Étude de faisabilité de mise en place d'une indexation semi-automatique avec un thésaurus spécialisé en archéologie

Anita Mazur

Anita MAZUR

Master DEFI

(Documents électroniques et flux d'informations)

RAPPORT de STAGE

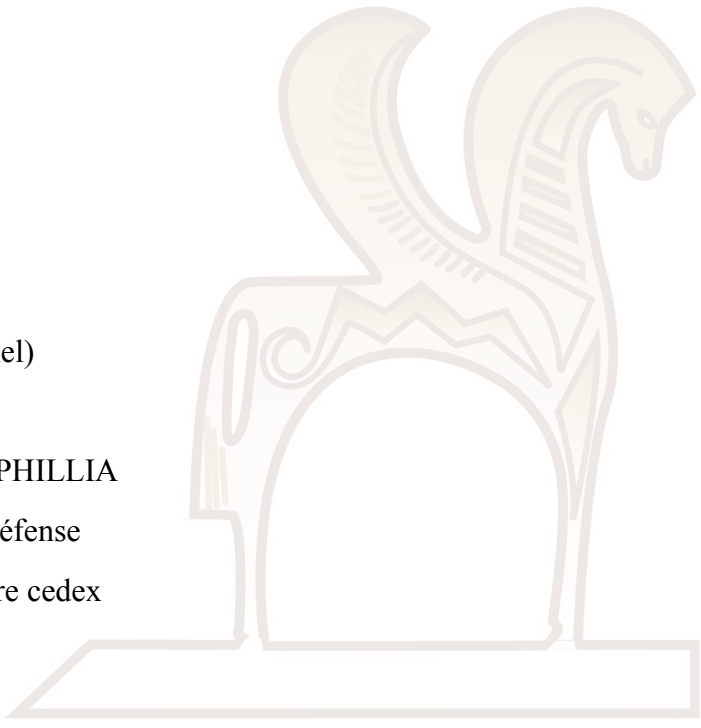
stage effectué du 02/04/2012 au 02/07/2012

à

Maison René Ginouvès

Nanterre

Étude de faisabilité de mise en place d'une indexation semi-automatique avec thésaurus



Sous la direction de :

Mme B. LEQUEUX (tuteur professionnel)

Soutenu le 24 Septembre 2012 à l'UFR PHILLIA

Université Paris Ouest - Nanterre – la Défense

200 av. de la République, 92001 Nanterre cedex

Année Universitaire 2011-2012

Remerciements

Je tiens tout particulièrement à remercier ma tutrice de stage Brigitte Lequeux, ainsi que les documentalistes Bénédicte MacGregor et Nathalie Le Tellier Becquart pour l'accueil et tous les conseils qu'elles m'ont apportés durant le stage.

Je remercie également le personnel de la bibliothèque d'archéologie, du pôle éditorial de la MAE ainsi que les enseignants du Master 2 DEFI.

Sommaire

• Introduction

Grâce à Internet, le nombre de ressources électroniques scientifiques ne cesse de s'accroître. Les revues exclusivement électroniques comme les revues papier proposent leurs articles sous forme de documents électroniques. Le réseau de bibliothèques et de laboratoires FRANTIQ[1] a entre autres, comme but de fédérer les documents d'archéologie et de sciences de l'Antiquité (papier ou électronique) dans un catalogue commun indexé afin de les retrouver plus facilement. Chaque document fait l'objet d'une notice et d'une indexation à l'aide du thésaurus spécialisé en archéologie PACTOLS[2]

Or, c'est une tâche très coûteuse en temps/homme. Alors qu'en archéologie, les informations sont cumulatives, c'est-à-dire que les informations ne deviennent jamais obsolètes mais s'ajoutent au fil du temps et les documents sont des ressources importantes pour le travail des chercheurs et des étudiants.

En 2009, pour alléger le travail d'indexation des documentalistes et bibliothécaires et étendre la couverture du catalogue, les premiers essais d'indexation semi-automatique à l'aide du thésaurus PACTOLS ont été réalisés par l'informaticien du réseau FRANTIQ aidé de stagiaires.

Le résultat n'étant pas très concluant, notamment à cause d'ambiguïtés dans les mots-clés proposés par l'algorithme, et par manque de temps et de moyens pour améliorer les résultats, ce projet a été temporairement abandonné. Des cours de programmation et de Traitement Automatique de Langue (TAL) étant dispensés dans le cursus du Master 2 DEFI[3], on m'a proposé de reprendre ce projet durant mon stage.

Ce mémoire présentera d'abord le contexte de la mission du stage, l'attente et les besoins en terme d'indexation automatique et le prototype existant. En second lieu sera proposé l'état de l'art de l'indexation automatique et semi-automatique. Enfin seront analysés les premiers essais d'indexation automatique réalisés durant ce stage ainsi que la réception du pré-prototype par les potentiels utilisateurs.

1 Fédération et ressources sur l'Antiquité (Frantiq) - http://frantiq.fr/
2 Thésaurus PACTOLS : Il s'agit d'un thésaurus multilingue spécialisé dans l'archéologie
3 Master 2 Documents électroniques et flux d'informations (DEFI) à l'université Paris Ouest Nanterre La Défense (92)

1. La Maison René-Ginouvès et FRANTIQ

Établissement de recherche et de formation à la recherche sous la triple tutelle du Centre national de recherche scientifique (CNRS), de l'université Paris I Panthéon-Sorbonne et de l'université Paris Ouest Nanterre La Défense, la Maison d'Archéologie et ethnologie René-Ginouvès (MAE) est une Maison des sciences de l'Homme regroupant des Unités de recherche dans ces domaines (plus de 300 chercheurs) et une Unité de services et de recherche mutualisant les équipements et les outils de travail communs.[4]

Parmi ceux-ci, un pôle éditorial publie plusieurs revues scientifiques de haut niveau dans les domaines de l'ethnologie, de l'archéologie et de la sociologie. La langue utilisée dans les revues est principalement le français.

Un pôle documentaire regroupe le service des Archives scientifiques, deux bibliothèques, le service Photo, et le Service d'ingénierie documentaire et réseaux pour l'archéologie (SIDRA). Composé de plusieurs documentalistes, et ponctuellement de stagiaires et de CDD, il gère FRANTIQ, réseau national de laboratoires et de bibliothèques du Centre national de recherche scientifique (CNRS) spécialisés en archéologie et sciences de l'Antiquité, de la Préhistoire au Moyen Âge. Nombreux outils et services sont proposés aux partenaires de ce réseau et au public (étudiants, enseignants-chercheurs). Citons par exemple, le catalogue collectif indexé (CCI) commun à toutes les bibliothèques membres, le thésaurus spécialisé multilingue PACTOLS[5], créé par la directrice du service Mme Lequeux, le gestionnaire de thésaurus OpenTheso et un site web mettant à disposition toutes ces ressources qui ont pour but de faciliter l'accès du public aux documents scientifiques que le réseau possède ou produit.

Les outils mis en place pour le réseau FRANTIQ s'insèrent dans le processus de la « chaîne documentaire ». Nous allons voir en premier lieu une définition de la chaîne documentaire, puis une présentation du corpus, des outils et de la méthodologie d'indexation manuelle par les bibliothécaires et documentalistes du réseau FRANTIQ.

4 MAISON RENÉ GINOUVÈS. Histoire de la Maison [en ligne]. Disponible sur : <http://www.mae.u-paris10.fr/usr3225/Histoire-de-la-Maison.html> (consulté le 28.06.2012)

5 Acronyme de Peuples, Anthroponymes, Chronologie, Toponymes, œuvres, Lieux et Sujets

2. La chaîne documentaire

Il s'agit d'une notion issue de la bibliothéconomie et désigne l'« ensemble des opérations successives de sélection/collecte, de traitement, de mise en mémoire et de stockage, et de diffusion de documents et d'information (Vocabulaire de la documentation, 2004) »[6].

Les trois grandes phases de la « chaîne documentaire » sont[7] :

- Entrée = acquisitions, collecte des documents
- Traitement = traitement matériel et traitement intellectuel (catalogage, analyse et indexation)
- Sortie = Services (prêt, reproduction), produits élémentaires et élaborés (recherche documentaire, dossiers...)

Ce processus pourra, selon Lamouroux, être appelé « chaîne de valeur du document » car le but du processus de « chaîne documentaire » est de mettre en valeur l'information et d'améliorer l'accès aux informations contenues dans les documents.

Au niveau de FRANTIQ, toutes les étapes sont prises en comptes. Intéressons-nous donc à l'entrée, au traitement et à la sortie du réseau FRANTIQ.

2.1. Entrée

a) Définition

Ce sont les bibliothèques qui font les acquisitions ou qui reçoivent des documents de différents types. Aujourd'hui, les bibliothèques prennent aussi en compte les documents au format numérique, qu'ils soient disponibles sur des CD/DVDs ou sur Internet. Afin de permettre aux usagers de la bibliothèque de retrouver facilement un document, les bibliothécaires procèdent souvent au traitement matériel et intellectuel.

Nous allons voir maintenant le type de corpus disponible dans la bibliothèque du réseau FRANTIQ.

6 LAMOUROUX Mireille. La chaîne documentaire. IN : Semaine de l'Inspection générale, 26 janvier 2011, 2011 ESEN, Poitiers Disponible sur : <http://www.cddp92.ac-versailles.fr/spip2/texteanim/La %20chaine_documentaire_Semaine_IG_ESEN.pdf> (consulté le 11.08.2012)

7 Ibid.

b) Corpus du réseau FRANTIQ

Les bibliothèques du réseau FRANTIQ possèdent des ressources variées, aussi bien des documents imprimés que numériques. Tous ont la particularité d'être des documents scientifiques aussi bien anciens que « up-to-date » puisque chaque producteur de FRANTIQ « est dans l'obligation de dépouiller la littérature que son laboratoire ou équipe produit »[8].

Voici une liste des types de documents[9] :

- ouvrages (monographies, actes de colloque ou tout autre réunion scientifique, exposition, catalogue de musée, etc.)
- extraits (articles de périodiques, contribution à des ouvrages collectifs : colloque, mélanges, etc., tirés-à-part)
- documents numériques (sur support matériel ou en ligne)
- documents cartographiques
- périodiques et collections
- la littérature grise (travaux universitaires et de rapports de fouilles ou d'activités)

La plupart des ressources sont donc des documents textuels. Ils sont soit disponibles physiquement dans les bibliothèques du réseau, soit disponibles au format numérique sur Internet ou des CD-ROMs/DVDs. Les textes nativement sous format numérique sont plus susceptibles d'être disponibles en texte intégral sur Internet[10].

En ce qui concerne la langue des documents, il n'y a pas seulement le français, mais aussi l'anglais, l'allemand, etc.

Pour ces documents seront ensuite signalés à l'aide du catalogue, des notices bibliographiques avec mots-clés pour permettre au public de retrouver les documents en faisant des recherches.

8 http://www.frantiq.fr/fr/CCI/TypeDocument

9 Ibid.

10 Par exemple, le portail de revues en sciences humaines et sociales Persée héberge des volumes de Gallia et Gallia Préhistoire au format PDF. Le portail est disponible sous l'adresse suivante : http://www.persee.fr

2.2. Traitement matériel et intellectuel

Le traitement matériel de base « commence par la réalisation d'un inventaire si celui-ci est inexistant et ensuite par la définition d'un système de classement et des conditions de rangement »[11].

Le traitement intellectuel comporte le catalogage et l'indexation.

a) Définition du catalogage

Selon la Bibliothèque nationale de France (BnF)[12], le catalogage consiste à créer des notices bibliographiques avec une description normalisée et complète du document.

b) Catalogage au sein du réseau FRANTIQ

Il existe plusieurs formats pour présenter une notice de façon normalisée, comme USMARC, UNIMARC, DUBLIN CORE, etc. Par souci d'interopérabilité avec les autres bibliothèques membres, le format UNIMARC (une variante européenne du format MARC) a été choisi. Ce format se fonde sur la normalisation internationale de l'échange de l'information bibliographique informatisée[13] et évolue régulièrement selon les normes internationales en vigueur. Créé initialement pour le catalogage des monographies et publications en séries imprimées, il s'est élargi à tous les types de ressources susceptibles d'exister dans les bibliothèques[14]. Ce format a donc une multitude de possibilités pour décrire les documents. On le voit aussi à travers les multiples champs à disposition lorsque l'on souhaite cataloguer un document. Dans KOHA, les champs classiques UNIMARC sont matérialisés par 7 pages thématiques[15] au plus. Cela dépend du type de document que l'on souhaite cataloguer. Une notice pour un ouvrage comportera d'autres champs qu'une notice de revues ou une ressource électronique.

11 POMMARET Sabine. Traitement documentaire et valorisation des fonds iconographiques anciens dans les bibliothèques : l'exemple de la collection d'estampes de la B.M. De Bourges [en ligne] . Diplôme de conservateur de bibliothèque. Villeurbanne : ENSSIB, 2002, 83 p. Disponible sur : <http://www.enssib.fr/bibliotheque-numerique/document-1032> (consulté le 25/08/2012).

12 BIBLIOTHÈQUE NATIONALE DE FRANCE (BnF). Politique de catalogage [en ligne]. Disponible sur : <http://www.bnf.fr/fr/professionnels/anx_catalogage_indexation/a.politique_catalogage.html> (consulté le 25/08/2012).

13 BIBLIOTHÈQUE NATIONALE DE FRANCE (BnF). Format UNIMARC [en ligne]. Disponible sur : <http://www.bnf.fr/fr/professionnels/f_um/s.format_unimarc_notices_bibliographie.html> (consulté le 25/04/2011).

14 BIBLIOTHÈQUE NATIONALE DE FRANCE (BnF). Format UNIMARC [en ligne]. Disponible sur : <http://www.bnf.fr/fr/professionnels/f_um/s.format_unimarc_notices_bibliographie.html> (consulté le 25/04/2012).

15 Voir figure 2 : Première page de champs UNIMARC pour une notice de type « ouvrage »

Figure 1: Les mots-clés de la vidéo « Faire parler le sol »dans le Catalogue Collectif Indexé (CCI)

Le catalogage se fait essentiellement à la main. Chaque notice comporte les caractéristiques principales d'un document (le titre, sous-titre, le ou les auteurs, etc.). Chaque information est inscrite dans le champ UNIMARC correspondant.

En plus de ces caractéristiques, il y a aussi des champs pour insérer des mots-clés[16] (appelés aussi Sujets ou descripteurs) caractérisant thématiquement le document. L'assignation de ces mots-clés est communément appelé indexation.

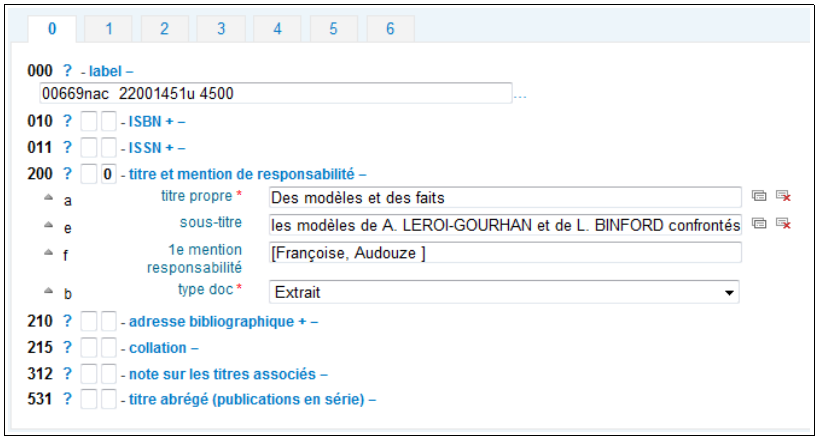


Figure 2: Première page de champs UNIMARC pour une notice de type « ouvrage »

Intéressons-nous maintenant à la méthodologie d'indexation.

16 Voir figure 1 : Les mots-clés de la vidéo « Faire parler le sol »dans le Catalogue Collectif Indexé (CCI)

c) Définition de l'indexation

L'AFNOR (Association Française de Normalisation) définit l'indexation comme suit:

"L'indexation est l'opération qui consiste à décrire et à caractériser un document à l'aide de représentations des concepts contenus dans ce document, c'est-à-dire à transcrire en langage documentaire les concepts après les avoir extraits du document par une analyse." (D'après la Norme AFNOR Z 47-102 (1978))

L'indexation permet de synthétiser les grandes idées des documents soit avec des mots-clés qui sont assignés librement, soit avec des descripteurs qui sont issus d'un vocabulaire contrôlé (thésaurus ou autre). Un mot-clé est un « mot ou expression choisi généralement dans le titre ou le texte d'un document pour en caractériser le contenu et en permettre la recherche. Il constitue un point d'accès. Il est à distinguer d'un descripteur, qui est un terme normalisé dans un thésaurus. »[17] Un descripteur est « un terme retenu dans un thésaurus pour représenter sans ambiguïté une notion contenue dans un document ou dans une demande de recherche documentaire. Ce peut être un nom commun ou un nom propre (nom géographique, de société, de personne, terme taxonomique, etc.), une locution, un mot composé ou un groupe de mots. »[18]

La finalité de l'indexation est de « faciliter l'accès au contenu d'un document ou d'un ensemble de documents à partir d'un thème, d'un sujet, d'un concept ou d'une combinaison de thèmes, sujets ou concepts »[19].

Au niveau de l'indexation, on peut distinguer plusieurs types d'indexation[20] :

- l'indexation par extraction automatique
- l'indexation humaine
- l'indexation en vocabulaire libre
- l'indexation en vocabulaire contrôlé

L'indexation automatique désigne l'indexation qui est essentiellement faite par l'ordinateur. Il s'agit d'une des méthodes les plus courantes dans le traitement automatique des langues (TAL). Souvent, des textes bruts sont analysés par l'ordinateur qui calculera ensuite la probabilité qu'un mot-clé ou

17 ADBS. mot clé [en ligne]. Disponible sur : <http://www.adbs.fr/mot-cle-17878.htm?RH=OUTILS_VOC> (consulté le 28.06.2012)

18 ADBS. descripteur [en ligne]. Disponible sur : <http://www.adbs.fr/descripteur-16756.htm> (consulté le 28.06.2012)

19 ASSAL S. Mots-clés d'auteurs et langages documentaires. Réflexions sur la valorisation des revues du pôle éditorial de la Maison René-Ginouvès [en ligne]. Mémoire pour obtenir leTitre professionnel « Chef de projet en ingénierie documentaire » INTD niveau I. Paris : INTD, 2009, 128 p. Disponible sur : <memsic.ccsd.cnrs.fr/docs/00/52/38/78/PDF/ASSAL.pdf> (consulté de 26.05.2012)

20 http://www.cddp92.ac-versailles.fr/spip2/texteanim/La%20chaine_documentaire_Semaine_IG_ESEN.pdf

descripteur soit pertinent pour un texte donné. L'ordinateur peut aussi bien faire une indexation avec ou sans vocabulaire contrôlé. Cette méthode est réputée rapide, mais peu fiable surtout par rapport à l'indexation humaine.

Dans le cas de l'indexation humaine, pour synthétiser un indexeur humain doit lire le document soit en entier, soit son résumé. Cette méthode apporte une grande précision, mais est très chronophage et coûteuse, car le nombre de documents disponibles s'est multiplié notamment avec l'Internet et l'obligation pour les chercheurs de plus en plus nombreux de publier. Aussi le choix des mots descripteurs peut varier de personne en personne. Deux indexeurs ne donneront rarement exactement les mêmes mots-clés pour un même texte. Cela est parfois dû au niveau de connaissances du sujet traité dans le document à indexer.

