HAL
open science

Vers un processus continu d'extraction de connaissances à partir de textes

My Thao Tang, Yannick Toussaint

To cite this version:

My Thao Tang, Yannick Toussaint. Vers un processus continu d'extraction de connaissances à partir de textes. IC - 24èmes Journées francophones d'Ingénierie des Connaissances, Jul 2013, Lille, France. hal-00861865

HAL Id: hal-00861865

https://inria.hal.science/hal-00861865v1

Submitted on 13 Sep 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Vers un processus continu d'extraction de connaissances à partir de textes

My Thao TANG[1], Yannick Toussaint[1]

Loria-Campus Scientifique BP 239 - 54506 Vandoeuvre-lès-Nancy Cedex
{My-Thao.Tang,Yannick.Toussaint}@loria.fr

Résumé : Nous posons les bases d'un système continu, itéractif et interactif, d'extraction de connaissances à partir de textes. Le cœur de notre approche repose sur l'analyse formelle de concept. Elle permet de préserver le lien entre les textes et la conceptualisation du domaine, *i.e.* le treillis construit à partir des objets et des propriétés identifiés dans les textes. Les propriétés formelles du treillis sont exploitées pour suggérer à l'expert différentes stratégies lui permettant de modifier l'état courant de la conceptualisation. Le système calcule alors les modifications à apporter aux annotations textuelles pour qu'à l'itération suivante, le treillis intègre la modification souhaitée.

Mots-clés : construction d'ontologies à partir de textes, analyse formelle de concepts, enrichissement d'ontologie, annotation sémantique

1 Introduction

Cet article présente les bases d'un *système continu* d'extraction de connaissances à partir de textes dont le noyau central repose sur l'analyse formelle de concepts (AFC). Nous définissons un *système continu* comme un système itératif et interactif dans lequel les liens entre chaque élément de connaissance et les sources (les textes) est assuré par des annotations dans les textes et ces annotations peuvent être modifiées, supprimées ou enrichies (création de nouvelles annotations) pour que le modèle de connaissances qui en résulte soit au plus proche des besoins des experts.

La construction d'une ontologie à partir de textes est un processus complexe (Aussenac-Gilles *et al.* (2000); Szulman *et al.* (2010)) et coûteux qui implique la collecte de ressources, la mise en œuvre d'outils automatiques de prétraitement et de traitement de ces ressources pour passer du niveau

linguistique au niveau conceptuel, identifiant en premier lieu la terminologie, puis construisant progressivement un ensemble de concepts. L'expert du domaine guide ce processus en privilégiant l'émergence de certains termes, de certains concepts, ou en introduisant des éléments de connaissance manquant. Son rôle est essentiel, il doit exploiter sa connaissance et sa perception du domaine mais il doit également prendre en compte la tâche pour laquelle l'ontologie est construite. Cependant, lorsque la tâche est réalisée par un moteur de raisonnement automatique, il est parfois difficile de savoir quelle est l'incidence des choix qu'il a faits lors de la conception sur la qualité du résultat. Plusieurs raisons peuvent donc amener un expert à envisager des changements ou des évolutions dans une ontologie : (1) la connaissance modélisée dans l'ontologie n'est pas en accord avec sa connaissance ; (2) la tâche ou application utilisant l'ontologie ne produit pas les résultats escomptés ; (3) de nouvelles ressources, notamment, de nouveaux textes, sont à prendre en compte.

L'analyse formelle de concepts (AFC), en tant que processus symbolique de classification, est la base de notre système continu d'extraction de connaissances. L'AFC est un outil de conceptualisation puissant, rigoureux et précis, pour lequel nous disposons d'algorithmes incrémentaux. Elle s'appuie sur une sémantique formelle et il est possible de transformer les résultats (le treillis) en des formules d'une logique de descriptions, le formalisme de représentation de connaissances que nous utilisons pour coder l'ontologie. De plus, il est possible d'introduire explicitement des connaissances externes pour pallier les connaissances implicites contenues dans les textes comme le suggère Bendaoud *et al.* (2008).

L'originalité de notre approche réside dans le fait de placer l'AFC au cœur d'un système dynamique, itératif, et d'exploiter les propriétés formelles du treillis pour guider les modifications des annotations dans les textes afin d'obtenir un nouveau treillis plus proche des attentes de l'expert ou des besoins d'une application. En effet, bien que l'extraction d'information soit un domaine très actif, l'annotation des textes pour identifier les termes et les relations entre ces termes est imparfaite. Le treillis s'appuyant sur ces annotations le sera donc aussi. Notre processus vise donc à améliorer conjointement l'annotation des textes et le treillis que nous considérons comme le noyau de connaissances à partir duquel sera construit l'ontologie finale. Dans le cas où le processus initial d'annotation des textes serait un processus automatique reposant sur une phase d'apprentissage (Conditional Random Fields par exemple...), le corpus final ainsi corrigé pourrait être réutilisé pour parfaire l'apprentissage de cet outil d'annotation.

