/documentation/reference/html/docbook.html

standard de l'édition numérique, était le parfait support des enrichissements prévus[6]. À ce stade de notre travail, nous n'avons pas exploité les TOPIC MAPS, norme destinée, entre autres, à représenter des index et fondée sur XML. Il semble cependant que ce soit un bon candidat pour la représentation des index numériques et nous envisageons de l'adopter à l'avenir.

Le premier défi consiste à tirer des informations de structure d'un document RTF qui ne possède aucune méta-information. La solution retenue exploite « les styles » typographiques de Word. Dans un document HTML, les balises indiquent le statut de l'élément auquel elle se rapporte. Des règles éditoriales peuvent être attachées à chaque type de balise. De même qu'une balise HTML définit certains cas l'apparence de l'élément auquel elle se rapporte (*i.e.* l'utilisation d'une balise <H1> implique que le contenu sera en gras à l'écran avec une police bien plus grande que le reste du texte). Dans notre cas, c'est le style du texte qui va donner du sens au texte (*i.e.* un titre de niveau 1 – en général, appelé Titre 1 –, nous permet de délimiter une frontière entre deux sections de niveau 1). Pour faire cela, après étude des propriétés de différents logiciels, nous avons choisi la version libre d'un logiciel de transformation de RTF, UPCAST[7], qui construit des fichiers respectant la DTD UPCAST. Ces mêmes fichiers XML ont une structure proche de ceux respectant la DTD DOCBOOK et peuvent donc être traduit pour respecter cette nouvelle grammaire via un programme XSL univoque[8].

La suite du travail a consisté à enrichir la représentation ainsi créée pour permettre de construire des fichiers HTML finaux intrumentés (tables de matières par fichiers, liens entre le texte et la bibliographie, repérage des auteurs, institutions, courriel, repérage des figures comme fichiers externes, etc.) à l'aide de modules logiciels écrits en PERL ou XSL. En parallèle, des fichiers « allégés » sont créés[9] pour être fournis en entrée de SYNTEX (*Cf.* figure 2).

4. Créer des candidats termes avec SYNTEX

SYNTEX [BOUR 00] est un analyseur syntaxique de corpus. Il existe actuellement deux versions, pour le français et pour l'anglais, qui ont été utilisées dans plusieurs projets [BOUR 04]. Le résultat de l'analyse effectuée par SYNTEX est un réseau de mots et de syntagmes : un syntagme verbal (resp. nominal, adjectival) est un groupe de mots dont la tête syntaxique est un verbe (resp. nom, adjectif). Par exemple, *révéler une lésion osseuse* est un syntagme verbal dont la tête syntaxique est le verbe *révéler* et l'expansion le syntagme nominal *lésion osseuse*. Dans le

[6] Une autre DTD sert de standard pour les documents numériques, c'est la DTD TEI de la *Text Electronic Initiative* mais elle correspond à des textes plus littéraires, au contraire de la DTD DocBook utilisée pour des documentations techniques (LINUX) et comportant d'origine des éléments de description des index.

[7] http://www.

[8] L'efficacité de cette transformation suppose que les auteurs aient respecté au mieux les indications éditoriales de styles et la syntaxe de tous les éléments pertinents pour le traitement comme, par exemple, les appels des références bibliographiques. Ce problème sera rediscuté en conclusion (*Cf.* § 8).

[9] Il s'agit de texte brut avec seulement un identifiant unique associé à chaque paragraphe permettant un repérage univoque de chaque partie du texte.

domaine du livre sur l'Ingénierie des connaissances, *modèle conceptuel de l'application* a pour tête syntaxique *modèle conceptuel* et pour expansion le nom *application*. Dans le réseau construit, dit « réseau terminologique », chaque syntagme est relié d'une part à sa tête (lien T) et d'autre part à ses expansions (lien E – *Cf.* figure 3). Les éléments du réseau (mots et syntagmes) sont appelés « candidats termes ». À chaque candidat terme est associé un certain nombre d'informations numériques, sur lesquelles l'utilisateur peut se baser pour organiser son dépouillement :

