Effects of Input Data Formalisation in Relational Concept Analysis for a Data Model with a Ternary Relation

Priscilla Keip, Alain Gutierrez, Marianne Huchard, Florence Le Ber, Samira Sarter, Pierre Silvie, Pierre Martin

Effects of Input Data Formalisation in Relational Concept Analysis for a Data Model with a Ternary Relation

Priscilla Keip[1][0000−0001−6542−3360], Alain Gutierrez, Marianne Huchard[2][0000−0002−6309−7503], Florence Le Ber[3][0000−0002−2415−7606], Samira Sarter[5][0000−0001−5115−0824], Pierre Silvie[1,4][0000−0002−3406−6230], and Pierre Martin[1][0000−0002−4874−5795]

[1] CIRAD, UPR AIDA, F-34398 Montpellier, France
AIDA, Univ Montpellier, CIRAD, Montpellier, France
{priscilla.keip,pierre.silvie,pierre.martin}@cirad.fr
[2] LIRMM, Université de Montpellier, CNRS, Montpellier, France
marianne.huchard@lirmm.fr
[3] ICube, Université de Strasbourg, CNRS, ENGEES, Illkirch-Graffenstaden, France
florence.leber@engees.unistra.fr
[4] IRD, UMR EGCE, F-91198 Gif-sur-Yvette, France
[5] ISEM, Univ Montpellier, CIRAD, CNRS, EPHE, IRD, Montpellier, France
samira.sarter@cirad.fr

Abstract. Today pesticides, antimicrobials and other pest control products used in conventional agriculture are questioned and alternative solutions are searched out. Scientific literature and local knowledge describe a significant number of active plant-based products used as bio-pesticides. The Knomana (KNOwledge MANAgement on pesticide plants in Africa) project aims to gather data about these bio-pesticides and implement methods to support the exploration of knowledge by the potential users (farmers, advisers, researchers, retailers, etc.). Considering the needs expressed by the domain experts, Formal Concept Analysis (FCA) appears as a suitable approach, due do its inherent qualities for structuring and classifying data through conceptual structures that provide a relevant support for data exploration. The Knomana data model used during the data collection is an entity-relationship model including both binary and ternary relationships between entities of different categories. This leads us to investigate the use of Relational Concept Analysis (RCA), a variant of FCA on these data. We consider two different encodings of the initial data model into sets of object-attribute contexts (one for each entity category) and object-object contexts (relationships between entity categories) that can be used as an input for RCA. These two encodings are studied both quantitatively (by examining the produced conceptual structures size) and qualitatively, through a simple, yet real, scenario given by a domain expert facing a pest infestation.

Keywords: Biopesticides · Data exploration · Formal Concept Analysis · Relational Concept Analysis

1 Introduction

Today pesticides, antimicrobials and other pest control products used in conventional agriculture are questioned and alternative solutions are searched out, including active plant-based products. The Knomana (KNOwledge MANAgement on pesticides plants in Africa) project aims to identify plants used as biopesticides, currently from the scientific literature, and to implement methods to support the exploration of knowledge by the potential users (farmers, advisers, researchers, retailers, etc.). About 30000 descriptions of plant uses have been collected and recorded according to a data model designed through meetings with domain experts. Each plant use is described using 36 attributes such as the plant taxonomy, the protected system (i.e. crop, animal and human being), or the preparation method.

Considering the data exploitation needs expressed by the domain experts, Formal Concept Analysis (FCA) appears as a suitable approach, due do its inherent qualities for structuring and classifying data through conceptual structures that provide a relevant support for data exploration. To exploit the recorded data, and extract knowledge about alternative protection systems, we rely on Relational Concept Analysis (RCA) [11], one of the possible extensions of Formal Concept Analysis [9] for relational data. RCA input is a so-called *relational context family*, composed with object-attribute (formal) contexts, which describe objects from various categories, and object-object (relational) contexts, which describe relationships between objects of several categories. It outputs a set of conceptual structures (such as concept lattices, AOC-posets or Iceberg lattices) connected through *relational attributes* which point to concepts. Each conceptual structure classifies the objects of one category according to the (initial) attributes and the *relational attributes*, thus according to the relations that the objects of this category have with objects and object groups (concepts) of another (or the same) category.

Considering a real-word context such as the Knomana project dataset, although a data model has been set for data collection purpose, there are still many ways to encode the data model into a *relational context family*. In this paper, we study the impact of the definition of a *relational context family* on the practicability of RCA on the Knomana project dataset. The first question is raised by the fact that relational contexts represent binary relations between objects, but the data model we deal with contains a ternary relation. The model has thus to be converted into a model with binary relations, while respecting the original semantics of the ternary relation. Two encodings are envisaged: the first one considers a reification of the ternary relation (i.e. a specific object-attribute context represents the 3-tuples), while the second one projects the ternary relation into three binary relations, one for each pair of the linked object sets. The second question is related with the possibility, for each relationship, to consider one direction only or both directions. It is connected to the potential explorations that the domain experts may have in mind. The proposed encodings are studied quantitatively (by examining the produced conceptual structures size

and running time) and qualitatively, through a simple, yet real, scenario given by a domain expert facing a pest infestation.

Section 2 presents Relational Concept Analysis and the RCAExplore tool which is used in our evaluations. In Section 3, we propose two encodings of the initial data model into *relational context families*. Then, in Section 4, we first show the size and computation time of several conceptual structures for the two encodings on an excerpt of the Knomana dataset restricted to a few key plants designated by the domain experts. Then we study a simple, yet real, exploration scenario using both encodings, to show their relevancy with respect to the studied question. Section 5 exposes related work. We conclude and give perspectives of this work in Section 6.