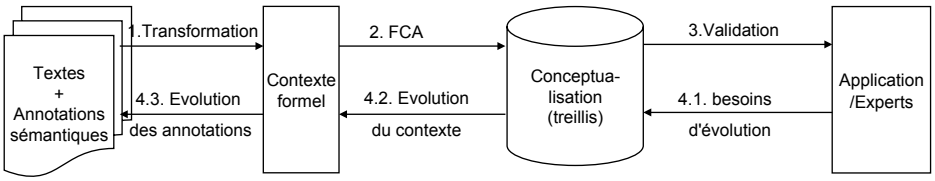

FIGURE 1 – Processus général de construction de l'ontologie

2 Schéma global et annotation des textes

2.1 Schéma global du processus

Le schéma global du processus est présenté en Figure 1. Les textes sont tout d'abord annotés sémantiquement pour identifier des objets et des propriétés associées à ces objets (voir section 2.2). L'annotation sémantique peut être manuelle, ou réalisée par un outil automatique. Dans le cadre de notre expérimentation, nous avons utilisé l'outil SEMREP pour annoter les textes initiaux. Cette annotation peut être entachée de bruit puisqu'un de nos objectifs est justement de corriger ces annotations en lien avec la construction de l'ontologie. Textes et annotations constituent les données d'entrée pour construire le contexte formel. Seules les annotations sémantiques sont utilisées pour construire le contexte formel. En revanche, les annotations ne sont pas décorrélées des textes pour permettre à l'expert d'identifier dans chaque texte l'information qui en a été extraite.

La seconde étape utilise l'AFC pour construire une conceptualisation du domaine. La conceptualisation nous est donc donnée par le treillis résultant de l'AFC. Elle est alors soumise à évaluation par l'expert ou par tout processus de raisonnement exploitant ces connaissances. Cette évaluation suggère des modifications des concepts du treillis. Pour réaliser ces modifications, le système propose différentes stratégies qui viseront à modifier le contexte formel par le biais d'une modification des annotations des textes. L'itération suivante construit alors le nouveau treillis.

Pour faire évoluer la conceptualisation, nous mettons à disposition de l'expert un ensemble d'opérations (par ex : *détruire le concept* c_i). Le système exploite alors les propriétés du treillis pour déterminer quelles sont les différentes stratégies possibles pour réaliser cet objectif. De même, pour une stratégie donnée, il détermine automatiquement les modifications qui doivent être appliquées sur les annotations textuelles pour ensuite recalculer le treillis. Ce processus assure donc une cohérence permanente entre les

Intracranial fibromuscular dysplasia in a six-year-old child: a rare cause of childhood stroke.

Intracranial fibromuscular dysplasia is a nonatheromatous angiopathy that most commonly affects adult women and is rarely recognized in children. Symptoms include stroke and headache, although the vasculopathy may be asymptomatic. Diagnosis is based on angiographic appearance, commonly described as a "string of beads." The etiology of intracranial fibromuscular dysplasia is not known, although possible causes include genetic predisposition, trauma, and underlying connective tissue disease. Treatment of intracranial fibromuscular dysplasia is largely supportive once symptoms become manifest. We report a 6-year-old girl who presented to our center for further evaluation of a large left middle cerebral artery distribution infarction. The patient was previously healthy, without known risk factors for stroke. Initial symptoms consisted of a dense global aphasia and a right hemiparesis. On arrival, the patient's aphasia had improved but she continued to have significant deficits in both receptive and expressive language as well as residual right hemiparesis. Magnetic resonance imaging and conventional angiographic studies demonstrated characteristic beading of the distal portion of the left internal carotid artery, as well as the proximal middle cerebral artery. Laboratory evaluation, echocardiogram, and renal ultrasound were normal. The renal vasculature did not demonstrate evidence of intracranial fibromuscular dysplasia. In conclusion, intracranial fibromuscular dysplasia should be considered in the differential diagnosis of childhood stroke. When recognized, other sites of vascular involvement should be sought, and consideration of underlying disorders is important, as connective tissue disorders have been associated with a propensity to develop this vascular abnormality. Careful follow-up is warranted, due to possible progression of disease.

FIGURE 2 – Exemple d'un document (titre et résumé) extrait de PUBMED (PMID 10961798).

données, le treillis et l'évaluation par l'expert.

2.2 Annotation sémantique des textes

Les données en entrée de notre processus sont des textes annotés sémantiquement. Notre expérimentation porte sur des textes dans le domaine de la médecine extraits de PUBMED [1]. La Figure 2 en donne un exemple. Ces textes sont alors annotés sémantiquement. Cette annotation doit permettre d'extraire des textes des objets du domaine et des propriétés. Dans cette expérience, nous avons utilisé SEMREP (Rindflesch & Fiszman (2003)).

