Identifying explicit and tacit knowledge in a life science knowledge base

Johanna Saoud, Alain Gutierrez, Marianne Huchard, Pierre Silvie, Pierre Martin

Identifying explicit and tacit knowledge in a life science knowledge base

Johanna Saoud[1,2], Alain Gutierrez[2], Marianne Huchard[2] , Pierre Silvie[3,4], and Pierre Martin[3]

1 : Master Sciences et Numérique pour la Santé, parcours Bioinformatique, Connaissances, Données, Université de Montpellier
2 : LIRMM, Université de Montpellier, CNRS, Montpellier, France
3 : CIRAD, UPR AIDA, F-34398 Montpellier, France
AIDA, Université de Montpellier, CIRAD, Montpellier, France
4 : PHIM Plant Health Institute, Université de Montpellier, IRD, CIRAD, INRAE, Institut Agro, Montpellier, France

Introduction

@Cirad - Pierre Martin

An alternative of the use of synthetic pesticides and antibiotics in agriculture is to spray local plants extracts, in aqueous or essential oil form. To this end, the Knomana (KNOwledge MANAgement on pesticidal plants in Africa for a safer food and a better environmental health) knowledge base [1] compiles various knowledge sets on plant use such as the 42000 descriptions of pesticidal plant uses for plant, animal, and public health presented in the literature. As the One Health approach dictates to be aware of the additional uses of these pesticidal plants to prevent their unintended effects on the animal, the human, and their environment, the challenge for the domain experts (e.g. entomologist, pathologist) is thus to identify the pesticidal plants in Knomana considering the One Health approach.

With the aim to present knowledge to the expert using a compact and comprehensive formalism, in [2], we computed the Duquenne-Guigues basis (DGB) of implications on an excerpt of Knomana, in which each plant is described using its taxonomy (i.e. species, genus, and family), to be consumed as food, and to be used in medical care.

The DGB method is based on Formal Concept Analysis (FCA) and provides a cardinality-minimal set of non-redundant implications. This poster describes the product line that formulates Knomana knowledge on pesticidal plants as implications, from which the implicit knowledge elements were removed and the side effects are highlighted to alert the expert. As an illustration, this poster presents the implications on *Spodoptera frugiperda*, a highly polyphagous insect that is close to invade South of Europe.

A perspective of this work is to identify pesticidal European plants species that share chemical components similarities with plants used to control this pest in its native area.

https://www.ecoco2.com/blog/le-4eme-plan-sante-environnement-en-construction/

Data

To be aware of the additional uses of pesticidal plants, 3 data sets were extracted from Knomana to conduct this work, i.e. indication on the consumption of pesticidal plant as food or drink, indication on the use of pesticidal plant as medical care, and the use of pesticidal plant to protect a crop against a pest. Fig 1 presents the data model resulting from the grouping of these 3 datasets.