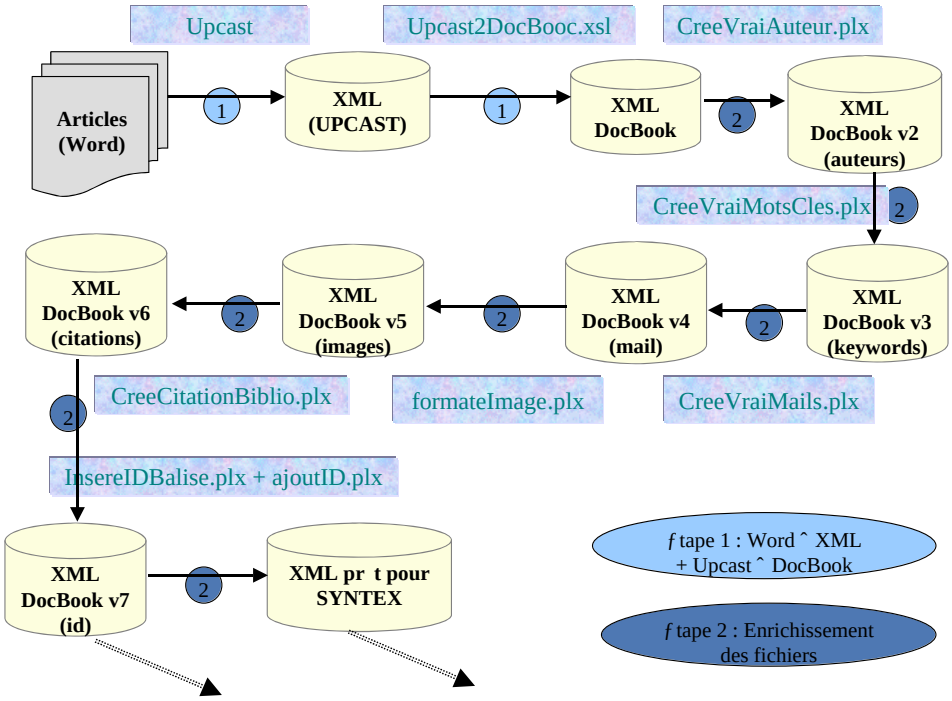

Fig. 2 – De Word/RTF à des fichiers XML enrichis – Étapes 2 et 3.

– **Fréquence** : c'est le nombre d'occurrences du candidat terme détectées par le logiciel dans le corpus. L'interface d'analyse des résultats permet à l'analyste d'accéder à l'ensemble des contextes d'apparition du candidat terme dans le corpus. Dans le travail exposé ici, cet accès, crucial, est assuré par le système INDDOC.

– **Productivité en tête** (resp. expansion) : c'est le nombre de « descendants en tête » (resp. « descendants en expansion ») du candidat terme, c'est-à-dire le nombre de candidats termes plus complexes qui ont le candidat terme en position tête (resp. expansion). À partir de ces informations, l'analyste peut visualiser des listes paradigmatiques de candidats termes partageant la même tête ou la même expansion (*Cf.* figure 3), ce qui le guide vers la constitution de taxinomies locales. La difficulté essentielle pour l'utilisateur vient de la masse des résultats issus de l'extraction. Même s'il existe de nombreux travaux fort intéressants sur le filtrage statistique de candidats termes extraits automatiquement de corpus [DAIL 94; MAYN 01; NAKA 01], l'expérience montre qu'aucune mesure statistique ne peut suppléer l'expertise de

l'analyste, en particulier parce qu'il y a toujours des candidats termes de fréquence 1 dont l'analyse est intéressante. De façon générale, sachant qu'il ne pourra analyser tous les candidats termes extraits du corpus, l'analyste doit adopter une stratégie optimale qui, étant donné le temps qu'il consacre à l'analyse terminologique et le type de la ressource à construire, lui garantit que, parmi les candidats qui auront échappé à son analyse, la proportion de ceux qui auraient pu être pertinents est faible. Ce constat demande, dans le cas de notre projet, que l'étape suivante, prise en charge par INDDOC, tienne compte de ces particularités (*Cf.* § 5).

modèle conceptuel
Tête de :
modèle conceptuel de l'application
modèle conceptuel des données
modèle conceptuel des traitements
Expansion de :
construction d'un modèle conceptuel
validation d'un modèle conceptuel

Fig. 3 - *Extrait du réseau terminologique construit par SYNTEX autour du candidat terme modèle conceptuel.*

5. Créer un index avec INDDOC

On élabore donc l'index à partir de la liste des candidats termes et du réseau syntaxique produits par Syntex. Pour cela, il faut :

1. Structurer cette liste en réseau sémantique : nous nous appuyons sur les liens tête-expansion produits par Syntex, mais il faut interpréter et enrichir ce réseau, *a minima* en introduisant des relations hiérarchiques entre les futures entrées et sous-entrées et des liens d'équivalence sémantique pour établir des références croisées d'une entrée à l'autre ;
2. Calculer les renvois au texte pour que l'index permette effectivement d'accéder aux passages pertinents, qui sont de taille variable ;
3. Introduire une mesure de pertinence pour permettre le filtrage et le tri des informations.