L'indexation par vocabulaire libre consiste à donner des mots-clés aux documents, mais sans avoir recours à une liste de mots. Cela se rapproche des tags que mettent les internautes aux articles de blogs. Nommé parfois *folksonomie*, cela permet d'indexer rapidement s'il y a beaucoup d'internautes qui taggent et peut traduire une tendance du moment sur les habitudes documentaires des internautes. L'inconvénient est que le vocabulaire n'est pas contrôlé. Cela signifie qu'un document peut être indexé sous plusieurs mots-clés qui s'avèrent être des synonymes ou des singuliers ou pluriels du même concept. Cela a comme conséquence qu'il peut y avoir beaucoup trop de mots-clés pour un seul document et qu'il sera difficile de retrouver un document si on ne possède pas le même vocabulaire ou la même conception du sujet que la personne qui l'a indexé. En corollaire, on ne sait pas toujours avec quels mots-clés on doit formuler sa requête de recherche pour être sûr de retrouver le plus de documents pertinents possible.

L'indexation en vocabulaire contrôlé consiste au contraire à donner des mots-clés issus d'un vocabulaire contrôlé aux documents. Ce vocabulaire contrôlé peut être sous la forme d'une liste de mots, d'un thésaurus ou tout autre vocabulaire.

d) Méthodologie d'indexation manuelle du réseau FRANTIQ

Les outils

Le thésaurus PACTOLS

D'après Chaumier, « un thésaurus est un dictionnaire de mots ou expressions du langage naturel, termes normalisés et préférentiels, organisé d'une manière conceptuelle présentant les termes groupés par affinité sémantique et complété d'indications de relations »[21].

Les termes sont normalisés, c'est-à-dire qu'habituellement on choisit le terme au masculin singulier. Les termes préférentiels sont des termes que l'on utilise au détriment d'autres termes qui sont le plus souvent des synonymes. Cela évite de multiplier les mots-clés lors de l'indexation. Par affinité sémantique, on peut comprendre qu'un terme peut avoir un terme générique et un terme spécifique. Par exemple, le terme *France* a comme terme spécifique *France du Nord* et comme terme générique *Europe*. Avec les nouvelles normes pour la conception d'un thésaurus, d'autres relations seront disponibles (tout, partie d'un tout, instance de, etc.).

Le thésaurus PACTOLS a la spécificité d'être multilingue et spécialisé sur « les sciences de l'Antiquité et l'archéologie de la préhistoire à la Seconde guerre mondiale »[22]. Il a le mérite de ne pas être élaboré par une seule personne ou équipe, mais grâce à l'aide des spécialistes des thématiques du réseau FRANTIQ. Il s'agit donc d'un outil créé par la communauté du réseau pour ses propres besoins.

Il est très régulièrement mis à jour et possède son propre outil de gestion OpenTheso[23] avec lequel il est facile de rajouter de nouveaux concepts et synonymes.

Les termes descripteurs préférentiels sont classés selon des 6 microthésaurus et une liste. Ces derniers reprennent les thématiques clés de l'archéologie[24].

Une des spécificités de ce thésaurus est qu'il est disponible au format SKOS (Simple Knowledge Organization System), une variété de XML au format RDF. Cela signifie que le thésaurus est disponible sous une forme structurée et donc plus facilement manipulable par un ordinateur dans la

21 CHAUMIER Jacques. Les techniques documentaires [en ligne]. Paris, France : PUF, 2002. (Que sais-je ?). Disponible sur : <http://www.cairn.info/feuilleter.php?ID_ARTICLE=PUF_CHAUM_2002_01_0001> (consulté le 25.04.2012) I.S.B.N. 9782130524243

22 FRANTIQ. Thésaurus PACTOLS [en ligne]. Disponible sur : <http://frantiq.mom.fr/fr/thesaurus-pactols> (consulté le 25.04.2012)

23 Crée par Miled Rousset, informaticien basé à la Maison de l'Orient et de la Méditerranée

24 Les 6 microthésaurus et la liste sont divisés en thématiques : Peuples et cultures préhistoriques, Anthroponymes, Chronologie relative, Toponymes (lieu-dit), Oeuvres artistiques et littéraires, Lieux (du continent à la commune) et Sujets

perspective du web sémantique. « Ce modèle est défini comme « simple » par opposition à d'autres modèles, comme OWL (Ontologic Web Language), plus à même de représenter des structures sémantiques plus riches telles que les ontologies, mais de ce fait également plus complexes à utiliser »[25].

La représentation des concepts SKOS « repose sur les graphes RDF »[26]. Cela permet d'utiliser des propriétés RDF. Par exemple, des « indications portant sur le concept lui-même »[27], comme des termes préférentiels, alternatifs et équivalents dans d'autres langues, des relations sémantiques, comme la hiérarchie et l'association ou encore la représentation du concept par une image.

Dans la figure suivante[28], on peut voir que le terme descripteur préférentiel (skos:prefLabel) est *Édifice* et que son synonyme est *Bâtiment* (skos:altLabel). Son terme spécifique (skos:narrower) est *Édifice public* et son terme générique (skos:broader) est *Architecture*.

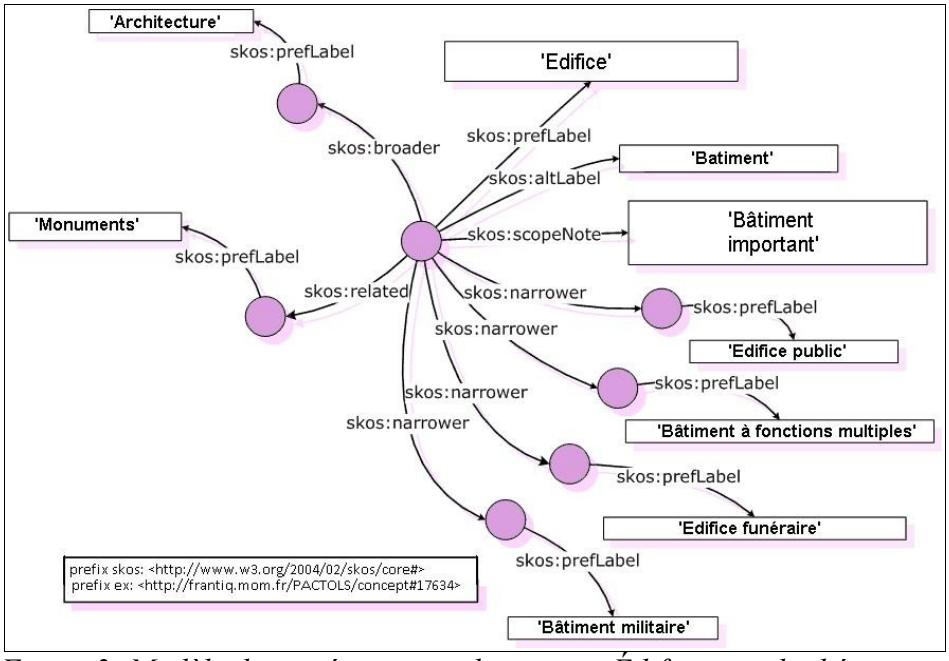


Figure 3: Modèle de représentation du concept Édifice issu du thésaurus PACTOLS (illustration prise du Livre blanc Indexation de corpus spécialisés avec ou sans référentiels de la MOM)

La transformation du thésaurus PACTOLS en ontologie a été commencée, mais malheureusement il s'est avéré que ce projet est très chronophage.

25 LÉNART Michèle. SKOS, un langage de représentation de schémas de concepts. Documentaliste-Sciences de l'Information [en ligne]. 2007, Vol. 44, pp. 75-75. Disponible sur : <www.cairn.info/revue-documentaliste-sciences-de-l-information-2007-1-page-75.htm> (consulté le 25.05.2012)

26 Ibid.

27 Ibid.

28 Voir figure 3 : Modèle de représentation du concept Edifice issu du thésaurus PACTOLS (illustration prise du Livre blanc Indexation de corpus spécialisés avec ou sans référentiels de la MOM)

OpenTheso

Ce projet a commencé en 2005 durant le stage d'un étudiant en Master informatique. Il s'agit d'un outil d'élaboration de thésaurus en ligne spécialement créé pour le réseau FRANTIQ car il n'en existait pas encore d'indépendant. Ce gestionnaire permet la poly-hiérarchie, ainsi que les traductions en plusieurs langues.

Pour pousser l'interopérabilité et l'accessibilité encore plus loin, des nouvelles fonctionnalités pour ce logiciel sont en cours de développement pour se conformer au format SKOS (Simple Knowledge Organisation System). »[29]

Grâce à OpenTheso de simples utilisateurs peuvent consulter tous les descripteurs du thésaurus et ainsi élaborer une stratégie de recherche, car ils peuvent voir quels mots sont utilisés pour l'indexation et quels mots sont juste des synonymes, par exemple. Comme on l'a vu dans le cours de « Systèmes de recherche et filtrage d'information »[30] de Madame Claverie, l'accès libre au thésaurus d'un fonds documentaire, aide l'utilisateur à élaborer sa propre stratégie de recherche et donc à retrouver plus d'ouvrages correspondant à ce qu'il recherche.

Le catalogue commun indexé

Ce catalogue est un catalogue commun à toutes les bibliothèques membres du réseau. Le logiciel libre système intégré de gestion de bibliothèque (SIGB) KOHA a été choisi car conforme aux formats MARC.

Ce catalogue a bien sûr une interface utilisateurs, appelée OPAC qui permet de faire des recherches et de créer des listes d'ouvrages favoris.

La méthodologie

Pour les membres du réseau FRANTIQ, il y a deux méthodes pour intégrer leurs notices bibliographiques dans le catalogue. Soit les membres rédigent leurs notices et assignent des mots-clés directement dans le catalogue (aussi appelé mode de catalogage partagé), soit ils envoient un fichier qui sera importé par les documentalistes du service de la MAE et par Miled Rousset, un informaticien basé dans une Maison partenaire située à Lyon (la Maison de l'Orient méditerranéen - Jean Pouilloux)[31].

29 TÊTES DES RÉSEAUX DOCUMENTAIRES (TRD). OpenTheso [en ligne]. Disponible sur : <http://trd.mom.fr/article.php3?id_article=81> (consulté le 25.06.2012)

30 LMSIC134 Systèmes de recherche et filtrage d'information

31 MAISON RENÉ GINOUVÈS. Présentation du service de documentation et Frantiq [en ligne]. Disponible sur : <http://www.mae.u-paris10.fr/usr3225/Presentation,32.html> (consulté le 25.05.2012)

La méthode de catalogage partagé a plusieurs avantages. Afin d'indexer les documents, le thésaurus PACTOLS est intégré au *Catalogue Collectif Indexé*, c'est-à-dire que l'indexeur peut directement dans le catalogue choisir les mots-clés qu'il juge utile et les intégrer directement dans les champs UNIMARC appropriés de la notice lors de sa création. Tout membre du réseau FRANTIQ peut soumettre des suggestions de « candidats descripteurs » via le Catalogue Collectif Indexé s'il ne trouve pas le mot clé approprié au document qu'il indexe.

Mais cela ne veut pas dire qu'ils seront tous acceptés dans le thésaurus. Une politique a été mise en place pour l'ajout de descripteurs. Il faut au minimum 3 bibliothèques membres utilisant ce descripteur au moins une dizaine de fois pour espérer voir leur candidat descripteur devenir un descripteur dans le thésaurus PACTOLS.

Cette procédure permet d'éviter les mots « à la mode » qui émergent parfois. Par exemple, le nom « ethnozooarchéologie ». Il est parfois utilisé dans des articles, mais il n'a pas été accepté dans le thésaurus car il existe déjà les descripteurs « ethnoarchéologie » et « archéozoologie ».

Il y a cependant des exceptions à cette règle, car s'il y a de nouvelles fouilles ou découvertes importantes, il y aura certainement un bon nombre d'articles qui seront publiés sur ce thème. Il s'agit d'anticiper en s'informant régulièrement sur ce qui se passe dans la communauté scientifique pour réussir à indexer le mieux possible les articles à venir.

Le thésaurus n'est pas le seul outil d'indexation dans le catalogue. Les indexeurs disposent de MACLES**.** Ce sont des mots-clés macroscopiques qui « sont constitués d'entités de trois chiffres obéissant à un découpage de la réalité selon quatre axes : espace-temps, disciplines-objets d'études, et méthodes »[32]. Ces mots-clés ont été élaborés par le Centre de recherches archéologiques (aujourd'hui intitulé le CEPAM) pour le plan de classement physique des ouvrages de sa bibliothèque.

Il ne suffit pas d'indexer pour indexer, mais il faut se mettre à la place de l'utilisateur. À quelles questions de chercheurs et d'étudiants l'indexation doit répondre ? Par exemple, pour un archéologue qui cherche un article, le mot clé « archéologie » ne suffit pas. Il faut être beaucoup plus précis.

Une autre règle est d'utiliser le terme générique, s'il y a plus de trois termes spécifiques dans le texte, sauf pour des termes comme « céramique », « méthodologie », « industrie osseuse », qui sont trop souvent utilisés pour l'indexation dans le catalogue. Donc, ces termes ne sont pas assez

32 FRANTIQ. Indexations dans FRANTIQ-CCI [en ligne]. Disponible sur : <http://www.frantiq.fr/fr/CCI/indexations> (consulté le 25.04.2012)

discriminants pour une recherche efficace dans le catalogue. Par exemple, pour le mot clé céramique, on retrouve 16789 notices. Une telle quantité de documents, peut rebuter le public.

Même si le thésaurus PACTOLS prévoit les thématiques importantes de l'archéologie, l'indexeur n'a aucune obligation de renseigner toutes les rubriques. La typologie des articles définit le nombre et les types de mots-clés, ainsi pour une comparaison entre des œuvres (par exemple, des stèles de types différents), on indexe par la période principale dont parle le texte, mais on ne retient pas les périodes avec lesquelles on compare la période principale. De plus, un seul thème peut être traité sous des angles différents dans des volumes de revues.

En général, on assigne :

- 8 à 10 mots-clés pour la thématique (termes spécifiques)
- 2 à 5 termes pour les lieux
- 1 à 3 pour la chronologie
- Pour l'anthropologie, le nombre de termes dépend du texte
- 0 à 3 pour les peuples
- 2 maximum pour les oeuvres
- pas de limite pour les toponymes

Peu importe la langue du document à indexer, les documents sont indexés avec des mots-clés en français, malgré le fait que le thésaurus soit multilingue.

En tout, il y a plus de 390 000 ressources actuellement référencées dans le catalogue CCI. Toutes ont été indexées manuellement par les documentalistes du réseau FRANTIQ selon ces critères.

2.3. Sortie

a) Définition

Il s'agit de valoriser la collection de documents présents dans un catalogue papier ou en ligne grâce au catalogage. Souvent pour les catalogues en ligne, on met une interface spécifique du catalogue à disposition du public. Elle permet, entre autre, de faire des recherches dans le catalogue sur des champs des notices tels que *auteur*, *titre*, *mots-clés* ou encore *résumé*.

b) Les services et produits du Réseau FRANTIQ

Le catalogue KOHA permet grâce à la partie OPAC (Online Public Access Catalogue) de faire des recherches documentaires et de créer des produits documentaires élaborés et la partie administrative permet de gérer les prêts. Le service de documentation de la MAE songe à instaurer des services pour créer des produits documentaires à la demande. Il s'agirait dans notre cas, de faire des

recherches documentaires très poussées à la demande de chercheurs et enseignants-chercheurs afin de créer des listes bibliographiques spécifiques à chaque demandeur dans le Catalogue Collectif Indexé.

3. Faisabilité d'une indexation semi-automatique

Ma mission de stage est de voir quels outils peuvent être susceptibles d'aider au mieux les indexeurs humains du réseau FRANTIQ dans leur travail d'indexation manuelle. Nous allons d'abord faire un état des lieux en analysant un prototype d'un outil d'indexation semi-automatique créé pour le réseau en 2009. Puis analyser des articles scientifiques pour comprendre pourquoi un outil d'indexation semi-automatique a été choisi par le réseau FRANTIQ.

Puis dans un second temps, nous allons voir en utilisation de la méthode Proof of concept (POC) si un outil d'indexation semi-automatique est réellement envisageable et quels avantages et inconvénients cela apportera au travail quotidien d'un indexeur professionnel au sein du réseau FRANTIQ.

3.1. L'état des lieux

a) Attentes du réseau FRANTIQ d'un outil d'aide à l'indexation

Aujourd'hui, l'indexation d'articles dans le catalogue commun indexé se fait exclusivement de façon manuelle. Or, il y a beaucoup de nouvelles ressources qui ne sont pas indexées, faute de temps et surtout de moyens humains. Il est connu que l'indexation manuelle d'une grande quantité de documents peut rapidement devenir un travail insurmontable. L'indexation automatique semble résoudre les problèmes de coût et de temps qu'implique l'indexation manuelle.

La grille de comparaison entre l'indexation manuelle et l'indexation automatique[33] donne des arguments pour et contre de ces deux méthodes[34].

33 D'après Jacques Chaumier et Martine Dejean

34 Voir figure 4: Grille de comparaison entre l'indexation manuelle et l'indexation automatique.

	Indexation humaine ANALYSTE	Indexation automatique ANALYSEUR
	Analyses réalisées	
Volume/Temps	Quelques documents ou dizaines de documents par jour	L'ensemble des documents en quelques minutes ou quelques jours (en fonction du volume traité)
Moyens humains	Un ou plusieurs documentalistes qualifiés	Aucun
Coût	Très élevé. À partir de 23 euros par document	Peu élevé car le retour sur investissement est rapide (en fonction du volume traité)
Qualité de l'indexation	Maximale	Peu pertinente dans son ensemble
Homogénéité / Uniformité de l'indexation	Subjective. Elle dépend du savoir du documentaliste, ainsi que de l'époque de l'indexation	Le système est totalement homogène et uniforme
Mise à jour de l'indexation	Jamais. Une indexation humaine ne suit pas l'évolution naturelle du vocabulaire, ce qui pose problème pour le rappel d'anciens documents	La ré-indexation de l'ensemble des documents est programmable. Elle peut se faire en temps réel ou différé

Figure 4: Grille de comparaison entre l'indexation manuelle et l'indexation automatique.

Le coût de l'indexation manuelle est certes élevé, mais l'indexation est d'une qualité supérieure à celle de l'indexation automatique. Par contre, en terme de volume/temps, l'indexation automatique est bien plus efficace. On peut se poser la question s'il n'existe pas un juste milieu entre la rapidité de l'indexation automatique et la qualité de l'indexation par un indexeur humain. Cette thématique a été déjà étudiée sous la forme du concept d'indexation semi-automatique. L'indexation semi-automatique n'est pas un outil qui indexe à la place de l'indexeur humain, mais un outil qui aide l'indexeur humain à indexer plus rapidement.

A l'initiative du réseau FRANTIQ, une équipe de 6 étudiants de Polytech Lyon a rédigé un rapport de recommandations[35] ainsi qu'un livre blanc[36]. Ils exposent que le réseau a des attentes complémentaires par rapport à une indexation automatique qui le porterait à choisir une indexation semi-automatique, en raison du taux d'erreur plus ou moins élevé qui existe en traitement automatique des langues.

Les futurs utilisateurs devraient donc valider un par un tous les mots-clés proposés par l'algorithme d'indexation semi-automatique. L'avantage est qu'un indexeur n'aura plus à chercher des descripteurs adéquats dans un grand thésaurus, mais de prendre des descripteurs dans la liste de mots-clés potentiels.