La Figure 3 donne l'annotation du texte de la Figure 2. SEMREP identifie des entités, que nous considérons par la suite comme des objets : `Fibromuscular Dysplasia` est ainsi typé comme `dsyn (disease or syndrom)`, associé au concept `C0016052` de l'UMLS (1ère ligne de la figure). D'autres informations sont également fournies sur la même ligne comme le terme préférentiel de l'UMLS et les indices de position dans le texte. Ces objets peuvent être ensuite impliqués dans des relations comme la relation `affect` (dernière ligne de la figure). Etant donnée une annotation $(objet_1, relation, objet_2)$, nous considérons $objet_1$ comme objet du domaine et $relation$ est concaténé avec $objet_2$ pour constituer un attribut de $objet_1$. L'annotation par SEMREP est bruitée : termes et relations peuvent être mal "délimités" ou pas reconnus mais aucun outil

1. http ://www.ncbi.nlm.nih.gov/pubmed

```
SE|00000000||tx|1|text|Intracranial fibromuscular dysplasia in a six-year-old
child: a rare cause of childhood stroke.
...

• SE|00000000||tx|1|entity|C0016052|Fibromuscular
  Dysplasia|dsyn|||fibromuscular dysplasia||||901|14|36
• SE|00000000||tx|1|entity|C0205452|Six|qnco|||six||||833|43|45
• SE|00000000||tx|1|entity|C0439508|/year|tmco|||year||||833|47|50
• SE|00000000||tx|1|entity|C0580836|Old|tmco|||old||||833|52|54
...

• SE|00000000||tx|1|relation|2|1|C0042373|Vascular
  Diseases|dsyn|dsyn|||angiopathy||||861|59|68|VERB|AFFECTS||89|95|2|1|C0043210
  |Woman|popg,humn|humn|||women||||888|103|107
```

FIGURE 3 – Exemples de quelques entités et relations extraites par SEMREP pour le texte donné en Figure 2.

automatique ne réalisera une annotation parfaite. Notre approche doit justement permettre de corriger les annotations en parallèle à la construction de la conceptualisation.

3 Analyse Formelle de Concepts

L'analyse formelle de concepts (AFC) Ganter (1999) est un formalisme mathématique pour construire un treillis de concepts à partir d'un contexte formel $\mathbb{K} = (G, M, I)$. La structure du treillis a souvent été mise en avant pour sa capacité à conceptualiser des données dans une démarche *bottom-up*, c'est-à-dire de partir de données décrivant des individus associés à des propriétés pour en proposer une organisation selon une hiérarchie de classes (appelées aussi concepts formels) de nature à faciliter l'analyse et la compréhension d'un jeu de données parfois important.

3.1 Quelques bases en analyse formelle de concepts

Dans l'analyse formelle de concepts, les données se présentent sous la forme d'un contexte formel $\mathbb{K} = (G, M, I)$ dans lequel G dénote un ensemble d'individus ou objets, M un ensemble d'attributs et I est une relation binaire définie sur le produit cartésien $G \times M$. Dans la table binaire représentant $I \subseteq G \times M$, une ligne correspond à un individu et une colonne correspond à un attribut (voir Table 1). L'analyse formelle de concepts construit le treillis de concepts composé d'un ensemble de *concepts formels* organisés selon un ordre partiel *i.e.* la relation de subsomption. Un concept est une paire (A, B) où $A \subseteq G$, $B \subseteq M$ et A est l'ensemble maximal des individus partageant l'ensemble des propriétés de B (et réciproquement). Étant donné un concept (A, B), A est appelé extension et B est

TABLE 1 – Contexte binaire $\mathbb{K}$.

Object	ISA_Diagnosis	ISA_Therapeutic procedure	ISA_Symptoms	ISA_Disease	ISA_Vascular Diseases	COEXISTS_WITH_Disease	CAUSES_Renal Artery Stenosis
Fibromuscular Dysplasia			x	x	x		x
Renal arteriography	x						
Fibrosis						x	x
Claudication			x				
Ischemia				x			
Aortography	x	x					
Arteriosclerosis							x
Hypertensive disease			x			x	
angiogram	x						

appelé intension. Le calcul des concepts formels se fait selon la connexion de Galois définie par les deux opérateurs de dérivation $'$:
$A' := \{m \in M \mid \forall g \in A, gIm\}$ $\qquad$ $B' := \{g \in G \mid \forall m \in B, gIm\}$

Un concept (A, B) doit vérifier la contrainte de clôture telle que $A' = B$ et $B' = A$. La relation de subsomption ($\sqsubseteq$) entre deux concepts est définie comme suit : $(A_1, B_1) \sqsubseteq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2$ (ou $B_2 \subseteq B_1$). Cette relation de subsomption permet d'organiser tous les concepts extraits du contexte $\mathbb{K}$ = (G, M, I) en un treillis noté $\underline{\mathfrak{B}}(\mathbb{K})$. Le treillis associé au contexte formel de la table 1 est donné par la figure 4. Ce contexte a été construit à partir des textes médicaux annotés par SEMREP : étant donné un triplet $(objet_1, relation, objet_2)$ les objets sont les entités apparaissant en partie droite des triplets et les attributs sont construit par concaténation du nom de la relation et de $objet_2$.

