HAL
archives-ouvertes.fr

Construction d'une ontologie à partir d'un corpus de textes avec l'ACF

Rokia Bendaoud, Amine Mohamed Rouane Hacene, Yannick Toussaint, Bertrand Delecroix, Amedeo Napoli

▶ **To cite this version:**

Rokia Bendaoud, Amine Mohamed Rouane Hacene, Yannick Toussaint, Bertrand Delecroix, Amedeo Napoli. Construction d'une ontologie à partir d'un corpus de textes avec l'ACF. IC 2007, Jul 2007, Grenoble, France. inria-00167678

HAL Id: inria-00167678
https://hal.inria.fr/inria-00167678

Submitted on 22 Aug 2007

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.

Construction d'une ontologie à partir d'un corpus de textes avec l'ACF

Rokia Bendaoud[1], Mohamed Rouane Hacene[1], Yannick Toussaint[1], Bertrand Delecroix[1], Amedeo Napoli[1]

Loria-Campus Scientifique BP 239 - 54506 Vandoeuvre-lès-Nancy Cedex
{Bendaoud,Rouanehm,Yannick,delecroix,Napoli}@loria.fr

Résumé : Nous présentons dans cet article une méthodologie semi-automatique de construction d'ontologie à partir de corpus de textes sur un domaine spécifique. Cette méthodologie repose en premier lieu sur la classification d'objets d'après les propriétés qu'ils partagent, en utilisant l'analyse de concepts formels (ACF) pour la construction d'un treillis de concepts. Ce treillis va servir à construire un noyau d'ontologie. Cependant, les objets sont aussi définis par les relations qu'ils entretiennent entre eux. Donc, en second lieu, nous proposons une méthode originale qui enrichit cette ontologie avec des relations transversales en utilisant une nouvelle méthode : l'analyse relationnelle de concepts (ARC). Chaque concept de l'ontologie résultante est défini puis représenté en Logique de Descriptions (LDs). Le domaine d'application de cette méthodologie est le domaine de l'astronomie.

Mots-clés : Construction d'ontologie à partir de textes, ACF : Analyse de concepts formels, relations transversales, ARC : Analyse Relationnelle de Concept

1 Introduction

Un système d'acquisition de connaissances est important pour n'importe quel domaine spécifique. Il permet aux experts de raisonner sur les connaissances du domaine et de les partager. Néanmoins, il faut savoir que ces systèmes souffrent tous de ce qu'on appelle "goulot d'étranglement dans l'acquisition de connaissances" (Cimiano *et al.*, 2005), c'est-à-dire la difficulté d'actualiser le modèle du domaine en question. Par exemple, dans le domaine de l'astronomie, la classification des objets célestes dans des classes prédéfinies (étoiles, galaxies, comètes, ...) est une base de connaissances très importante. Cette classification est faite manuellement d'après les propriétés avec lesquelles les objets apparaissent dans les textes : l'astronome lit les articles qui traitent d'un objet particulier et essaye de déterminer la classe qui lui convient le mieux. Jusqu'à présent, plus de 3 millions d'objets ont été classifiés de la sorte dans la base de données SimBad[1], mais il reste des milliards d'objets à classifier, d'où l'apparition du goulot d'étranglement dans l'acquisition des connaissances. De plus, les astronomes

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

remettent en cause cette classification et veulent une définition précise de chacune des classes afin d'être certains de l'affectation d'une classe à un objet.

Cet article a pour objectif de construire une classification des objets célestes, puis de donner une définition formelle à chaque classe d'objets. Pour cela, nous proposons une méthodologie de construction d'une ontologie dans le domaine de l'astronomie. Cette méthodologie est semi-automatique car supervisée à chaque étape par les experts du domaine. "Une ontologie est une spécification formelle des concepts et des relations entre ces concepts" (Gruber, 1993). Elle permet de définir les classes d'objets célestes puis de les représenter en Logique de Descriptions (LDs) et ainsi de répondre formellement à des requêtes appelées questions de compétence (*competency questions*). Ces questions sont d'abord écrites en langage naturel puis traduites dans le langage formel utilisé.

Nous choisissons d'utiliser un corpus de textes comme source de connaissances pour la détection des objets célestes et de leurs propriétés. Dans le domaine de l'astronomie, les textes sont disponibles sous format électronique. Nous pouvons donc utiliser des outils d'analyse terminologique afin de les exploiter et de repérer les objets célestes et leurs propriétés automatiquement (Aussenac-Gilles, 2005).

