Text-based ontology construction using relational concept analysis

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

UMR 7503 LORIA, BP 239, 54506 Vandœuvre-lès-Nancy, FRANCE

Abstract. We present a semi-automated process that constructs an ontology based on a collection of document abstracts for a given domain. The proposed process relies on the formal concept analysis (FCA), an algebraic method for the derivation of a conceptual hierarchy, namely '*concept lattice*', starting from data context, i.e., set of individuals provided with their properties. First, we show how various contexts are extracted and then how concepts of the corresponding lattices are turned into ontological concepts. In order to refine the obtained ontology with transversal relations, the links between individuals that appear in the text are considered by the means of a richer data format. Indeed, Relational Concept Analysis (RCA), a framework that helps FCA in mining relational data is used to model these links and then inferring relations between formal concepts whose semantic is similar to roles between concepts in ontologies. The process describes how the final ontology is mapped to logical formulae which can be expressed in the Description Logics (DL) language $\mathcal{FLE}$. To illustrate the process, the construction of a sample ontology on the astronomical field is considered.

1 Introduction

Knowledge systems are of great importance in many fields, since they allow knowledge representation, sharing and reasoning. However, the knowledge acquisition process is complex and can be seen as a "*bottleneck*" [12]. The difficulty is to acquire knowledge (especially from experts) and then to maintain knowledge in a given domain. For example, in the area of astronomy, assigning classes to the growing number of celestial objects is a difficult task and leads to a large number of classes. Traditionally, this classification task is performed manually according to the object properties appearing in the astronomy documents. The task consists in reading articles of various sources that deal with a given celestial objects and finding the corresponding class. At present, more than three million celestial objects were classified in this way and made available through the SIMBAD database[1], but considerable work has to be done in order to classify the billion remaining objects. Moreover, human experts are not confident with the resulting classification as the classes lack precise definitions to be examined when a new object must be classified.

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

The spread of languages and frameworks for building ontologies, mainly within the Semantic Web initiative, has turned current trends in classification towards the construction of classification in the form of ontologies [15]. Ontologies are an explicit specification of a domain conceptualization, developed for the purpose of sharing and reuse. It comprises a set of concepts and a set of taxonomic and transversal relations. In attempt to bring a formal representation to the ontology components (concepts, roles, etc.), several studies [8] have documented the mapping of an ontology into DL formulae. Such translation is crucial as it makes the domain knowledge encoded by the means of ontology at the disposal of DL reasoners which in turn enables sharing and reasoning on a clear semantic basis.

The aim of this paper is to introduce a semi-automated process for the construction of classifications in the form of ontologies [15] and the derivation of expressions in Description Logics (DL) that formally describes the resulting classes. Several approaches were proposed for ontology construction, such those relying on Formal Concept Analysis (FCA) [3]. FCA is a mathematical approach for abstracting conceptual hierarchies from set of individuals (e.g., celestial objects, telescopes, etc.) and the set of their properties (e.g., emitting, collimated, mass, etc). These individuals and their properties are extracted from text corpora using NLP tools. Applying FCA with the aim of ontology construction brings forward two main benefits. First, the formal characterization of the FCA-powered concept hierarchy provides a basis for a formal specification to the derived ontology. Moreover, many efficient operations have been designed in FCA to maintain the concept hierarchy over data evaluation, such as those performing an incremental update of the hierarchy by adding either a formal object or a formal attribute and those operations for lattice assembly from parts [13]. These various operations could be used to solve the 'bottleneck' problem in knowledge acquisition. Indeed, when the concept hierarchy changes, the ontology will evolve and still be correct and consistent.

However, in order to deal with complex descriptions of individuals that go beyond a mere conjunction of properties, an extended FCA framework, namely 'Relational Concept Analysis' (RCA) is used to derive conceptual hierarchies where, beside property sharing, formed concepts reflect commonalties in object links [5]. RCA approach lifts up links between individuals to the rank of relations between concepts whose meaning is similar to roles in ontologies. RCA output — concepts organized by a partial order relation — is translated in a very obvious way to an ontology components [9]. Moreover, recent advances in combining RCA and DL languages have shown how RCA output, in particular concepts provided with relational descriptions, can be expressed in the form of DL formulae ranging in the $\mathcal{FLE}^2$ language family [7].

The proposed process is fed with astronomy data to classify celestial objects. The translation of the ontology into a DL knowledge base (KB) allows querying the KB through a DL reasoner and thus answering to '*competency questions*'.

[2] DL language that comprises the following constructors: conjunction $\sqcap$, universal quantification $\forall$ and existential quantification $\exists$.

