Analyse statistique lexicale d'articles pour la construction d'une base de connaissances

Nicolas FAURE
Université Lyon 3

Jean-François THIMUS
Université catholique de Louvain, Belgique

René-Michel FAURE
ENTPE

Le nombre d'articles issus des congrès de tunnels ou de mécanique des sols augmente chaque année. Afin de rester maître de cet afflux, un nouvel outil (MKD[1]) a été mis au point en utilisant les techniques innovantes d'analyse du langage et du management des connaissances pour gérer facilement, rapidement et efficacement la connaissance contenue dans ces articles. Une base de données, accessible via internet, est construite automatiquement en analysant titre, auteurs, bibliographies et texte de chaque article. Un lexique est créé avec les mots ou groupes de mots définis par des patrons morphosyntaxiques*.
Une approche bibliographique (qui écrit quoi, qui cite qui et qui est cité par qui, etc…) est proposée. Le lexique permet l'indexation des textes pour les classer suivant un indice de ressemblance ou les regrouper par thèmes en fonction de leur contenu.
Cette thématisation permet d'aborder le recueil des connaissances et de son utilisation à l'aide d'une ontologie construite à partir des textes analysés et d'un composant nommé granule de connaissance. L'usage de MKD pour la veille technologique, l'aide à l'enseignement ou au projet est brièvement présenté.
L'article présente les nombreuses possibilités offertes par MKD en s'appuyant, par exemple, sur le traitement des 2544 articles des dix derniers congrès ITA/AITES.

English abstract

For managing numerous papers a new tool, called MKD[1], is presented. It uses technics from language processing and ontological approach. All titles, authors, references are collected in an ad-hoc format to feed the data base. Using a morphosyntaxic approach all syntagms (words and set of words) are extracted and sorted following an id-itf score as to propose lexicon representing the topics of the set of papers.
A loglikelihood measure is proposed as to compare papers and sort them following the similarity with a given paper.
Knowledge management is done with the help of a specific ontology and a new feature called grain of knowledge. The use of MKD for technology watch, teaching help and computer aided design is briefly presented.
This paper shows some possibilities of MKD using all papers (2544) from the ten last congresses ITA/AITES; some results are presented.

1 - Introduction

Dans cet article nous présentons les premiers résultats de l'outil MKD (Modélisation de la Connaissance par Domaine), outil élaboré pour répondre aux objectifs suivants :

- maîtriser le flux documentaire et scientifique abondant, qui devient de plus en plus envahissant et difficilement exploitable,
- utiliser facilement l'expertise 'mondiale' contenue dans les documents,
- maintenir à jour la base de connaissance, support de l'application.

Pour réussir ces challenges, la prise en compte des contraintes suivantes est nécessaire :

- proposer une gestion directe par les experts du domaine afin de conserver le contexte d'emploi des notions et des concepts,
- accepter les contradictions, les divergences théoriques. En effet, d'un

1 - MKD : Modelising Knowledge by Domain ou Modèlisation de la Connaissance par Domaine.

2 - Magnan (Magnan, 2007) nous rappelle la dimension « artisanale » du projet « géotechnique »

contexte à l'autre, des vérités peuvent être remises en cause ce qui impose à MKD une contextualisation des faits énoncés[2].

MKD doit donc être un outil facile d'emploi, collectif, capable de gérer efficacement et rapidement la connaissance contenue dans de nombreux documents et de la restituer précisément en fonction du contexte affiné par chaque question de l'utilisateur.

2 - Historique

MKD est l'aboutissement d'une suite de projets orientés *«management des connaissances»* et rappelés succinctement ci-après.

Le projet XPENT (France - Québec) Système expert en stabilité des pentes (1988 - 1992) a montré le peu d'efficacité de l'interview d'expert et la nécessité d'une base de cas. (Faure et al, 1988,1992).

Cette constatation a donné lieu de 1995 à 2001 au projet WASSS (World Area Slope Stability Server, développé au sein du TC 11 de l'ICSMFE) dont les principes ont été repris par l'USGS pour la construction de leur base de cas. (Faure et al.1999)

De 2003 à 2009, le projet RAMCESH du CETu *«Recueil Assisté et Maniement de la Connaissance des Espaces souterrains Habités»* (Faure et al.2006, 2007) a permis la définition du granule de connaissance (Faure N., 2004, 2006, 2007) et l'application CALFAT (réalisée pour la Drast à l'Inria) a permis de tester des mesures de ressemblance entre textes. (Faure et al. 2008)

Les travaux récents ont porté sur la partie gestion des documents et mise au point d'une interface dédiée au spécialiste du domaine, ici le géotechnicien.

3 - Présentation de MKD

Le prototype de MKD est un ensemble de fonctions activées par un ingénieur du domaine grâce à une interface très conviviale. Ces fonctions sont regroupées en cinq parties : mise en forme des données, élaboration d'un lexique, traitement de la connaissance, bibliométrie, maniement de la connaissance. Les deux dernières parties correspondent à l'interrogation de la base de connaissance créée.

3.1 - Données bibliométriques*

MKD met en œuvre :

- une interface pour le dialogue homme machine,
- une base de connaissance qui stocke toute l'information,
- des modules de traitement de fichiers intermédiaires pour la mise en forme des données et leur vérification.

Les données de départ sont des documents sous forme de fichiers au format pdf. Ce format est très utilisé et tous les congrès (dont ceux d'ITA/AITES) sont sous cette forme. MKD gère des ensembles d'articles (EArt) composés, soit d'articles d'une source unique (cas des congrès), soit d'articles de provenances différentes, d'une collection personnelle par exemple. La taille de ces EArt est de quelques centaines d'articles.

Les sommaires de congrès, les bibliographies de thèse ou de documents de synthèse sont aussi des documents gérés par MKD pour augmenter sa base bibliographique.

Les fichiers au format pdf sont placés dans un répertoire et vont être transformés au format txt, format plus facile à manier. Ils forment un nouveau répertoire.

Le traitement de ce nouveau répertoire conduit à deux fichiers, S1 et S2. Le premier (S1) contient l'entête des fichiers, à savoir : nom des fichiers des articles, titre de l'article, auteur(s), adresses des auteurs si elles existent, langue (français ou anglais) du texte automatiquement déterminée, année et référence (dans ce cas le nom du congrès). Le second (S2) contient le nom des fichiers des articles et l'ensemble des références bibliographiques de chaque article.

Les fichiers S1 et S2 sont vérifiés visuellement, un module de MKD mettant en évidence par la couleur les anomalies qui sont alors corrigées par un lecteur.

La vérification du fichier S2 qui contient toutes les références citées dans les bibliographies s'impose car de nombreuses inexactitudes sont mises en évidence et proviennent principalement du non-respect par les auteurs des règles d'édition demandées par les organisateurs du congrès. Le copier-coller augmente ce non-respect, car d'une référence à l'autre, provenant de deux articles d'origine différente, les règles d'écriture sont souvent différentes. Une relecture attentive par les auteurs n'est donc pas toujours faite. A cela s'ajoutent des références souvent invérifiables car impubliées, des références incomplètes, des *« communications personnelles »* ou des citations de quatrième main (*« cité par X dans l'article Y, qui citait lui-même Z dans l'article W, … »*). Ces références sont alors rejetées vu leur peu d'intérêt. De même les références comportant la locution latine *« et al »*, les mémoires d'élèves, les ouvrages non diffusables, les listes de livres bien connus, etc… sont rejetés. L'information attachée au nom de chaque auteur est très variable. Doit-on trouver la liste de leurs titres, (Docteur, Président, etc…) et/ou la hiérarchie complète de leurs postes, du ministère au laboratoire ?[3]

Lors de l'introduction en base des références (traitement des fichiers S1 et S2), MKD vérifie si elles sont déjà dans la base de façon à avoir des références uniques. Cependant, un lien est gardé entre la référence citant et la référence citée, ce qui permettra de savoir qui cite qui et qui est cité par qui. Les titres des références peuvent présenter des différences typographiques mineures d'une référence à l'autre ; une fonction de MKD permet à l'utilisateur de retrouver et éventuellement fusionner les données qui sont attachées à deux titres ressemblants, jugés identiques.

