L'Analyse Formelle de Concepts au service de la construction et l'enrichissement d'une ontologie

Rokia Bendaoud*, Yannick Toussaint*
Amedeo Napoli*

*UMR 7503 LORIA, BP 239, 54506 Vandœuvre-lès-Nancy, FRANCE
{Rokia.Bendaoud,Yannick.Toussaint,Amedeo.Napoli}@loria.fr,
http://www.loria.fr/~bendaoud/

Résumé. Dans cet article, nous proposons une méthodologie appelée PACTOLE «Property And Class Caracterisation from Text to OntoLogy Enrichment» qui permet de construire une ontologie dans un domaine spécifique et pour une application donnée. PACTOLE fusionne et combine différentes ressources à l'aide de l'Analyse Formelle de Concepts (AFC) et de son extension l'Analyse Relationnelle de Concepts (ARC). Les expressions produites par AFC/ARC sont représentées en expressions d'une Logique de Descriptions LD (ici $\mathcal{FLE}$) puis implémentées en OWL. Il est ensuite possible de raisonner sur ces expressions. Cette méthodologie est appliquée au domaine de l'astronomie. Nous montrons aussi comment nous avons formalisé et répondu à certaines questions que se posent les astronomes.

1 Introduction

Les ontologies sont des éléments essentiels du Web sémantique. Ce sont «des spécifications explicites de la conceptualisation d'un domaine ; elles sont représentées par des hiérarchies de concepts reliés par des relations transversales» Gruber (1993). Les ontologies permettent de partager, de diffuser et d'actualiser les connaissances du domaine. Différentes ressources peuvent être utilisés pour les construire, telles que des thésaurus, bases de données, dictionnaires, corpus de textes,... Chacune de ces ressources peut être considérée comme étant un point de vue du domaine spécifique qu'elle traite mais aucune d'elles n'est considérée comme complète. Afin de construire une ontologie du domaine la plus complète possible, il faut combiner ces différentes ressources hétérogènes.

Dans notre domaine d'application, la construction d'une hiérarchie de concepts et l'identification/classification des objets célestes (i.e. attribuer un concept à un objet) est une tâche très difficile. Traditionnellement, la classification des objets est faite manuellement par les astronomes. Cette tâche consiste d'abord à lire les articles scientifiques où apparaissent les objets (première ressource), puis à trouver la classe de cet objet dans un ensemble de classes prédéfinies dans la base SIMBAD[1] (deuxième ressource). Dans cet article, une classe représente

[1] http ://simbad.u-strasbg.fr/simbad/sim-fid

un ensemble d'objets comme en analyse des données alors qu'un concept regroupe un ensemble d'objets d'après des attributs binaires et des attributs relationnels qu'ils entretiennent avec d'autres objets comme dans l'analyse formelle de concepts. Cette méthode a permis de classifier plus de trois millions d'objets célestes dans la base SIMBAD. SIMBAD est la plus importante base en astronomie pour la classification des objets célestes et de leurs attributs. Mais SIMBAD n'est pas une ontologie, car en particulier elle ne contient pas de définition formelle des classes d'objets et n'a pas de représentation explicite des relations entre les objets.

Cet article présente une méthodologie et un processus nommés PACTOLE «Property And Class Characterization from Text to OntoLogy Enrichment» pour construire et enrichir une ontologie dans un domaine spécifique. PACTOLE construit semi-automatiquement une ontologie à partir d'une hiérarchie de classes construites manuellement par les experts du domaine, puis l'enrichit avec des définitions extraites de toutes les ressources du domaine spécifique. L'ontologie résultante permet de répondre à des questions complexes des astronomes telles que «*Quel est la classe de l'objet* `TWA`*? Existe-il une classe qui contient à la fois* `Loop 1` *et* `honeycomb nebula`*? Quelle est la classe de* `T Tauri` *et existe-il des télescopes qui l'observent?*»

PACTOLE s'appuie sur trois grandes étapes. D'abord, l'étape de prétraitement des ressources (thésaurus, bases de données, corpus de textes, ...) d'où sont extraits les objets du domaine ainsi que leurs différentes descriptions. La deuxième étape est l'application des méthodes AFC/ARC. Le «schéma d'ontologie» est obtenu après évaluation puis validation, par les experts du domaine, des hiérarchies de concepts et des relations entre ces concepts résultant des méthodes AFC/ARC. Dans cet article, un schéma d'ontologie (par analogie au schéma de base de données) est un ensemble de hiérarchies de concepts et de relations entre ces concepts avant leur représentation en un langage formel. La troisième étape transforme le schéma d'ontologie en un ensemble d'expressions en un langage des Logiques de Descriptions (LD). Ces expressions sont ensuite codées en (OWL) pour donner l'ontologie finale.

Le processus semi-automatique PACTOLE s'appuie l'Analyse Formelle de Concepts (AFC)présentée dans Ganter et Wille (1999) qui permet de construire une hiérarchie de concepts d'objets à partir d'un ensemble d'objets et d'un ensemble de descripteurs (attributs binaires) de ces objets. Cette méthode possède aussi plusieurs propriétés et extensions afin d'enrichir les descriptions des concepts. Elle permet aussi d'enrichir la définition de ces concepts avec la prise en compte de relations entre concepts. Cet enrichissement est possible grâce à l'«Analyse Relationnelle de Concepts» (ARC), présentée dans Rouane-Hacene et al. (2007).

L'utilisation de l'AFC pour la construction d'une ontologie présente plusieurs intérêts. Premièrement, l'AFC est incrémentale, car elle permet l'ajout (ou la suppression) d'objets ou d'attributs au contexte. Deuxièmement, si les hiérarchies de concepts changent (parce que par exemple, le corpus de textes a changé), l'ontologie évolue de façon correcte et consistante. Ainsi, la représentation du schéma de l'ontologie en une logique de descriptions est faite sans souci d'inconsistance. La difficulté liée à la méthode de fouille (AFC) est due au fait qu'elle n'est applicable que sur des tableaux binaires (objets/ attributs). Ainsi il faut transformer les ressources hétérogènes (textes, bases de données, thésaurus) du domaine, en tableaux binaires capables d'être traités par l'AFC. Pour extraire ces tableaux, la méthodologie PACTOLE s'appuie sur des outils de Traitement Automatique de la Langue Naturelle (TALN) pour traiter les corpus de textes, ainsi que sur des outils d'interrogation de bases de données et de thésaurus.

Les sections de cet article sont organisées comme suit. La prochaine section présente un état de l'art des méthodes de construction d'ontologies. La section 3 détaille les différents descripteurs d'objets utilisés dans notre méthodologie de construction d'ontologie. La section 4 introduit les bases de l'Analyse Formelle de Concepts, ainsi que celles de son extension l'ARC. La section 5 présente les différentes étapes de la méthodologie PACTOLE et les opérations d'extraction des connaissances à partir des différentes ressources, ainsi que la représentation en logiques de description des hiérarchies résultantes. PACTOLE est appliquée dans le domaine de l'astronomie. L'évaluation ainsi que l'interaction avec les experts du système PACTOLE sont présentées avec une synthèse dans la section 6. Enfin, une conclusion et les perspectives de ce travail terminent l'article.

2 État de l'art

Afin d'être réactif aux évolutions du domaine, il faut que les processus de construction d'une ontologie reposent sur l'apprentissage mais aussi qu'ils soient guidés par des analystes. Car, ces processus sont destinés à répondre à un besoin d'analystes (experts du domaine). Ces processus de construction d'ontologies reposent sur l'Extraction de Connaissances à partir de Données (ECD) Fayyad et al. (1996). Le processus d'ECD s'articule autour de trois grandes étapes : le prétraitement des ressources, l'application d'une ou de plusieurs méthodes de fouille de données puis l'interprétation et la validation des éléments extraits par l'analyste.

Nous présentons dans cette section les travaux de construction d'une ontologie reposant sur le processus d'ECD et nous divisons cette section en deux sous-sections : la construction d'une ontologie à partir de corpus de textes et la construction d'une ontologie à partir de thésaurus, de bases de données ou ontologies déjà existantes.

2.1 Construction d'ontologie à partir de corpus de textes

Maedche et Staab (2000) présente la construction d'une ontologie à partir de corpus de textes comme étant une modélisation du domaine à partir des données textuelles. Cette modélisation est obtenue par l'application du processus d'Extraction de Connaissance à partir de Textes (ECT) Feldman et Sanger (2007), c'est-à-dire l'adaptation du processus d'ECD aux données textes. Dans les paragraphes suivants, nous présentons un panorama sur les deux premières étapes du processus d'ECT avec : la construction des hiérarchies de concepts et l'extraction des relations transversales. Une relation transversale est une relation binaire entre deux concepts autre que la relation d'ordre partiel.

Prétraitement du corpus de textes. Le prétraitement du corpus de textes consiste à identifier les termes ou unités terminologiques du corpus. Ces termes sont les unités significatives constituées d'un mot (terme simple) ou plusieurs mots (terme complexe) qui désignent une notion univoque à l'intérieur d'un domaine Dubois et al. (1994). Dans les dix dernières années, plusieurs logiciels d'extraction de termes à partir de corpus de textes ont vu le jour : par exemple les logiciels LEXTER Bourigault (1994) et FASTR Jacquemin (1997), qui reposent sur des critères morpho-syntaxiques pour extraire des groupes nominaux à partir de corpus de textes en langues française et anglaise. Les deux logiciels permettent l'utilisation de patrons syntaxiques et de thésaurus existant pour extraire ces termes.

Construction de la hiérarchie de concepts à partir de corpus de textes. Un premier type de travaux repose sur l'extraction de patrons syntaxiques, comme Hearst (1992) qui propose de construire une hiérarchie de termes en détectant la relation d'hyponymie entre les concepts à partir des textes. Les patrons suivants sont extraits : «*X est un Z*» ou «*X, Y, et autres Z*» (par exemple, «*les chiens, les chats et autres animaux...*»), puis définit «*Z*» comme étant un hyponyme de «*X et Y*». Cette méthode est efficace pour des corpus de textes comme les dictionnaires ou les supports pédagogiques mais pas pour des corpus de textes scientifiques comme les textes d'astronomie, car il est difficile de trouver une définition pour chaque type d'objets.

Un second type de travaux repose sur l'hypothèse de distribution de Harris : «*des termes sont similaires s'ils partagent une similarité linguistique*» Harris (1968). Pour la langue anglaise, le travail de Grefenstette (1994) regroupe dans un même concept à l'aide du système SEXTANT (qui effectue l'analyse syntaxique de chaque phrase d'un corpus) tous les noms apparaissant derrière un même ensemble d'expressions (verbe + préposition). Faure et Nedellec (1999) généralisent cette idée dans le système d'apprentissage appelé «Acquisition of Semantic knowledge Using Machine learning methods» (ASIUM). ASIUM regroupe dans un même concept, tous les termes apparaissant comme arguments (sujets, objets, compléments, préposition, ...) du même ensemble de verbes. Pour le français, les travaux de Habert et Nazarenko (1996) regroupent avec le système ZELLIG des termes apparaissant dans des patrons du type «*N prep N*» (nom apparaissant après ou avant un nom et une préposition) ou «*N Adj*» (nom ayant les mêmes adjectifs) dans un même concept. Le travail de Bourigault (1994) avec le système LEXICLASS décompose les termes complexes T_1 T_2 (par exemple, «*star formation*») en T_1 : terme de tête et T_2 : terme d'expansion (par exemple, «*star*» terme de tête - «*formation*» terme d'expansion) puis regroupe les termes T_1 ayant les mêmes termes T_2 dans un même concept et les termes T_2 ayant les mêmes termes T_1 dans un autre concept. Tous ces travaux regroupent successivement les concepts de termes en une hiérarchie de concepts en appliquant des méthodes d'agglomération statistique, comme la méthode du K-Means. Dans ces méthodes d'agglomération statistique, les hiérarchies de concepts changent à chaque application.

Cimiano et al. (2005) s'appuient aussi sur l'hypothèse de Harris mais ils utilisent l'AFC afin d'obtenir à chaque application du processus de construction de la hiérarchie de concepts la même hiérarchie de concepts. Ils montrent que l'AFC donne de meilleurs résultats que plusieurs méthodes d'agglomération statistique. L'AFC permet d'associer un ensemble d'attributs binaires (intension) à un ensemble d'objets (extension).

Les méthodes de construction de hiérarchies de concepts présentées ici ne prennent pas en compte les relations qu'entretiennent les concepts entre eux. Dans ce qui suit, nous présentons des méthodes qui extraient de telles relations afin de proposer des définitions de concepts les plus complètes et caractéristiques possibles.