These questions are first written in natural language and then translated into the DL language. Competency questions look like '*do objects* `M87` *and* `PSRA` *belong to the same class*?', '*Which objects can be observed with an Xray telescope*?', or '*What are the objects that* `MXX-Newton` *observes*?', etc.

The paper starts with an overview of the proposed methodology that builds a domain ontology based on free text. The next section introduces the processing texts with NLP tools that are used to collect RCA data. Section 4 recalls the FCA method, its extended framework RCA, and their application to the domain of astronomy. Section 5 presents the translation of the RCA output into DL KB. First, general rules are listed and then applied to the result of the previous step. We present in the section 6 the related work and conclude with brief discussion on the learned facts and the remaining open issues.

2 Methodology

Our methodology (described in figure 1) is based on "Methontology" [1]. The "Methontology" is a semi-automatic methodology, that builds an ontology from a set of terms extracted from resources (the resources are not specified). The objective is to find the exhaustive definition for each concept and each relation of the ontology in DL language. The four steps of the "Methontology" are adapted on proposed methodology.

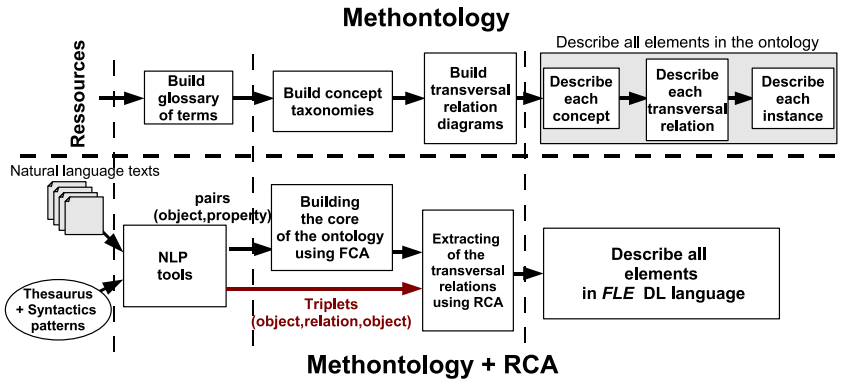

Fig. 1. Mapping between the "Methontology" and Methodology + RCA

Resources: They are represented by the texts corpora, the thesaurus of astronomy[3] and the syntactic patterns[4] such as: all `NGC nnnn` where n is a number represents one celestial object.

[3] http://msowww.anu.edu.au/library/thesaurus/
[4] http://simbad.u-strasbg.fr/simbad/sim-fid

Build glossary of terms: The extraction of the terms is done from the texts corpora using the existing resources in the astronomical domain. We extract also in this step the pairs (object,property) and the tuples (object,relation,object) using Natural language processing (NLP) tools.

Build concept taxonomies: We propose in this step to use the FCA. The FCA is the mathematical tool (presented in the section 4) that builds the hierarchy of concepts by grouping the terms sharing the same properties.

Build transversal binary relation diagrams: The extraction of the transversal relations is done in the same time as the construction of new hierarchy of concepts taking into account their properties and also their links with other objects. This step is done with RCA (see the section 4).

Describe all elements of the ontology: The representation of all concepts, relations and instances is done with $\mathcal{FLE}$ language. The representation in a DL language is done to support reasoning, i.e. classification, instantiation and consistency checking (see the section 5).

3 Processing texts with NLP tools

We want to extract the pairs (object,property) and the tuples ($object_1$,link,$object_2$) from the text corpora. The links, in the tuples, are used to define the set of the relations in the ontology (see section 4.2). We choose to use the Faure's approach [4] based on the Harris hypothesis [16]. This hypothesis studies the syntactic regularities in the text corpora of sub-languages (or specific languages), allowing to identify the syntactic schema to build classes. There classes are grouping the terms (celestial objects) that are arguments of the same set of verbs, i.e., the subject of the same set of verbs and the complement of the same set of verbs. For example: The set {HR5223, PRSA, SS433} are in the same class because they are appearing as subject with the verb {to emit} and as complement with the set of verbs {to observe,to locate}. The set of verbs is translated to the set of properties, like for example if one term are subject of the verb "to emit", it has a property "emitting" and if one term are complement of the verb "to observe", it has the property "observed". We use the same approach to extract the set of links, if $object_1$ is the subject of the verb V and the $object_2$ the complement of the verb V then we extract the tuple ($object_1$,VP,$object_2$) where VP is the verb phrase which represent the link between ($object_1$,$object_2$).