2 Background

RCA extends the purpose of Formal Concept Analysis (FCA, [9]) to relational data. RCA applies iteratively FCA on a Relational Context Family (RCF), that is a pair $(\mathcal{K}, \mathcal{R})$, where $\mathcal{K}$ is a set of object-attribute contexts and $\mathcal{R}$ is a set of object-object contexts. $\mathcal{K}$ contains n object-attribute contexts $K_i = (G_i, M_i, I_i), i \in \{1, ..., n\}$. $\mathcal{R}$ contains m object-object contexts $R_j = (G_k, G_l, r_j)$, $j \in \{1, ..., m\}$, where $r_j \subseteq G_k \times G_l$ is a binary relation with $k, l \in \{1, ..., n\}$, $G_k = dom(r_j)$ the domain of the relation, and $G_l = ran(r_j)$ the range of the relation. RCA relies on a relational scaling mechanism that is used to transform a relation r_j into a set of *relational attributes* that extends the object-attribute context describing the objects of $dom(r_j)$. A relational attribute $\exists r_j(C)$, where $\exists$ is the existential quantifier, $C = (X, Y)$ is a concept, and $X \subseteq ran(r_j)$, is owned by an object $g \in dom(r_j)$ if $r_j(g) \cap X \neq \emptyset$. Other quantifiers can be found in [11]. RCA process consists in applying FCA first on each object-attribute context of an RCF, and then iteratively on each object-attribute context extended by the relational attributes created using the concepts from the previous step. The RCA process stops when the families of lattices of two consecutive steps are isomorphic and the extended object-attribute contexts are unchanged.

Pests	Chrysomelidae	Noctuidae	Liposcelididae	Trichocomaceae
CallosobruchusC	×			
SpodopteraL		×		
LiposcelisB			×	
AspergillusF				×
PenicilliumE				×
CallosobruchusM	×			

Plants	Rhizome	Root	Leaf	Seed	Fruit
AcorusC	×	×			
AlpiniaO	×				
AgeratumC			×		
LaphangiumL			×		
PelargoniumR			×		
HelianthusA				×	
CitrusL					×

treatedBy	AcorusC	AlpiniaO	AgeratumC	LaphangiumL	PelargoniumR	HelianthusA	CitrusL
CallosobruchusC	×						
SpodopteraL	×				×		×
LiposcelisB		×					
AspergillusF			×				
PenicilliumE				×			
CallosobruchusM						×	

Table 1. A relational context family about (a) `Pests` and (b) `Plants` able to treat them, as indicated in `treatedBy` relation (c)

In the following, we consider a small example from the Knomana dataset. Object-attribute contexts `Plants` and `Pests` (Table 1, (a) and (b)) respectively represent a set of plants and a set of pests. Pests are described by their family (e.g. *Callosobruchus maculatus* –`CallosobruchusM` in Table 1– is a member of the leaf beetle family, *Chrysomelidae*), while plants are described by the parts (e.g. fruit or leaf) that are used for treating the pests. The object-object context `treatedBy` (Table 1, (c)) represents the link between pests and plants they can be treated by. At the first step of the RCA process, FCA is applied on `Plants` and `Pests` and results in two lattices $\mathcal{L}_{\texttt{Plants}}$ and $\mathcal{L}_{\texttt{Pests}}$ (Figure 1).

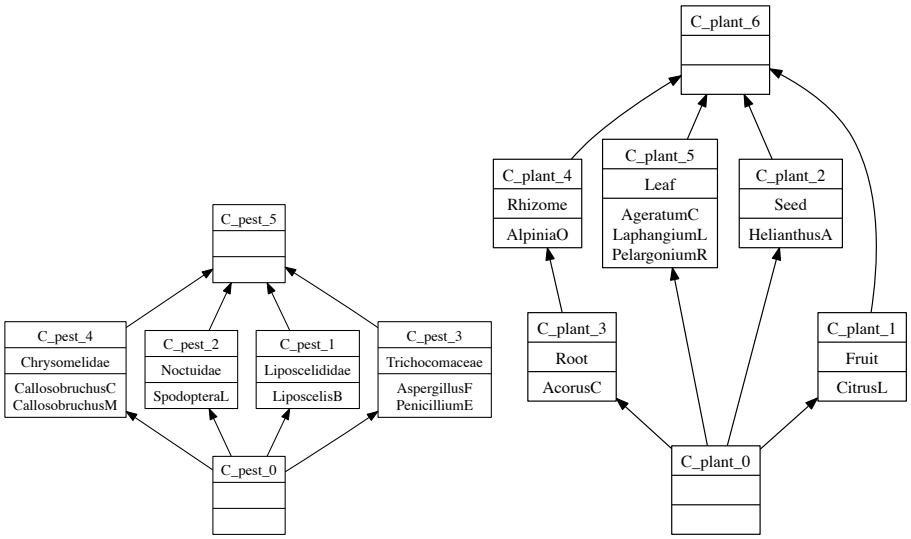

Fig. 1. Lattices $\mathcal{L}_{\texttt{Pests}}$ (left) and $\mathcal{L}_{\texttt{Plants}}$ (right)

At the second step of the process, `Pests` context is extended with relational attributes built from context `isTreatedBy` and concepts of $\mathcal{L}_{\texttt{Plants}}$ (Table 2). For instance, the relational attribute $\exists$`treatedBy`(`C_plant_2`) is added to `CallosobruchusM` since this pest is related to `HelianthusA` which is an object of `C_plant_2`. A new lattice is built, that is represented in Figure 2. In this last lattice, we can observe that pests grouped in `C_pest_8` are both treated by *Acorus Calamus* `AcorusC` (sweet flag), using its root or rhizome (see `C_plant_3`).

The tool RCAexplore[6] was developed during project ANR 11 MONU 14 Fresqueau, in order to explore relational hydroecological data. RCAExplore is an implementation of the RCA process where several choices can be made before each iteration: the algorithm to be used, the scaling operator, and the considered contexts.

