Combining Implications and Conceptual Analysis to Learn from a Pesticidal Plant Knowledge Base

Lina Mahrach, Alain Gutierrez, Marianne Huchard, Priscilla Keip, Pascal Marnotte, Pierre Silvie, Pierre Martin

Combining Implications and Conceptual Analysis to Learn from a Pesticidal Plant Knowledge Base

Lina Mahrach[1], Alain Gutierrez[2], Marianne Huchard[2], Priscilla Keip[1], Pascal Marnotte[1], Pierre Silvie[1,3], and Pierre Martin[1]

[1] CIRAD, UPR AIDA, F-34398 Montpellier, France
AIDA, Univ Montpellier, CIRAD, Montpellier, France
lina.mahrachlm@gmail.com,{firstname.lastname}@cirad.fr
[2] LIRMM, Univ Montpellier, CNRS, Montpellier, France
{firstname.lastname}@lirmm.fr
[3] PHIM Plant Health Institute, Montpellier University, IRD, CIRAD, INRAE, Institut Agro, Montpellier, France

Abstract. Supporting organic farming aims to find alternative solutions to synthetic pesticides and antibiotics, using local plants, to protect crops. Moreover, in the One Health approach (OHA), a pesticidal plant should not be harmful to humans, meaning it cannot be toxic if the crop is consumed or should have a limited and conscious use if it is used for medical care. Knowledge on plant use presented in the scientific literature was compiled in a knowledge base (KB). The challenge is to develop a KB exploration method that informs experts (including farmers) about protection systems properties that respect OHA. In this paper, we present a method that extracts the Duquenne-Guigues basis of implications from knowledge structured using Relational Concept Analysis (RCA). We evaluate the impact of three data representations on the implications and their readability. The experimentation is conducted on 562 plant species used to protect 15 crops against 29 pest species of the Noctuidae family. Results show that consistently splitting data into several tables fosters less redundant and more focused implications.

Keywords: Relational Concept Analysis · Duquenne-Guigues basis · Implication Rules · Life Sciences Knowledge Base · One Health · Formal Concept Analysis

1 Introduction

Reducing the use of synthetic pesticides and antibiotics is a major challenge for the environment and living organisms. Moreover, for the Global South countries, it is also crucial to preserve biodiversity and design sustainable production systems (SPS) that respect the One Health approach (OHA) [4]. OHA calls for an interdisciplinary and intersectoral action in the public management of health problems at the interface between humans, animals, and their shared environment. An alternative solution to synthetic pesticides and antibiotics accepted by

OHA is the use of local plants, in the form of essential oil or aqueous solution, with a pesticidal or anti-parasitic effect. Using such plants requires ensuring that they are not harmful to humans. Some plants can indeed be toxic to humans when inhaled during their spray in the field or ingested through crop consumption. Other plants, also used by humans for medical care, can induce a resistance to certain molecules through excessive absorption. One challenge for the scientific experts and for the farmers is to understand the properties and constraints of the already known **protection systems**, composed of a **crop** to be protected against a **pest** using a protecting **plant**, with respect to OHA.

A significant number of protection systems have been extracted in the scientific literature and gathered in the Knomana knowledge base [13]. Knomana includes several datasets. Among them, PPAf (Pesticide Effect Plant of Africa) currently gathers 44270 descriptions of plants used for plant, animal, human, and public health. In PPAf, each use is described using 70 data, such as the protected organism (e.g. crop, fish, human being), the target organism (e.g insect, fungus, bacterium), the location, and the usage domain (plant, animal, environmental, human, or public health). Knomana also includes PAL (Edible plants), which informs whether plants are consumed by humans as food or drink.

In this paper, we make the assumption that implications are a relevant formalism for delivering information on protection systems relative to OHA. We choose to build the Duquenne-Guigues basis (DGB) of implications for its quality of being a non redundant implication set of minimal cardinality. Besides, we assess the impact of three data representations on the implication form and readability. These three representations reconcile the two datasets and split them into one or several data tables. When the representation has several data tables, we build the DGB of implications from the extended formal contexts computed by Relational Concept Analysis (RCA) [7] with AOC-posets. An experimentation is conducted on a Knomana excerpt composed of 562 plants species used to protect 15 crops against 29 pest species of the Noctuidae family. Results show that consistently reconciling datasets and splitting the data into several tables fosters less redundant and more focused implications.

Section 2 introduces the background and outlines the approach. Section 3 describes the Knomana excerpt and the three studied representations. Section 4 reports and discusses the experiment. Section 5 exposes related research and Section 6 concludes and draws future work.

2 Approach

This section introduces the approach, which combines RCA and the computation of the DGB of implications.

RCA. RCA is designed to analyze a dataset conforming to the entity-relationship model [7]. RCA is an extension of Formal Concept Analysis (FCA) [5]. FCA seeks to extract *formal concepts* from a formal context (FC) $\mathcal{K} = (G, M, I)$ where G is an object set, M is an attribute set and $I \subseteq G \times M$. Two operators, both