3.2 - Usage des données bibliométriques

MKD permet de regrouper dans des LRA* (Liste de Références d'Ar-

3 - Que penser aussi, quand un article de 6 pages possède 25 auteurs ? (trouvé dans un congrès international).

ticles), les références répondant à des critères communs. (articles de tel auteur, ayant tel(s) mot(s) dans le titre, de tel congrès, de telle année, de tel Eart (ensemble d'articles), etc…). Ces LRA peuvent être combinées (somme, différence) pour des interrogations plus élaborées (auteurs de cette LRA, articles cités par les articles de cette LRA, édition de cette LRA, etc…).
Cette approche autorise des recherches par corpus constitué, permettant des requêtes précises pour des contextes particuliers.

3.3 - Données lexicales*

A partir du répertoire des fichiers lisibles (format txt) un module de MKD effectue une analyse morphosyntaxique pour retrouver dans chaque fichier, les concepts du texte. Ces concepts sont représentés par des mots ou des ensemble de mots, jusqu'à 5 mots, qui sont détectés à l'aide de patrons*. L'analyse morphosyntaxique détermine d'abord le rôle syntaxique de chacun des mots du texte, c'est-à-dire si ce mot est un nom (N), une préposition (P), un adjectif (A), un verbe, etc…Puis dans une seconde lecture, elle recherche les groupes de mots définis par les patrons. Pour MKD, nous avons choisi les patrons nominaux suivants, à savoir N, NA, AN, NPN, NPAN, NPNA, NPNPN, Nom Propre, et pour chaque groupe de mots correspondant à un patron, l'analyse compte le nombre de fois que ce groupe de mots, appelé syntagme*, apparaît dans l'article [4].
Pour chaque fichier du répertoire une liste de syntagmes est ainsi créée, comportant pour chaque syntagme son nombre d'occurrences et son patron morphosyntaxique. Les résultats de l'analyse morphosyntaxique forment les fichiers S3.
Avec ces syntagmes un lexique est alors établi pour chaque EArt. Chaque syntagme qui apparaît plus de deux fois [5] dans la liste des syntagmes d'un article, est sélectionné dans le lexique qui correspond à l'EArt que l'on traite. Pour chaque EArt (par exemple un congrès) on obtient un lexique dont le nombre de syntagmes est assez important. MKD permet de comparer ces lexiques et surtout de les additionner dans un lexique plus vaste associé à un ensemble d'EArt. Par exemple, pour les dix congrès ITA/AITES (de 2004 à 2013) le nombre de syntagmes détectés approche 200000. La construction des lexiques de chaque EArt permet d'attacher à chaque syntagme une donnée importante : le nombre d'articles dans lesquels il est utilisé. Cette donnée va permettre, combinée avec l'occurrence de chaque syntagme, d'ordonner par *« usage décroissant »* tous les syntagmes du lexique. Cette probabilité d'usage calculée s'appelle le tf-idf* [6] et par simple seuillage, on retient dans le lexique attaché au domaine couvert par l'ensemble des EArt, les "n" syntagmes les plus employés [7]. Ce nombre "n" est fourni par l'utilisateur. Plus "n" est grand, plus les réponses aux questions seront précises. L'usage montre qu'un lexique d'environ 2000 syntagmes est assez bien adapté aux recherches de MKD. Cependant la sauvegarde des syntagmes les moins utilisés permettra, à la demande, de les ajouter au lexique si le besoin existe [8].
La détermination d'un lexique va permettre l'indexation automatique de chaque texte, c'est-à-dire lui affecter un certain nombre de *« mots-clés »* qui sont les syntagmes communs à l'article et au lexique. Nous appellerons cette liste de syntagme, la signature de l'article.

3.4 - Utilisation des données lexicales

Postulant que deux textes écrits avec les mêmes syntagmes se ressemblent, MKD permet de calculer un degré de ressemblance entre texte en rapportant en pourcentage le nombre de syntagmes communs aux deux textes (comparaison de deux signatures). Appliqué à la base de données de MKD, pour un article donné, le calcul et la présentation par degré de ressemblance décroissant permettent de savoir quels sont les articles d'une LRA* qui *« ressemblent »* le plus à l'article donné.
MKD propose aussi à l'utilisateur une thématisation des articles basée sur une partition du lexique [9]. L'utilisateur crée des sous domaines qu'il remplit grâce à une interface très conviviale qui sélectionne des syntagmes du lexique [10]. Chaque sous domaine est ainsi très précisément caractérisé et la liste des syntagmes de chaque sous domaine peut être utilisée pour une recherche de ressemblance, ce qui conduit à la thématisation des articles de la base.
Appliqué à un flux régulier de nouveaux documents dont le lexique a été soigneusement partitionné, MKD apparaît comme un outil de veille technologique signalant tout concept nouveau dans un document traitant d'un sous domaine bien identifié.

3.5 - Eléments de comparaison avec d'autres systèmes de recherche documentaire

Le tableau suivant montre la complémentarité des systèmes documentaires usuels. Le prétraitement qui extrait l'information pertinente et la stocke de façon non redondante rend MKD efficace et précis en limitant le volume stocké.

4 - Exemple de patrons morphosyntaxiques : N : Nom = glissement, NA : Nom + Adjectif = glissement lent, AN : Adjectif + Nom = grand glissement, NPN : Nom + Préposition + Nom = glissement de terrain, NPAN : Nom + Préposition + Adjectif + Nom = glissement de grande ampleur, NPNA : Nom + Préposition + Nom + Adjectif = glissement de terrain rocheux, NPNPN : Nom + Préposition + Nom + Préposition + Nom = glissement de terrain dans l'argile.

5 - Ce qui élimine les erreurs typographiques ou les fautes de frappe.

6 - C'est une fonction simple du nombre d'occurrence du syntagme (combien de fois ce syntagme est utilisé dans tous ces articles) et du nombre d'articles dans lequel est utilisé ce syntagme.

7 - La fonction «log-vraisemblance» est aussi utilisée par MKD (loglikelihood measure*).*

8 - MKD construit la base de connaissance au fil des textes, ce qui entraîne des adaptations, comme lors de la création d'un granule de connaissance (voir plus loin).

9 - Le lexique partitionné ressemble alors à un thésaurus. (Voir plus loin)

10 - Pour ces sélections, une aide est proposée par MKD, basée sur le calcul des co-occurrences du second ordre*

	Gestion électronique de documents (GED)	Système documentaire	Moteurs de recherche (Google)	MKD
Usage	Recherche de documents	Recherche et classement de documents	Recherche d'information	Recherche de connaissances par contexte, suivant causes ou conséquences. Classement des documents
Origine des documents	Ceux de l'entreprise	cf DIRR, Pascal Nécessite une organisation	Tout Internet	Ceux du domaine - dignes d'intérêt - tous si veille technologique
Mise en base	Quelques documents	Variable	Tout ce qui est publiable	Automatique par paquets choisis par les experts du domaine
Traitement lexical	Quelques mots clés	Thésaurus adapté au thématiques de la structure (travail important)	Non	Création automatique de lexiques et de thésaurus
Approche ontologique	Non	Non	Non	Oui
Bibliométrie	Faible	Bonne, réponses souvent abondantes	Non	Etendue, comme la classification par degré de ressemblance des documents à un autre
Coûts, y compris mise à jour	Logiciel + une personne	Abonnement + prix par document	Accès Internet + temps passé par chacun	Logiciel + temps partagé des experts du domaine via un Intranet
Pertinence des réponses	Limitée à l'entreprise	Bonne	Faible	Elevée avec comparaison de documents
Volume occupé	Variable	Important	Réparti, diffus, immense	Faible : MKD ne retient que des extraits de texte, ce qui l'exempte de droits d'auteur

Tableau 1 - Eléments de comparaison des systèmes documentaires.

