Ontology Learning from Text using Relational Concept Analysis

Mohamed Rouane Hacene*, Amedeo Napoli*, Petko Valtchev**, Yannick Toussaint*, and Rokia Bendaoud*

* INRIA-LORIA Laboratory
Campus Scientifique, BP 239
Vandoeuvre, Nancy Cedex, France
{rouanehm,napoli,toussaint,bendaoud}@loria.fr

** LATECE Laboratory
Department of Computer Science, UQAM
CP 8888, succ. Centre-Ville, Montréal, H3C 3P8, Canada
valtchev.petko@uqam.ca

Abstract

We propose an approach for semi-automated construction of ontologies from text whose core component is a Relational Concept Analysis (RCA) framework which extends Formal Concept Analysis (FCA), a lattice-theory paradigm for discovering abstractions within objects × attributes tables, to the processing of several sorts of individuals described both by own properties and inter-individual links. As a pre-processing, text analysis is used to transform a document collection into a set of data tables, or contexts, and inter-context relations. RCA then turns these into a set of concept lattices with inter-related concepts. A core ontology is derived from the lattices in a semi-automated manner, by translating relevant lattice elements into ontological concepts and relations, i.e., either taxonomic or transversal ones. The ontology is further refined by abstracting new transversal relations from the initially identified ones using RCA. We discuss as well the results of an application of the method to astronomy texts.

1. Introduction

Ontologies have become an important means for sharing, reusing and reasoning about domain knowledge. An ontology formally represents the knowledge about a domain in terms of classes of entities, their features, relationships and instances [16]. The construction of an ontology, typically carried out by a human expert, is a time-consuming and an error-prone task. Hence it has been targeted with a variety of semi-automated approaches whose bottom line is the discovery of the ontology elements from descriptive documents about the domain. In this respect, techniques for the construction of conceptual hierarchies have proved particularly suitable as they naturally output the backbone of an ontology, i.e., the taxonomy of concepts.

Formal concept analysis (FCA) [9] is an approach for abstracting conceptual hierarchies from sets of individuals (e.g., celestial bodies) described by properties (e.g., located, emitting, etc.) rooted in lattice-theory. FCA has been advocated to support ontology construction [3, 17]. Indeed, a major benefit of using FCA lays in the mathematical characterization of its output conceptual hierarchy, the concept lattice. It implies a number of valuable properties, e.g., the FCA concepts being described both intensionally and extensionally whereby extensions are maximal sets of individuals matching the corresponding intensions, that eases the translation into a formal ontology.

Although concept lattices are easily translated into a set of ontology concepts and their taxonomic relations, this is not enough for general ontologies featuring non-taxonomic (transversal, semantic) relations. The latter cannot be determined since core FCA does not cover relations in the data. The need for concept analysis of richer data formats has motivated the design of the relational concept analysis (RCA) framework [5]. Besides classical contexts, RCA enables the abstraction from inter-individual links while admitting several sorts of individuals. Thus, within an ontology construction process, it helps discovering transversal relations between concepts based on links between their respective instances. To that end, RCA turns links into binary attributes.
using quantifier-based selectors similar to role restriction in description logics.

In this paper, we introduce a RCA-based approach for ontology construction that combines distinct, yet complementary, research in FCA and ontology learning. The approach extracts a formal contexts from document corpus using linguistic analysis and derives two is-a hierarchies, of ontological concepts and of the associated transversal relations. The paper starts with a motivating example that is based on the actual experiments followed by a background on FCA and a brief presentation of the RCA framework. The core of our approach is then presented. We conclude with a summary of what was accomplished and the remaining open issues.

2. Motivating example

One of the main purposes of astronomy is assigning classes to the growing number of celestial bodies in attempt to explain the behavior of the universe. Traditionally, this classification task is performed manually according to the body properties mentioned in the scientific publications. The task consists in reading articles of various sources that deal with a given celestial body and then find the appropriate class. At present, more than three million celestial bodies are classified in this way and made available through the SIMBAD database\(^1\), but work has to continue for classifying the billion remaining bodies. Moreover, experts are not confident with the resulting classification as class definitions lack precision.