denoted by $'$, associate object sets with attribute sets. For $O \subseteq G$, the set of attributes shared by the objects of O is $O' = \{m | \forall g \in O, (g,m) \in I\}$. For $A \subseteq M$, the set of objects that share the attributes of A is $A' = \{g | \forall m \in A, (g,m) \in I\}$. A formal concept $\mathcal{C} = (Extent(\mathcal{C}), Intent(\mathcal{C}))$ associates a maximal object group (extent) with their maximal shared attribute group (intent): $Extent(\mathcal{C}) = Intent(\mathcal{C})'$. More generally, we denote by $\preceq_{\mathcal{C}}$ the concept order: $\mathcal{C}_1 \preceq_{\mathcal{C}} \mathcal{C}_2$ when $Intent(\mathcal{C}_2) \subseteq Intent(\mathcal{C}_1)$ and $Extent(\mathcal{C}_1) \subseteq Extent(\mathcal{C}_2)$. The set of all concepts, provided with $\preceq_{\mathcal{C}}$, forms the concept lattice. The lowest (w.r.t. $\preceq_{\mathcal{C}}$) concept owning one object is its introducer concept. The highest (w.r.t. $\preceq_{\mathcal{C}}$) concept owning one attribute is its introducer concept. The suborder of the concept lattice restricted to these introducer concepts is called the AOC-poset (Attribute-Object Concept poset). For instance, in Table 1, the FC *OrganismInfo* describes plant (pl_i), crop ($prot_i$), and pest ($pest_i$) organisms using their genus ($genus_i$) and their non-use in medical care (*no-medical*). Plants $pl1$ and $pl2$ are grouped as a concept being both from $genus1$ and not used in medical care. $pl1$ and $pl2$ can as well be grouped with $prot1$ and $prot2$ as they are not used in medical care. As presented in Fig. 1, these two concepts, respectively named C_Org_15 and C_Org_22, are ordered by inclusion of their object sets from bottom to top, or equivalently by inclusion of their attribute sets from top to bottom. This figure shows that C_Org_15 is a subconcept of C_Org_22 where C_Org_15 introduces $genus1$, C_Org_22 introduces *no-medical*, C_Org_15 inherits *no-medical* from C_Org_22, and C_Org_22 inherits $pl1$ and $pl2$ from C_Org_15.

Table 1. Example of RCF made of 2 FCs (i.e. *OrganismInfo* and *ProtSystem*) on the top and 3 RCs (i.e. *uses*, *protects*, and *treats*) on the bottom. The attribute set of FC *ProtSystem* is empty.

OrganismInfo	genus1	...	genus4	...	genus6	...	no-medical	...	***ProtSystem***
plant1 (pl1)	x	...		...		...	x	...	1
plant2 (pl2)	x	...		...		...	x	...	2
prot1		...		...	x	...	x	...	3
prot2		...		...	x	...	x	...	4
pest1		...	x	...		...		...	5
pest2		...	x	...		...		...	6
...	...	...	...	...	...	...	...	...	7

uses	pl1	pl2	pl3	pl4	...
1				x	...
2			x		...
3	x				...
4	x				...
5		x			...
6		x			...
7				x	...

protects	...	prot1	prot2	...
1	...			...
2	...			...
3	...	x		...
4	...	x		...
5	...		x	...
6	...		x	...
7	...			...

treats	...	pest1	pest2	...
1	...			...
2	...			...
3	...	x		...
4	...	x		...
5	...		x	...
6	...		x	...
7	...			...

RCA takes a *Relational Context Family* (RCF) as input. A RCF is a pair $(\mathbf{K}, \mathbf{R})$ where $\mathbf{K}$ is a set of FCs ($\mathbf{K} = \{\mathcal{K}_i = (G_i, M_i, I_i)\}_{i=1,2,\ldots,n}$), and each FC describes an object category. $\mathbf{R}$ is a set of relational contexts (RC) between the objects of the FCs. $\mathbf{R} = \{r_j\}_{j=1,2,\ldots,p}$ and $r_j \subseteq G_k \times G_l$ for $k, l \in \{1, 2, \ldots, n\}$. To compute the concepts for each FC considering the RCs, RCA builds *relational attributes* $qr(\mathcal{C})$, where q is a quantifier (e.g. the existential quantifier

$\exists$ or the universal quantifier $\forall$), r is a RC, and $\mathcal{C}$ is a concept on the objects of the co-domain of r. These attributes thus group the individual-to-individual relationships into individual-to-concept relationships. To compute the final conceptual structure family, RCA alternates between building conceptual structures associated with FCs (such as a concept lattice or an AOC-poset) and extending the FCs with relational attributes, including the concepts of these structures, until a fix-point is reached. Table 1 presents a RCF, composed of the FCs *OrganismInfo* and *ProtSystem* and 3 RCs, i.e. *uses*, *protects*, and *treats*. These 3 RCs respectively indicate the plant, the crop, and the pest for each protection system. In this example, the FC *ProtSystem* is finally extended with relational attributes formed with the quantifier $\exists$, a RC (i.e. *uses*, *protects*, or *treats*) and a concept of *OrganismInfo* as shown in Table 2. In Fig. 1, the concepts built on the extended FC (EFC) *ProtSystem* group and organize protection systems by considering the relational attributes.

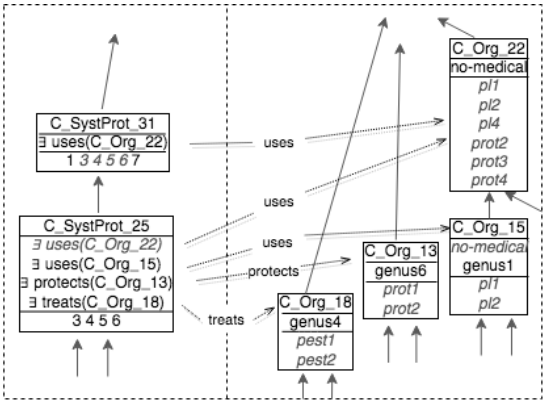

Fig. 1. Partial view of a lattice family, of the RCF presented in Table 1, with the protection system lattice to the left and an organism one to the right. A plain or dashed arrow represents respectively a subconcept-superconcept relation or a cross-lattice link materialized by a relational attribute. Concept $C\_SystProt\_31$ groups 6 protection systems $(1, 3, 4, 5, 6, 7)$ using a plant from concept C_Org_22, i.e. $pl1$, $pl2$, or $pl4$, not used in medical care. $C\_SystProt\_25$, which is a subconcept of $C\_SystProt\_31$, groups 4 protection systems $(3, 4, 5, 6)$, informing that they use a plant from $genus1$ ($\exists uses(C\_Org\_15)$), not used in medical care ($\exists uses(C\_Org\_22)$) to protect a crop from $genus6$ ($\exists protects(C\_Org\_13)$) against a pest of $genus4$ ($\exists treats(C\_Org\_18)$).