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

Pests*	**Chrysomelidae**	**Noctuidae**	**Liposcelididae**	**Trichocomaceae**	∃ treatedBy(C_plant_0)	∃ treatedBy(C_plant_1)	∃ treatedBy(C_plant_2)	∃ treatedBy(C_plant_3)	∃ treatedBy(C_plant_4)	∃ treatedBy(C_plant_5)	∃ treatedBy(C_plant_6)
CallosobruchusC	×							×	×		×
SpodopteraL		×				×		×	×	×	×
LiposcelisB			×						×		×
AspergillusF				×						×	×
PenicilliumE				×						×	×
CallosobruchusM	×						×				×

Table 2. Extended context from `Pests`.

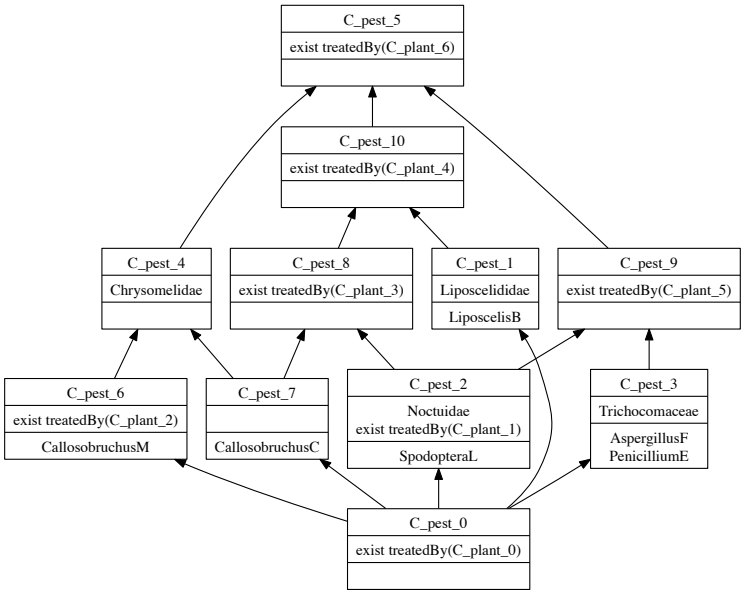

Fig. 2. Lattice built on the extended context `Pest*` of Table 2

3 From the Knomana Model to a Relational Context Family and Conceptual Structures

The Knomana database gathers descriptions of plant uses, each one characterized using 36 data types, including the protecting plant, the protected organism, the controlled aggressor (also called pest and disease), the method adopted to prepare the product to be applied, or the reference to the document describing the use. Currently, the Knomana database comprises 28700 plant use descriptions manually entered from 250 documents (mainly scientific publications). The

descriptions include 966 plant species, originated from 60 territories, used to protect 39 species of organism (animal, vegetal, and human) against 253 species of aggressors (Bacteria, Chromista, Eukaryota, Fungi, Insecta, and Virus).

In the data model, data types are grouped as data classes to represent the three main entity categories of the system: biopesticide, protected system and targeted organism. To represent the biological system, these three main entity categories (or data classes) are linked through a ternary relationship. As the relational contexts of RCA are binary relationships, two different data models have been designed. The first one, called M1 (see left-hand side of Figure 3), consists in reifying the ternary relationship as a specific data class, the latter supporting binary relationships with each of the three main data classes of the biological system. The second one, called M2 (see right-hand side of Figure 3), consists in establishing binary relationships between the data classes of the biological system, corresponding to projections of the ternary relationship. The relation directions have been determined in order to obtain classifications and propagation of relations according to the first question (Q1) asked by the experts (introduced below).

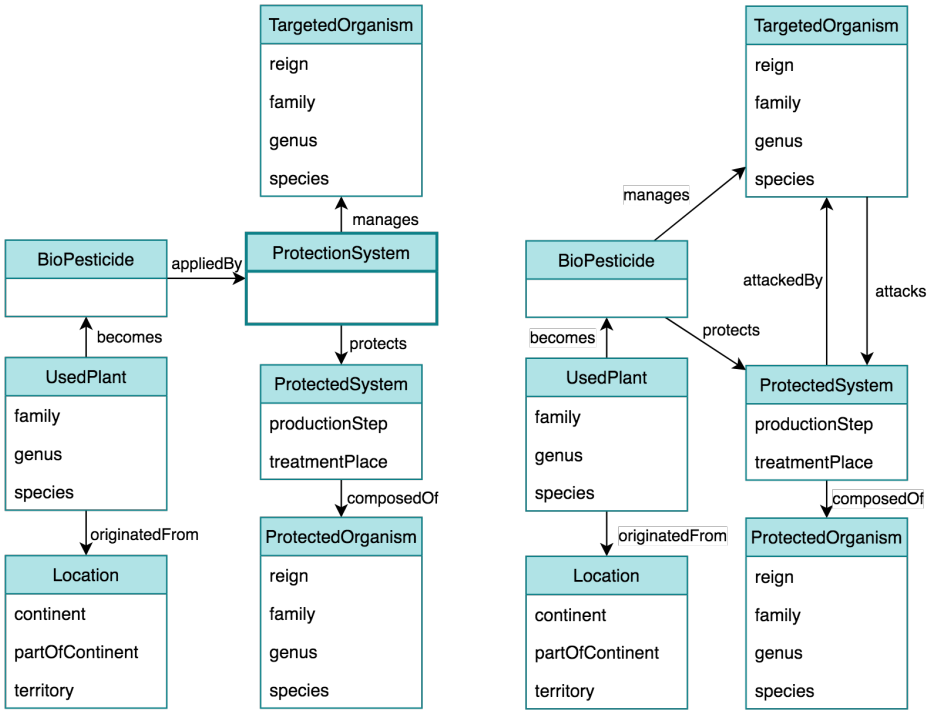

Fig. 3. Models M1 (left-hand side) and M2 (right-hand side). Implementations of the data model : (M1) with the ternary relationship as a data class, (M2) without the ternary relationship, which is transformed by establishing binary relationships between the main data classes

The encoding of M1 and M2 as relational context families (RCF) consists in converting each data class as a formal context (object-attribute context) and each arrow as a relational context (object-object context). To measure the effect of the encodings on the resulting conceptual structures, two encodings of the M1 and M2 arrows are considered: the "original" encoding implements the arrows presented in Figure 3, while the "enhanced" encoding includes the arrows and their opposite (making a sort of symmetric closure at the model level).