Dans la figure 4, de par la construction de la relation de subsomption, les attributs d'un concept subsumé hérite des attributs du concept subsumant et inversement, les objets du concept subsumé sont hérités par le concept subsumant. Cependant, pour en simplifier la lecture, on ne visualise dans le treillis que les attributs (resp. les objets) locaux à un concept (affichage réduit), c'est à dire ceux qui ne sont pas hérités. Par exemple, le concept C_4 subsume le concept C_5. Un certain nombre d'autres propriétés découlent de cette construction : (1) deux concepts ne peuvent avoir ni le même ensemble d'attributs, ni le même ensemble d'objets. L'ensemble des attributs est donc définitoire pour le concept. C'est une propriété importante que nous exploitons pour la définition des concepts dans l'ontologie ; (2) deux

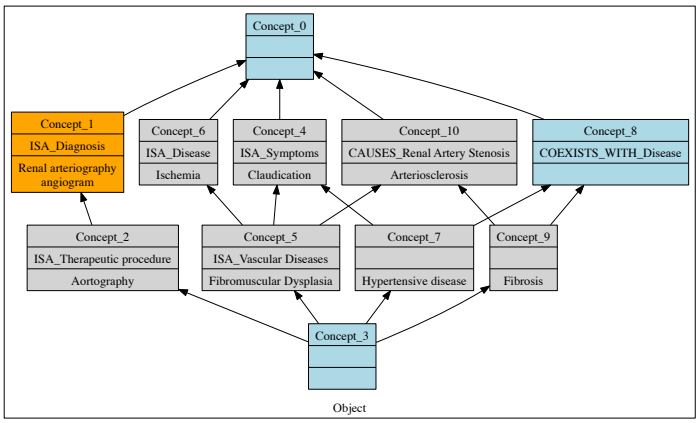

FIGURE 4 – Le treillis $\underline{\mathfrak{B}}(\mathbb{K})$ en affichage réduit.

attributs locaux à un concept sont équivalents au sens où ils sont associés au même ensemble d'objets ; (3) il est possible d'extraire des règles d'association à partir du treillis, notamment les règles d'implication du type *attribut_local* $\longrightarrow$ *attributs_hérités*.

3.2 Analyse formelle de concepts et ontologie

Il existe de nombreux travaux exploitant l'analyse formelle de concepts pour les ontologies. Ainsi, Stumme & Maedche (2001) exploite l'AFC pour fusionner deux ontologies en s'appuyant sur le contexte d'apparition des termes et ainsi rapprocher les termes ayant des contextes similaires. Sur un plan plus formel, Baader *et al.* (2007) et Sertkaya (2008) se sont intéressés aux liens entre l'analyse formelle de concepts et les logiques de descriptions, formalisme souvent utilisé pour la représentation de connaissances. Maedche (2002) et Cimiano & Völker (2005) s'intéressent à construire une hiérarchie de concepts à partir des propriétés extraites des textes et suggèrent une méthode d'évaluation basée sur une mesure de similarité qui permet de rapprocher les concepts du treillis d'un thésaurus. Bendaoud *et al.* (2008) fait une présentation assez détaillée de l'intérêt de l'AFC pour la construction d'ontologie. Nous y montrons que l'opération d'apposition permet d'introduire explicitement dans le processus de conceptualisation des connaissances expertes qui sont implicites dans les textes. Il est également possible d'exploiter les termes d'un thésaurus pour

nommer les classes ou, autrement dit, de proposer une définition formelle pour des termes d'un thésaurus. Enfin, Bendaoud *et al.* (2008) exploite les contextes relationnels introduits dans la RCA (Relationnal Concept Analysis, Huchard *et al.* (2007)) pour que les relations entre objets puissent contribuer à la construction des concepts et propose une transformation de ces concepts en logique de descriptions.

Dans cet article, nous ne prendrons pas en compte la RCA qui introduit un niveau de complexité supplémentaire par rapport à nos objectifs de placer l'AFC au cœur d'un processus continu. Des travaux que nous avons cités, il ressort que l'AFC est un outil puissant pour la construction d'ontologies :

- L'AFC peut exploiter tous les éléments nécessaires à la représentation de connaissances : les individus, les attributs et les relations entre ces individus ;
- l'AFC propose une organisation des concepts sous forme hiérarchique et un ensemble d'opérations pouvant exploiter des hiérarchies (thésaurus...) existantes ;
- Il existe des algorithmes incrémentaux permettant de mettre à jour un treillis quand l'ensemble initial d'objets ou d'attributs augmente ;
- Le treillis peut être transformé en une base de connaissances exprimée en logique de descriptions pour être exploitée par un raisonneur.

Cependant, l'AFC a également des faiblesses et assimiler le treillis à l'ontologie est un raccourci dont il s'agit d'estimer le biais :

- L'AFC est sensible au bruit dans les données. La présence d'un attribut peu pertinent pour la modélisation du domaine peut entraîner une prolifération de concepts sans intérêt pour l'expert ;
- Deux concepts ne peuvent pas avoir la même intention. Il est donc nécessaire d'introduire ou d'identifier de nouvelles propriétés pour que deux concepts différents puissent alors exister dans le treillis ;
- L'AFC produit un treillis unique pour un jeu de données en entrée. Pour tout couple de concepts, il existe dans le treillis un concept *sup* et un concept *inf*. Cependant, parmi tous ces concepts du treillis, tous ne sont pas de nature "ontologique" mais aucun concept ne peut être supprimé. Ainsi, on peut imaginer qu'un médicament et une maladie se trouvent regroupés dans un même concept du treillis parce qu'ils partagent, par exemple, la propriété d'être liés à une même partie du corps. Cependant, si un expert ne souhaite pas valider ce concept formel comme un concept de son ontologie, il ne peut pas pour autant le supprimer du treillis.