Extraction des relations transversales entre concepts à partir de corpus de textes. L'extraction de relations transversales permet d'avoir une définition plus complète et plus fine des concepts, car les concepts ne sont plus définis seulement par des attributs binaires mais aussi par des attributs relationnels (relations qui les relient à d'autres concepts).

Certains travaux s'intéressent à l'extraction de relations (des patrons) dans de très grands corpus de textes hétérogènes sur des thèmes généraux (par exemple des journaux). Les linguistes doivent alors définir des formules, puis des schémas à l'aide d'observations du langage. Parmi ces travaux, le «Système Expert d'Exploration Contextuelle» (SEEK) Jouis (1993) vise à

extraire des relations de causalité entre des termes dans un corpus de textes en langue française. SEEK utilise des listes de schémas et des règles morphologiques du type : Si <Conditions> Alors <Actions> OU <Conclusions>. La méthode COATIS de Garcia (1998) reprend l'idée de SEEK en rajoutant une liste de verbes exprimant la relation de causalité comme «créer, faciliter ou encore pousser à». Ces travaux ont été adaptés à la langue anglaise par Goujon (1999). Ces travaux sont très intéressants, mais n'ont été testés que sur la relation de causalité.

D'autres travaux se sont intéressés à l'extraction «semi-automatique» de schéma de relation à partir de corpus de textes. Le logiciel STARTEX développé par Rousselot et al. (1996) extrait à partir d'un verbe v_i (exprimant une relation donnée et choisi par des linguistes) une liste de tous les termes T_1T_2 liés par le verbe v_i, $i.e.$ $T_1v_iT_2$ (par exemple, pour le verbe «containing», il extrait la liste de termes «*galaxy stars*» ou «*associations_of_stars stars*»...), puis, il regroupe dans un concept l'ensemble des termes T_1 d'une même liste et dans un autre concept l'ensemble des termes T_2. Ensuite, il regroupe dans une même relation, les verbes ayant les mêmes listes de termes T_1T_2. Le but de cette méthode est d'extraire à la fois une hiérarchie de termes et une hiérarchie de verbes (hiérarchie de relations). Le travail de Barriere (2002) utilise le même principe pour proposer une étude des différents patrons syntaxiques, en langue anglaise, exprimant une causalité entre deux termes dans un corpus de textes. Par exemple, les patrons du type «T_1 resulting in T_2» ou «T_1 can result in T_2», ... expriment une relation de causalité entre les deux termes T_1 et T_2. Ces méthodes reposent beaucoup sur la personne qui fixe et valide la liste des verbes. Le système PROMETHEE développé par Morin et Martienne (2000) procède à une première extraction de contextes de cooccurrences de termes (supervisée), qu'il analyse afin d'y retrouver des patrons lexico-syntaxiques similaires. Les expressions lexico-syntaxiques dites similaires sont regroupées ce qui permet de proposer des candidats patrons qui, une fois validés, permettent d'extraire de nouveaux termes. Le travail d' Aussenac-Gilles et al. (2000) reprend l'idée de PROMETHEE et propose d'extraire des instances de relations manuellement puis de les généraliser en regroupant les termes qui partagent les mêmes relations avec les autres termes, sans prendre en compte les attributs binaires des termes. Le travail de Maedche et Staab (2000) extrait les relations entre termes en utilisant l'extraction de règles d'associations pour passer de relations entre termes à des relations entre concepts. Cette méthode nécessite une hiérarchie de concepts préalablement construite et n'étiquette pas les relations entre les concepts. L'affectation des noms aux relations est effectuée manuellement par les experts du domaine.

Aucune des méthodes présentées dans cette sous section n'extrait simultanément des hiérarchies de concepts et des relations entre ces concepts. La méthode que nous allons présenter permet de regrouper des objets d'après des attributs binaires mais aussi d'après des attributs relationnels (des relations entre objets) et ainsi d'obtenir des hiérarchies de concepts reliés par des relations transversales.

2.2 Construction d'une ontologie à partir de thésaurus, de bases de données ou ontologies déjà existantes

Hahn et Schulz (2004) proposent une méthode de construction d'ontologies s'appuyant sur la logique de descriptions ALC à partir du vocabulaire intégré «Unified Medical Language System (UMLS)». La méthode construit une hiérarchie de concepts à partir des termes de la ressource et extrait des relations entre ces concepts. Chaque concept est décrit formellement

en ALC afin de pouvoir appliquer des raisonnements sur la structure résultante. Enfin, l'expert est mis dans la boucle de construction pour vérifier et valider l'ontologie résultante. Stumme et Maedche (2001) fusionnent deux ontologies déjà existantes en s'appuyant sur un corpus de textes. Ils utilisent des techniques de TALN pour extraire des termes d'un corpus de textes, puis l'AFC pour relier chacune des deux ontologies existantes aux termes extraits. Les instances, concepts et relations des deux ontologies qui ont pu être reliés aux mêmes termes des textes constituent l'ontologie de fusion. Ces deux méthodes utilisent différentes ressources pour construire une ontologie, mais n'extraient pas de nouvelles connaissances pour enrichir l'ontologie résultante.

Faatz et Steinmetz (2004) enrichissent une ontologie déjà existante, construite manuellement par les experts du domaine d'intérêt. Leur méthode consiste à extraire des propositions d'un corpus de textes (une proposition peut être un mot ou une phrase) puis à demander aux experts de placer ces propositions en tant que nouvelles instances ou nouveaux concepts dans l'ontologie.

Notre méthodologie prend en compte plusieurs types de ressources, comme les corpus de textes, les thésaurus, les bases de données… et construit simultanément des hiérarchies de concepts et des relations transversales entre ces concepts. Dans la section suivante, nous allons définir les différents descripteurs d'objets utilisés pour construire cette ontologie.

3 Les différentes ressources pour la construction de l'ontologie

Les ontologies ne sont pratiquement jamais construites à partir de zéro et plusieurs ressources du domaine sont généralement utilisées pour les construire. Le choix de ces ressources se fait par rapport aux types d'éléments qu'elles contiennent. Dans la suite, nous nommons ces types d'éléments Descripteurs d'Objets (*DO*), qui sont choisis en collaboration avec des experts du domaine (ici, les astronomes) pour définir des classes d'objets. Afin de choisir les descripteurs d'objets, nous donnons une définition d'une ontologie décrivant ses constituants. Une ontologie $\mathcal{O}$ est constituée de :

- un ensemble de concepts C organisé dans une hiérarchie H,
- dans H, les concepts sont reliés hiérarchiquement par une relation de spécialisation $\sqsubseteq$, réflexive, transitive et anti-symétrique (ordre partiel), où $C_1 \sqsubseteq C_2$ veut dire que le concept C_1 subsume le concept C_2,
- un ensemble de relations binaires R spécifiant des paires (C_1, C_2) où C_1 sont les domaines et C_2 les co-domaines des relations R (C_1 et C_2 appartiennent à C).

Soient G un ensemble fini d'objets d'un domaine et M un ensemble d'attributs de ces objets du domaine. Dans ce qui suit, tous les espaces dans les noms d'objets et de leurs attributs sont remplacés par la chaîne «_». Nous définissons trois types de Descripteurs d'Objets (*DO*) sur l'ensemble G.

Descripteur d'objets 1 : Les classes d'objets. Le premier descripteur d'objets (*DO1*) décrit le lien qui peut unir un objet élémentaire à une classe prédéfinie du domaine. L'ensemble de ces classes est noté M_1. Une classe C est définie comme un ensemble d'objets de G ou encore

un élément de 2^G (l'ensemble des parties de G). L'ensemble des objets d'une classe C_1 est appelé «extension de la classe» et noté $ext(C_1)$.

Soit M_1 un ensemble de classes d'objets $g_i \in G$ et $\sqsubseteq$ un ordre partiel défini sur M_1 de la façon suivante : $\forall\ C_1, C_2 \in M_1$, $C_1 \sqsubseteq C_2$ si et seulement si $ext(C_1) \subseteq ext(C_2)$ (relation d'inclusion). L'ensemble ordonné (M_1,$\sqsubseteq$) est appelé «hiérarchie source». Les objets représentent les «feuilles» de cette hiérarchie (nœux terminaux).

Tous les objets d'une classe sont aussi dans les superclasses de celle-ci. Par exemple, `3C_273` est une `Quasar`, `Galaxy` est une sur classe de `Quasar` donc `3C_273` est aussi une `Galaxy`. Dans le contexte de l'astronomie, la hiérarchie de la base SIMBAD joue le rôle de la hiérarchie source du domaine.

Mais les experts du domaine ne donnent pas de définitions aux classes d'objets, ils ne font qu'affecter une classe à un objet. Ainsi, pour définir ces classes nous extrayons un deuxième type d'éléments qui sont les attributs binaires notés (*DO2*).

Descripteur d'objets 2 : Les attributs binaires Dans certaines ressources, les objets du domaine sont décrits par des attributs binaires qui décrivent des caractéristiques propres des objets. Ces attributs peuvent être extraits d'une base de données ou d'un corpus de textes. Dans un corpus de textes, il faut définir ce qui peut être considéré comme l'attribut d'un objet. Par exemple, dans la phrase «*We report the discovery of strong flaring of the object* `HR2517`», le fait que l'objet `HR2517` possède l'attribut `isFlaring` (i.e. `flare` signifie avoir une éruption de plasma à la surface de l'objet) permet aux experts de classifier cet objet dans un type particulier d'étoile. Autre exemple : *«Absolute dimensions for the detached type eclipsing* `Y Cygni` *are derived, based on analyses of new Reticon spectra and existing photometric data.»* Puisque l'objet `Y_Cygni` possède l'attribut `isEclipsing`, permet aux experts de classifier l'objet `Y_Cygni` dans un autre type particulier d'étoile. Dans les textes, l'attribut d'un objet est donné par le verbe que nous renommons sous la forme «`isAttribut`». Une approche similaire a été utilisée par Faure et Nedellec (1999) s'appuyant sur l'hypothèse de Harris (1968) qui considère que dans un corpus, les termes sont similaires s'ils partagent une similarité linguistique. Ici cette similarité est donnée par les dépendances verbes-arguments.

Le deuxième descripteur d'objets (*DO2*) est composé d'un ensemble d'attributs binaires noté M_2. L'ensemble M_2 regroupe les verbes reliés aux objets célestes par des dépendances verbes-arguments dans le corpus de textes et jugés pertinents par les astronomes.

Enfin, comme une ontologie est aussi constituée de relations transversales entre concepts, nous avons besoin d'attributs relationnels (*DO3*). Les attributs relationnels sont des instances de relations.

Descripteur d'objets 3 : Les attributs relationnels. Les objets du domaine peuvent être reliés à d'autres objets. Ces relations sont extraites d'un corpus de textes ou de bases de données. Par exemple, la phrase *«The* `BeppoSAX and ROSAT` *observations of the edge-on spiral galaxy* `SMC`*.»* Indique qu'il existe une relation d'`observation` entre l'objet `SMC` et les télescopes `BeppoSAX` et `ROSAT`. Les télescopes sont extraits des textes à l'aide d'un thésaurus fourni par les astronomes. Les relations entre objets célestes et télescopes sont extraits sous la forme et renommés par notre système sous la forme «`objet_céleste,isObservedBy,télescope`».

Le choix de différencier les attributs binaires des attributs relationnels dépend tout d'abord des experts (ce sont les astronomes qui proposent la relation d'observation entre les objets célestes et les télescopes comme étant intéressante), mais aussi des ressources et des techniques de TALN utilisées. Parfois, le domaine ou le co-domaine d'une relation n'est pas donné ou ne peut pas être extrait et donc l'attribut relationnel n'est considéré que comme un attribut binaire. Un attribut relationnel peut se voir comme la composition de deux attributs binaires. Par exemple, le triplet (`SMC`,`isObservedBy`,`BeppoSAX`) peut être considéré comme un attribut relationnel ou comme la composition de deux attributs binaires (`SMC`,`isObserved`) et (`BeppoSAX`,`isObserving`).

Avant de présenter notre méthodologie, nous détaillons les principes de l'Analyse Formelle de Concepts (AFC) et de son extension l'Analyse Relationnelles de Concepts (ARC).

4 Analyse formelle de concepts et analyse relationnelle de concepts

4.1 Analyse formelle de concepts

L'*Analyse Formelle de Concepts* (AFC) Ganter et Wille (1999) est un formalisme mathématique qui permet de dériver un treillis de concepts à partir d'un contexte formel $\mathbb{K} = (G, M, I)$. L'AFC est utilisée pour modéliser, acquérir et traiter des connaissances pour la construction d'une ontologie, la recherche d'information et la fouille de données.

