Expérience de création et d'utilisation d'ontologie dans un domaine d'ingénierie

Experiment and use of ontology in an engineering domain.

Nicolas Faure* --- René-Michel Faure**

**Chercheur associé,*
CR Magellan, IAE Lyon 3
mkd.onto@gmail.com

***Professeur ENTPE, Ingénieur*
Centre d'Etude des Tunnels.
25 Avenue F. Mitterrand, 69500 Bron
rm.faure@pentes-tunnels.eu

RESUME. L'article rapporte l'expérience issue de tests d'un outil d'assistance à la représentation des connaissances dans le domaine de la géotechnique. Ce domaine nécessite un tel outil en raison d'une mutation technologique et normative perpétuelle. Les acteurs du domaine souhaitant garder la maîtrise totale de l'outil, la construction de l'ontologie sur laquelle il repose a été scindée en deux parties, une partie lexicale et une partie relationnelle fondée sur le granule de connaissance. Les règles de construction du référentiel de connaissances sont simples et souples ce qui évite aux acteurs du domaine de rechercher un consensus absolu en utilisant celui issu du traitement scolaire des publications. Plusieurs grands corpus (plus de 3000 articles) ont été traités.

ABSTRACT. This paper reports the trial of the geotechnical branch of civil engineering asking for a tool for transmitting knowledge in perpetual evolution. As people from this domain ask for a total mastering, the used ontology is cut in two parts. A lexical part and a relational part using knowledge grain which solves language ambiguities. Building rules are easy to use and avoids predefined consensus. Several large corpus, more than 3000 papers, have been treated.

MOTS-CLÉS : Ontologies métiers, Conception d'ontologies, Géotechnique

KEYWORDS: Domain ontologies, Ontology building, Geotechnics

1. Introduction

La géotechnique est le domaine de la construction qui détermine la meilleure solution pour une bonne interaction entre un ouvrage et le sol, comme la fondation d'un pont ou d'un bâtiment, la réalisation d'un tunnel, …

Du fait de la difficulté à connaître l'état du sol, la construction d'un ouvrage peut rencontrer de nombreuses surprises, auxquelles l'expert géotechnicien doit faire face sans pour autant en avoir une connaissance exhaustive.

Afin de faire face à cette nouveauté récurrente, l'expert peut s'appuyer sur l'abondante littérature du domaine, pour lequel il existe de nombreux congrès et publications. La plupart des publications concernent un projet ou un contexte géotechnique précis, relatant les spécificités d'un chantier ou d'une zone géographique. Mais qui peut lire et assimiler toute cette littérature ? Ce caractère particulier des publications géotechniques illustre également le fait que la pratique géotechnique est essentiellement liée à l'acquisition de connaissances par analogie de cas, la validité de l'analogie étant soumise à l'appréciation du géotechnicien, dont la compétence et le savoir-faire sont particulièrement prégnants, plus encore que dans d'autres domaines ([Magnan, 2002] comparant même à ce propos le géotechnicien à un artisan).

Pour faciliter l'accès de l'expert à ces connaissances documentaires, une réflexion a été menée qui a conduit à développer un outil permettant de synthétiser la littérature du domaine (la fraction la plus utile à l'expert, en tous cas) en une ontologie. Cette ontologie doit pouvoir répondre à la question : « dans tel cas donné, que dois-je savoir ? » Ce qui a nécessité un mode de description de cas (le contexte) et un outil pour accéder aux bonnes réponses dans la base de connaissance structurée par l'ontologie.

Cet outil est le logiciel MKD[1], développé de façon incrémentale et dont nous présentons ici quelques éléments et résultats.

2. Constitution de la base de connaissances

2.1. *ORIENTATIONS MAJEURES DU PROJET.*

La démarche habituelle en termes de constitution d'ontologies est une démarche que nous pourrions qualifier d'onomasiologique : en s'appuyant sur un modèle de domaine décrivant formellement des concepts, on y relie les éléments lexicaux relevés dans le domaine. Cette démarche peut emprunter des formes très distinctes (en termes

[1] Model, Knowledge, Domain

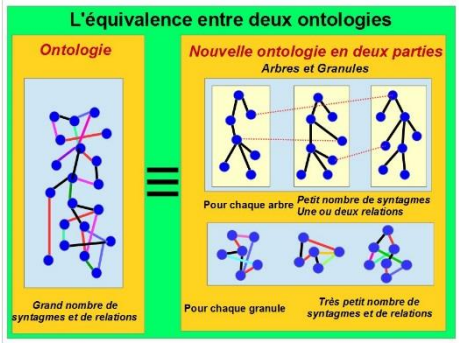

Figure 1 : Equivalence des ontologies.

de description de concepts notamment), mais elle suppose la définition formelle, le plus souvent par consensus d'experts, d'une conceptualisation explicite (la définition la plus couramment retenue de l'ontologie, celle de Grüber, est du reste formulée ainsi). Cette conceptualisation peut intervenir à différentes étapes du processus de construction d'ontologie, y compris après le traitement lexical d'un corpus documentaire surtout destiné à faciliter la mise au point du modèle conceptuel.

