Performance-friendly rule extraction in large water data-sets with AOC posets and relational concept analysis

Xavier Dolques, Florence Le Ber, Marianne Huchard, Corinne Grac

To appear in the *International Journal of General Systems*
Vol. 00, No. 00, Month 20XX, 1–22

Performance-friendly rules extraction in large water datasets with AOC-posets and Relational Concept Analysis

Xavier Dolques [a*], Florence Le Ber[a],

Marianne Huchard[b], and Corinne Grac[c]

[a] *ICube, Université de Strasbourg/ENGEES, CNRS, Strasbourg*;
[b] *LIRMM, Université de Montpellier, CNRS, Montpellier*;
[c] *LIVE, Université de Strasbourg/ENGEES, CNRS, Strasbourg*

(*AUTHOR VERSION / URLs and references updated 2023/09/24*)

In this paper, we consider data analysis methods for knowledge extraction from large water datasets. More specifically, we try to connect physico-chemical parameters and the characteristics of taxons living in sample sites. Among these data analysis methods we consider Formal Concept Analysis (FCA), which is a recognized tool for classification and rule discovery on object-attribute data. Relational Concept Analysis (RCA) relies on FCA and deals with sets of object-attribute data provided with relations. RCA produces more informative results but at the expense of an increase in complexity. Besides, in numerous applications of FCA, the partially ordered set of concepts introducing attributes or objects (AOC-poset, for Attribute-Object-Concept poset) is used rather than the concept lattice in order to reduce combinatorial problems. AOC-posets are much smaller and easier to compute than concept lattices and still contain the information needed to rebuild the initial data. This paper introduces a variant of the RCA process based on AOC-posets rather than concept lattices. This approach is compared with RCA based on iceberg lattices. Experiments are performed with various scaling operators, and a specific operator is introduced to deal with noisy data. We show that using AOC-poset on water datasets provides a reasonable concept number and allows us to extract meaningful implication rules (association rules which confidence is 1), whose semantics depends on the chosen scaling operator.

Keywords: Formal Concept Analysis, Relational Concept Analysis, implication rules, water dataset

1. Introduction

Relational Concept Analysis (RCA) is a method for exploring relational data which has already proved its relevance in several applications (Hacène et al. 2013a). It is based on the iterative use of the classical Formal Concept Analysis (FCA) algorithm (Huchard et al. 2007): formal objects are described with formal attributes as in FCA, and with their relationships with other formal objects. The resulting relational concepts can be used to extract implication rules linking objects or attributes from different contexts. However, since RCA is an iterative process and each concept created from an iteration step can be used to generate concepts at the following step, it often comes with a combinatorial explosion, and relevant rules are difficult

*Corresponding author. Email: xavier.dolques@engees.unistra.fr

To appear in the *International Journal of General Systems*
Vol. 00, No. 00, Month 20XX, 2–22

to extract from the huge set of built concepts. Various strategies can be used to cope with this complexity, including separating the initial formal object sets into smallest ones after a first analysis, or introducing queries (Azmeh et al. 2011b). A promising strategy consists in using AOC-posets, or Galois sub-hierarchies, i.e. the sub-posets of concept lattices induced by concepts introducing objects or attributes. AOC-posets are smaller and easier to compute than concept lattices (Godin and Mili 1993), and their sets of concepts have interesting properties for extracting implication rules. In a previous work (Dolques, Le Ber, and Huchard 2013), we have adapted the RCA process into an RCA-AOC process which uses AOC-posets rather than full concept lattices at each step of the process. Furthermore we have presented preliminary results obtained on an application concerning the assessment of watercourses quality.

In this paper, we further refine this preliminary work. We show that RCA-AOC can handle large relational datasets and can easily extract relevant rules (implication rules, i.e. association rules which confidence is 1) with a premise composed of one attribute, in the case of a concrete application. We also introduce a general scaling operator (the percent scaling operator) and show that it can be used to represent fuzzy information better suited to our data and corresponding to the expert expectations. We explore and discuss the effect of various scaling operators (existential, universal strict and percent scaling operators) on the number and interest level of rules obtained from the RCA-AOC process. Results are given from the analysis of the FRESQUEAU dataset[1] about watercourses.

The paper is organized as follows. A state of the art is presented in Section 2. Section 3 gives some useful definitions about FCA and AOC-posets. Section 4 details the RCA process and its variant based on AOC-posets. Section 5 presents the dataset and the principles of our analysis. The choice of RCA-AOC to perform this analysis relies on a comparison with RCA based on iceberg lattices. Numerical results obtained with various scaling operators, and several rules are presented and discussed in Section 6. Section 7 concludes the paper by opening some perspectives of this work.

2. State of the art

In this paper, we aim at presenting to biological scientists implication rules extracted by exploring large relational datasets. Implication rules extraction is closely connected to Galois lattices and Formal Concept Analysis (Bertet, Guillas, and Ogier 2007). This brings us to consider relational dataset exploration in the context of FCA (Section 2.1), with techniques for dealing with the combinatorial dimension of the problem (Section 2.2) and implication rules (Section 2.3).

2.1. *Integrating relations in FCA*

Several approaches to integrate relations in FCA have been proposed, including *power context family* (Prediger and Wille 1999), Relational Boolean Factor Analysis (Krmelova and Trnecka 2013) and approaches connected to logics (Ferré, Ridoux, and Sigonneau 2005; Baader and Distel 2008).

Prediger and Wille (1999) deal with many-valued contexts that are transformed into a family of formal contexts under the guidance of a user objective. This family

[1] `http://dataqual.engees.unistra.fr/fresqueau_presentation_gb`

To appear in the *International Journal of General Systems*
Vol. 00, No. 00, Month 20XX, 3–22

of formal contexts is called the power context family, a notion that has been introduced by Wille (1997). It represents all the k-ary relations on the object set. From the concept lattices built on the formal contexts of the power context family, concept graphs are extracted which, in turn, are organized into a lattice. Wolff (2009) proposed another approach for obtaining concept graphs that relies on temporal concept analysis, where conceptual scales are used instead of the concept lattices of the k-ary relations. Kötters (2013) also considers objects connected by relations. They introduce a Galois connection (and the derived concept lattice) which associates a table (variables and the corresponding tuples) to a description that takes the form of a *windowed s-structure*. Such a windowed s-structure (designed to be a form of a query) is roughly a graph with edges labelled by the relations and with some nodes labelled by variables.

Ferré, Ridoux, and Sigonneau (2005) transform relational data into logical formulae within the framework of logical concept analysis. Object contexts are combined with relational contexts and equipped with a combined logic. Relational attributes are defined as follows: $(\exists r.f)(x) =_{def} \exists x'.(r(x,x') \wedge f(x))$. The concepts' intents of the resulting lattice contain either classical attributes (f) or relational attributes $(\exists r.f)$. Meta-relations are also built for navigating from one concept to another. Baader and Distel (2008) propose a method for computing a basis of general concept inclusions in Description Logics $\mathcal{EL}_{gfp}$ where cyclic concept definition has close connections with RCA.

Boolean Factor Analysis is applied to multi-relational data by Krmelova and Trnecka (2013, 2014). The relational factors are tuples of boolean factors extracted from the various data tables in an independent way. Several connection schemas can be applied that act similarly to the scaling operators of RCA. Compared to RCA which builds initially upon all formal concepts that can be extracted from the object-attribute tables and iterates, the boolean factors are only a part of the formal concepts.

Hacène et al. (2013a) propose RCA whose originality is to compute in an iterative manner (with a possible stop at each step) several concept lattices from data represented in relational format. The concept lattices are connected by links that abstract the relations between objects (the "relational attributes"). Several operators, borrowed to Description Logics, build the links between concepts. In RCA, the scaling operation allows us to consider different kinds of granularity for the relations between the concepts.

The initial RCA framework, using whole concept lattices, has been used for the analysis and modernization of UML models (Dao et al. 2004; Arévalo et al. 2006; Dolques et al. 2012), namely in class diagrams and in use case diagrams. Moha et al. (2008) use concept lattices to exploit relations, between methods and between methods and attributes, to detect and fix design defects. Model transformations are learned from transformation examples thanks to several kinds of relations between model elements (e.g. between elements inside a model, transformation links between source elements and target elements) by Saada et al. (2012). Azmeh et al. (2011a) use the relations between abstract tasks in an abstract orchestration to classify relevant Web services to instantiate the tasks. Other applications can be found in ontology engineering (Bendaoud, Napoli, and Toussaint 2008; Rouane-Hacène, Valtchev, and Nkambou 2011; Shi et al. 2011). In these applications, the datasets are medium-size guarantying the feasibility of the approach. In the FRESQUEAU project, we have larger sets of data and it is necessary to reduce the lattice size.

To appear in the *International Journal of General Systems*
Vol. 00, No. 00, Month 20XX, 4–22

2.2. *Reducing lattice size*

The size of a concept lattice can be exponentially bigger than the size of its context, the size of a concept lattice being bounded by $2^{min(|\mathcal{G}|,|\mathcal{M}|)}$. Iceberg lattices and AOC-posets are two ways of reducing the number of concepts, which benefits and drawbacks will be presented in more details in Section 5.