– Identifier les concepts dits primitifs (au sens de la logique de descriptions) et les rôles ou encore nommer les concepts restent un travail que l'expert doit prendre en charge.

Placer l'AFC au sein d'un processus itératif et interactif peut pallier un certain nombre de ces limites. Les propriétés formelles de l'AFC guideront l'interaction avec l'expert. De plus, le lien entre les données et la conceptualisation est ainsi préservé, conformément à notre définition de la continuité. L'idée est donc de faire interagir l'expert sur le treillis pour qu'au final, l'expert apporte le moins de valeur ajoutée possible lors du passage du treillis à l'ontologie. Cette valeur ajoutée "résiduelle" est un point que nous devrons évaluer par la suite. Nous avons défini un ensemble d'opérations sur les treillis qui nous permette de répondre aux besoins des experts pour faire évoluer conjointement le treillis et les données en entrée de la FCA et ainsi accorder au plus proche le treillis de la vision de l'expert.

4 Evolution et enrichissement des ontologies

Le treillis obtenu à partir des annotations textuelles doit être évalué par l'expert. L'expert peut alors suggérer des changements pour que le treillis s'accorde mieux avec la connaissance qu'il souhaite représenter. Cependant, comme nous l'avons souligné, nous ne souhaitons pas donner la possibilité à l'expert de modifier directement les concepts du treillis comme il pourrait le faire pour modifier les concepts d'une ontologie sous PROTÉGÉ : la structure du treillis serait perdue et avec elle, les propriétés définitoires des attributs, mais également, le lien avec les annotations textuelles originales. Pour faire évoluer le treillis, nous mettons à disposition de l'expert un ensemble d'opérations. Il peut ainsi formuler les modifications qu'il souhaite apporter, le système lui proposera alors plusieurs alternatives s'il y en a, et selon sa décision, le système lui suggérera des modifications dans les annotations textuelles de sorte que le calcul du nouveau treillis tienne compte de la modification souhaitée. L'accord entre l'annotation des textes et le treillis est ainsi préservé, et donc aussi le lien entre l'ontologie finale et les élements de connaissance identifiés dans les textes.

4.1 Les changements dans une ontologie

L'évolution d'une ontologie peut se définir comme une adaptation dans le temps de l'ontologie à des changements qui ont surgis dans le temps et la

propagation de cette adaptation aux artéfacts qui en dépendent (Stojanovic (2004)).

Stojanovic *et al.* (2002) définit l'évolution d'une ontologie par comme un cycle composé de six phases :

1. la capture du changement : la capture du chagement se fait à partir de la formulation d'une exigence implicite ou explicite. Plusieurs travaux se sont intéressés à la nature du changement (Stojanovic (2004); Cimiano & Völker (2005); Castano *et al.* (2006)) et à l'évolution des connaissances pouvant conduire à l'évolution de l'ontologie (Klein (2004); Enkhsaikhan *et al.* (2007)) ;
2. la représentation du changement : une fois les besoins identifiés, il faut donner à l'expert les moyens de le formuler (Stojanovic (2004); Klein (2004); Luong (2007); Tissaoui (2009); Reymonet *et al.* (2010)) ;
3. la sémantique du changement porte sur l'impact d'un changement sur l'ontologie, et vise notamment à assurer la préservation de la cohérence de l'ontologie (Stojanovic (2004); Haase *et al.* (2004)) ;
4. la mise en œuvre du changement vise à effectuer le changement et à garder dans le système une trace de l'évolution (Stojanovic (2004); Haase *et al.* (2004); Klein (2004); Flouris (2006); Luong (2007)) ;
5. la propagation de ce changement doit permettre d'identifier les conséquences d'un changement sur les autres artéfacts (Luong *et al.* (2006); Köpke & Eder (2011); Tissaoui *et al.* (2011)) ;
6. la validation est l'étape durant laquelle l'expert doit valider le changement et c'est aussi la fin d'un cycle et, éventuellement, le début d'un nouveau cycle.

De façon similaire à ces travaux, nous devons définir un processus dans lequel un expert peut évaluer un treillis, demander des modifications, et transposer ces modifications sur les données (les annotations dans les textes) pour que le nouveau treillis réponde aux exigences. Il n'y a pas à proprement parler de vérification de la cohérence, le nouveau treillis étant, par construction, correct. En revanche, les modifications dans les données doivent être suggérées à l'expert et si les conséquences de ces modifications impactent plusieurs concepts du treillis, il faut également le lui signaler.

4.2 Définition d'opérations sur le treillis

Les opérations sur le treillis correspondent d'une certaine façon à du rétro-engineering : l'expert formule une opération sur le treillis et le système doit suggérer les différentes alternatives, doit évaluer l'incidence de telle ou telle modification et calculer quels sont les changements nécessaires sur les données pour que le nouveau treillis réponde aux exigences.