The spread of languages and frameworks for constructing ontologies, in particular, within the Semantic Web initiative, has turned current trends in classification towards ontology-shaped classifications. Here we present a way to use RCA in the construction of a celestial bodies ontology. The target ontology is intended to support answering competency questions whose goal could be to classify bodies, retrieve their properties or simply recognize the most specific class that holds a given set of bodies, e.g., ‘do celestial bodies M87 and PSRA belong to the same class?’; ‘Which bodies can be observed with an Xray telescope?’; etc. For illustration purposes, Table 1 provides a comprehensive example drawn from our actual experiments.

3. Formal concept analysis

FCA is a lattice theory-based data analysis paradigm. The basic data format in FCA [9] is an objects × attributes table called (formal) context. A context is a triplet \( \mathcal{C} = (O, A, I) \), where \( O \) is a set of individuals, called formal objects, \( A \) is a set of properties called formal attributes, and \( I \) is an incidence relation representing pairs \( oIa \) which stand for “object \( o \) has the attribute \( a \)”. For instance, Table 2 illustrates a context derived from text documents (in a way described in Sec.5.2). Here \( O \) is a set of celestial bodies and \( A \) their properties.

<table>
<thead>
<tr>
<th>Celestial bodies</th>
<th>emitting</th>
<th>accreting</th>
<th>observed</th>
<th>located</th>
<th>grouping</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSRA</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M87</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Andromeda</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HR2</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HR5223</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SS433</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Sample sentences of an abstract from the Astronomy and Astrophysics main Journal.

Two derivation operators, both denoted ‘ link objects and attributes. Let \( X \subseteq O, Y \subseteq A: X' = \{a \in A|\forall o \in X, oIa\}, Y' = \{o \in O|\forall a \in Y, oIa\}. \) For example, following the encoding of Table 2, \( \{PSRA, M87\}' = \{\text{observed, located}\}. \) The compound operators ” are closure operators over \( 2^O \) and \( 2^A \) respectively. A pairs of ‘-connected sets is called a formal concept. Formally, \( c = (X,Y) \in O \times A \) is a concept of \( \mathcal{K} \) if \( X' = Y \) and \( Y' = X \) (here \( X \) and \( Y \) are called the extent and the intent of \( c \), respectively). For instance, \( \{\text{Andromeda,NGC3507}\}, \{\text{observed,grouped}\}\) is a concept (see Figure 1). The set \( \mathcal{K}_C \) of all concepts of the context \( \mathcal{K} \) is partially ordered by extent inclusion also called the specialization between concepts. The structure \( \mathcal{L} = (\mathcal{K}_C, \leq \mathcal{K}) \) is a complete lattice, called the concept lattice. For instance, Table 2 illustrates a formal context whereas Figure 1 presents the corresponding concept lattice. A simplified (or reduced) labeling scheme is often used where each object and each attribute appear only once on the diagram as illustrated in Figure 1. 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. For example, the extent of the concept with the object label ‘NGC3507’ is \( \{\text{NGC3507, Andromeda}\} \).

\(^1\)http://simbad.u-strasbg.fr/simbad/sim-fid
As many practical applications manipulate richer data formats, many-valued contexts have been introduced in FCA. In such a context \( K = (O, A, V, J) \), an object \( o \) is described by a set of attribute value pairs \((a, v)\), meaning that \( J \) is a ternary relation that binds the objects from \( O \), the attributes from \( A \) and the values from \( V \). For instance, context depicted in Table 3 represents three telescopes with their main features such as the orbital period and the mass.

The construction of a lattice for \( K \) requires a pre-processing step, called conceptual scaling [9], that derives a binary context out of \( K \). Scaling turns a non-binary attribute \( a \) into a set of binary ones using a scale context \( K_a = (V_a, P_a, J_a) \) where scale attributes \( P_a \) are abstractions of the values on \( V(a) \). For instance, the values of the attribute \textit{perigee} of the context in Table 3 could be distributed on the ranges low and high, each of them expressed as a predicate (e.g., \textit{perigee} \( \leq \) 1000 km for low one). Non-binary attributes \textit{orbitalPeriod} and \textit{mass} are distributed on the ranges short or long, and heavy or lightweight, respectively. Figure 2 illustrates the lattice derived from many-valued context given in Table 3.