L'ontologie est construite en plusieurs étapes. Tout d'abord, le noyau de l'ontologie est construit. Ce noyau représente la hiérarchie des concepts en s'appuyant sur la méthode de Cimiano (Cimiano *et al.*, 2005) qui utilise l'Analyse de Concepts Formels (ACF) (Ganter, 1999), méthode formelle qui regroupe un ensemble d'objets d'après les propriétés qu'ils partagent et, de façon duale, regroupe un ensemble de propriétés d'après les objets qui les possèdent. Notre apport dans cette partie est de représenter les concepts en LDs FL^-[2] (Baader, 2003), puis de donner la méthode de raisonnement utilisée pour répondre aux questions de compétences du type :

– *Quels objets sont de même classe que l'objet* `CAL83` *?*
– *Est-ce que les objets* `CAL83` *et* `PSRA` *sont dans la même classe ?*
– *Est-ce que l'objet* `GRO` *est une étoile ? Si oui, à quel type d'étoiles appartient-il ?*

Par ailleurs, les objets ne sont pas seulement définis par leurs propriétés, ils sont aussi définis par les relations qu'ils entretiennent avec d'autres objets. La deuxième étape de cette construction consiste donc à prendre en compte des relations transversales entre objets. Nous proposons, lors de cette étape, une nouvelle approche fondée sur une extension de l'ACF, l'Analyse Relationnelle de Concept (ARC), qui classifie un ensemble d'objets célestes par rapport aux relations qu'ils entretiennent avec d'autres objets. Cette extension d'ontologie nous oblige à utiliser un langage en LDs plus expressif qui permet de prendre en compte les restrictions de relations. Nous choisissons d'exprimer cette ontologie dans la LDs ALC[3] (Baader, 2003). Les questions de compétences auxquelles répond cette extension sont :

– *Qu'observe par infra-rouge un* Infra-red-Telescope *?*
– *Est-ce que les telescopes* XMM-Newton *et* Chandra *observent les mêmes objets ?*

Cet article se décompose comme suit. Tout d'abord la section 2 introduit la méthodologie suivie pour construire notre ontologie. Puis, la section 3 détaille le pro-

[2] FL^- inclut les opérateurs suivants : A : concept primitif, R : rôle, C ⊓ D : conjonction de concepts, ∀R.C : restriction universelle, ∃R : quantification existentielle, ⊤ : Top et ⊥ : Bottom.

[3] ALC est une extension de la FL^- + l'opérateur ¬C, ce qui permet de déduire la restriction des rôles : ∃R.C

cessus de fouille de textes qui nous permet de passer du corpus de textes aux données d'entrée pour la construction d'ontologie. La section 4 décrit la hiérarchie de concepts construite avec l'ACF. Elle formalise également le passage entre l'ACF et le noyau de l'ontologie. La section 5 présente l'enrichissement du noyau de l'ontologie avec des relations transversales entre concepts en utilisant l'ARC. Enfin la section 6 conclut ce papier en introduisant quelques perspectives.

2 Méthodologie

Notre méthodologie (décrite dans la figure 1) s'appuie sur la "Methontology" (Gómez-Pérez *et al.*, 2004). La "Methontology" construit une ontologie à partir des termes extraits des ressources (les ressources ne sont pas spécifiées) et a pour objectif de donner une définition à chaque concept et à chaque relation de l'ontologie dans un langage de LDs. La "Methontology" nous spécifie les étapes à suivre pour construire une ontologie.

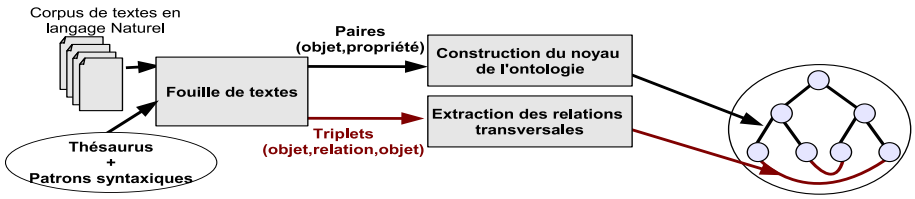

FIG. 1 – **Méthodologie de construction de l'ontologie**

Cette méthodologie se compose de trois étapes :

- Extraction de termes : consiste à extraire les termes et leurs propriétés avec une méthode de fouille de textes qui prend en entrée le corpus de textes et des ressources pour détecter les objets célestes (le thésaurus de l'astronomie[4] et des patrons syntaxiques[5]). Puis, elle analyse le corpus avec un analyseur syntaxique. Ensuite, elle extrait des paires et des triplets présents dans les mêmes syntagmes syntaxiques (Voir section 3).
- Construction du noyau de l'ontologie : consiste à utiliser les paires (objet,propriété) pour la construction d'une hiérarchie de concepts avec l'ACF (Voir section 4).
- Extraction des relations transversales : prend en entrée les triplets extraits du texte, puis utilise l'ARC pour extraire les relations transversales (Voir section 5),
- Regroupement des deux modules pour obtenir l'ontologie complète.