The parsing of the corpus is done with the shallow parser "Stanford Parser"[5] [6]. We give two examples in the astronomic domain:

1. *"One HR2 candidate was* `detected` *and* `regrouped` *in each of the galaxies NGC 3507 and CygnusA"*. We extract the pairs: (HR2, `regrouped`), (HR2, `detected`), (NGC 3507, `regrouping`), (CygnusA, `regrouping`).

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

2. *'The XMM-Newton X-ray telescope observed the bursting pulsar M87"*, the extraction process will first identify `XMM-Newton X-ray` as a Telescope, and `M87` as a `celestial object`. We extract the tuple : (`M87`, ObservedByXRay,`XMM-Newton X-ray`).

4 Background on concept lattices

4.1 Basics of FCA

FCA is a mathematical approach to data analysis based on lattice theory. The basic data format in FCA [3] is a binary table $\mathcal{K} = (G, M, I)$ called *formal context*, where G is a set of individuals (called *objects*), M a set of properties (called *attributes*) and I the relation "has" on $G \times M$. Table in the left-hand side of Fig. 2 represents an example of context. Here, G is the set of `celestial objects` and M the set of their properties. A pair (X, Y) where X is a maximal set of individuals (called *extent*) and Y is a maximal set of shared properties (called *intent*), is called a *formal concept*. For instance, $(\{Andromeda, NGC3507\}, \{observed, grouping\})$ is a concept (see diagram in the right hand side of Fig. 2).

Celestial objects	emitting	accreting	collimated	observed	located	grouping
PSRA	X			X	X	
NGC3507				X		X
Andromeda		X		X		X
M87			X	X	X	
HR2				X	X	
NGC2018			X	X	X	
HR5223	X			X	X	
SS433	X			X	X	

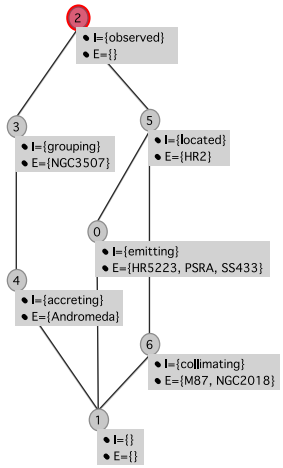

Fig. 2. The binary context of celestial objects and the corresponding concept lattice.

Furthermore, the set $\mathcal{C}_\mathcal{K}$ of all concepts of the context $\mathcal{K} = (G, M, I)$ is partially ordered by extent inclusion also called the *specialization* (denoted $\leq_\mathcal{K}$) between concepts. $\mathcal{L} = \langle \mathcal{C}_\mathcal{K}, \leq_\mathcal{K} \rangle$ is a complete lattice, called the *concept lattice*. Fig. 2 illustrates a context and its corresponding lattice. A simplified (or reduced) labeling schema is often used where each object and each attribute appear only once on the diagram. The full extent of a concept is made up of all objects

whose labels can be reached along a descending path from the concept while its full intent can be recovered in a dual way (ascending path). For details on the construction of concept lattices, see [3].

As many practical applications involve non-binary data, *many-valued contexts* has been introduced in FCA where individuals have value associated to properties. The construction of a lattice for this kind of contexts requires a preprocessing step, called *conceptual scaling* [3], that derives a binary context out of many-valued one. Scaling turns a non-binary attribute into a set of binary ones representing abstractions of values on the domain of the underlying non-binary attribute. For instance, the values of non-binary attribute *orbitalPeriod* in the context illustrated in Tab. 1 could be distributed on the ranges *short* and *long*, each of them expressed as a predicate (e.g., *orbital period* $\leq$ 24 *hours* for short one). Observe that the definition of the predicates precedes the scaling task and is usually in charge of a domain expert.

4.2 From FCA to RCA

Relational Concept Analysis (RCA)[5] was introduced as an extended FCA framework for extracting formal concepts from sets of individuals described by '*local*' properties and links. In RCA data are organized within a structure called '*relational context family*' (RCF). RCF comprises a set of contexts $\mathcal{K}_i = (G_i, M_i, I_i)$ and a set of binary relations $r_k \subseteq G_i \times G_j$, where G_i and G_j are the object sets of the contexts $\mathcal{K}_i$ and $\mathcal{K}_j$, called *domain* and *range*, respectively. For instance, table in Fig. 2 and Tab. 1 depict a sample RCF made of two contexts, `celestial objects` context and `telescopes` context.Two inter-context relations, '*Observed By Xray*' (OBXray) and '*Observed By Infrared*' (OBInfrared) indicate the observation links between telescopes and objects.