Il est nécessaire de proposer à l'expert un ensemble d'opérations pour modifier le treillis. Dans la mesure où l'ajout d'objet ou de propriétés dans un contexte formel a été étudié et a motivé le développement d'algorithmes de construction de treillis incrémentaux (Valtchev *et al.* (2003)), nous nous sommes intéressés jusqu'à présent à la suppression d'un concept du treillis et nous en détaillons le fonctionnement sur un exemple. Etant donné le treillis en Figure 4, supposons que l'expert souhaite supprimer le concept c_8 de $\underline{\mathfrak{B}}(\mathbb{K})$. Deux solutions s'offrent à lui. Soit il glisse le concept $c_8$vers un des concepts immédiatement supérieur (selon la subsomption), ici, cela ne peut être que le concept c_0. Soit il le glisse vers un des concepts inférieurs c_7 ou c_9.

Pour déplacer le concept c_8 vers son concept supérieur c_0, trois stratégies se présentent à l'expert. La première – *Stratégie 1.1* – préserve dans le contexte formel tous les objets et tous les attributs mais retire l'attribut *COEXISTS_WITH_Disease* pour chacun des objets de l'extension de c_8 : *Hypertensive disease*, et *Fibrosis*. On se reportera donc aux textes pour retrouver les occurences des triplets RDF (*Hypertensive disease*, *COEXISTS_WITH*, *Disease*) et (*Fibrosis*, *COEXISTS_WITH*, *Disease*). Dans l'optique d'associer un coût aux changements, on considère que ce changement implique deux modifications *primitives*. Les occurences de ces annotations dans les textes sont alors marquées comme invalides et ne seront plus prises en compte pour la construction du nouveau contexte formel.

La seconde stratégie – *Stratégie 1.2* – retire l'attribut *COEXISTS_WITH_Disease* du contexte formel. Là encore, cette stratégie implique deux modifications primitives. La figure 5 correspond au nouveau treillis construit une fois la stratégie 1.2 appliquée.

Si on estime que l'identification des objets dans le texte peut être aussi bruitée que l'annotation des propriétés ou attributs, de par la dualité du treillis entre extension et intension, une troisième stratégie pourrait envisager de supprimer les deux objets appartenant au concept c_8.

La seconde solution consiste à déplacer le concept c_8 vers un de ses

TABLE 2 – Stratégies pour supprimer un concept c_8 associées au nombre de modifications impliquées(NOM), au nombre de concepts modifiés (NOCC) et leur effet.

	Strategie	**NOM**	**NOCC**	**Effets**
1.1	Move c_8 up to c_0	2	2	c_9 jumps to c_{10} c_7 jumps to c_4
1.2	Move c_8 up to c_0 (remove attribute *COEXISTS_WITH_Disease* out of the context)	2	2	c_9 jumps to c_{10} c_7 jumps to c_4
2.1	Move c_8 down to c_7	1	1	c_9 jumps to c_{10}
2.2	Move c_8 down to c_9	1	1	c_7 jumps to c_4

TABLE 3 – Nouveau contexte formel suivant la *Stratégie 1.2*.

Object	ISA_Diagnosis	ISA_Therapeutic procedure	ISA_Symptoms	ISA_Disease	ISA_Vascular Diseases	CAUSES_Renal Artery Stenosis
Fibromuscular Dysplasia			x	x	x	x
Renal arteriography	x					
Fibrosis						x
Claudication			x			
Ischemia				x		
Aortography	x	x				
Arteriosclerosis						x
Hypertensive disease			x			
angiogram	x					

fils. Là encore deux stratégies sont possibles. La *Stratégie 2.1* déplace le concept c_8 vers le concept c_7. Il faut donc garder les objets et attributs du contexte formel mais supprimer l'attribut *COEXISTS_WITH_Disease* pour l'objet *Fibrosis*. Cette stratégie ne suppose donc qu'une modification primitive.

La *Stratégie 2.2* déplace c_8 vers c_9. Il faut également garder tous les objets et attributs du contexte formel mais supprimer l'attribut *COEXISTS_WITH_Disease* de l'objet *Hypertensive disease*. Cette stratégie ne suppose elle aussi une seule modification primitive.

Au final, quatre stratégies peuvent répondre aux attentes de l'expert. Elle sont résumées dans la table 2. Une fois la stratégie choisie par l'expert, les modifications qu'elle suppose sont propagées vers les textes, les annotations correspondantes sont *invalidées* et le contexte formel peut alors être mis à jour en fonction de ces modifications. Ainsi si l'expert choisit la *Stratégie 1.2* pour détruire le concept c_8, le nouveau contexte formel sera celui de la Table 3. Dans ce contexte, l'attribut *COEXISTS_WITH_Disease*

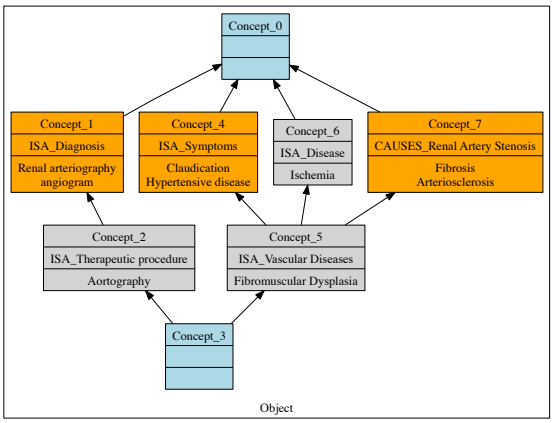

FIGURE 5 – Treillis résultant de la *Stratégie 1.2* après mise à jour des annotations et du contexte formel.

n'est plus associé aux objets *Hypertensive disease* et *Fibrosis* et le nouveau treillis et donné en Figure 5.