Iceberg lattices have been introduced by Stumme et al. (2002) for their use in Knowlegdge Discovery databases. An iceberg lattice is induced by the subset of "frequent" concepts, i.e., concepts which have an extent support greater than a given threshold. It is described by the authors as a visualization method for large databases, a condensed representation for frequent itemsets and a visualization tool for association rules.

To our best knowledge, the AOC-posets have been firstly used by Godin and Mili (1993) in the domain of software engineering (object-oriented programming). AOC-posets have also been used in applications of FCA to non-monotonic reasoning and domain theory (Hitzler 2004) and to produce classifications from linguistic data (Osswald and Pedersen 2002; Petersen 2001), because of their capability to structure knowledge. Osswald and Petersen (2003) argue that AOC-posets are more compact than concept lattices, and that the price of recovering missing rules is not too high. Aboud et al. (2014) compare efficiency of concept lattices and AOC-posets for classifying components in subtyping-based directories. The conclusion is that in real-case studies (dynamic environments), AOC-posets provide more realistic classification structures for storage and execution time.

Specific parts of the AOC-poset (mainly the attribute-concepts part, namely AC-poset) are used in several works. AC-posets help optimizing class inheritance hierarchy in object-oriented programming (Godin et al. 1998; Huchard, Dicky, and Leblanc 2000). In this context, the concepts are interpreted as classes and the concept specialization is interpreted as class inheritance. AC-posets are more appropriate than concept lattices because in the general case, there is little interest to create classes empty of attributes and methods. In software product line applications of FCA (Ryssel, Ploennigs, and Kabitzsch 2011; Al-Msie'deen et al. 2013), the concepts are extracted from the formal context mapping products to features. Ryssel, Ploennigs, and Kabitzsch (2011) use the attribute concepts for inferring the 1-1 implication rules, as well as a part of feature diagrams and, with more computation, an implication base. Al-Msie'deen et al. (2013) consider concepts of the AOC-poset as blocks of code entities that are variation points in a software system.

2.3. *Extracting rules*

Many approaches exist to extract logical rules from data either in a supervised context for building decision tree (Quinlan 1986) or classification rules (Clark and Boswell 1991), or in an unsupervised context for association rules learning (Agrawal, Imieliński, and Swami 1993). In our context, we are interested in proposing to biological scientists sets of implication rules (association rules of confidence 1). Implication rules have been extensively studied, and defining implicational bases is studied in Bertet and Monjardet (2010). Other work of Adaricheva, Nation, and Rand (2011) identifies in the basis of a reduced closure system the *binary part* composed of the rules such that both the premise and the conclusion are singletons.

Here implication rules are built upon the relational dataset. Their extraction is done by considering a main lattice and the relational attributes that connect this lattice to the others. For extracting a significant subset of the implication rules, AOC-posets are relevant because they contain all the irreducible elements and more

To appear in the *International Journal of General Systems*
Vol. 00, No. 00, Month 20XX, 5–22

precisely all the concepts that introduce attributes. Besides, the poset may be used to explore rules that are in the neighborhood of a point of interest. For a reduced context [2], the concepts of the AOC-poset are exactly the irreducible elements. In this case, for implication rules where premise is a single attribute, we can consider only the meet-irreducible elements[3] (attribute-concepts). In our case, the formal context is not reduced and we use all attribute-concepts (which include the meet-irreducible elements) for building implication rules with premise containing one single attribute a and conclusion containing a set of attributes S which is roughly the intent of the concept, with some attribute removal as it is explained in Section 6. These rules can be easily transformed into binary ones by writing all rules $a \rightarrow s$ with $s \in S$.

The relational aspect of the approach permits the extraction of rules from complex relational data which would require to be transformed by propositionalization approaches (Lachiche 2010) for many learning approaches. It also could be compared to Inductive Logic Programming (Muggleton and de Raedt 1994) in some ways as it will result in first order logic formulas. ILP being a supervised approach, its goal differs as it will try to find the right premise for a given conclusion. The expressivity of the results of our approach is far more restricted than with ILP, even with all the scaling operators that can be imagined, leading to better performances but with a restricted output language.

3. From FCA Basics to AOC-posets

Formal Concept Analysis (Ganter and Wille 1999) aims at extracting an ordered set of concepts from a dataset, called a Formal Context, composed of objects described by attributes. A formal context $\mathcal{K}$ is a 3-tuple $(\mathcal{G}, \mathcal{M}, \mathcal{I})$, where $\mathcal{I} \subseteq \mathcal{G} \times \mathcal{M}$.

Table 1. Formal Context $\mathcal{K}_{Taxons}$ of animals (taxons) described by their life traits

Animal	MH1	MH2	MH3	MH4	MH5	MH8	MH9
Aeschnidae						×	
Agabus					×	×	×
Agraylea					×		
Agriotypus	×	×	×				
Ancylus	×	×					
Anisus	×				×		×
Anodonta			×	×			×
Anthomyiidae	×				×	×	

Table 1 is a Formal Context $\mathcal{K}_{Taxons} = (\mathcal{G}_{Taxons}, \mathcal{M}_{Taxons}, \mathcal{I}_{Taxons})$ which describes taxons, i.e., animals or plants, by characteristics they may own. The considered taxons here are macro-invertebrates that can be found in rivers. We took the examples of the following kinds of animals : *Aeschnidae* (`Aes.`), *Agabus* (`Agb.`), *Agraylea* (`Aga.`), *Agriotypus* (`Agi.`), *Ancylus* (`Anc.`), *Anisus* (`Ani.`), *Anodonta* (`Ano.`), *Anthomyiidae* (`Ant.`). They are described by their microhabitats: MH1

[2]Which has no identical rows, no identical columns, no row which is the intersection of several other rows, and no column which is the intersection of several other columns (Ganter and Wille 1999).
[3]Elements of the lattice with a unique successor, while considering ascending order in our diagram representations: lowest elements are below greatest elements.

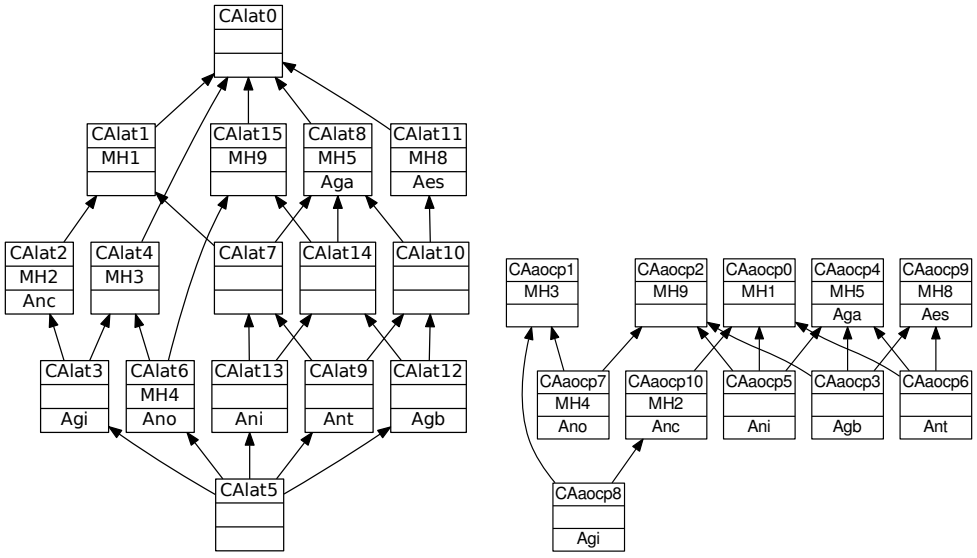
To appear in the *International Journal of General Systems*
Vol. 00, No. 00, Month 20XX, 6–22

(flags/boulders/cobbles/pebbles), MH2 (gravel), MH3 (sand), MH4 (silt), MH5 (macrophytes), MH8 (organic detritus/litter), MH9 (mud).

Given a $\mathcal{K} = (\mathcal{G}, \mathcal{M}, \mathcal{I})$ formal context, a formal concept associates a maximal set of objects with the maximal set of attributes they share. It is thus a $C = (Extent(C), Intent(C))$ pair where:

- $Extent(C) = \{g \in \mathcal{G} | \forall m \in Intent(C), (g, m) \in \mathcal{I}\}$ is the extent of the concept (objects covered by the concepts),
- $Intent(C) = \{m \in \mathcal{M} | \forall g \in Extent(C), (g, m) \in \mathcal{I}\}$ is the intent of the concept (shared attributes).

Given two formal concepts $C_1 = (E_1, I_1)$ and $C_2 = (E_2, I_2)$ of $\mathcal{K}$, the concept specialization order $\leq_s$ is defined by $C_1 \leq_s C_2$ if and only if $E_1 \subseteq E_2$ (and equivalently $I_2 \subseteq I_1$).



(a) Concept lattice of taxons $\mathcal{L}^0_{Taxons}$ (b) AOC poset AOC_{Taxons}

Figure 1. Concept lattice vs. AOC poset for $\mathcal{K}_{Taxons}$ formal context

Let $\mathcal{C}_\mathcal{K}$ be the set of all concepts of a $\mathcal{K}$ formal context. This set of concepts provided with the specialization order $(\mathcal{C}_\mathcal{K}, \leq_s)$ has a lattice structure, and is called the concept lattice associated with $\mathcal{K}$.