M1 and M2 encodings are evaluated on a dataset reduced to the descriptions associated to six protecting plant species, i.e. *Cymbopogon citratus*, *Hyptis suaveolens*, *Lantana camara*, *Moringa oleifera*, *Ocimum gratissimum*, and *Thymus vulgaris*. These plants have been selected by domain experts of the Knomana project for their first investigations, according to these plant efficacy in contrasted situations, diversity of applications, and high presence in most of the West African territories. The dataset comprises 225 descriptions composed of 16 pieces of information: the protecting plant (name of the species, genus and family), the plant origin (territory, part of continent, and continent), the targeted organism (name of the species, genus, family, and reign), and the protected system. The latter is described using the protected organism (name of the species, genus, family, and reign), the production step (e.g. crop, or cattle) and the treatment place (e.g. field, or stock). Table 3 presents the number of objects and attributes of the formal contexts of M1 and M2 on the reduced dataset, and details the number of binary attributes generated for each data type of the model.

Formal context (Class name)	Number of objects	Number of attributes	Included data types	Number of values
BioPesticide	38	0	(empty class)	-
UsedPlant	6	16	family	4
			genus	6
			species	6
Location	20	32	continent	5
			partOfContinent	7
			territory	20
ProtectedSystem	14	19	productionStep	7
			treatmentPlace	12
ProtectedOrganism	21	48	reign	3
			family	11
			genus	15
			species	19
TargetedOrganism	111	234	reign	5
			family	43
			genus	78
			species	108
ProtectionSystem	225	0	ternary relation	-

Table 3. Description of formal contexts of M1 and M2

RCAExplore software enables to evaluate four kinds of conceptual structures, and algorithms that allow to build them: concept lattices built with addIntent/addExtent [16] (FCA), AOC-posets built with Ares [4] (ARES), and ICEBERG lattices [21] for support 30 and 40 (ICEBERG30, ICEBERG40). An AOC-poset is a restriction of the concept lattice to the concepts introducing objects or attributes. ICEBERG is a restriction of the concept lattice to the concepts having a minimal support (i.e. extent size), here to concepts with a minimal support of 30% and 40%. In the next section, we conduct an evaluation of the dataset along two dimensions: quantitatively, by assessing the possibility of building the conceptual structures and their size, if appropriate, and qualitatively, by analyzing the ability of M1 and M2 to answer a specific case of the following generic biological question Q1 raised by our domain experts: "Given a plant able to protect an organism against an aggressor, which other plants can alternatively be used with the same benefits?". This question corresponds to a so-called "replacement" scenario: replace a plant by another one with supposed similar ability to deal with the observed aggressor on the attacked organism.

4 Results

In this section, we first present the effects of the two proposed encodings on the conceptual structure construction for our six key plant dataset (Section 4.1), and then on a real replacement scenario for an *Aspergillus*[7] attack (Section 4.2).

4.1 Six Key Plant Dataset: Conceptual Structure Variants

Tables 4 and 5 respectively show for model M1 and model M2 the numbers of concepts that were built for each algorithm, and for the enhanced case (where we take the symmetric closure of the model) and the original cases (as shown in Figure 3). Tables 6 and 7 respectively show for model M1 and model M2 the numbers of relational attributes that were built. Step numbers until RCA stops and execution times are compared in Table 8.

As a first remark, some computations failed (see italics figures between parentheses) on a laptop in the enhanced case. For enhanced M1 (resp. enhanced M2), concept lattices and Iceberg30 (resp. concept lattices) could not be computed because of lack of memory. Computing AOC-posets and Iceberg40 was always possible, but Iceberg40 shows very few concepts in the original models, thus we suspect it will not be very useful for experts in this case.

In Tables 4 and 5, the AOC-posets for both enhanced models M1 and M2 show similar concept numbers, except for `ProtectionSystem`, which is specific to M1. The concept number of `ProtectionSystem` AOC-poset can be roughly obtained from the sum of the concept numbers of the AOC-posets of the neighbour contexts (`Biopesticide`, `ProtectedSystem` and `TargetedOrganism`). The concept numbers of `Location` and `ProtectedOrganism` do not change between

[7] *Aspergillus* genus groups several species of microscopic fungi.

	M1 enhanced				M1 original			
	FCA	ARES	ICEBERG30	ICEBERG40	FCA	ARES	ICEBERG30	ICEBERG40
ProtectionSystem	*(1660066)*	1151	*(2750032)*	619	415	238	9	3
BioPesticide	*(16212)*	359	*(3625)*	57	576	113	15	3
UsedPlant	*(51)*	36	*(16)*	3	51	37	12	3
Location	*(191)*	63	*(31)*	5	29	27	4	3
ProtectedSystem	*(1585)*	154	*(771)*	20	42	32	5	4
ProtectedOrganism	*(300)*	95	*(38)*	17	33	29	4	3
TargetedOrganism	*(380084)*	560	*(9386)*	61	153	151	5	2
TOTAL	*(2058489)*	2418	*(2763899)*	782	1299	627	54	21

Table 4. M1 model: number of concepts for each algorithm (italics figures between parentheses are for failed computations because of lack of memory)

	M2 enhanced				M2 original			
	FCA	ARES	ICEBERG30	ICEBERG40	FCA	ARES	ICEBERG30	ICEBERG40
BioPesticide	*(307618)*	354	9563	84	23650	178	60	6
UsedPlant	*(55)*	36	16	3	57	38	12	3
Location	*(191)*	63	31	5	29	27	4	3
ProtectedSystem	*(2270)*	153	316	20	747	81	21	5
ProtectedOrganism	*(495)*	93	42	108	33	29	4	3
TargetedOrganism	*(1363817)*	555	48556	108	7186	216	35	4
TOTAL	*(1674446)*	1254	58524	328	31702	569	136	24

Table 5. M2 model: number of concepts for each algorithm (italics figures between parentheses are for failed computations because of lack of memory)

original models M1 and M2 because they are sinks in the model graph. In the enhanced case, `Location` and `UsedPlant` concept numbers are almost the same so that we can assume that they do not influence each other too much. For both enhanced and original models and considering the cases where the computation finished, M2 Iceberg40 lattices contain more concepts, what suggests that M2 concepts are more populated than M1 concepts; nevertheless there is no significant change in scaling factor. Whole concept lattices are built only for the original M1 and M2 models. For enhanced M2 model, the number of concepts for `Biopesticide` and `TargetedOrganism` explodes, likely due to the circuit between `TargetedOrganism` and `ProtectedSystem`.