The relationnal and non relationnal attributes in both contexts list the features of objects such as the orbit height (perigee) and the orbital period for telescopes and emitting or grouping faculty for the `celestial objects`.

Telescopes			
	perigee	orbitalPeriod	mass
BeppoSAX	600 *km*	96 *min*	1400 *kg*
XMM-Newton	114000 *km*	48 *hours*	3800 *kg*
Chandra	26300 *km*	66 *hours*	1790 *kg*

OBXray			
	BeppoSAX	XMM-Newton	Chandra
M87		X	
NGC2018			X

OBInfrared			
	BeppoSAX	XMM-Newton	Chandra
HR5223	X		
SS433	X		

Table 1. Sample RCF encoding astronomy data.

RCA uses the mechanism of '*relational scaling*' which translates domain structures (concept lattices) into binary predicates describing individual subsets. Thus, for a given relation r which links formal objects from $\mathcal{K}_i = (G_i, M_i, I_i)$

to those from $\mathcal{K}_j = (G_j, M_j, I_j)$, new kind of attributes, called '*relational attributes* are created and denoted by r:c, where c is concept in $\mathcal{K}_j$. For a given object $g \in G_i$, relational attribute $r : c$ characterizes the correlation of $r(g)$ and the extent of $c = (X, Y)$. Many levels of correlation can be considered such as the 'universal' correlation $r(g) \subseteq X$ and the 'existential' correlation $r(g) \cap X$. Due to correlation constraint, existential encoding of object links yields to richer link sharing among objects and thus a wider conceptual structure to explore when mining relevant concepts. In the present work, we consider only existential scaling.

	perigeeIsLow	perigeeIsHigh	shortOrbitalPeriod	longOrbitalPeriod	light	heavy
BeppoSAX	X		X		X	
XMM-Newton		X		X		X
Chandra		X		X	X	

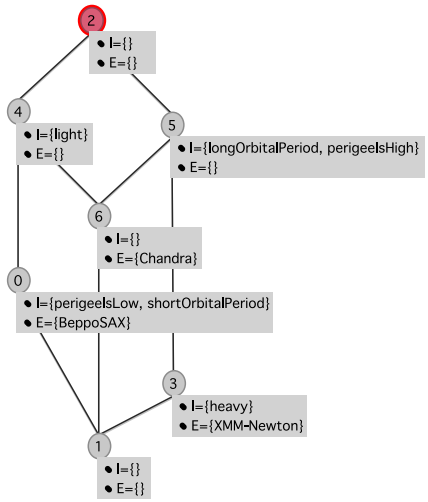

Fig. 3. The derived context of telescopes and the corresponding lattice.

For example, suppose that the context of celestial objects has to be scaled along the relation OBXray with respect to the lattice given in Fig. 3. As OBXray(M87) = {XMM_Newton} and the telescope XMM_Newton is present in the extent of concepts c_2, c_3 and c_5 (see Fig. 3), the celestial objects context is extended by relational attributes of the form $r : c_i$, where $i = \{2, 3, 5\}$. Tab. 2 depicts the extended context of celestial objects after the scaling of both relations OBXray and OBInfrared. It can be noticed that beside local attributes, new relational attributes encode object links that have been assigned to objects. For instance, in Figure 4, objects HR5223 and SS433 in the concept c_9 share the attribute OBInfrared:c0 which is interpreted as a common link with telescope BeppoSAX (the only object in the extent of concept c_0 of Figure 3).

	Local attributes						Relational attributes														
	emitting	accreting	collimated	observed	located	grouping	OBxray:c_0	OBxray:c_1	OBxray:c_2	OBxray:c_3	OBxray:c_4	OBxray:c_5	OBxray:c_6	OBinfrared:c_0	OBinfrared:c_1	OBinfrared:c_2	OBinfrared:c_3	OBinfrared:c_4	OBinfrared:c_5	OBinfrared:c_6
HR5223	X			X	X									X		X		X		
M87			X	X	X				X	X		X								
SS433	X			X	X									X		X		X		
NGC2018			X	X	X				X		X	X	X							

Table 2. The result of scaling of celestial objects context along its relations. Formal objects that are not affected by relational scaling are not displayed.

4.3 Qualitative interpretation of RCA

The relational scaling is the key step in a process which, given an RCF, derives a relational lattice family (RLF), one lattice by context. A relational attribute is interpreted as a relation between two concepts, on the first side the concept whose intent owns this attribute (i.e. the domain), and, on the other side, the concept indicated in the relational attribute expression (i.e. the range). The RLF extraction process is iterative since relational scaling modifies contexts and thereby the corresponding lattices, which in turn, implies a re-scaling of all the relations that use these lattices as source of predicates. This iterative process stops when a fixed point is reached, i.e., additional scaling steps do not involve any more context extension.