L'acquisition de l'index source procède en deux étapes :

– La création automatique d'une ébauche d'index. Cette étape part du texte et produit un index source. Elle permet de construire le contenu de l'index (la liste structurée des descripteurs et les renvois au texte). Elle repose sur des techniques de structuration de terminologie pour construire le réseau, sur des techniques de segmentation de texte pour établir les renvois au texte et sur des mesures de pertinence pour trier et sélectionner l'information dans l'index. Nous détaillons les différentes étapes de ce processus ci-après
– La validation de l'ébauche d'index. Cette étape interactive repose sur une interface qui permet à l'auteur de l'index de visualiser l'ébauche, de la modifier et de l'enrichir. L'index peut-être visualisé sous différentes formes. L'interface de valisation permet d'ajouter, de supprimer une fiche, ou de modifier un descripteur ou un renvoi. . En cas de doute, l'indexeur

peut, à tout instant, consulter les segments de textes associés à une entrée. Le travail de validation reste coûteux mais l'interface permet de l'organiser et d'en assurer la cohérence.

Dans ce qui suit, nous ne présentons pas l'interface de validation [AITEL02] mais nous décrivons le processus qui permet de construire automatiquement un index aussi « bon » que possible. La construction de l'ébauche d'index consiste notamment à filtrer les candidats termes, à les organiser en réseau, à calculer leurs occurrences et à mesurer le poids des termes et de leurs occurrences (*Cf.* figure 4).

Le filtrage

Dans un premier temps, la liste des candidats termes produite par SYNTEX est filtrée. Seuls les termes nominaux sont conservés comme entrées potentielles (les termes verbaux peuvent être pris en compte mais seulement comme variantes d'un terme nominal). L'application d'un antidictionnaire permet d'éliminer une partie des termes non pertinents pour former des entrées d'index. L'application de règles approximatives de racinisation permet de réduire encore la liste initiale en regroupant certains termes.

Mise en forme numéros

La structuration

INDDOC tente de repérer différents types de relations entre les termes : variation morphosyntaxique, synonymie, hyperonymie notamment. Il s'appuie pour ce faire sur les résultats obtenus en structuration de terminologie (FASTER [JACQ 96], PROMETHE [MORI 99], SYNOTERM [HAMO 01]), la principale difficulté vient de ce