Le contexte formel $\mathbb{K} = (G, M, I)$ est constitué de G, un ensemble d'objets, de M, un ensemble d'attributs et de I une relation binaire définie sur le produit Cartésien $G \times M$. Dans une table binaire représentant $I \subseteq G \times M$, les lignes correspondent à des objets et les colonnes à des attributs. Le treillis résultant du processus d'AFC est composée de *concepts formels* (ou simplement *concepts*) ordonnés par une relation d'ordre partiel entre concepts appelée *relation de subsomption*. Un concept est une paire (A, B) où $A \subseteq G$ et $B \subseteq M$, A est l'ensemble maximal d'objets partageant l'ensemble des attributs de l'ensemble B (et vice-versa). Dans un concept (A, B), A est appelée l'$extension$ et B l'$intension$ du concept. Les concepts dans un treillis de concepts se définissent par rapport à une *connexion de Galois* qui repose sur deux opérations de dérivation notées par $'$:

$$A' = \{m \in M \mid gIm \; for \; all \; g \in A\}$$
$$B' = \{g \in G \mid gIm \; for \; all \; m \in B\}$$

Un concept (A, B) vérifie que $A' = B$ et $B' = A$, c'est-à-dire,

A' est l'ensemble de tous les attributs de B possédés par les objets de A et de façon duale B' est l'ensemble de tous les objets possédant les attributs de B. plus précisément, pour un concept (A, B), A et B sont des ensembles de fermés pour les opérateurs de dérivation. La relation de subsomption ($\sqsubseteq$) entre un concept et un super concept est définie comme suit : $(A_1, B_1) \sqsubseteq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2$ (ou de façon duale $B_2 \subseteq B_1$). L'ensemble des concepts extraits du contexte formel $\mathbb{K} = (G, M, I)$ et organisés par la relation de subsomption $\sqsubseteq$ est appelé *treillis de concept* et noté $\underline{\mathfrak{B}}(G, M, I)$.

Dans cet article, tous les treillis de concepts que nous présentons, nous utilisons la notation des concepts dite réduite : c'est-à-dire qu'elle s'appuie sur l'héritage à la fois des attributs et des objets entre les concepts du treillis. Les attributs sont placés au plus haut dans le treillis, ce qui veut dire qu'à chaque fois qu'un concept C est étiqueté par un attribut m, tous les descendants de C dans le treillis héritent l'attribut m. De façon duale, les objets sont placés au plus bas dans le treillis, ce qui veut dire qu'à chaque fois qu'un concept C est étiqueté par un objet g, g est hérité «vers le haut» et tous les ancêtres de C le partagent. Ainsi l'extension A d'un concept (A, B) est obtenue en considérant toutes les extensions des descendants du concept C dans le treillis et son intension B est obtenue en considérant toutes les intensions des ascendants du concept C dans le treillis Ganter et Wille (1999).

4.2 Analyse Relationnelle de Concepts

L'AFC permet de traiter des attributs binaires mais pas de prendre en compte les relations entre objets. L'Analyse Relationnelle de Concept (ARC) est une extension de l'AFC qui permet de prendre en compte les attributs binaires et des attributs relationnels qui proviennent de relations entre objets Rouane-Hacene et al. (2007). Les concepts formés sont dits «concepts relationnels» car les intensions des concepts sont définies par des attributs binaires et relationnels Rouane-Hacene et al. (2008).

Les données en ARC sont organisées dans une Famille de Contextes Relationnels (FCR) composé d'un ensemble de contextes $\mathbf{K} = \{\mathbb{K}_i\}$ et d'un ensemble de relations $\mathbf{R} = \{r_k\}$ où $r_k \subseteq G_i \times G_j$ correspondant à des relations entre les objets G_i et G_j. Plus formellement :

Définition 1 (Famille de contextes relationnels) *Une famille de contextes relationnels (*FCR*) est un couple (*$\mathbf{K}, \mathbf{R}$*) où :*

- $\mathbf{K}$ *est un ensemble de contextes* $\mathbb{K}_i = (G_i, M_i, I_i)$,
- $\mathbf{R}$ *est un ensemble de relations* $r_k \subseteq G_i \times G_j$ *où* G_i *et* G_j *sont des ensembles d'objets de contextes dans* $\mathbf{K}$.

Les relations $\mathbf{R}$ sont orientées et peuvent se voir comme des fonctions ensemblistes, c'est-à-dire, r : $G_i \rightarrow 2^{G_j}$. Ainsi, pour toute relation $r_k \subseteq G_i \times G_j$:

- l'ensemble G_i est l'ensemble d'objets du contexte $\mathbb{K}_i$ appelé *domaine* de la relation r_k et noté : *dom :* $\mathbf{R} \rightarrow \mathbf{G}$*, dom(r) =* G_i,
- l'ensemble G_j est l'ensemble d'objets du contexte $\mathbb{K}_j$ appelé *co-domaine* de la relation r_k et noté : *cod :* $\mathbf{R} \rightarrow \mathbf{G}$*, cod(r) =* G_j,

L'ARC propose une approche qui consiste à injecter les liens inter-objets au même temps que les attributs binaires. Ainsi, la construction des treillis se fait simultanément d'après les attributs binaires qu'ils partagent mais aussi d'après les liens inter-objets. La partie relationnelle des concepts est construite par rapport aux liens existant entre les objets. Intuitivement, un attribut relationnel $Q.r.D$ est associé à un concept C dés qu'un individu instance de C est en relation avec un individu de D Rouane-Hacene et al. (2008). Ainsi, il faut déterminer pour un concept C d'un contexte donné et une relation r dont le domaine est ce même contexte, l'ensemble des concepts cibles par la relation r appliquée à l'extension de $C(r(Ext(C)))$.

L'ARC utilise le mécanisme d'*échelonnage relationnel* pour définir les attributs relationnels. Pour une relation $r : G_i \longrightarrow G_j$, qui lie les objets de G_i aux objets de G_j, un attribut

relationnel est crée et noté $Q.r.D$, où Q est un quantificateur et D est un concept du contexte $\mathbb{K}_j$.

Ainsi, pour un objet $g \in G_i$, l'attribut relationnel $r.D$ caractérise la «corrélation» entre g et $r(g) = h$ qui est instance du concept $D = (X, Y)$ dans $\mathbb{K}_j$. Il existe plusieurs niveaux d'échelonnage, un «échelonnage existentiel» où $r(g) \cap X \neq \emptyset$ et un «échelonnage universel» où $r(g) \subseteq X$. Ainsi, à partir de la famille de contextes relationnelles (FCR), l'ARC dérive une famille relationnelle de treillis (FRT), un pour chaque contexte. Un attribut relationnel est interprété comme une relation entre deux concepts, le concept où apparaît l'attribut relationnel est le domaine de la relation et le concept vers lequel pointe l'attribut relationnel est le co-domaine de la relation. La famille relationnelle de treillis est extraite par un processus itératif car l'*échelonnage relationnel* modifie les contextes et par conséquence les treillis correspondants, ce qui entraîne un nouvel échelonnage pour tous les contextes contenant une relation qui possède comme co-domaine les treillis qui ont été modifiés. Ce processus itératif s'arrête quand un point fixe est atteint, c'est-à-dire qu'un nouvel échelonnage ne modifie plus le contexte Rouane-Hacene et al. (2008).

Des exemples de quantification universelle et existentielle sont les suivants :

$\mathtt{C}_1 := \forall\ \mathtt{IsObservedBy.T}_1$ et $\mathtt{C}_2 := \exists\ \mathtt{IsObservedBy.T}_1$.

La première proposition est vraie si tous les objets de l'extension du concept $\mathtt{C}_1$ ne sont observés que par des instances du concept $\mathtt{T}_1$. La deuxième proposition est vraie si pour toute instance x de $\mathtt{C}_2$, il existe au moins une instance y de $\mathtt{T}_1$ tel que y est en relation avec x. Dans ce travail, nous ne nous intéressons qu'à la quantification existentielle.

5 La méthodologie PACTOLE

5.1 Le processus PACTOLE

PACTOLE est une méthodologie de construction semi-automatique d'ontologies à partir de différentes ressources d'un domaine spécifique (thésaurus, bases de données, dictionnaires, corpus de textes,...) Bendaoud et al. (2008b). PACTOLE s'inspire de différentes méthodologies notamment «SENSUS» de Valente et al. (1999), «METHONTOLOGY» de Maedche et Staab (2004) et «TEXT-TO-ONTO» de Maedche et Staab (2004). Des méthodologies «TEXT-TO-ONTO» et «METHONTOLOGY», PACTOLE reprend plusieurs idées telles que l'introduction de l'expert du domaine dans la boucle de construction de l'ontologie afin de valider chaque étape, l'extraction d'un ensemble de termes à partir d'un corpus de textes et la définition des concepts résultant en logique de descriptions, mais aussi l'idée de s'appuyer sur des méthodes symboliques pour extraire les hiérarchies des concepts et les relations entre ces concepts. De «SENSUS» PACTOLE reprend l'idée de s'appuyer sur des ressources du domaine déjà existantes pour construire l'ontologie au lieu de la reconstruire à partir de zéro.

Le processus PACTOLE (voir figure 1) se décompose en six étapes, où chaque étape nécessite la validation de l'expert.

La première étape utilise des outils de Traitement Automatique de la Langue Naturelle (TALN) et d'Extraction d'Information (EI) pour extraire des textes trois types d'éléments d'objets. Le premier type éléments d'objets est constitué par les objets du domaine. Dans le domaine de l'astronomie, ces derniers sont les objets célestes et les télescopes. Le deuxième type d'éléments d'objets est constitué par des paires

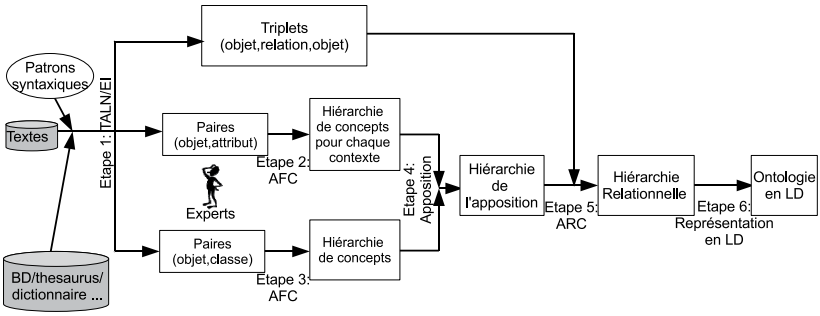

FIG. 1 – ***Le processus*** PACTOLE

(objet_céleste,attribut_binaire) où les objets célestes possèdent une dépendance syntaxique avec les attributs binaires du type : verbe/sujet, verbe/objet et verbe/phrase prépositionnelle. Le troisième type d'éléments d'objets est constitué par la relation «isObservedBy» entre les objets célestes et les télescopes (cette forme de la relation, n'est pas sa forme textuelle mais elle est renommée ainsi par le logiciel GATE). Pour chaque instance de cette relation, un triplet (objet_céleste,isObservedBy,télescope) est extrait.

L'AFC est utilisée sur les éléments (*DO1*) (deuxième étape) et (*DO2*) (troisième étape). A partir des tableaux binaires fournis par les descripteurs, l'AFC produit un treillis de concepts. Ainsi, un treillis de concepts est construit pour chacun des types d'éléments extraits.

La quatrième étape «Apposition» fusionne les deux contextes formels résultant des étapes deux et trois. L'idée ici est d'enrichir les connaissances associées à une ressource du domaine avec les connaissances associées à une autre ressource du domaine. Les deux ressources sont combinées pour donner un seul treillis de concepts grâce à l'apposition de contextes.

La cinquième étape utilise l'ARC pour construire à partir des triplets (objet_céleste,isObservedBy, télescope) (*DO3*) et des contextes issus de la quatrième étape, des treillis où les concepts ne sont pas seulement définis par des attributs binaires mais aussi par les relations qu'ils entretiennent avec les autres concepts.

Enfin, la sixième et dernière étape permet de produire une représentation des treillis résultant de la cinquième étape dans un formalisme de LD $\mathcal{FLE}$. Toutes ces étapes sont détaillées dans les sous-sections suivantes.

5.2 Traitement automatique de la langue et extraction d'information

Cette étape est composée de trois sous-étapes : l'extraction d'objets, l'extraction de paires (objet, attribut) qui est faite par une analyse syntaxique des textes et l'extraction de triplets ($objet_1$,isObservedBy,$objet_2$) qui nécessite l'utilisation d'outils d'extraction d'information et de TALN.

Le système a été utilisé sur un corpus de 11591 résumés d'articles scientifiques dans le domaine de l'astronomie, sélectionnés à l'aide des experts de l'observatoire de Strasbourg[2].