3 Fouille de textes

Nous cherchons le moyen d'exploiter le corpus de textes pour construire une hiérarchie de concepts. La premiere approche étudiée est celle de (Malaisé, 2005). Malaisé pro-

[4]http ://msowww.anu.edu.au/library/thesaurus/

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

pose d'utiliser des patrons définitoires pour extraire les définitions de chaque terme à partir du corpus de textes. Cette méthode donne de très bons résultats quand elle est utilisée pour un corpus de type dictionnaire (`terme : définition ou terme`$_1$ `est_un terme`$_2$), mais pas pour un corpus de textes tel que celui de l'astronomie, car il est difficile de trouver une définition à chaque objet céleste. La deuxième approche étudiée est fondée sur l'hypothèse de Harris (Harris, 1968), selon laquelle l'étude des régularités syntaxiques dans un corpus de sous-langage (ou langage spécialisé) permet d'identifier des schémas syntaxiques formés de combinaisons de classes qui reflètent les connaissances du domaine traité. L'une des méthodes fondées sur cette hypothèse est la méthode de (Faure & Nedellec, 1998) qui regroupe des termes en classes d'après leur présence dans des syntagmes syntaxiques avec le même groupe de verbes. Cette méthode nous permet, d'une part, de regrouper les objets célestes d'après l'ensemble des verbes dont ils sont sujets ou compléments. Par exemple, les objets {PSRA,SN437} sont regroupés dans la même classe car ils apparaissent comme sujet du verbe {to emit} et comme complément pour l'ensemble des verbes {to observe, to locate}. D'autre part, cette méthode nous permet aussi d'extraire des relations entre des objets apparaissant comme sujet (ou comme complément) du même verbe et un ensemble de compléments (ou un ensemble de sujets). Par exemple, les objets célestes {PSRA,SN437} sont des complements reliés aux télescopes {BeppoSax} qui sont les sujets par le verbe {to observe}.

La méthode de fouille de textes consiste à prendre en entrée le corpus de textes et les ressources pour la détection des objets célestes et en sortie, à extraire de chaque phrase des paires (sujet,verbe), (complément,verbe) et des triplets (sujet,verbe,complement). L'analyse syntaxique des textes est effectuée par un analyseur partiel et robuste le "Stanford Parser[6]"(de. Marneffe *et al.*, 2006).

Le verbe est différent selon qu'il définit les sujets ou les compléments, c'est-à-dire, que pour chaque verbe, nous extrayons deux classes différentes : la classe des sujets et la classe des compléments. Les paires et les triplets sont présentés à un expert afin de ne garder que les plus pertinents. Nous présentons deux exemples en astronomie :

1. "One HR2 candidate was `detected` and `regrouped` in each of the galaxies NGC 3507 and CygnusA". D'où on extrait les paires : (HR2, `regrouped`), (HR2, `detected`), (NGC 3507, `regrouping`), (CygnusA, `regrouping`).
2. "A total of 2 stars PSRA and SN437 `have been observed by` $\texttt{Infra} - \texttt{Red}$ with the W Field Cameras telescopes on board BeppoSAX during a monitoring observation of the SgrA region in August-September 1996". D'où on extrait les triplets : (PSRA, $\texttt{observed\_by\_Infra} - \texttt{Red}$,BeppoSAX), (SN437, $\texttt{observed\_by\_Infra} - \texttt{Red}$,BeppoSAX).

4 Construction du noyau de l'ontologie

Il existe deux types de travaux pour la construction d'une hiérarchie de concepts à partir de corpus de textes. Le premier type repose sur la co-occurrences de termes dans les textes (Sanderson & Croft, 1999) puis, utilise des mesures de similarité pour créer la

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

hiérarchie de classes d'objets. Ces travaux répondent au besoin d'avoir une hiérarchie de concepts, mais pas à celui de proposer une définition des classes d'objets. Le second type est de nature symbolique, c'est-à-dire qu'il construit la hiérarchie de classes d'objets d'après les propriétés qu'ils partagent. Nous reprenons l'idée de Cimiano (Cimiano *et al.*, 2005) qui propose de construire une ontologie formelle avec l'ACF. Nous reformulons ensuite le passage entre le treillis et l'ontologie, puis une représentation en FL^- est proposée pour chaque concept de l'ontologie. Ces définitions vont permettre de répondre aux questions de compétences présentées dans la section 1.

L'ACF est une méthode qui fait appel à la notion de concept formel. La hiérarchie résultante qui regroupe des objets partageant les mêmes propriétés est appelée treillis de concepts. La notion centrale de l'ACF est le contexte formel.

Définition 1 (Contexte formel)
Un triplet $\mathbb{K}=(\mathtt{G},\mathtt{M},\mathtt{I})$ *est appelé* ***contexte formel*** *si* $\mathtt{G}$ *et* $\mathtt{M}$ *sont des ensembles disjoints et* $\mathtt{I} \subseteq \mathtt{G} \times \mathtt{M}$ *est une relation binaire entre* $\mathtt{G}$ *et* $\mathtt{M}$*. Les éléments de* $\mathtt{G}$ *sont appelés* ***objets*** *et ceux de* $\mathtt{M}$ ***propriétés****.*

Définition 2
Soit $\mathbb{K}=(\mathtt{G},\mathtt{M},\mathtt{I})$ *un contexte formel. Pour tout* $A \subseteq \mathtt{G}$ *et* $B \subseteq \mathtt{M}$*, on définit :*
- $\mathtt{A}' = \{\mathtt{m} \in \mathtt{M} \mid \forall\, \mathtt{g} \in \mathtt{A},\ \mathtt{gIm}\}$
- $\mathtt{B}' = \{\mathtt{g} \in \mathtt{G} \mid \forall\, \mathtt{m} \in \mathtt{B},\ \mathtt{gIm}\}$