Tables 6 and 7 show the numbers of relational attributes and inform us about the grouping factor provided by the conceptual structures. The formal contexts that are sinks in the model (no outgoing relation) have no relational attributes. While observing the enhanced models, we can notice that M2 gives rise to less relational attributes (but the difference is more significant for Iceberg40 than for AOC-posets). For the original models, this is the reverse, there are more relational attributes for M2 than for M1, which could be explained by the fact that M2 original contains a circuit, which is not the case of M1 original. We also observe a significant difference between Iceberg30 and Iceberg40 in M2.

	M1 enhanced				M1 original			
	FCA	ARES	ICEBERG30	ICEBERG40	FCA	ARES	ICEBERG30	ICEBERG40
ProtectionSystem	*(3330)*	1087	*(2025)*	138	195	183	10	6
BioPesticide	*(47959)*	1173	*(423074)*	622	415	238	9	3
UsedPlant	*(1881)*	428	*(457)*	78	605	140	19	6
Location	*(51)*	36	*(16)*	35	0	0	0	0
ProtectedSystem	*(48126)*	1230	*(23091)*	655	33	29	4	3
ProtectedOrganism	*(289)*	156	*(219)*	68	0	0	0	0
TargetedOrganism	*(47908)*	1137	*(423058)*	853	0	0	0	0
TOTAL	*(149544)*	5247	*(871940)*	2449	1248	590	42	18

Table 6. M1 model: number of relational attributes added at each formal context for each algorithm (italics figures between parentheses are for failed computations because of lack of memory)

	M2 enhanced				M2 original			
	FCA	ARES	ICEBERG30	ICEBERG40	FCA	ARES	ICEBERG30	ICEBERG40
BioPesticide	*(24777)*	744	48888	131	7933	297	56	9
UsedPlant	*(5323)*	417	9594	89	23679	205	64	9
Location	*(51)*	36	16	3	0	0	0	0
ProtectedSystem	*(29560)*	1002	58161	209	7219	245	39	7
ProtectedOrganism	*(585)*	153	316	20	0	0	0	0
TargetedOrganism	*(5789)*	507	9879	104	747	81	21	5
TOTAL	*(66085)*	2859	126854	556	39578	828	180	30

Table 7. M2 model: number of relational attributes added at each formal context for each algorithm (italics figures between parentheses are for failed computations because of lack of memory)

Table 8 shows the running times and the step numbers. The step numbers are similar in original M1 and M2. Computing concept lattices for original M2 needs more steps due to the existing circuit and the creation of many non-introducer concepts. The 16 steps for obtaining the AOC-poset of enhanced M1 are noticeable and correspond to a running time relatively high, compared to the others. The different running time for original M1 and M2 (from 127ms to about 9s) allows to envisage online work for experts. For enhanced models (AOC-posets), it can be preferable to compute them offline.

	enhanced models		original models	
	M1	M2	M1	M2
FCA	*(5)*	*(4)*	5	9
ARES	16	10	5	6
ICEBERG30	*(7)*	10	6	6
ICEBERG40	11	8	5	6

	enhanced models		original models	
	M1	M2	M1	M2
FCA	-	-	351	9722
ARES	311149	28288	1195	1677
ICEBERG30	-	29864	137	144
ICEBERG40	796	223	127	166

Table 8. (left) Final step number and (right) computation time (milliseconds) for each algorithm and each model

In the light of the above evaluation, AOC-poset and Iceberg40 are appropriate for the dataset on both M1 and M2 original models. They will be used in the next section on a real question raised by the experts. Iceberg40 gives incomplete information, but allows us to focus on frequent situations. AOC-poset, as it holds all the introducer concepts, contains the whole initial information. It can be used to build the entire concept lattice. A concept which appears in the concept lattice and not in the AOC-poset represents a group Ext of objects and a group Int of attributes such that (1) each object of Ext is introduced in a sub-concept because it has an attribute which is not in Int, and (2) each attribute of Int is introduced in a super-concept because it is owned by an object which is not in Ext. These concepts are useful to reveal data regularities. In our context, they could be connection points between different exploration paths. In the future, we will evaluate in which extent they are useful during exploration, as they could be built on the fly based on the introducer concepts. Let us notice that the algorithm running time does not cover all the needed time for a concrete analysis. In a real scenario, the analyst also needs to select and extract or focus on presumed relevant data.

4.2 Aspergillus Attack: Answering a Concrete Replacement Scenario

To assess the pertinence of our approach, we have selected a smaller dataset from Knomana base and have explored it with both models M1 and M2 with AOC-posets. This smaller dataset contains the same 6 plants, but only targeted organisms of *Aspergillus* family. The aim is then to answer an instantiation of general question Q1: "knowing recognized benefits of *Hyptis suaveolens* in the management of *Arachis hypogaea* against *Aspergillus parasiticus*, which other plants could alternatively be used?".

Figure 4 gives a simplified version of an excerpt from the AOC-posets built from `BioPesticide` and `ProtectionSystem` contexts, according to M1 model. In this figure, arrows represent the navigation links (the relational attributes) which allowed to find and highlight the concepts and the hierarchy we want to give to our domain expert to explore data around their question. The bold numbers are the concept numbers, the italic text is for objects of the AOC-poset and normal text is for relational attributes.