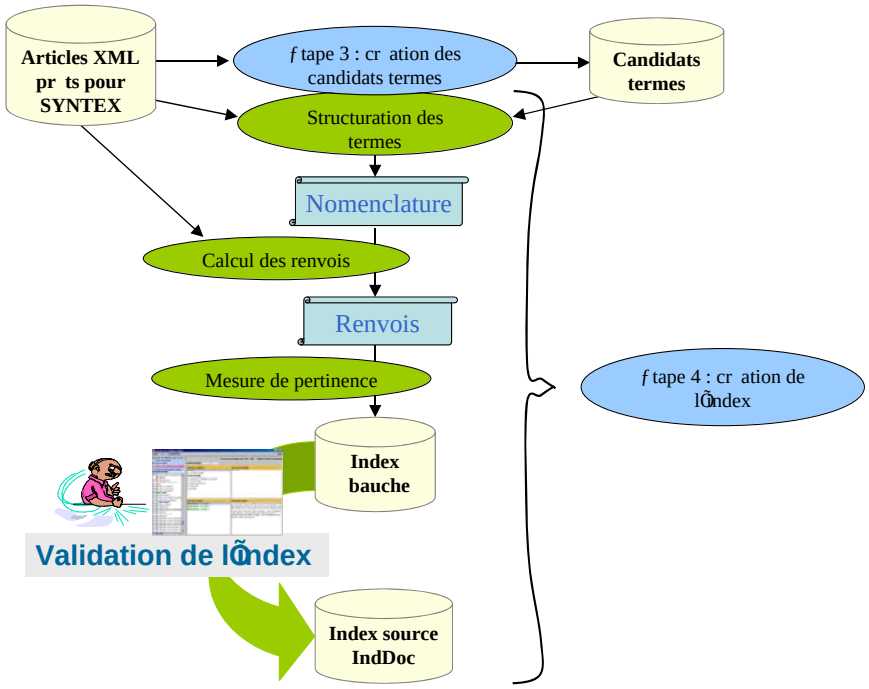

Fig. 4 – *Génération de l'index source – Étapes 3 et 4.*

que les outils existants ont été conçus pour acquérir un type particulier de relation et du fait qu'il semble impossible d'avoir une méthode unique de détection de relations.

En pratique, l'intégration ne doit pas se faire en fonction du type de relation visée mais selon la méthode utilisée. INDDOC comporte ainsi 2 sous-modules de structuration [AITE 03]. Le premier repose sur la structure interne des termes et exploite directement les résultats de FASTER et SYNOTERM. Le second module exploite l'information contextuelle. Pour cela, nous avons développé un module de recherche de relations à base de patrons. Nous avons écrit une base de patrons pour l'hyperonymie et la méronymie qui s'appliquent à tout type de corpus. INDDOC exploite ainsi à la fois les relations tête-expansion et des patrons génériques[10] (ex. « SN1 être SN2 », « SN1 Verbe-Composition SN2 »...) pour établir les relations hiérarchiques[11]. Nous intégrons l'ensemble des liens produits par les différents modules dans un réseau commun. Les résultats sont parfois redondants et nous conservons l'union des ensembles de relations produits par chaque outil.

Mise en forme
numéros

Le calcul des renvois

Le calcul des renvois consiste, pour chaque entrée et chaque sous-entrée, à établir la liste de ses occurrences dans le texte, la difficulté étant de sélectionner les plus pertinentes et de définir la taille de l'empan de texte auquel il est pertinent de renvoyer[12].. Pour identifier les segments de renvoi, nous partons d'une segmentation absolue qui ne dépend que du document. À ce stade, on segmente le document en unités documentaires minimales[13] (UDMs), puis on élargit ces UDMs en unités documentaires élargies (UDEs) en fonction des marqueurs linguistiques et typographiques tout en respectant la structure logique du document (une UDE ne peut pas être à cheval sur deux sections de document). Cette segmentation permet donc de découper le document en UDEs linguistiquement ou typographiquement homogènes. À l'issue de cette phase, le document est représenté comme une liste d'UDEs. Ensuite, nous procédons à une segmentation relative qui dépend d'un descripteur donné. Cette phase est nécessaire pour établir la liste des segments de renvoi (liste des renvois) d'un descripteur. Elle comporte trois étapes : (1) identification des segments de renvoi (les UDEs qui contiennent le descripteur ou une de ses variantes); (2) regroupement des segments de l'étape 1 qui sont adjacents dans le texte du document, ce qui permet d'obtenir une liste simplifiée de segments de renvoi; (3) généralisation de la séquence des segments d'une même section et sous-section en un unique renvoi à la section, lorsqu'une partie suffisamment grande de la section figure dans la liste des segments établie à l'étape 2. Une partie de ce regroupement est également laissée à la charge de l'auteur qui a la possibilité dans une interface de validation de regrouper plusieurs occurrences qui lui semblent

[10] SN pour syntagme nominal.

[11] À l'avenir, on pourrait envisager de laisser la possibilité à l'auteur de l'index d'introduire des patrons spécifiques.

[12] Précisons que les pages ne sont pas les bonnes unités pour un index dédié à un ouvrage numérique.

[13] Un extrait de texte de type paragraphe, phrase ou même de mots. Le type de l'extrait dépend de la tâche à réaliser (indexation, résumé automatique etc.). Dans notre cas, c'est le paragraphe

proches sous un seul renvoi. Il peut ainsi compléter et corriger le regroupement qui est basé sur des heuristiques génériques.