	Emitting	Accreting	Collimated	Observed	Located	grouping
PSRA	x			x	x	
NGC3507				x		x
Andromeda		x		x		x
M87			x	x	x	
HR2				x	x	
CygnusA		x		x		x
SN437	x			x	x	
NGC2018			x	x	x	

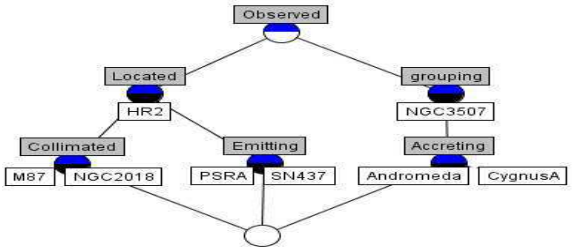

FIG. 2 – **Contexte formel $\mathbb{K}$=(G,M,I)**

Intuitivement, $\mathtt{A}'$ est l'ensemble des propriétés communes à tous les objets de $\mathtt{A}$, et $\mathtt{B}'$ est l'ensemble des objets possédant toutes les propriétés de $\mathtt{B}$. L'opérateur ' est appelé opérateur de dérivation et s'applique aussi bien aux sous-ensembles de $\mathtt{G}$ qu'aux sous ensembles de $\mathtt{M}$. Cet opérateur peut se composer avec lui même, pour partir d'un sous-ensemble d'objets $\mathtt{A}$, produire $\mathtt{A}'$, et à partir de $\mathtt{A}'$ produire le sous-ensemble d'objets $\mathtt{A}''$ (la notation $''$ est utilisée pour marquer la composition et elle définit la fermeture sur l'ensemble des parties de $\mathtt{G}$ et sur l'ensemble des parties de $\mathtt{M}$). Nous définissons maintenant la notion de concept formel :

Définition 3 (Concept formel)
Une paire $(\mathtt{A},\mathtt{B})$ *est un* ***concept formel*** *ssi* $\mathtt{A} \subseteq \mathtt{G}$, $\mathtt{B} \subseteq \mathtt{M}$, $\mathtt{A}'=\mathtt{B}$ *and* $\mathtt{B}' = \mathtt{A}$.

En d'autres termes, `(A,B)` est un **concept formel** si et seulement si l'ensemble de toutes les propriétés partagées par les objets dans `A` est identique à `B`, et de façon duale, `A` est l'ensemble de tout les objets qui ont en commun les propriétés de `B`. `A` est appelé **extension** et `B` est appelé **intension** du concept `(A,B)`. Les concepts sont ordonnés par la relation de subsomption définie par :

– $(A_1, B_1) \leq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2 \ (\Leftrightarrow B_2 \subseteq B_1)$

La relation de subsomption permet d'organiser les concepts formels en un treillis complet, (C (G,M,I),$\leq$), appelé treillis de concepts, qui est noté $\underline{\mathfrak{B}}(G, M, I)$.

Nous présentons un exemple dans le domaine de l'astronomie dans la figure 2. Avec le contexte formel $\mathbb{K}$=(G,M,I), où : G est l'ensemble des objets célestes, M l'ensemble des propriétés de ces objets. Ici l'ensemble des propriétés représente l'ensemble des verbes significatifs dans le domaine de l'astronomie. I est la relation binaire entre G et M, tel que `I(g,m)` signifie que `g` apparaît en tant que sujet ou de complément du verbe `m` dans le corpus de textes.

Le treillis résultant est présenté dans la même figure. Cette représentation des treillis s'appuie sur l'héritage à la fois des attributs et des objets entre les nœuds représentant les concepts du treillis. Les attributs sont placés au plus haut dans le treillis : à chaque fois qu'un nœud `n` est étiqueté par un attribut `m`, tous les descendants de `n` dans le treillis héritent de l'attribut `m`. De façon duale, les objets sont placés au plus bas dans le treillis : à chaque fois qu'un nœud `n` est étiqueté par un objet `g` tous ses ancêtres héritent de `g`. Ainsi l'extension `A` d'un concept `(A,B)` est obtenue en considérant tous les objets qui apparaissent sur les descendants du nœud `n` dans le treillis et son intension `B` est obtenue en considérant tous les attributs qui apparaissent sur les ancêtres du nœud n dans le treillis (Messai *et al.*, 2006). Cette présentation permet d'identifier les propriétés et les objets propres de chaque concept.

4.1 Passage du treillis à l'ontologie et étiquetage par des experts

Nous définissons la fonction de transformation $\alpha : \underline{\mathfrak{B}}(G, M, I) \rightarrow$ TBox $\cup$ ABox où : $\underline{\mathfrak{B}}(G, M, I)$ est le treillis de concepts obtenu par l'ACF, la TBox représente le noyau de l'ontologie et l'ABox est la base de Connaissances (BC) (Cimiano *et al.*, 2005). La TBox et la ABox sont définies comme suit :

Définition 4 (noyau d'ontologie)
Un noyau d'ontologie est représenté par untriplet $\mathcal{O} := (C, \sqsubseteq_C, A)$, *où* C *est un ensemble de concepts,* $\sqsubseteq_C$ *la relation de subsomption entre les concepts, elle est transitive, rèflexive et anti-symétrique (ordre partiel) et* A *est l'ensemble des attributs des concepts.*

Définition 5 (Base de Connaissances (BC))
Une BC pour une ontologie $\mathcal{O} := (C, \sqsubseteq_C, A)$ *est la structure :*
$\mathcal{CB} := (I, i_C, i_A)$ *où : I est l'ensemble des instances,* $i_C : C \rightarrow 2^I$ *est une fonction appelée instantiation de concepts et* i_A *est une fonction appelée instantiation d'attributs.*

La fonction de transformation α qui formalise ce passage (figure 3) est présentée dans le tableau 1 .