La notion de *modification primitive* est utilisée pour privilégier dans la présentation à l'expert les stratégies les moins coûteuses. Elle correspond au nombre de cellules modifiées dans le contexte formel. Cependant, d'autres mesures pourraient être expérimentées. Ainsi, le nombre d'occurences modifiées dans les annotations pourrait également être significatif.

4.3 Expérimentation en corpus

Nous avons expérimenté notre approche sur un corpus réel sur la dysplasie fibromusculaire artérielle. Le corpus est constitué de 400 textes et l'annotation par SEMREP produit 2402 objets, dont seulement 668 apparaissent en partie droite ou gauche d'un triplet de relation, et 481 en partie droite ; 36 relations différentes sont identifiées dans ces textes.

Le contexte final contient donc 481 objets, 545 attributs constitués par l'association de la relation et de l'objet sur laquelle elle porte. Le treillis contient 523 concepts et le plus grand chemin entre le Top et le Bottom est de 7. Le temps nécessaire à l'expert pour supprimer les concepts n'a pas encore pu être évalué. Notre propre expérience montre qu'un choix judicieux dans l'ordre de suppression des concepts et dans la stratégie impacte fortement sur la vitesse de convergence vers une conceptualisation satisfaisante. Au niveau du document, nous gardons une trace des annotations

initiales mais elles sont validées ou invalidées au gré des modifications demandées par l'expert. Parmi ces annotations, un certain nombre d'entre elles étaient manifestement incorrectes si on se réfère à la syntaxe et la sémantique de la phrase. Cette proportion reste cependant à mesurer précisément.

5 Perspectives et conclusions

Nous avons posé les bases d'un système continu de construction d'ontologies à partir de textes en plaçant l'analyse formelle de concepts au cœur du processus. Un tel système permet de profiter des propriétés formelles de l'AFC tout en réduisant les limites liées au formalisme. Ces bases ouvrent de nombreuses perspectives. Ainsi, il serait intéressant d'étudier comment mieux articuler ces travaux avec les travaux menés précédemment pour rendre progressif la construction d'ontologies, comme cela est suggéré dans la plateforme Dafoe (Szulman *et al.* (2010)). L'incrémentalité dans la construction de treillis a fait l'objet de nombreux travaux dont nous pourrions mieux tirer profit pour guider l'expert dans son choix mais également, sur un plan algorithmique, pour optimiser les coûts de calcul. Enfin, le coût actuellement associé à une modification peut également être affiné. Il est possible de prendre en compte le nombre d'occurrences des triplets dans les textes conjointement au nombre de modifications dans le contexte formel. La propagation des modifications pourrait également intervenir. Nous envisageons d'introduire au niveau du treillis des mesures statistiques qui elles aussi pourraient aider l'expert à choisir. De ce point de vue, la stabilité nous semble une mesure intéressante à expérimenter.

Références

AUSSENAC-GILLES N., BIÉBOW B. & SZULMAN S. (2000). Modélisation du domaine par une méthode fondée sur l'analyse de corpus. In P. TCHOUNIKINE, Ed., *Actes de la 9e Conférence Francophone d'Ingénierie des Connaissances IC 2000*, p. 93–104, Toulouse, France : Université Paul Sabatier. 12 pages.

BAADER F., GANTER B., SERTAKAYA B. & SATTLER U. (2007). Completing descriprion logics knowledge bases using formal concepts analysis. In M. M. VELOSO, Ed., *Proceedings of the 20th International Joint Conférence on Artificial Intelligence (IJCAI'2007)*, p. 230–235, Hyderabad, India.

BENDAOUD R., NAPOLI A. & TOUSSAINT Y. (2008). Formal Concept Analysis : A unified framework for building and refining ontologies. In ALDO GANGEMI & JÉRÔME EUZENAT, Eds., *16th International Conference on Know-*

ledge Engineering and Knowledge Management - EKAW 2008, volume 5268, p. 156–171, Acitrezza, Catania Italie : Springer Berlin / Heidelberg.

CASTANO S., FERRARA A. & HESS G. N. (2006). Discovery-driven ontology evolution. In *SWAP 2006 Semantic Web Applications and Perspectives - Proceedings of the 3rd Italian Semantic Web Workshop*.

CIMIANO P. & VÖLKER J. (2005). Text2onto - a framework for ontology learning and data-driven change discovery. In A. MONTOYO, R. MUNOZ & E. METAIS, Eds., *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB)*, volume 3513 of *Lecture Notes in Computer Science*, p. 227–238 : Springer.