Fig. 1(a) shows the concept lattice associated with the formal context of Table 1. For simplicity's sake, the lattice representation shows attributes (*resp.* objects) only in the concept where they are introduced. An attribute is introduced in the highest concept (for $\leq_s$) where it appears, and it is (top-down) inherited in the subconcepts. For example, in Fig. 1(a), `MH2` attribute is introduced in `CAlat2` concept. It is inherited by `CAlat2` subconcepts: `CAlat3` and `CAlat5`. Symmetrically, an object is introduced in the lowest concept (for $\leq_s$) where it appears, and it is (bottom-up) inherited in the super-concepts. For example, in Fig. 1(a), `Anc.` object is introduced in `CAlat2` concept. It is inherited by `CAlat2` super-concepts: `CAlat0` and `CAlat1`. In other words, the representation shows the simplified intents (intents restricted to introduced attributes) and simplified extents (extents restricted to introduced objects) which will be denoted respectively $Intent_S(C)$ and $Extent_S(C)$ for a given concept C.

Given the potential complexity of the lattice computing in time and space, as the

To appear in the *International Journal of General Systems*
Vol. 00, No. 00, Month 20XX, 7–22

size of the lattice can rise up to $2^{min(|\mathcal{G}|,|\mathcal{M}|)}$ concepts, several FCA applications (see Section 2) use only a sub-order of $(\mathcal{C}_\mathcal{K}, \leq_s)$ built only from the *object concepts* (which introduce at least one object) or the *attribute concepts* (which introduce at least one attribute). In Fig. 1(a), `CAlat13` and `CAlat2` are examples of object concepts; `CAlat1` and `CAlat2` are examples of attribute concepts; `CAlat7`, `CAlat14` and `CAlat10` are examples of concepts that do not introduce any object or any attribute.

The specific sub-order of the concept lattice restricted to object concepts and attribute concepts is called an *AOC-poset* (for Attribute-Object-Concept poset) or sometimes Galois Sub-Hierarchy, but we consider the latter term less explicit. Fig. 1(b) shows the AOC-poset for the context of Table 1. The size of an AOC-poset can be significantly smaller than the size of the lattice built from the same context, the number of concepts of an AOC-poset being bounded by $|\mathcal{G}| + |\mathcal{M}|$. On our small example, the AOC-poset has five concepts less than the corresponding lattice. However, an AOC-poset is sufficient to recover the entire context. In the following sections, AOC-posets will be part of a novel approach using their properties to reduce the set of concepts generated to a more reasonable number.

4. RCA-AOC: A variant of Relational Concept Analysis with AOC-poset

This section describes the classical RCA approach and the proposed variant of the process for a new approach.

4.1. *Relational Concept Analysis: the lattice-based approach*

In the RCA input dataset, objects of several categories are described by attributes and by relations to objects. This kind of dataset is called a Relational Context Family. For more details about RCA, the reader is invited to read Hacène et al. (2013b) where we borrow notations that refine these of Hacène et al. (2013a).

Definition 1 (Relational Context Family (RCF))**.** *A Relational Context Family (denoted RCF) is a* $(\mathbf{K}, \mathbf{R})$ *pair where:*

- $\mathbf{K} = \{\mathcal{K}_i\}_{i=1,\ldots,n}$ *is a set of* $\mathcal{K}_i = (G_i, M_i, I_i)$ *formal contexts (object-attribute relations), where* G_i *is the set of objects,* M_i *is the set of attributes and* $I_i \subseteq G_i \times M_i$*.*
- $\mathbf{R} = \{r_j\}_{j=1,\ldots,p}$ *is a set of* r_j *object-object relations where* $r_j \subseteq G_k \times G_l$ *for some* $k, l \in \{1, \ldots, n\}$*.*

In the following we will be using the notation $dom(r_j)$ to denote the domain of the relation r_j and $ran(r_j)$ to denote the range of r_j.

To illustrate RCA, we use an RCF composed of the `Taxons` context denoted $\mathcal{K}_{Taxons}$ (Table 1), a `Sites` context, denoted $\mathcal{K}_{Sites}$ (Table at the left-hand side of Fig. 2), and a `contains` relation, denoted by $r_{contains}$ (Table 2).

RCA integrates object-object relations as new attributes (called *relational attributes*) in formal contexts. Object-object relations link source objects to target objects. They are different from object-attribute relations by the fact that target objects are classified in concepts which are used to produce relational attributes. Using simple attributes, the FCA approach is able to make a one-step concept discovery, where a concept such as "sites containing `Agb` individuals" will emerge. The RCA objective is to compose several relations. In our simple case, RCA will allow us to recognize the group of sites (`site3` and `site5`) which contain animals owning

To appear in the *International Journal of General Systems*
Vol. 00, No. 00, Month 20XX, 8–22

Site	NH4	SO4
site0	×	×
site1		×
site2		×
site3	×	
site4	×	
site5	×	

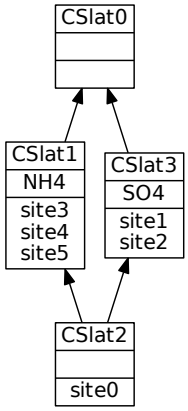


Figure 2. Formal Context of Sites $\mathcal{K}_{Sites}$ (left), Concept lattice of Sites (right, $\mathcal{L}^0_{Sites}$)

Table 2. Relation $r_{contains}$

contains ↗	Aes	Agb	Aga	Agi	Anc	Ani	Ano	Ant
site0	x		x					
site1				x	x		x	
site2				x	x			
site3		x						
site4						x		
site5								x

`MH5` and `MH8` attributes, that live on macrophytes or on organic detritus.

The principle followed by RCA consists of transforming an object-object relation r into a relation between objects of one category (the domain of r), and *relational attributes* that involve concepts formed on objects of the other category (the range of r) using *scaling operators*. These relational attributes have the general syntactic form $q\ r(C)$, where q is a *quantifier*, r is the relation and C is a concept whose extent contains objects of the category which is the range of r. Quantifiers are chosen within a set $\mathbf{Q}$ which includes the $\exists$ quantifier, which is the most used up to now and which is used in this paper to illustrate the approach. Two other quantifiers are defined below. In the next definitions, for $r \subseteq G_1 \times G_2$ a relation, and $o_1 \in G_1$, we denote the image set of o_1 by $r(o_1) = \{o_2 \in G_2 | (o_1, o_2) \in r\}$.

Definition 2 (Existential scaling)**.** *Let $\mathcal{K} = (G, M, I)$ and $\mathcal{K}_r = (G_r, M_r, I_r)$ be two contexts, and r a relation, where G is the domain of r, and G_r is the range of r. Let $\mathcal{C}_r$ be the concept set built on on $\mathcal{K}_r$. For every object $o \in G$ and every concept $C_r \in \mathcal{C}_r$, if $r(o)$ has a non-empty intersection with Extent(C_r) then the relational attribute $\exists r(C_r)$ is added to the attributes of o. This operation is called existential scaling on $\mathcal{K}$, $\mathcal{C}_r$ and r.*

The existential scaling applied to the $\mathcal{K}_{Sites}$ context (right-hand side of Fig. 2), for the $r_{contains}$ relation (Table 2) and the concept set of animal lattice (Fig. 1(a)) is presented in Table 3 after the vertical triple bar (the two columns that precede the triple bar correspond to the initial site context). In this table, relational attributes (with $\exists$ scaling) encode relations between sites and groups of taxons. For example, `site0` owns relational attribute $\exists$cont(`CAlat11`), expressing that `site0` contains one of the taxons (`Aes`) grouped in concept `CAlat11`.

There are several scaling operations for dealing with relational contexts (Hacène

To appear in the *International Journal of General Systems*
Vol. 00, No. 00, Month 20XX, 9–22

Table 3. Existential Scaling of Formal Context of sites. `cont` is the abbreviation of `contains`

Site	NH4	SO4	∃cont(CAlat0)	∃cont(CAlat1)	∃cont(CAlat2)	∃cont(CAlat3)	∃cont(CAlat4)	∃cont(CAlat5)	∃cont(CAlat6)	∃cont(CAlat7)	∃cont(CAlat8)	∃cont(CAlat9)	∃cont(CAlat10)	∃cont(CAlat11)	∃cont(CAlat12)	∃cont(CAlat13)	∃cont(CAlat14)	∃cont(CAlat15)
site0	×	×	×								×			×				
site1		×	×	×	×	×	×		×									×
site2		×	×	×	×	×	×											
site3	×		×								×		×	×	×		×	×
site4	×		×	×						×	×					×	×	×
site5	×		×	×						×	×	×	×	×				

et al. 2013a). We define here the universal strict scaling operator $\forall\exists$ and a general percent operator $S_{>n\%}$, an operator we newly introduce to handle noisy results. This last scaling comes in various levels, according to application needs.