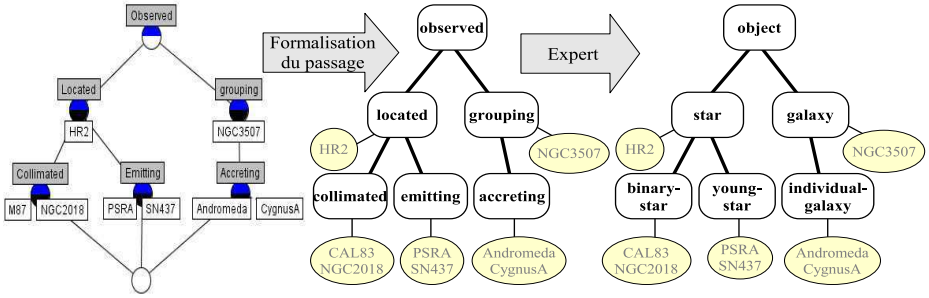

FIG. 3 – **Transformation du treillis en ontologie et étiquetage des concepts**

TAB. 1 – **Formalisation du passage Treillis-Ontologie**

Treillis des concepts	ontologie
Contexte $\mathcal{K}$	Concept atomique c $\equiv \alpha(\mathcal{K})$
Propriété propre $m \in M$	Concept atomique $\alpha(m) \equiv \exists m.\top$ dans la TBox
Objet propre $g \in G$	Instance $\alpha(g)$ dans la ABox
Element $(g, m) \in I$	Assertion $\alpha(m)(\alpha(g))$ dans la ABox
Concept $c = (X, Y) \in \mathcal{C}$	Concept défini $\alpha(c)$ dans la TBox, s.t., $\alpha(c) \equiv \sqcap_{m \in Y} \alpha(m)$
$\forall (c, \bar{c}) \in \mathcal{C} \times \mathcal{C}$, tel que, $c \prec \bar{c}$	Inclusion d'axiomes $\alpha(m) \sqsubseteq \alpha(\bar{m})$ dans la TBox
Infimum d'attributs de concepts ($\bigwedge_{i=1}^{n} c_i$)	conjunction de concepts $\alpha(c_1) \sqcap ... \sqcap \alpha(c_n)$

Nous proposons l'ontologie résultante aux experts qui doivent donner une étiquette à chaque concept d'après les propriétés que partagent les instances de ces concepts. Par exemple, la classe des objets qui possèdent l'ensemble des propriétés {`observed`,`localised`} est étiquetée par `star`, ou encore, la classe des objets qui possèdent l'ensemble des propriétés {`observed`,`regrouping`,`accreting`} est étiquetée par `individual galaxy`, etc, jusqu'à ce que tout les concepts soient étiquetés. Cette représentation n'est faite que pour affecter une étiquette à un ensemble d'objets célestes et pour aider l'expert à lire l'ontologie.

4.2 Représentation des concepts en Logique de Descriptions

La définition d'un concept en LDs FL^- est obtenue par la conjonction de ses attributs avec le quantificateur existentiel "$\exists$". Le tableau 2 présente la définition de chaque concept de l'ontologie présentée dans la figure 3.

Cette représentation permet d'utiliser les propriétés de la LDs pour répondre à des requêtes de type :

Population d'ontologie : Soit o_1 un objet dont les propriétés sont {`a`,`b`}. *Quelle*

Concept définit	Concept définit
Object := ∃observed	Star := ∃observed ⊓ ∃located
Young-Star := ∃observed ⊓ ∃located ⊓ ∃emitting	Galaxy := ∃observed ⊓ ∃grouping
Binary-Star := ∃observed ⊓ ∃located ⊓ ∃collimated	Individual-Galaxy := ∃observed ⊓ ∃grouping ⊓ ∃accreting

TAB. 2 – Définition des concepts de l'ontologie en FL^-

est la classe dont l'objet o_1 est instance ? La réponse est donnée par la classe la plus générale qui possède l'ensemble {a,b}, c'est-à-dire, la classe $C_1 \sqsubseteq \exists \texttt{a}.\top \sqcap \exists \texttt{b}.\top$ (Baader, 2003). Par exemple, pour la question *"Est-ce que l'objet* GRO *est une étoile, sachant qu'il a les propriétés* {observed, located, emitting} ? *Si oui, à quel type d'étoile appartient-il ?"* Réponse : La classe la plus générale $C_1 \sqsubseteq \exists$observed ⊓ ∃located ⊓ ∃emitting est la classe young-star, et la classe young-star ⊑ star. Donc GRO est bien une étoile de type young-star.