Starting from the introducer concept of *Hyptis suaveolens*, in `UsedPlant` AOC-poset (not shown), we can navigate to concepts 0, 10, 9, 7 and 5 of `BioPesticide` AOC-poset (see left of Figure 4). The most specific concept among them is concept 5 which introduces H_B, i.e. the biopesticide produced from *Hyptis suaveolens* coming from Benin. Following the relational attributes of the concept 5, we can navigate to concepts 9 and 10 of ProtectionSystem AOC-poset (see right of Figure 4). Concepts 9 and 10 are together non comparable; their relational attributes show that concept 9 groups plant uses that protect *Arachis hypogaea* (`PeS 0`) from *Aspergillus ochraceus* (`TO 1`), while concept 10 groups plant uses that protect *Arachis hypogaea* from *Aspergillus parasiticus* (`TO 2`), the last biological system being the one we want to manage.

In `BioPesticide` AOC-poset, we also notice that concept 5 owns a subconcept, concept 1, that introduces OG_B, i.e. a biopesticide produced from *Ocimum gratissimum* coming from Benin. According to the lattice order, which is preserved in AOC-posets, it can be deduced, thanks to the inheritance of the attributes, that the biopesticide produced from *Ocimum gratissimum* coming from Benin allows to manage at least one of the same biological systems as presented in concepts 9 and 10 of `ProtectionSystem` AOC-poset. *Ocimum gratissimum* can thus be used instead of *Hyptis suaveolens* in order to protect *Arachis hypogaea* against *Aspergillus parasiticus*, but also against *Aspergillus ochraceus*, and *Aspergillus flavus*. These facts can be checked in Knomana knowledge base.

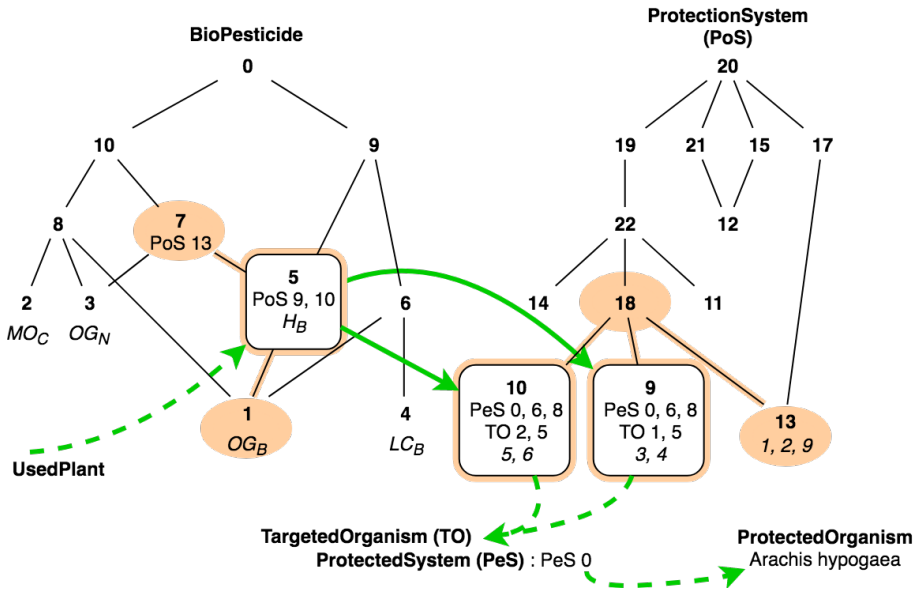

Fig. 4. Simplified version of an excerpt from the AOC-posets built from `BioPesticide` and `ProtectionSystem` contexts with M1 to answer Q1 in the case study: navigated concepts and their links are highlighted

Besides, concept 5 inherits a relational attribute PoS 13 from concept 7, that leads to concept 13 in `ProtectionSystem` AOC-Poset. This concept 13 is not comparable with concepts 9 and 10, but all these three concepts are subconcepts of concept 18. Concept 13 groups same plant uses as concepts 9 and 10 (protecting *Arachis hypogaea* against *Aspergillus flavus*), but also a different use (protecting *Oryza sativa* against *Aspergillus flavus*) due to the chosen encoding. Actually, following model M1, protected systems are first classified with respect to the production step and the treatment location, and then with respect to the protected organism.

Furthermore, concept 1 is a subconcept of concept 5 but also of other concepts in `BioPesticide` AOC-poset. Based on these hierarchical links we can infer that the biopesticide produced from *Ocimum gratissimum* coming from Benin can protect other biological systems than the ones previously described. The hierarchical organization highlights these facts for the domain experts.

Let us now consider the analysis based on M2 model; a simplified excerpt of the resulting AOC-posets is shown in Figure 5. Starting from the introducer concept of *Hyptis suaveolens* in UsedPlant AOC-poset (not shown), we can navigate to concept 5 of BioPesticide AOC-poset (see middle of Figure 5) that introduces H_B, i.e. the biopesticide produced from *Hyptis suaveolens* coming from Benin. As for model M1, concept 5 has a subconcept introducing OG_B. Both models give currently the same result. Going further, we see that relational attributes of concept 5 are of two types: eight of them lead to concepts of `ProtectedSystem` AOC-poset (see right of Figure 5) while the seven others lead to concepts of `TargetedOrganism` AOC-poset (see left of Figure 5). The attributes leading to `TargetedOrganism` concepts reveal, by looking at the most specific concepts (number 1, 2 and 0), that the biopesticides produced from *Hyptis suaveolens* and *Ocimum gratissimum* are used to fight against *Aspergillus ochraceus*, *Aspergillus parasiticus* and *Aspergillus flavus*. The attributes leading to `ProtectedSystem` concepts allow to find again *Arachis hypogaea* that is introduced by the most specific among the targeted concepts, concept 0, thanks to relational attribute `PeO 0`.