Definition 3 (Universal strict Scaling)**.** *Let $\mathcal{K} = (G, M, I)$ and $\mathcal{K}_r = (G_r, M_r, I_r)$ be two contexts, and r a relation, where G is the domain of r, and G_r is the range of r. Let $\mathcal{C}_r$ be the concept set built on on $\mathcal{K}_r$. For every object $o \in G$ and every concept $C_r \in \mathcal{C}_r$, if $r(o)$ is included in Extent(C), then the relational attribute $\forall\exists r(C)$ is added to the attributes of o.*

Definition 4 (Percent Scaling)**.** *Let $\mathcal{K} = (G, M, I)$ and $\mathcal{K}_r = (G_r, M_r, I_r)$ be two contexts, and r a relation, where G is the domain of r, and G_r is the range of r. Let $\mathcal{C}_r$ be the concept set built on on $\mathcal{K}_r$. Let n be a value, $n \in [1, 100]$. For every object $o \in G$ and every concept $C_r \in \mathcal{C}_r$, if more than n percent of $r(o)$ is included in Extent(C) (i.e., $|r(o) \cap Extent(C)| > n|r(o)|/100$), then the relational attribute $S_{>n\%}r(C)$ is added to the attributes of o.*

For defining more precisely the scaling operators, a generic function κ is introduced which maps a scaling operator, a relation r and an object subset in the range of r to a subset from the domain of r.

$$\kappa : \mathbf{Q} \times \mathbf{R} \times \bigcup_{i=1,\ldots,n} 2^{G_i} \rightarrow \bigcup_{i=1,\ldots,n} 2^{G_i}$$

In the case of the existential scaling operator, for a relation $r \subseteq G_k \times G_l$, the function κ is instantiated as:

$$\begin{aligned} \kappa_\exists : \mathbf{R} \times 2^{G_l} &\rightarrow 2^{G_k} \\ (r, S_l) &\rightarrow \{o | r(o) \cap S_l \neq \emptyset\} \end{aligned}$$

The generic notion of scaling operator can now be defined as follows.

Definition 5 (Scaling operator)**.** *Let $\mathcal{K}_k = (G_k, M_k, I_k)$ and $\mathcal{K}_l = (G_l, M_l, I_l)$ be two contexts, $r \subseteq G_k \times G_l$ a relation, and $\mathcal{C}_l$ a concept set on $\mathcal{K}_l$; q denotes a scaling quantifier. The scaling operator $\mathrm{S}_{(r,q),\mathcal{C}_l}$ over $\mathcal{K}_k$ yields the derived context $(G^+, M^+, I^+) = \mathrm{S}_{(r,q),\mathcal{C}_l}(\mathcal{K}_k)$, where:*

- $G^+ = G_k$,
- $M^+ = \{'q\ r(c)' \mid c \in \mathcal{C}_l\}$,
- $I^+ = \bigcup_{c \in \mathcal{C}_l} \kappa(q, r, Extent(c)) \times \{'q\ r(c)'\}$.

Definition 6 (Relational extension of a context $\mathcal{K}_k = (G_k, M_k, I_k)$)**.** *Under Def. 5,*

To appear in the *International Journal of General Systems*
Vol. 00, No. 00, Month 20XX, 10–22

let $R_k = \{r_j, 1 \leq j \leq p_k | dom(r_j) = G_k\}$ be the set of relations with domain G_k. Let ρ be a mapping from $\mathbf{R}$ to $\mathbf{Q}$ which associates a scaling operator to each object-object relation r_j. Let a set of concept lattices $\mathbf{L}$ where each lattice $\mathcal{L}_{G_l}$ corresponds to a context $\mathcal{K}_l = (G_l, M_l, I_l)$ from $\mathbf{K}$ with $1 \leq l \leq p_k$. Each G_l is the range of a relation $r_j \in R_k$. The relational extension of the $\mathcal{K}_k$ context, consists of apposing to $\mathcal{K}_k$ the respective scaling upon each $r_j \in R_k$:

$$\mathrm{E}_{\rho,\mathbf{L}}(\mathcal{K}_k) = \mathcal{K}_k \mid \mathrm{S}_{(r_1,\rho(r_1)),\mathcal{L}_{ran(r_1)}}(\mathcal{K}_k) \mid \ldots \mid \mathrm{S}_{(r_{p_k},\rho(r_{p_k})),\mathcal{L}_{ran(r_{p_k})}}(\mathcal{K}_k)$$

Table 3 shows the relational extension of $\mathcal{K}_{Sites}$, when considering $\rho(r_{contains}) = \exists$ and the concept set of animal lattice of Fig. 1(a). If an additional relation connecting sites to another kind of objects, for example, $r_{inHEregion}$, connecting sites to hydro-ecoregions (e.g. Alsace plain, Parisian Basin, ...) were in the dataset, then the relational extension of $\mathcal{K}_{Sites}$ would include the scaling upon $r_{inHEregion}$ too.

The relational extension of the whole $\mathbf{K}$ is composed of all the relational extensions of all $\mathcal{K}_i$ in $\mathbf{K}$.

Definition 7 (Relational extension of an RCF)**.** *Under the previous definitions, the relational extension of $\mathbf{K}$ is:*

$$\mathrm{E}^*_{\rho,\mathbf{L}}(\mathbf{K}) = \{\mathrm{E}_{\rho,\mathbf{L}}(\mathcal{K}_1), \ldots, \mathrm{E}_{\rho,\mathbf{L}}(\mathcal{K}_n)\}$$

In our example, if we consider only the existential scaling and the animal and site lattices of Fig. 1(a) and right-hand side of Fig. 2, the relational extension of $\mathbf{K}$ would be composed of the relational extensions of $\mathcal{K}_{Taxons}$ and $\mathcal{K}_{Sites}$. The relational extension of $\mathcal{K}_{Taxons}$ is simply $\mathcal{K}_{Taxons}$, because there is no outgoing relation. The relational extension of $\mathcal{K}_{Sites}$ has been shown in Table 3.

Now a whole construction process consists in building a finite sequence of contexts and concept lattices associated with $(\mathbf{K}, \mathbf{R})$ and ρ. The first set of contexts (step 0) is $\mathbf{K}^0 = \mathbf{K}$. The contexts of step p are used to build the associated concept lattices. The $\mathbf{L}_p$ set composed of the lattices at step p is used to calculate the relational extension. The set of contexts at step $p+1$ is defined using the relational extension: $\mathbf{K}^{p+1} = \mathrm{E}^*_{\rho,\mathbf{L}_p}(\mathbf{K})$. The last sequence is obtained when the fix point is reached, i.e., when the obtained lattice family is isomorphic to the one from the previous step and the contexts are unchanged. The RCA process guarantees that such a fix point exists.

For our example, the fix point is obtained after two steps. The lattice for taxons is the same during all the process (see Fig. 1(a)). The final lattice for sites is shown in Fig. 3. In $\mathcal{L}^1_{Sites}$ lattice, we observe:

- `CSlat11` represents the group of sites (`site0`, `site3` and `site5`) which own relational attribute ∃cont(`CAlat11`), meaning that they contain at least one individual from `CAlat11` extent (individuals owning property `MH8`).
- `CSlat10` represents the group of sites (`site3` and `site5`) which own relational attribute ∃cont(`CAlat10`), meaning that they contain at least one individual from `CAlat10` extent (individuals owning properties `MH5` and `MH8`).
- `CSlat15` represents the group of sites (`site1` and `site4`) which own relational attribute ∃cont(`CAlat1`) and relational attribute ∃cont(`CAlat15`), meaning that they contain at least one individual from `CAlat1` extent (individuals owning property `MH1`) and an individual from `CAlat15` extent (individuals owning property `MH9`).

To appear in the *International Journal of General Systems*
Vol. 00, No. 00, Month 20XX, 11–22

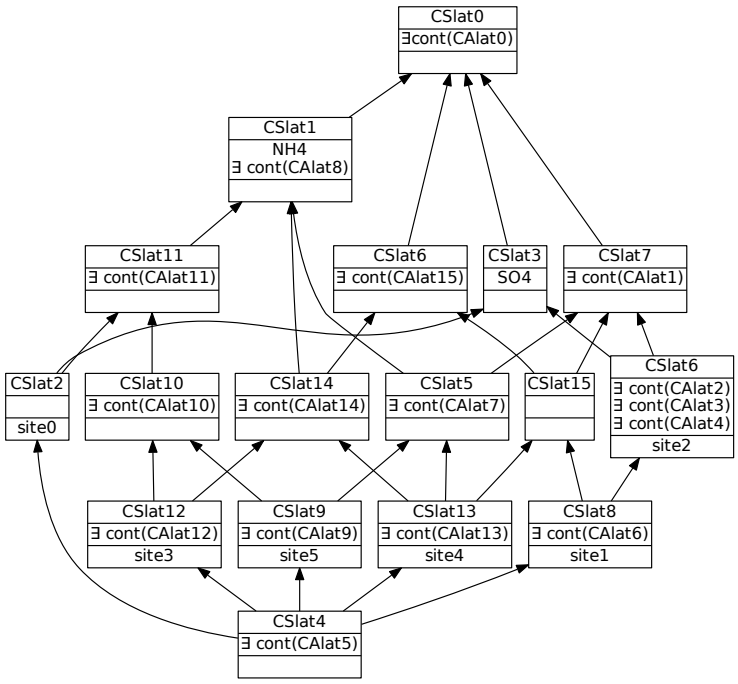


Figure 3. Lattice of sites ($\mathcal{L}^1_{Sites}$) (step 1 of RCA)

4.2. *A variant of RCA based on AOC-posets*

As explained in Section 2, AOC-posets are used as a scalable alternative to concepts lattices in some application domains where the remaining concepts give the needed information or structure. We call RCA-AOC the variant of RCA which uses AOC-posets instead of lattices. The principle of RCA-AOC is roughly the same as this of classical RCA. A consequence of using AOC-posets, is that, with some (rare) specific circular data schemas, RCA-AOC may diverge. The precise convergence conditions are under study. Nevertheless, in the current application we can assume the convergence of this method.