Mise en forme : numéros

La mesure de pertinence

Une fois identifiés les renvois associés à une entrée, il reste à les classer par ordre de pertinence. Nous nous inspirons de l'approche TF/IDF pour l'évaluation de la pertinence de différentes clefs d'indexation dans une base documentaire mais notre mesure [AITE 04] permet de prendre en compte, outre le poids d'un mot dans l'ensemble du document et sa fréquence dans le segment de renvoi, le poids d'une occurrence particulière (qui peut être mise en valeur typographiquement, par exemple) et le poids des segments où il cité (lequel dépend en retour du poids des termes qu'il comporte). Notre critère de pertinence tient ainsi compte des paramètres qui sont traditionnellement exploités par les indexeurs [NANC 93; THEC 03] : la typographie, la présence d'une occurrence dans un titre, une mise en relief discursive... Notre mesure repose ainsi sur différentes marqueurs ce quila rend plus robuste aux variations de genre, de domaine et des style.

Supprimé : ¶

Mise en forme : numéros

6. Instrumenter un fonds documentaire indexé

À ce stade du processus, l'index source est créé au format INDDOC (DTD). Il faut le traduire dans un format respectant la DTD DOCBOOK et faire un certain nombre de traitements pour permettre sa traduction en HTML, en particulier sur la formes des termes dans les articles et des renvois dans l'index devant respecter la syntaxe des ancres HTML. C'est durant cette traduction qu'a lieu la génération de l'index dérivé en fonction des contraintes éditoriales que l'on s'est données : affichage de l'index hiérarchique sur 2 niveaux, conservation du typage de certains liens (« voir aussi », hyperonymie), abandon des autres (projetés dans « voir aussi »).La fin de l'instrumentation est la génération de l'index HTML, la création des articles HTML et la mise en place de l'interface de navigation qui comporte, dans cette première version, 3 « frames », (1) table des matières des articles, (2) articles et (3) index. C'est à ce niveau qu'est visualisé l'empan des renvois. Cette visualisation pose problème : si nous visualisons les renvois d'index dans les textes avec de la couleur, ce que nous faisons, nous pouvons utiliser pour cela des fonctions JAVASCRIPT mais avec un défaut majeur : la non portabilité de ce langage d'un navigateur à l'autre. La solution choisie est donc de générer autant de fichiers HTML coloriés qu'il y a d'entrées d'index pour un article. Cela augmente le volume de l'ensemble des fichiers mais rend l'interface totalement portable (*Cf.* figure 5).

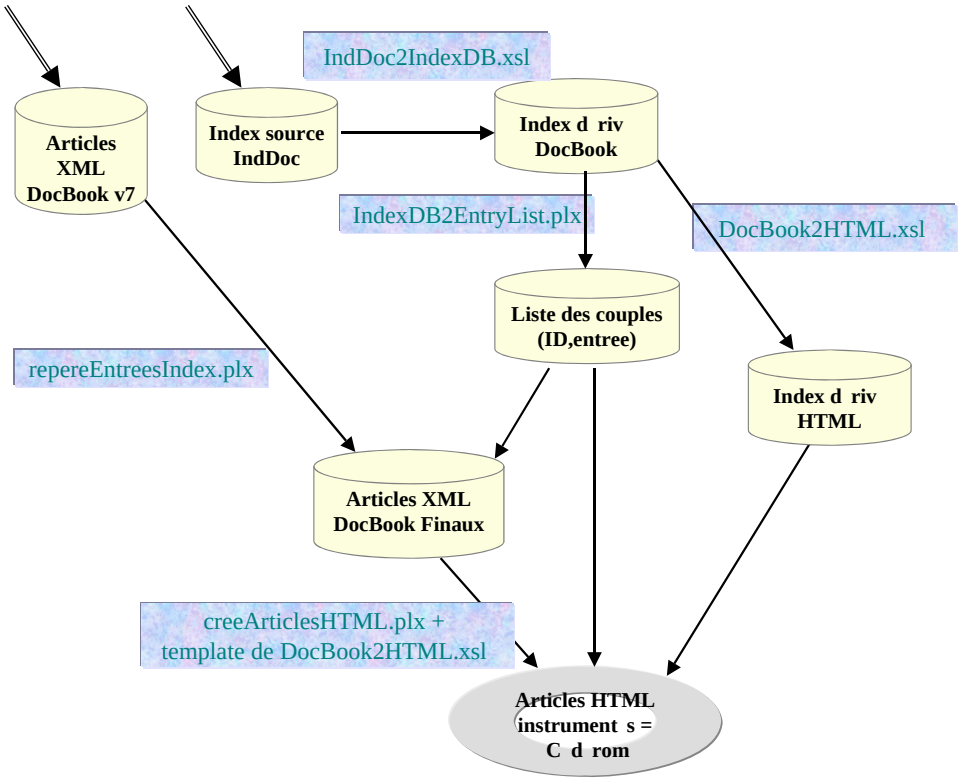

Fig. 5 - Génération des fichiers HTML instrumentés – Étape 5.