Implications An implication, denoted by $A \implies B$, is a pair of attribute sets $(A, B), A, B \subseteq M$ where all the objects that own the attributes of A (premise) also own the ones of B (conclusion): $A' \subseteq B'$. For example, the implication (I1) indicates that no plant of $genus1$ is used in medical care:

$\{genus1\} \implies \{no - medical\}$ ***(I1)***

Table 2. Excerpt of the EFC *ProtSystem* presenting relational attributes formed with the universal quantifier ∃, a relation (*uses*, *protects*, or *treats*), and a concept from FC *OrganismInfo*.

ProtSystem	...	$\exists uses(C\_Org\_15)$	$\exists uses(C\_Org\_22)$	...	$\exists protects(C\_Org\_13)$	...	$\exists treats(C\_Org\_18)$	...
1	...		x	...		...		...
2	...			...		...		...
3	...	x	x	...	x	...	x	...
4	...	x	x	...	x	...	x	...
5	...	x	x	...	x	...	x	...
6	...	x	x	...	x	...	x	...
7	...		x	...		...		...

There are several types of implication sets and bases [1] that can be computed from a FC. Binary implications such as (I1) can also be obtained from $\preceq_{\mathcal{C}}$ and the introducing attributes' concepts, e.g. in Fig.1: C_Org_15 introduces $genus1$, while its superconcept C_Org_22 introduces *no-medical*. The Duquenne-Guigues Basis (DGB) of implications can be defined upon pseudo-intents [6]. A pseudo-intent is an attribute set $P_i \subseteq M$ such that: P_i is not an intent ($P_i'' \neq P_i$); for any other pseudo-intent $P_j \subset P_i$, $P_j'' \subset P_i$. The DGB is the implication set $\{P_i \implies P_i'' | P_i \text{ is a pseudo-intent}\}$. It is canonical and a cardinality minimal set of non redundant implications, from which all implications can be produced.

Our approach. In our work, we compute the DGB of implications, that is usually built for an FC. When using RCA, implications are extracted when the fix-point is reached. For a FC which is not extended, because it is not the object set of a RC, the DGB of implications is directly computed on itself. For a FC which is extended, the DGB is built from its extension (EFC). In our approach, AOC-posets are built at each RCA step. For an easier interpretation of the implications extracted from the EFCs, the concepts in the relational attributes are recursively replaced by the 'non-relational' attributes that serve as seeds for these concepts [15, 16]. For instance, the implication *(I2)* becomes *(rewritten I2)*:

$\{\exists treats(C\_Org\_18)\} \Longrightarrow \{\exists uses(C\_Org\_22)\}$ ***(I2)***
$\{\exists treats(genus4)\} \Longrightarrow \{\exists uses(no-medical)\}$ ***(rewritten I2)***

Both *(I2)* and *(rewritten I2)* stipulate that for the protection systems treating a pest of $genus4$ ($pest1$ or $pest2$ grouped in C_Org_18), we then observe the use of one of the plants ($pl1, pl2, pl4$) grouped in concept C_Org_22, these plants not being used in medical care as indicated by C_Org_22 intent. Implication (I2) can also be read from: $C\_SystProt\_25 \preceq_{\mathcal{C}} C\_SystProt\_31$; $C\_SystProt\_25$ introduces $\exists treats(C\_Org\_18)$; and $C\_SystProt\_31$ introduces $\exists uses(C\_Org\_22)$. The *scope* (S) of an implication informs on the number of objects verifying the implication premise, the *support* being the proportion of EF or EFC objects verifying the implication premise: Let $Imp = A \implies B$, we have $S(Imp) = |A'|$. $Support(Imp) = S(Imp)/|G|$. Fig. 2 summarizes this computation process for the running example.

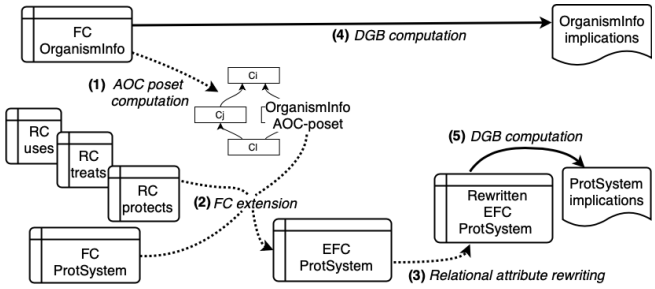

Fig. 2. Overview of the process for the running example. (1) The AOC-poset is built from FC *OrganismInfo*. (2) The EFC *ProtSystem* is built using the relational attributes $\exists r(C)$, where r is $uses$, $treats$ or $protects$, and C is a concept from the *OrganismInfo* AOC-poset. (3) The relational attributes are rewritten for easier reading. (4) and (5) The DGBs of implications are built for FC *OrganismInfo* and EFC *ProtSystem*.

3 Three Representations of the Datasets

This section presents the datasets and their combination through three representations splitting the data differently. Our objective is to assess the impact of the splitting on the form and the readability of the implications.

The datasets. The datasets concern 29 pest species belonging to 15 genera of the Noctuidae family [12]. To control these species on 15 crops (e.g. tomato, maize, cotton) belonging to seven families, 562 plant species, belonging to 352 genus and to 94 families, are identified.