5.2.1 Détection des objets

Il n'existe pas de réelle normalisation pour nommer un objet céleste en astronomie. Ainsi, l'identification des noms des objets à partir du corpus de textes nécessite l'utilisation de deux stratégies complémentaires suggérées par la base SIMBAD. La première stratégie consiste à se servir des noms d'objets déjà répertoriés dans la base SIMBAD (par exemple `Orion`). Ainsi, une simple recherche de chaînes de caractères permet de les localiser dans les textes. La deuxième stratégie utilise des patrons syntaxiques issus d'un dictionnaire de nomenclature de la base SIMBAD pour repérer les noms des objets. Par exemple, l'objet `NGC_6994` est détecté par le patron `NGC_NNNN` où `N` est un chiffre.

Le système a extrait 1382 objets célestes à partir du corpus de textes, ce qui représente 90% des objets présents dans les textes (cette évaluation a été faite par les astronomes). Trois nouveaux objets ont été extraits : `HH 24MMS, S140 IRS3, M33 X-9`, qui n'étaient pas dans la base SIMBAD.

En revanche, quelques objets détectés ne sont pas des objets célestes. Ces erreurs peuvent être expliquées comme suit :

- Des patrons non discriminants : quelques objets dans le domaine de l'astronomie possèdent les mêmes patrons que les objets célestes, comme par exemple, le patron `IRA_X` de SIMBAD extrait l'objet céleste `IRAS_16293` mais aussi l'objet `IRAM_30` qui est un télescope,
- Des abréviations dans les textes : quelques auteurs utilisent des abréviations dans les textes, comme par exemple, `S_180` au lieu de `Sand_180`,
- Des erreurs de typographie dans SIMBAD : quelques noms d'objets sont mal écrits, comme par exemple, `Name_Lupus_2` au lieu de `Lupus_2`.

5.2.2 Extraction des attributs binaires

Les attributs binaires sont extraits par une analyse syntaxique du corpus de textes à l'aide d'un analyseur partiel et robuste, le «Stanford Parser»[3] de Marneffe et al. (2006). Le Stanford Parser extrait, à partir de chaque phrase du corpus de textes, son arbre syntaxique, puis les dépendances entre les verbes et leurs sujets, leurs objets, leurs compléments et leurs phrases propositionnelles. Prenons par exemple la phrase : *«We report results from two COMPTEL observations, in June and October 1991, of the quasars `3C 273` containing binary star `M 83`.»*

Puis, l'analyseur affecte à chaque mot sa catégorie grammaticale : *«We/PRP report/VBP results/NNS from/IN two/CD COMPTEL/JJ observations/NNS ,/, in/IN June/NNP and/CC October/NNP 1991/CD ,/, of/IN the/DT quasars/NN `3C_273`/CD containing/VBG binary/JJ star/NN `M_83`/CD./.»*

Puis l'analyseur syntaxique dérive l'arbre syntaxique de la phrase :

```
(ROOT
  (S
```

[2] http ://astro.u-strasbg.fr/observatoire/

[3] http ://nlp.stanford.edu/software/lex-parser.shtml

```
(NP(PRP We))(VP(VBP report)(NP(NNS results))
 (PP(IN from)(NP(CD two)(JJ COMPTEL)(NNS observations)))
  (, ,)
  (PP(IN in)
   (NP(NP(NNP June))(CC and)(NP(NNP October)(CD 1991)))
  (, ,)
  (PP(IN of) (NP(NP(DT the)(NN quasars)(QP(CD 3C_273))
  (VP(VBG containing)(NP(DT the)(JJ binary)(NN stars)(QP(CD M_83)))
))))))
(. .))))
```

Ensuite, l'analyseur extrait les dépendances entre termes. Nous avons ajouté ici un programme qui ne garde que dépendances du type `(objet_céleste,attribut)` en s'appuyant sur la liste des objets extraits à l'étape précédente :

`det(3C_273-24,the-22)`[4]
`amod(3C_273-18,quasars-17)`[5]
`nsubj(containing-20,3C_273-18)`[6]
`dobj(containing-20,M_83-23)`[7]

Ensuite, une sélection des dépendances les plus significatives est faite par les experts du domaine :

`subject(containing-20,3C_273-18)`
`direct_object(containing-20,M_83-23)`

Enfin, les dépendances sont transformées en paires. Ainsi, la paire `(isContaining,3C_273)` est dérivée de la dépendance `subject(containing-20,3C_273-18)` et la paire `(isContained,M_83)` est dérivée de la dépendance `direct_object(containing-20,M_83-23)`. A noter que le verbe apparaît de deux façons différentes, la première lorsqu'il est associé à son sujet et la deuxième lorsqu'il est associé à son complément d'objet.

La plupart des dépendances (objet,attribut) ne sont que des artefacts linguistiques et ne permettent pas réellement de décrire l'objet. Pour ne sélectionner que les dépendances pertinentes, le système définit des filtres avant l'AFC : le premier filtre ne retient que les attributs qui apparaissent au moins deux fois avec un objet (ceci réduit aussi le bruit introduit par les erreurs de l'analyseur syntaxique). Le deuxième filtre regroupe les synonymes, par exemple, les verbes `consists`, `contains`, `includes`... sont regroupés dans un seul attribut noté «`isIncluding`». Enfin, l'expert filtre manuellement les attributs binaires pour ne garder que les attributs significatifs. Cette dernière étape peut être revue tout au long du processus, car les astronomes peuvent considérer un attribut comme étant intéressant et s'apercevoir qu'il ne l'est pas. Par exemple les attributs «`isPerforming`» ou «`isOscillating`», ou au contraire découvrir qu'un attribut est intéressant après l'étape de construction de la hiérarchie des concepts, par exemple l'attribut «`isRotating`».

Cette méthode d'extraction d'attributs binaires a conduit à découvrir certaines corrélations entre les objets célestes et des attributs qui n'existaient pas dans SIMBAD et que les astronomes ont définis comme de nouvelles connaissances. Par exemple le fait

[4]det : determinant, 3C_273-24 veut dire que le terme 3C_273 est à la position 24 dans la phrase
[5]amod : adjectival modifier
[6]nsubj : noun_subject
[7]dobj : direct_object

que les objets «59_Aurigae, V1208_Aql» possède l'attribut «isPulsing», l'objet «MM_Herculis» possède l'attribut «isEclipsing» ou les objets «AB_Dor, OJ_287» possèdent l'attribut «isFlaring».

5.2.3 Extraction des relations transversales

L'extraction des relations transversales à partir du corpus de textes est faite à l'aide du système GATE «General Architecture for Text Engineering»[8]. GATE est un outil de TALN et d'Extraction d'Information (EI) qui permet d'extraire des relations entre des objets dans un corpus de textes en utilisant des méta-règles. Par exemple, dans le domaine de l'astronomie, une des relations qui peut caractériser les liens entre les objets célestes et les télescopes est la relation isObservedBy. Ainsi, à partir de la phrase suivante : «The BeppoSAX and ROSAT observations of the edge-on spiral galaxy SMC» et la liste des objets célestes et la liste des télescopes extraites du corpus de textes, GATE utilise la méta-règle suivante (que nous avons écrit) pour annoter le texte avec la relation isObservedBy :

```
Rules : IsObservedBy (
{telescope}
({Token.string == «and»}{telescope}?)*
{Token.string == «observation»}
{Token.category == OF}
{object}
) :observation_label
```

Le résultat de l'application de cette règle sur la phrase précédente produit les deux triplets (BeppoSAX,isObservedBy,SMC) et (ROSAT,isObservedBy,SMC).

5.3 Construction d'un treillis à partir de la hiérarchie source

Le premier type de Descripteur d'Objets (*DO1*) est représenté par la hiérarchie source (M_1,$\sqsubseteq$). Les objets sont classifiés dans des classes prédéfinies et ces classes sont organisées dans un arbre, c'est-à-dire qu'il n'y a pas d'héritage multiple dans cette hiérarchie. Nous souhaitons transformer cet arbre en un treillis de concepts et pour cela, il faut construire un contexte formel.

Soit G l'ensemble des objets et $\subseteq$ la relation d'inclusion sur l'ensemble des parties de G. Le couple (G,$\sqsubseteq$) dénote un ensemble ordonné ainsi que la hiérarchie source (M_1,$\sqsubseteq$).

Nous transformons la hiérarchie source en un contexte formel $\mathbb{K}_1 := (G, M_1, I_1)$ comme suit : G est l'ensemble des objets du domaine, M_1 est l'ensemble des classes prédéfinies dans la hiérarchie source et I_1 est la relation qui assigne à chaque objet sa classe et toutes ses superclasses. Un exemple est donné en figure 2, détaillant un contexte $\mathbb{K}_1$ d'objets célestes, les classes extraites de SIMBAD ainsi que le treillis de concepts correspondant.

5.4 Construction d'un treillis à partir d'attributs binaires

Le deuxième type de Descripteur d'Objets *(DO2)* est donné par les attributs extraits des textes, qui permet de construire le contexte formel $\mathbb{K}_2 := (G, M_2, I_2)$ où : G est l'ensemble

[8] http ://gate.ac.uk/

	Classes SIMBAD					
	Quasar	Galaxy	Asso._of_Stars	T_Tau_type_Star	Eclipsing_Binary	Star
Cen_A		×				
3C_273	×	×				
TWA			×			
Per_OB2			×			
T_Tauri				×		×
Y_Cygni					×	×
V773_Tau				×		×
Algol					×	×

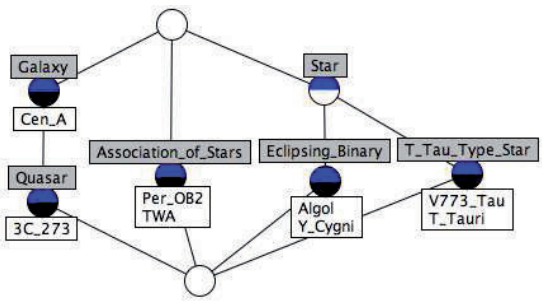

FIG. 2 – *Le contexte* $\mathbb{K}_1=(G, M_1, I_1)$ *représentant les objets et leurs classes extraites de la base* SIMBAD *et son treillis de concepts associé. La classe* `Association_of_Stars` *est écrit sous l'abréviation* `Asso._of_Stars`

des objets, M_2 l'ensemble des attributs extraits des textes et $I_2 \subseteq G \times M_2$ est la relation qui exprime que l'objet g possède l'attribut m_2 (ici l'ensemble G des objets est le même ensemble dans les deux contextes $\mathbb{K}_1$ et $\mathbb{K}_2$). Un exemple de contexte $\mathbb{K}_2$ est présenté à la figure 3.

	Attributs binaires						
	isObserved	isIncluding	isEmitting	isEclipsing	isExpanding	isIncluded	isFlaring
Cen_A	×	×					
3C_273	×	×	×				
TWA	×	×			×		
Per_OB2	×	×					
T_Tauri	×		×			×	×
Y_Cygni	×		×	×		×	
V773_Tau	×		×			×	×
Algol	×		×	×		×	

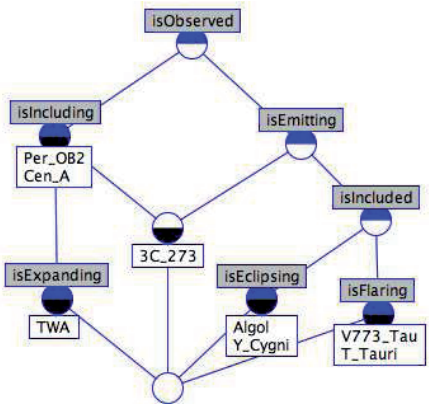

FIG. 3 – *Le contexte* $\mathbb{K}_2=(G, M_2, I_2)$ *représentant les objets et leurs attributs binaires extraits des textes et son treillis de concepts associé*

5.5 Affectation d'attributs binaires à des classes d'objets

Dans cette partie, nous proposons une méthode d'affectation des attributs binaires extraits du corpus de textes aux classes de la hiérarchie source de la base SIMBAD. Cette méthode utilise une propriété de l'AFC qui est l'«apposition de contextes» définie par Ganter et Wille (1999).

Définition 2 (Apposition de contextes) *Soit* $\mathbb{K}_1 = (G_1, M_1, I_1)$ *et* $\mathbb{K}_2 = (G_2, M_2, I_2)$ *deux contextes formels. Si* $G = G_1 = G_2$ *et* $M_1 \cap M_2 = \emptyset$ *alors* $\mathbb{K} := \mathbb{K}_1 | \mathbb{K}_2 := (G, M_1 \cup M_2, I_1 \cup I_2)$. $\mathbb{K}$ *est appelé l'*apposition *des deux contextes* $\mathbb{K}_1$ *et* $\mathbb{K}_2$.