7. Retour d'expérience

Mise en forme : numéros

Résultats de IndDoc

Filtrage

Le corpus comporte environ 177 000 mots. SYNTEX fournit une liste de 32 334 candidats termes. De cette liste, nous retenons 17 521 candidats termes nominaux. Le filtrage à base d'antidictionnaire et de racilnisation ramène cette liste à quelques 2 800 candidats termes qui sont autant d'entrées potentielles de l'index.

Structuration

Pour la structuration, nous prenons en compte toutes les variantes de termes. Nous travaillons donc sur la liste complète de 10 008 candidats termes fournie par SYNTEX. À partir de cette liste, le module de structuration de INDDOC calcule 4 440 relations sémantiques. Comme les termes ne sont pas validés au préalable, certaines relations relient en fait des termes non pertinents (par ex. *différents membres : partie différente*, *haut niveau : niveau supérieur*). Cette proportion de relations s'avère relativement faible si on la compare avec des monographies [AITE 04]. Cela peut être dû à la nature de l'ouvrage qui est un recueil d'articles d'auteurs différents, à

forte diversité stylistique.

Calcul des renvois

Nous avons constaté que la segmentation réduit effectivement le nombre de renvois (*Cf.* tableau 1), même si là encore cette réduction est moins forte que sur d'autres corpus. On observe en effet que le facteur de réduction de la segmentation dépend de la nature du document (monographie *vs* collection) et du style de la rédaction (le style littéraire emploie plus de marqueurs linguistiques ; ce qui augmente le facteur de réduction dans le passage des UDMs aux UDEs). L'intérêt global de cette étape apparaît quand on compare le nombre de renvois obtenus (2 997) est avec le nombre d'occurrences des différentes entrées de l'index (28 342)[14] .

Par comparaison avec d'autres corpus étudiés, on observe aussi (*Cf.* tableau 1) que les étapes de simplification et de généralisation sont moins marquées dans notre corpus dont les articles sont généralement fortement structurés.

Tab. I - *Nombre d'UD et de segments aux différentes étapes.*

Nb d'UDMs	1 085	Segmentation absolue
Nb d'UDEs	907	
Nb de seg. De renvois	3 097	Segmentation relative
Nb de seg. après simplification	3 008	
Nb de seg. après génération	2 997	
Nb de UDMs occurrences	28 342	

Mesure de pertinence

Nous avons appliqué la mesure de pertinence de INDDOC sur notre corpus. Il est difficile d'évaluer cette mesure en tant que telle et globalement le tri obtenu mais on peut observer sur des exemples le bon comportement de notre mesure de pertinence.

Considérons le descripteur « Modélisation » qui a 4 renvois : le premier dans l'ordre du corpus, S1, apparaît dans une introduction. Le deuxième segment S2 regroupe une sous-section qui traite de la « modélisation ». Dans le troisième segment S3, le descripteur apparaît en début de segment mais le segment lui-même est inclus dans une conclusion. Le troisième segment S4 correspond à une section qui traite de la « modélisation ».

Le système a ordonné les renvois en privilégiant S4 pour la quantité de l'information apportée, devant S2 dont l'apport d'information est plus faible. Le renvoi S1 est placé en dernière position parce qu'il s'agit d'un paragraphe de l'introduction. S3 apparaît dans une conclusion mais il apporte davantage d'information et le descripteur apparaît au début du segment, ce qui lui confère plus d'importance.

[14] Il s'agit en réalité du nombre d'UDMs occurrences, *i. e.* de la somme, pour chaque entrée, du nombre de paragraphes dans lesquels elle figure (un paragraphe donné peut donc figurer deux fois, associés à deux entrées différentes), résultat qu'on obtiendrait par un calcul naïf des renvois.