4 - Possibilités de MKD appliquées aux congrès ITA/AITES

La base de MKD constituée après traitement des articles des congrès ITA/AITES, fournit des statistiques, une certaine analyse bibliométrique et des approches par contenu innovantes que nous présentons ci-après.

4.1 - Statistiques

Sur plusieurs années [11], on peut suivre l'évolution des congrès. Les deux tableaux ci-dessous indiquent le nombre des éléments traités. Après avoir traité les 20 fichiers (10 S1 et 10 S2) la base comporte 8588 références d'articles et 12887 auteurs identifiés. Les 10 fichiers S3 apportent 920709 syntagmes, y compris les doublons, à la base ce qui permet le calcul du lexique (tableau 2).

Les 32210 syntagmes différents sont ordonnés suivant leurs tf-idf et seuls les 1762 premiers sont retenus (tableau 3).
L'évolution du vocabulaire est sensible montrant que le domaine couvert par les journées s'élargit.

Congrès	Nom Eart	Fichiers initiaux txt Nombre de lignes								
		S1	S2	S3	Nb articles	Nb ref lues	Nb ref nouvelles	Nb auteurs lus	Nb auteurs nouveaux	Nb syntagmes lus
Singapour 2004	Ita_Singa	1925	3434	71123	173	722	595	1501	928	69023
Istanbul 2005	Ita_Ist	2469	4578	79542	207	894	787	1912	992	78101
Séoul 2006	Ita_Seoul	2286	10318 (**)	69316	198	1070	1015	2289	1776	68305
Prague 2007	Ita_Prague	3443	6548	58648	305	1411	1091	2840	1272	57487
Aggra 2008	Ita_Aggra	2127	4497	97268	201	854	747	1844	1051	95929
Budapest 2009	Ita_Buda	3044	5833	108219	266	1195	865	2710	1138	106617
Vancouver 2010	Ita_Vanc	2259	4774	103034	207	888	671	1980	779	101758
Helsinski 2011 (*)	Ita_HLK	1491	3564	87289	134	662	531	1476	672	86183
Bangkok 2012	Ita_BGK	4013	7905	140781	355	1465	1195	3275	2351	138862
Genève 2013	Ita_Gen	3436	7366	120095	298	1404	1091	3228	1928	118444
Totaux		26493		935315	2344	10565	8588	23055	12887	920709

Tableau 2.
(*) du fait d'un fichier pdf unique pour tout le congrès, tous les fichiers d'articles n'ont pu être récupérés
(**) avec l'abstract

11 - Depuis 2004, congrès de Singapour, premier congrès avec des actes sous forme numérique. Pour les congrès plus anciens ou d'autres documents, une scannerisation sera nécessaire préalablement au traitement par MKD.

Congrès	Nb syntagmes lus	Nb de syntagmes retenus	Nb syntagmes sommés	Nb syntagmes nouveaux
Singapour 2004	69023	5669	5669	5669
Istanbul 2005	78101	5444	8802	3133
Séoul 2006	68305	5861	11925	3123
Prague 2007	57487	4537	13799	1874
Aggra 2008	95929	6648	16857	3058
Budapest 2009	106617	7286	20127	3270
Vancouver 2010	101758	6668	22693	2566
Helsinski 2011	86183	7244	25594	2901
Bangkok 2012	138862	9119	29413	3819
Genève 2013	118444	7730	32210	2797
Totaux	920709	66206		32210

Tableau 3.

4.2 - La recherche bibliographique.

Elle commence par des tris sur les mots du titre et des noms des auteurs pour obtenir une LRA (Copie d'écran 1). En combinant les LRA* obtenues (Copie d'écran 2), on obtient d'autres LRA, que l'on affine ou non, pour éditer les articles, rechercher les articles ou les auteurs cités par les articles de la LRA.

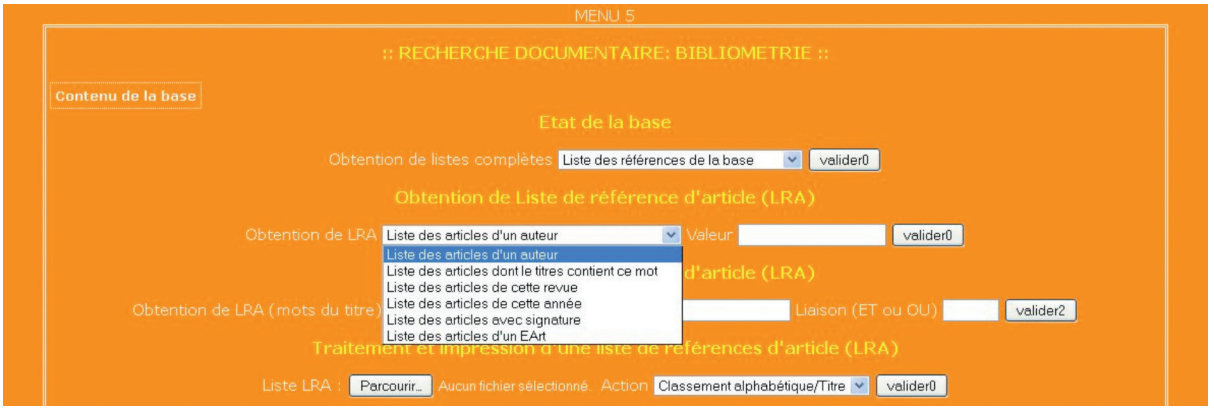
Une fonction de MKD, permet de retrouver les articles qui citent ceux d'une LRA. Par exemple en prenant les LRA de deux congrès on obtient la liste des articles du congrès 1 cités dans le congrès 2, ce qui permet d'évaluer l'impact d'un congrès sur un autre.

Ces fonctions permettent de mettre en évidence des communautés d'intérêt, ceux qui se citent entre eux. Dans le cas des congrès ITA/AITES, on a remarqué la faible interpénétration des références et les communautés s'identifient à quelques pays leaders. Les références externes ne sont que quelques ouvrages fondamentaux.

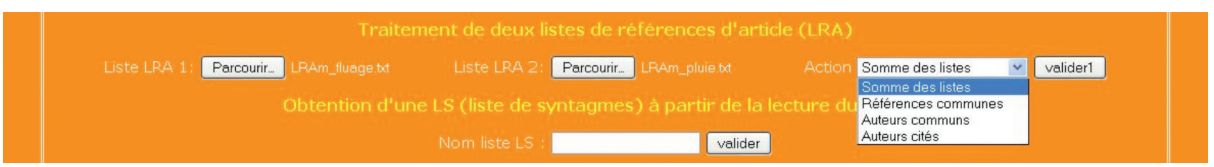
Le tableau 4 liste une LRA, très courte pour des raisons de place, réponse à la question quels sont les articles ayant les mots 'segmental' ET 'joint' dans le titre. L'index est celui de MKD et le titre celui interprété par MKD lors de l'analyse textuelle des fichiers de chaque article. Seuls trois articles du congrès de Bangkok répondent à la question. L'adjectif 'segmental' est peu usité, est-il une mode ou une tendance locale ? Parfois apparaît un article cité deux fois, du fait d'une écriture différente des titres. MKD permet de retrouver ces titres semblables et laisse à l'utilisateur la décision de fusionner les deux titres. Cela illustre encore la nécessité du respect des consignes d'édition.

Index MKD	Titre de l'article	Référence EArt
1758	Developments in Joint Design for Steel Fibre Reinforced Concrete Segmental Tunnel Linings	Ita_BGK
2000	Deformation analysis of a tunnel with concrete segmental lining subjected to ground surface loading using novel joint model *	Ita_BGK
8646	Influence of segmental joints on tunnel lining *	Ita_BGK

Tableau 4 - exemple de LRA (Liste de Références d'Article).



Copie d'écran 1 - exemple de questions prédéfinies.



Copie d'écran 2 - exemple de traitements de MKD.