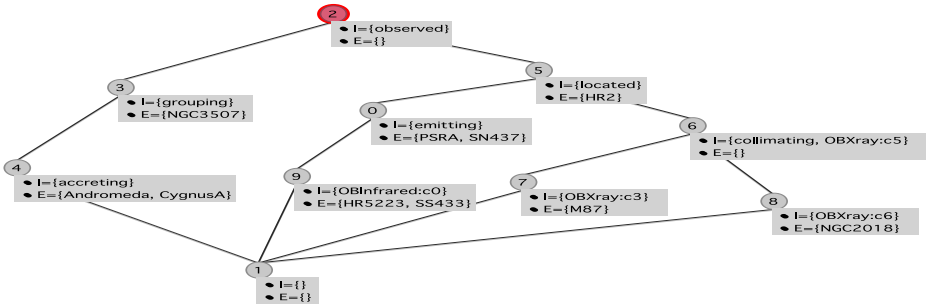

Fig. 4. The final relational lattice of celestial objects context

The analysis of the sample RCF using RCA process yields to the concept lattices illustrated in Fig. 3 and Fig. 4. Relational attributes in concept intents are associated to the most specific concepts in the corresponding lattice. Telescope

context is not a domain of relation in the running RCF. Therefore, the final lattice corresponds to the initial one shown in Fig. 3. By contrast, the lattice of celestial objects context has changed. The resulting concepts trigger yet further sharing, at the object links level. Indeed, the intents of various formal objects are enriched with relational attributes encoding inter-object links. These attributes lift up object link to relations between concepts. For example, the concept c_6 in Fig. 2 represents the celestial objects `M87` and `NGC2018`, that are both binary stars as they are observed, located and collimated. The intent of the former concept is encoded with the relational attribute $\texttt{OBXray:c}_5$, meaning binary stars are also observable by XRay telescopes. Moreover, new concept are discovered. For example, even if the two celestial objects `HR5223` and `SS433` have already composed a formal concept in the initial lattice (concept c_0 in Fig. 2) with an additional object, namely `PSRA` they let a new concept emerge in the final lattice (concept c_9 in Fig. 4), due to the common link they share with the telescope `BeppoSAX`. The new concept represents the stars that are observable with an Infrared telescope such as `BeppoSAX`.

5 Ontology derivation

The ontology resulting from the RCA process is represented with the DL $\mathcal{FLE}$.

The TBox		
RCA entity	Ontology	Example
Context $\mathcal{K}$	Atomic concept c $\equiv \alpha(\mathcal{K})$	$\alpha(\texttt{Telescope}) \equiv$ Telescope
Formal attribute $m \in M$	Defined concept c $\equiv \alpha(m) \equiv \exists m.\top$	$\alpha(\texttt{observed}) =$ Object $\equiv \exists$observed.$\top$
Concept $c = (X, Y) \in \mathcal{C}$	Defined concept $\alpha(c)$, i.e. $\alpha(c) \equiv \sqcap_{m \in Y} \alpha(m)$	$\alpha(\mathsf{C}_5) \equiv \exists$observed.$\top \sqcap \exists$located.$\top$
$\forall (c, \bar{c}) \in \mathcal{C} \times \mathcal{C}$, i.e. $c \prec \bar{c}$	Inclusion axiom $\alpha(c) \sqsubseteq \alpha(\bar{c})$	$\alpha(\mathsf{C}_8) \sqsubseteq \alpha(\mathsf{C}_6)$
Relation $r \in R$	primitive role $\alpha(r)$	OBXray is a primitive role in the TBox
Relational attribute $r.C$	Atomic concept c $\equiv \alpha(r) \equiv \exists r.\alpha(c)$	α(OBXray.XMM-Newton)$\equiv \exists$OBXray.XMM-Newton
The ABox		
RCA entity	Ontology	Example
Formal object $g \in G$	Instance $\alpha(g)$	Andromeda is an instance
Element $(g, m) \in I$	Assertion $\alpha(m)(\alpha(g))$	Object(HR2)
Let $c = (X, Y)$, $\forall g \in X$	Concept instantiation $\alpha(c)(\alpha(g))$	HR2 is an instance of the concept `Star`

Table 3. Mapping between lattice and DL knowledge base

The translation between the RCA formal concepts and relations and the DL $\mathcal{FLE}$ is carried on using a function α defined as follows: $\alpha : (\mathbf{K}, \mathbf{R}) \rightarrow$ TBox $\sqcup$

ABox, where: $(\mathbf{K}, \mathbf{R})$ is a family RCF, TBox and ABox being the components of the ontology. The function α is presented in the Tab. 3. The application of the function α in the two lattices (Fig. 3 and Fig. 4) results in the ontology in the Fig.5.