Comparaison d'objets célestes : *Soient les deux objets o_1 et o_2, est-ce que o_1 est de la même classe que o_2* ? Répondre à cette question revient à déterminer : quelle est la classe C_1 de o_1 et quelle est la classe C_2 de o_2, et enfin vérifier si $C_1 \equiv C_2$. Par exemple, pour les questions :

- *"Quels objets sont de la même classe que l'objet* CAL83 *?"*
  Réponse : l'ensemble des instances de même classe dont CAL83 est instance. CAL83 est instance de binary-star. L'ensemble des autres instances de binary-star est {NGC2018}
- "*Est-ce que les objets* CAL83 *et* PSRA *sont dans la même classe ?*"
  Réponse : CAL83 est instance de la classe binary-star
  binary-star := ∃observed ⊓ ∃located ⊓ ∃collimated, PSRA est instance de young-star := ∃observed ⊓ ∃located ⊓ ∃emitting et young-star ⊓ binary-star = ⊥. Donc, non CAL83 et PSRA ne sont pas de la même classe.

Pour l'instant, les objets ne sont représentés que par des propriétés. Or, en astronomie, les objets sont aussi définis par les relations qu'ils entretiennent avec d'autres objets. Ainsi, nous avons besoin d'enrichir le noyau de l'ontologie avec des relations transversales entre concepts. Nous proposons alors d'utiliser une méthode complémentaire à l'ACF, l'analyse relationnelle de concepts (ARC), afin de prendre en compte les relations transversales dans l'ontologie.

5 Extraction des relations transversales

L'extraction de relations transversales permet de donner une définition plus précise d'un concept. Le concept n'est plus seulement défini par les propriétés que partagent ses instances, mais également par les relations qu'ils entretient avec les autres concepts. Seuls quelques travaux se sont intéressés à extraire des relations transversales à partir de corpus de textes. Aussenac-Gilles (Aussenac-Gilles *et al.*, 2000) propose d'utiliser une méthode d'apprentissage pour des patrons syntaxiques. *À* partir de tri-

plets extraits manuellement du corpus de textes $(\texttt{terme}_1, \texttt{relation}_1, \texttt{terme}_2)$, on cherche tous d'abord à généraliser la $\texttt{relation}_1$ en extrayant les triplets du type $(\texttt{terme}_1, \texttt{relation}_k, \texttt{terme}_2)$. On obtient une relation plus générale R entre $(\texttt{terme}_1, \texttt{terme}_2)$, Ensuite, on extrait tous les termes reliés par la relation R en extrayant des triplets de la forme $(\texttt{terme}_i, \texttt{R}, \texttt{terme}_j)$. Cette méthode permet de regrouper des instances d'une relation en une relation plus générique, mais elle n'utilise pas le noyau de l'ontologie pour généraliser les termes. Une seconde approche est celle de (Maedche & Staab, 2000) qui s'appuie sur l'extraction de règles d'association. On extrait des paires $(\texttt{terme}_1, \texttt{terme}_2)$ du corpus de textes, puis les règles d'association $(\texttt{terme}_1 \Rightarrow \texttt{terme}_2)$ afin de ne garder que les paires les plus fréquentes et avec la meilleure confiance (Agrawal & Srikant, 1995). Cette méthode place une relation le plus haut possible dans le noyau de l'ontologie, mais la relation n'est pas définie. On sait que le concept C_1 est relié au concept C_2, mais on ne sait pas par quelle relation.

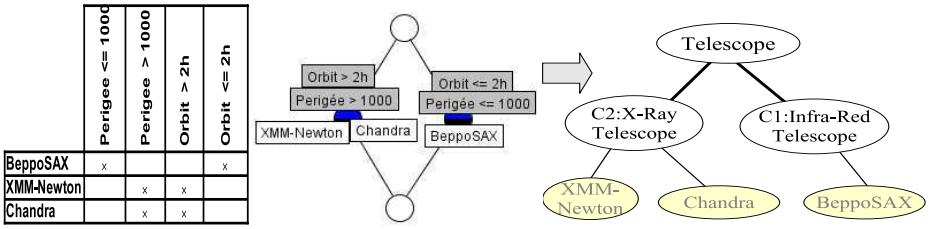

FIG. 4 – **Contexte formel des telescopes**

5.1 L'Analyse Formelle de concepts

Nous proposons une méthode formelle qui extrait des relations transversales entre les concepts. Cette méthode assigne une étiquette à chacune des relations extraites et elle utilise le noyau de l'ontologie pour généraliser les termes. C'est une extension de l'ACF qui regroupe un ensemble d'objets, non pas d'après les propriétés qu'ils partagent, mais d'après les relations qui les relient à d'autres objets. L'idée est de construire un contexte formel pour chaque relation transversale extraite du textes, tel que l'ensemble des individus représente le domaine de cette relation et l'ensemble des propriétés représente son co-domaine et la relation binaire du contexte est la relation qui les relie. Cette extension est l'Analyse relationnelle de concept (ARC) (Valtchev *et al.*, 2003; Rouane *et al.*, 2007). La notion centrale de l'ARC est la Famille Relationnelle de Contextes (FRC) :