### 4. Relational concept analysis

Main concern of RCA is to infer relations between formal concepts based on links among formal objects. For instance, we would like to have an observation relation between the concept \( c_0 \) in Figure 1 representing celestial bodies HR5223 and SS433 and the concept \( c_0 \) in Figure 2 representing the telescope BeppoSAX as both bodies are observed by this telescope. RCA comes up with formal concepts that are connected in same way as ontological concepts are connected by means of role restrictions involving logical quantifiers. This allows the mapping of concept lattice into an ontology to be done in a simple yet natural way. In RCA, the data are organized within a structure called relational context family (RCF).

**Definition 1** A relational context family \( \mathcal{R} \) is a pair \((K, R)\), where \( K = \{K_i\} \) is a set of formal contexts \( K_i = (O_i, A_i, I_i) \), and \( R = \{r_k\} \) is a set of binary relations \( r_k \subseteq O_i \times O_j \), where \( O_i \) and \( O_j \) are the object sets of \( K_i \) (domain) and \( K_j \) (range), respectively.

<table>
<thead>
<tr>
<th>Telescopes</th>
<th>perigee</th>
<th>orbitalPeriod</th>
<th>mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>BeppoSAX</td>
<td>600 km</td>
<td>96 min</td>
<td>1400 kg</td>
</tr>
<tr>
<td>XMM-Newton</td>
<td>114000 km</td>
<td>48 hours</td>
<td>3800 kg</td>
</tr>
<tr>
<td>Chandra</td>
<td>26300 km</td>
<td>66 hours</td>
<td>1790 kg</td>
</tr>
</tbody>
</table>

Table 3. The ‘many-valued’ context of telescopes.

For instance, the relations ‘\textit{Observed with Xray}’ (OWxray) and ‘\textit{Observed with Infrared}’ (OWinfrared) model observation links between telescopes and celestial bodies (see Table 4). Both relations together with the contexts of celestial bodies (Table 2) and of telescopes (Table 3) form our sample RCF.

<table>
<thead>
<tr>
<th>OWxray</th>
<th>BeppoSAX</th>
<th>XMM-Newton</th>
<th>Chandra</th>
</tr>
</thead>
<tbody>
<tr>
<td>M87</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>NGC2018</td>
<td>X</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OWinfrared</th>
<th>BeppoSAX</th>
<th>XMM-Newton</th>
<th>Chandra</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR5223</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>SS433</td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

Table 4. Sample relations encoding astronomy data.
A dedicated scaling mechanism is used to translate links into concept attributes. To that end, relations are interpreted as attributes whose values are object sets, hence the target attributes are predicates describing these sets. The predicates are themselves derived from the available conceptual structures on the underlying context. Thus, for a given relation seen as a function \( r : O_i \rightarrow \phi(O_j) \), new attributes, called relational, of the form \( r:c \), are added to \( K_j \), where \( c \) is concept on \( K_j \). An object \( o \in O_i \) owns an attribute \( r:c \) depending on the relationship between its link set \( r(o) \) and the extent of \( c \), denoted \( Ext(c) \). It can be either inclusion, i.e., \( r(o) \subseteq X \) (called universal scaling schema), or non-empty intersection, i.e., \( r(o) \cap X \) (called existential scaling schema). In the following, only existential scaling is considered. Formally:

**Definition 2** Given a context \( K_i = (O_i, A_i, I_i) \), a relation \( r \subseteq O_i \times O_j \) and the lattice \( L_j \) of context \( K_j \), the image of \( K_i \) for the existential scaling operator \( sc_\exists \) is:

- \( sc_\exists(K_i) = (O_i, A_i^+, I_i^+) \), where:
  - \( A_i^+ = A_i \cup \{ r : c | c \in L_j \} \)
  - \( I_i^+ = I_i \cup \{(a, r : c) | a \in O_i, c \in L_j, r(a) \cap Ext(c) \neq \emptyset \} \)

For example, assume celestial bodies are scaled along OWxray regarding the lattice in Figure 2. As OWxray(NGC2018)={Chandra} and the telescope Chandra belongs to the extent of concepts \( c_2, c_4, c_5 \) and \( c_6 \), the celestial bodies context is extended by relational attributes of the form \( r:c_i \), where \( i = \{2, 4, 5, 6\} \). Table 5 presents the scaling of both relations OWxray and OWinfrared and their integration to the context of bodies in the form of new relational attributes.