Dans cette optique, et pour diminuer le risque d'un algorithme complexifiant l'indexation, l'équipe

35 HADJEB Sonia, BEDOK Kévin, BERMOND Ameline, et al. Rapport de recommandations - Indexation de corpus spécialisés avec ou sans référentiels. Lyon : Maison de l'Orient et de la Méditerranée (MOM), 2012, 53 p.

36 HADJEB Sonia, BEDOK Kévin, BERMOND Ameline, et al. Livre blanc - Indexation de corpus spécialisés avec ou sans référentiels. Lyon : Maison de l'Orient et de la Méditerranée (MOM), 2012, 53 p.

d'étudiants préconise une interface conviviale et facile à utiliser. Dans ce même rapport, ils proposent aussi quelques algorithmes et logiciels prêts à l'emploi. Nous allons nous intéresser d'abord à leurs préconisations, puis aussi à d'autres solutions dans un chapitre suivant.

Il existe un moyen de savoir comment les mots-clés ont été trouvés afin que les utilisateurs puissent plus facilement désambiguïser les mots-clés proposés par le programme. Le texte devrait être consultable et que les mots-clés potentiels devraient être visualisés en contexte. Par exemple, les mot-clés potentiels sont soulignés directement dans le texte à indexer et les synonymes des mots-clés potentiels sont aussi repérables.

Les utilisateurs pourraient proposer des mots-clés candidats, s'ils ne trouvent pas les mots-clés appropriés dans le thésaurus.

Une étude sur l'indexation semi-automatique du Laboratoire PSI – INSA et de l'Université de Rouen d'un corpus de documents liés à la santé avec le thésaurus MeSH a montré des besoins similaires. L'objectif d'un développement d'un système d'indexation automatique est de « permettre d'étendre la couverture du catalogue tout en maintenant une indexation de qualité, et en assurant des taux de précision et de rappel élevés lors de la recherche d'information dans le [catalogue] CISMeF ». Pour arriver à combler les besoins, « le système d'indexation automatique doit impérativement intégrer les normes d'indexation manuelle en vigueur. En effet, ce système est destiné à aider l'indexation manuelle, afin de réduire les délais actuels, tout en validant la qualité de l'indexation automatique proposée. Le système ne constituera pas une alternative à l'indexation manuelle. »[37].

Comme on a pu le voir dans la grille de comparaison, les indexeurs humains (souvent faute de temps), ne ré-indexent que rarement les anciens documents.

Un des besoins exprimés par le réseau FRANTIQ est d'utiliser un programme pour indexer ou ré-indexer des anciens articles pour améliorer le catalogue documentaire commun[38] des bibliothèques du réseau. Le thésaurus évolue constamment et au moment où les textes anciens ont été indexés, il n'y avait peut-être pas encore les mots-clés plus appropriés dans le thésaurus et ils peuvent avoir été inscrits en mots libres (champs spécifiques). Or, beaucoup d'anciens mots libres font maintenant partie du thésaurus, mais comme ils n'étaient pas disponibles au moment de l'indexation, ces mots libres ne pointent en conséquence pas sur le thésaurus.

37 NÉVÉOL Aurélie. Indexation automatique de ressources de santé à l'aide d'un vocabulaire contrôlé. Actes des Rencontres des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues [en ligne]. 2004, pp. 105-14. Disponible sur : <http://aune.lpl.univ-aix.fr/jep-taln04/proceed/actes/recital2004/Neveol.rec04.pdf> (consulté le 15.05.2012)

38 Le catalogue commun indexé (CCI Frantiq) sous le logiciel KOHA - http://koha.mom.fr/

b) Le prototype d'un outil d'indexation semi-automatique

Suite aux demandes du réseau, en 2009 un premier prototype[39] a été élaboré tenant compte des besoins de leurs indexeurs.

L'algorithme développé prend en entrée, le thésaurus PACTOLS et des textes au format MODS[40] ou XML, puis compare les textes avec les concepts du thésaurus. La sortie est une liste de mots-clés potentiels affichés par ordre décroissant d'occurrences, le titre du document et un signalement des synonymes[41].

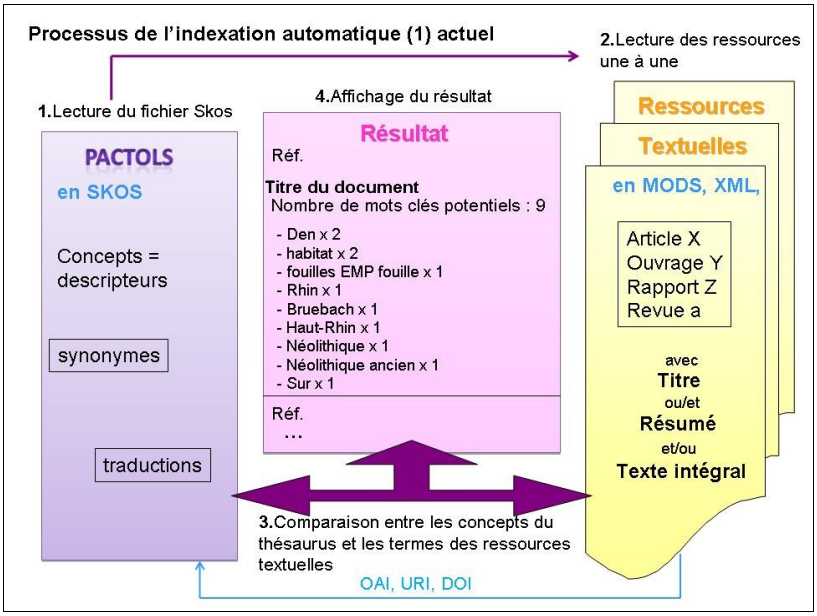


Figure 5: Schéma du fonctionnement de l'indexation automatique

Après la première phase de développement, il y a eu des modifications pour améliorer les résultats. En effet, il y a parfois des mots vides comme *sur* qui sont détectés comme mots-clés potentiels.

39 MARCHEIX Loraine, ROUSSET Miled, FERHOD Djamed. Document interne au Service FRANTIQ n° FRQ-Doc 09-04. - Bilan d'une première indexation automatique. 2009, 4 p.

40 MODS (Metadata Object Description Schema) est un modèle destiné au traitement de données bibliographiques, en particulier dans le contexte des bibliothèques, mais peut être élargi à d'autres usages. Il est particulièrement intéressant dans le cadre de projets de description de documents numérisés. Source : BnF http://www.bnf.fr/fr/professionnels/f_mods/s.mods_presentation.html

41 Voir figure 5: Schéma du fonctionnement de l'indexation automatique

Réflexions

Le schéma suivant tente de résumer ces réflexions[42].

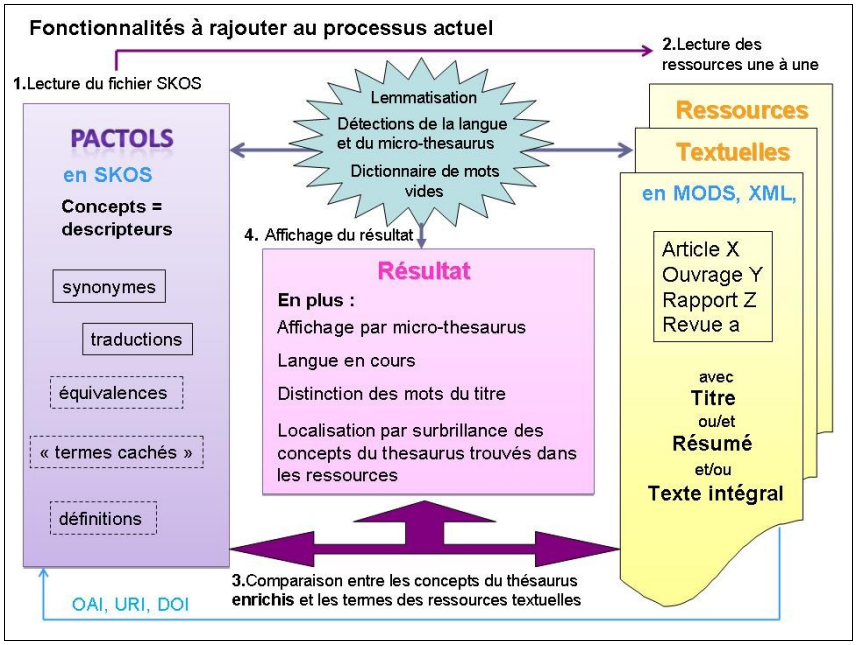


Figure 6: Fonctionnalités à rajouter après réflexion

On voit donc, que les membres du réseau souhaitent un classement par microthésaurus des mots-clés pour mieux comprendre la sémantique des mots-clés potentiels.

Afin de minimiser les erreurs d'indexation, mais aussi l'ambiguïté des mots dans le texte, il y a plusieurs méthodes du traitement automatique des langues qui sont suggérés. La lemmatisation[43] du texte et l'utilisation d'un dictionnaire de mots vides[44], par exemple.

La lemmatisation permet de réduire le nombre de mots à traiter, car toutes les formes conjuguées du verbe être par exemple, seront vues comme le verbe être à l'infinitif.

Il est intéressant d'enlever les mots vides, car statistiquement ce sont les mots les plus nombreux dans un texte. Or, si on se base sur la méthode classique qui propose que les mots les plus fréquents sont les plus importants dans le texte, il faut alors éliminer les mots vides. Après leur élimination, idéalement, il ne reste que des mots porteurs de sens dans le texte. Et c'est à partir de ce point, que l'on pourra efficacement classer les mots par leur fréquence.

Ensuite, il est aussi prévu d'indexer de façon semi-automatique des textes en d'autres langues que le français actuellement. Une détection de la langue sera souhaitable dans le cas.

42 Voir figure 6: Fonctionnalités à rajouter après réflexion

43 La lemmatisation étant de ramener un mot à sa forme canonique (lemme). C'est-à-dire, par exemple de convertir tous les verbes conjugués en des verbes à l'infinitif.

44 Un dictionnaire de mots vides est une liste de mots qui ne sont pas considérés comme porteurs de sens. Les mots comme le, la, etc.

Il est aussi important de voir souligner les mots-clés potentiels du texte. Cela permettrait à l'indexeur d'un coup d'œil de voir les mots-clés faux ou inappropriés proposés par l'algorithme.

En effet, en traitement automatique des langues, il y a toujours un taux d'erreur, car il est difficile d'apprendre à un programme toute la méthodologie de réflexion humaine. C'est dans cette difficulté que réside l'intérêt d'un outil d'indexation semi-automatique.

Maquette

Dans cette optique, une première maquette d'interface pour l'outil[45] a été élaborée[46].

Il est prévu que l'outil d'indexation semi-automatique soit un outil indépendant. C'est-à-dire qu'il ne sera pas couplé ni avec le catalogue, ni avec le logiciel de gestion de thésaurus. L'accès à l'outil se fera cependant à partir du site web FRANTIQ.

La page d'accueil (Annexe 1), permet à l'utilisateur de choisir la langue du document et les documents à indexer. Une fois que l'utilisateur a validé son choix, les documents sont traités par l'algorithme derrière l'interface.

Puis, une autre page affiche les textes que l'algorithme a traités (Annexe 2). L'utilisateur peut choisir le texte pour lequel il souhaite voir la liste de mots-clés potentiels, ainsi qu'un extrait (Annexe 3) avec les mots-clés potentiels soulignés. L'utilisateur peut ensuite décocher les mots-clés qu'il pense inappropriés et ensuite valider les mots-clés appropriés et ajouter des mots-clés à la main au cas où l'algorithme les a oubliés.

La possibilité d'ajouter les mots-clés à la main est aussi intéressante au cas où le mot-clé n'existe pas dans le thésaurus afin de le proposer comme candidat descripteur.

Ce projet est resté à l'état de maquette, car il est nécessaire de faire une étude sur les besoins du public visé.

Nous allons voir à travers d'autres exemples, qu'il existe d'autres outils et méthodes pour fournir une aide aux indexeurs humains.

c) Autres méthodes et outils

Un des premiers constats était qu'il n'y a que très peu d'études sur l'indexation semi-automatique à l'aide de thésaurus. Surtout quand il s'agit de documents électroniques, l'indexation par un langage contrôlé est « souvent opposée à l'indexation automatisée sur le texte intégral, alors qu'il serait préférable de voir leurs complémentarités »[47]. L'indexation automatique indexe tous les mots

45 MARCHEIX L., ROUSSET M., FERHOD D, Document interne au service FRANTIQ n° FRQ-Doc 09-10.- Projet d'interface pour l'assistance à l'indexation. 2009, 4 p.

46 Voir Annexe 1, 2 et 3

47 ASSAL S. Mots-clés d'auteurs et langages documentaires. Réflexions sur la valorisation des revues du pôle éditorial de la Maison René-Ginouvès [en ligne]. Mémoire pour obtenir le Titre professionnel « Chef de projet en ingénierie documentaire » INTD niveau I. Paris : INTD, 2009, 128 p. Disponible sur : <memsic.ccsd.cnrs.fr/docs/00/52/38/78/PDF/ASSAL.pdf> (consulté de 26/05/2012)

pleins[48] du texte intégral et la recherche des utilisateurs se fait sur les mots contenus dans le texte intégral et non sur des thèmes qui pourraient qualifier le contenu.

Pourtant, utiliser un thésaurus pour l'indexation automatique évite d'extraire des mots-clés qui sont peut-être très souvent présents dans le texte, mais qui ne sont peut-être pas assez pertinents.

Rares sont les disciplines qui ont leur propre thésaurus spécialisé. La mise en place d'un thésaurus est coûteux et à ce jour il n'existe qu'un seul logiciel libre de gestion de thésaurus OpenTheso. Cela pourrait expliquer pourquoi l'indexation automatique à l'aide de thésaurus n'est pas très répandue.

Il y a quand même de bons exemples sur ce sujet.

ISIDORE

Le portail de ressources scientifiques spécialisé en sciences humaines et sociales Isidore[49] est un bon exemple d'indexation totalement automatique. L'équipe fait un enrichissement automatique, c'est-à-dire qu'elle propose pour chaque ressource[50], des mots-clés issus soit du thésaurus PACTOLS de FRANTIQ soit du langage d'indexation matière RAMEAU de la Bibliothèque nationale de France[51]. Comme dans toutes les applications de TAL, il y a un risque d'erreur dans l'attribution des mots-clés. Par contre, ce risque est atténué en affichant un extrait du document à côté des mots-clés, ainsi on peut se rendre compte des mots-clés pertinents et de ceux qui ne le sont pas. Pour plus de sécurité, ils prennent aussi en compte les mots-clés d'origine du document. Ainsi, on multiplie les façons avec lesquelles les utilisateurs peuvent retrouver les informations recherchées.

Projet d'indexation semi-automatique pour le catalogue médical CISMeF

Un autre exemple est le système d'indexation semi-automatique de ressources dans le domaine de la santé développé par l'équipe autour d'Aurélie Névéol pour les indexeurs du Catalogue et Index des Sites Médicaux Francophones (CISMeF). Ce système « intègre la plateforme linguistique INTEX (Silberztein 1993), un outil d'analyse de corpus puissant, dont les différentes fonctionnalités peuvent être intégrées dans d'autres systèmes ». La méthodologie d'indexation de leurs indexeurs est la suivante. Elle se sert de deux types de ressources, le thésaurus MeSH[52], et « trois bases de

48 Mot plein : opposé à mot vide. Les mots pleins sont tous les mots qui portent du sens dans une langue donnée, au contraire les mots vides ne portent peu ou pas de sens (les déterminants, les conjonctions, etc.)

49 http://www.rechercheisidore.fr/

50 Ce sont des données non structurées (texte brut intégral, d'un article scientifique par exemple ...) et des données structurées (par exemple, des articles scientifiques au format XML, RDF, ...)

51 RAMEAU (Répertoire d'autorité-matière encyclopédique et alphabétique unifié) http://rameau.bnf.fr/

52 Medical Subject Headings (MeSH)

connaissances contenant des informations sur les relations hiérarchiques entre les termes MeSH, la liste des descripteurs obligatoires (check tags), et un historique des associations mot clé/qualificatif tiré des notices CISMeF. »[53] Les règles d'indexation sont fondées sur les spécificités du MesH. Ce thésaurus ne propose pas seulement des descripteurs, mais des qualificatifs pour chacun. Ainsi, par exemple, les indexeurs n'indexeront pas seulement par le mot-clé simple *sida*, mais attribueront également le qualificatif *prévention & contrôle.* Cela donne la paire *sida/prévention & contrôle*, qui est beaucoup plus pertinente pour la recherche d'informations ultérieurement. C'est sur cette particularité qu'est basé aussi le système d'indexation semi-automatique. Leur système extrait grâce à des transducteurs INTEX (UNITEX) des paires de mots-clés simples/qualificatifs qui seront ensuite validés ou non par leurs indexeurs. Ces transducteurs sont crées entièrement à la main par des experts[54].

Des méthodes statistiques sont aussi prises en compte. Un score est attribué à chaque paire, selon le « nombre d'occurrences de chaque mot clé, de la longueur de la ressource, de l'historique des notices CISMeF, de la liste des descripteurs obligatoires, et de la structure hiérarchique du MeSH ».

Un poids supplémentaire est attribué aux mots-clés (ou paires) apparaissant dans les titres des paragraphes ou dans les résumés. Cela est justifié par le fait que leurs indexeurs « consultent principalement les titres pour vérifier l'indexation d'une ressource »[55].

Étude d'indexation semi-automatique pour un corpus spécialisé en droit

Une autre étude[56], utilisant également INTEX (UNITEX), préconise aussi une indexation semi-automatique pour aider l'indexeur à choisir plus rapidement les mots-clés appropriés. Le corpus traité concerne essentiellement des textes en français spécialisés dans le domaine législatif et parlementaire.

Avec cette méthode, l'indexeur humain aura idéalement juste à lire le titre, le résumé et éventuellement le premier paragraphe du texte, puis à choisir des mots-clés potentiels dans une liste

53 NÉVÉOL Aurélie. Indexation automatique de ressources de santé à l'aide d'un vocabulaire contrôlé. Actes des Rencontres des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues [en ligne]. 2004, pp. 105-14. Disponible sur : <http://aune.lpl.univ-aix.fr/jep-taln04/proceed/actes/recital2004/Neveol.rec04.pdf> (consulté le 15.05.2012)

54 KEVERS Laurent. Indexation semi-automatique de textes : thésaurus et transducteurs . IN : CORIA 2009 - Conférence en Recherche d'Information et Applications, 5 – 7 mai 2009, Presqu'île de Giens [en ligne]. Disponible sur : <asso-aria.org/coria/2009/151.pdf> (consulté le 20.05.2012)

55 NÉVÉOL A., Indexation automatique de ressources de santé à l'aide d'un vocabulaire contrôlé, RECITAL 2004, Fès, 21 avril 2004

56 KEVERS Laurent. Indexation semi-automatique de textes : thésaurus et transducteurs . IN : CORIA 2009 - Conférence en Recherche d'Information et Applications, 5 – 7 mai 2009, Presqu'île de Giens [en ligne]. Disponible sur : <asso-aria.org/coria/2009/151.pdf> (consulté le 20.05.2012)

générée par l'algorithme.

Cette liste utilise exclusivement les mots-clés présents dans un thésaurus. Cela permet de restituer chacun des mot-clés par rapport aux termes génériques etc. L'originalité de cette méthode réside dans le fait qu'elle part du thésaurus pour trouver les mots-clés potentiels dans le texte. De manière automatique, elle génère à la volée des transducteurs utilisables dans INTEX (UNITEX) qui seront ensuite appliquées aux textes afin d'extraire des mots-clés potentiels.