L'apposition de contextes $\mathbb{K}_o = (G_o, M_o, I_o)$ des deux contextes $\mathbb{K}_1 = (G, M_1, I_1)$ et $\mathbb{K}_2 = (G, M_2, I_2)$ peut se voir comme suit : G_o est l'ensemble des objets (le même ensemble pour $\mathbb{K}_1$ et $\mathbb{K}_2$), $M_o := M_1 \cup M_2$ où $M_1 \cap M_2 \neq \emptyset$ et où M_1 est l'ensemble des classes du contexte $\mathbb{K}_1$ (extraits de la hiérarchie de la base SIMBAD) et M_2 est l'ensemble des attributs du contexte $\mathbb{K}_2$ (extraits du corpus de textes) et $I_o := I_1 \cup I_2$. L'apposition de contextes $\mathbb{K}_o$ est présentée dans la table 1 et le treillis de concepts résultant est présenté à la figure 4.

TAB. 1 – *L'apposition de contextes* $\mathbb{K}_{\bowtie} = (G_o, M_o, I_o)$

	Attributs binaires							Classes SIMBAD					
	isObserved	isIncluding	isEmitting	isEclipsing	isExpanding	isIncluded	isFlaring	Quasar	Galaxy	Association_of_Stars	T_Tau_Type_Star	Eclipsing_Binary	Star
Cen_A	×	×							×				
3C_273	×	×	×					×	×				
TWA	×	×			×					×			
Per_OB2	×	×								×			
T_Tauri	×		×			×	×				×		×
Y_Cygni	×		×	×		×						×	×
V773_Tau	×		×			×	×				×		×
Algol	×		×	×		×						×	×

Dans le treillis $\underline{\mathfrak{B}}(\mathbb{K}_o)$, les classes se voient associer des attributs binaires. Par exemple le concept {Star, Eclipsing_Binary, isObserved, isEmitting, isIncluded, isEclipsing} {Algol, Y_Cygni} permet d'associer à la classe Eclipsing_Binary, de la hiérarchie SIMBAD les attributs binaires {isObserved, isEmitting, isIncluded, isEclipsing}. Les experts peuvent interpréter ces éléments de deux façons différentes. Le premier est de considérer que les attributs binaires

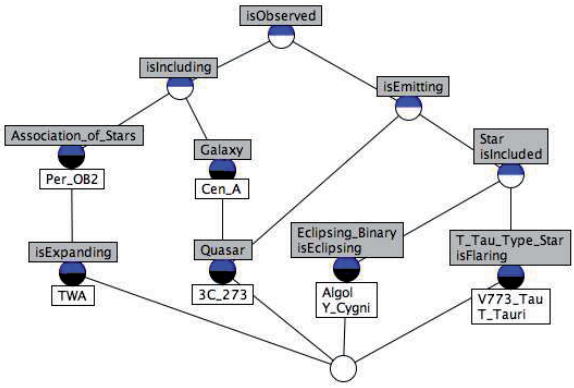

FIG. 4 – *Le treillis de l'apposition de contextes* $\mathbb{K}_o = (G_o, M_o, I_o)$

{`isObserved, isEmitting, isIncluded, isEclipsing`} définissent la classe `Eclipsing_Binary` une sous classe de la classe `Star`. La deuxième est de considérer que les attributs binaires {`isObserved, isEmitting, isIncluded, isEclipsing`} définissent une sous classe de la classe `Eclipsing_Binary` et ainsi créent une nouvelle classe dans la hiérarchie source. L'apposition de contextes permet donc non seulement d'enrichir la hiérarchie source avec des attributs binaires mais aussi d'enrichir cette hiérarchie source avec de nouvelles classes.

5.6 Construction d'un treillis relationnel

Le troisième type de Descripteur d'Objets (*DO3*) recouvre les relations entre les objets qui sont prises en compte par l'Analyse Relationnelle de Concepts (ARC) introduite dans Rouane-Hacene et al. (2007). Un objet est décrit par des attributs binaires mais aussi par les relations qu'il entretient avec d'autres objets (attributs relationnels). Ainsi, des objets célestes peuvent être regroupés dans la même classe parce qu'ils sont observés par le même télescope.

Dans notre exemple, la FCR $(\mathbf{K}, \mathbf{R})$ est composée de deux contextes et d'une relation :

- $\mathbb{K}_O = (G_O, M_O, I_O)$ le contexte des objets célestes,
- $\mathbb{K}_T = (G_T, M_T, I_T)$ le contexte des télescopes,
- $r_1 \subseteq G_O \times G_T$ la relation `IsObservedBy` entre l'ensemble des objets célestes et l'ensemble des télescopes.

La première étape consiste à construire les deux contextes formels $\mathbb{K}_O$ et $\mathbb{K}_T$. Nous avons déjà le contexte des objets célestes $\mathbb{K}_O$ et le treillis correspondant à la figure 4, reste à construire le contexte et le treillis des télescopes.

5.6.1 Construction du treillis des télescopes

Le contexte formel $\mathbb{K}_T = (G_T, M_T, I_T)$ se compose qu'un ensemble de télescopes G_T (ces télescopes sont extraits du corpus de textes à l'aide d'un thésaurus), un ensemble d'attri-

buts binaires M_T (extraits d'une base de données) et une relation binaire $I_T \subseteq G_T \times M_T$ où $I_T(g_T, m_T)$ veut dire que le télescope g_T possède l'attribut m_T.

	hasPerigeeIsLow	hasPerigeeIsHigh	hasAShortOrbitalPeriod	hasALongOrbitalPeriod	isLight	isHeavy
BeppoSAX	×		×		×	
XMM-Newton		×		×		×
Chandra		×		×	×	

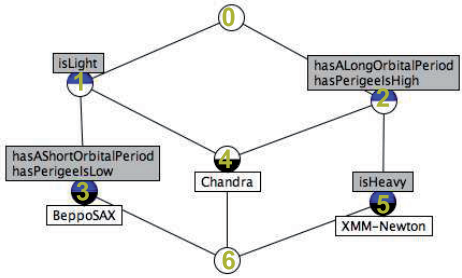

FIG. 5 – *Le contexte des télescopes $\mathbb{K}_T = (G_T, M_T, I_T)$ et son treillis de concepts associé*

isObservedBy	BeppoSAX	XMM-Newton	Chandra
Cen_A	×		
3C_273	×		
TWA	×		
T_Tauri		×	
V773_Tau			×

TAB. 2 – *La relation* `isObservedBy`

L'ARC va mettre en jeu la relation `isObservedBy` (table 2), l'apposition des contextes d'objets célestes (table 4), et le contexte des télescopes (figure 5). Les treillis finaux obtenus par le processus de l'ARC sont donnés à la figure 6.

La table 3 présente le contexte des objets célestes et explique le phénomène d'échelonnage utilisé pour construire le treillis de concepts à gauche de la figure 6. Par exemple, l'objet `3C_273` est relié au télescope `BeppoSax` par la relation `isObservedBy`, cela signifie que l'objet `3C_273` possède un attribut relationnel dont le co-domaine mentionne tous les concepts ayant `BeppoSax` comme instance. Ainsi, `3C_273` possède les attributs relationnels `isObservedBy:T0`, `isObservedBy:T1` et `isObservedBy:T3`.

A la figure 6, les objets sont regroupés par les attributs binaires qu'ils partagent mais aussi par les relations qu'ils entretiennent avec les télescopes. Par exemple, l'ensemble des objets {`3C_273`, `Cen_A`} (concept C5) possède en commun les attributs binaires {`Galaxy`, `isObserved`, `isIncluding`} mais aussi les relations {`isObservedBy:T0`, `isObservedBy:T1`, `isObservedBy:T3`}. Cette méthode a aussi été appliquée dans un autre domaine la microbiologie dans Bendaoud et al. (2008a).

TAB. 3 – *Le contexte de l'échelonnage* $\mathbb{K}_o = (G_o, M_o, I_o)$

	Attributs binaires							Classes SIMBAD						Attributs relationnels					
	isObserved	isIncluding	isEmitting	isEclipsing	isExpanding	isIncluded	isFlaring	Quasar	Galaxy	Association_of_Stars	T_Tau_Type_Star	Eclipsing_Binary	Star	isObservedBy:T0	isObservedBy:T1	isObservedBy:T2	isObservedBy:T3	isObservedBy:T4	isObservedBy:T5
Cen_A	×	×							×					×	×		×		
3C_273	×	×	×					×	×					×	×		×		
TWA	×	×			×					×				×	×		×		
Per_OB2	×	×								×									
T_Tauri	×		×			×	×				×		×	×		×			×
Y_Cygni	×		×	×		×						×	×						
V773_Tau	×		×			×	×				×		×	×		×			×
Algol	×		×	×		×						×	×						

5.7 Représentation des concepts formels de l'ARC en logique de descriptions

Pour représenter les concepts formels de l'ARC, nous devons choisir un langage de représentation des connaissances. Ici nous avons fait le choix de la LD $\mathcal{FLE}$ comme suggéré dans les travaux de Rouane-Hacene et al. (2007).

Le langage des LD $\mathcal{FLE}$ inclut les constructeurs $\top$ (top), $\bot$ (bottom), $C \sqcap D$ (conjonction de concepts), $\forall r.C$ et $\exists r.C$ (quantificateurs universel et existentiel). Cet ensemble de constructeurs représente l'ensemble minimal permet de représenter les concepts formels du treillis en particulier : les objets, les attributs binaires, les attributs relationnels, les concepts primitifs, les concepts définis et la relation de subsomption entre concepts.

5.7.1 La représentation des concepts formels en $\mathcal{FLE}$

La transformation des treillis de l'ARC en une base de connaissances en $\mathcal{FLE}$ se schématise par une transformation appelée τ.

$\tau : \underline{\mathfrak{B}}(G_f, M_f, I_f) \longrightarrow TBox \cup ABox$, où $\underline{\mathfrak{B}}(G_f, M_f, I_f)$ sont les treillis de l'ARC (figure 6) et la TBOX et l'ABOX sont les composants en LD de l'ontologie finale. Le détail de τ est donné avec des exemples pour chaque type de transformation :

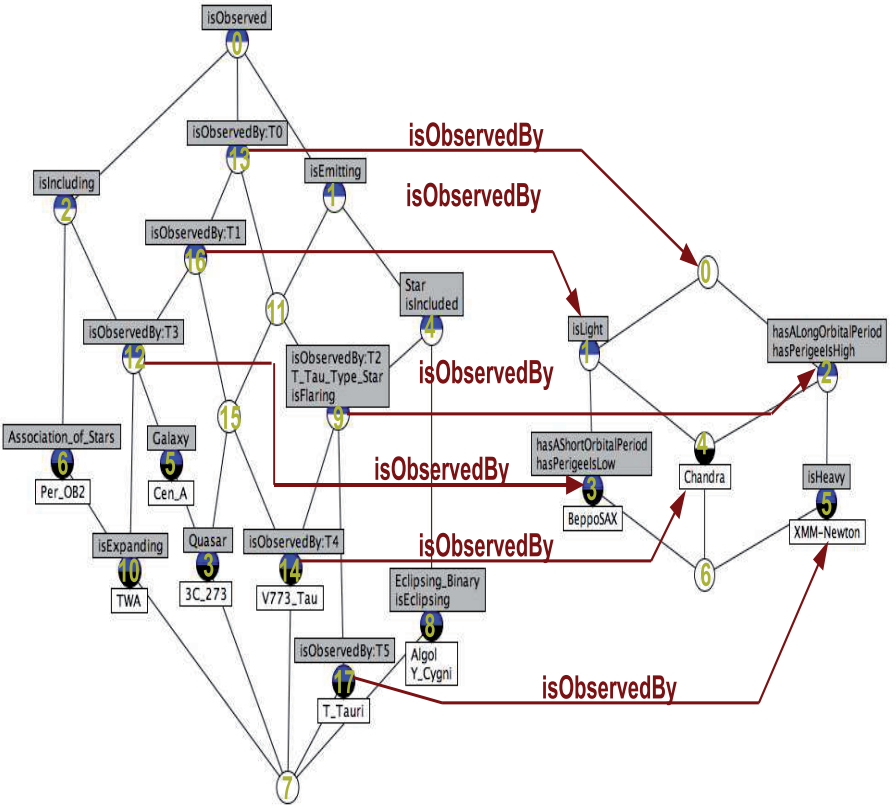

FIG. 6 – *Les treillis résultant de l'application de l'*ARC *avec une notation réduite appliqué aux descripteurs d'objets (DO3).*