We illustrate the variant with the same example. Table 4 shows, after the triple bar, $\mathrm{S}_{(r_{contains},\exists),\mathcal{C}_{Taxons}}(K_{Sites})$ where $\mathcal{C}_{Taxons}$ is the set of concepts of the AOC-poset of Fig. 1(b). For a purpose of notation, we now use the concepts of the AOC-poset $\mathcal{C}_{Taxons}$ instead of the lattice $\mathcal{L}_{Taxons}$ when we apply the definitions of the previous sections.

The whole Table 4 shows the relational extension of the site context. The relational extension of our relational context family is composed of Table 1 (no outgoing relation), and Table 4. The AOC-poset built from this extended context is shown on Fig. 4. Comparing this AOC-poset and the lattice presented in Fig. 3 we observe that 6 concepts have disappeared. The other concepts are as follows.

- 6 Concepts introducing objects are similar in the two hierarchies, i.e., they have same simplified-extents and same simplified-intents' cardinality.
- 2 Concepts introducing simple attributes are similar in the two hierarchies.
- 3 Concepts `CSlat6`, `CSlat7`, and `CSlat11` correspond respectively to three concepts of the AOC-poset, `CSaocp3`, `CSaocp5`, and `CSaocp4` : they have the same extents (including inheriting objects).
- 1 Concept `CSlat15` from $\mathcal{L}^1_{Sites}$ disappears since both its simplified-intent and its simplified-extent are empty.
- 5 Concepts from $\mathcal{L}^1_{Sites}$ have no corresponding concepts in the AOC-poset since

To appear in the *International Journal of General Systems*
Vol. 00, No. 00, Month 20XX, 12–22

Table 4. Existential Scaling of Formal Context of sites $S_{(r_{contains},\exists),\mathcal{C}_{Taxons}}(K_{Sites})$

Site	NH4	SO4	∃cont(CAaoc0)	∃cont(CAaoc1)	∃cont(CAaoc2)	∃cont(CAaoc3)	∃cont(CAaoc4)	∃cont(CAaoc5)	∃cont(CAaoc6)	∃cont(CAaoc7)	∃cont(CAaoc8)	∃cont(CAaoc9)	∃cont(CAaoc10)
site0	×	×					×					×	
site1		×	×	×	×					×	×		×
site2		×	×	×							×		×
site3	×				×	×	×					×	
site4	×		×		×		×	×					
site5	×		×				×		×			×	

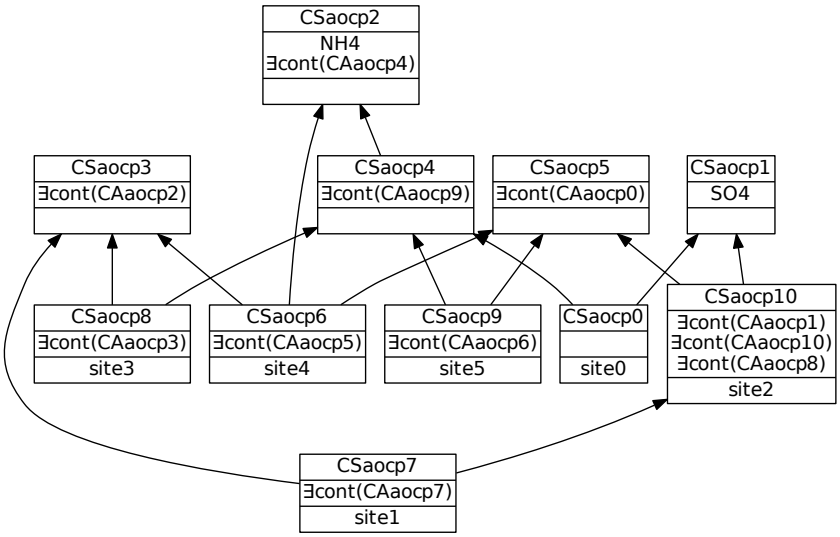


Figure 4. AOC-poset AOC^2_{Sites} for the K^2_{Sites} formal context

the relational attributes they introduce do not correspond to any concepts of the AOC-poset AOC_{Taxons} (Fig. 1(b)).

The whole construction process consists in building a (possibly infinite) sequence of contexts and AOC-posets. In this simple example, the following steps do not produce any new concept (a fix-point is reached), but, as we said, in the general case the process may diverge. Both approaches are implemented in a single tool called RCAExplore [4]. RCAExplore was originally created to implement an exploration process described in Dolques et al. (2013) which permits the redefinition of the construction configuration at each iteration of the process. This tool is developed in Java and implements the RCA process with a modular architecture: new concept generation algorithms and new scaling operators can easily be implemented to be used by the RCA process. The tool proposes a few concept generators among which are a concept lattice building using an adapted version of the AddIntent algorithm (van der Merwe, Obiedkov, and Kourie 2004) and an AOC-poset building using

[4]http://dataqual.engees.unistra.fr/logiciels/rcaExplore

To appear in the *International Journal of General Systems*
Vol. 00, No. 00, Month 20XX, 13–22

Hermes algorithm (Berry et al. 2012).

5. Analysing Water datasets

We present in this section an application of our approach to FRESQUEAU project. We first describe the used dataset and the addressed problems. We show that AOC-posets are relevant with respect to iceberg lattices.

5.1. *The dataset*

We rely on a large database collecting data on Alsatian streams and water areas (North-east of France) which is described by Grac et al. (2011), and more data are available through the current FRESQUEAU project, concerning larger areas and periods. The data, which have been collected during 3 years over 40 sites in the Alsace plain, come either from samples (e.g., physical, physico-chemical and biological data collected on stream sites), synthetic data (e.g., biological indices, land cover) or from the literature (e.g., information about the aquatic species living in the streams that were analysed with FCA in a previous work (Bertaux et al. 2009)). More precisely in this paper we work with three tables. The first one gives values of 27 physical (e.g., state of the river bed, presence of hydraulic structures) and physico-chemical parameters (e.g., temperature, pH, sulfite SO4, nitrite NH4, organic matters) collected on 49 stream sites[5]. The second table gives the level of population for 197 macro-invertebrates (e.g., *Ancylus*, *Anisus*, *Anodonta*) collected on the same 49 sites. The third one describes the macro-invertebrates with 18 different life traits, i.e., their characteristics and functioning, e.g., life cycle, reproduction mode, etc. Each life trait is represented by several modalities (e.g., for the life trait *life cycle* there are two possible modalities: *less than a year* or *more than a year*). The number of modalities over all life traits is 116. We look for rules combining life traits and physico-chemical parameters, e.g., *"the M modality of the T life trait is associated with a high value of the C physico-chemical parameter"*.

Data are modeled within 4 formal contexts: *stream sites*, *physico-chemical parameters*, *life traits* and *macro-invertebrates* and we consider the three relations between them that are described by our tables: level of physico-chemical parameter, population of macro-invertebrates and life trait of macro-invertebrates. We separate values into several classes by preprocessing the numerical values of the different relations. So the *level of physico-chemical parameter* relation has been split into 5 binary relations describing 5 different levels, the *population of macro-invertebrates* relation has been split into 3 different binary relations and the *life trait of macro-invertebrates* relation has been split into 6 binary relations. This binarisation process has been accomplished under the guidance of a domain expert. The whole relational context family, denoted RCF_w in the following, is represented in Figure 5.

5.2. *Principles of the approach*

The initial need of the domain experts is to find relations between the physico-chemical state of a watercourse and the life trait modalities of the taxons living there. The output result has to be limited in size and complexity as it will be analyzed by hand and some metrics should be provided to highlight the most relevant results.

[5]In the following both physical and physico-chemical parameters are called physico-chemical parameters.

To appear in the *International Journal of General Systems*
Vol. 00, No. 00, Month 20XX, 14–22

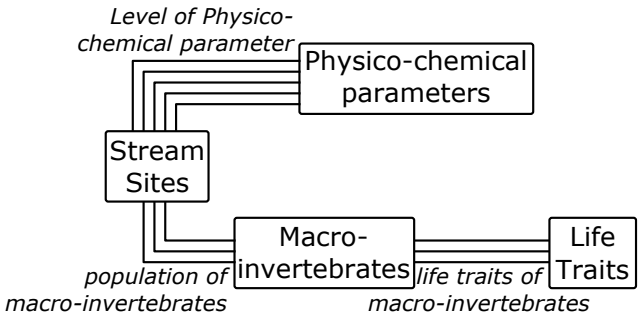


Figure 5. Relational schema of RCF_w.