Le défi fouille de texte (DEFT)

Créé en 2005, le *défi fouille de texte* (DEFT) propose chaque année un défi en fouille de textes francophones sur des thématiques et corpus différents. L'intérêt d'un défi est « de permettre de confronter, sur un même corpus, des méthodes et logiciels d'équipes différentes »[57]. La thématique de l'année 2012 était l'« identification automatique des mots-clés indexant le contenu d'articles scientifiques ayant paru en revues de Sciences Humaines et Sociales, avec l'aide de la terminologie des mots-clés [et] sans terminologie. » Dix-huit équipes ont participé à ce défi.

Toutes les équipes reçoivent un même corpus d'apprentissage qu'ils doivent traiter pour ensuite pouvoir proposer une indexation automatique qui sera ensuite testée sur un corpus de test (différent du corpus d'apprentissage). Cette indexation automatique doit se faire avec et sans vocabulaire. Ce vocabulaire est issu des articles du corpus, ce sont en fait tous les mots-clés auteurs présents dans le corpus d'apprentissage. Le travail de chaque équipe consiste soit « à identifier, parmi [la liste des mots-clés auteurs], quels sont [les mots-clés] qui se rapprochent le plus de ceux que l'on aurait naturellement eu tendance à choisir, soit dans le cadre d'une indexation libre […] à déterminer quels sont les meilleurs candidats à l'indexation, généralement en usant de méthodes statistiques éventuellement complétées par d'autres approches »[58].

Tout au long du défi, on a pu voir qu'il y a différentes techniques, technologies et méthodologies pour l'indexation automatique et qu'il n'y a pas de consensus. Il y a des résultats de qualité différente, mais il n'y a aucune méthode universelle. Surtout cela montre qu'il n'y a pas pour la langue française de logiciels tout prêts pour l'indexation automatique ou semi-automatique avec un vocabulaire contrôlé (thésaurus ou liste de mots contrôlée).

57 http://deft.limsi.fr/

58 PAROUBEK Patrick, ZWEIGENBAUM Pierre, FOREST Dominic, et al. Indexation libre et contrôlée d'articles scientifiques Présentation et résultats du défi fouille de textes DEFT2012. IN : DEFT2012, 2012, Paris [en ligne]. Disponible sur : <http://jeptaln2012.org/actes/DEFT2012/pdf/DEFT201201.pdf> (consulté le 20.05.2012)

Étude d'un logiciel d'annotation sociale

Cette étude dont les auteurs ne sont pas favorables à une indexation par des indexeurs professionnels propose plutôt un outil d'indexation assistée par et pour les chercheurs en archéologie. Les chercheurs lisent brièvement les articles susceptibles de les intéresser un jour, mais « doivent être capables d'y accéder plus tard par rapport à un thème spécifique ». Selon Aurélien Benel, Sylvie Calabretto et Jean-Marie Pinon, le problème n'est pas de trouver l'information, mais de la *retrouver*[59]. En effet, comme dans beaucoup de disciplines, le document est un outil important pour les chercheurs. Chaque chercheur a sa propre vision des thématiques principales d'un corpus.

Le projet Porphyry[60] propose un outil dans lequel un chercheur peut annoter un corpus, puis confronter ces idées avec ceux des autres chercheurs qui ont également annotés ce même corpus. « Par exemple, pour comparer deux structurations de corpus, un expert pourra affirmer que le site archéologique « Shisma Eloundas » est équivalent au site appelé par un autre auteur « Schisma » partie de « la région d'Elounda » (car dans « Eloundas » le « s » est la marque d'un génitif grec). »

Si, les chercheurs arrivent à un consensus, ils pourront publier le résultat.

On notera que ce projet n'est valable que pour quelques chercheurs qui ont la même optique à un moment T ; il ne peut donc y résulter d'universalité.

L'indexation manuelle vs mots-clés d'auteurs

Comme nous avons vu précédemment, les auteurs sont invités à donner des mots-clés à chacun de leurs articles scientifiques. On peut se demander quels rôles ils jouent dans l'indexation.

Sophie Assal a mené une enquête auprès des auteurs afin de savoir comment ils élaborent leurs mots-clés et si les secrétaires de rédaction et directeurs de publications portent de l'attention à ces mots-clés. D'après son étude, les auteurs ont du mal à attribuer des mots-clés qui soient pertinents pour tout le monde et surtout pour leurs lecteurs potentiels.

Pour l'archéologie, les thématiques des mots-clés attribués par l'auteur devraient idéalement et au minimum couvrir l'aire géographique, la période et le nom du site archéologique. Mais les revues ne leur donnent pas de consignes pour le choix de ces mots-clés. Les auteurs sont donc libres du choix de l'orthographe d'un site (translittération française ou anglaise d'un lieu au Proche Orient) ou du pluriel ou du singulier d'un terme. Cela pose un problème aux secrétaires de rédaction qui ont des

59 BÉNEL Aurélien, CALABRETTO Sylvie, PINON Jean-Marie. Indexation "sémantique" de documents archéologiques. IN : Deuxième colloque du chapitre français de l'ISKO, 1999, Lyon [en ligne]. Disponible sur : <http://benel.tech-cico.fr/publi/benel_ISKO_99.pdf > (consulté le 10.05.2012)

60 http://www.porphyry.org/

attentes en terme de mots-clés (périodes, lieux, etc).

En revanche, certains auteurs se fixent leurs propres règles : par exemple, de n'utiliser que deux mots-clés généraux et le reste étant des mots-clés plus spécifiques et surtout de ne pas reprendre les mots du titre, car perçus comme trop redondants. Ou encore de se mettre à la place d'un lecteur potentiel, afin d'essayer d'imaginer avec quels mots-clés il pourrait chercher son article. Ou encore d'autres qui veulent cibler un très large public et mettent donc des mots-clés généraux.

Pour que les chercheurs et doctorants aient quand même quelques repères, Mme Lequeux a donné quelques cours sur l'indexation et les règles d'assignation de mots-clés. Dans ces cours, il est préconisé pour les mots-clés de donner des informations minimales comme la localisation géographique, le groupe ou le site étudié, la chronologie, l'époque et les thèmes principaux et éventuellement les thèmes secondaires. Il faut aussi tenir compte du contexte dans lequel on publie[61]. Selon le type de revues et le niveau scientifique, les mots-clés peuvent être plus ou moins pointus et spécifiques.

Selon Sophie Assal, les auteurs doivent fournir une liste de mots-clés avec leurs articles depuis le « développement de l'informatique documentaire, des bases de données et des langages contrôlées ». Pourtant, rares sont les utilisations de ces mots-clés dans l'informatique documentaire, car d'une part des vocabulaires contrôlés se sont développés et d'autre part les mots-clés des auteurs sont très similaires au taggage libre de texte à la volée qui demande un énorme travail de normalisation pour être rationalisé.

Selon Sophie Assal, les auteurs doivent fournir une liste de mots-clés avec leurs articles depuis le « développement de l'informatique, des bases de données et des langages contrôlées ». Pourtant, rare sont les utilisations de ces mots-clés dans l'informatique documentaire. On peut quand même citer l'exemple de la version électronique de la revue scientifique *Journal de la société des Americanistes* (JSA). L'index « appelé *mots-clés* a été créé à partir des mots- clés d'auteurs et ne subit aucun contrôle »[62].

61 LEQUEUX B., ASSAL S., Résumé – Mots-clés – Indexation, Atelier Ecole doctorale « Milieux, cultures et sociétés du passé et du présent » de Paris Ouest-Nanterre-La Défense

62 ASSAL S. Mots-clés d'auteurs et langages documentaires. Réflexions sur la valorisation des revues du pôle éditorial de la Maison René-Ginouvès [en ligne]. Mémoire pour obtenir le Titre professionnel « Chef de projet en ingénierie documentaire » INTD niveau I. Paris : INTD, 2009, 128 p. Disponible sur : <memsic.ccsd.cnrs.fr/docs/00/52/38/78/PDF/ASSAL.pdf> (consulté de 26/05/2012)

3.2. Les concepts en traitement automatique des langues

Nous avons vu à travers divers exemples, qu'il y a plusieurs études et projets pour aider les indexeurs dans leurs tâches d'indexation. En ce qui concerne les projets d'indexation semi-automatique, ils peuvent ou non partir d'une étude préalable des méthodologies d'indexation manuelle des indexeurs auxquels sont destinés ces outils. Les méthodes peuvent diverger, mais ce qui est toujours similaire est le résultat. C'est-à-dire, la sortie est toujours une liste de mots-clés potentiels, à partir de laquelle l'indexeur humain doit choisir les mots-clés les plus appropriés.

Intéressons-nous maintenant aux méthodes du traitement automatique des langues (TAL) pour arriver à cette liste de mots-clés potentiels.

En traitement automatique des langues, il y a plusieurs méthodes plus ou moins modernes pour l'indexation. Il y a d'abord la méthode moderne d'apprentissage supervisé ou non et aussi une méthode plus classique sans apprentissage (classification) qui peut-être également supervisée. Supervisé signifie qu'au préalable, une personne donne à un programme des règles ou des modèles d'apprentissage et exemples validés à la main. Non supervisé signifie que c'est le programme qui génère lui-même des règles ou modèles de données (clustering). Les statistiques viennent compléter ces méthodes.

a) Traitement du texte

Avant d'appliquer des méthodes d'apprentissage ou de règles, il faut pré-traiter le texte afin que les programmes reconnaissent les mots ou même des expressions, phrases ou paragraphes et d'autres phénomènes linguistiques dans une langue donnée.

En traitement automatique des langues, on distingue « classiquement (pour la langue écrite) six niveaux de traitement » [63]:

- le niveau de la segmentation en mots et en phrases ;
  - la tokenization (transformation du texte en unités lexicales)
- le niveau morphologique qui traite de la manière dont sont constituées les unités lexicales (flexion, dérivation, composition, etc.) et vise à déterminer la catégorie de discours de l'unité considérée ;
  - la lemmatisation (suppression des marques du pluriel, féminin, etc)
  - le tagging (catégoriser chaque mot d'un texte par rapport à son appartenance grammaticale. Verbe, nom, pronom, etc.)
  - le stemming (conservation de la racine lexicale des mots)

63 CHAUDIRON Stéphane. Technologies linguistiques et modes de représentation de l'information textuelle. Documentaliste-Sciences de l'Information [en ligne]. 2007, vol. 44, n°1, pp. 30-39. Disponible sur : <www.cairn.info/revue-documentaliste-sciences-de-l-information-2007-1-page-30.htm> (consulté le 16.08.2012)

    - la suppression de mots vides (déterminants, etc)
    - la normalisation des mots ambigus (transformation notamment des abréviations par l'écriture en toutes lettres)
- le niveau syntaxique qui détermine la structure des phrases en fonction de la grammaire de référence ;
- le niveau sémantique qui traite du sens des mots et des phrases ;
- le niveau du discours qui vise à identifier la structure discursive et argumentative du document ;
- le niveau pragmatique qui traite du monde de connaissance de référence, c'est-à-dire qui prend en compte les informations extra-linguistiques qui peuvent contribuer à la compréhension du texte.

Tous les niveaux ne font pas toujours partie d'un projet de TAL. En effet, « le niveau pragmatique est rarement pris en compte en tant que tel mais des connaissances de nature pragmatique peuvent être intégrées dans les dictionnaires de référence, en particulier les connaissances métiers »[64].

b) Les systèmes

Systèmes à apprentissage supervisé

L'algorithme a besoin d'exemples concrets afin d'entraîner un modèle d'apprentissage. Mieux on choisit les exemples d'apprentissage qu'on propose à l'algorithme meilleur sera le modèle d'apprentissage. Souvent, pour les algorithmes d'indexation, les exemples se composent de textes du corpus et des mots-clés validés par un indexeur humain. Pour Larochelle, « la tâche d'un algorithme d'apprentissage est alors d'entraîner un modèle qui puisse imiter le processus d'étiquetage par un humain, i.e., qui puisse prédire pour une entrée x quelconque la valeur de la cible y qui aurait normalement été donnée par un humain »[65].

Ce modèle d'apprentissage sera ensuite utilisé pour trouver des mots-clés potentiels, en suivant au plus près les exemples donnés lors de l'entraînement.

64 CHAUDIRON Stéphane. Technologies linguistiques et modes de représentation de l'information textuelle. Documentaliste-Sciences de l'Information [en ligne]. 2007, vol. 44, n°1, pp. 30-39. Disponible sur : <www.cairn.info/revue-documentaliste-sciences-de-l-information-2007-1-page-30.htm> (consulté le 16.08.2012)

65 LAROCHELLE Hugo. Étude de techniques d'apprentissage non-supervisé pour l'amélioration de l'entraînement supervisé de modèles connexionnistes [en ligne]. Thèse Philosophiæ Doctor (Ph.D.) en informatique . Montréal : Université de Montréal, 2008, 237 p. Disponible sur : <http://www.cs.toronto.edu/~larocheh/publications/thesis.pdf> (consulté le 20.06.2012)

Système à apprentissage non-supervisé

Ce type d'apprentissage ne prend pas d'exemples validés par des humains. Il prend par contre souvent des textes bruts en entrée et entraîne un modèle d'apprentissage selon ce qu'il trouve dans les textes. On utilise cette méthode, si on ne sait pas à l'avance quels résultats on souhaite trouver. Dans le modèle d'apprentissage supervisé, on veut que le modèle apprenne à sortir des mots-clés provenant d'un thésaurus par exemple. Or, si on n'a pas de thésaurus, on peut utiliser la méthode d'apprentissage non supervisée afin que celle-ci génère une catégorisation selon ce qu'il trouve dans les textes d'apprentissage.

Systèmes à base de règles ou transducteurs (graphes)

On utilise cette méthode dans le cas où le corpus d'entraînement n'est pas suffisant pour alimenter un algorithme d'apprentissage supervisé. On utilise des règles ou des transducteurs et d'autres ressources comme un dictionnaire pour analyser les corpus. Que l'on utilise des règles ou des transducteurs[66], on crée souvent les règles de grammaire à la main pour retrouver ce que l'on veut analyser dans le corpus. Par exemple, pour retrouver toutes les années dans un corpus, on doit rédiger la règle suivante : *une suite de quatre chiffres*. On doit l'expliciter ainsi car l'ordinateur ne connaît pas nécessairement le concept *année*.

Systèmes hybrides

Ces systèmes utilisent les règles écrites à la main et des algorithmes pour compléter automatiquement ces règles à l'aide d'un corpus d'apprentissage.

66 Ce sont des graphes, qui avec des symboles, permettent de représenter des phénomènes linguistiques complexes

c) Les étapes classiques de l'indexation automatique

Peu importe le système que l'on utilisera, à apprentissage, à base de règles ou encore hybride, les six étapes classiques du processus d'indexation automatique suivantes s'appliquent presque toujours [67]:

- segmentation des mots de la phrase ;
- élimination des mots vides ;
- lemmatisation des formes fléchies ;
- identification des syntagmes comme candidats-descripteurs ;
- pondération des mots, syntagmes ou descripteurs retenus ;
- éventuellement, remplacement des candidats-descripteurs (mots et syntagmes) par les termes du thésaurus du domaine (dans ce cas, l'indexation redevient contrôlée)

De manière classique, on segmente alors le texte en phrases, puis en mots. On lemmatise et élimine les mots vides (déterminants, etc.) pour trouver le plus de mots-clés potentiels. Puis, pour trier les mots-clés potentiels selon leur pertinence, on leur attribue des scores et une pondération.

Ces scores et pondérations permettent de filtrer les mots-clés. Une des méthodes les plus répandues est de faire ressortir les mots les plus fréquents du texte afin de générer des mots-clés. Pour éliminer dès le début un bon nombre de termes non pertinents, on peut aussi utiliser un thésaurus pour l'indexation. À partir de cette liste de mots les plus fréquents, on peut utiliser la méthode de TF-IDF[68]. Les mots-clés potentiels sont ensuite pondérés selon leur fréquence d'apparition dans le texte et aussi grâce aux indices TF-IDF (term frequency / inverse term frequency). TF-IDF permet d'attribuer un score aux mots-clés, selon leur présence ou absence dans *tout* le corpus et pas uniquement dans un seul texte. Plus ils sont fréquents dans le corpus, moins ils seront pertinents pour l'indexation d'un texte du corpus.

Une autre pondération consiste à prendre en compte la position des termes trouvés dans le texte, plus un terme est trouvé au début du texte, plus il a la chance d'être pertinent pour l'indexation.

Nous allons voir maintenant quels outils et algorithmes permettent d'utiliser les concepts vus auparavant.

67 CHAUDIRON Stéphane. Technologies linguistiques et modes de représentation de l'information textuelle. Documentaliste-Sciences de l'Information [en ligne]. 2007, vol. 44, n°1, pp. 30-39. Disponible sur : <www.cairn.info/revue-documentaliste-sciences-de-l-information-2007-1-page-30.htm> (consulté le 16.08.2012).

68 Term Frequency / inverse document frequency

3.3. Les algorithmes et outils disponibles sur le marché

Ils doivent être disponibles librement et gratuitement sur Internet, et supporter au minimum le français.

a) NLTK (Natural Language ToolKit)

Il s'agit d'une suite de bibliothèques[69] pour le langage de programmation python. Assez complet, il permet entre autre d'extraire des bi-grams[70], la tokenization, un interfaçage avec des tagger[71] et donne une liste de mots-vides en français.

b) Stanford Tagger

Il s'agit d'un algorithme qui permet d'annoter des textes. Il reconnaît des verbes, noms propres, …, et les annote pour qu'un autre programme puisse les reconnaître comme tels. Récemment, le français a été ajouté comme langue à traiter. Entièrement sous licence libre[72], il est une bonne alternative à TreeTagger.

c) TreeTagger

Il s'agit également d'un algorithme pour annoter des textes. En plus de l'annotation du type de mots, il permet une lemmatisation de la langue française. Par contre, il n'est pas en licence libre[73] et est principalement dédié à la recherche scientifique et l'utilisation à des fins pédagogiques.

d) Yatea

Il s'agit d'un extracteur de termes[74], spécialisé dans le français et l'anglais. Il prend en entrée un texte annoté par un tagger, puis en sortie donne une liste en XML de mots-clés potentiels. Ce programme s'utilise en priorité avec le langage de programmation Perl.

e) Snowball

Il s'agit d'un « stemmer »[75] pour plusieurs langues dont le français.[76] Il peut être utilisé avec n'importe quel langage de programmation pour traiter des textes.

69 http://nltk.org/
70 Bi-grams : Ce sont des paires de mots qui sont trouvés très souvent ensemble
71 Tagger : Ce sont des algorithmes qui peuvent annoter un texte. Ces annotations servent à différencier les types de mots. (Déterminants, pronoms, verbes, etc)
72 http://nlp.stanford.edu/software/tagger.shtml
73 http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
74 http://search.cpan.org/~thhamon/Lingua-YaTeA-0.5/
75 Conservation de la racine lexicale des mots
76 http://snowball.tartarus.org/

f) UNITEX (INTEX)

Un puissant programme d'analyse de corpus[77], écrit en JAVA. Il a la particularité de prendre en entrée des graphes (transducteurs) afin de représenter les phénomènes linguistiques que l'on souhaite analyser.

Ce logiciel est un système à base de règles.

g) KEA++

Un exemple d'apprentissage supervisé est le logiciel KEA[78] (Keyword Extraction Algorithm) de l'*University of Waikato*, une université néo-zélandaise. Il s'agit d'un algorithme très complexe[79] qui prend, ou non, en entrée soit un vocabulaire en forme de liste, ou soit un thésaurus au format SKOS (dérivé du XML) et un corpus de textes déjà indexés afin de générer un modèle d'apprentissage.