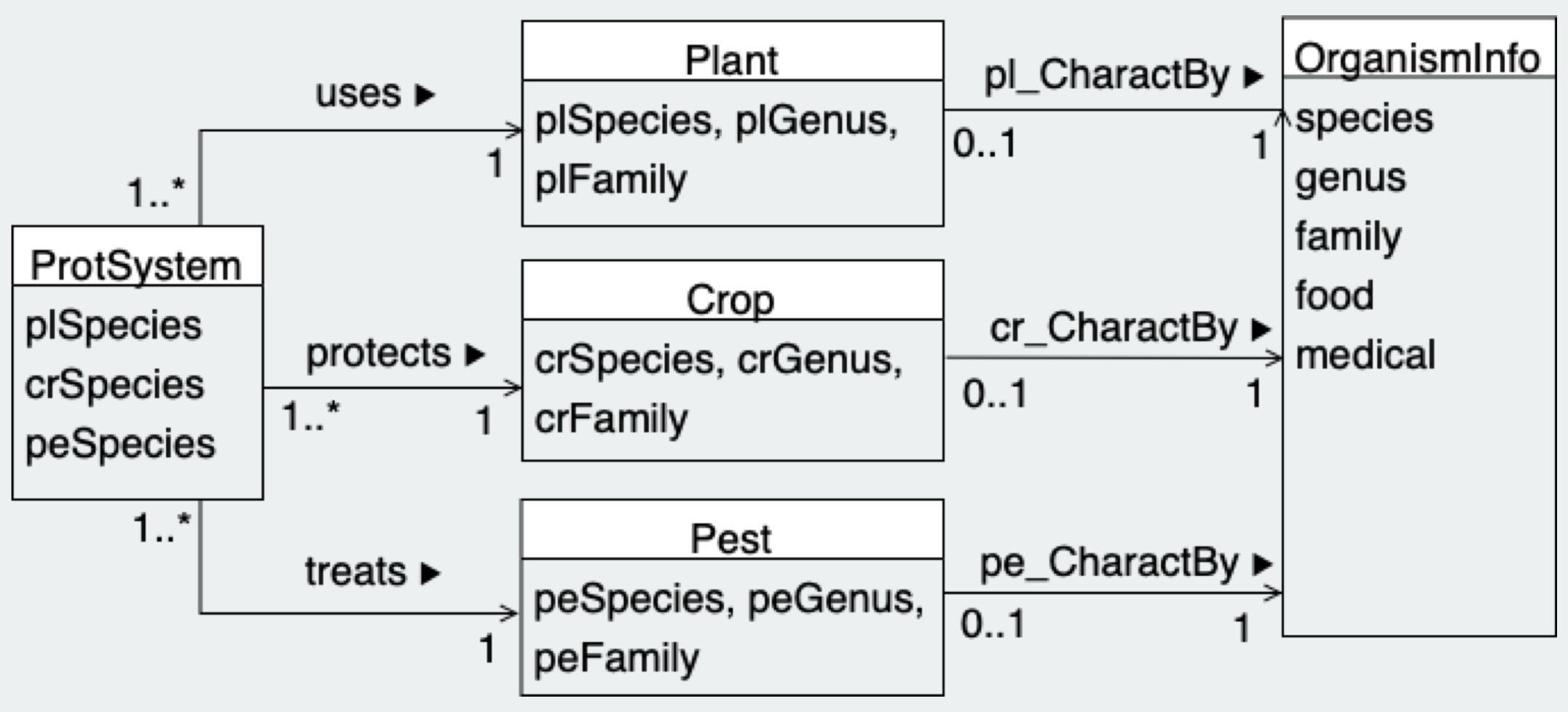


Fig 1: Data model resulting from the the grouping of the 3 datasets

Background

FCA [4] is a mathematical framework based on lattice theory which aims to group objects and their attributes as concepts and order them relatively. While FCA handles a Boolean data table (called a Formal Context), its extension devoted to the discovery of knowledge within this kind of data model is the Relational Concept Analysis (RCA). To proceed, RCA builds the concepts and orders them relatively taking into account the relations existing between the formal contexts.

As an alternative to compute lattice, FCA comes with the DGB of implication method.

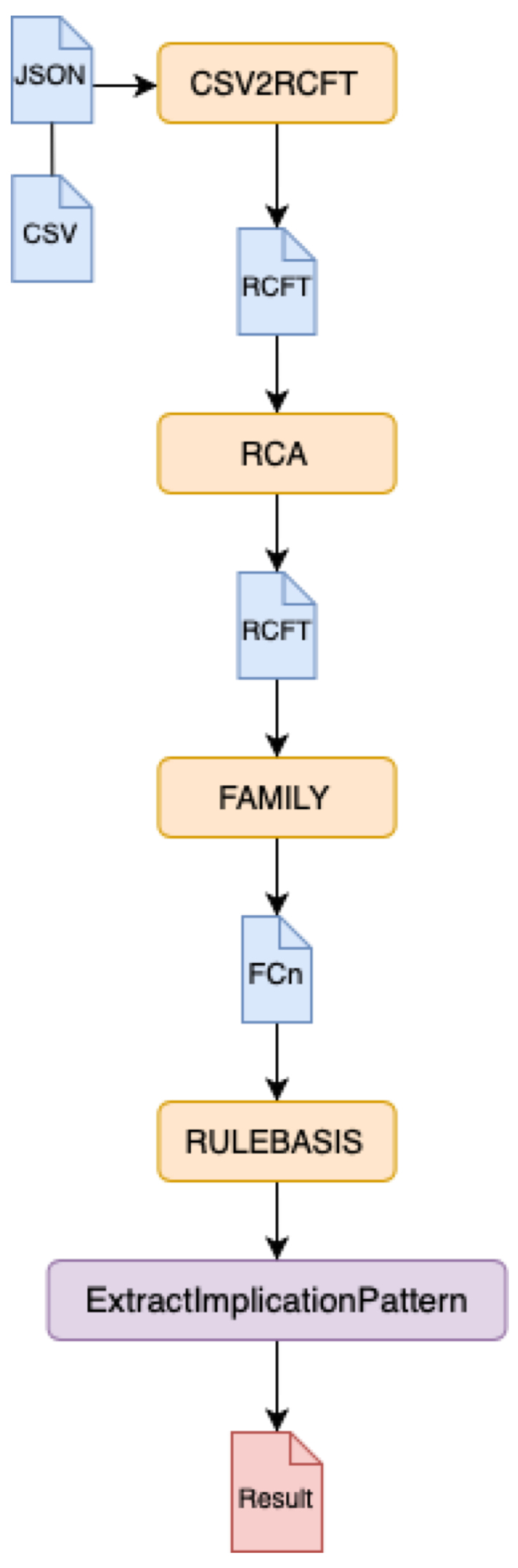
About implications :

An implication A implies B where A and B are attribute sets holds when the set of objects owning A attributes also own B attributes.

In this work, we will call scope of an implication the number of objects owning the attributes of A.

As the DBG of implications computation theory is not yet developed for RCA, its adaptation to RCA consisted in computing the DGB of implications for each formal context using the final ordering of the concepts by RCA, that conducts to add relational concepts within the initial formal contexts.