<table>
<thead>
<tr>
<th>( HR5223 )</th>
<th>( SS433 )</th>
<th>( SSS333 )</th>
<th>( NGC2018 )</th>
<th>( OWxray:0 )</th>
<th>( OWxray:2 )</th>
<th>( OWxray:4 )</th>
<th>( OWxray:5 )</th>
<th>( OWxray:6 )</th>
<th>( OWinfrared:0 )</th>
<th>( OWinfrared:2 )</th>
<th>( OWinfrared:5 )</th>
<th>( OWinfrared:6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

**Table 5.** Scaling of bodies context along its relations.

Bodies that are not affected by scaling are not mentioned.

The relational scaling is only one step in the global analysis process which, given an RCF, yields a set of lattices, one per concept, called relational lattice family (RLF). The RLF is defined as the set of lattices whose concepts jointly reflect all the shared attributes and links among objects of the RCF. Its construction is an iterative process since a relational scaling step modifies contexts and thereby the corresponding lattices which in turn may require a new scaling to reflect the newly formed concepts and the link sharing they provoke. Iterations stop whenever a fixed point is reached, i.e., further scaling leaves the lattice set unchanged. In the final lattices, a relational attribute is interpreted as an association between two concepts, the one whose intent it belongs to and the one it refers to explicitly. The analysis of the astronomy RCF using RCA process yields the concept lattices illustrated in Figure 2 and Figure 3. The final lattice of telescopes coincides with the initial one as the corresponding context is not the domain of any RCF relation. In contrast, the final lattice of celestial bodies is different from the initial one due to the relational information inserted into the scaled version of the context. Indeed, the formal objects are assigned relational attributes that lead to additional attribute sharing among objects and hence generate new concepts. By factoring out the new attributes into concept intents, object links are lifted up to the concept level, yielding relations between concepts⁵.

![Figure 3. The final lattice of celestial bodies.](image)

Thus, in Figure 3, previously existing concepts can be seen getting new attributes while completely new concepts emerge. For example, the concept \( c_6 \) represents the celestial bodies M87 and NGC2018 which are recognized as binary stars by the expert since they are observed, located and collimating. In the final lattice it has been assigned the relational attribute OWxray:c5 which basically means that some binary stars are also observable by x-ray telescopes. Furthermore, HR5223, SS433 and PSRA are grouped into the extent of \( c_0 \) in the initial lattice (see Figure 1). However, their respective links with telescope BeppoSAX being revealed through scaling, HR5223 and SS433 form a new concept \( c_9 \) (see Figure 3) which represents the stars that are observable with an infrared telescope such as BeppoSAX.

⁵Observe that for compactness reasons, only non-redundant relational attributes are visualized in concept intents, i.e., the ones referring to the most specific concepts.
5. Ontology construction

5.1. Overall process

We define a four-step RCA-based ontology construction process. First, text analysis extracts terms representing relevant entities and their properties as well as phrases representing inter-entity links from text documents. To that end, linguistic analysis tools are applied [8, 6, 4]. The extracted data are then encoded into a RCF where each sort of entity (resp. type of link) is associated to a formal context (resp. binary relation). Next, the input RCF is transformed into a lattice family by a RCA process. Thereafter, a core ontology is generated from the lattices of the family. Finally, potentially useful abstractions of the transversal relations are targeted with a new round of RCA, this time regarding concepts and relations as first-class objects, i.e., as meta-data individuals. Figure 4 depicts the workflow of the process.

![Figure 4. Ontology building methodology using RCA.](image)

5.2. Text analysis and RCF design

In the field of cognitive semantics, Langacker states that a lexieme refers to a concept in the mind according to previous experiences with the entity or the relation associated with this lexieme [11]. Consequently, nouns and verbs co-occurrence in texts coincide with concepts and semantic relations. In the present study of astronomy paper abstracts (see 2), we focus on verb phrases where verbs are predicates (unary or binary) and entities constitute the respective arguments. To reduce the volume of the acquired data, we define name patterns for celestial bodies and restrict entity extraction to terms defined both in Thesaurus of Astronomy 3 and in the patterns. For instance, Table 1 shows two sample sentences. Terms HR2, NGC3507 and M87 are identified as named entities of celestial body type whereas XMM-Newton is identified as named entity of telescope type. Entity properties are extracted using the Standford shallow parser [6] and assigned to the appropriate entities following the rules:

(i) if the entity \( S \) is the subject (in the active voice) of a verb \( V \), then the property "has the capability to \( V \)" is assigned to \( S \);
(ii) if the entity \( O \) is the object (in the active voice) of a verb \( V \), then the property "being \( V \) -- able" is assigned to \( O \);
(iii) if entity \( S \) is the subject (in the active voice) of a verb \( V \) and \( O \) is its object, then the link (\( O, V \rightarrow O \)) is stated.