5.1 The translation of the concepts lattice into the ontology

The translation of each context represents an atomic concept, that express the top $\top$ of the hierarchy in this context. Each formal attribute is translate in defined concept. For example, attribute `observed` is translated into the concept $\mathsf{c} \equiv \exists\, \mathsf{observed}.\top$. Each relational attribute r.C is translated in defined concept in the TBox. For example, the relational attribute if the form OBXray.BeppoSAX is translated into $\mathsf{c} \equiv \exists\, \mathsf{OBXray.BeppoSAX}$, etc.

The design of the ontology is carried out in collaboration with astronomers. The astronomers have to give a label to each concept in the ontology according to the properties and the links associated to the instances of a concept. For example, the class of objects having the set of properties {`observed,located,collimating`} and the link {`Observed-By-Xray`} with the range `X-Ray-Telescope` is labeled by `Binary-Star`. The class of objects having the set of properties {`observed,located,emitting`} and the relation {`Observed-By-Infra-Red`} with the range `Infra-Red-Telescope` is labeled by `Pulsing-Variable-Star`: `Infra-Red-Telescope` observes `Young-Star` that has a large emission compared with the `X-Ray-Telescope` that observes older stars like `Binary-Star`. This representation is done only to give one label for each set of celestial objects and to help the experts to read the ontology.

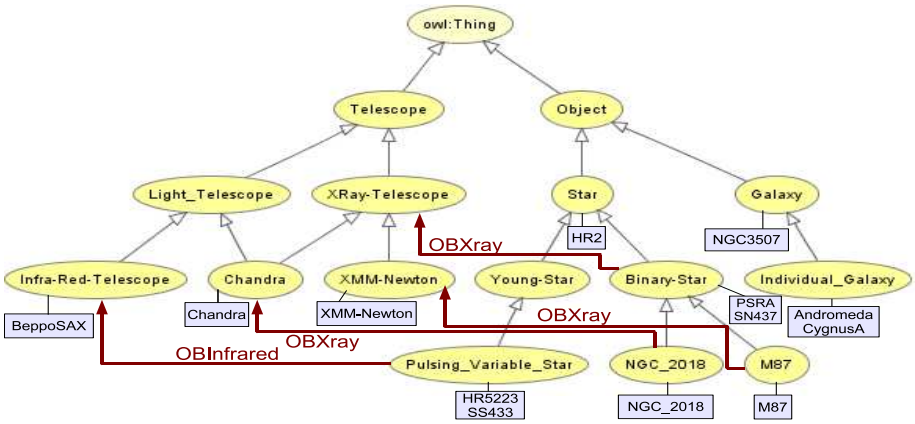

Fig. 5. Complete Ontology

5.2 Representation of the concepts in the DL language $\mathcal{FLE}$

The ontology is represented within the $\mathcal{FLE}$ language. Tab. 4 presents the definition of each concept in the ontology presented in the figure (Fig. 5). The ontology can be used for three kinds of tasks :

N° in the lattice	Concept Name	Defined Concept
C_2	Object	$\exists$observed.$\top$
C_5	Star	$\exists$observed.$\top \sqcap \exists$located.$\top$
C_0	Young-Star	$\exists$observed.$\top \sqcap \exists$located.$\top \sqcap \exists$emitting.$\top$
C_9	Pulsing-Variable-Star	$\exists$observed.$\top \sqcap \exists$located.$\top \sqcap \exists$emitting.$\top \sqcap \exists$OBInfrared. Infra-Red-Telescope
C_6	Binary-Star	$\exists$observed.$\top \sqcap \exists$located.$\top \sqcap \exists$collimated.$\top \sqcap \exists$OBXray.Xray_Telescope
C_7	M87	$\exists$observed.$\top \sqcap \exists$located.$\top \sqcap \exists$collimated.$\top \sqcap \exists$OBXray.XMM-Newton
C_8	NGC_2018	$\exists$observed.$\top \sqcap \exists$located.$\top \sqcap \exists$collimated.$\top \sqcap \exists$OBXray.Chandra
C_3	Galaxy	$\exists$observed.$\top \sqcap \exists$grouping.$\top$
C_4	Individual-Galaxy	$\exists$observed.$\top \sqcap \exists$grouping.$\top \sqcap \exists$accreting.$\top$
T_2	Telescope	Telescope
T_4	light_Telescope	$\exists$light.$\top$
T_5	XRay-Telescope	$\exists$longOrbitalPeriod.$\top \sqcap \exists$perigeeIsHight.$\top$
T_0	Infra-Red-Telescope	$\exists$shortOrbitalPeriod.$\top \sqcap \exists$perigeeIsLow.$\top$
T_6	Chandra	$\exists$longOrbitalPeriod.$\top \sqcap \exists$perigeeIsHight.$\top \sqcap \exists$light.$\top$
T_3	XMM-Newton	$\exists$longOrbitalPeriod.$\top \sqcap \exists$perigeeIsHight.$\top \sqcap \exists$heavy.$\top$