Définition 6 (FRC)
Une Famille Relationnelle de Contextes est une paire $(\mathbf{K}, \mathbf{R})$ *avec :*

- $\mathbf{K}$ *ensemble de contextes* $\mathcal{K}_i = (G_i, M_i, I_i)$*, un contexte pour chaque catégorie d'individus,*
- $\mathbf{R}$ *ensemble de relations* $r_k \subseteq G_i \times G_j$*, où* G_i *(domaine) et* G_j *(co-domaine) sont deux ensembles d'individus de* $\mathbf{K}$*.*

Observed_by_X-Ray			
	BeppoSAX	XMM-Newton	Chandra
M87		X	
NGC2018			X

Observed_by_Infra-Red			
	BeppoSAX	XMM-Newton	Chandra
PSRA	X		
SS433	X		

TAB. 3 – Les deux Relations Observed_by_X-Ray et Observed_by_Infra-Red

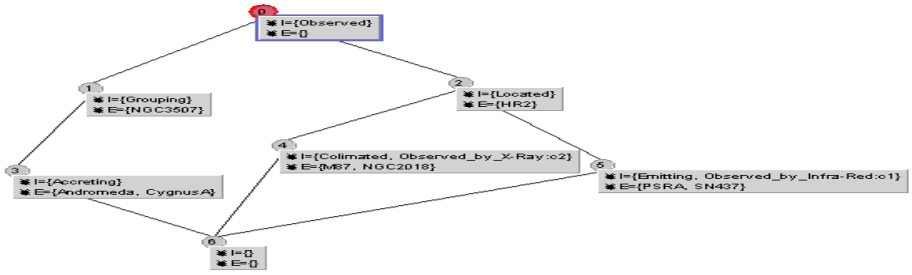

FIG. 5 – **Treillis final des objets célestes**

Nous présentons un exemple dans le domaine de l'astronomie. Soient :

- deux contextes construits avec l'ACF : $\mathcal{K}_1 = (G_1, M_1, I_1)$ présenté dans la figure 2, le contexte des objets célestes et $\mathcal{K}_2 = (G_2, M_2, I_2)$ présenté dans la figure 4, où G_2 est l'ensemble des télescopes, M_2 l'ensemble des propriétés des télescopes et I_2 la relation binaire qui les relie.
- deux relations r_1 = `observed_by_X-Ray`, r_2 = `observed_by_Infra-Red`. Ces deux relations sont présentées dans le tableau 3.

L'intégration des relations r_1 = `"observed_by_X-Ray"`, r_2 = `"observed_by_Infra-Red"` dans le treillis des objets celestes est faite par le processus de scaling (présenté en détail dans (Rouane *et al.*, 2007)). Le treillis resultant présenté dans la figure 5 met en évidence deux relations transversales `observed_by_X-Ray` entre les deux concepts {Young-Star} et {C2 :X-Ray Telescope}, et la relation `observed_by_Infra-Red` entre les deux concepts {Binary-Star} et {C1 :Infra-Red Telescope}.

Avec ces relations transversales le noyau d'ontologie peut être étendu à une ontologie plus complète (voir figure 6) exprimée dans la LDs *ALC* :

Définition 7 (Ontologie complète)
Une ontologie complète est représentée par un quintuplet $\mathcal{O} := (C, \sqsubseteq_C, A, R, \sigma)$*, où* $(C, \sqsubseteq_C, A)$ *représente le noyau de l'ontologie,* R *un ensemble de relations et* σ *est la signature des relations.*

Définition 8 (Base de Connaissances)
la BC associée à cette ontologie $\mathcal{O} := (C, \sqsubseteq_C, A, R, \sigma)$ *est la structure* $\mathcal{CB} := (I, i_C, i_A, i_R)$
où : I est l'ensemble des instances, $i_C : C \rightarrow 2^I$ *est une fonction appelée instantiation*

de concepts, i_A est une fonction appelée instantiation d'attributs et $i_R : C \rightarrow 2^{I^+}$ est une fonction appelée instantiation de relations.

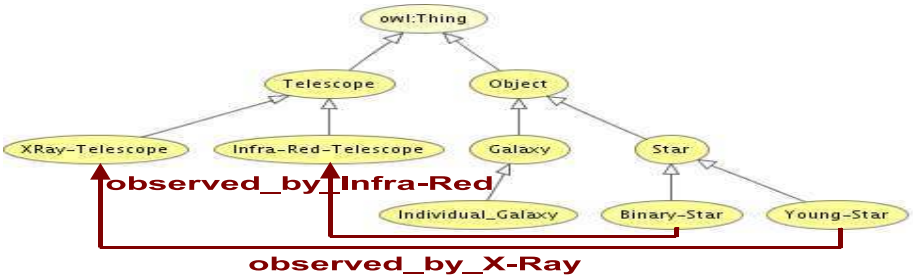

FIG. 6 – **L'ontologie enrichie avec les relations transversales**

5.2 Représentation des concepts en Logique de descriptions

L'extraction des relations transversales permet d'enrichir la définition des concepts dans la LDs *ALC*. Les deux concepts `binary-star` et `young-star` sont redéfinis avec les deux relations `observed_by_X-Ray` et `observed_by_Infra-Red` :