The first dataset, which is an excerpt of PPAf, contains 721 protection systems, i.e. triplets *(plant, pest, crop*) describing the use of a plant to protect a crop against a Noctuidae pest at the species taxonomic level. The modeling interest of the Noctuidae family raises on the polyphagous or highly polyphagous nature of some of its pest species' diet. A polyphagous pest, such as *Trichilia pallida*, attacks crops from various genera of the same family, while a highly polyphagous pest, such as *Spodoptera frugiperda*, attacks crops from various families. In this first dataset, some publications do not specify the crop but the plant and the pest, mainly because of the polyphagous nature of the pest diet. To obtain a triplet in this case, a generic name was provided to the crop. Five generic species names were adopted, namely CropBrasS, CropFabaS, CropMalvS, CropPoacS, and CropS. The first four correspond to a crop attacked by a polyphagous pest, respectively from Brassicaceae, Fabaceae, Malvaceae, and Poaceae family. CropS corresponds to a crop attacked by a highly polyphagous pest. To the best of our knowledge, and to be as cautious as possible, we consider that they are all consumed by humans, and that only CropMalvS and CropS are used for medical care. Finally, in this PPAf excerpt, six organism species (e.g. pepper, chickpea, and castor bean) are both described as a crop and a protecting plant.

The second dataset is another excerpt of PPAf and informs on the plants used for medical care. This information was extracted for each protecting plant

and each crop listed in the first dataset. None of the pests is used for human and public health.

The last dataset, an excerpt of PAL, informs on the consumption of plants and crops by humans. This excerpt includes only plants and crops present in the first dataset. In this work, we consider that none of the pests are consumed by human or used in medical care.

The three representations. Combining the three datasets enables to representing SPSs that respect OHA. Three representations, leading to three different RCFs, were developed according to the reification of different entities and roles.

The *Relational* representation (Fig. 3a) considers five different entities. The three first represent the biological organisms. The first entity is *crop*. It is described using three attributes, i.e. $crSpecies$, $crGenus$, and $crFamily$, which respectively correspond to its species, its genus, and its family. The second entity is $Pest$, i.e. an aggressor of a crop. It contains three attributes, i.e. $peSpecies$, $peGenus$, and $peFamily$, which respectively correspond to its species, its genus, and its family. The third entity is $Plant$. Plants are described using three attributes, i.e. $plSpecies$, $plGenus$, and $plFamily$, which respectively correspond to its species, its genus, and its family. The fourth entity represents the protection systems ($ProtSystem$). $ProtSystem$ reifies the ternary relation linking $plSpecies$, $crSpecies$, and $peSpecies$. The last entity is $OrganismInfo$ in which each organism is described using its name at the species, genus, and family taxonomic levels using respectively the attributes *species*, *genus*, and $family$. In addition $OrganismInfo$ indicates whether the organism is consumed (attribute $food$) and whether it is used for medical care (attribute *medical*). $ProtSystem$ includes the data from PPAf knowledge set, and $OrganismInfo$ compiles the two other knowledge sets. The RCF for this representation is thus composed of five FCs ($ProtSystem$, $Plant$, $Crop$, $Pest$, $OrganismInfo$). Boolean attributes are obtained through a nominal scaling of the attributes [5]. The RCF also contains six RCs: *uses*, *protects*, *treats*, $pl\_CharactBy$, $cr\_CharactBy$, and $pe\_CharactBy$.

The $TwoTables$ representation (Fig. 3b) comports two entities. $ProtSystem$ and $OrganismInfo$ respectively represent the protection systems and the organisms, as in the *Relational* representation. This representation does not reify the role of the organisms in the protection systems, as does the *Relational* representation. The RCF for this representation is thus composed of two FCs (i.e. $ProtSystem$, $OrganismInfo$). The native (Boolean) attributes are obtained through a nominal scaling of the attributes. The RCF also contains three RCs: *uses*, *protects*, and *treats*.

The $OneTable$ representation (Fig. 3c) reifies protection systems in an entity named $CombinedSystem$. This entity includes the attributes of entities $Plant$, $Crop$, and $Pest$ of the *Relational* representation. It also contains the medical and food attributes related to the protecting plant and to the crop, respectively named $medicalPlant$, $foodPlant$, $medicalCrop$, and $foodCrop$. Additional attributes were included to express relationships between data not formalized by this representation. $plSpIsCrEw$ and $plGeIsCrEw$ indicate respec-

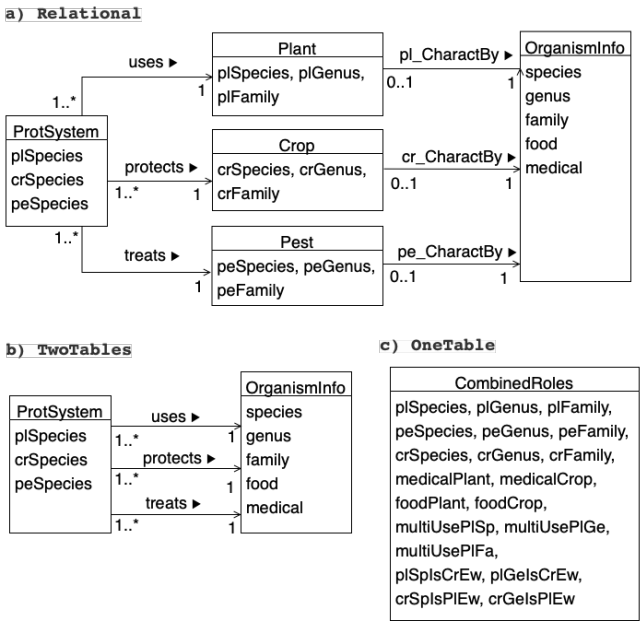

Fig. 3. Data model of the three datasets' representations.

tively that a protecting species is a crop species in another triplet, and a protecting genus is a crop genus in another triplet. The attributes $crSpIsPlEw$ and $crGeIsPlEw$ indicate respectively that a crop species is a protecting species in another triplet, and a crop genus is a protecting genus in another triplet. $multiUsePlSp$, $multiUsePlGe$, and $multiUsePlFa$ indicate whether the protecting plant, respectively at the species, genus, and family taxonomic levels, is both consumed and used for medical care. The RCF is here reduced to a single FC $CombinedSystem$, with attributes obtained by a nominal scaling of the $CombinedSystem$ entity attributes.