Table 4. Definition of each concept of the Fig 5 in $\mathcal{FLE}$

1. **Ontology population:** Let o_1 an object with the properties {a,b}, and the relations {r_1.c_1,r_2.c_2}. A first task is instantiation,i.e. to find the class of an object such as o_1. The class of o_1 is the most general class X such that $X \sqsubseteq \exists a.\top \sqcap \exists b.\top \sqcap \exists r_1.c_1 \sqcap \exists r_2.c_2$. For example, let us consider the question *"What is the class of the object GRO, that has the properties* {observed,located,emitting} and the relation OBInfrared with the range Infra-red-Telescope? The answer is: the most general class $X \sqsubseteq \exists$observed.$\top \sqcap \exists$located.$\top \sqcap \exists$emitting.$\top \sqcap \exists$OBInfrared. Infra-red-Telescope. This class in the ontology is the concept Pulsing-Variable-Star.
2. **Comparison of celestial objects:** Let us consider two objects o_1 and o_2. A second task consists in comparing o_1 and o_2 and determining whether o_1 and o_2 have the same class. One way for checking that is to find the

class of o_1, then the class of o_2, and then to test whether the two classes are equivalent. For example, let us consider the two objects `M87` and `PSRA`. `M87` is an instance of the class `M87` and `PSRA` is an instance of the class `Young-Star`. Knowing that `M87` $\sqcap$ `Young-Star` $= \bot$, it can be inferred that both objects do not belong to the same class.

3. **Detection of the domain or the range of relation:** Let us consider the relation r_1 with the range C_1. A third task consists in finding the domain of the relation r_1. The domain of r_1 is the most specific class X such that X is the most specific class, union of all the classes linked to the class C_1 by the relation r_1. For example *Which objects can be observed by Xray with a Xray telescope*? The most specific class domain of the relation `observed by Xray` where `Xray telescope` is the range, is the concept `Binary-star`.

6 Related work

6.1 Building the core ontology

There are two main approaches for building ontologies from text corpora. The first one is based on the co-occurrence of terms in text and on the use of similarity measures for building the hierarchy of the objects classes [10]. This approach can not satisfy our needs to give a definition to each concept of the hierarchy, because every concept is represented by numeric vector and it is difficult to find an interpretation for each vector. The second approach is symbolic, and is based on the use of a syntactic structure to describe an object by the verb with which it appears. Faure uses this structure for building the object classes and the statistic measures for building the hierarchy of the classes [4]. Cimiano uses the same approach but builds the hierarchy of classes using FCA, without taking into account the relations between objects [12].

6.2 Extracting the transversal relations

The extraction of transversal relations allows us to have a better definition of each concept. The concepts are not only defined by their properties but also by their relations with other concepts. We cite two related approaches in the extraction of relations. The first one is the work of Aussenac-Gilles [11], who proposes to use a learning method to extract syntactic patterns. Tuples manually extracted from the texts (`term`$_1$`, relation`$_1$`, term`$_2$`)` are the inputs. All the tuples (`term`$_1$`, relation`$_k$`, term`$_2$`)` are searched to build a general relation R, such that $R = relation_1 \sqcup \ldots \sqcup relation_n$. Then, tuples of the form $(term_i, R, term_j)$ are extracted. This method groups the set of objects according to the relations that they share, and extracts the general relations between two concepts. It does not use the hierarchy of the concepts to make a generalization. A second approach by Maedche and Staab [2] consists in extracting the association rules [14] ($term_1 \Rightarrow term_2$) and in keeping only those rules having a given support and frequency. This method finds all the pairs (C_1, C_2) linked by one relation but does not specify the name of the relation between these pairs.