MKD détermine les articles les plus cités, ce qui permet d'aborder la complétude de la base. En effet, la base de connaissance bâtie avec des articles numérisés, ne contient pas les articles du passé. Connaître ceux dont l'intérêt est attesté par les nombreuses citations implique de numériser ceux-ci, et seulement ceux-ci, pour les ajouter à la base et en traiter le texte.
Dans la base MKD, deux genres d'articles cohabitent, ceux dont les textes sont connus, et ceux, issus des bibliographies dont seule la référence est connue. L'approche lexicale du chapitre suivant, n'est donc possible qu'avec les articles du premier genre.
A partir de l'index de l'article, MKD propose des fonctions, comme l'édition complète de sa référence ou l'obtention de sa signature sous forme de liste de syntagmes (LS*).

4.3 - L'approche lexicale*

Si l'approche bibliographique décrite ci-avant utilise les données des références des articles, l'approche lexicale développée ci-après utilise les mots et groupes de mots, les syntagmes*, du texte des articles, ce qui va permettre des recherches sur les concepts, désignés par les syntagmes, utilisés par l'auteur.
Un lexique pour chacun des congrès est construit à partir de l'analyse lexicale de chaque article du congrès. Puis ces lexiques sont rassemblés dans un lexique général et le tf-idf* de chacun des syntagmes est calculé, ce qui permet d'ordonner ceux-ci et de retenir comme lexique actif les premiers syntagmes de la liste (ici 1762). Les autres syntagmes sont sauvegardés et pourront, à la demande, faire partie du lexique actif.
Les syntagmes du lexique actif sont répartis en sous domaines [12]. Une première classification est faite avec des sous domaines comme : Geology, Water, Soils, Risk, Works, Tunnel, Slopes, Rock, etc… Cette liste est appelée à évoluer mais se stabilisera très vite avec l'usage et correspondra à une approche plus élaborée du domaine. C'est pourquoi ce travail doit être collégial. Le prototype actuel de MKD confirme toutes ces possibilités et une base de connaissance construite avec plusieurs dizaines de congrès géotechniques est en cours d'élaboration.
La détermination du lexique et de sa partition permet le maniement du contenu des articles grâce aux notions de Signature* et de Liste de syntagmes* (LS).
Chaque article possède un ensemble de syntagmes détectés par l'analyse morphosyntaxique. Cet ensemble peut être assez grand et contient des syntagmes peu porteurs de sens dans le domaine géotechnique ou géologique. La signature* réduit cet ensemble en ne gardant que les syntagmes qui appartiennent à l'article et au lexique. La comparaison de ces signatures permet de définir facilement un indice de ressemblance entre deux textes et de les ordonner suivant cet indice.
Par exemple, pour un participant qui a écrit un article dans un congrès le plus simple est de comparer son article à ceux du congrès et de lire en priorité les plus ressemblants. Il connaîtra très vite l'état de l'art du micro domaine que sous-tend son article.
Pour être plus pertinent (précis), on peut réduire encore la taille de cette signature en ne gardant que les syntagmes appartenant à un ou plusieurs sous domaines. Ceci est fait dynamiquement par MKD.
MKD permet aussi de construire des LS, à partir d'un article, d'un ensemble d'articles défini par une LRA, d'un sous thème défini dans le lexique. Des opérations d'addition, de soustraction entre listes, de sélection dans une liste permettent, sans écrire un seul syntagme, d'obtenir un ensemble de syntagmes représentant une idée, un concept, une théorie, un ouvrage très particulier, etc…
Un module de MKD propose une comparaison croisée des articles d'une LRA, chaque article étant comparé aux autres articles (ce qui fait n x (n-1)/2 comparaisons pour n articles). Le classement par degré de ressemblance permet d'identifier les articles qui se ressemblent le plus.
A partir d'une LS issue d'un sous thème, la même comparaison triera les articles les plus centrés sur le sous thème, sorte de thématisation dynamique. Le tableau 5 liste les syntagmes du sous thème *« tunnel »* pour accéder par exemple à une thématisation de l'ensemble des articles d'une LRA.

bolt _ burst _ bypass _ canopy _ ceiling _ chamber _ confinement _ convergence _ conveyance _ conveyor _ cooling _ crown _ culvert _ curtain _ curvature _ cut _ cutter _ cutting _ cycle _ cylinder _ deck _ depth _ detonation _ diameter _ diaphragm _ disc _ disk _ door _ dowel _ drift _ driver _ duct _ dust _ edge _ entrance _ entry _ escalator _ exit _ face _ fan _ floor _ front _ frost _ gallery _ gasket _ gate _ heading _ headrace _ heat _ heating _ hole _ inclination _ ingress _ inlet _ input _ inrush _ inside _ intake _ intersection _ interval _ intrusion _ invert _ issue _ jack _ joint _ junction _ liner _ lining _ lock _ metro _ muck _ niche _ opening _ outcome _ outlet _ outside _ packer _ pair _ panel _ paste _ pattern _ pavement _ pin _ pit _ plug _ portal _ portion _ post _ radius _ rib _ ridge _ rig _ ring _ rise _ roof _ segment _ shaft _ sheet _ shell _ shield _ shoulder _ side _ sidewall _ sign _ skin _ slab _ slate _ slot _ spot _ station _ strike _ strip _ stroke _ subway _ sump _ support _ traffic _ tube _ tunnel _ umbrella _ vault _ vehicle _ ventilation _ wear_ bored tunnel _ circular tunnel _ concrete lining _ concrete segment _ concrete slab _ conventional method _ deep tunnel _ existing tunnel _ final lining _ first tunnel _ fresh air _ full face _ initial support _ inner diameter _ inner lining _ internal diameter _ joint set _ large diameter _ large tunnel _ long tunnel _ longitudinal direction _ longitudinal section _ main tunnel _ mixed face _ new tunnel _ open mode _ primary lining _ primary support _ second tunnel _ secondary lining _ segmental lining _ shallow tunnel _ single tube _ tail void _ twin tunnel _ vertical shaft

Tableau 5 : liste brute des 150 syntagmes rangés dans le sous domaine « tunnel ».

4.4 - Analyse temporelle

Les syntagmes apparaissent dans la base suivant les articles qui sont datés. Cette indexation temporelle permet de retrouver l'évolution dans le temps du vocabulaire du domaine. MKD applique cette recherche sur une LS et une LRA. Pour chaque syntagme de la LS, il recherche les années quand ce dernier est cité dans les articles de la LRA et ne retient que la première date sur l'axe du temps. Avec une base bien plus large que celle faite avec les10 congrès ITA/AITES, on peut ainsi dater l'apparition de concept ou des effets de mode.

4.5 - Le multilinguisme. (français, anglais, et autre)

La langue de chaque article est déterminée par un score sur le titre et, pour les quelques articles où l'identification n'est pas probante, l'utilisateur intervient. Les lexiques sont construits en fonction de la langue. Un appariement des syn-

12 - Une classe particulière est nécessaire, celle qui correspond aux syntagmes de sens 'général', sans rapport avec un sous-domaine.

tagmes de deux langues différentes est alors fait par un petit groupe d'experts. Cet appariement des syntagmes étant fait, la traduction automatique des questions est possible et l'interrogation de la base est indépendante de la langue et retrouve les données cherchées dans tous les corpus. L'extension vers d'autres langues est facilement envisageable à partir d'un corpus de textes dans la nouvelle langue.

4.5 - Aide à l'enseignement

Pour illustrer les concepts évoqués dans un cours ou demander comme travail une synthèse d'articles, la recherche par thème (ou sous thème) fournit les articles les plus représentatifs, ceux qui parlent le plus du thème (Faure et al.2012). Cette forme de validation d'un cours de géotechnique est très bien accueillie par les élèves qui découvrent à cette occasion des aspects recherche ou d'innovation jamais abordés dans un cours.

4.6 - Veille technologique

Cette approche lexicale permet de mettre en œuvre une veille technologique. MKD traite un flux de document, compare les syntagmes utilisés et alerte en cas de nouveaux syntagmes. Si le flux est traité avec l'approche recueil des connaissances, présenté au chapitre suivant, c'est l'innovation qui est alors détectée.