We propose here to meet these needs by extracting implication rules between physico-chemical characters and life trait modalities. The rules will be proposed to the experts sorted by support. Tests are made using 3 different scaling operators on the relational context family RCF_w:

- the universal strict scaling operator $\forall\exists$ described in Definition 3.
- the percent scaling operator $S_{>n\%}$ described in Definition 4. Three values of n are used.
- the existential scaling operator $\exists$, described in Definition 5.

Those operators are applied on the relation *population of macro-invertebrates* while other relations, due to the fact they are mostly targeting singleton concepts (Physico-chemical parameters and life traits are attributes for relational classification but they are not classified by other attributes), will use the existential scaling operator.

Classical RCA leads to a combinatorial explosion of the number of concepts that forces us to try other methods of classification that reduce the number of concepts. The use of iceberg lattices has been considered but the choice of the threshold is arbitrary. On our case study, the number of concepts obtained by computing iceberg lattices on one object-attribute relation, while computing AOC-posets on others, depending on the chosen threshold is represented in Figure 6 and is compared to the number of concepts obtained with AOC-posets on all object-attribute relations. The threshold is here described in percentage of the number of objects in the concept, e.g., iceberg(90) will only compute the concepts with an extent size greater than 90% of the number of objects. The number of concepts grows exponentially and there are more than 10,000 concepts by lattice with a threshold around 80%.

Iceberg lattices limit the concepts computed to the most general ones, which is useful to extract only the most frequent behaviors found in the data, but makes it impossible to extract less frequent but interesting behaviors that could be found with an AOC-poset. The figure also illustrates the impracticality of classical RCA (which is equivalent to iceberg lattices with a threshold of 0%) as it is not computable (in time and in memory) on a desktop configuration[6] and the number of concepts is far too large to be manually analyzed by the expert.

6. Extracting Implication Rules

Implication rules are implications that are verified by the whole considered dataset. Some implication rules can be extracted from the AOC-poset concepts by considering

[6]Experiments were done on a Processor Intel®Core™2 Duo CPU L9600 @ 2.13GHz × 2 and 1.7 GiB of RAM.

To appear in the *International Journal of General Systems*
Vol. 00, No. 00, Month 20XX, 15–22

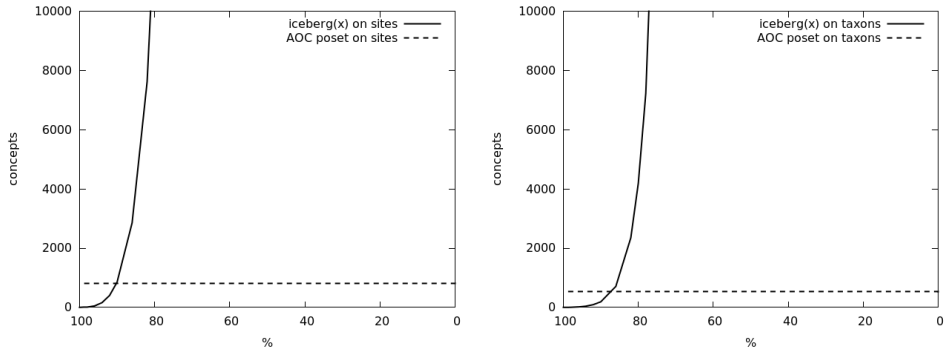


Figure 6. Concept number explosion with Iceberg-lattices compared to AOC-posets. For the streamsite lattice on the LHS and for the taxons lattice on the RHS.

their simplified intent.

An attribute a from the simplified intent of a concept C is an attribute that is not contained in the intent of any concept more general than C, i.e., the set of all the objects sharing a is the extent of C. The objects from the extent also all share the attributes of the full intent of C, but they may not be the only ones. By consequence, for every object o in the dataset, the presence of a implies the presence of all the attributes from the intent of C. Thus, rules are calculated as follows:

for *Concept* C **do**
  if $simplifiedIntent(C) \neq \emptyset$ **then**
    for *Attribute* $a \in simplifiedIntent(C)$ **do**
      $a \rightarrow \wedge\{b | b \in fullIntent(C) \wedge a \neq b\}$;

For instance, in Fig. 1(b) we can consider the concept `CAaocp7`. $Intent(\text{CAaocp7}) = \{MH4, MH3, MH9\}$ and $Intent_S(\text{CAaocp7}) = \{MH4\}$ which means that in all the dataset the rule $MH4 \rightarrow MH3 \wedge MH9$ is verified. In natural language this rule says that taxons living in silt (MH4) also live in sand (MH3) and mud (MH9). Note that this rule cannot be taken as relevant considering the small size of the example.

All the concepts with non empty simplified intent can be found in the AOC-poset meaning that all the rules having one element in the premise can be found with an AOC-poset. Extracting rules from an existing AOC-poset is straightforward as it consists in reading the simplified intent and the full intent of each concept. The concept order permits to extract rules ordered by support, the rules extracted from the most general concepts having a larger support than more specific rules. The use of RCA permits to use relational attributes in implication rules and thus obtain more expressiveness.

Furthermore, when the attribute number in the rule conclusion is important, the rule is difficult to interpret. We thus propose to extract simplified rules, following to the application needs. The following simplifications are applied.

- First, if the implication rule has a relational attribute as premise pointing to a concept C_p, then we will only keep in the conclusion relational attribute pointing to concepts that are from a different AOC-poset than the one containing C_p. This is motivated by the fact that we want to see how some variables can be influenced by variables from different types, e.g., we would like to find how the physico-chemical state of water influences the taxons living in it.
- In the conclusion, for a given relation and a given scaling operator, only the

To appear in the *International Journal of General Systems*
Vol. 00, No. 00, Month 20XX, 16–22

most specific of the pointed concepts are kept. Indeed, with the scaling operators we use, we have the property that if a concept contains in its intent a relational attribute pointing to a concept C, then it will also contains the relational attributes with the same relation and the same scaling operator pointing to all the concepts more general than C. This property is not necessarily shared by every scaling operator.

- Finally we consider attributes in the intent of C that are inherited from all the concepts that are more general than C and that are separated by at most 3 generalization relations to the current concept C (this can be modified according to the application needs). This is to prevent from having the most shared attributes of the dataset in the conclusion of every computed rule.

6.1. *Experiments*

Experience shows that the full rules describing the most specific concepts can be really large as the size of their intent is large. However, some attributes of these concepts were found to be irrelevant. For instance, the dataset describes fresh water stream sites, so most of the taxons have a high level of affinity with the modality *fresh water* of the trait *salinity*. Hence a high affinity with the modality *fresh water* will appear in most of the concepts and is not considered as relevant to experts since all considered watercourses are fresh water. Seeing that, we produced the previously described simplified rules according to experts request.

Table 5. Description of the results depending on the scaling operator

Scaling Operator	Number of Concepts	Full Rules describing		Simplified Rules describing	
		Life traits	*Physico-chemical parameters*	*Life traits*	*Physico-chemical parameters*
$\exists$	954	1428	134	170	134
$S_{>25\%}$	696	917	134	285	134
$S_{>50\%}$	502	708	134	301	134
$S_{>75\%}$	412	627	134	312	134
$\forall\exists$	390	589	134	305	134

From the obtained AOC-posets we extracted two sets of rules: rules with physico-chemical parameters in the premise and rules with life traits in the premise. Table 5 denotes the number of concepts in the *site* AOC-poset obtained for each scaling operator and the number of simplified and full rules. The different scaling operators have a direct influence on the number of generated concepts: the $\exists$ scaling operator is the less constrained and produces the biggest number of concepts, which are found too generic as the simple occurrence of a modality will add a new attribute to an object. The $\forall\exists$ scaling operator is the most constrained and produces the lowest number of concepts, which are found too specific. Indeed the $\forall\exists$ scaling operator requires that all the taxons found share a characteristic to add this characteristic as an attribute to a stream site. In between we introduce 3 instances of the percent operator, respectively $S_{>25\%}$, $S_{>50\%}$ and $S_{>75\%}$. The $S_{>50\%}$ operator, which we call the majority operator, is found to be a good compromise as it requires several occurrences of a modality for it to be considered but allows some exceptions. Exact implications can be hard to find, especially on real world data with a lot of external parameters influencing the results. The percent operator helps us coping with noise in the dataset that makes the use of $\forall\exists$ inefficient; furthermore it provides stronger

To appear in the *International Journal of General Systems*
Vol. 00, No. 00, Month 20XX, 17–22

rules than the ∃ scaling operator that only reveals that ubiquitous taxons are present everywhere.

We obtained 1428 rules describing life traits with the existential operator, and around half of it with the majority operator. Simplifying the rules also leads to a reduction of their number, which means that many rules were built from general attributes. The reduction of the rule number after simplification is more important with existential scaling as relational attributes describing life traits are more general and are introduced higher in the AOC poset. However this reduction is not necessarily following the value of the percent operator as we can note that the number of simplified rules describing life traits is higher with $S_{>75\%}$ than with $\forall\exists$, which would be equivalent to $S_{=100\%}$. The number of rules describing physico-chemical parameters is quite surprising as it depends neither on the operator used nor on the simplification process. This is partly explained by the following properties of the dataset.