In the example of Table 1, the extracted properties are as follows: detectable(HR2), regroupable(HR2), regroup (NGC3507) and regroup(CygnusA). The Stanford shallow parser could be used to extract links as well. However, sentences are generally too complex for the parser to properly extract the main assertion of these sentences. Thus, we have used GATE [4], a general architecture for text engineering, to extract a subset of selected verbs and their arguments.

GATE requires a pool of lexical-grammatical patterns to assist the identification of known domain relations. For example, in astronomy, the observation relation between celestial bodies and telescopes as illustrated by Table 6 can be expressed through a set of regular expressions. GATE looks up the corpus and annotates all bodies and telescopes that match the expressions with tags 'CelestialBody' and 'Telescope', respectively. Thus, the link observed by X-ray is established between the celestial body M87 and the telescope XMM-Newton from the second sentence in Table 1.

![Table 6. GATE extraction rule for observation links.](image)

In order to filter relevant lexical entries that may indicate ontological concepts, the frequencies of the selected terms within the corpus are computed. An additional filtering aims at eliminating irrelevant terms by operating a word sense disambiguation. Thus, to detect mutually similar terms, measures are computed and reported to a domain expert for validation. The similarity measure we currently use explores the atomic term distances within WORDNET. The method, originally devised for the ontology alignment tool OLA, is an extension of the Wu and Palmer measure [20] to compound terms of variable length. Thus, the total dissimilarity between two compound terms combines the pairwise semantic distances within WORDNET of the pairs in an optimal matching between the respective atomic term sets (see [7] for details). For example, the terms active spectroscopic binary and eccentric binary yield a matching \{\{binary, binary\}, \{eccentric, active\}\} whose total dissimilarity is 0.614. The resulting frequent

---


4http://www.iro.umontreal.ca/~owlola/
and disambiguated terms are first submitted to a domain expert for validation, and then compiled into a RCF together with their links while following the mapping rules in Table 7.

<table>
<thead>
<tr>
<th>formal object</th>
<th>subject, verb object</th>
</tr>
</thead>
<tbody>
<tr>
<td>formal attribute</td>
<td>intransitive verb and transitive verb verifying the first two rules</td>
</tr>
<tr>
<td>inter-context relation</td>
<td>transitive verb</td>
</tr>
<tr>
<td>inter-object link</td>
<td>pair (subject, object).</td>
</tr>
</tbody>
</table>

Table 7. RCF elements v.s. syntactic constituents.

5.3. Core ontology modeling

From the viewpoint of domain conceptualization, an ontology can be straightforwardly generated from a RLF. However, there will probably be many irrelevant abstractions in the RLF due to the exhaustive nature of the lattice. The problem can be addressed by a heuristic filtering procedure whose output is validated by a domain expert. Moreover, to make ontology concepts intelligible, the expert assigns a name to each reflecting the underlying description. For instance, celestial bodies that are observed, located, emitting are known to be pulsing variable stars. Thus, the underlying class, called PulsingVariableStar, is a subclass of YoungStar as illustrated in Figure 5. In our example, the pulsing variable star individuals are grouped in concept $c9$ in Figure 3). The $c9$ concept is linked to the telescope concept $c0$ (comprising uniquely BeppoSAX) by the relation OWinfrared.

In [10] we have defined a set of translation rules that map RCA constructs to corresponding ontological components. The target of the translation is a DL base $KB = (TBox, ABox)$, where the TBox hosts conceptual expressions and the ABox the set of ground facts. For instance, RCA elements such as local attributes, objects, inter-context relations, concepts and specialization links are mapped into primitive concepts, instances, primitive roles, defined concepts and subsumption relations in the generated ontology, respectively. Relational attributes are translated into DL role expressions with universal or existential quantification, depending on the scaling schema for inter-context relations. Table 8 depicts some of the mapping from the lattices in Figure 2 and Figure 3 into ontology elements in the DL language $\mathcal{FL}$. In Table 8 is given in Figure 5 where all ontological concepts are given labels.