Table 3 presents the size of the different representations, in terms of number of objects and attributes, number of relational attributes, and size of the AOC-posets at the initial and at the last steps of RCA process.

4 Evaluation

This section presents (Sect. 4.1) and discusses (Sect. 4.2) the results obtained for the three data structures. The experiments were conducted using Cogui software platform[4], which includes Java implementations of RCA and LinCbO [8]. Running times for the Java LinCbO implementation remain below 3229 ms for

[4] http://www.lirmm.fr/cogui/

Table 3. Quantitative description of the RCFs and AOC-posets.

Representation	Formal context	#objects	#attributes	#relational attributes	#concepts AOC Poset (initial step)	#concepts AOC Poset (last step)
OneTable	FC CombinedRoles	721	1113	0	1005	1005
TwoTables	*All formal contexts*	1321	1078	2250	751	1750
	EFC ProtSystem	721	0	2250	1	1000
	FC OrganismInfo	600	1078	0	750	750
Relational	*All formal contexts*	1927	2169	3017	1507	2517
	EFC ProtSystem	721	0	767	1	1000
	EFC Plant	562	1008	750	700	705
	EFC Pest	29	45	750	36	37
	EFC Crop	15	38	750	20	25
	FC OrganismInfo	600	1078	0	750	750

Table 4. Implications (implic.) from the Duquenne-Guigues basis per scope (S) and maximum scope (Smax).

Representation	Formal context	#implic. S = 0	#implic. S = 1	#implic. S= 2	#implic. S = 3	#implic. S = [4-10] (avg)	#implic. S >10 (avg)	#total implic. S > 0	Smax (#implic.)
OneTable	FC CombinedRoles	3360	1105	281	125	236 (33.71)	173 (2.34)	1920	721 (1)
TwoTables	EFC ProtSystem	4891	827	234	95	165 (23.57)	74 (1.90)	1395	721 (1)
	FC OrganismInfo	6069	1007	76	37	42 (7.00)	6 (1.2)	1168	35 (1)
	All FCs	10960	1834	310	132	207	80	2571	
Relational	EFC ProtSystem	3414	825	234	95	164 (23.42)	73 (1.87)	1391	721 (1)
	EFC Plant	5698	1509	132	64	85 (14.17)	25 (2.08)	1815	87 (2)
	EFC Pest	855	67	8	0	4 (2)	1	80	29 (1)
	EFC Crop	740	58	8	4	0	0	70	3 (4)
	FC OrganismInfo	6069	1007	76	37	42 (7)	6 (1.2)	1168	35 (1)
	All FCs	16776	3466	450	208	295	105	4532	

the most complex case (relational data model), summing the running times for all the EFCs.

4.1 Analysis of the implications obtained for the 3 representations

In this section, we analyze the DGB of implications for the three representations (cf. Table 4). For each one, we present a quantitative and a qualitative analysis describing the main implication patterns, and provide selected examples. To consider implications applicable to OHA, we focus on the ones with scope > 0.

Implications in *Relational* representation

OrganismInfo. The DGB contains 1168 implications: 1007 are held by one object ($S = 1$) and thus are very specific. Four types of implications are observed. The first one informs about the uses in medical care and food care for a species, a genus, or a family, e.g. the Meliaceae are not consumed (with $S = 35$)[5]:

$$Family\_Meliaceae \implies Food\_$$

The second type gives more specific information about subsets of species in families and genus, e.g. the species of Annonaceae, which are not consumed, are also

[5] $Food\_X$ means *is consumed*; $Food\_$ means *is not consumed*, and similarly for $Medical\_X$ and $Medical\_$.

not used in medical care. The third type reflects taxonomy: a genus implies a family or a species implies a genus, e.g. Genus Salvia implies Family Lamiaceae (with $S = 18$):

$$Genus\_Salvia \implies Family\_Lamiaceae$$

The fourth implication type reveals data variety in the dataset. For instance species of Lythraceae family are not consumed and not used for care, and are exclusively from Genus Lythrum.

Crop. The DGB contains 70 implications. The small value of $Smax$ (3), indicates that the implications are rather specific. The implications focus on the role of the organisms as crop. A first implication type describes the taxonomy. They come from FC $OrganismInfo$. A second implication type describes the bijection between the taxonomic information encoded in FCs $Crop$ and $OrganismInfo$, as the attributes are duplicated in both contexts, completed by information on food and medical care if appropriate. For instance, the following implication (with $S = 2$) indicates that a crop belonging to family Fabaceae ($CrFamily\_Fabaceae$) is connected to the $OrganismInfo$ objects representing this family, and is also consumed and not used in medical care:

$$CrFamily\_Fabaceae \implies \exists cr\_CharactBy(Medical\_), \exists cr\_CharactBy(Food\_X), \exists cr\_CharactBy(Family\_Fabaceae)$$