Il fonctionne entièrement en ligne de commande avec une machine virtuelle JAVA approprié. On peut donc en ligne de commande, choisir si on veut utiliser KEA++ avec un vocabulaire contrôlé ou pas.

En tout cas, il faut donner un corpus d'entraînement à l'algorithme. Il s'agit de donner au programme un corpus d'au moins 30 textes. Puis pour chacun des textes du corpus, on crée un nouveau fichier reprenant le nom du fichier de chaque texte mais avec une extension .key (clé en français). Ces fichiers .key contiennent les mots-clés validés manuellement pour chacun des textes.

Une fois installé, on doit d'abord entraîner cet algorithme afin qu'il produise un modèle d'apprentissage qui contiendra toutes les règles qu'il a déduites grâce aux textes d'apprentissage, les mots-clés validés à la main pour chacun de ces textes et éventuellement le thésaurus en SKOS. Cela est fait grâce à la classification naïve bayésienne, une classification probabiliste. C'est-à-dire qu'il compare les mots-clés potentiels trouvés avec les fichiers .key contenant les mots-clés validés manuellement. Il considérera alors tous les mots-clés potentiels qui sont identiques à ceux dans les fichiers .key comme exemples positifs et ceux qui ne se trouvent pas dans les fichiers .key comme des exemples négatifs. Il pourra ainsi apprendre quels mots-clés ont une grande importance et ceux qui en ont moins.

L'algorithme en appliquant le modèle d'apprentissage généré sur des textes de test, calculera d'après les exemples positifs (les mots-clés validés à la main), quels mots-clés pourraient être pertinents. Puis, une liste de mots-clés potentiels est générée en sortie.

77 http://igm.univ-mlv.fr/~unitex/

78 http://www.nzdl.org/Kea/

79 Voir Annexe 4 pour un schéma complet de fonctionnement de l'algorithme KEA++

Comme la méthode de classification décrite auparavant, cet algorithme d'apprentissage utilise les méthodes de TF-IDF, de lemmatisation, de stemming, la pondération des termes selon leur position dans le texte.

h) Maui

C'est un algorithme dérivé de KEA++. Il a quasiment le même fonctionnement de base que KEA++, mais rajoute d'autres variables pour l'extraction de mots-clés potentiels. On peut ajouter des statistiques sur les fichiers de mots-clés pour l'entraînement. C'est-à-dire que l'on peut spécifier le nombre de personnes qui ont approuvé ces mots-clés pour un texte donné.

3.4. Choix techniques

Comme nous voulons rester proche de la méthodologie des indexeurs du réseau FRANTIQ, nous privilégierions une méthode qui prend en compte le fait que le thésaurus sert de base pour l'indexation. Aussi, comme nous voulons savoir si une méthode d'indexation semi-automatique ou automatique comme suggéré par le livre blanc pour la MOM et le prototype réalisé par FRANTIQ peut convenir, j'ai choisi d'utiliser deux types d'outils d'indexation automatique qui seront ensuite remaniés pour devenir des outils d'indexation semi-automatique.

3.5. Preuve de concept (POC)[80]

Il s'agit de minimiser les risques en montrant une première implémentation qui pourrait résoudre le problème initial. À savoir, si un outil d'indexation automatique voire semi-automatique peut réellement correspondre aux besoins et aux attentes des indexeurs du réseau FRANTIQ. Pour cela, nous voulons aller plus loin que le premier prototype d'outil d'indexation semi-automatique du réseau. Dans ce dernier, des étapes classiques du traitement automatique des langues manquent. Par exemple, la tokenization et la lemmatisation.

Avec ce prototype, on peut montrer aux futurs utilisateurs à quoi pourrait ressembler l'application finale et tester son adhésion auprès du public cible.

J'ai choisi cette méthode pour pouvoir mesurer concrètement un taux de pertinence des mots-clés issus d'un outil d'indexation semi-automatique, tout en prenant en compte les attentes, les besoins et la méthode manuelle d'indexation utilisée par les documentalistes et bibliothécaires du réseau FRANTIQ. Cela est très important, car l'outil s'adresse à eux en priorité.

Pour débuter un prototype, il faut d'abord créer et analyser un corpus de test, puis choisir une méthodologie et faire des tests concrets.

a) Création et analyse d'un corpus de test

Le corpus se compose essentiellement d'articles des revues Gallia et Gallia Préhistoire qui appartiennent au même domaine scientifique que le thésaurus. Cela est important, car le thésaurus PACTOLS constitue une des ressources linguistiques les plus importantes dans l'indexation manuelle et dans le prototype d'outil d'indexation semi-automatique de FRANTIQ.

Tous les articles sont écrits en langue française, avec parfois des résumés en anglais et/ou en autres langues. Une des difficultés était de recevoir ces articles dans un format facilement exploitable par

80 En français québécois « démonstration de faisabilité »

des langages informatiques. En effet, souvent, on ne pouvait me fournir des textes qu'au format Word ou PDF. Or, un meilleur format pour pleinement exploiter les textes sont des formats ouverts comme des formats de texte brut (.txt) ou des formats de fichiers balisés (XML, HTML, etc). Même si il y a de plus en plus d'informatisation des documents (numérisation de textes anciens, des revues électroniques, …), il n'y a pas forcément partout des textes en format structuré ni des thésaurus ou ontologies pour chaque discipline. Pour les fichiers qui n'étaient pas au format texte brut (txt), je les ai converti manuellement ou automatiquement au format txt avec le logiciel libre pdftotxt[81].

Pour les articles de revues, le style d'écriture varie selon la langue. En sciences humaines, habituellement, les documents en langue française sont écrits dans un style littéraire et obéissent ainsi aux règles de ce style. C'est-à-dire qu'un auteur doit éviter de répéter le même terme plusieurs fois dans un même paragraphe, d'où l'utilisation fréquente de synonymes d'un même terme. Cela est différent pour les textes écrits en langue anglaise, par exemple. Dans cette langue, la répétition de termes dans un même paragraphe est répandue.

Les textes présentent généralement un résumé et une liste de mots-clés donnés par l'auteur. Ces résumés et liste de mots-clés sont écrits dans la langue du document, mais aussi souvent en anglais. C'est à partir des années 1970 que les revues scientifiques demandent aux auteurs d'articles de fournir des mots-clés qualifiant l'article ainsi que de traduire les résumés de leur article en anglais ou en plusieurs langues[82].

81 http://www.foolabs.com/xpdf/home.html

82 ASSAL S. Mots-clés d'auteurs et langages documentaires. Réflexions sur la valorisation des revues du pôle éditorial de la Maison René-Ginouvès [en ligne]. Mémoire pour obtenir le Titre professionnel « Chef de projet en ingénierie documentaire » INTD niveau I. Paris : INTD, 2009, 128 p. Disponible sur : <memsic.ccsd.cnrs.fr/docs/00/52/38/78/PDF/ASSAL.pdf> (consulté de 26.05.2012)

b) Méthodologie

Choix des algorithmes et application de la théorie

Au travers les cours de traitement automatique des langues, nous avons surtout vu des systèmes à base de règles et un peu moins les systèmes à apprentissage supervisés. Lors de ce stage, je voulais voir quels résultats je peux avoir en utilisant les deux méthodes et surtout voir les avantages et les inconvénients de chaque méthode.

J'ai donc choisi de d'utiliser le langage de programmation python avec NLTK pour le système à base de règles et KEA++ pour le système à apprentissage supervisé. La suite NLTK étant une suite de traitement automatique des langues très puissant et complet pour l'application sur des corpus en langue française, mon choix a donc porté sur le langage de programmation utilisée par cette suite, python. J'ai choisi aussi KEA++, car il s'agit d'un algorithme qui supporte les thésaurus en SKOS comme entrée. Cela est bien pratique pour l'utilisation du thésaurus PACTOLS qui lui est aussi en SKOS.

J'ai fait le choix de ne pas continuer le prototype d'indexation semi-automatique de FRANTIQ écrit en php, car il n'existe pas de programme de traitement automatique de langue (TAL) pour le français dans ce langage de programmation. Or, d'après les réflexions de l'équipe qui a créé le prototype, il était très important d'ajouter des outils TAL afin d'améliorer le prototype.

Système à base de règles

Pour ce système, je me suis inspirée du travail de Kervers[83] qui part du thésaurus et transforme les descripteurs et synonymes en transducteurs pour trouver les mots-clés potentiels puis qui filtre ces mots-clés selon des statistiques (TF-IDF notamment) et des étapes classiques en TAL d'une indexation automatique. Les transducteurs peuvent être comparés à des expressions régulières. Ces transducteurs sont des graphes dérivés des descripteurs et synonymes. Cela permet, quand on applique les transducteurs au corpus, de trouver un maximum de mots dérivés des descripteurs présents dans le thésaurus, en plus des synonymes. En effet, il est assez improbable qu'un auteur utilise exactement la même formulation d'un concept tel qu'il est écrit dans le thésaurus.

Je trouvais ce travail très original et en plus, il permet d'utiliser des règles que nous avons pu appliquer en cours. Aussi, comme le stage se déroule sur une courte période, il était plus judicieux

83 KEVERS Laurent. Indexation semi-automatique de textes : thésaurus et transducteurs . IN : CORIA 2009 - Conférence en Recherche d'Information et Applications, 5 – 7 mai 2009, Presqu'île de Giens [en ligne]. Disponible sur : <asso-aria.org/coria/2009/151.pdf> (consulté le 20.05.2012)

d'utiliser la méthode de l'équipe de Kervers, car elle fait générer automatiquement la plupart des règles. Cela fait gagner du temps, surtout par rapport à la méthode utilisée par le projet CISMeF pour lequel les règles ont étés écrites à la main par des experts[84]. Cela donne certainement une meilleure précision dans l'extraction de mots-clés potentiels, mais cela est surtout chronophage.

Par contre, je n'utiliserai pas UNITEX contrairement à l'équipe de Kervers, car je ne maîtrise pas assez ce programme pour l'utiliser pleinement. Donc, je vais utiliser le deuxième type de règles, les expressions régulières. Elles sont facilement implémentables dans python.

Prétraitement du thésaurus

En premier lieu, on analyse le thésaurus pour ensuite le transformer en expressions régulières. On va tenter d'appliquer tous les pré-traitements que l'on fait habituellement pour un texte, car le but est d'avoir des dérivées des descripteurs et synonymes.

Par exemple, pour retrouver le terme chasseur-cueilleur ou son dérivé chasseurs-cueilleurs dans un texte. Ce terme sera représenté en tant qu'expression régulière sous la forme suivante :

[ ]chasseur['.-a-zéëêèôàç]*cueilleur['.-a-zéëêèôàç][ ]

Dans cet exemple, on cherche un espace suivi de chasseur suivi de toute lettre alphabétique et des signes de ponctuation, puis suivi de cueilleur suivi de toute lettre alphabétique et des signes de ponctuation, et d'un espace.

Afin de pouvoir arriver à ce résultat là, il faut pré-traiter tous les descripteurs et les synonymes.

On remarque qu'on ne peut pas utiliser la racinisation française avec l'outil *snowball*, car beaucoup de mots sont trop spécifiques pour que nous gardions seulement la racine (les noms de personnes, villes, lieux, et certains mots comme « parement » qui ont pour racine « par »)

Afin de quand même retrouver des mots-clés au pluriel ainsi que d'autres variations minimales, on peut envisager un stemming minimal : supprimer le *s* du pluriel pour les sujets, car les lieux, par exemple, sont invariables et il est inutile de les pré-traiter. À la place de la marque du pluriel, on peut ainsi rajouter une expression régulière qui permet de trouver le mot ainsi qu'une variation du mot avec au maximum 2 caractères à la fin du mot.

Exemple : avec « dé » on trouvera « dés », mais pas « début ».

Les mots vides et signes de ponctuation seront supprimés. On termine donc l'analyse avec des mots-

84 KEVERS Laurent. Indexation semi-automatique de textes : thésaurus et transducteurs . IN : CORIA 2009 - Conférence en Recherche d'Information et Applications, 5 – 7 mai 2009, Presqu'île de Giens [en ligne]. Disponible sur : <asso-aria.org/coria/2009/151.pdf> (consulté le 20.05.2012)

composés composés de mots pleins. Les mots-vides sont remplacés par des expressions régulières permettant de retrouver dans un texte au maximum 3 mots entre chaque mot plein d'un mot composé. Cela donne l'expression régulière suivante pour le descripteur *main-d'œuvre* :

[ ]main['.a-zéëêèôàç]*(?: ['.a-zéëêèôàç]* | ['.a-zéëêèôàç]* ['.a-zéëêèôàç]* | ['.a-zéëêèôàç]* ['.a-zéëêèôàç]* ['.a-zéëêèôàç]* ['.a-zéëêèôàç]* |)oeuvre['.a-zéëêèôàç]{0,3}[ ]

Entre le mot *main* et *oeuvre* on cherche soit

- un espace suivi de plusieurs caractères alphabétiques
- un espace suivi de plusieurs caractères alphabétiques suivi d'un espace suivi de plusieurs caractères alphabétiques
- un espace suivi de plusieurs caractères alphabétiques suivi d'un espace suivi de plusieurs caractères alphabétiques suivi d'un espace suivi de plusieurs caractères alphabétiques

Les mots permettant de différencier les mots homonymes seront supprimés également. Cela est crucial, car plusieurs homonymes se trouvent dans le thésaurus. Par exemple, il y a *Vienne (Autriche)* et *Vienne (Drôme)*. Or, habituellement dans un texte, l'auteur ne mentionne qu'une seule fois dans son texte la localisation de la ville. Si on part du postulat que les mots les plus fréquents sont les plus importants, cela pose un problème. Le programme ne trouvera sûrement pas le *bon* Vienne, car la localisation n'est mentionnée qu'une seule fois, alors que le terme *Vienne* tout seul est mentionné nombreuses fois tout au long du texte. C'est pour cela que j'ai préféré enlever les distinctions entre les homonymes pour que le futur programme d'indexation retrouve tous les homonymes afin que l'indexeur humain fasse le choix du mot le plus approprié. Par contre, on pourra essayer, comme le suggère le livre blanc de la MOM, d'utiliser des bi-grams pour savoir quels autres termes sont proches des termes ambigus, afin de déterminer leur type. C'est-à-dire que si on trouve « Minerve » avec ville ou tout autre mot proche de l'urbanisation, on saura plus facilement qu'il s'agit de la ville « Minerve » et non de la déesse.

Pré-traitement du corpus

Afin d'alléger le nombre de mots à traiter dans le corpus, on peut envisager d'enlever les mots qui ne sont habituellement pas présents dans le thésaurus. Il y a d'abord les types de mots qui seront éliminés dans le thésaurus (les déterminants, …) et les types de mots qui ne sont quasiment jamais présents dans le thésaurus. Il est préférable de les enlever dans le texte, afin que le programme ait moins de mots à analyser. Cela présente un gain de temps lors de l'indexation. Pour cela on peut annoter le texte avec un tagger, puis ne pas prendre en compte les types de mots non désirés.

Par contre, on ne pourra pas appliquer sur le texte tous les pré-traitements réalisés sur le thésaurus.

On peut faire quelques désambiguïsations en remplaçant *l'* par *le* et un certain nombre de ponctuations ainsi que les retours à la ligne.

On ne peut pas toujours enlever les mots vides classiques du français car dans notre cas, par exemple « On » et « Grand », désignent des lieux. Cela pose un problème, car pour désambiguïser, on ne peut pour l'instant que choisir de laisser le texte tel quel (le texte n'est pas converti en minuscule, comme cela est habituel en TAL, et on n'enlève pas les accents). Ainsi, si dans le texte on trouve « On » avec une majuscule au début, l'expression régulière trouvera le descripteur pour Tell Hasan (« On » étant le synonyme), par contre il y a parfois des phrases en français qui commencent par un « On » pour désigner le pronom personnel.

L'utilisation d'un tagger français standard est envisageable, mais ce n'est pas pour autant une solution parfaite, car « On » sera soit toujours taggé comme étant un pronom personnel. Le même problème se pose aussi avec le descripteur « gratia » dont le synonyme est « grâce ». Il peut donc être facilement confondu avec la préposition « grâce à ». Pour ces derniers on pourrait envisager d'enlever toutes les expressions telles que « grâce à », « on a » etc.

Une autre méthode sera d'utiliser un tagger personnalisé qui désignera ce type de mots comme noms et non comme des adjectifs ou pronoms personnels. Avec un corpus annoté spécialement conçu pour l'archéologie et l'ethnologie, on pourra dire par exemple que *On* et *Grand* sont des noms propres.

A ce stade, on voit déjà qu'utiliser une solution standard de traitement du français n'est pas vraiment adapté au domaine de l'archéologie, même s'il s'agit de travaux en langue française.

La désambiguïsation va finalement être faite en laissant les mots du texte en l'état. On ne convertit pas le texte en tout minuscules, comme cela est souvent fait dans un traitement TAL. Cela permet, d'avoir une désambiguïsation simple et efficace.

Les statistiques

On souhaite disposer d'une liste de mots-clés potentiels une fois les expressions régulières appliquées sur le texte pré-traité.

Or, on peut déjà imaginer qu'il faudra un autre traitement afin de retrouver les mots-clés les plus pertinents pour le texte.

Comme dans l'article de Kevers, j'ai souhaité implémenter la méthode TF-IDF. Aussi, j'ai souhaité séparer le pré-traitement du thésaurus, qui doit être fait une seule fois, du pré-traitement et du traitement du corpus.

Puis, une fois le score calculé pour chaque mot-clé potentiel, on peut essayer de donner des coefficients différents selon des critères précis. Par exemple, on peut donner un plus grand score aux mots-clés qui se trouvent au début du texte ou diminuer le score s'il s'agit d'un mot-clé non désiré car trop fréquent dans le catalogue commun.

Les règles d'indexation

Aussi, les règles d'indexation doivent être prises en compte, comme le fait de choisir comme mot-clé le terme générique, dès qu'il y a plus de trois termes spécifiques qui sont trouvés ou encore d'éliminer les mots-clés les plus fréquents et peu significatifs pour le public (les mots-clés archéologie, céramique, etc).

Le tri des mots-clés selon les micro-thésaurus est une étape importante. On peut le faire si on extrait toutes les informations concernant chaque descripteur dans le thésaurus.

Schéma d'enchaînement dans python

Tous les techniques évoquées sont réunies dans un enchaînement afin d'avoir un programme qui pré-traite le thésaurus, puis le ou les textes et sort une liste de mots-clés potentiels.

Pour permettre de ne pas avoir à pré-traiter le thésaurus à chaque fois que l'on souhaite indexer un texte, on crée un programme de pré-traitement à part. Tant que le thésaurus ne change pas, on n'aura pas besoin de relancer ce programme.

Voici l'explication du fonctionnement du fichier intitulé thesaurus.py, qui pré-traite le thésaurus :

1. Le programme lit le thésaurus au format SKOS
2. Ensuite sont extraits les descripteurs, les synonymes, les identifiants des concepts et du microthésaurus, la langue
3. Puis avec des règles différentes selon le microthésaurus, on transforme les descripteurs et les synonymes en expressions régulières
4. Finalement le programme écrit les expressions régulières, les descripteurs à l'état original, la langue et les identifiants de concepts et du microthésaurus dans un nouveau fichier texte.

On extrait tous les informations possibles contenues dans le thésaurus pour chaque descripteur, afin de pouvoir facilement retrouver avec le programme leur appartenance à une langue, un microthésaurus, etc. Cela permettra ensuite un tri des mots-clés potentiels.