Validation de l'index

La validation de l'index vient de s'effectuer avec le système décrit. À partir du corpus disponible, l'interface IndDoc nous a proposé 2700 termes comme possibles entrées d'index. Ces termes étant à valider, structurer en niveaux et à rattacher au corpus. Nous avons retenu un peu moins de 1000 termes comme entrées d'index et environ le double de liens vers les textes. Avant de discuter plus avant des critiques et questions en suspens, on peut noter que :

- À l'inverse des constitutions d'index sur papier, et cette remarque est aussi valable pour le travail précédent, le choix des entrées d'index se fait par suppression au sein d'une liste « large », à l'inverse d'un travail standard, repérant les entrées d'index dans n texte en partant de zéro. Cela amène à la constitution d'un index très riche.
- Vis-à-vis de cet index très riche, l'usage via une interface Web doit être observé et évalué, peu d'expériences ayant été faites dans ce domaine, sauf pour des documentations techniques. On peut penser, mais ce doit être validé, que la richesse de l'index est compensée par la facilité de navigation.

8. Conclusion et perspectives

Revenons sur les points critiques de l'expérience précédente en 1999 [BOUR 99] :

– **L'empan d'un renvoi**. C'est l'un des aspects originaux ; la question de la prise en charge est prise en compte dès le départ dans le système et l'interface et aide l'indexeur à choisir cet empan. Par ailleurs, les mesures de pertinence du système INDDOC proposent les renvois dans un ordre, justement pertinent, pour prendre en compte cette question.

– **La difficulté de choisir ce qui est une bonne entrée d'index**. La méthodologie et le système INDDOC en particulier sont une réponse partielle à cette difficulté mais cette question recevra toujours une réponse en termes de choix humain.

– **La structuration de l'index**. Le système INDDOC et la méthodologie mise en œuvre ici est beaucoup plus riche que précédemment : le système et l'interface permettent en plus de typer les relations entre les entrées de l'index, beaucoup plus même que ce que nous avons choisi de faire (*Cf.* infra).

– **La validation.** Rien n'a changé. Puisque le choix des entrées d'index est un choix de l'indexeur alors que les articles sont écrits par d'autres, on pourrait faire valider ce choix avec les auteurs des articles. Cette procédure, testée précédemment, n'a pas été mise en œuvre ici. En revanche, comme dans l'expérience précédente, l'indexation se fait à 2 pour éviter l'idiosyncrasie d'un travail solitaire.

Pour le reste, deux enseignements peuvent déjà être tirés de ce travail :

– Construire des fichiers XML semi-structurés à partir de Word est un travail sans fin : venant d'un éditeur WYSIWYG utilisant de nombreux caractères spéciaux pour la mise en page ou la possibilité de générer des figures en interne, nous

avons été obligés à des corrections incessantes. Ainsi, exemple parmi tant d'autres, un caractère qui semble être le « caractère blanc » peut être un « blanc collant » ou un « blanc italique ». Pour les programmes qui enrichissent les fichiers XML, cela veut dire l'obligation de prendre en charge de nouvelles contraintes d'expression à chaque nouvel article ou nouveau morceau d'article. Sur la question des figures, nous avons été obligés de redéfinir (ce qui a amené à refaire des figures mais c'est un choix qui avait été fait de toute façon au départ) toutes les figures comme des fichiers externes, Sinon il était impossible de faire des fichiers XML cohérents avec les articles « papier » et affichant correctement les figures. Enfin, un éditeur comme celui précité, ne pousse pas les auteurs à respecter les styles que nous leur demandions d'utiliser très précisément pour pouvoir créer des fichiers semi-structurés avec UPCAST (*Cf.* § 3). Il en aurait été autrement à partir d'un formateur de texte de type LATEX.

– Les travaux des auteurs de INDDOC permettaient de générer des index avec un accès graphique beaucoup plus riche que l'index textuel que nous avons décidé de construire de prime abord. On pouvait par exemple. envisager trois vues différentes qui correspondent à trois stratégies de recherche différentes : (1) la recherche par descripteur qui privilégie l'exhaustivité et la précision de l'information en permettant de visualiser l'ensemble des informations qui se rattachent à un descripteur sous la forme d'une étoile ; (2) la recherche par réseau qui donne une vue globale sur la nomenclature de l'index ; et (3) la recherche thématique qui permet d'accéder à un terme puis à l'ensemble des descripteurs qui relèvent de ce thème. Ce mode de structuration de l'information se rapproche des approches thésauriques et des pratiques lexicographiques anglosaxones.