A third implication type informs on the organisms role as crop, such as the following implication (with $S = 1$) named *Rel1*, which indicates that crops, used in medical care and not consumed, are from the Ricinus Communis species:

$$\exists cr\_CharactBy(Food\_), \exists cr\_CharactBy(Medical\_X) \implies CrSpecies\_RicinusCommunis, CrGenus\_Ricinus, CrFamily\_Euphorbiaceae, \exists cr\_CharactBy(Family\_Euphorbiaceae), \exists cr\_CharactBy(Species\_RicinusCommunis\&Genus\_Ricinus) \quad \textbf{(Rel1)}$$

The next example of implication (with $S = 1$), indicates that Malvaceae crops, not used in medical care, are restricted to Gossypium Genus and not consumed:

$$CrFamily\_Malvaceae, \exists cr\_CharactBy(Medical\_), \exists cr\_CharactBy(Family\_Malvaceae) \implies CrSpecies\_GossypiumHirsutum, CrGenus\_Gossypium, \exists cr\_CharactBy(Food\_), \exists cr\_CharactBy(Species\_GossypiumHirsutum\&Genus\_Gossypium)$$

Pest. The DGB contains 80 implications. Some implications reflect the taxonomy, already highlighted in $OrganismInfo$, and add no information for the experts. The Smax implication ($Smax = 29$) indicates that all pests are from the Noctuidae family, not consumed, and not used in medical care.

Plant. The DGB contains 1815 implications. Most of the implications (1509) hold for a single plant. As for crops and pests, the implications either reflect taxonomy or information about human consumption and medical care usage (restricted to organisms that play the role of protecting plant). Some other implications are true for protecting plants only, such as the following one, indicating that family Poaceae plants not used in medical care, are also not consumed (with $S = 2$):

$$PlFamily\_Poaceae, \exists pl\_CharactBy(Medical\_), \exists pl\_CharactBy(Family\_Poaceae) \implies \exists pl\_CharactBy(Food\_)$$

ProtectionSystem. The DGB contains 1391 implications, among which 566 held by more than one object. This result informs the expert on the numerous combinations of information existing in the datasets. The implication with $Smax = 721$, i.e. held by all objects, indicates that all systems treat Noctuidae. Within the 1391 implications, many implications types are present. They gather knowledge on the various roles of the organisms. We present some representative examples with diverse S values. The following implication (with $S = 380$), named *Rel2*, informs that when the studied protection systems treat Spodoptera Genus (Noctuidae Family), with a plant not consumed and not used in care, then the crop is used in medical care:

$\exists treats(PeFamily\_Noctuidae), \exists treats(PeGenus\_Spodoptera), \exists uses(pl\_CharactBy(Food\_)),$
$\exists uses(pl\_CharactBy(Medical\_)) \implies \exists protects(cr\_CharactBy(Medical\_X))$ ***(Rel2)***

The next implication (with $S = 8$) indicates that when studied protection systems treat Noctuidae Family with Genus Cymbopogon plants, then this is with Poaceae plants on consumed crops and the plants are used in medical care. Poaceae are also crops, and thus subject to implications for both roles:

$\exists treats(PeFamily\_Noctuidae), \exists uses(PlGenus\_Cymbopogon) \implies \exists uses(PlFamily\_Poaceae),$
$\exists protects(cr\_CharactBy(Food\_X)), \exists uses(pl\_CharactBy(Medical\_X))$

The next implication (with $S = 4$) indicates that when the protection systems protect Poaceae crops consumed and not used in medical care, to treat Noctuidae pests, using non consumed plants, then this is with Meliaceae plants used in medical care:

$\exists protects(CrFamily\_Poaceae), \exists treats(PeFamily\_Noctuidae), \exists uses(pl\_CharactBy(Food\_))$
$\exists protects(cr\_CharactBy(Food\_X)), \exists protects(cr\_CharactBy(Medical\_)),$
$\implies \exists uses(PlFamily\_Meliaceae), \exists uses(pl\_CharactBy(Medical\_X))$

Implications in *TwoTables* representation As the FC $OrganismInfo$ is similar to the one of *Relational*, it thus provides the same implication set. The DGB contains 1395 implications for the FC *ProtSystem*. This implication number is very similar to the one of the *ProtSystem Relational* representation. As an illustration, two implications are compared. The first one, *TT1*, focuses on the crop role:

$\exists protects(Food\_), \exists protects(Medical\_X), \exists treats(Family\_Noctuidae)$
$\exists treats(Food\_), \exists treats(Medical\_) \implies \exists protects(Family\_Euphorbiaceae),$
$\exists protects(Species\_RicinusCommunis\&Genus\_Ricinus), \exists treats(Genus\_Spodoptera),$
$\exists treats(Species\_SpodopteraLitura), \exists uses(Food\_), \exists uses(Medical\_), \exists uses(Family\_Asteraceae),$
$\exists uses(Species\_WollastoniaDentata\&Genus\_Wollastonia)$ ***(TT1)***

It is one of the 5 implications that mention *Ricinus Communis*. Compared to its *Relational* representation formulation, i.e. *Rel1*, it mixes information proper to *Ricinus Communis* as a crop with additional information on the protection systems, in particular the usage of *Wollastonia Dentata* as the protecting plant. In this case, implications of the *Relational* representation are easier to read, as they focus on organism roles. The second implication, *TT2*, is held by Noctuidae that are not consumed and not used in medical care:

$\exists treats(Food\_), \exists treats(Medical\_), \exists treats(Family\_Noctuidae), \exists treats(Genus\_Spodoptera),$
$\exists uses(Food\_), uses(Medical\_) \implies \exists protects(Medical\_X)$ ***(TT2)***

This information is not provided in its corresponding *Relational* representation formulation *Rel2* because it is not needed: in *Relational* representation it in-

deed appears in a separate and more precise way through the *Pest* implication indicating that Noctuidae are never consumed, nor used in medical care. *Rel2* is more focused and more synthetic.