- un attribut $m_1 \in M_1$, où $\underline{\mathfrak{B}}(G, M_1, I_1)$ dénote le treillis associé à la hiérarchie source, est transformé en un concept atomique dans la TBOX. Par exemple : $\tau(\texttt{Quasar}) = \texttt{Quasar}$,
- un attribut du contexte $\underline{\mathfrak{B}}(G, M_2, I_2)$, qui dénote le contexte des objets et de leurs attributs binaires, est transformé en une expression conceptuelle : $\tau(\texttt{m}_2) = \exists \texttt{m}_2 . \top$. Par exemple : $\tau(\texttt{isEmitting}) = \exists \texttt{isEmitting} . \top$,
- un attribut relationnel $r \in R$ est transformé dans la TBOX en un rôle atomique $\tau(r)$. Par exemple : $\tau(\texttt{isObservedBy}) = \texttt{isObservedBy}$,
- un concept formel $C = (X, Y)$ en un concept défini par une conjonction de concepts primitifs et de quantificateurs existentiels de relations. Par exemple : Le concept $\texttt{C}_9$ de la figure 6 a une intension comprenant les attributs

{Star,T_Tau_Star,isObserved,isEmitting,isFlaring,isIncluded, isObservedBy:T0,isObservedBy:T2}. Sa forme en $\mathcal{FLE}$ est : C_9 = Star ⊓ T_Tau_Star ⊓ ∃isObserved.⊤ ⊓ ∃isEmitting.⊤ ⊓ ∃isFlaring.⊤ ⊓ ∃isIncluded.⊤ ⊓ ∃IsObservedBy.T0 ⊓ ∃IsObservedBy.T2

– la relation de subsomption entre concepts (du treillis) est transformée en une inclusion générale de concepts. Par exemple : la relation de subsomption $C_0 \sqsubseteq C_2$ dans le treillis devient $\tau(C_0) \sqsubseteq \tau(C_2)$,
– un objet $g \in G$ est transformé en un individu $\tau(g)$. Par exemple : l'objet 3C_273 devient l'instance τ(3C_273).

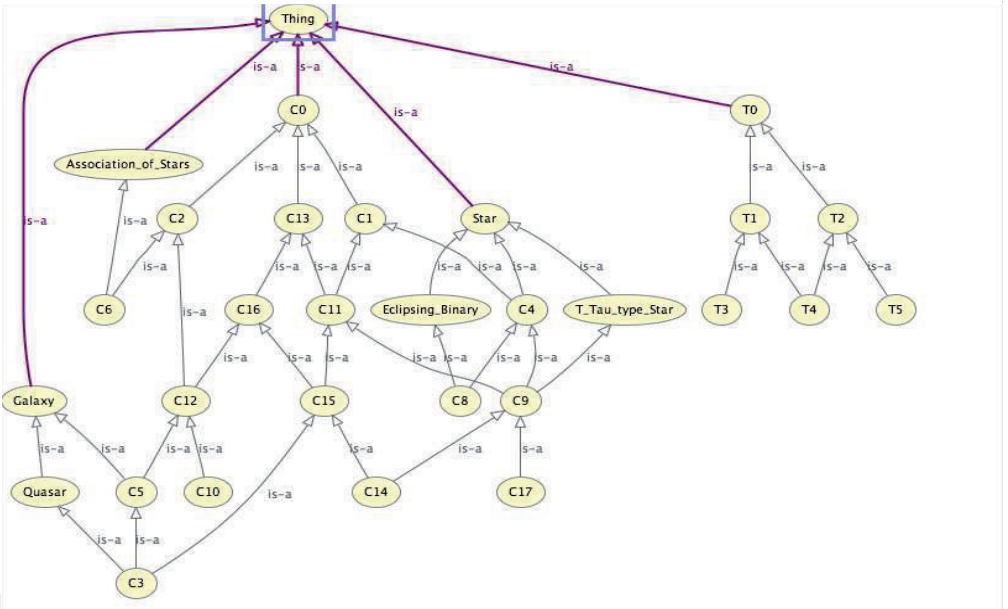

FIG. 7 – *L'ontologie provenant des treillis de l'*ARC *et représentée en logique de descriptions et éditée avec le logiciel* PROTÉGÉ.

L'ensemble de ces transformations produit l'ontologie finale présentée à la figure 7 éditée avec le logiciel PROTÉGÉ[9]. Nous donnons ici quelques exemples de concepts définis extraits de la figure 6 :

C1 = ∃isObserved.⊤ ⊓ ∃isEmiting.⊤

C3 = Galaxy ⊓ Quasar ⊓ ∃isObserved.⊤ ⊓ ∃isEmitting.⊤ ⊓ ∃isIncluding.⊤ ⊓ ∃isObservedBy.T0 ⊓ ∃isObservedBy.T1 ⊓ ∃isObservedBy.T3

Dans le treillis des télescopes :

T5 = ∃isHeavy.⊤ ⊓ ∃hasPerigeeIsHigh.⊤ ⊓ ∃hasALongOrbitalPariod.⊤

[9]http ://protege.stanford.edu/

5.7.2 Raisonnement avec les concepts de l'ontologie

Les opérations de raisonnement associées à l'ontologie sont l'instanciation, la subsomption de concepts, la comparaison de concepts et la détection du co-domaine d'une relation. Les détails de ces opérations de raisonnement sont donnés sur des exemples.

Instanciation de concepts. L'instanciation de concepts consiste à trouver le concept d'un objet dans la hiérarchie des concepts. Cette opération permet de répondre à des questions comme «*trouver le concept* $C(o_1)$ *de l'objet* o_1 *dans l'ontologie*». La réponse à cette question utilise la $classification\ progressive$ qui consiste à partir du $\top$ dans la hiérarchie et de descendre progressivement dans la hiérarchie jusqu'à arriver au concept de l'objet o_1. Prenons par exemple l'instance o_1 ayant les attributs $\{a, b\}$ et appartenant à la classe C_3 dans la hiérarchie source, c'est-à-dire C_3 est un concept atomique dans l'ontologie. Le concept de o_1 est défini par $C(o_1) \sqsupseteq C_3 \sqcap \exists a.\top \sqcap \exists b.\top$:

1. Si $C(o_1) = C_3 \sqcap \exists a.\top \sqcap \exists b.\top$ existe, alors l'objet o_1 est instance du concept $C(o_1)$,
2. Si $C(o_1) = C_3 \sqcap \exists a.\top \sqcap \exists b.\top$ n'existe pas, alors l'objet o_1 est instance des concepts C appartenant à l'ensemble des parties de l'ensemble $E = \{C3, \exists a.\top, \exists b.\top\}$ noté $P(E)$.

L'application de cette opération dans un exemple d'astronomie consiste à répondre à des questions telle que «*Quelle est le concept de l'instance* `TWA` *qui possède les attributs* {`isObserved`,`isExpanding`, `isIncluding`} *et appartient à la classe* {`Association_of_stars`} *dans la hiérarchie de la base* SIMBAD». La réponse est le concept $X \sqsupseteq$ `Association_of_Stars` $\sqcap \exists$ `isObserved`.$\top \sqcap \exists$ `isExpanding`.$\top \sqcap \exists$ `isIncluding`.$\top$. Ce concept est le concept `C10` dans l'ontologie de la figure 7.

Comparaison de concepts. Dans les domaines scientifiques tels que l'astronomie, les hiérarchies sources sont généralement des arbres. Ce type de structure est insuffisant pour classifier des objets aussi complexes que les objets célestes. Prenons par exemple les deux objets `Loop_1` et `Honeycomb_nebula` en astronomie. L'objet céleste `Loop_1` est classifié en tant que `SuperNova_Remnant_Candidate` dans SIMBAD (`SuperNova_Remnant_Candidate` est la classe des objets dont les experts supposent que c'est des `SuperNova_Remnant` mais ils ne sont pas sures) et `Honeycomb_nebula` est classifié en tant que `SuperNova_Remnant` dans SIMBAD. Les deux objets ne se retrouvent qu'à la racine ($\top$) de SIMBAD alors qu'ils sont supposés être dans des classes presque identiques.

La deuxième opération de raisonnement consiste à rechercher un concept dont deux objets sont instances. Cette opération est aussi faite par classification progressive, car pour trouver le concept des deux instances o_1 et o_2, il faut trouver les concepts $C(o_1)$ et $C(o_2)$, puis chercher le plus petit subsumant commun (LCS) de $C(o_1)$ et $C(o_2)$.

Plusieurs cas sont possibles pour $LCS(CPS(o_1), CPS(o_2))$ avec les conditions que $CPS(o_1) \neq \top$ et $CPS(o_2) \neq \top$:

1. Si $C(o_1) \sqsubseteq C(o_2)$ alors $LCS(C(o_1), C(o_2)) = C(o_2)$,
2. Si $C(o_2) \sqsubseteq C(o_1)$ alors $LCS(C(o_1), C(o_2)) = C(o_1)$,

3. Si C(o_1) = C(o_2) alors LCS(C(o_1),C(o_2)) = C(o_1) = C(o_2),
4. Si C(o_1) $\sqcup$ C(o_2) $\neq \top$ alors LCS(C(o_1),C(o_2)) existe,
5. Si C(o_1) $\sqcup$ C(o_2) = $\top$ alors LCS(C(o_1),C(o_2)) n'existe pas,

Dans les quatre premiers cas, les deux objets sont instances d'un même concept LCS(C(o_1),C(o_2)) qui sera présenté aux experts comme potentiellement une nouvelle classe d'objets. Dans le cinquième et dernier cas, il n'existe pas de concept dont les deux objets sont instances.

Nous appliquons cette méthode sur l'exemple des deux objets Loop_1 et Honeycomb_nebula.
C(Loop_1) = C_{367}, C(Honeycomb_nebula) = C_{368} et $C_{367} \sqcup C_{368} \neq \top$. Ainsi nous sommes dans le quatrième cas où LCS(C(o_1),C(o_2)) existe et il est égal à C267 dans l'ontologie :
C267 = Stars $\sqcap \exists$ isObserved.$\top \sqcap \exists$ isIncluding.$\top \sqcap \exists$ isEmitting.$\top$. Le concept C267 est présenté aux experts comme nouvelle classe potentielle d'objets.

Détection du co-domaine d'une relation. La troisième opération nous permet de détecter le co-domaine d'une relation, par exemple de trouver avec quels télescopes on peut observer l'objet o_i. Considérons pour exemple la question «Quel est le concept de l'objet T_Tauri et avec quel télescope est-il observé ?». La première étape consiste à instancier l'objet T Tauri, en utilisant l'opération d'instanciation, le C(T_Tauri) = C17. Ensuite il faut trouver les concepts des télescopes qui sont en relation avec le concept C17. D'après la définition de C17 dans l'ontologie, il existe une relation isObservedBy entre les concepts C17 et T5 ce qui signifie que T_Tauri est observé par les télescopes instances du concept T5. Puisque le concept T_5 possède comme instance le télescope XMM-Newton, alors l'objet T_Tauri est observé par le télescope XMM-Newton.

6 Interprétation et évaluation

6.1 L'interaction entre l'expert et le système PACTOLE

L'expert est invité à interpréter les résultats du système et à détecter les problèmes dans l'ontologie. Les raisons de ces problèmes peuvent être de deux natures différentes : (1) les ressources du domaine peuvent contenir du bruit ou alors ce bruit est issu de la première étape d'extraction d'informations et d'analyse des textes, (2) les experts ne sont pas satisfaits par l'ontologie résultante et ils veulent l'adapter à leurs besoins. Quelles que soient les raisons de ces problèmes, une étape de prétraitement est mise à disposition des experts. Cette étape définit des filtres qui peuvent être appliqués sur les ressources par les experts afin de réappliquer l'AFC/ARC et avoir une nouvelle version de l'ontologie. Cette étape de prétraitement enregistre toutes les opérations effectuées par les experts pour garder une trace des actions des experts et ne pas les obliger à refaire le même travail pour chaque nouveau corpus par exemple. Nous définissons différents types d'opérations qui dépendent du type des descripteurs d'objets (*DO1*, *DO2* ou *DO3*).

Les opérations sur la hiérarchie source (*DO1*).

- *Ajouter une nouvelle classe.* Une nouvelle classe peut être insérée dans le thésaurus. Cette opération nécessite l'ajout d'une colonne dans le contexte formel représentant la hiérarchie source et l'expert doit assigner les objets appropriés à cette classe.
- *Changer la classe d'un objet.* Pour changer la classe d'un objet, la ligne décrivant cet objet dans le contexte formel doit être considérée : l'ancienne classe et toutes ses superclasses doivent être supprimées comme attributs de l'objet. La nouvelle classe et toutes ses superclasses doivent être ajoutées comme attributs de l'objet.
- *Supprimer une classe.* Cette opération est définie mais n'est jamais utilisée dans notre expérimentation. Supprimer une classe revient à supprimer la colonne de cette classe dans le contexte formel de la hiérarchie source.