Observe that concepts M87 and NGC2018 are connected to concepts LightweightTelescope and HeavyTelescope, respectively (see Figure 5). However, the first pair of concepts are both sub-concepts of BinaryStar and the second one of XrayTelescopes. Therefore, one would expect a transversal relation to generalize OWxray observations between binary stars and the x-ray telescopes. The final step of our process targets this sort of abstractions (impossible to detect at the previous step).

5.4. Core ontology refactoring

To motivate this step, observe that the core ontology in Figure 5 states that binary stars and pulsing stars are observed by x-ray and infrared telescopes, respectively. One can further deduce a general observation relation between a specific sort of stars and the telescope type that provides for its observation. The search for such generalized transversal relations can be automated by using the same RCA technique, but applying it differently so that shared specifications between abstractions are factored out. It is noteworthy that RCA has been successfully used to refactor UML conceptual models [5]. Traditionally, the meta-model of the conceptual model is used to guide the design of the RCF. Thus, to encode the ontology obtained so far into a new RCF that is prone to further abstraction, we rely on a popular ontology metamodel. In ODM [14], which is an ontology definition model for RDF(S)/OWL, an ontological concept refers to an OWL class (owl:class). A class may have restrictions (owl:restriction) referring to property (rdfs:property) whose domain may either be a primitive type (datatype), e.g., string or double, or a complex (object)}
type, i.e., a class. A property domain is the class the property is defined on whereas its range is the declared type of its value. Figure 6 shows a UML class diagram representing a simplified view of the ontology meta-model in ODM.

![UML class diagram](image)

**Figure 5.** Part of the ontology classifying celestial bodies.

As to ODM, an ontology can be encoded into two contexts, for ontology concepts and transversal relations, respectively. Relations coming from the grammar, i.e., domain-independent, such as domain and range, connect the contexts. In this way, abstractions can be discovered on each context and then used on the other one, i.e., by scaling upon the transversal relations, to trigger further abstraction discoveries. Furthermore, the attribute name, defined in the ODM metamodel, is assigned to both contexts. In the ontological concept set, where generality links already exist, these are translated into the corresponding context so that they can be preserved in the final ontology. Thus, each concept is assigned not only its own name but also the names of its subsuming concepts. Further formal attributes can be added to the contexts, e.g., symmetric, transitive, etc., for relations, enabling a better factorization of common features. Such translations is even more attractive if the previous step is tuned to infer not only property definitions but also property restrictions (cardinality, quantifier, etc.). For instance, Figure 7 illustrates the RCF encoding the core ontology depicted in Figure 5.

Running RCA process produces a first lattice (Figure 8) factorizing shared specifications among ontology concepts and a second lattice (Figure 10) representing transversal relation abstractions that can be turned into relation hierarchy. Figure 10 suggests that relations oWxray and oWinfrared can be abstracted by new relation, labeled observesWith, whose domain is formal concept c4 (Star in Figure 8) and range is formal concept c3 (Telescope in Figure 8). The translation of the lattices in Figure 8 and Figure 10 following the previously described mapping rules yields an ontology that is partially depicted in Figure 9. Notice that oWxray and oWinfrared have been abstracted by observesWith in this new ontology.
6. Preliminary Experimental Results

In order to evaluate the proposed RCA framework for ontology construction, we conducted a preliminary experimental study using the public domain tools for text analysis and an implementation of the RCA framework within the GALICIA platform.

The study used a corpus sample of 830 abstracts of papers taken from the astronomy journal A&A. The text analysis has yielded 67 extracts pertaining to observation, e.g., 'XMM-Newton X-ray observatory performed a pointed observation of the bursting pulsar GRO J1744-28' and 'Using the IRAM 30m telescope we have observed a dense core in the cirrus cloud MCLD 123.5+24.9'. The entity extraction process has identified 10 different telescopes and 60 different celestial bodies.

A prototype ontology was presented to astronomers who were asked to evaluate the plausibility and interestingness of the proposed classes of celestial bodies. A large group of classes were found to correspond to known categories of celestial bodies hence they brought little new knowledge yet witnessed the robustness of our method. Another subset of classes were assessed as implausible since corresponding to meaningless combinations of attributes (due to sharing of less significant features). Finally, a small residual of classes seemed to be both plausible and surprising for astronomers as they presented combinations of attributes that although typical have not been focused on previously.