5 - Le recueil des connaissances

Les articles se recoupent et portent souvent la même connaissance. Pour éviter cette redondance il faut donc détecter cette connaissance et ne la ranger qu'une seule fois dans la base. Deux rappels éclaireront le lecteur.

Rappel 1 : Qu'est-ce la connaissance ?

La connaissance est un élément dont la définition est souvent discutée, notamment dans sa définition par rapport aux autres éléments que sont les données et l'information.

Dans notre approche, nous considérons que les données sont agrégées au sein de l'information et que la connaissance est une information assortie d'une finalité, c'est-à-dire qui permet d'aboutir à un état envisagé par un opérateur humain ; d'où la difficulté de stocker *« de la connaissance »* de manière statique, la connaissance n'émergeant que dans le contexte défini par une tâche précise. Le savoir est stocké et MKD présentera le *« ce qu'il faut savoir sur un contexte donné »* et l'utilisateur en fera de la connaissance.

Au sein du système MKD, on peut ramener concepts et relations entre concepts à des données, ontologie et granules à de l'information. La connaissance n'émerge qu'en fonction d'une requête spécifique pilotée

Pratiquement, pour l'ingénieur la connaissance prend quatre formes.

- La nomination des concepts, une chose, une théorie, un état….,
- Les relations entre ces concepts, généralement sous forme déductive : si ceci alors cela,
- Les algorithmes ou méthodes de calcul,
- Les données : mesures et enquêtes.

MKD doit donc rassembler les concepts trouvés dans l'analyse lexicale ci-dessus, rechercher les relations entre concepts, toujours à partir des mêmes textes et ranger ces informations le plus efficacement possible. Une ontologie correspond à ce type de rangement. Ceci étant fait, des jeux de règles (rangés dans des granules de connaissance*) permettront de piloter les algorithmes avec les données nécessaires et de faire apparaître ce qu'il faut savoir en fonction d'un but : la question posée.

Rappel 2 : Qu'est-ce une ontologie.

Depuis les années 90, de nombreux travaux expliquent et présentent les ontologies comme incontournables pour la gestion de la connaissance. Il existe de nombreuses définitions des ontologies, la plus célèbre étant sans doute celle de Thomas Gruber : *« une ontologie est la spécification explicite d'une conceptualisation »*. (Gruber, 1993)

Pour le projet MKD, on se contentera d'indiquer qu'une ontologie est le recensement des concepts majeurs d'un domaine (la géotechnique, par exemple) et de la manière de les désigner, ainsi que des relations que ces concepts entretiennent entre eux dans ce domaine. Autrement dit, une ontologie est donc une vision pragmatique (dans le sens où elle est validée par son adéquation à un projet) de l'organisation des concepts d'un domaine.

L'une des limites des ontologies, en fonction des choix qui président à leur construction, est qu'il faut effectuer un arbitrage entre pertinence des concepts et relations et taille de l'ontologie – les domaines sont fréquemment très restreints pour conserver à l'ontologie une complexité raisonnable. La notion de sous domaine introduite dans MKD constitue une réponse à ce problème.

Plus simplement, une ontologie est une représentation de la connaissance correspondant à un graphe dont les nœuds sont des mots ou des syntagmes (nommant des concepts) et les arcs les relations entre ces nœuds. Les relations sont de tout type et sont très nombreuses [13] et le graphe qui en découle difficilement maniable et représentable.

Grâce à l'interface conviviale de MKD, une ontologie de domaine va être construite à l'aide de la notion de sous domaines.

5.1 - L'ontologie de MKD

MKD représente le domaine des tunnels, qui s'élargit naturellement à la géotechnique, avec une ontologie qui se compose d'une partie arborescence conceptuelle, qui sert simultanément à recenser les concepts du domaine et à les classer au sein d'une hiérarchie de subsomption, et d'une partie plus conséquemment sémantique modélisant les relations entre concepts relevées dans les documents traités. Ce sont les granules de connaissance* qui supportent cette modélisation.

La partie arborescence conceptuelle rassemble les 2500 à 5000 syntagmes mentionnés plus haut au sein d'un ensemble d'arborescences (20 à 25) composées chacune au maximum d'environ 300 syntagmes [14]. Chacune de ces arborescences concerne plus spécifiquement un sous domaine géotechnique (matériels, géolo-

13 - La relation la plus universelle est la relation de subsomption (une pelle mécanique est un engin de chantier), mais bien d'autres existent comme le conducteur conduit la pelle ou le mécanicien graisse la pelle. Il y a une quasi infinité de relations et de concepts.

14 - Le nombre 300 a été retenu pour des raisons pratiques. C'est ce que peut englober facilement le regard de l'expert sur un graphe conceptuel plan édité sur une feuille A3.

gie, tunnels, …). Les syntagmes utilisés dans une arborescence sont uniques, ils peuvent cependant se retrouver dans d'autres arborescences.

L'objectif de cette différentiation en sous domaines est double :

1. Autoriser les variations éventuelles dans les hiérarchies dues aux hétérogénéités conceptuelles entre spécialités. Par exemple, le *« sol »* d'un géologue consiste essentiellement en roche mère et est situé plusieurs mètres sous la surface, tandis que le *« sol »* d'un expert en stabilité des pentes consiste essentiellement en couches superficielles détritiques et excède rarement les 30 mètres de profondeur sous la surface. Le mot sol aura donc des significations différentes suivant le *« sous domaine »* à laquelle il appartient.
2. Partitionner l'ensemble de concepts en sous-ensembles manipulables par les utilisateurs. Non qu'un système informatique ne puisse stocker des dizaines de milliers de concepts, mais il paraît inenvisageable à un utilisateur de le faire et de les manipuler. Donc, en-dehors de tout modèle formel qui permette de classer automatiquement des nouveaux concepts, la maintenance et l'augmentation de telles arborescences est envisageable dans le cadre d'une utilisation orientée métier avec l'aide de graphes heuristiques (Jésus M., 2009)

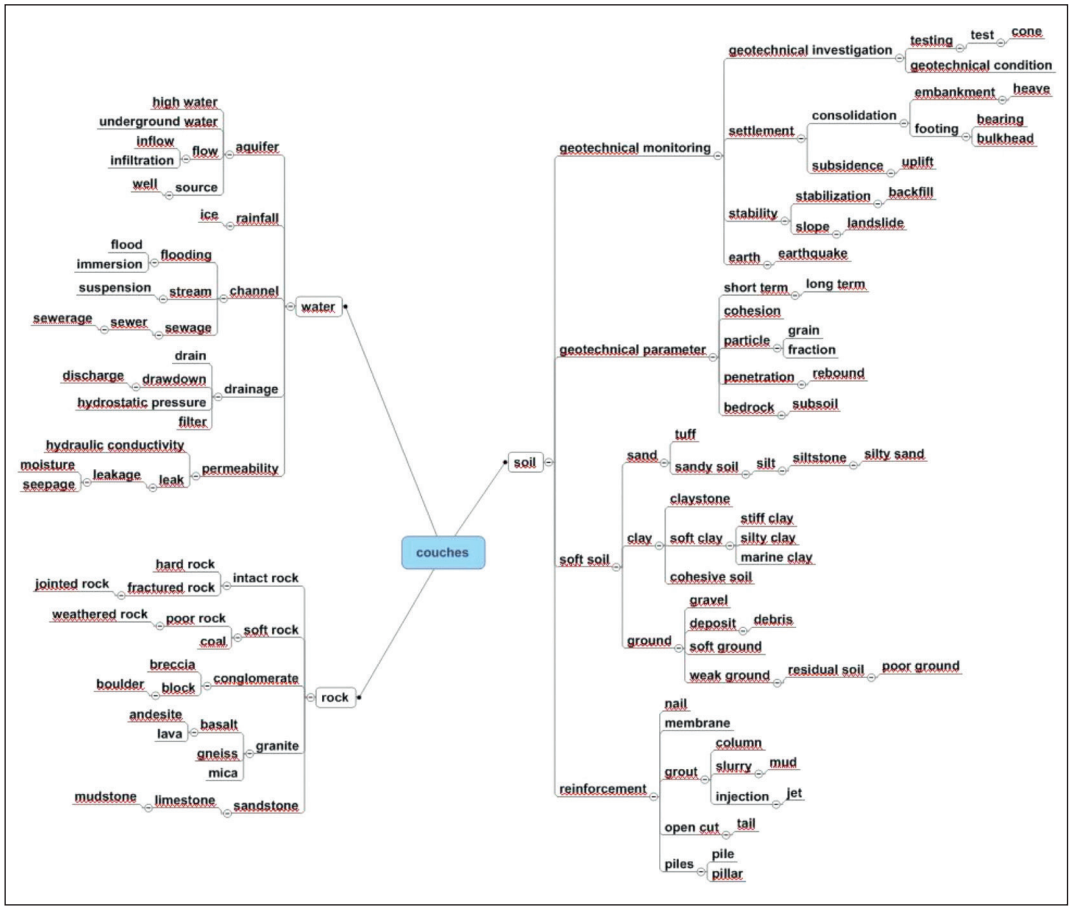
Pour construire ces arborescences locales (nommée *« couche métier »*) l'expert exporte la partition du lexique vers un outil, ici Mind Manager (copie d'écran 3) et dessine par glisser-coller le graphe. Il a la possibilité d'ajouter des syntagmes qui doivent appartenir au lexique global et n'apparaître qu'une seule fois dans le graphe, ce qui est vérifié par MKD. Si le graphe devient trop grand, il peut créer de nouvelles couches plus spécialisées ou reprendre la partition du lexique.