Puis, une fois toutes ces informations écrites dans un nouveau fichier intitulé descripteursPACTOLSfr.txt, il est prévu qu'un autre programme prenne ce fichier en entrée.

Voici l'explication brève du fonctionnement[85] du fichier intitulé regex_thesaurus.py, qui reprend les expressions régulières et les informations du thésaurus, pré-traite le ou les textes du corpus, et sort une liste triée de mots-clés potentiels :

1. Lecture du fichier descripteursPACTOLSfr.txt et stockage des informations en mémoire
2. Décompte du nombre de textes à indexer et du nombre de mots pour chaque texte, utile pour le calcul des scores TF-IDF
3. Pré-traitement du ou des textes
4. Application des expressions régulières au(x) texte(s)
5. Enregistrement des mots trouvés, leur fréquence et position
6. Élimination des termes les moins fréquents, selon la taille du texte
7. Calcul des scores TF-IDF pour chaque mot-clé potentiel
8. Enfin, tri des mots-clés potentiels selon leur appartenance au texte et aux micro-thésaurus et affichage de la liste

Résultats

Le résultat d'une indexation avec le programme python se présente ainsi :

Anthroponymes : Europe (femme)

Chronologie : Paléolithique

Lieux : Europe, Provence

Peuples : NMB

Sujets : approvisionnement, production, minéral, diffusion, grotte, matière première, lamelle, silex, matériaux, assemblage, éclat, population, matière, exploitation, foyer, lame, stade, débitage, artisanat, plancher, plaquette, nucleus, cour, objet, nombre

titre : Pasquini.txt

Pour l'instant cette indexation est seulement disponible en ligne de commande, mais présente un des critères importants voulus pour l'indexation semi-automatique, à savoir la classification selon les micro-thésaurus.

Après des tests, sur des résumés et des textes entiers, les résultats ont été vérifiés par ma tutrice de stage qui est à la fois documentaliste et archéologue.

On calcule d'abord la précision, qui est la proportion de mots-clés pertinents retrouvés par rapport au nombre total de mots-clés retrouvés puis le rappel, qui est la proportion de mots-clés pertinents retrouvés par rapport au nombre total de mots-clés pertinents attribués à la main par un indexeur[86]. Ces calculs attribuent un score de pertinence compris entre 0 et 1 à chacun des textes. Plus le taux

85 Une explication plus détaillée est disponible dans un schéma de l'annexe 5

86 CLAVERIE. C. LMSIC134 Systèmes de recherche et filtrage de l'information. Université Paris Ouest Nanterre La Défense, 2012

se rapproche de 1, plus il y a de mots-clés pertinents retrouvés par les expressions régulières.

Voici quelques statistiques de résultats.

5 résumés

Tableau de statistiques sur 5 résumés de textes provenant de Gallia et Gallia Préhistoire

	Texte 1	**Texte 2**	**Texte 3**	**Texte 4**	**Texte 5**
Précision :	32/38 = 0,84	19/23 = 0,82	15/22 = 0,68	21/30 = 0,70	21/25 = 0,84
Rappel :	1/7 = 0,14	9/39 = 0,23	8/32 = 0,25	7/22 = 0,31	4/47 = 0,08

5 textes entiers

Tableau de statistiques sur 5 textes entiers provenant de Gallia et Gallia Préhistoire

	Texte 1	**Texte 2**	**Texte 3**	**Texte 4**	**Texte 5**
Précision :	24/30 = 0,80	21/24 = 0,87	13/16 = 0,68	15/19 = 0,78	21/24 = 0,87
Rappel :	2/7 = 0,28	6/39 = 0,15	3/32 = 0,09	4/22 = 0,18	9/47 = 0,19

Réflexions

Beaucoup de mots-clés pertinents ont été trouvés. Par contre, le taux de rappel est fortement biaisée. On compare les mots-clés trouvés par le programme avec ceux qui ont été donnés auparavant par un indexeur humain. Or, même si les mots-clés potentiels trouvés ne correspondent pas exactement à ceux qui ont été attribués par l'indexeur humain, cela ne veut pas dire qu'ils ne sont pas pertinents. De plus, les mots-clés donnés par l'indexeur humain peuvent varier d'un indexeur à un autre.

Le taux de précision est beaucoup plus fiables, car un indexeur humain a vérifié directement la pertinence de chaque mot-clé et peut ainsi dire si le mot est vraiment pertinent ou pas.

En regardant les statistiques et les différences entre l'indexation sur 5 résumés et 5 textes entiers, on ne peut pas vraiment dire qu'il y ait une méthode meilleure que l'autre. Parfois, l'indexation sur les résumés a eu un meilleur résultat, mais cela dépend fortement de la qualité du résumé. Si le résumé parle de tout ce qui est important dans l'article, on va retrouver les concepts dans la liste des mots-clés potentiels.

Cela nous fait voir une des limites du programme. On trouve uniquement les termes et expressions qui sont écrites explicitement dans le texte, même si on a recours à des synonymes, du stemming et de la lemmatisation pour trouver les variations des mots dans le texte. C'est-à-dire que l'on ne

pourra pas trouver des expressions qui résument le texte, si cette expression n'a pas été donnée dans le texte lui-même. Le programme n'a pas les connaissances extra-textuelles qu'un indexeur humain pourrait avoir.

L'utilisation de synonymes est une aide précieuse pour le programme, mais là aussi, il faudrait être très exhaustif pour permettre au programme de reconnaître toutes les formes possibles d'un descripteur.

La méthode TF-IDF, permet d'atténuer les erreurs d'indexation, car si un mot se trouve dans tous les textes du corpus, il sera considéré comme peu discriminant. Par exemple, le descripteur « don » est mal attribué dans les trois textes d'exemple (les expressions régulières trouvent aussi bien « don » que « dont »). Le descripteur apparaît dans tous les textes et sera donc moins bien classé.

Pour améliorer la méthode TF-IDF, on est tenté de penser que l'idéal serait d'utiliser tous les mots-clés présents dans le catalogue afin de minimiser le score pour les mots-clés souvent attribués dans la totalité du catalogue (389160 notices dans le CCI au 11 mars 2012). Par contre avec cette quantité de notices, cela peut devenir difficilement réalisable du côté technique, car à chaque indexation, il faudrait recalculer les taux et scores pour chaque mot.

Deux causes sont principalement responsables des catégories non trouvées : l'absence de synonymes dans le thésaurus et l'utilisation, dans le texte, de termes trop ou pas assez concrets en regard du thésaurus. La polysémie de certains termes génère quant à elle du bruit[87].

Les termes du thésaurus ne se trouvent pas toujours tels quels dans les textes, même si les expressions régulières permettent de chercher à un niveau plus large. Il y a beaucoup de synonymes qui ne sont pas référencés dans le thésaurus.

Si les descripteurs sont aussi bien présents sous un terme unique que sous d'un terme composé, l'algorithme trouve les deux. Par exemple, ll trouve aussi bien « lieu » que « lieu de culte ».

Même si on utilise la pondération, le seul moyen de faire ressortir des termes par rapport à d'autres est leur fréquence d'apparition alors qu'un indexeur humain a bien d'autres critères pour choisir les mots-clés pertinents.

À part les statistiques, on voit aussi des erreurs issues d'expressions régulières, car celles-ci ne sont pas écrites à la main, mais générées automatiquement à la volée. On voit qu'il y a des erreurs, notamment pour les noms des auteurs, qui sont reconnus comme des noms de Saints. On pourra

87 KEVERS Laurent. Indexation semi-automatique de textes : thésaurus et transducteurs. IN : CORIA Conférence en Recherche d'Information et Applications, 5 - 7 mai 2009, Presqu'île de Giens [en ligne]. Disponible sur <http://asso-aria.org/coria/2009/151.pdf> (consulté le 10.05.2012)

songer à prétraiter le texte en coupant les lignes où se trouvent les noms d'auteurs. Cela est assez difficile, car cette ligne ne se trouve pas toujours au même endroit dans un texte. L'utilisation d'un corpus en HTML pourrait faciliter le découpage, mais nécessite ensuite un autre traitement, celui d'enlever les balises afin de ne pas apporter du bruit supplémentaire dans les résultats. Bruit, parce que les balises elles-mêmes pourraient être confondues avec le texte et donc être interprétées comme mots-clés potentiels.

Au niveau du temps d'exécution du programme, on pourra préférer traiter seulement les résumés, car ils présentent beaucoup moins de mots à analyser.

Ce prototype fonctionne et pourrait être amélioré avec un élargissement du corpus. Pour compléter la mission, il faudrait aller plus loin et proposer une interface graphique, si on veut que des personnes qui n'ont pas de compétences particulières en informatique puissent utiliser ce programme.

Avantages

La génération automatique des expressions régulières est rapide, mais nécessite l'écriture de règles de génération.

Il n'est pas indispensable d'avoir un corpus d'apprentissage annoté, comme avec les systèmes à apprentissage.

Les bibliothèques de NLTK sont riches et évoluent constamment. Elles permettent l'évolution du prototype avec des règles et calculs plus complexes.

Le prototype en python ne fait pas beaucoup de lignes et a l'avantage de pouvoir être lu dans un éditeur de texte pour être modifié.

Si on regarde bien le code, on peut facilement éditer et ajouter des règles d'indexation.

Inconvénients

Il faut continuer à développer le programme en python, car le langage python évolue aussi.

Les résultats trouvés dépendent fortement du nombre de synonymes et descripteurs présents dans le thésaurus. Il n'y a pas de calculs de probabilités permettant de déduire des mots-clés potentiels : les mots-clés potentiels doivent être écrits tels quels (avec quelques dérivés comme le pluriel ou des mots entre deux mots) pour être retrouvés par les expressions régulières. Cette procédure peut permettre de compléter sémantiquement le thésaurus.

Les règles pour générer les expressions régulières font qu'elles induisent parfois des erreurs. Ainsi

tous les mots-clés potentiels ne sont pas forcément trouvés car les expressions régulières ne détectent pas les mots dans le corpus à cause de ces erreurs. On peut y remédier en écrivant les expressions régulières à la main, mais cela est coûteux en temps.

Système à apprentissage

Pour l'indexation à l'aide d'un système à apprentissage supervisé, l'algorithme KEA++ a été retenu.

Une fois installé, on doit d'abord entraîner cet algorithme afin qu'il produise un modèle d'apprentissage qui contiendra toutes les règles qu'il a déduites grâce aux textes d'apprentissage, les mots-clés validés à la main pour chacun de ces textes et le thésaurus en SKOS.

Puis, l'entraînement fini, on peut utiliser KEA++ avec notre modèle d'apprentissage afin de pouvoir sortir une liste de mots-clés potentiels sur le corpus test de textes.

C'est une démarche simple et complexe à la fois. Comme tous les systèmes à apprentissage, les textes d'exemples et leurs mots-clés validés à la main prennent du temps à être élaborés. D'après le manuel d'utilisation de KEA++, il faut au moins 30 textes d'exemples avec mots-clés validés par un expert. En théorie, plus le corpus d'entraînement est conséquent, meilleur sera le résultat lors de l'indexation.

Avec l'aide de ma tutrice de stage, nous avons produit un corpus d'entraînement d'une centaine de textes.

Résultats

Le résultat d'une indexation avec l'algorithme KEA++ se présente soit comme une liste de mots-clés avec statistiques[88] dans une ligne de commande, soit comme une liste de mots-clés dans un fichier *.key* portant le nom du fichier texte traité.

Après des tests, sur des résumés et des textes entiers, les résultats ont été vérifiés également par ma tutrice de stage.

On calcule d'abord la précision puis le rappel. Ces calculs attribuent un score de pertinence compris entre 0 et 1 à chacun des textes. Plus le taux se rapproche de 1, plus il y a de mots-clés pertinents qui ont été retrouvés avec le modèle d'apprentissage.

Voici quelques statistiques de résultats.

88 Voir Annexe 6

5 résumés

Tableau de statistiques sur 5 résumés des 5 mêmes textes provenant de Gallia et Gallia Préhistoire que ceux testés avec le programme python.

	Texte 1	**Texte 2**	**Texte 3**	**Texte 4**	**Texte 5**
Précision :	16/ 25 = 0,64	18/ 25 = 0,72	16/ 25 = 0,64	20/ 25 = 0,80	20/25 = 0,80
Rappel :	2/7 = 0,28	6/39 = 0,15	4/32 = 0,12	3/22 = 0,13	3/47 = 0,06

5 textes entiers

Tableau de statistiques sur 5 textes entiers des 5 mêmes textes provenant de Gallia et Gallia Préhistoire que ceux testés avec le programme python

	Texte 1	**Texte 2**	**Texte 3**	**Texte 4**	**Texte 5**
Précision :	22/25 = 0,88	23/25 = 0,92	18/25 = 0,72	18/25 = 0,72	25/25 = 1
Rappel :	3/7 = 0,42	8/39 = 0,20	7/32 = 0,21	5/22 = 0,22	8/47 = 0,17

Réflexions

En comparant les résultats de précision et de rappel avec ceux du programme python, on voit que l'algorithme KEA++, donne souvent de meilleurs résultats, surtout pour le rappel, qui est une comparaison entre les mots-clés attribués par l'algorithme et les mots-clés attribués manuellement. Cela est peut-être dû au fait que l'algorithme a « appris » en s'entraînant la méthodologie d'indexation des indexeurs humains et a donc su assez souvent calculer une probabilité proche du choix de l'indexeur humain.

Par contre, des comportements étranges ont pu être observés. Le modèle d'apprentissage indexe tous les textes qui contiennent le mot « rue » avec le descripteur « Rue (Somme) », même si le texte parle d'urbanisation. On peut penser que comme plusieurs textes d'apprentissage avaient le mot-clé « Rue (Somme) », qui désigne la ville de « Rue » dans le département de la Somme, le modèle a pris comme modèle universel ces textes là. De même, le terme « Le Temple (Gironde) » est souvent utilisé à tort. On peut aussi seulement essayer de déduire l'origine de ce comportement en analysant le corpus, en effet, dans 14 textes du corpus d'apprentissage, on retrouve des expressions Le Temple de …, et l'algorithme a du méprendre cette expression pour le mot clé de la ville de Le Temple. Vu que cette expression est apparue dans 14 textes, il a du prendre cela comme exemple pour l'indexation de temples de tout genre. Une autre explication pourrait être dû au fait que l'identifiant de Le Temple (20702) est supérieur à celui de « temple » et que le système recherche les termes les

plus récents dans le thésaurus.

Il n'est pas facile d'interpréter ou de « désapprendre » ce comportement là, car le modèle d'apprentissage qui est généré par l'algorithme est sous forme binaire (Java serialization data, version 5). Il n'est donc pas facile de voir pour une personne qui n'a pas de compétences pointues en informatique quelles règles ont été apprises, à moins de « décompiler » le modèle d'apprentissage. Les modifications de programmes ne seront donc pas faciles à réaliser.

Dans le corpus on remarque que certaines périodes historiques sont comprises comme des dates importantes. Or, l'algorithme ne peut savoir cela, sauf si on le lui précise peut-être en mettant des périodes en synonyme. On peut d'ailleurs songer à ce sujet que l'algorithme peut peut-être être amélioré en prenant comme source une ontologie au lieu d'un thésaurus, afin que l'algorithme sache aussi faire des déductions.

Comme avec le traitement des résumés avec l'algorithme sous python, KEA++ confond souvent les noms et prénoms des auteurs avec des noms de Saints. La liste de mots-clés potentiels générée à partir des résumés semble aussi être moins intéressante que celle générée à partir de textes entiers. C'est-à-dire que dans la liste des mots-clés générée à partir de résumés comporte beaucoup de mots-clés génériques pour un article spécialisé en archéologie, comme *histoire*, *fouille* ou encore *édifice*. Cela peut-être dû au fait que dans un texte plus long, KEA++ peut faire des calculs de pertinence sur des fréquences de mots plus élevées. Or, dans un résumé, beaucoup de mots n'apparaissent qu'une seule fois dans le résumé. KEA++ peut donc difficilement appliquer la règle statistique qui dit que les mots les plus fréquents sont les plus importants, vu que la plupart des mots ont la même fréquence.

Globalement, on trouve beaucoup de termes corrects. Par contre, l'algorithme ne respecte pas la règle d'indexation selon laquelle on indexe par la période principale dont parle le texte, mais on ne retient pas les périodes avec lesquelles on compare la période principale du texte. Par exemple, le texte parle du Bronze Final, puis compare cette période au Néolithique moyen. Dans ce cas, un indexeur humain ne retiendra que la période principale du texte, à savoir Bronze Final et ignorera le Néolithique moyen lors de l'indexation.

On pourra peut-être améliorer le résultat en ajoutant des textes parlant de la Préhistoire. En effet, la majorité de notre corpus d'apprentissage est constitué de textes de la période Classique.

On note les avantages et inconvénients suivants :

Avantages

C'est un logiciel facile d'utilisation, sous licence libre et qui donne d'assez bons résultats si on l'entraîne bien.

Par des calculs de probabilités, on peut trouver des concepts même s'ils ne sont pas écrits tels quels dans le texte.

Il est rapide et portable. C'est-à-dire qu'on peut l'utiliser ce programme sous n'importe quel système d'exploitation, car il est écrit en java.

Il utilise des technologies du TAL moderne et des méthodes d'apprentissage complexes qu'il n'est pas facile de reproduire avec un système à base de règles.

L'algorithme est encore en développement et il est donc amélioré constamment et dispose d'une aide active sur des forums.

Inconvénients

La méthode d'apprentissage peut paraître obscure : le modèle d'apprentissage est généré en binaire java qu'il faut décompiler afin de pouvoir en lire les informations.

L'algorithme en lui-même est modifiable selon les besoins, mais il faut des connaissances en java, linguistique et les algorithmes d'apprentissage supervisé.

Un « mauvais » apprentissage n'est pas très facile à corriger. Comme avec tous les systèmes à base d'apprentissage. En effet, si on en donne trop peu, le programme trouve à peu près les mêmes mots descripteurs que le programme python qui n'utilise pas d'apprentissage.

L'algorithme ne possède pas les connaissances hors texte et la méthodologie complexe qu'un indexeur humain doit posséder pour bien indexer

Les mots-clés trouvés par l'algorithme sont, si on le souhaite, issus exclusivement d'un thésaurus, ce qui permet d'avoir un vocabulaire contrôlé, mais l'algorithme est incapable de faire des déductions.

Aussi, KEA++, on ne peut pas l'utiliser tel quel pour des utilisateurs finaux. Il faut encore faire du développement, comme le tri des mots-clés par micro-thésaurus et proposer une interface graphique adaptée.

c) Choix du système

Si on regarde les avantages et inconvénients et les résultats des deux systèmes (à base d'apprentissage et à base de règles), c'est le système à base d'apprentissage qui est le plus performant au niveau des résultats. Quel que soit le système choisi, il faut prévoir du temps pour chaque système car il faut créer des règles pour le système à base de règles et un corpus d'apprentissage pour le système d'apprentissage. De plus, les deux systèmes nécessitent du développement et des mises à jour.

d) Quelques idées pour améliorer les résultats

Utiliser un vocabulaire contrôlé pour l'indexation est un grand pas vers la sémantisation des données, c'est-à-dire vers la liaison des concepts entre eux et vers la création d'identifiant propre[89] pour chaque concept. Cependant, l'utilisation d'ontologies pour l'indexation est à l'étude, car une ontologie a une plus grande structuration et des relations entre les données sont plus complexes que dans un thésaurus..