- The scaling operators are not applied on the relation *level of physico-chemical parameters* which means that the relational attributes extracted from this relation are always the same associated to the same sets of objects. The number of relational attributes pointing to a concept describing a physico-chemical parameter is always the same: $25 \times 5 = 135$.
- There exist in life traits some modalities that are shared by nearly all the taxons studied here, e.g., as we study only freshwater watercourses all the taxons will have a strong affinity with freshwater. This means that in every sample, for every physico-chemical parameter at every value, we can create a rule implying strong affinity to freshwater.
- In our dataset, only one relational attribute pointing to a concept describing a physical parameter does not describe a sample, i.e., a rule can not be created from it. Actually this physical parameter admits only four value classes in the current dataset.

Besides, the simplification process we used defines arbitrarily, considering the expert feedback, to put in the rule conclusions only attributes introduced by concepts distant from at most 3 by the generalization relation. When reducing this limit, experiments show that the rule number may be lower.

6.2. *Rule examples and discussion*

The rule presented below is an example of rules extracted from the experimental data and selected as a relevant one by the expert:

$$
\begin{array}{ll}
S_{>50\%} & \text{high\_population}(\\
 & \quad \exists \text{strong\_affinity}(\text{transversaldistribution} : \text{banks and sidearms})) \\
\rightarrow & \quad \exists \text{good\_state}(\text{banks})
\end{array}
$$

This rule means that if a *site* of the dataset has more than half of its highly represented taxons having a transversal distribution mainly on banks and sidearms, then the state of its banks is considered as good. The following rule represents another knowledge extracted from the dataset:

To appear in the *International Journal of General Systems*
Vol. 00, No. 00, Month 20XX, 18–22

$$\begin{array}{ll} S_{>50\%} & \text{high\_population}(\\ & \quad \exists \text{strong\_affinity}(\text{slow current})) \\ \rightarrow & \quad \exists \text{bad\_state}(\text{hydrology}) \end{array}$$

This rule means that if a site of the dataset has more than half of its highly represented taxons preferring slow current, then the hydrological state of the site is considered as bad. Each rule is extracted from a concept and as such a partial order can be defined on the rules, following the partial order of the concepts that generate them. The experts can follow this partial order to analyze the rules.

The levels that are referred here are defined specifically to each physico-chemical parameter, each taxon, and each biological character. For the physico-chemical parameters, chemical parameters are usually attached to a level of concentration, physical parameters have different kinds of values describing the state of the site. The same goes for the taxon presence. For the biological characters it refers to an affinity of the species to the modality of a particular biological trait (e.g., for its habitat, an animal may have a stronger affinity with *sand* than *gravel*, but may be found in both of them).

Although the modeling can have several variants for analysis purpose (the current one being established with domain experts) and several approaches could be used to limit the number of concepts, we saw with the current modeling, that classical RCA reveals scale issues while we are only working on a small part of the data. Using AOC-poset allows us to handle large data without exploding the number of concepts. Besides, reducing the number of concepts means that the extracted information is also reduced. For the stream site concepts, from which we extract the rules, the AOC-poset keeps the relevant data as we still have the concepts where each attribute is introduced and the concept order. For the macro-invertebrates however, several concepts are lost that represent combinations of shared life traits. We still have to measure the impact of the missing concepts on the results but it would seem appropriate to combine the RCA-AOC approach with a more traditional lattice building approach that would compute (using the AOC-poset) a few additional relevant concepts for the macro-invertebrates while keeping the number of stream site concepts low with AOC-poset.

The rule format, with premises of size 1, is a small set of all the rules that can be generated, but the premise being a relational attribute we obtain more expressivity than could be obtained for instance with classical FCA. By varying scaling operators rules can be more general or specific and the concept pointed by the relational attribute can be a rich description with several attributes, e.g., with a relational attribute pointing to a taxon concept, we can have a premise including several life traits modalities.

7. Conclusion

This paper has detailed and refined the RCA-AOC process that we introduced in Dolques, Le Ber, and Huchard (2013). This process is based on AOC-poset in order to deal with computational complexity over big datasets. Actually AOC-posets reduce the number of concepts, with no information lost as the context can still be retrieved from an AOC-poset. It allows us to compute implication rules linking attributes from different tables. According to the chosen scaling operator, various rules can be obtained. In this paper we have introduced a specific operator for ex-

To appear in the *International Journal of General Systems*
Vol. 00, No. 00, Month 20XX, 19–22

ploring relational data about watercourses and aquatic species characteristics. The approach proved to be efficient and revealed relevant rules.

In the future, we will work on specifying the convergence conditions of AOC-poset based RCA. Indeed, more complex datasets may include cycles between objects. Convergence is ensured with the RCA specification from Hacène et al. (2013a) where the set of concepts used at each step is the set of concepts of the whole lattice. With AOC-posets, the convergence is not guaranteed when there are cycles between objects.

Regarding our application aim, we will study how to extract bases of binary rules, and how to calculate some of the non-binary rules by reconstructing from the AOC-poset parts of the lattice that experts have identified as significant. We also plan to evaluate with the experts whether they prefer to interpret reduced rule sets (implication rule bases) or at the contrary more complete rule sets (containing inferences).

A tool for vizualising the results and assisting their exploration is under development. Finally the approach will be tested on other datasets of the FRESQUEAU project, and compared with other approaches for relational data mining such as statistical approaches (Doledec et al. 1996) or propositionalization (Lachiche 2010). A first attempt is described in Dolques et al. (2014).

Acknowledgements

This work was funded by ANR11_MONU14. We warmly acknowledge Karell Bertet for sharing part of her knowledge on implication rules and Alain Gutierrez (LIRMM) for the implementation of the AOC-poset algorithm (`https://www.lirmm.fr/aoc-poset-builder/`).

References

Aboud, Nour, Gabriela Arévalo, Olivier Bendavid, Jean-Rémy Falleri, Nicolas Haderer, Marianne Huchard, Chouki Tibermacine, Christelle Urtado, and Sylvain Vauttier. 2014. "Building Hierarchical Typed Component Directories using Formal Concept Analysis." Submitted to Journal of Object Technologies.

Adaricheva, Kira V., James B. Nation, and Robert Rand. 2011. "Ordered direct implicational basis of a finite closure system." *CoRR* abs/1110.5805. `http://arxiv.org/abs/1110.5805`.

Agrawal, Rakesh, Tomasz Imieliński, and Arun Swami. 1993. "Mining Association Rules Between Sets of Items in Large Databases." *SIGMOD Rec.* 22 (2): 207–216. `http://doi.acm.org/10.1145/170036.170072`.

Al-Msie'deen, Ra'Fat, Abdelhak-Djamel Seriai, Marianne Huchard, Christelle Urtado, and Sylvain Vauttier. 2013. "Mining features from the object-oriented source code of software variants by combining lexical and structural similarity." In *IRI,* 586–593. IEEE.

Arévalo, Gabriela, Jean-Rémy Falleri, Marianne Huchard, and Clémentine Nebut. 2006. "Building Abstractions in Class Models: Formal Concept Analysis in a Model-Driven Approach." In *MoDELS 2006,* 513–527.

Azmeh, Zeina, Maha Driss, Fady Hamoui, Marianne Huchard, Naouel Moha, and Chouki Tibermacine. 2011a. "Selection of Composable Web Services Driven by User Requirements." In *ICWS 2011,* 395–402.

Azmeh, Zeina, Marianne Huchard, Amedeo Napoli, Mohamed Rouane-Hacène, and Petko Valtchev. 2011b. "Querying Relational Concept Lattices." In *CLA'11,* 377–392.

Baader, Franz, and Felix Distel. 2008. "A Finite Basis for the Set of $\mathcal{EL}$-Implications Holding

To appear in the *International Journal of General Systems*
Vol. 00, No. 00, Month 20XX, 20–22

in a Finite Model." In *Formal Concept Analysis, 6th Int. Conf., ICFCA, LNCS 4933,* 46–61.

Bendaoud, Rokia, Amedeo Napoli, and Yannick Toussaint. 2008. "Formal Concept Analysis: A unified framework for building and refining ontologies." In *EKAW 2008,* LNCS 5268. 156–171.

Berry, Anne, Marianne Huchard, Amedeo Napoli, and Alain Sigayret. 2012. "Hermes: an efficient algorithm for building Galois Sub-hierarchies." In *CLA 2012,* 21–32.

Bertaux, A., F. Le Ber, A. Braud, and M. Trémolières. 2009. "Identifying ecological traits: a concrete FCA-based approach." In *ICFCA 2009,* LNAI 5548. 224–236. Springer-Verlag.

Bertet, Karell, Stéphanie Guillas, and Jean-Marc Ogier. 2007. "Extensions of Bordat's Algorithm for Attributes." In *Proceedings of the Fifth International Conference on Concept Lattices and Their Applications, CLA 2007, Montpellier, France, October 24-26, 2007,* `http://ceur-ws.org/Vol-331/Bertet.pdf`.

Bertet, K., and B. Monjardet. 2010. "The multiple facets of the canonical direct unit implicational basis." *Theoretical Computer Science* 411 (22–24): 2155 – 2166. `http://www.sciencedirect.com/science/article/pii/S0304397510000034`.