7 Conclusion

A method for building an ontology from text corpora was proposed. The method uses the RCA framework that extends standard FCA for mining relational data. RCA derives a structure that is compatible with an ontology. We have shown how RCA output could be represented in terms of DL expressions ranging in the $\mathcal{FLE}$ DL family. The proposed method was applied to the astronomy domain in order to extract knowledge about celestial objects that can be used through a DL reasoner for problem-solving such as celestial objects classification and comparison. The construction of a first prototype ontology from astronomy data proved that RCA-based ontology construction is a promising method allowing to data mining and knowledge representation techniques.

On going work consists in improving the RCA input data gathering process by considering alternate syntactic patterns in the extraction of object pairs such as (subject, verb), (complement, verb), (subject, adjective), etc. These new sorts of pairs will provide a contexts with additional formal attributes that make formal object descriptions richer as well as a new inter-context relations. Eventually, the construction of hierarchy of relations need to be addressed. The principle consists of using once again the RCA abstraction process to introduce abstract relations between concepts based on the transversal relations —originally inferred from instances links— that hold among their subsumers. Once the derived relation hierarchy merged with concept hierarchy, the resulting structure forms a complete ontology that fully captures the domain knowledge.

References

1. Gómez-Pérez A., M. Fernàndez-López, and O. Corcho. *Ontological Engineering*. Springer Verlag, 2004.
2. Maedche A. and S. Staab. Discovering conceptual relation from text. In *Proceeding of the 14th European Conference on artifical intelligence*, pages 321–325, Berlin, Germany, 2000.
3. Ganter B. and R. Wille. *Formal Concept Analysis Mathematical Foundations*. Springer Verlag, 1999.
4. Faure D. and C. Nedellec. 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, 1998.
5. M. Dao, M. Huchard, M. Hacene Rouane, C. Roume, and P. Valtchev. Improving generalization level in uml models: Iterative cross generalization in practice. In *Proceedings of the 12th International Conference on Conceptual Structures (ICCS'04)*, volume 3127 of *Lecture Notes in Computer Science*, pages 346–360, Huntsville, AL, July 2004. Springer-Verlag.
6. Marneffe M.C. de., B. MacCartney, and C.D. Manning. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC-06*, GENOA, ITALY, 2006.
7. Baader F. Description logic terminology. In Baader F., D. Calvanese, D. McGuinness, N. Daniele, and P.F. Patel-Schneider, editors, *The Description Logic Handbook: Theory, Implementation, and Applications*, pages 485–495. Cambridge University Press, 2003.

8. Baader F., I. Horrocks, and U. Sattler. Description logics as ontology languages for the semantic web. In Hutter D. and W. Stephan, editors, *Mechanizing Mathematical Reasoning: Essays in Honor of Jörg H. Siekmann on the Occasion of His 60th Birthday*, volume 2605 of *Lecture Notes in Artificial Intelligence*, pages 228–248. Springer-Verlag, 2005.
9. Rouane M. H., M. Huchard, A. Napoli, and P. Valtchev. 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*, Lecture Notes in Computer Science. Springer Verlag, 2007.
10. Sanderson M. and B. Croft. Deriving concept hierarchies from text. In *Research and Development in Information Retrieval*, pages 206–213, 1999.
11. Aussenac-Gilles N., B. Biébow, and S. Szulman. Revisiting ontology design: A method based on corpus analysis. In Dieng R. and O. Corby, editors, *12th Int. Conference in Knowlegde Engineering and knowledge Management (EKAW'00)*, volume 1937, pages 172–188, 2000.
12. Cimiano P., A. Hotho, and S. Staab. Learning concept hierarchies from text corpora using formal concept analysis. In *Journal of Artificial Intelligence Research (JAIR)*, volume Volume 24, pages 305–339, 2005.
13. Valtchev P., M. Rouane Hacene, and R. Missaoui. A generic scheme for the design of efficient on-line algorithms for lattices. In A. de Moor, W. Lex, and B. Ganter, editors, *Proceedings of the 11th Intl. Conference on Conceptual Structures (ICCS'03)*, volume 2746 of *Lecture Notes in Computer Science*, pages 282–295, Berlin - Germany, 2003. Springer.
14. Agrawal R. and R. Srikant. Mining generalized association rules. In *21st VLDB Conference*, Zurich, Swizerland, 1995.
15. Gruber T.R. Toward principales for the design of ontologies used for knowledge sharing. In *Formal Analysis in Concepttual Analysis and Knowledge Representation*, 1993.
16. Harris Z. *Mathematical Structure of Language*. Wiley J. and Sons, 1968.