On peut citer par exemple, le projet DYNAMO (DYNAMic Ontology for Information Retrieval) qui propose d'expérimenter l'indexation sémantique. Elle diffère d'une indexation semi-automatique avec un thésaurus, car « l'utilisation d'une ontologie permet d'accéder à la connaissance et de la rendre manipulable par des systèmes.»[90]. Elle permet aussi une pondération à partir de similarité conceptuelle, c'est-à-dire que le « pouvoir représentatif d'un concept prend en compte la fréquence d'apparition des termes désignant le concept dans le texte, mais également ses relations avec les autres concepts du domaine. Plus un concept a des relations avec les autres concepts présents dans le texte, plus il est représentatif de la page »[91]. On peut donc aller plus loin que ce que le permet le concept de TF – IDF.

89 DALBIN Sylvie et al.. Approches documentaires : priorité aux contenus. Documentaliste-Sciences de l'Information [en ligne]. 2011, vol. 48, n°4, pp. 42-59. Disponible sur <www.cairn.info/revue-documentaliste-sciences-de-l-information-2011-4-page-42.htm> (consulté le 30.05.2012)

90 LAUBLET P., AUSSENAC-GILLES N., CAMPS V., et al. DYNAMO, DYNAMic Ontology for information retrieval. Rapport d'évaluation de la version 1.0 Validation – Evaluation du modèle [en ligne]. Disponible sur : < www.irit.fr/dynamo/uploads/Rapports/Rapport-Lot9.pdf (consulté le 04.06.2012)

91 Ibid.

e) Une interface graphique

Quel que soit le système choisi et pour aller plus loin que le simple affichage de mots-clés, l'idéal serait un site web avec lequel les utilisateurs pourraient interagir afin d'avoir une liste de mots-clés potentiels, mais aussi afin de pouvoir éditer cette liste et l'exporter.

Avec nos deux systèmes, on peut imaginer le cas d'utilisation suivant[92] :

Cas d'utilisation indexation semi-automatique par un utilisateur connecté

1. L'utilisateur se connecte au site
2. Le site affiche un formulaire pour envoyer des fichiers sur le serveur afin que le site affichera ensuite des mots-clés potentiels
3. L'utilisateur envoie un ou plusieurs fichiers au format texte ou dans d'autres formats structurés
4. Le site stocke les fichiers dans un dossier Corpus
5. Le site affiche la liste des fichiers contenus dans le dossier Corpus
6. L'utilisateur choisi quels textes il souhaite voir traités en les cochant
7. L'utilisateur valide son choix
8. Le site donne la liste des fichiers à traiter au programme (python ou KEA)
9. Le programme sort une liste de mots-clés potentiels ainsi que les positions des mots-clés dans le texte d'origine dans des fichiers texte qui sont stockés dans un dossier Corpus_traité
10. Dès la fin du traitement qui peut durer plus ou moins longtemps selon la taille des textes, le site envoie un e-mail à l'utilisateur et lui donne un lien
11. L'utilisateur clique sur ce lien
12. Le site affiche une liste cliquable des textes qui ont été traités et qui se trouvent dans le dossier Corpus_traité
13. L'utilisateur clique sur le lien du texte désiré
14. Le site affiche un formulaire avec la liste des mots-clés permettant à l'utilisateur de décocher les mots-clés non voulus et d'en ajouter d'autres manuellement et le texte avec les mots-clés et synonymes soulignés
15. L'utilisateur valide ses choix
16. Le site exporte les données dans un format structuré

En ayant un compte sur le site, l'utilisateur peut revoir à chaque moment les mots-clés qui ont été générés et ses propres modifications. Aussi, il peut recevoir par e-mail une notification quand le traitement par l'algorithme est fini. Cela permet à l'utilisateur de ne pas avoir à attendre trop longtemps devant une interface, sans indication du temps de traitement, ni indication de l'exécution ou du plantage de l'algorithme.

Il faudrait, cependant, trouver une autre manière de présenter les résultats que celle qui était suggéré par la maquette d'origine. En effet, KEA++ comme le script python cherchent avec des synonymes,

92 Voir Annexe 7 pour un diagramme de séquence basé sur ce cas d'utilisation

avec les liens sémantiques du thésaurus et avec des expressions entières. Il est donc difficile de se contenter de souligner dans le texte les mots-clés potentiels trouvés, car les programmes vont au-delà de la simple correspondance entre mots du thésaurus et mots du texte.

Voici des idées de logiciels et langages qui pourraient aider à donner une interface graphique à chacun des méthodes d'indexation explorées.

Le programme en python

On peut créer un site web autour du programme python avec Django[93]. Django est une plateforme de développement web écrit en python. Il permet de créer ses propres pages et applications web en utilisant le langage python. On pourrait créer un site autour d'un script python d'indexation semi-automatique afin de l'utiliser dans un environnement web.

L'algorithme KEA++

Pour KEA, on pourra aussi directement créer un site web autour de cet algorithme, car il est facile d'interconnecter java et php. De plus, en guise d'exemple les développeurs de Maui (un autre algorithme se basant sur KEA++) ont mis en place une version web[94] de leur algorithme qui prend comme ressource linguistique différents thésaurus en SKOS (comme agrovoc[95]) et comme entrée un texte idéalement du même domaine que les thésaurus utilisés. On pourra s'inspirer de cet exemple pour notre futur logiciel.

L'informaticien à la MOM, Miled Rousset, m'a aidée à élaborer un plan pour utiliser KEA++ avec une page web. Pour développer avec java des sites Web, il faut installer Tomcat qui interprète les pages JavaServer Pages (JSP). Cela permet alors d'afficher les résultats dans des pages HTML. Ensuite pour structurer les pages web autour de l'algorithme KEA++, il faut suivre le modèle MVC (Modèle Vue Contrôleur).

Cela comporte une :

- couche Modèle : du code java avec connexion à la base de données et la base de données elle-même
- couche Vue : les pages html ou les JSP
- couche Contrôleur : ce sont les servlets ou (un FrameWork comme Struts2) qui font le lien entre Java et JSP ou HTML

93 http://www.django-fr.org/
94 http://maui-indexer.appspot.com/
95 Un thésaurus spécialisé en agriculture

On peut aussi, comme le suggère le livre blanc de la MOM, utiliser un logiciel tout prêt qui utilise KEA++ comme moteur d'indexation. Il s'agit de Constellio[96], qui est plutôt un moteur de recherche qui permet d'indexer tout type de document à l'aide d'un thésaurus et de KEA++. Il faudra voir plus tard, en testant ce produit, si cela correspond aussi aux besoins des indexeurs.

Avec KOHA

Vu que chaque notice comporte déjà des champs UNIMARC pour insérer des mots-clés, on pourra imaginer utiliser un outil d'indexation semi-automatique pour pré-remplir ces champs. Pour cela, il faudra développer une *moulinette* qui communiquera directement avec le catalogue KOHA.

96 http://www.constellio.com/

3.6. Premiers avis sur les résultats des systèmes d'indexation

Les listes de mots-clés ont été soumises à des futurs utilisateurs, afin qu'ils émettent leurs avis. Auparavant, je me suis posé les questions suivantes :

Est-ce qu'ils accepteront qu'une partie des mots-clés donnés par les logiciels soit complètement fausse ?

Est-ce qu'ils veulent une liste de mots-clés toute prête même s'ils sont obligés de relire le texte afin de voir quels mots-clés sont pertinents ?

Ou est-ce qu'ils veulent plutôt voir le texte en question et que le logiciel leur souligne les mots dans le texte qui se trouvent dans le thésaurus (mot descripteur et synonyme) ?

Les avis suivant ont été donnés :

On ne peut pas laisser utiliser un algorithme sans correction manuelle d'ajustement postérieur à l'exécution, car il y aura toujours des ambiguïtés dans les textes et donc des erreurs dans l'attribution automatique de mots-clés. Il y aura toujours un auteur qui emploiera un mot spécifique sous un autre contexte que celui que l'on utilise dans le thésaurus. Il faut élargir le contexte afin de déterminer si cela est une tendance générale ou s'il s'agit d'un cas isolé. On ne peut pas tout prévoir dans un algorithme, car la langue, les connaissances scientifiques, les termes et les concepts sont en constante évolution. C'est pour cela qu'il est utile que le thésaurus évolue parallèlement à la discipline.

En général, les mots-clés permettent de proposer aux indexeurs des mots-clés spécifiques qu'ils attribuent rarement.

D'autres questions ont été soulevées. Même si les mots-clés donnés par les logiciels sont présents dans le texte, est-ce que les professionnels de l'indexation les auraient choisi dans un contexte normal ? Il ne faut pas non plus que l'indexation semi-automatique leur complique le travail d'indexation, en proposant beaucoup de mots-clés non pertinents.

De plus, il ne faut pas oublier les utilisateurs : est-ce que les mots-clés donnés par les logiciels sont pertinents pour eux et leurs besoins de recherche ? Il faudra étudier l'habitude de recherche des utilisateurs, pour savoir quels mots-clés ils utilisent souvent pour retrouver un certain type de texte. En effet, le but d'une indexation, en général, est de rendre l'information plus facilement accessible aux utilisateurs.

Dans mon cas, ne connaissant pas du tout l'archéologie, il est indéniable que l'indexation semi-

automatique m'apporte quelque chose. En effet, surtout le logiciel en python, en triant les mots-clés potentiels par microthésaurus, m'a fait découvrir le sens de quelques concepts clés de l'archéologie. En effet, on lisant une simple liste de mots-clés, je ne savais pas à quoi correspondaient certains de ces mots-clés. Par exemple, je ne savais pas que *NMB (Néolithique moyen bourguignon)* ou *Alains* désignent des cultures ou peuples.

Il est donc possible que les personnes ne connaissant pas très bien le thésaurus vont découvrir des termes liés au texte. Cela n'empêche pas qu'ils doivent quand même prendre connaissance du thésaurus, car les logiciels font des erreurs et sont souvent incapables de retrouver des concepts complexes qui nécessitent une connaissance hors texte.

En ce qui concerne le taux d'erreur dans l'indexation semi-automatique, il est tolérable d'avoir un petit pourcentage d'erreur qui sera vérifié à la main avant validation. Sauf s'il y a une grande quantité de documents à indexer. Dans ce cas, la rapidité de l'indexation prime par rapport aux erreurs.

Cette méthode semble être pratiquée par le portail ISIDORE qui utilise aussi le thésaurus PACTOLS comme référentiel d'indexation, mais qui ne vérifie pas manuellement si les mots-clés attribués sont corrects ou pas. Il y a juste un réglage global pour minimiser le taux d'erreurs.

D'après L. Marcheix[97], le thésaurus « PACTOLS est utilisé par deux catégories de personnes, d'une part par des gestionnaires [...] et d'autre part, par des utilisateurs qui se divisent eux même en deux catégories : les indexeurs et personnels des bibliothèques, et les étudiants et chercheurs – « le public ». ». Les indexeurs des plateformes comme Revues.org qui n'utilisent pas de mots-clés pour indexer les revues pourraient également utiliser PACTOLS pour indexer les articles archéologiques. Cela permettrait de donner une autre possibilité de recherche aux utilisateurs de cette plateforme. On pourrait aussi ajouter les secrétaires de rédaction de toutes les revues archéologiques françaises.

La recherche en texte intégral est très utile, mais une recherche avec mots-clés permet idéalement de retrouver aussi des textes dont le mot-clé recherché n'existe pas explicitement dans le texte. En effet, un thésaurus peut prendre en compte les synonymes et traductions des termes et donc indexer tous les textes sous une seule thématique. Or, dans une recherche en texte intégral, l'utilisateur doit reformuler sa requête pour trouver aussi les synonymes du mot recherché, ainsi que les termes dans d'autres langues (fibule = fibula = Fibeln, etc.). Donc, l'utilisateur doit bien connaître les synonymes

97 MARCHEIX L. Conception d'une ontologie a partir d'un thesaurus specialisé dans le domaine de l'archéologie et des sciences de l'antiquité : rapport de stage [en ligne]. Master II professionnel de Gestion de l'Information et du Document Spécialité Gestion des Connaissances. Saint-denis : Université Vincennes, 2009, 157 p. Disponible sur : <memsic.ccsd.cnrs.fr/docs/00/35/59/20/PDF/Memoire_ontologie_Marcheix-vDefinitive.pdf> (consulté le 26.05.2012)

courants et ceci en plusieurs langues. Avec un thésaurus, ce problème est moins important car l'utilisateur a souvent accès à l'intégralité du vocabulaire et peut donc choisir les termes les plus appropriés à sa recherche avant de formuler sa requête.

• Conclusion

L'inflation du nombre de publications archéologiques et le manque de moyens humains entraîne la nécessité de mettre en place une de ces solutions d'indexation semi-automatique, tout en réalisant des tests d'adéquation avec les textes et en questionnant leur apport auprès des utilisateurs professionnels et chercheurs.

Il n'y a pas d'algorithmes parfaits, un logiciel ne va jamais remplacer de vrais indexeurs, car ils ont des méthodologies de métier qui ne sont pas totalement transposables en langage informatique. Les indexeurs physiques ont surtout des connaissances hors-texte qui permettent d'indexer de manière plus pertinente que le programme qui n'a que les informations du corpus pour trouver les bons concepts. Cependant, le logiciel développé en python ainsi que le logiciel KEA++ constituent déjà une bonne base vers une indexation semi-automatique.

Il serait bien à l'avenir d'étudier les habitudes et les méthodes d'indexation des indexeurs professionnels dans le domaine de l'archéologie, afin de prendre encore mieux en considération leurs besoins qu'ils soient chercheurs et donc rédacteur d'un texte qu'ils connaissent bien ou documentalistes, bibliothécaires ou secrétaires de rédaction qui doivent découvrir le texte pour l'indexer.

Bibliographie

- ADBS. descripteur [en ligne]. Disponible sur : <http://www.adbs.fr/descripteur-16756.htm> (consulté le 28.06.2012)
- ADBS. mot clé [en ligne]. Disponible sur : <http://www.adbs.fr/mot-cle-17878.htm?RH=OUTILS_VOC> (consulté le 28.06.2012)
- ASSAL S. Mots-clés d'auteurs et langages documentaires. Réflexions sur la valorisation des revues du pôle éditorial de la Maison René-Ginouvès [en ligne]. Mémoire pour obtenir le Titre professionnel « Chef de projet en ingénierie documentaire » INTD niveau I. Paris : INTD, 2009, 128 p. Disponible sur : <memsic.ccsd.cnrs.fr/docs/00/52/38/78/PDF/ASSAL.pdf> (consulté de 26/05/2012)
- BÉNEL Aurélien, CALABRETTO Sylvie, PINON Jean-Marie. Indexation "sémantique" de documents archéologiques. IN : Deuxième colloque du chapitre français de l'ISKO, 1999, Lyon [en ligne]. Disponible sur : <http://benel.tech-cico.fr/publi/benel_ISKO_99.pdf > (consulté le 10.05.2012)
- BIBLIOTHÈQUE NATIONALE DE FRANCE (BNF). Politique de catalogage [en ligne]. Disponible sur : <http://www.bnf.fr/fr/professionnels/anx_catalogage_indexation/a.politique_catalogage.html > (consulté le 25.08.2012).
- BIBLIOTHÈQUE NATIONALE DE FRANCE (BNF). Format UNIMARC [en ligne]. Disponible sur : <http://www.bnf.fr/fr/professionnels/f_um/s.format_unimarc_notices_bibliographie.html> (consulté le 25.06.2012)
- CHAUDIRON Stéphane. Technologies linguistiques et modes de représentation de l'information textuelle. Documentaliste-Sciences de l'Information [en ligne]. 2007, vol. 44, n°1, pp. 30-39. Disponible sur : <www.cairn.info/revue-documentaliste-sciences-de-l-information-2007-1-page-30.htm> (consulté le 16.08.2012)
- CHAUMIER Jacques. Les techniques documentaires [en ligne]. Paris, France : PUF, 2002. (Que sais-je ?). Disponible sur : <http://www.cairn.info/feuilleter.php?ID_ARTICLE=PUF_CHAUM_2002_01_0001> (consulté le 25.04.2012). I.S.B.N. 9782130524243
- DALBIN Sylvie et al.. Approches documentaires : priorité aux contenus. Documentaliste-Sciences de l'Information [en ligne]. 2011, vol. 48, n°4, pp. 42-59. Disponible sur <www.cairn.info/revue-documentaliste-sciences-de-l-information-2011-4-page-42.htm> (consulté le 30.05.2012)
- FRANTIQ. Indexations dans FRANTIQ-CCI [en ligne]. Disponible sur : <http://www.frantiq.fr/fr/CCI/indexations> (consulté le 25.04.2012)
- FRANTIQ. Thésaurus PACTOLS [en ligne]. Disponible sur : <http://frantiq.mom.fr/fr/thesaurus-pactols> (consulté le 25.04.2012)
- FRANTIQ. Types de documents [en ligne]. Disponible sur : <http://www.frantiq.fr/fr/CCI/TypeDocument> (consulté le 25.06.2012)
- HADJEB Sonia, BEDOK Kévin, BERMOND Ameline, et al. Rapport de recommandations - Indexation de corpus spécialisés avec ou sans référentiels. Lyon : Maison de l'Orient et de la Méditerranée (MOM), 2012, 8 p.