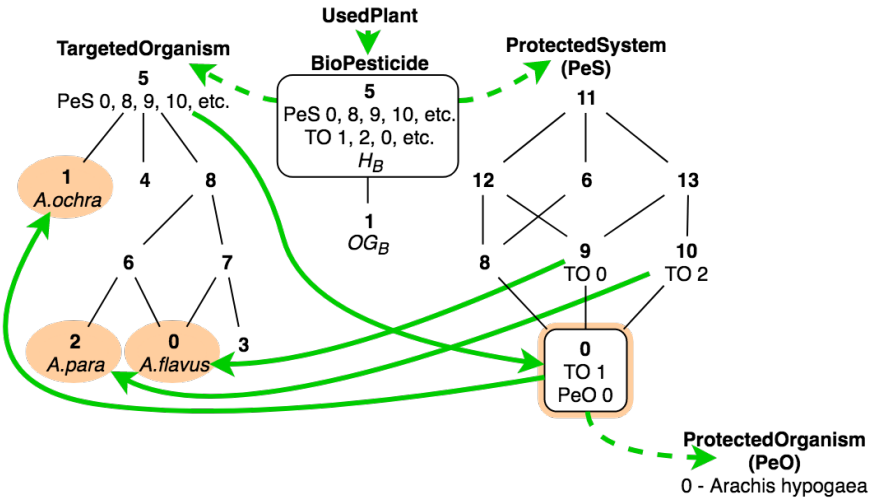

Fig. 5. Simplified version of an excerpt from the AOC-posets built from `TargetedOrganism` and `ProtectedSystem contexts` with M2 to answer Q1 in the case study: navigated concepts and their links are highlighted

To summarize, for the case study of this small dataset, both models M1 and M2 allow to find *Ocimum gratissimum* as an alternative plant to *Hyptis suaveolens* for protecting *Arachis hypogaea* against *Aspergillus parasiticus*. In addition, a query will not give more information, contrary to RCA. The formed concepts not only give one or more answers to the initial question, but they also show how these answers are classified, and which additional (not included in the initial question) description they share. Indeed, both models also show that *Hyptis suaveolens* can be replaced by *Ocimum gratissimum* to protect *Arachis hypogaea* against *Aspergillus ochraceus*, and *Aspergillus flavus*, which is an additional information. Besides, by examining the neighborhood of the concepts which give the searched answer, the experts can formulate hypotheses for new research. E.g., if they notice that a plant protects a targeted organism against a specific aggressor, they may design experiments for evaluating if other plants with similar characteristics (grouped in the same concept) may also have the same effect. However, a set of three binary relations is not equivalent to one ternary relation. Model M2 will thus be sometimes less precise because it lacks the ternary relation. Furthermore, the navigation is more difficult in M2 than in M1 lattice family because of the greater number of concepts and relational attributes in M2 lattices.

5 Related Work

As for any data analysis method, studying the data encoding for Formal Concept Analysis and its impact on the analysis results is an essential phase. In our case, we need to take into account two specific features of our dataset: multi-relational information and ternary relations.

Multi-relational information can be encoded through different and complementary schemes, according to the envisaged analysis. In the FCA domain, several approaches highlight the graph nature of relational data [12, 15], and pattern structures [8] are used to classify graphs describing objects (or tuples). Other approaches [1, 7] rely on logical formula for relational data encoding, providing features equivalent to the RCA scaling quantifiers.

With the RCA scheme, the objective is to classify the objects themselves in several conceptual structures (one per object category), according to the relations that the objects of one category have with objects of another (or the same) category. The encoding scheme is rooted in an entity-relationship model, highlighting the categories (entities, encoded through object-attribute formal contexts) and the relationships (encoded through object-object/relational contexts).

Graph-FCA (G-FCA) [5] proposes to consider knowledge graphs based on n-ary relationships as formal contexts. The intent of a G-FCA concept is a projected graph pattern and the extent is an object relation. In the same vein, triadic concept analysis [14] (resp. more generally polyadic concept analysis [22]) has been introduced to deal with 3-dimensional (resp. n-dimensional) formal contexts. Both proposals could be a solution for giving additional views and

highlighting more specific information on our data and we will consider them as future work.

Reading and interpreting RCA structures is known to be difficult. To facilitate this interpretation, [19] proposed to synthesize the concepts of a main lattice and their related concepts from the other lattices within a hierarchy of closed partially-ordered (sequential) patterns, i.e. directed acyclic graphs. This idea has been generalized by [6], where a family of concept lattices built by RCA is summarized through a hierarchy of concept graphs. Each concept graph is a set of concepts (potentially coming from several lattices) whose intents are mutually dependent, allowing to highlight relational patterns. Concept graphs are then organized according to the specialization order between concepts they include.

Overlays on RCAExplore have been proposed in [20] to help the analyst choices, e.g. by forecasting the number of concepts and rules resulting from a relational concept family and a quantifier or an algorithm, and several configurations are studied on an environmental dataset. Encoding legal document description in an RCF is presented in [17], where a relation links legal documents representing orders to documents representing legislative texts. The resulting conceptual structures are analyzed through relational queries and exploration strategies. The effect of several encodings of the UML meta-model in a relational context family (RCF), that includes or not the navigability and unnamed roles, has been studied in [10], allowing to conclude which RCFs are practicable, and in general about the applicability of RCA in class model normalization. Later on, using concept lattices versus AOC-posets has been studied on 15 UML class models and 15 Java code models in [18], to conclude to the superiority of AOC-posets in performance and relevancy of the produced structures.

6 Conclusion and Perspectives

The Knomana project provides a valuable collection of information about biopesticide plants in Africa. The project comes with many challenges, including information gathering, moving from raw information to knowledge associated with a stable vocabulary and ontology, and exploitation of the gathered information. In this paper, we investigate the information exploitation dimension through the application of relational concept analysis. We analyze variants for encoding the initial data model into a relational context family, and the effect of several encoding options, both quantitatively and qualitatively on a few key plants designated by the domain experts of the Knomana project as their first investigation focus.