Clark, Peter, and Robin Boswell. 1991. "Rule Induction with CN2: Some Recent Improvements." In *EWSL,* 151–163.

Dao, Michel, Marianne Huchard, Mohamed Rouane Hacène, Cyril Roume, and Petko Valtchev. 2004. "Improving Generalization Level in UML Models Iterative Cross Generalization in Practice." In *Proceedings of Conceptual Structures at Work: 12th International Conference on Conceptual Structures, ICCS 2004, Huntsville, AL, USA, July 19-23, 2004,* 346–360. `http://dx.doi.org/10.1007/978-3-540-27769-9_23`.

Doledec, S., D. Chessel, C.J.F. ter Braak, and S. Champely. 1996. "Matching species traits to environmental variables: a new three-table ordination method." *Environmental and Ecological Statistics* 3: 143–166.

Dolques, Xavier, Florence Le Ber, Marianne Huchard, and Clémentine Nebut. 2013. "Relational Concept Analysis for Relational Data Exploration." In *Advances in Knowledge Discovery and Management - Volume 5 [Best of EGC 2013, Toulouse, France],* Vol. 615 of *Studies in Computational Intelligence* edited by Fabrice Guillet, Bruno Pinaud, Gilles Venturini, and Djamel Abdelkader Zighed. 57–77. Springer. `https://doi.org/10.1007/978-3-319-23751-0\_4`.

Dolques, Xavier, Marianne Huchard, Clémentine Nebut, and Philippe Reitz. 2012. "Fixing Generalization Defects in UML Use Case Diagrams." *Fundam. Inform.* 115 (4): 327–356.

Dolques, Xavier, Florence Le Ber, and Marianne Huchard. 2013. "AOC-posets: a scalable alternative to Concept Lattices for Relational Concept Analysis." In *CLA 2013: 10th International Conference on Concept Lattices and Their Applications, Oct 2013, La Rochelle, France,* 129–140. `http://ceur-ws.org/Vol-1062/`.

Dolques, Xavier, Kartick Chandra Mondal, Agnès Braud, Florence Le Ber, and Marianne Huchard. 2014. "RCA as a data transforming method: a comparison with propositionalisation." In *ICFCA 2014: International Conference on Formal Concept Analysis, June 2014, Cluj, Romania,* LNCS 8478. 112–127.

Ferré, Sébastien, Olivier Ridoux, and Benjamin Sigonneau. 2005. "Arbitrary Relations in Formal Concept Analysis and Logical Information Systems." In *ICCS'05,* Vol. 3596 of *LNCS*166–180. Springer.

Ganter, B., and R. Wille. 1999. *Formal Concept Analysis: Mathematical Foundations.* Springer Verlag.

Godin, R., and H. Mili. 1993. "Building and Maintaining Analysis-Level Class Hierarchies using Galois Lattices." In *OOPSLA '93,* Vol. 28Washington, DC, USA. 394–410.

Godin, R., H. Mili, G. W. Mineau, R. Missaoui, A. Arfi, and T.-T. Chau. 1998. "Design of Class Hierarchies based on Concept (Galois) Lattices." *Theory and Application of Object Systems* 4 (2): 117–134.

Grac, C., F. Le Ber, A. Braud, M. Trémolières, A. Bertaux, A. Herrmann, S. Manné, and M. Lafont. 2011. *Programme de recherche-développement* Indices – *Rapport scienfique final.* Contrat pluriannuel 1463 de l'Agence de l'Eau Rhin-Meuse. LHYGES – LSIIT – ONEMA – CEMAGREF.

To appear in the *International Journal of General Systems*
Vol. 00, No. 00, Month 20XX, 21–22

Hacène, Mohamed Rouane, Marianne Huchard, Amedeo Napoli, and Petko Valtchev. 2013a. "Relational concept analysis: mining concept lattices from multi-relational data." *Ann. Math. Artif. Intell.* 67 (1): 81–108.

Hacène, Mohamed Rouane, Marianne Huchard, Amedeo Napoli, and Petko Valtchev. 2013b. "Soundness and Completeness of Relational Concept Analysis." In *ICFCA,* Vol. 7880 of *Lecture Notes in Computer Science* edited by Peggy Cellier, Felix Distel, and Bernhard Ganter. 228–243. Springer.

Hitzler, P. 2004. "Default Reasoning over Domains and Concept Hierarchies." In *Proc. of KI 2004,* Vol. 3238 of *LNCS*351–365. Springer Verlag.

Huchard, M., H. Dicky, and H. Leblanc. 2000. "Galois Lattice as a Framework to specify Algorithms Building Class Hierarchies." *Theoretical Informatics and Applications* 34: 521–548.

Huchard, Marianne, Mohamed Rouane-Hacène, Cyril Roume, and Petko Valtchev. 2007. "Relational concept discovery in structured datasets." *Ann. Math. Artif. Intell.* 49 (1-4): 39–76.

Kötters, Jens. 2013. "Concept Lattices of a Relational Structure." In *20th Int. Conf. on Conceptual Structures, ICCS 2013, Mumbai, India,* LNCS 7735. 301–310.

Krmelova, Markéta, and Martin Trnecka. 2013. "Boolean Factor Analysis of Multi-Relational Data." In *10th Int. Conf. on Concept Lattices and Their Applications, CLA 2013, La Rochelle, France,* CEUR Workshop Proceedings 1062. 187–198.

Krmelova, Markéta, and Martin Trnecka. 2014. "An algorithm for the Multi-Relational Boolean Factor Analysis based on essential elements." In *Proceedings of the Eleventh Int. Conf. on Concept Lattices and Their Applications, CLA 2014, Kosice, Slovakia,* CEUR Workshop Proceedings 1252. 107–118.

Lachiche, Nicolas. 2010. "Propositionalization." In *Encyclopedia of Machine Learning,* edited by Claude Sammut and Geoffrey I. Webb. 812–817. Springer.

Moha, Naouel, Amine Rouane Hacène, Petko Valtchev, and Yann-Gaël Guéhéneuc. 2008. "Refactorings of Design Defects Using Relational Concept Analysis." In *ICFCA 2008,* 289–304.

Muggleton, S., and L. de Raedt. 1994. "Inductive logic programming: Theory and methods." *The Journal of Logic Programming* 19 (20): 629–679.

Osswald, R., and W. Pedersen. 2002. "Induction of Classifications from Linguistic Data." In *Proc. of ECAI'02 Workshop,* July.

Osswald, Rainer, and Wiebke Petersen. 2003. "A Logical Approach to Data-Driven Classification." In *Proceedings of the 26th Annual German Conference on Advances in Artificial Intelligence, KI 2003,* Vol. 2821 of *LNCS*267–281. Springer.

Petersen, W. 2001. "A Set-Theoretical Approach for the Induction of Inheritance Hierarchies." In *ENTCS,* Vol. 51Elsevier. July.

Prediger, Susanne, and Rudolf Wille. 1999. "The Lattice of Concept Graphs of a Relationally Scaled Context." In *ICCS'99,* 401–414. Springer.

Quinlan, J.R. 1986. "Induction of decision trees." *Machine Learning* 81–106.

Rouane-Hacène, Mohamed, Petko Valtchev, and Roger Nkambou. 2011. "Supporting Ontology Design through Large-Scale FCA-Based Ontology Restructuring." In *ICCS 2011,* 257–269.

Ryssel, Uwe, Joern Ploennigs, and Klaus Kabitzsch. 2011. "Extraction of feature models from formal contexts." In *SPLC Workshops,* edited by Ina Schaefer, Isabel John, and Klaus Schmid. 4. ACM.

Saada, Hajer, Xavier Dolques, Marianne Huchard, Clémentine Nebut, and Houari A. Sahraoui. 2012. "Generation of Operational Transformation Rules from Examples of Model Transformations." In *MoDELS 2012,* 546–561.

Shi, Lian, Yannick Toussaint, Amedeo Napoli, and Alexandre Blansché. 2011. "Mining for Reengineering: an Application to Semantic Wikis using Formal and Relational Concept Analysis." In *ESWC'11,* Vol. 6644 of *LNCS*421–435. Springer.

Stumme, Gerd, Rafik Taouil, Yves Bastide, Nicolas Pasquier, and Lotfi Lakhal. 2002. "Computing Iceberg Concept Lattices with TITANIC." *Data Knowl. Eng.* 42 (2): 189–222. `http://dx.doi.org/10.1016/S0169-023X(02)00057-5`.

To appear in the *International Journal of General Systems*
Vol. 00, No. 00, Month 20XX, 22–22

van der Merwe, Dean, Sergei A. Obiedkov, and Derrick G. Kourie. 2004. "AddIntent: A New Incremental Algorithm for Constructing Concept Lattices." In *ICFCA,* Vol. 2961 of *Lecture Notes in Computer Science* edited by Peter W. Eklund. 372–385. Springer.

Wille, Rudolf. 1997. "Conceptual Graphs and Formal Concept Analysis." In *5th Int. Conf. on Conceptual Structures, ICCS'97,* LNCS 1257. 290–303.

Wolff, Karl Erich. 2009. "Relational Scaling in Relational Semantic Systems." In *17th Int. Conf. on Conceptual Structures, ICCS 2009,* LNCS 5662. 307–320.