Le graphe ci-dessous est obtenu par exportation de trois sous domaines de la partition du lexique de 1762 syntagmes. On remarque la pauvreté terminologique de la couche *« rock »*. Ceci est dû au fait que le lexique limité à 1762 syntagmes n'a pas pris en compte bien d'autres termes de mécanique des roches du fait d'une faible représentation dans le corpus traité. Il y a plusieurs façons d'améliorer cet état de chose.

- Soit chercher dans la littérature des arborescences déjà établies pour les roches et les intégrer au lexique. Le choix n'est pas facile et sera incomplet ou surabondant du fait de nombreuses appellations locales et des révisions régulières de ces classifications, c'est cependant tout à fait envisageable.
- Soit augmenter la taille du lexique à sa création et gérer plus de termes. Mais de combien ?
- Soit ajouter des lexiques obtenus par traitement d'autres EArt.
- Soit augmenter le lexique lorsque l'usage de MKD le demande, et MKD propose cette approche lors de la mise en forme des granules de connaissance* (voir ci-après).

Une propriété importante de ces graphes est l'unicité du syntagme. Mais deux graphes différents peuvent contenir le même syntagme, ce qui va permettre au programme de construire une arborescence globale du domaine [15]. MKD utilise cette arborescence globale lors de l'élaboration des granules de connaissance et pour le glissement sémantique quand l'utilisateur veut *« élargir »* sa question. Ces graphes permettent l'usage de voisinages sémantiques définis par les arcs reliant deux syntagmes. On a ainsi les voisinages *« père-fils »* ou *« père-fils-petit fils »* par exemple.

Dans la mesure où les syntagmes sont nominaux et désignent des concepts du domaine partagés par les différentes spécialités, l'alignement des couches métier peut s'appuyer sur les nombreux concepts communs entre couches (plusieurs centaines avec le corpus actuel). Ce sont ces syntagmes qui permettent la réunion des *« petits réseaux »* correspondant aux couches métiers pour réaliser, lors de son traitement, l'ontologie globale.



Copie d'écran 3 - Exemple de portion de graphe heuristique.*

15 - On appelle cette opération « alignement des ontologies ».

Il a été évoqué ci-dessus que les concepts classés dans ces arborescences ne sont pas formellement modélisés. Habituellement, un concept est formellement défini comme une catégorie aristotélicienne implémentée dans un modèle ou un autre, c'est-à-dire par un ensemble de propriétés considérées collectivement comme essentielles [16] au concept. Cette modélisation du concept a fait ses preuves pour le calcul automatique, mais l'on sait depuis les travaux de Wittgenstein que l'esprit humain ne fonctionne pas de cette manière, et les experts géotechniciens jugent très artificiel le fait de tenter de définir ainsi les concepts qu'ils emploient. (Wittgenstein, 1993).
Pratiquement, l'approche de MKD repose sur le fait que la représentation des connaissances issues de textes s'apparente plus à représenter une connaissance communicationnelle que langagière, et s'inscrit donc dans une semiosis secondaire dynamique [17], réactualisée au fil du discours. (Chingareva-Slavine, 2003).
C'est pourquoi, plus que la représentation formelle d'un concept, on s'attache à représenter les relations entre concepts (tels que désignés par des syntagmes nominaux) dans un fragment de texte. Le formalisme utilisé pour cette représentation dérive de la DRT (Discourse Representation Theory) et introduit un ensemble de relations entre concepts à un niveau local. (Kamp et Reyle, 1993). Il s'agit du granule de connaissance*, cité plus avant dans d'autres documents et rappelé ci-après. (Faure, 2007).
Le schéma suivant représente l'équivalence entre une ontologie classique et la structure utilisée dans MKD. Les arborescences « couches métiers » définissent le lexique du domaine et les « granules » sont des ontologies locales définies par le « fragment de texte » qui s'apparente à une règle de production. L'automatisation de l'extraction des fragments de texte et de leur transformation en granule est en cours d'amélioration. Près de 1000 fragments de texte ont été identifiés par lecture attentive des textes et un premier automate réussit la transformation ces fragments de texte en granule, à plus de 50 %. (Faure, 2008)
L'ontologie de MKD est équivalente à une ontologie classique, les couches lexicales étant liées entre elles par l'usage de syntagmes identiques. (figure 1).

5.2 - Le granule de connaissances.

Le granule de connaissance est une structure informatique (genre graphe conceptuel) élaboré pour contenir la connaissance élémentaire portée par une partie du texte, le fragment de texte.
Le cœur du granule est composé de deux parties, les prémisses et les conclusions, ce qui rappelle les règles de production des systèmes experts. La relation d'implication entre ces deux composants peut être différenciée suivant : obligation, recommandation, possibilité, négation et risque.
Prémisses et conclusions sont un ensemble de phrases reliées par des ET. Une phrase est formée d'un thème (ce dont on parle), d'un prédicat (ce qu'on en dit) avec entre ces deux composants, une relation métier*. La phrase peut ne comporter que le thème. Le thème, unique dans une phrase peut être associé à plusieurs prédicats. Thème et prédicat sont des syntagmes de l'ontologie. D'un granule, on peut faire référence à un autre granule, en s'assurant qu'une simple extension (réécriture) du premier granule ne suffit pas. La référence ne s'applique qu'aux prémisses.
Le granule comporte aussi la définition d'un univers et d'un modèle de cet univers ce qui permet d'assigner des valeurs numériques à certaines grandeurs et de faire ainsi des comparaisons. L'ensemble des concepts présents dans un granule forme cet univers du granule (appelé aussi signature) et les informations relatives aux individus afférents à ces concepts sont comprises dans ce modèle d'univers : informations numériques pour certaines grandeurs, cardinalité ou équivalence entre concepts. C'est dans cette phase d'élaboration que des syntagmes peuvent être rajoutés au lexique, qui ainsi se complète pour une traduction plus précise des fragments de texte.
D'autres éléments informatifs sont aussi attachés à chaque granule : les références des textes dont est issu le granule, le fragment de texte initial, la date de création, l'auteur, et