Implications in *OneTable* representation The DGB contains 1920 implications. Its *Smax* value is identical to the one of FCs *ProtSystem* of $TwoTables$ and *Relational*. Its *S* value is lower, and the total number of implications is low (1920), compared to respectively 2571 and 4532 for $TwoTables$ and *Relational*. This representation thus provides less implications, but with a higher diversity of implication formulations. As illustration, let us consider the implication *OT1* that corresponds to *Rel1* and *TT1*:

$PeFamily\_Noctuidae, MedicalCrop\_X, FoodCrop\_ \implies PeGenus\_Spodoptera, FoodPlant\_,$
$MedicalPlant\_, MultiUsePlSp\_, MultiUsePlGe\_, MultiUsePlFa\_, PlSpIsCrEw\_, PlGeIsCrEw\_,$
$PlFamily\_Asteraceae, PeSpecies\_SpodopteraLitura, CrSpIsPlEw\_X,$
$CrGeIsPlEw\_X, PlSpecies\_WollastoniaDentata, PlGenus\_Wollastonia,$
$CrSpecies\_RicinusCommunis, CrGenus\_Ricinus, CrFamily\_Euphorbiaceae$ ***(OT1)***

The role has been encoded in the attribute name (e.g. $MedicalCrop$), rather than in the relations. Compared to *Rel1*, attributes about the protection system are included, e.g. $PeGenus\_Spodoptera$. Compared to both *Rel1* and *TT1*, additional attributes indicate multi-use purpose, e.g. that the crop *Ricinus Communis* is used elsewhere as a protecting plant ($CrSpIsPlEw\_X$), and the protecting plant *Wollastonia Dentata* is not used as a crop ($PlSpIsCrEw\_$). Another example is *OT2*, where roles appear as attributes rather than through relations:

$PeGenus\_Spodoptera, PeFamily\_Noctuidae, FoodPlant\_, MedicalPlant\_, MultiUsePlSp\_,$
$MultiUsePlGe\_, MultiUsePlFa\_ \implies MedicalCrop\_X$ ***(OT2)***

Information on food and medical care has not been encoded for pests in $OneTable$ representation to simplify, being identical for all Noctuidae. Compared to both *Rel2* and *TT2*, additional attributes complete the premise to indicate that the plant has no multiple uses, e.g. $MultiUsePlGe\_$.

4.2 Discussion

Lessons Learned. There is many taxonomic information in the implications, and some are duplicated in several tables. Although this duplication helps reading separately the implications (not considering several FCs at the same time), it complicates the reading of implications. Some other taxonomic information, such as indicating species, genus, and family may seem redundant too, as the latter two can be deduced from the species. Nevertheless, it may be useful for the readers who are not totally familiar with the taxonomy. In addition, some implications only precise the taxonomy, such as *species implies genus*. These implications could be automatically discarded, as they correspond to initial data encoding. Different settings of the implication formulation could be proposed to the user depending on the expected information.

Effect of Splitting the representations. As shown in Table 4, dividing data into separate FCs, which introduces RCs, produces more implications. This may be explained by the fact that, for a relation (e.g. *uses* in $TwoTables$), concepts

grouping target objects (e.g. $OrganismInfo$) induce concepts grouping source objects (e.g. $ProtSystem$) via the relational attributes. The implications include the result of this propagation schema. As a counterpart, in *Relational* representation, the implications are divided into coherent subsets, i.e. one per FC, simplifying their analysis. As the examples show, having few or no separate roles limits the relational attribute number and complexity. E.g., an advantage of $TwoTables$ over *Relational* may be that the $TwoTables$ implications contain one-level relational attributes (one RC), when the *Relational* ones contain relational attributes composing two RCs. But in return, when reducing the splitting, technical attributes, such as $plSpIsCrEw$ in $OneTable$ (plant species is crop elsewhere), have to be added to express multi-use and role, giving longer implications. This building and the formulation are not easier to understand by the expert. Moreover, as it has been highlighted by $TwoTables$ and $OneTable$, the less the dataset is split, the more the implications mix information. This situation occurs with organism roles and protection systems, that are mixed in the implications of $CombinedRole$ in $OneTable$, and of $ProtSystem$ in $TwoTables$.

Threats to Validity. With regard to internal validity, the Knomana knowledge sets have been manually collected by many participants but controlled by two domain experts that are co-authors of this paper. The software used in this evaluation, i.e. RCA algorithms implemented in Cogui and Conexp, have already been used in other case studies with validated results. LinCbO has been implemented in Java and inserted in the Cogui framework. To confirm the correctness of this implementation, the results have been compared with those of Conexp. Construct validity can be appreciated through the metrics and the qualitative analysis adopted to evaluate the effect of representation splitting. The metrics have been chosen in order to evaluate the feasibility in terms of structure size and implication number. The running time obtained thanks to LinCbO is very low. The obtained implications have been exhaustively examined, a task made easier by the substitution of the concept number by the seed attributes. Some recurring schemes and representative implications have been reported in the paper as a result of this analysis. Conclusion validity is concerned with the possibility of generalizing the observations. Knomana knowledge set has its own particularities, such as being organized around a ternary relation $Plant \times Pest \times Crop$ (protection system). Other secondary relations gravitate around this central relation. This has the effect of centralizing, for protection systems, the information coming from the other contexts. The implications reflect this organization. Using another dataset, not organized this way, conclusions may be different.

5 Related Work

Modeling complex data with the objective of extracting knowledge is part of the Knowledge Discovery and Data Mining processes (KDD) [3]. This issue is addressed in FCA through various encoding schemes and extensions, starting with conceptual scaling [5]. In the case of RCA, data modeling includes choosing

a kind of entity-relationship model with binary relationships and Boolean attributes. This requires deciding how data are separated in formal and relational contexts, and how to represent n-ary relations, e.g. ternary relations, a topic we studied in [10]. Life sciences data raise other issues, such as indeterminate species [9].