Par rapport à cette dernière recherche thématique, il faut noter que, comme nous le remarquons dans [BOUR 99], nous avons fait le choix de construire un index qui rend compte des usages des auteurs, à l'inverse des index thématiques. Ces derniers correspondent à des ressource beaucoup plus normalisées, qui peuvent aller jusqu'à des ontologies [CHAR 02]. Enfin, en construisant l'index source, nous nous sommes réservés la possibilité de construire un index dérivé plus riche en répertoriant plus de relations que celles utilisées pour le présent travail.

9. Références bibliographiques

[AITE 02] Ait El Mekki T., Nazarenko A., « Comment aider un auteur à construire l'index d'ouvrage ? » *Colloque International sur la Fouille de Texte*, Tunis, 2002, p. 141- 157.

[AITE 03] Ait El Mekki T., Nazarenko A., « Le réseau terminologique, un élément central pour les index de fin de livre » *actes des cinquièmes rencontres Terminologie et Intelligence*, Strasbourg, 2003, p. 1-10.

[AITE 04] Ait El Mekki T., Nazarenko A., « Une mesure de pertinence pour le tri de l'information dans un index de fin de livre » TALN04, Fès, 2004 (Soumis)

[NANC 93] Nancy C. Mulvany, « *Indexing Books* » The University of Chicago Press, 1993.

[BOUR 99] Bourigault D., Charlet J. Construction d'un index thématique de l'Ingénierie des connaissances. Actes de la conférence IC'99, Massy-Palaiseau/Polytechnique, 1999.

[BOUR 00] Bourigault D., Fabre C., « Approche linguistique pour l'analyse syntaxique de corpus », Cahiers de Grammaires. Univ. Toulouse - Le Mirail, n° 25, 2000, p. 131-151.

[BOUR 04] Bourigault D., Aussenac-Gilles N., Charlet J. Construction de ressources terminologiques ou ontologiques àpartir de textes : un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artifcielle*, 2004.

[CHAR 00] Charlet J., Zacklad M., Kassel G. & Bourigault D. [2000] (éd.), Ingénierie des connaissances. Évolutions récentes et nouveaux défis, « Coll. technique et scientifique des télécommunications », Eyrolles, Paris, 632 p.

[CHAR 02] Charlet J., *L'ingénierie des connaissances : développements, résultats et perspectives pour la gestion des connaissances médicales*. Mémoire d'habilitation à diriger des recherches. Paris VI, décembre 2002.

[DAIL 94] Daille B., *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*, thèse en Informatique Fondamentale, Univ. de Paris 7, Paris, 1994.

[JACQ 96] Jacquemin, C., « A symbolic and surgical acquisition of terms through variation. » ,in S. Wermter, E. Riloff, and G. Scheler, ed., Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing, Springer, Heidelberg, 1996.

[HAMO 01] Hamon T., Nazarenko A., « Detection of synonymy links between terms: experiment and results », in Bourigault D, L'Homme M.-C., Jacquemin C. (éd.), *Recent advances in computational terminology*, John Benjamins Publishing Company, Amsterdam, 2001 p. 185-208.

[MAYN 01] Maynard D., Ananiadou S., Term extraction using similarity-based approach, in Bourigault D, L'Homme M.-C., Jacquemin C. (éd.), *Recent advances in computational terminology*, John Benjamins Publishing Company, Amsterdam, 2001, p. 261-78.

[MORI 99] Morin E., « Des patrons lexico-syntaxiques pour aider au dépouillement terminologique », *Traitement Automatique des Langues*, 1999, 40(1): 143-166.

[NAKA 01] Nakagawa H., Experimental evaluation of ranking and selection methods in term extraction, in Bourigault D, L'Homme M.-C., Jacquemin C. (éd.), *Recent advances in computational terminology*, John Benjamins Publishing Company, 2001, p. 303-26.

[TEUL 04] Teulier R.,Charlet J., Tchounikine P. Ingénierie des connaissances, L'harmattan, Paris, 2004. À paraître.

[THEC 03] *The chicago Manuel of Style*, chapter 18, fifteenth Edition, *T*he University of Chicago Press Staff, 1993.