Les opérations sur *DO2* ou *DO3*. Les outils de TALN ou ceux de l'extraction d'informations introduisent beaucoup de bruit car le niveau linguistique est beaucoup plus détaillé que le niveau de l'ontologie. Les experts doivent supprimer le bruit introduit par ces outils. Les opérations suivantes vont leur permettre d'y arriver :

- *Fusionner des attributs.* Cette opération permet aux experts de fusionner les attributs qu'ils considèrent comme synonymes. Par exemple : si l'expert veut fusionner les attributs `isIncluding` et `isContaining`, alors l'opération consiste à fusionner les deux colonnes de ces deux attributs. Cette opération est équivalente au constructeur logique «ou» dans les colonnes correspondantes dans le contexte formel.
- *Supprimer un attribut d'un objet.* Un attribut peut être mal affecté à un objet à cause par exemple du processus de TALN ; l'expert peut donc le supprimer. Dans le contexte formel, cela revient à supprimer la corrélation (objet, attribut) correspondante.
- *Supprimer un attribut pour tous les objets.* Lorsque les experts interprètent l'ontologie, ils peuvent s'apercevoir qu'un attribut n'est pas très significatif ou inutile. La colonne avec cet attribut est alors supprimée dans le contexte formel.
- *Ajouter un attribut à un ensemble d'objets.* Si l'expert considère qu'un attribut intéressant n'est pas utilisé pour décrire un ensemble d'objet, il peut décider de le rajouter pour tous les objets concernés (d'après les paires extraites du processus de TALN). Cette opération revient à ajouter une colonne dans le contexte (objets,attributs_binaires).

Les opérations sur les attributs relationnels sont similaires à celles qui sont appliquées sur les attributs binaires.

6.2 Évaluation du système

Dans cette partie, une évaluation de chaque étape du système PACTOLE est présentée. Le système PACTOLE a été appliqué sur un corpus de 11591 résumés du journal A&A «Astronomy and Astrophysics» de 1994 à 2002.

Extraction des descripteurs d'objets. L'analyseur syntaxique Stanford Parser a analysé 68.5% des phrases du corpus, où la taille maximale des phrases analysées se situe entre 31 et 36 mots (d'après la complexité syntaxique de la phrase). Le système extrait trois ensembles différents de dépendances syntaxiques entre le verbe et ses arguments, nommés respectivement : SO, SOC et SOCP (voir la table 4) où :

– SO : subject(object,verb) + object(object,verb),
– SOC : SO + complement(object,verb),
– SOCP : SOC + preposition_X(object,verb), où X peut être (in, of, ...).

TAB. 4 – *Les résultats de l'analyse syntaxique du corpus de textes*

	SO			SOC			SOCP		
	Paires	Objets	Attributs	Paires	Objets	Attributs	Paires	Objets	Attributs
11591 résumés	384	209	14	401	211	14	1709	470	23

L'ensemble des dépendances SOCP permet d'extraire plus de paires, plus d'attributs et plus de concepts que les autres ensembles de dépendances (voir table 4). Ce qui paraît naturel : plus il y a de dépendances prises en compte, plus il y a de paires extraites et par conséquent plus d'objets, d'attributs et de concepts.

Le logiciel d'extraction d'information GATE a permis d'extraire plus de 200 phrases représentant des instances de la relation `isObservedBy` entre 10 télescopes et 64 objets célestes.

Les treillis de concepts de chaque descripteur d'objets. A partir du contexte formel $\mathbb{K}_1 = (G, M_1, I_1)$, où : G l'ensemble des objets célestes, M_1 l'ensemble des classes de la hiérarchie source et I_1 la relation binaire où $I_1(g, m_1)$ veut dire que l'objet g est attribué à la classe m_1 dans la hiérarchie source. L'AFC produit un treillis de 94 concepts, avec un contexte formel de 470 objets et 92 attributs (nous avons affecté la classe «`Object_of_unknown_nature`» aux objets qui n'apparaissaient pas dans SIMBAD) .

A partir du contexte formel $\mathbb{K}_2 = (G, M_2, I_2)$, où : G l'ensemble des objets célestes, M_2 l'ensemble des attributs binaires et I_2 la relation binaire où $I_2(g, m_2)$ veut dire que l'objet g possède l'attribut m_2. Si nous utilisons les dépendances SO ou SOC, l'AFC produit un treillis de 30 concepts. Si nous utilisons toutes les dépendances SOCP, l'AFC produit un treillis de 70 concepts.

Évaluation de la correspondance entre les deux hiérarchies résultant de l'AFC. La correspondance entre les deux hiérarchies est effectuée pour déterminer si la hiérarchie extraite semi-automatiquement du corpus de textes produit les mêmes regroupements que la hiérarchie source construite manuellement par les experts (la base SIMBAD). En d'autres termes, pouvons nous associer les classes de la base SIMBAD (classes de validation) aux classes définies par attributs binaires extraits du corpus de textes (classes d'expérimentation) ?

Plusieurs travaux ont proposé des méthodes d'évaluation de correspondance entre deux hiérarchies. Par exemple, les approches de Hearst (1992) et de Carpineto et Romano (2000) consistent à calculer le nombre de relations hiérarchiques de la hiérarchie source, qu'on retrouve dans la hiérarchie construite semi-automatiquement. Mais la méthodologie PACTOLE comme celle de Cimiano et al. (2005) ne produit pas de label (nom de concept) pour les concepts générés, ce qui rend impossible l'utilisation de cette approche.

Une autre évaluation possible est de mesurer la similarité entre les ensembles d'instances des classes des deux hiérarchies. Les mesures de précision et de rappel sont utilisées pour évaluer cette similarité. Pour chaque classe d'expérimentation, il faut retrouver la classe de validation la plus proche, puis une précision et un rappel locaux sont calculés entre ces deux classes.

La précision globale (Precision_G) et le rappel global (Rappel_G) représentent la moyenne de toutes les précisions et de tous les rappels locaux respectivement.

Calcul de la précision et du rappel. La précision est le nombre d'instances communes entre C_{E_i} (la *i*-ème classe d'expérimentation) et C_{V_j} (la *j*-ème classe de validation) divisé par le nombre d'instances de la classe C_{E_i}. Autrement dit, la précision est le rapport entre le nombre de vrais positifs (les objets bien placés) sur le nombre total des objets deC_{E_i}. Le rappel est le nombre d'instances communes entre C_{E_i} et C_{V_j} divisé par le nombre d'instances de C_{V_j}. En d'autres termes, le rappel est le rapport entre les vrais positifs (les objets bien placés) et le nombre total des objets deC_{E_i}. Soit N est le nombre de classes de C_E.

$$Precision_i = \frac{|C_{E_i} \cap C_{V_j}|}{|C_{E_i}|}, \;\; Recall_i = \frac{|C_{E_i} \cap C_{V_j}|}{|C_{V_j}|}$$

$$Precision\_G = \frac{\sum_{i=1..N}(Precision_i)}{N}, \;\; Recall\_G = \frac{\sum_{i=1..N}(Recall_i)}{N}$$

Détection de la classe la plus proche. Pour chaque classe C_{V_j} de la hiérarchie source, la classe la plus proche C_{E_i} de la hiérarchie extraite des textes est la classe qui partage le plus instances avec la classe C_{V_j}.

$$\forall C_{E_k} \in C_E, (C_{E_i} \cap C_{V_j}) = max(|C_{E_k} \cap C_{V_j}|) \wedge min(|C_{E_k} \setminus C_{V_j}|)$$

Par exemple, soit G l'ensemble des objets G = {`3C_273, TWA, Cen_A, T_Tauri, V773_Tau, Per_OB2, Y_Cygni, Algol`} (voir figures 2 et 3). La classe la plus proche de la classe C_{E_1} avec les instances {`Per_OB2, Cen_A, TWA`} (figure 3) est la classe C_{V_1} de SIMBAD avec les instances {`Per_OB2, TWA`} (figure 2).

	SO		SOC		SOCP	
	Precision_G	Rappel_G	Precision_G	Rappel_G	Precision_G	Rappel_G
AFC	58.33%	05.03%	58.91%	05.94%	74.71%	30.22%

TAB. 5 – *Résultats des mesures de précision et de rappel pour l'application de l'*AFC

L'ensemble SOCP donne de meilleurs résultats (voir table 5) que les autres dépendances avec une bonne précision (74.71%) ce qui veut dire que les objets ont été classés convenablement. En revanche, le rappel est très bas (30.22%), cela peut être expliqué de différentes façons.

Premièrement, le nombre d'attributs associés aux objets n'est pas suffisant. Le système PACTOLE n'a extrait que 23 attributs car la plupart des verbes dans les textes ne sont pas significatifs et il est très difficile de trouver des verbes discriminants dans un domaine spécifique (1600 attributs binaires proposés et seulement 23 validés).

Deuxièmement, les classes ne sont pas seulement définies par les verbes, mais d'autres attributs binaires pourraient être considérés. Par exemple : les adjectifs, les adverbes, les mesures,...Le système extrait ces dépendances du corpus de textes mais ne les traite pas. Soit

de la phrase «*Analysis of the gas velocity structure within GF 17 and GF 20 reveals evidence for smooth large-scale streaming motions along the filamentary structures with magnitude = 0.5 kmpc.*» La dépendance `conj_and(GF_17-18, GF_20-20)` qui exprime la conjonction pourrait permettre de déduire que les deux objets `GF_17` et `GF_20` ont des attributs communs. Ou encore, la mesure de magnitude «`magnitude = 0.5 kmpc`» pourrait être considérée comme un attribut des deux objets. D'autres dépendances sont intéressantes aussi, comme par exemple dans la phrase «*The highly inclined spiral NGC 4258 has been observed in X-rays with the PSPC of the Roentgen observatory ROSAT.*» La dépendance `nn(NGC_4258-5, spiral-4)`, entre les deux noms `NGC_4258` et `spiral`, nous permet de considérer `spiral` comme attribut de l'objet `NGC_4258`. Le choix de ne pas prendre en compte toutes les dépendances est dû au fait qu'elles introduisent aussi beaucoup de bruit. En effet, passer des dépendances (SO) aux dépendances (SOCP) augmente de façon considérable le nombre de paires (de 209 pour l'ensemble SO à 1709 pour l'ensemble SOCP). La validation de ces paires représente un travail considérable pour les astronomes.

Troisièmement, certains attributs binaires sont implicites et ne peuvent pas être extraits par un analyseur syntaxique, ce qui rend impossible la définition de toutes les classes proposées par les experts du domaine.

Évaluation la hiérarchie résultant de l'ARC. Pour évaluer la hiérarchie de concepts résultant de l'ARC et mesurer la correspondance entre les classes de SIMBAD et les concepts résultant de l'ARC, nous appliquons la méthode utilisée pour évaluer la hiérarchie résultant de l'AFC qui s'appuie sur le calcul de la précision et du rappel.

L'ARC ne donne pas de meilleurs résultats que l'AFC (voir table 6), mais elle produit plus de classes (125 classes), ce qui augmente le travail d'évaluation des experts. Néanmoins, cette extension nous a permis d'observer des classes que l'AFC n'a pas pu révéler.

Par exemple : La classe `Quasar` qui n'avait pas d'équivalence en l'AFC a pu être définie avec l'ARC. La distance entre la classe `Quasar` et la classe résultant de l'ARC {`isObservedBy:T0, isObservedBy:T1, isObservedBy:T3, isObservedBy:T7, isIncluding, isEmitting, isObserved, isOutbursting, isRedshift`} est égale à zéro.

	Precision_G	Rappel_G
ARC	52,19%	25,07%

TAB. 6 – *Résultats de l'application de l'*ARC *avec des mesures de précision et de rappel*

6.2.1 Évaluation de l'affectation des attributs binaires aux classes d'objets

Le treillis de concepts résultant de l'apposition de contextes a été présenté aux astronomes afin de déterminer si cette opération a permis de définir les classes d'objets ou d'enrichir la hiérarchie des classes avec de nouvelles classes. Cette opération a fait émerger de nouvelles unités de connaissances, comme par exemple, le concept ({`Orion, TWA`}, {`Association_of_Stars, isExpanding, isObserved`}). Ce concept représente les «`Association_of_Stars`» qui peuvent s'étendre (`isExpanding`). Ce concept a

été considéré comme intéressant par les experts et a servi à définir une nouvelle classe «Association_of_Young_Stars».