The Knomana project is intended to extend its geographical scope to the whole world. The information collection is a continuous task, involving master students and researchers from several countries. Answering the expert questions will benefit from other approaches, such as using an on-demand algorithm [2, 3], exploring other scaling quantifiers, as well as applying metrics evaluating the interest of formal concepts [13]. We envisage to define strategies for formaliz-

ing the expert questions and automatize, at least partially, the construction of appropriate relational context families. To save execution time, we plan to implement in RCAExplore other AOC-poset building algorithms, that we previously implemented in a more specific tool[8]. Besides, RCAExplore is currently moving to the COGUI platform[9] in order to pool knowledge processing activities.

Acknowledgement. This work was supported by the French National Research Agency under the Investments for the Future Program, referred as ANR-16-CONV-0004 and by INRA-CIRAD GloFoodS metaprogram (KNOMANA project).

References

1. Baader, F., Distel, F.: A finite basis for the set of $\mathcal{EL}$-implications holding in a finite model. In: ICFCA, LNCS 4933. pp. 46–61 (2008)
2. Bazin, A., Carbonnel, J., Huchard, M., Kahn, G.: On-demand relational concept analysis. CoRR **abs/1803.07847** (2018), `http://arxiv.org/abs/1803.07847`
3. Bazin, A., Carbonnel, J., Huchard, M., Kahn, G., Keip, P., Ouzerdine, A.: On-demand relational concept analysis. In: Proc. of ICFCA'19 (to appear) (2019)
4. Dicky, H., Dony, C., Huchard, M., Libourel, T.: Ares, adding a class and restructuring inheritance hierarchy. In: Onzièmes Journées Bases de Données Avancées, Nancy, France (Informal Proceedings). pp. 25–42 (1995)
5. Ferré, S.: A Proposal for Extending Formal Concept Analysis to Knowledge Graphs. In: 13th Int. Conference, ICFCA 2015, Nerja, Spain. pp. 271–286. LNCS 9113 (2015)
6. Ferré, S., Cellier, P.: How hierarchies of concept graphs can facilitate the interpretation of RCA lattices? In: 14th Int. Conference CLA 2018, Olomouc, Czech Republic. pp. 69–80 (2018)
7. Ferré, S., Ridoux, O., Sigonneau, B.: Arbitrary relations in formal concept analysis and logical information systems. In: 13th Int. Conference ICCS'05, Kassel, Germany. pp. 166–180 (2005)
8. Ganter, B., Kuznetsov, S.O.: Pattern structures and their projections. In: 9th Int. Conference ICCS'01, Stanford, CA, USA. pp. 129–142 (2001)
9. Ganter, B., Wille, R.: Formal concept analysis - mathematical foundations. Springer (1999)
10. Guédi, A.O., Huchard, M., Miralles, A., Nebut, C.: Sizing the underlying factorization structure of a class model. In: 17th IEEE Int. Conference EDOC 2013, Vancouver, BC, Canada. pp. 167–172 (2013)
11. Hacene, M.R., Huchard, M., Napoli, A., Valtchev, P.: Relational concept analysis: mining concept lattices from multi-relational data. Ann. Math. Artif. Intell. **67**(1), 81–108 (2013)
12. Kötters, J.: Concept Lattices of a Relational Structure. In: 20th Int. Conf. ICCS 2013, Mumbai, India. pp. 301–310. LNCS 7735 (2013)
13. Kuznetsov, S.O., Makhalova, T.P.: On interestingness measures of formal concepts. CoRR **abs/1611.02646** (2016), `http://arxiv.org/abs/1611.02646`

[8] http://www.lirmm.fr/AOC-poset-Builder/
[9] https://www.lirmm.fr/cogui/

14. Lehmann, F., Wille, R.: A triadic approach to formal concept analysis. In: 3rd Int. Conference ICCS'95, Santa Cruz, California, USA. pp. 32–43 (1995)
15. Liquière, M., Sallantin, J.: Structural Machine Learning with Galois Lattice and Graphs. In: ICML, Madison, Wisconsin. pp. 305–313 (1998)
16. van der Merwe, D., Obiedkov, S.A., Kourie, D.G.: Addintent: A new incremental algorithm for constructing concept lattices. In: 2nd Int. Conference ICFCA 2004, Sydney, Australia. pp. 372–385 (2004)
17. Mimouni, N., Nazarenko, A., Salotti, S.: A conceptual approach for relational IR: application to legal collections. In: 13th Int. Conference, ICFCA 2015, Nerja, Spain. pp. 303–318 (2015)
18. Miralles, A., Molla, G., Huchard, M., Nebut, C., Deruelle, L., Derras, M.: Class model normalization - outperforming formal concept analysis approaches with aoc-posets. In: 12th Int. Conf. CLA 2015, Clermont-Ferrand, France. pp. 111–122 (2015), http://ceur-ws.org/Vol-1466/paper09.pdf
19. Nica, C., Braud, A., Dolques, X., Huchard, M., Le Ber, F.: Extracting hierarchies of closed partially-ordered patterns using relational concept analysis. In: 22nd Int. Conf. ICCS 2016, Annecy, France. pp. 17–30 (2016)
20. Ouzerdine, A., Braud, A., Dolques, X., Huchard, M., Le Ber, F.: Régler le processus d'exploration dans l'analyse relationnelle de concepts. le cas de données hydroécologiques. In: Actes de la 19e conférence sur l'extraction et la gestion de connaissances (EGC 2019). Nouvelles Technologies de l'Information (2019)
21. Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., Lakhal, L.: Computing iceberg concept lattices with Titanic. Data & Knowledge Engineering **42**(2), 189–222 (2002)
22. Voutsadakis, G.: Polyadic concept analysis. Order **19**(3), 295–304 (Sep 2002). https://doi.org/10.1023/A:1021252203599