Association and Implication extraction is closely connected to FCA [1, 11]. Implications with premises restricted to one attribute, have been extracted from the result of RCA combined with AOC-posets [2]. More recently, M. Wajnberg et al. extracted implications together with RCA, using generators [15, 16]. The approach is applied to detect anomalies in manufacturing by aluminum die casting. The relational context is composed of machined parts, problems and the relation *generates* between parts and problems. Relational attributes and then concepts are built using the existential quantifier. Then in relational attributes, concepts are rewritten using their initial intent (the intent they had at their creation). This rewriting is made recursively.

In this paper, we build the DGB, and we use AOC-posets rather than concept lattices. We rewrite the relational attributes, as inspired by [15, 16], to analyze the implications. In addition, we compare several encodings of our data to investigate the impact of this encoding on the implication sets.

6 Conclusion

This paper explores the combination of RCA and the Duquenne-Guigues basis of implications on an environmental knowledge set in order to render knowledge suitable to experts. Our case study gathers information on plants that can replace synthetic pesticides and antibiotics, and be consumed or used in medical care. The guiding research question was to assess whether splitting the datasets could have a positive or negative impact on the implications' readability by the experts. We identified advantages of this splitting to enable the separate analysis of coherent, simpler, implication subsets, not mixing information types. This is strengthened by the relational attribute rewriting that makes the implication easier to read and to interpret.

As future work, we plan to evaluate the impact of using concept lattices and Iceberg rather than AOC-posets for building the implications, as well as using other quantifiers provided by RCA. We will analyze the complete Knomana knowledge base, which includes additional descriptors such as location and plant chemical compounds. Finally, we will post-process the implications. In particular, we plan to present implications by categories and order them by relevance, using standard metrics or metrics specific to the experts' questions. A preliminary work [14] investigates the potential of using patterns on implication premise and conclusion for categorizing the implications. These patterns are based on multi-valued attributes (before nominal scaling) describing species, genera and families, and on a 'meta-attribute' representing the presence of information on *medical* or $food$.

Acknowledgments. This work was supported by the French National Research Agency under the Investments for the Future Program, referred as ANR-16-CONV-0004.

References

1. Bertet, K., Demko, C., Viaud, J.F., Guérin, C.: Lattices, closures systems and implication bases: A survey of structural aspects and algorithms. Theoretical Computer Science **743**, 93–109 (2018)
2. Dolques, X., Ber, F.L., Huchard, M., Grac, C.: Performance-friendly rule extraction in large water data-sets with AOC posets and relational concept analysis. Int. J. General Systems **45**(2), 187–210 (2016)
3. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: The KDD process for extracting useful knowledge from volumes of data. Commun. ACM **39**(11), 27–34 (1996)
4. Frank, D.: One world, one health, one medicine. The Canadian Veterinary Journal **49**(11), 1063–1065 (2008)
5. Ganter, B., Wille, R.: Formal Concept Analysis - Mathematical Foundations. Springer (1999)
6. Guigues, J.L., Duquenne, V.: Famille minimale d'implications informatives résultant d'un tableau de données binaires. Math. et Sci. Hum. **24**(95), 5–18 (1986)
7. 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)
8. Janostik, R., Konecny, J., Krajča, P.: Pruning techniques in LinCbO for computation of the Duquenne-Guigues basis. To appear in ICFCA 2021 (2021)
9. Keip, P., Ferré, S., Gutierrez, A., Huchard, M., Silvie, P., Martin, P.: Practical comparison of FCA extensions to model indeterminate value of ternary data. In: CLA 2020. CEUR Works. Proc., vol. 2668, pp. 197–208 (2020)
10. Keip, P., Huchard, M., Ber, F.L., Sarter, S., Silvie, P., Martin, P.: Effects of Input Data Formalisation in RCA for a Data Model with a Ternary Relation. In: ICFCA 2019,. LNCS, vol. 11511, pp. 191–207. Springer (2019)
11. Kuznetsov, S.O., Poelmans, J.: Knowledge representation and processing with formal concept analysis. Wiley Interd. Rev. Data Min. Knowl. Disc. **3**(3), 200–215 (2013)
12. Martin, P., Gutierrez, A., Marnotte, P., Huchard, M., Keip, P., Mahrach, L., Silvie, P.: Dataset on noctuidae species used to evaluate the separate concerns in conceptual analysis: Application to a life sciences knowledge base (2021). https://doi.org/10.18167/DVN1/HTFE8T
13. Martin, P., Silvie, P., Sarter, S.: Knomana - usage des plantes à effet pesticide, antimicrobien, antiparasitaire et antibiotique (patent APP IDDN.FR.001.130024.000.S.P.2019.000.31235) (2019)
14. Saoud, J., Gutierrez, A., Huchard, M., Marnotte, P., Silvie, P., Martin, P.: Explicit versus Tacit Knowledge in Duquenne-Guigues Basis of Implications: Preliminary Results. In: Workshop Analyzing Real Data with Formal Concept Analysis (RealDataFCA@ICFCA'2021, Strasbourg, June, 29. To appear (2021)
15. Wajnberg, M.: Analyse relationnelle de concepts : une méthode polyvalente pour l'extraction de connaissance. Ph.D. thesis, Université du Québec à Montréal (2020)
16. Wajnberg, M., Valtchev, P., Lezoche, M., Massé, A.B., Panetto, H.: Concept analysis-based association mining from linked data: A case in industrial decision making. In: Proc. of the Joint Ontology Works. 2019 Episode V: The Styrian Autumn of Ontology. CEUR Workshop Proceedings, vol. 2518. CEUR-WS.org (2019)