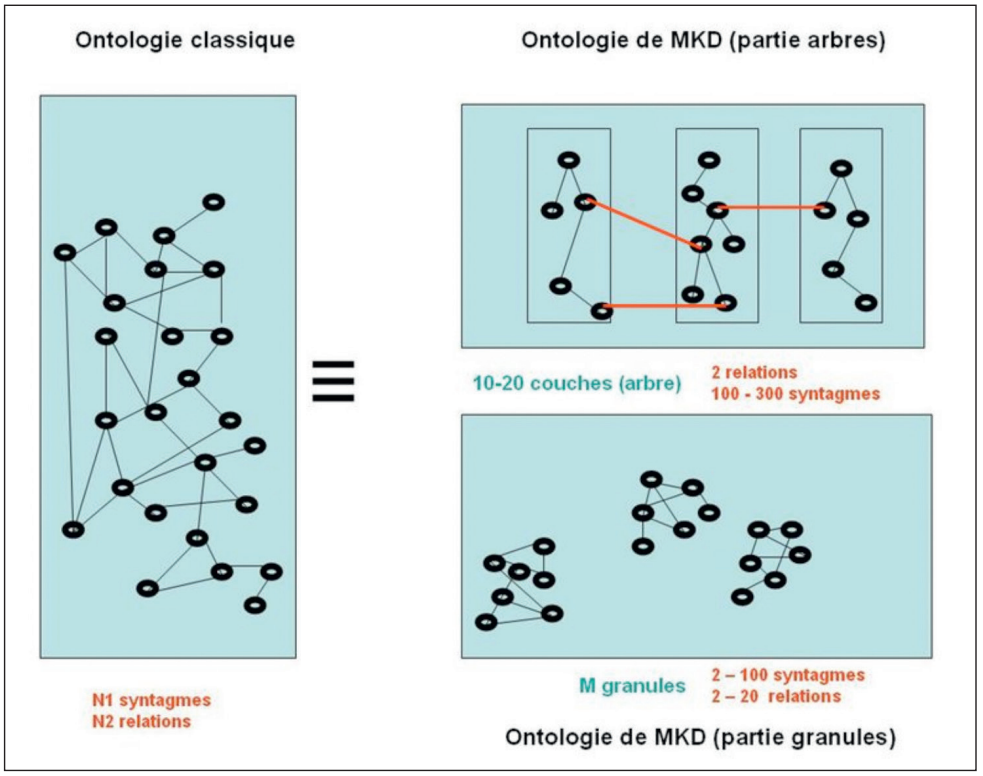


Figure 1 - Equivalence ontologie classique et ontologie MKD
N1, N2 et M sont des grands nombres, plusieurs milliers.

16 - *Par opposition aux propriétés accidentelles*

17 - *« Par semiosis, on entend habituellement le processus de dénotation des objets par des signes. Donc, l'outil principal de la sémiosis est le signe grâce auquel la semiosis se réalise : dans la langue (semiosis primaire stable) et dans le discours (semiosis secondaire dynamique) ». (Chingareva-Slavine, 2003).*

d'autres éléments qui facilitent son maniement. (Figure 2)
L'article présenté à Monaco (Faure et al.,2008) contient de nombreux exemples de granules.

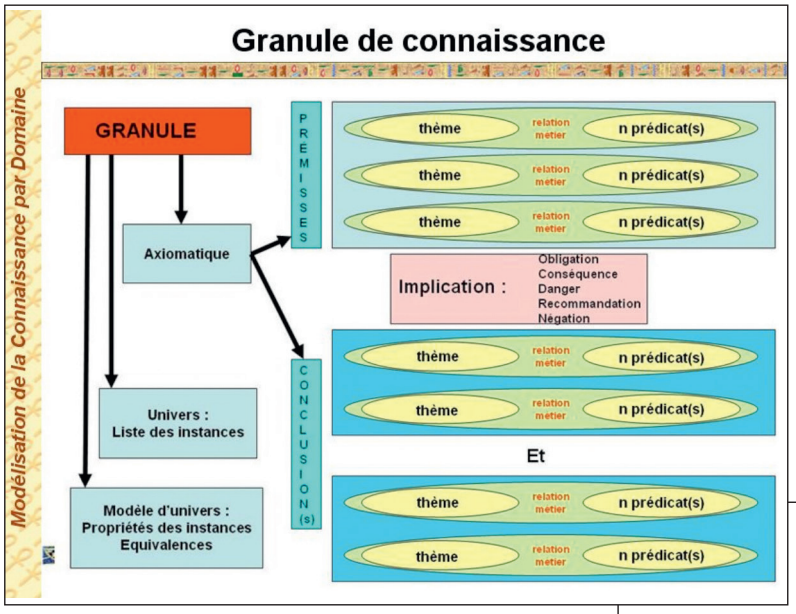


Figure 2 - Schéma du granule de connaissance.

En adoptant le granule de connaissance comme représentation de la connaissance, MKD vise à garantir une représentation unique (non redondante) de cette connaissance lors de sa formalisation. En effet, lors de la mise en forme du granule une recherche des granules voisins est faite (comparaison de signatures de granule), conduisant à un choix entre

- l'ajout du granule à la base de connaissance,
- son rejet (avec ou non addition de sa référence dans le granule le plus voisin),
- modification du granule le plus proche pour compléter prémisse ou conclusion.

Ce choix garantit l'unicité de la connaissance stockée tout en sauvegardant les références des différentes sources où elle a été trouvée.

La continuité sémantique mise en œuvre va de la possibilité de décrire un simple concept dans l'ontologie, et grâce au formalisme du granule qui est un élément fortement associatif, à celle de décrire une situation très compliquée, comme un projet d'ouvrage.

Le glissement sémantique est contenu dans l'ontologie construite sous forme d'arbres, ce qui permet de définir des voisinages et des distances sémantiques qui seront utilisés à la fois pour la construction du granule et pour l'élargissement automatique des questions.

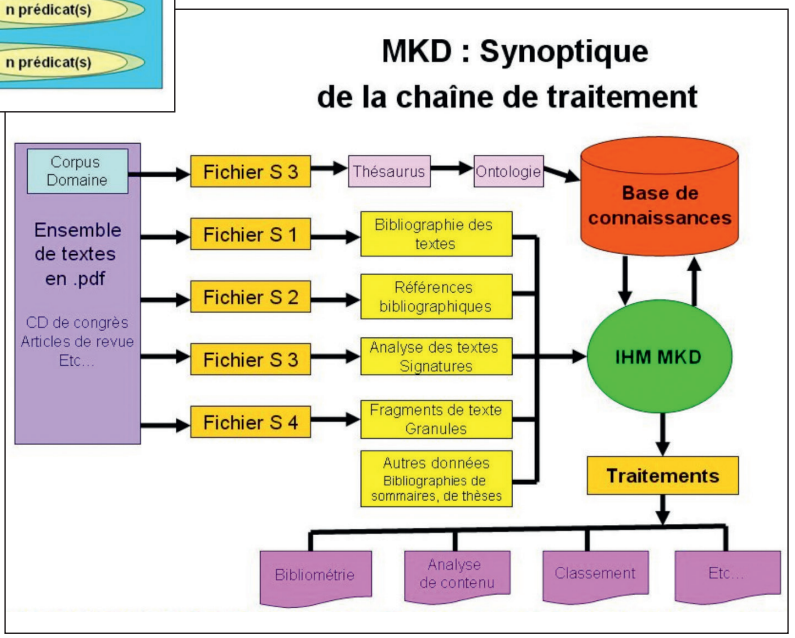
6 - La mise en œuvre

MKD peut être accessible en ligne via un Intranet ou Internet. Les parties alimentation et interrogation de la base étant distinctes, on peut donc imaginer un serveur géré par quelques personnes mettant à disposition la base de connaissance et des utilisateurs venant interroger cette base. Les modalités d'accès restent à définir.

Le traitement de nombreux textes issus de congrès ou de revues permettra de construire une base de connaissance pertinente pour de nombreux usages. MKD n'utilise que les références des textes et une extraction minimale des contenus pour être en accord avec les règles de la propriété intellectuelle.

La figure 3 représente la chaîne de traitement de MKD avec le rappel des fichiers utilisés.

Figure 3 - Schéma synoptique de MKD.