Nous nous positionnons différemment, dans la mesure où l'importance de l'expertise personnelle prime en géotechnique et rendrait coûteux comme peu utile un consensus préalable (forcément issu de compromis) où chaque utilisateur se verrait amené à adopter un « engagement ontologique » qui l'éloignerait de ses conceptions usuelles. En conséquence, nous adoptons une démarche de type sémasiologique, où le sens (et, au-delà, la conceptualisation) se construit pour chaque utilisateur par la combinaison d'ensembles terminologiques.

Pour cela, on distingue deux éléments constitutif de notre système : une arborescence terminologique, constituée de représentations terminologiques de sous-domaines (les cartes métier) à partir de l'analyse d'un corpus (et potentiellement d'éléments ajoutés par l'utilisateur) et un ensemble de représentations plus restreintes d'extraits de textes, le granule de connaissance, fondé sur la représentation du discours trouvé dans les textes et en partie comparable aux micro-theories (Lenat, 1995) ou aux micro-ontologies (Gruzitis et Barzdins, 2015).

Les cartes de sous-domaine rassemblent entre 300 et 600 syntagmes reliés par des relations sémantiques simples (subsomption, voisinage), les granules rassemblent jusqu'à 50 syntagmes au sein d'un formalisme reposant sur des relations sémantiques plus précises (appelées « relations métier »). L'objectif final est de permettre à chaque utilisateur de reconstruire, par agrégation de granules, une représentation de domaine plus complète et limitée à chaque projet, incluant l'actualisation d'une axiomatique simple supportée par les granules. Cette représentation de domaine est comparable à une ontologie (fig 1).

C'est ce système, composé de modules écrits en PHP et en PERL adossés à une base de données SQL afin qu'il soit implémentable dans n'importe quel service, que nous avons testé dans différents contextes géotechniques.

L'objectif premier, dont quelques résultats sont présentés ici, se situe dans le domaine de la recherche d'informations et le protocole de test retenu était fort simple : un utilisateur a une requête spécifique, liée à un cas concret, et effectue sa recherche d'information à son habitude. En parallèle, la recherche est effectuée avec MKD. Les résultats sont alors transmis à l'utilisateur, qui peut comparer ses résultats à ceux du système.

2.2. *CHOIX DU CORPUS.*

Bien que chaque groupe d'utilisateurs puisse déterminer à fin d'analyse le corpus documentaire qui correspond le mieux à ses activités, les essais ont été réalisés en se fondant sur des séries de congrès spécialisés (dans notre cas, les tunnels principalement), afin de conserver une certaine homogénéité des connaissances, ainsi qu'un certain nombre de documents considérés comme fondamentaux (normes, cours, synthèses), pour balayer l'ensemble du domaine et s'assurer a priori d'une certaine complétude. Cette approche assez informelle de la constitution du corpus correspond à une certaine réalité du terrain, les ressources documentaires utilisées par les experts du domaine étant fréquemment de ce type[2].

2.3. *DETERMINATION D'UN LEXIQUE*

Trois corpus, deux en français et un en anglais, ont été traités.

Le premier est francophone et académique (normes, cours, recommandations), le second est anglophone et orienté métier. Il comprend l'ensemble des publications présentes dans les actes des 10 derniers congrès de l'AITES [3] depuis 2004 (congrès de Singapour) et jusqu'en 2013 (congrès de Genève) soit 2344 articles.

Le troisième corpus est francophone et comprend l'ensemble des communications des 6 congrès des Journées Nationales de Géotechnique et Géologie, (JNGG) soit près de 600 articles. (voir tableau 1)

Corpus	Volume	Synt. retenus	Arbres construits
1-Académique	>4000 pages	6039	23
2-ITA/AITES	2344 articles	>5000	25
3-JNGG	561 articles	>4000	23

Tableau 1 - Constitution des corpus et résultats du traitement

Le traitement est effctué par l'analyseur syntaxique Treetagger (Schmid, 1994), puis par MKD qui détecte tous les syntagmes et les ordonne suivant leur tf-idf[4] après une analyse morphosyntaxique (7 puis 5 patrons nominaux ont été utilisés[5]) et une désambiguïsation par cooccurrences du deuxième ordre (Bertels et Speelman, 2012).

---

[2] On peut ajouter que le domaine de la recherche d'information tend à privilégier ces éléments. Voir par exemple (Zacklad, 2007)

[3] AITES : Association Internationale des Travaux En Souterrain ou ITA International Tunneling Association. (73 nations affiliées, 50 entreprises majeures du BTP) dont l'AFTES : Association Française des Travaux En Souterrain, est un membre important.

[4] tf-idf est une fonction prenant en compte la fréquence d'un syntagme dans un corpus et la fréquence de sa présence dans les textes d'un corpus.