Moreover, whereas links to telescope classes provided valuable information about the observability of celestial bodies, these do not seem to have generated meaningful classes on their own. A possible reason for that is the limited number of telescopes that adversely impacts the number of purposeful telescope classes. This outcome suggests a different modeling of the relational structure of the astronomy domain, by possibly including relations between celestial bodies (neighborhood links, satellite links, etc.).

Recently, we have been applying the proposed approach to further areas where the relational is more explicit and rich such as biology/biochemistry and pharmacovigilence. In the latter case, the goal is to classify patients, drugs and adverse drug reactions according to their own features and connections. As to the tool, a slightly more recent version thereof implements pruning functions that discard spurious concepts from the lattices. The underlying mechanism assesses the relevance degree of each concept using criteria such as the presence of key domain information in the corresponding intent, e.g., the wavelength of a telescope as opposed to its mass.

7. Related work

A comprehensive survey of state-of-the-art in ontology learning from text was proposed by Buitlaar et al. in [2]. We focus here on approaches relying on FCA [3, 17, 13] and those automatically deriving transversal relations from texts [12, 19, 15, 18]. Cimiano et al. [3] propose to extract terms referring to relevant domain concepts as individuals and to explore syntactic dependencies between the verbs
and their grammatical arguments in deriving term properties. Both sets then form a formal context which yields a concept hierarchy. Regarding this study, our RCA-centered framework rely on similar text analysis techniques while covering relations extraction. A different approach was proposed by Stumme et al. [17] who explored ontology construction by merging two existing ontologies provided with a corpus of textual documents. NLP techniques are used to capture the relationships between documents and concepts from an ontology and organize them into a dedicated context. The two contexts are then merged and a pruned concept lattice is constructed which is further transformed into a merged ontology by a human expert. In [13], Nanda et al. introduced a FCA-based methodology for the design of formal ontologies for product families. Terms describing the components in a product family along with the required properties are captured in a lexicon set and put into context. A concept hierarchy is then derived from the concept lattice and exported into OWL. Unlike our own approach, the entire process relies heavily on human involvement while proposing no specific assistance for the extraction of transversal relations.

Existing approaches for transversal relation acquisition from text either rely on distributional properties of words [12, 18], or apply extraction pattern [15]. In [12], Maedche et al. present the TEXTTOONTO tool that acquires non-taxonomic relations from text. The approach assumes a concept taxonomy which is to be decorated with transversal relations. Terms and linguistic dependency relations between terms are extracted using a shallow text processor. Terms are then attached to taxonomy concepts and, based on term co-occurrence, associations between pairs of taxonomy concepts are mined using a generalized association rule model. Finally, each strongly associated concept pair is interpreted as the endpoints of a new transversal relation. Wang et al. [18] have investigated support vector machines classification as an approach for relation extraction from text. Various components of the GATE infrastructure contribute to the derivation of classification features (e.g., syntactic parse tree, part-of-speech tags, etc.) whereas background knowledge is fed into the process in form of concept (entity) and relation type hierarchies and WORDNET is used as additional semantic resource. Ruiz-Casado et al. [15] have presented a method for completing WORDNET with new relationships extracted from Wikipedia articles. To that end, they study Wikipedia entry definitions and, on the one hand, match them with synsets from WORDNET while on the other hand, record occurrences of other Wikipedia entries in them. A set of general patterns based on part-of-speech tags is extracted while existing relations between entries within WORDNET are used to assign each of them to a general semantic relation (hyperonymy, hyponymy, meronymy, etc.). Finally, the patterns are used in inferring new relations between previously unrelated WORDNET synsets. Compared to concurrent methods, though using similar NLP tools, our approach presents the important advantage of a systematic construction of both the hierarchy of concepts and of transversal relations according to a single computational model, the RCA.
8. Discussion

The presented method for learning ontologies from textual documents relies on an extension of classical FCA to integrate the processing of domain concepts and transversal relations into a single step. In that it outperforms concurrent approaches that typically use FCA to merely construct the concept hierarchies. We appraise that two factors could jeopardize the success of the proposed approach and hence need to be addressed in the near future. First, the ontology derived directly from a concept lattices is usually large. We hypothesize that the appropriate answer to that challenge is a combination of automated pruning and refactoring to help obtain focused ontology. Next, the RCA input data has a strong impact on the quality of the resulting ontology. This fact motivates the strengthening of text analysis components in our approach by introducing alternate syntactic patterns in the extraction of individuals and their properties.

References