7 - Conclusion

Le management des connaissances se heurte à la grande dimensionnalité du langage, nombre de syntagmes, synonymie, polysémie, métonymie. Au commencement de cette recherche en 1990, le regretté François Tavenas s'en inquiétait, mais nous avons cru à l'augmentation continue de la puissance des machines et aujourd'hui, manier 100 000 syntagmes se fait en quelques secondes. Cependant, le chemin pour faire MKD fut long et tortueux ; finalement un prototype permet de présenter les résultats évoqués dans ce document et de montrer la puissance de l'outil.

MKD facilite grandement l'approche et la maîtrise d'un corpus de textes comme il est montré ci-dessus. Des développements se poursuivent pour automatiser et contrôler toutes les étapes décrites avec une ergonomie simple permettant à l'ingénieur métier de travailler sans recours à un spécialiste d'un autre domaine (informaticien, cogniticien, ergonome…).

De cette connaissance formalisée en granules (Faure et al., 2008), MKD qui en assure l'unicité et qui garde les diverses origines de ce savoir, peut répondre à la question: dans ce cas, que dois-je savoir et qui a dit cela? Et en plus, suivant la structure du granule, on peut interroger sur la cause ou sur les effets d'une situation.

Pour l'enseignement initial, mais surtout continué, et pour l'aide au projet cette nouvelle mise en forme de la connaissance est une avancée certaine.

Bibliographie

- *Chingareva-Slavine E., 2003, Sémiotique, linguistique et modélisation ; Hermès-Lavoisier, Paris*
- *Faure R.M., Leroueil S., Rajot J.P., Larochelle P., Sève G., Tavenas F. 1988, XPENT, système expert en stabilité des pentes. Int. Symp. On Landslides, ISL5, Bonnard Ed., Lausanne, Suisse. pp625-630*
- *Faure R.M., Mascarelli D., Vaunat J., Leroueil S., Tavenas F. 1992, Present state of development of XPENT, expert system for slopes stability problems. ISL6, Bell Ed., Christchurch, pp1671-1678*
- *Faure R-M. 1999, Data-bases and the management of landslides. Int. Symp. on Landslides. Shikoku (Japan). IS Shikoku'99, pp1317-1330*
- *Faure N. 2004, Le granule de connaissances, Actes Inforsid, Biarritz, 4p.*
- *Faure R.M., Faure N., Hémond G. 2006, The use of knowledge management in the management of tunnels and tunnel projects. ITA-AITES World Tunnel Congress. Séoul, 11p.*
- *Faure N. 2006, Représentation des connaissances géotechniques, Actes Inforsid, Hammamet, 12p.*
- *Faure R.M., Faure N., Hémond G. 2007, Recueil Assisté et Maniement des Connaissances des Espaces Souterrains Habités. Tunnels et Ouvrages Souterrains, n°202, pp351-363*
- *Faure N. 2007, Un système d'aide à la modélisation des connaissances en géotechnique. Thèse, Université Lyon 3, pp1-151*
- *Faure N. 2007, Ontologies, thesauruses and information system for tunnels. ITA-AITES World Tunnel Congress. Prague, Bartok ed. pp65-68*
- *Faure N., Faure R.M., Balasch M.A., Cottaz Y. 2008, Mise en forme de granules de connaissances à l'aide de l'outil RAMCESH, Congrès International AFTES, Monaco. Pp 505-514*
- *Faure N., Hémond G., Gress J.C., Faure R.M. 2012, Course validation using an automatic system for sorting similar technical papers. SFGE conference, Galway*
- *Gruber T., 1993, A Translation Approach to Portable Ontology Specifications, Knowledge Acquisition*
- *Hearst M., 1997, TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages, Computational Linguistics, 1997.*
- *Jésus M., 2009, Le Mind Mapping pour l'entreprise avec MindManager 8, Dunod.*
- *Kamp J., Reyle U., 1993, From discourse to logic, Kluwer*
- *Magnan J.P., 2002, L'organisation du travail en géotechnique : normalisation, développement et artisanat ; Lettre de la géotechnique (26-27), Société Internationale de la Mécanique des Sols et de la Géotechnique, 2002 (http://www.geotechnique.org)*
- *Site: www.pentes-tunnels.eu*
- *Wittgenstein L., 1953, Philosophical Investigations*

Glossaire

Les expressions du glossaire sont repérées dans le texte par un astérisque (*)

MKD : nom du projet et du prototype mis au point signifiant Modélisation de la connaissace par domaine ou Modelising Knowledge by Domain.

EArt : Ensemble d'article auquel on donne un nom. Les articles sont au format pdf. La taille d'un EArt varie de quelques centaines de pages à quelques milliers.

LRA : Liste de références d'article. Ces listes sont construites par MKD en réponse à des questions. Des opérations sur les listes sont possibles.

Analyse morphosyntaxique: détermination de la fonction grammaticale du mot (nom, adjectif, préposition, etc..) et recherche des syntagmes correspondant à des patrons morphosyntaxique.*

Patron morphosyntaxique: suite de fonctions grammaticales des mots, par exemple NPN signifiera Nom suivi d'une Préposition puis d'un Nom. Ce qui permet de définir la strcture d'un syntagme.*

Syntagme: mot ou groupe de mots signifiant un concept. MKD recherche les syntagmes d'un texte suivant des patrons.*

LS : liste de syntagmes, un ensemble de syntagmes pour définir un thème ou définir une recherche,

Graphe heuristique : construit suivant l'expérience de l'expert. Dans MKD l'expert construit un arborescence des syntagmes qui sont alors liés par les relations de subsomption (tout syntagme doit avoir un père) et éventuellement de voisinage.

Lexique : ensemble des syntagmes d'un domaine, présentés alphabétiquement.

Thésaurus ; ensemble de syntagmes regroupés en sous-domaines. Peut être représenté par un graphe.

Ontologie : forme la plus élaborée pour représenter la connaissance. Peut être représenté par un graphe tri-dimensonnel avec de multiples liaisons.

Tf-idf : fonction statistique qui permet de classer les syntagmes suivant les fréquences d'usage.

Log-vraissemblance : fonction statistique semblable à tf-idf.

Données bibliométriques : les données qui permettent de manier, classer, retrouver un article.

Données lexicales : les mots et syntagmes d'un article.

Couche lexicale : ensemble des syntagmes associés à un sous-domaine.

Fragment de texte : partie d'un texte porteuse de connaissance, une forme déductive.

Granule de connaissance : partie élémentaire de connaissance dont la forme est définie dans MKD.

Relation métier : relation utilisée dans l'ontologie et la construction du granule de connaissance qui relie deux syntagmes. Elles caractérisent le métier des gens du domaine.

Exemple de granule

Fragment de texte issu des recommandations AFTES

S'il peut exister une poussée au vide du profilé alvéolé préfabriqué en élastomère par mise en charge hydrostatique accidentelle du drain ou sous l'effet de la dépression consécutive au passage des camions, alors il est conseillé de rajouter à l'intrados du dispositif de drainage des languettes de fixation démontables en acier inoxydables de 5cm de largeur, généralement espacées de 0.33 m en piédroit, et de 0,25 m en voûte.

Granule de connaissance correspondant

Thème du Granule : Matériau

Implication : Recommandation

Prémisses

Thème	Relation métier	Prédicat
Profilé alvéolé élastomère	Préfabriqué en	Elastomère
Risque	Associé à	Poussée au vide
Poussée au vide	Associé à	Mise en charge
Mise en charge	Associé à	Dépression
Dépression	Cause	Passage de camion

Conclusion

Thème	Relation métier	Prédicat
Profilé alvéolé élastomère	Préfabriqué en	Elastomère
Risque	Associé à	Poussée au vide
Poussée au vide	Associé à	Mise en charge
Mise en charge	Associé à	Dépression
Dépression	Cause	Passage de camion

Univers : ensemble des syntagmes du granule
Modèle d'univers

Languette de fixation démontable		
Largeur	=	5 cm
Espacement	=	33 cm en piédroit et 25 cm en voûte