[5] La prise en compte d'adjectif dans les patrons de syntagme génère un bruit important difficile à gérer avec des résultats peu utiles.

Les premiers syntagmes du classement sont utilisés pour construire la partie cartes métier, via des graphes heuristiques. Ils permettent de définir les sous domaines des cartes métier qui sont une partition du lexique. Pour l'utilisateur géotechnicien, la première mise en œuvre a souvent semblé au départ fastidieuse, mais la structuration de ce lexique entraîne rapidement un questionnement qui en souligne l'intérêt. Comme il peut ajouter ses propres syntagmes la structure se complète rapidement. Il peut alors redéfinir le lexique avec les syntagmes des arbres construits et introduire dans ces arbres les syntagmes suivants du classement.

L'expérience montre qu'un lexique d'au moins 4000 syntagmes rend le système performant, et il est possible d'ajouter encore tous les syntagmes jugés intéressants pour la description du domaine en utilisant glossaires ou thésaurus de la profession. Cette partition en cartes correspond à une thématisation du domaine, dont l'arbitraire est laissé à l'appréciation de chaque groupe d'utilisateurs[6].

Les syntagmes utiles donc significatifs sont les premiers syntagmes du classement. Les répartir dans les arbres formant la première partie de l'ontologie, est le résultat d'une analyse d'expert. Ce travail d'initialisation renforce l'intérêt du travail collaboratif, surtout quand il a été fourni dans le cadre de la constitution ex-nihilo des lexiques comme pour les corpus 1 et 2. L'extension du lexique lors de l'ajout du corpus 3 s'est fait facilement.

En traitant à nouveau chaque texte, le logiciel établit la « signature » du texte, qui est la liste des syntagmes du texte qui ont été retenus dans le lexique. Cette signature sert de fondement à la comparaison de deux articles.

3. Résultats

MKD permet la classification d'un ensemble d'articles par degré de ressemblance avec un article donné ce qui oriente la lecture et est très prisé par les utilisateurs, dont les résultats recoupent partiellement ceux de l'outil mais se voient conseiller des articles pertinents avec lesquels ils n'auraient pas fait le lien.

C'est aussi une façon de faire de la veille technologique quand MKD retrouve très rapidement dans les éléments d'un corpus additionnel ceux qui correspondent à un thème nouveau.

L'usage des granules correspond à la recherche des granules contenant, en majorité, les syntagmes décrivant le contexte que l'on veut connaître. L'élargissement du champ de recherche se fait en ajoutant les syntagmes d'un voisinage que l'utilisateur choisit parmi plusieurs propositions. Ces voisinages sont définis par les relations entre syntagmes portés par la structure des arbres, et permettent un glissement sémantique pour augmenter le nombre des granules répondant à la recherche.

---

[6] Les partitions préalables, même réalisées par consensus, tendent à être vite négligées par les groupes d'utilisateurs, au profit de partitions propres.

4. Conclusion

Cette approche correspond à la demande de petites et moyennes entités (laboratoire, petite entreprise) qui cherchent à pérenniser leur savoir. Cette expérience montre qu'il est possible de sensibiliser les professionnels d'un domaine à formaliser facilement et rapidement les connaissances de leur domaine, ce que l'expérience malheureuse des systèmes experts avait rendu malaisé.

Cette approche permet des usages attendus, mais aussi nouveaux. Parmi des usages attendus, la transmission du savoir est simplifiée, la comparaison de textes rend performante l'appropriation de l'information en sélectionnant rapidement les textes pertinents à lire en priorité. La qualité du projet est améliorée, moins d'oublis, pas de contre sens, pas de redondance et des échanges entre métiers plus sûrs du fait d'un vocabulaire précis. Chaque « sociolecte » est bien délimité et les liens entre eux sont correctement déterminés. Parmi les usages nouveaux, une expérience d'aide à la pédagogie a été appréciée par les élèves (Faure et al., 2012). La recherche de nouvelles connaissances par la combinaison de granules est pour l'instant à l'étude et semble fort possible.

5. Bibliographie

Bertels A., Speelman D., 2012, La contribution des cooccurrences de deuxième ordre à l'analyse sémantique, revue Corpus. URL : http://corpus.revues.org/2184

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 Int. 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

Faure N., Thimus J.F., Faure R.M., 2014, Analyse statistique lexicale pour la construction d'une base de connaissances, Tunnels et Espaces Souterrains, n°244, pp317-328

Gruzitis N., Barzdins G., 2015, Polysemy in Controlled Natural Language Texts, CoRR

Kamp J., Reyle U., 1993, From discourse to logic, Kluwer

Lenat D., 1995 Cyc: A Large-Scale Investment in Knowledge Infrastructure, Communications of the ACM, 1995

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)

Schmid H. 1994, Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.

Zacklad M. 2007, Classification, thésaurus, ontologies, folksonomies : comparaisons du point de vue de la recherche ouverte d'information (ROI), ACSI, Montreal, 2007