6.3 Discussion

Nous avons montré dans notre expérimentation sur le corpus d'astronomie que l'utilisation de toutes les dépendances syntaxiques (SOCP) donnait de meilleurs résultats que la prise en compte d'une partie de ces dépendances, même si ces dépendances ne sont pas suffisantes pour définir toutes les classes. Nous avons aussi présenté d'autres dépendances dont nous pouvons extraire du corpus de textes. Toutefois, il faut limiter le nombre de dépendances à extraire car plus nous extrayons de dépendances plus le travail de validation et d'interprétation des experts du domaine s'accroît.

Nous avons aussi choisi l' AFC qui offre plusieurs avantages, tels que le fait d'avoir des hiérarchies de concepts incrémentales, bien fondées mathématiquement et assez faciles à représenter dans une logique de descriptions comme $\mathcal{FLE}$. L'AFC nous a permis d'associer des définitions (ensemble d'attributs binaires) à des classes (ensemble d'objets) prédéfinies par les experts, ainsi que d'enrichir la hiérarchie des classes en proposant de nouvelles classes. L'Analyse Relationnelle de Concepts (ARC), nous permet de proposer des définitions incluant des relations avec d'autres types d'objets (attributs relationnels).

Néanmoins, l'AFC et ARC présentent aussi quelques désavantages, comme la production d'un très grand nombre de concepts. La taille du treillis peut atteindre ($2^{min\{n,m\}}$) concepts (n : est le nombre d'objets et m : le nombre d'attributs du contextes) Ganter et Wille (1999). Pour traiter ce problème d'explosion combinatoire, certains travaux ont proposé des solutions pour contrôler le nombre de concepts. Par exemple les travaux de Stumme et al. (2002) ou ceux de Messai et al. (2008). Stumme et al. (2002) proposent de limiter la construction du treillis de concepts aux concepts les plus généraux. Il s'agit alors de fixer un seuil minimal contrôlant la taille de l'extension des concepts extraits (treillis d'iceberg). Messai et al. (2008) dans le cadre des treillis de concepts multivalués, apportent une solution double permettant de contrôler l'évolution du nombre de concepts. D'une part, les treillis de concepts multivalués sont directement déduits des contextes multivalués sans avoir à recourir à l'échelonnage ce qui réduit la taille du contexte. D'autre part, la génération des concepts est fonction de la similarité entre les données dans le contexte. La variation de la similarité entraîne la variation du nombre de concepts dans le treillis obtenu.

Pour terminer, soulignons encore que PACTOLE a extrait des unités de connaissances dans le domaine de l'astronomie et a permis d'enrichir la hiérarchie de la base SIMBAD. Ces unités de connaissances peuvent être divisées en quatre types. Le premier type de connaissances est l'identification de nouveaux objets célestes (voir 5.2.1) qui a servi aux astronomes pour enrichir les entrées de la base SIMBAD. Le deuxième type de connaissance est la mise en évidence de nouvelles corrélations entre les objets célestes et leurs attributs binaires (voir 5.2.2). Le troisième type de connaissances consiste à prendre en compte la relation «isObservedBy» entre les objets célestes et les télescopes afin d'enrichir les définitions de concepts (voir 6.2). Enfin, le quatrième type de connaissances recouvre la proposition de nouvelles classes dans la hiérarchie source de la base SIMBAD (voir 6.2.1). Ces trois derniers types de connaissances permettent aux astronomes de définir de nouvelles classes et ainsi d'enrichir la hiérarchie source de la base SIMBAD.

7 Conclusion

Dans cet article, nous avons présenté la méthodologie PACTOLE pour construire semi-automatiquement une ontologie à partir de ressources textuelles hétérogènes. Cette méthodologie fusionne plusieurs hiérarchies extraites de différentes ressources telles que des bases de données, des dictionnaires, des corpus de textes ... En s'appuyant sur l'analyse formelle de concepts et l'analyse relationnelle de concepts, elle combine trois types de descripteurs d'objets : des classes de la base SIMBAD, des attributs binaires et des attributs relationnels. La hiérarchie de concepts produite est représentée avec la logique de descriptions $\mathcal{FLE}$. Cette méthodologie a été appliquée dans le domaine de l'astronomie pour l'extraction d'unités de connaissances sur les objets célestes et pour résoudre des problèmes de classification et de comparaison entre objets. Pour évaluer la correspondance entre la hiérarchie résultant de la base SIMBAD et la hiérarchie résultant du corpus de textes, nous avons proposé une définition des mesures de précision et de rappel.

La prochaine étape de ce travail consiste à utiliser la méthodologie PACTOLE pour les objets annotés «Object_of_unknown_nature» dans SIMBAD et d'essayer de suggérer des classes aux experts. Un autre travail consistera aussi à tester PACTOLE dans un autre domaine, par exemple la classification des bactéries dans le domaine de la microbiologie.

Références

Aussenac-Gilles, N., B. Biébow, et S. Szulman (2000). Revisiting ontology design : A method based on corpus analysis. In R. Dieng et O. Corby (Eds.), *12th Internationnal Conference in Knowlegde Engineering and knowledge Management (EKAW'00)*, Volume 1937, pp. 172–188.

Barriere, C. (2002). Investigating the causal relation in informative texts. *Terminology 7*, 135–154.

Bendaoud, R., A. Napoli, et Y. Toussaint (2008a). Formal concept analysis : A unified framework for building and refining ontologies. In A. Gangemi et J. Euzenat (Eds.), *16th International Conference on Knowledge Engineering and Knowledge Management Knowledge Patterns (EKAW'08)*, Volume 5268, Acitrezza, Catania, Italy, pp. 156–171. Springer.

Bendaoud, R., Y. Toussaint, et A. Napoli (2008b). Pactole : A methodology and a system for semi-automatically enriching an ontology from a collection of texts. In P. Eklund et O. Haemmerlé (Eds.), *16th International Conference on Conceptual Structures (ICCS'08) - Conceptual Structures : Knowledge Visualization and Reasoning*, Volume 5113, Toulouse, France, pp. 203–216. Springer.

Bourigault, C. (1994). *LEXTER, Un logiciel d'Extraction de TERminologie. Application à l'acquisition de connaissances à partir de textes*. Thèse d'informatique, Ecole des Hautes Etudes en Sciences Sociales, Paris.

Carpineto, C. et G. Romano (2000). Order-theoretical ranking. *Journal of the American Society for Information Science (JASIS'00) 51*(7), 587–601.

Cimiano, P., A. Hotho, et S. Staab (2005). Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research (JAIR'05) 24*,

305–339.

de Marneffe, M., B. MacCartney, et C. Manning (2006). Generating typed dependency parses from phrase structure parses. In *5th International conference on Language Resources and Evaluation (LREC'06)*, GENOA, ITALY.

Dubois, J., L. Guespin, M. Giacomo, C. Marcellesi, J. Marcellesi, et J. Mével (1994). *Dictionnaire de linguistique et des sciences du langage*. Collection Trésors du Français, Larousse.

Faatz, A. et R. Steinmetz (2004). Ontology enrichment evaluation. In E. Motta, N. Shadbolt, A. Stutt, et N. Gibbins (Eds.), *14th International Conference on Engineering Knowledge in the Age of the Semantic Web (EKAW'04)*, Volume 3257/2004, Whittlebury Hall, UK, pp. 497–498. Springer.

Faure, D. et C. Nedellec (1999). Knowledge acquisition of predicate argument structures from technical texts using machine learning : The system asium. In *11th International Conference in Knowledge Acquisition, Modeling and Management (EKAW'99)*, Dagstuhl Castle, Germany, pp. 329–334. Springer.

Fayyad, U., G. Piatetsky-Shapiro, et P. Smyth (1996). From data mining to knowledge discovery : An overview. In *Advances in Knowledge Discovery and Data Mining*, pp. 1–34.

Feldman, R. et J. Sanger (2007). *The Text mining handbook*. Cambridge.

Ganter, B. et R. Wille (1999). *Formal Concept Analysis, Mathematical Foundations*. Springer.

Garcia, D. (1998). *Analyse automatique des textes pour l'organisation causale des actions. Réalisation du système informatique COATIS*. Thèse d'informatique, Université de Paris-Sorbonne.

Goujon, B. (1999). Extraction d'informations techniques pour la veille par l'exploitation de notions indépendantes d'un domaine. *Terminologies nouvelles 19*, 33–42.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Norwell, MA, USA : Kluwer Academic Publishers.

Gruber, T. (1993). Toward principales for the design of ontologies used for knowledge sharing. In N. Guarino et R. Poli (Eds.), *Formal Analysis in Conceptual Analysis and Knowledge Representation*, The Netherlands. Kluwer Academic.

Habert, B. et A. Nazarenko (1996). La syntaxe comme marche-pied de l'acquisition des connaissances : Bilan critique d'une expérience. In *Actes des septièmes Journées Acquisition des Connaissances (JAC'96)*, Sète, pp. 137–148.

Hahn, U. et S. Schulz (2004). Building a very large ontology from medical thesauri. In S. Staab et R. Studer (Eds.), *Handbook on Ontologies*, pp. 133–150. Springer.

Harris, Z. (1968). *Mathematical Structure of Language*. Wiley, J. and Sons.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In *International Conference on Computational Linguistics (COLING'92)*, pp. 539–545.

Jacquemin, C. (1997). *Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus*. Mémoire d'habilitation à diriger des recherches en informatique, Université de Nantes.

Jouis, C. (1993). *Contributions à la conceptualisation et à la modélisation des connaissances à partir d'une analyse linguistique de textes. Réalisation d'un prototype : le système SEEK*.

Thèse d'informatique, Ecole des Hautes Etudes en Sciences Sociales, Paris.

Maedche, A. et S. Staab (2000). Discovering conceptual relation from text. In *14th European Conference on Artifical Intelligence (ECAI'00)*, Berlin, Germany, pp. 321–325.

Maedche, E. et S. Staab (2004). Ontology learning. In *Handbook on Ontologies*, pp. 173–189. Springer.

Messai, N., M. Devignes, A. Napoli, et M. Smaïl-Tabbone (2008). Many-valued concept lattices for conceptual clustering and information retrieval. In M. Ghallab, C. Spyropoulos, N. Fakotakis, et N. Avouris (Eds.), *18th biennial European Conference on Artificial Intelligence, (ECAI'2008), 21-25 July,*, Volume 178, Patras, Greece, pp. 127–131. IOS Press.

Morin, E. et E. Martienne (2000). Using a symbolic machine learning tool to refine lexico-syntactic patterns. In R. de Mántaras et E. Plaza (Eds.), *11th European Conference on Machine Learning (ECML'00)*, pp. 292–299. Springer.

Rouane-Hacene, M., M. Huchard, A. Napoli, et P. Valtchev (2007). A proposal for combining formal concept analysis and description logics for mining relational data. In S. Kuznetsov et S. Schmidt (Eds.), *5th International Conference on Formal Concept Analysis (ICFCA'07)*, LNAI 4390, pp. 51–65. Springer.

Rouane-Hacene, M., M. Huchard, A. Napoli, et P. Valtchev (2008). Extraction de connaissances à partir de données relationnelles avec l'analyse formelle de concepts. In P. Marquis et I. Bloch (Eds.), *Conférence sur la reconnaissance des formes et l'intelligence artificielle (RFIA 2008), Amiens*, pp. 143–152. AFRIF–AFIA.

Rousselot, F., P. Frath, et R. Oueslati (1996). Extracting concepts and relations from corpora. In *Proceedings of ECAI Workshop on Corpus-Orientated Semantic analysis*, Budapest.

Stumme, G. et A. Maedche (2001). Fca-merge : Bottom-up merging of ontologies. In *International Joint Conference on Artificial Intelligence (IJCAI'01)*, pp. 225–234.

Stumme, G., R. Taouil, Y. Bastide, N. Pasquier, et L. Lakhal (2002). Computing iceberg concept lattices with t. *Data and Knowledge Engineering 42*(2), 189–222.

Valente, A., T. Russ, R. MacGregor, et W. Swartout (1999). Building and (re)using an ontology of air campaign planning. *IEEE Intelligent Systems 14*(1), 27–36.

Summary

In this article, we propose a methodology named PACTOLE «Property And Class Caracterisation from Text to OntoLogy Enrichment» that allows to build an ontology in a specific domain. PACTOLE merges and combines different ressources using the Formal Concept Analysis (FCA) and its extension Relational Concept Analysis (RCA). This resulting target ontology can then be encoded within OWL or a description logic formalism, allowing classification-based reasoning. A real-world example in astronomy is detailed and shows how experts may interact with the system. We have also formalized ans answered to some questions asked by the astronomers.