CIMIANO P. & VÖLKER J. (2005). Text2onto - a framework for ontology learning and data-driven change discovery. In SPRINGER, Ed., *Proceedings of Natural Language Processing and Information Systems (NLDB 2005)*, number 3513 in Lecture Notes in Computer Science (LNCS), p. 227–238.

ENKHSAIKHAN M., WONG W., LIU W. & REYNOLDS M. (2007). Measuring data-driven ontology changes using text mining. In *Proceedings of the sixth Australasian conference on Data mining and analytics - Volume 70*, AusDM '07, p. 39–46, Darlinghurst, Australia, Australia : Australian Computer Society, Inc.

FLOURIS G. (2006). *On Belief Change and Ontology Evolution*. PhD thesis, University of Crete.

GANTER B. (1999). *Formal Concept Analysis - Mathematical Foundations*. Springer Verlag.

HAASE P., SURE Y. & VRANDECIC D. (2004). *Ontology Management and Evolution – Survey, Methods and Prototype*. SEKT formal deliverable D3.1.1, Institute AIFB, University of Karlsruhe.

HUCHARD M., NAPOLI A., ROUANE M. H. & VALTCHEV P. (2007). A proposal for combining formal concept analysis and description logics for mining relational data. In S. KUZNETSOV & S. SCHMIDT, Eds., *proceeding of the 5th International Conference Formal Concept Analysis (ICFCA'07)*, LNAI 4390, p. 51–65, Clermond-Ferrand, France : Springer, Berlin.

KLEIN M. (2004). *Change Management for Distributed Ontologies*. PhD thesis, Amsterdam : Vrije Universiteit.

KÖPKE J. & EDER J. (2011). Semantic invalidation of annotations due to ontology evolution. In *Proceedings of the 2011th Confederated international conference on On the move to meaningful internet systems - Volume Part II*, OTM'11, p. 763–780, Berlin, Heidelberg : Springer-Verlag.

LUONG P. H. (2007). *Gestion de l'évolution d'un Web sémantique d'entreprise*. These, École Nationale Supérieure des Mines de Paris.

LUONG P.-H., DIENG R. & BOUCHER A. (2006). Managing semantic annotations evolution in the coswem system. In *Third national symposium on Research, Development and Application of Information and Communication*

Technology (ICT.rda), Hanoi (Vietnam).

MAEDCHE A. (2002). *Ontology Learning for the Semantic Web*. Springer.

REYMONET A., THOMAS J. & AUSSENAC-GILLES N. (2010). Ontologies et recherche d'information : une application au diagnostic automobile. In S. DESPRÈS, Ed., *Journées Francophones d'Ingénierie des Connaissances (IC 2010*, p. 283–294.

RINDFLESCH T. C. & FISZMAN M. (2003). The interaction of domain knowledge and linguistic structure in natural language processing : Interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics*, **36**(6), 462–477.

SERTKAYA B. (2008). *Formal Concept Analysis Methods for Descriptions Logics*. PhD thesis, Dresden university.

STOJANOVIC L. (2004). *Methods and Tools for Ontology Evolution*. PhD thesis, University of Karlsruhe.

STOJANOVIC L., MAEDCHE A., MOTIK B. & STOJANOVIC N. (2002). User-driven ontology evolution management. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, EKAW '02, p. 285–300, London, UK : Springer-Verlag.

STUMME G. & MAEDCHE A. (2001). Fca-merge : Bottom-up merging of ontologies. In *acte de 17th International Joint Conferences on Artificial Intelligence (IJCAI'01)*, p. 225–234,, San Francisco, CA : Morgan Kaufmann Publishers, Inc.

SZULMAN S., CHARLET J., AUSSENAC-GILLES N., NAZARENKO A., TEGUIAK V. & SARDET E. (2010). Dafoe : an ontology building platform from texts or thesauri. In *Proc. of International Conference on Knowledge Engineering and Knowledge Management (EKAW 2010)*.

TISSAOUI A. (2009). Typologie de changements et leurs effets sur l'évolution de ressources termino-ontologiques (poster). In F. GANDON, Ed., *IC 2009 : Posters des 20es Journées Francophones d'Ingénierie des Connaissances, Hammamet (Tunisie)*.

TISSAOUI A., AUSSENAC-GILLES N., HERNANDEZ N. & LAUBLET P. (2011). Evonto - joint evolution of ontologies and semantic annotations. In J. DIETZ, Ed., *International Conference on Knowledge Engineering and Ontology Development (KEOD)*, p. 1–6 : INSTICC - Institute for Systems and Technologies of Information, Control and Communication.

VALTCHEV P., HACENE M. R., HUCHARD M. & ROUME C. (2003). Extracting Formal Concepts out of Relational Data. In E. SANJUAN, A. BERRY, A. SIGAYRET & A. NAPOLI, Eds., *Proceedings of the 4th Int. Conference Journées de l'Informatique Messine (JIM'03) : Knowledge Discovery and Discrete Mathematics, Metz (FR), 3-6 September*, p. 37–49 : INRIA.