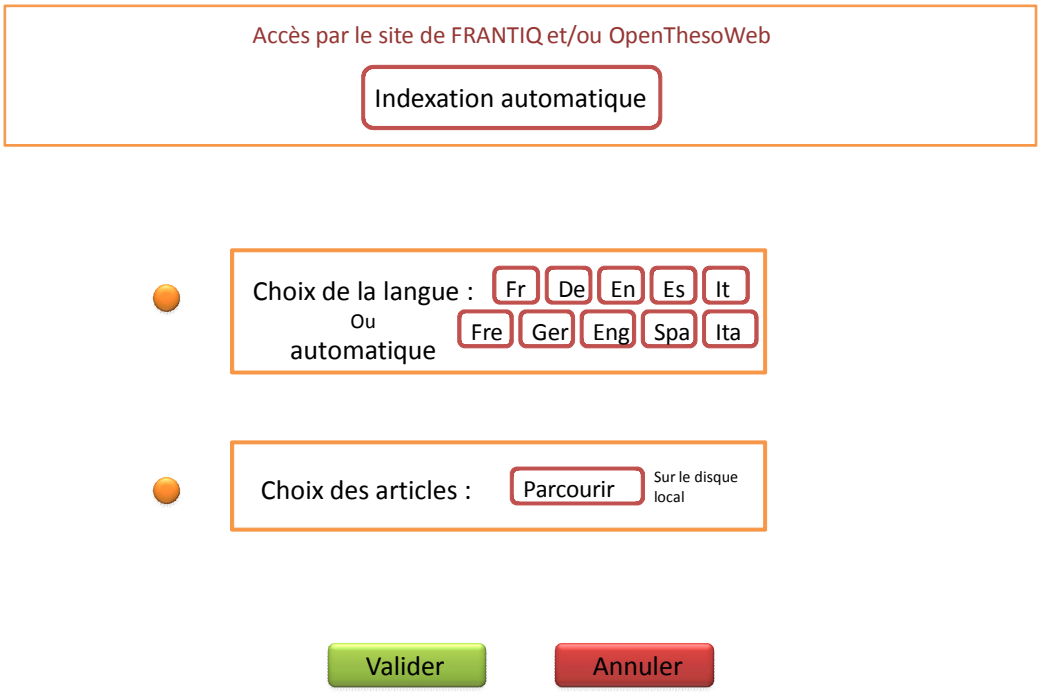
- HADJEB Sonia, BEDOK Kévin, BERMOND Ameline, et al. Livre blanc - Indexation de corpus spécialisés avec ou sans référentiels. Lyon : Maison de l'Orient et de la Méditerranée (MOM), 2012, 53 p.
- KEVERS Laurent. Indexation semi-automatique de textes : thésaurus et transducteurs . IN : CORIA 2009 - Conférence en Recherche d'Information et Applications, 5 – 7 mai 2009, Presqu'île de Giens [en ligne]. Disponible sur : <asso-aria.org/coria/2009/151.pdf> (consulté le 20.05.2012)
- LAMOUROUX Mireille. La chaîne documentaire. IN : Semaine de l'Inspection générale, 26 janvier 2011, 2011 ESEN, Poitiers Disponible sur : <http://www.cddp92.ac-versailles.fr/spip2/texteanim/La%20chaine_documentaire_Semaine_IG_ESEN.pdf> (consulté le 11.08.2012)
- LÉNART Michèle. SKOS, un langage de représentation de schémas de concepts. Documentaliste-Sciences de l'Information [en ligne]. 2007, Vol. 44, pp. 75-75. Disponible sur : <www.cairn.info/revue-documentaliste-sciences-de-l-information-2007-1-page-75.htm> (consulté le 25.04.2011)
- LEQUEUX B., ASSAL S. Résumé – Mots-clés – Indexation. Atelier Ecole doctorale « Milieux, cultures et sociétés du passé et du présent ». Nanterre : Université Paris Ouest Nanterre La Défense, 2009
- MAISON RENÉ GINOUVÈS. Histoire de la Maison [en ligne]. Disponible sur : <http://www.mae.u-paris10.fr/usr3225/Histoire-de-la-Maison.html> (consulté le 28.06.2012)
- MAISON RENÉ GINOUVÈS. Présentation du service de documentation et Frantiq [en ligne]. Disponible sur : <http://www.mae.u-paris10.fr/usr3225/Presentation,32.html> (consulté le 25.05.2012)
- MARCHEIX L. Conception d'une ontologie a partir d'un thesaurus specialisé dans le domaine de l'archéologie et des sciences de l'antiquité : rapport de stage [en ligne]. Master II professionnel de Gestion de l'Information et du Document Spécialité Gestion des Connaissances. Saint-denis : Université Vincennes, 2009, 157 p. Disponible sur : <memsic.ccsd.cnrs.fr/docs/00/35/59/20/PDF/Memoire_ontologie_Marcheix-vDefinitive.pdf> (consulté le 26.05.2012)
- MARCHEIX L., ROUSSET M., FERHOD D. Document interne au Service FRANTIQ n° FRQ-Doc 09-04. - Bilan d'une première indexation automatique. 2009, 4 p.
- MARCHEIX L., ROUSSET M., FERHOD D, Document interne au service FRANTIQ n° FRQ-Doc 09-10.- Projet d'interface pour l'assistance à l'indexation. 2009, 4 p.
- NÉVÉOL Aurélie. Indexation automatique de ressources de santé à l'aide d'un vocabulaire contrôlé. Actes des Rencontres des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues [en ligne]. 2004, pp. 105-14. Disponible sur : <http://aune.lpl.univ-aix.fr/jep-taln04/proceed/actes/recital2004/Neveol.rec04.pdf> (consulté le 15.05.2012) (consulté le 20.05.2012)
- PAROUBEK Patrick, ZWEIGENBAUM Pierre, FOREST Dominic, et al. Indexation libre et contrôlée d'articles scientifiques Présentation et résultats du défi fouille de textes DEFT2012. IN : DEFT2012, 2012, Paris [en ligne]. Disponible sur : <http://jeptaln2012.org/actes/DEFT2012/pdf/DEFT201201.pdf> (consulté le 20.05.2012)

- POMMARET Sabine. Traitement documentaire et valorisation des fonds iconographiques anciens dans les bibliothèques : l'exemple de la collection d'estampes de la B.M. De Bourges [en ligne] . Diplôme de conservateur de bibliothèque. Villeurbanne : ENSSIB, 2002, 83 p. Disponible sur : <http://www.enssib.fr/bibliotheque-numerique/document-1032> (consulté le 25/08/2012).
- TÊTES DES RÉSEAUX DOCUMENTAIRES (TRD). OpenTheso [en ligne]. Disponible sur : <http://trd.mom.fr/article.php3?id_article=81> (consulté le 25.06.2012)

• Annexes

- **Annexe 1**

Maquette d'interface d'accès du prototype d'outil d'indexation semi-automatique crée par le réseau FRANTIQ.[98]



98 MARCHEIX L., ROUSSET M., FERHOD D, Document interne au service FRANTIQ n° FRQ-Doc 09-10.- Projet d'interface pour l'assistance à l'indexation. 2009, 4 p.

- **Annexe 2**

Maquette d'interface de la liste des textes traités par le prototype d'outil d'indexation semi-automatique crée par le réseau FRANTIQ.[99]

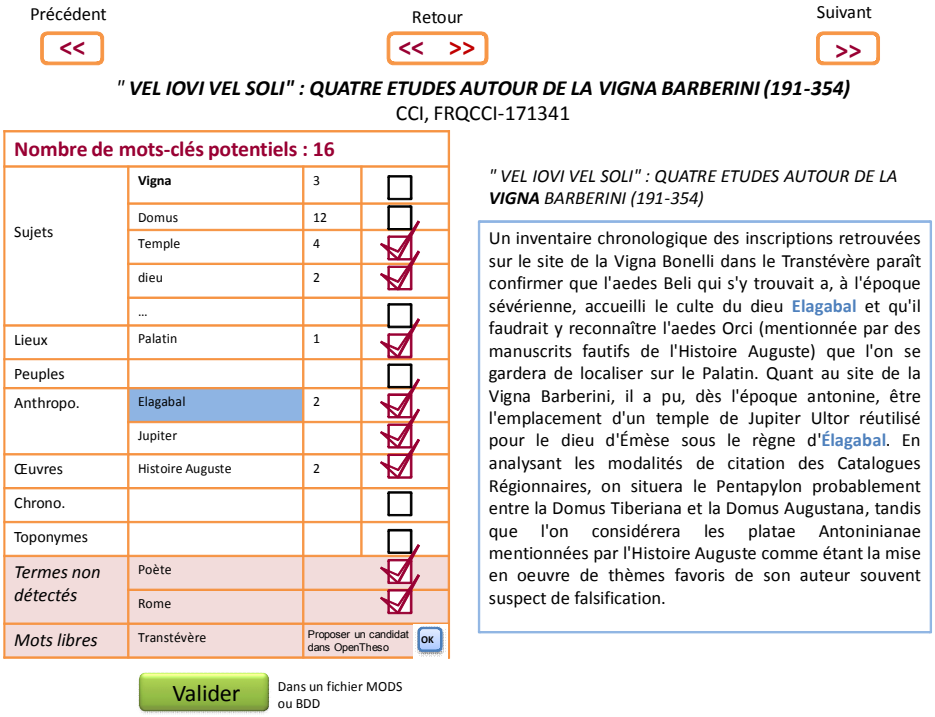
Nombre de documents traités : **07**
Langue des documents : **français**

Liste des documents traités		
1	titre + origine de la numérisation (CCI, revues.org, Persée, HAL-SHS…) + ID ou DOI	Voir
2	titre + origine de la numérisation (CCI, revues.org, Persée, HAL-SHS…) + ID ou DOI	Voir
3	***" VEL IOVI VEL SOLI" : QUATRE ETUDES AUTOUR DE LA VIGNA BARBERINI (191-354)*** CCI, FRQCCI-171341	Voir
4	titre + origine de la numérisation (CCI, revues.org, Persée, HAL-SHS…) + ID ou DOI	Voir
5	titre + origine de la numérisation (CCI, revues.org, Persée, HAL-SHS…) + ID ou DOI	Voir
6	titre + origine de la numérisation (CCI, revues.org, Persée, HAL-SHS…) + ID ou DOI	Voir
7	titre + origine de la numérisation (CCI, revues.org, Persée, HAL-SHS…) + ID ou DOI	Voir

99 MARCHEIX L., ROUSSET M., FERHOD D, Document interne au service FRANTIQ n° FRQ-Doc 09-10.- Projet d'interface pour l'assistance à l'indexation. 2009, 4 p.

- **Annexe 3**

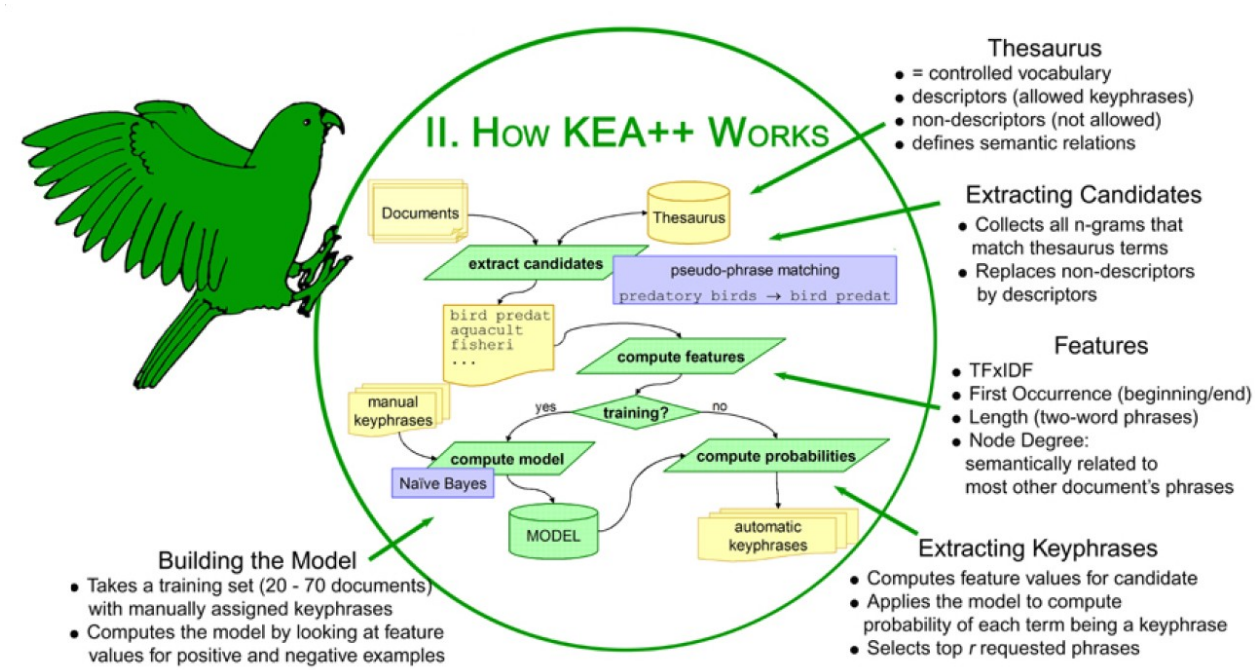
Maquette d'interface du résultat d'indexation pour une ressource textuelle du prototype d'outil d'indexation semi-automatique crée par le réseau FRANTIQ.[100]



100MARCHEIX L., ROUSSET M., FERHOD D, Document interne au service FRANTIQ n° FRQ-Doc 09-10.- Projet d'interface pour l'assistance à l'indexation. 2009, 4 p.

- **Annexe 4**

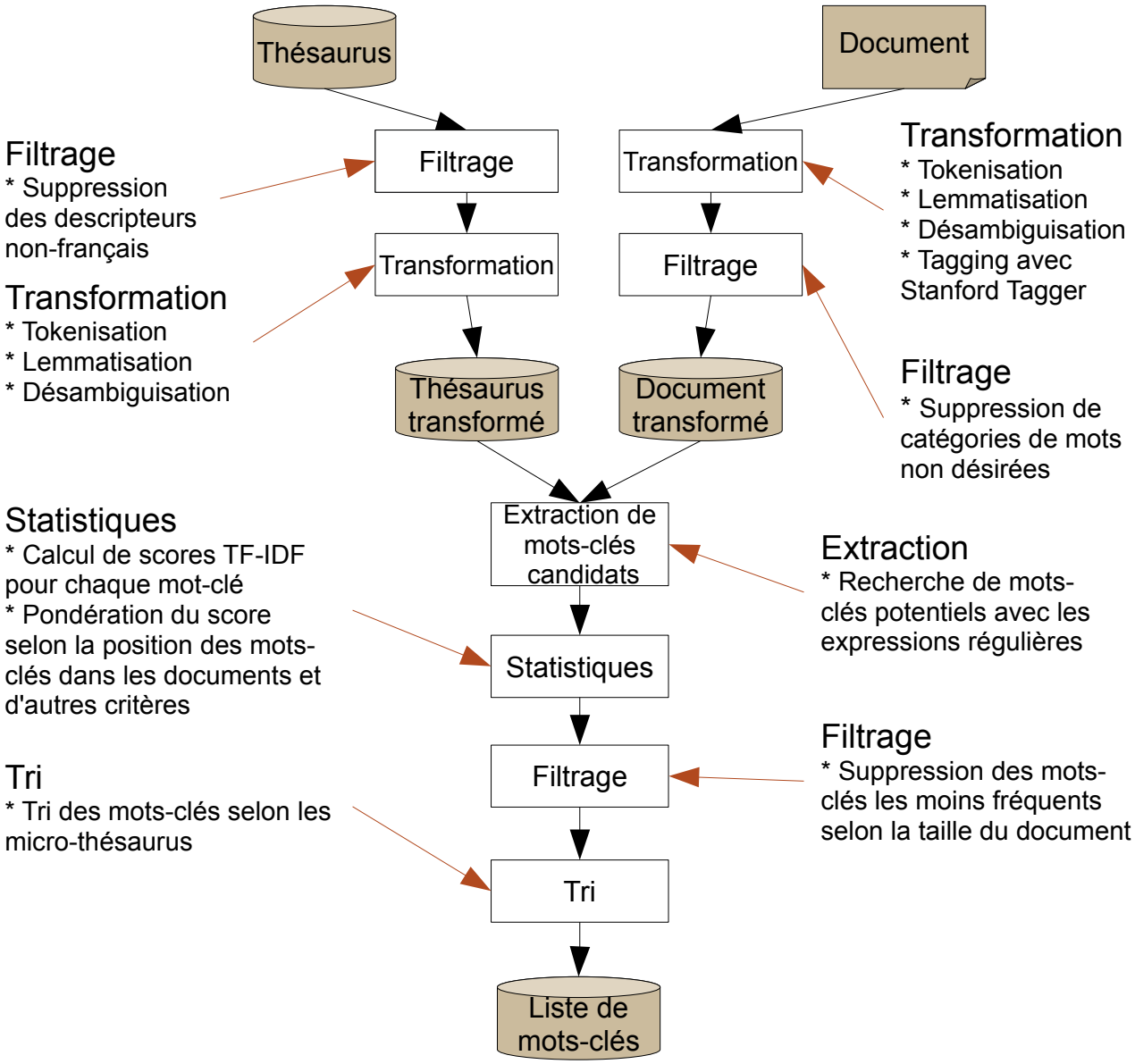
Schéma complet de fonctionnement de l'algorithme KEA++[101].



101 http://www.cs.waikato.ac.nz/~olena/publications/kea_wiml06_poster.pdf

- **Annexe 5**

Schéma de fonctionnement du programme python.



- ## Annexe 6

Résultat d'un traitement d'un texte avec KEA++.

Extracting keyphrases with options: -l testdocs/fr/pactols_test/ -m pactols_model -v pactols -f skos -e default -i fr -n 25 -t kea.stemmers.SremovalStemmer -s kea.stopwords.StopwordsFrench -d
-- Loading the Model...
-- Loading the Index...
-- Building the Vocabulary index from SKOS file
-- Extracting Keyphrases...
-- Reading instance
-- Converting instance
-- Document: TEXTEGOFFAUX
-- Keyphrases and feature values:
http://opentheso.frantiq.fr/PACTOLS/concept#16808,sanctuaire,0.038019,0.024876,23,0,1,0.931143,1,True
http://opentheso.frantiq.fr/PACTOLS/concept#17689,épigraphie,0.009577,0.159204,31,0,1,0.73043,2,True
http://opentheso.frantiq.fr/PACTOLS/concept#24765,'Vienne R',0.016934,0.049751,14,0,2,0.702889,3,True
http://opentheso.frantiq.fr/PACTOLS/concept#11107,Tours,0.011144,0.034826,15,0,1,0.668815,4,True
http://opentheso.frantiq.fr/PACTOLS/concept#11406,Vendeuvre-du-Poitou,0.051315,0.044776,2,0,1,0.623338,5,True
http://opentheso.frantiq.fr/PACTOLS/concept#17515,vicus,0.023321,0.144279,10,0,1,0.602789,6,True
http://opentheso.frantiq.fr/PACTOLS/concept#21561,Mirande,0.046232,0.039801,0,0,1,0.537025,7,True
http://opentheso.frantiq.fr/PACTOLS/concept#16599,religion,0.012075,0.149254,6,0,1,0.387077,8,True
http://opentheso.frantiq.fr/PACTOLS/concept#3998,Aquitaine,0.010355,0.134328,10,0,1,0.387077,9,True
http://opentheso.frantiq.fr/PACTOLS/concept#15577,marbre,0.008312,0.199005,12,0,1,0.387077,10,True
http://opentheso.frantiq.fr/PACTOLS/concept#12880,Pictaves,0.059002,0.074627,0,0,1,0.361219,11,True
http://opentheso.frantiq.fr/PACTOLS/concept#14528,dédicace,0.011761,0.517413,5,0,1,0.187689,12,True
http://opentheso.frantiq.fr/PACTOLS/concept#16282,plaque,0.007829,0.189055,1,0,1,0.146817,13,True
http://opentheso.frantiq.fr/PACTOLS/concept#300,Apollon,0.014914,0.975124,4,0,1,0.126176,14,True
http://opentheso.frantiq.fr/PACTOLS/concept#16396,princeps,0.020192,0.41791,0,0,1,0.123803,15,False
http://opentheso.frantiq.fr/PACTOLS/concept#15940,objet,0.002583,0.348259,7,0,1,0.112793,16,True
http://opentheso.frantiq.fr/PACTOLS/concept#17374,travail,0.005178,0.368159,1,0,1,0.076785,17,False
http://opentheso.frantiq.fr/PACTOLS/concept#14811,fouille,0.00173,0.283582,5,0,1,0.066048,18,True
http://opentheso.frantiq.fr/PACTOLS/concept#1482,Jean,0.009079,0.109453,0,0,1,0.055535,19,False
http://opentheso.frantiq.fr/PACTOLS/concept#16094,'panthéon (édifice)',0.017993,0.99005,0,0,2,0.05489,20,False
http://opentheso.frantiq.fr/PACTOLS/concept#13905,cité-ville,0.005098,0.830846,1,0,1,0.033056,21,True
http://opentheso.frantiq.fr/PACTOLS/concept#15045,histoire,0.004689,0.781095,1,0,1,0.033056,22,False
http://opentheso.frantiq.fr/PACTOLS/concept#9829,Rome,0.006793,0.124378,0,0,1,0.027634,23,False
http://opentheso.frantiq.fr/PACTOLS/concept#14371,dieu,0.011097,0.970149,0,0,1,0.023599,24,False
http://opentheso.frantiq.fr/PACTOLS/concept#13230,années,0.003535,0.318408,0,0,1,0.018363,25,False
-- 17.0 correct

- **Annexe 7**

Diagramme de séquence d'une interaction grâce à une interface web d'un utilisateur et le prototype d'indexation semi-automatique en python.