- `Young-Star := ∃observed ⊓ ∃located ⊓ ∃emitting⊓ ∃observed_by_X-Ray.X-Ray-Telescope,`
- `Binary-Star := ∃observed ⊓ ∃located ⊓ ∃collimated⊓ ∃observed_by_Infra-Red.Infra-Red-Telescope`

Cette extension nous permet de répondre à de nouvelles requêtes telles que : *"Qu'observe par infra-rouge un* Infra-red-Telescope *?"* Pour répondre à cette question, il faut trouver C_1 qui est le domaine de la relation `observed_by_Infra-Red` et où `Infra-Red-Telescope` représente le co-domaine. Ce concept est donné par la figure 6, C_1 = `Binary-Star`.

Nous donnons un autre exemple de requête : *"Les telescopes* `XMM-Newton` *et* `Chandra` *observent-ils les mêmes objets ?"* La réponse est fournie par le contexte formel de la figure 5. Les objets `XMM-Newton` et `Chandra` sont dans le co-domaine de la relation `observed_by_X-Ray` et donc ils sont reliés à la même classe d'objets `young-star`.

6 Conclusion

Dans cet article nous avons présenté deux méthodes permettant de construire une ontologie à partir d'un corpus de textes dans le domaine de l'astronomie. La premiere méthode construit la hiérarchie de concepts (noyau d'ontologie) avec une méthode formelle, l'ACF, puis elle représente chaque classe d'objets célestes dans la LDs FL^-. La seconde méthode, extrait les relations transversales entre les concepts avec une nouvelle méthode, l'ARC, afin d'enrichir le noyau de l'ontologie, puis elle représente des concepts en *ALC*. Notre méthodologie est indépendante du corpus de textes et du domaine d'application.

La prochaine étape de notre travail va consister, tout d'abord, à trouver d'autres syntagmes syntaxiques plus riches que les triplets (Sujet,Verbe,Complément), afin d'avoir

des classes d'objets plus pertinentes. Ensuite, il faudra compléter notre ontologie en proposant une méthode de hiérarchisation des relations transversales entre concepts.

Références

AGRAWAL R. & SRIKANT R. (1995). Mining generalized association rules. In *21st VLDB Conference*, Zurich, Swizerland.

AUSSENAC-GILLES N. (2005). *Methodes ascendantes pour l'ingenierie des connaissances*. Habilitation à diriger des recherches, Université Toulouse III - Paul Sabatier.

AUSSENAC-GILLES N., BIÉBOW B. & SZULMAN S. (2000). Revisiting ontology design : A method based on corpus analysis. In D. R. & O. CORBY, Eds., *12th Int. Conference in Knowlegde Engineering and knowledge Management (EKAW'00)*, volume 1937, p. 172–188.

BAADER F. (2003). Description logic terminology. In B. F., D. CALVANESE, D. MCGUINNESS, D. NARDI & P. PATEL-SCHNEIDER, Eds., *The Description Logic Handbook : Theory, Implementation, and Applications*, p. 485–495 : Cambridge University Press.

CIMIANO P., HOTHO A. & STAAB S. (2005). Learning concept hierarchies from text corpora using formal concept analysis. In *Journal of Artificial Intelligence Research (JAIR)*, volume Volume 24, p. 305–339.

DE. MARNEFFE M., MACCARTNEY B. & MANNING C. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC-06*, GENOA, ITALY.

FAURE D. & NEDELLEC C. (1998). A corpus-based conceptual clustering method for verb frames and ontology acquisition. In *The LREC workshop on Adapting lexical and corpus reesources to sublanguages and applications*, Granada, Spain.

GÓMEZ-PÉREZ A., FERNÀNDEZ-LÓPEZ M. & CORCHO O. (2004). *Ontological Engineering*. Springer Verlag.

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

GRUBER T. (1993). Toward principales for the design of ontologies used for knowledge sharing. In *Formal Analysis in Concepttual Analysis and Knowledge Representation*.

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

MAEDCHE A. & STAAB S. (2000). Discovering conceptual relation from text. In *Proceeding of the 14th European Conference on artifical intelligence*, p. 321–325, Berlin, Germany.

MALAISÉ V. (2005). *Méthodologie linguistique et terminologique pour la structuration d'ontologies différentielles à partir de corpus textuels*. Thèse d'informatique, Université de Paris 7.

MESSAI N., DEVIGNES M., NAPOLI A. & SMAIL-TABBONE M. (2006). Treillis de concepts et ontologies pour interroger l'annuaire de sources de données biologiques bioregistry. In *Ingénierie des Systèmes d'Information : Systèmes d'information spécialisés*, volume 11/1, p. 39–60.

ROUANE M., HUCHARD M., NAPOLI A. & VALTCHEV P. (2007). Proposal for combining formal concept analysis and description logics for mining relational data. In *Int. Conference on Formal Concept Analysis, ICFCA 2007, Clermont-Ferrand, France* : Springer Verlag.

SANDERSON M. & CROFT B. (1999). Deriving concept hierarchies from text. In *Research and Development in Information Retrieval*, p. 206–213.

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.