Process



- **Step 1:** Using a JSON file referencing multivalued .csv files an entry, the CSV2RCFT function builds the formal and relational contexts and stores them in an RCFT file.
- **Step 2**: The RCA function uses the RCFT file as input to compute the relational contexts. As output, it creates an extended RCFT file (RCFTextend) containing the original formal and relational contexts enriched with new relational attributes.
- **Step 3:** The FAMILYexport function extracts the formal contexts from the RCFTextend file and stores them into a new file.
- **Step 4:** The RULEBASIS function computes the Duquenne-Guigues basis of implications for each formal context.
- **Step 5:** The ExtractImplicationPattern function computes the implication patterns.

The functions of steps 1 to 4 are provided by the library FCA4J, from Cogui software (http://www.lirmm.fr/fca4j/).

Example of Results

The implication patterns were computed for each formal context. Table 1 presents the ones obtained from OrganismInfo. A pattern is a combination of the symbols S, F, P, and p corresponding respectively to a Species, a Genus, a Family, and a plant property (i.e. use).

ID	Premise	Conclusion	knowledge elements	Example of implication	#implications	Max scope
1	F	p	KU	F_Meliaceae ⇒ no-food	35	35
2	G	F	KT	G_Salvia ⇒ F_Lamiaceae	12	18
3	Fp	p	KU	no-food,F_Annonaceae ⇒ no-medical	10	16
4	G	Fp	KU, KT	G_Trichilia ⇒ no-food,F_Meliaceae	84	10
5	Fp	G	KU, KD	food,medical,F_Rutaceae ⇒ G_Citrus	6	5
6	F	G	KD	F_Piperaceae ⇒ G_Piper	1	5
7	GFp	p	KU, KT	medical,F_Asteraceae,G_Artemisia ⇒ no-food	7	4
8	Fp	Gp	KU, KD	medical,F_Annonaceae ⇒ food,G_Annona	1	3
9	F	Gp	KU, KD	F_Lythraceae ⇒ no-food,no-medical,G_Lythrum	5	2
10	S	GFp	KU, KT	S_ZygophyllumAlbum ⇒ no-food,no-medical,G_Zygophyllum,F_Zygophyllaceae	600	1

Table 1: implications pattern identified from the OrganismInfo data structure

The analysis of the patterns of Table 1 enabled to identify 3 types of knowledge elements: knowledge on plant use at diverse taxonomy levels when p is present (KU), knowledge on plant taxonomy (KT) when S or G is present in the premise and G or F is present in the conclusion (resp.), and side effect of the knowledge set (KD) when F or G is present in the premise and G or S in the conclusion (resp.). Effectively KD is not in accordance with taxonomic referential, e.g. a family may contain more that one species, and thus informs on the extend of knowledge inserted in Knomana. KD corresponds therefore to tacit knowledge. Moreover, plant taxonomy is known by the experts.

Removing KT from the implications eases the implication reading but makes KT tacit knowledge. As an illustration, the implication eq 1 (support 7) was extracted from ProtSystem as it dealt with the pest *Spodoptera frugiperda*. The removal of KT from eq 1 provided eq 2. The latter can be interpreted as follow: "Protecting *Zea mays* (a plant consumed as food and used in medical care) against a pest from the genus Spodoptera using a Meliaceae (a plant not consumed and not used in medical care) implies that the treated pest species is *Spodoptera frugiperda".*

protects(CrFamily_Poaceae),protects(CrSpecies_ZeaMays&CrGenus_Zea),treats(PeFamily_Noctuidae),treats(PeGenus_Spodoptera),uses(PlFamily_Meliaceae),protects(cr_CharactBy(no-food)),protects(cr_CharactBy(medical)),uses(pl_CharactBy(food)),uses(pl_CharactBy(no-medical)) => treats(PeSpecies_SpodopteraFrugiperda) *(eq 1)*

protects(CrSpecies_ZeaMays),treats(PeGenus_Spodoptera),uses(PlFamily_Meliaceae),protects(cr_CharactBy(food)),protects(cr_CharactBy(medical)),uses(pl_CharactBy(no-food),uses(pl_CharactBy(no-medical)) => treats(PeSpecies_SpodopteraFrugiperda) *(eq 2)*

References

[1] Silvie P.J., Martin P., Huchard M., Keip P., Gutierrez A., and Sarter S. Prototyping a knowledge based system to identify botanical extracts for plant health in sub-Saharan Africa. Plants, 10(5), 2021

[2] Saoud J., Gutierrez A., Huchard M., Marnotte P., Silvie M., and Martin P. Explicit versus Tacit Knowledge in Duquenne-Guigues Basis of Implications: Preliminary Results. Analyzing Real Data with Formal Concept Analysis, RealDataFCA'2021, an ICFCA workshop https://icfca2021.sciencesconf.org/., pp. 6, 2021

[3] Mahrach L., Gutierrez A, Huchard M, Keip P, Marnotte P, Silvie P, Martin P. Combining Implications and Conceptual Analysis to Learn from a Pesticidal Plant Knowledge Base. To appear in Proceedings of International Conference on Conceptual Structure (ICCS), pp. 15, 2021.

[4] Ganter B., Wille R. Formal Concept Analysis - Mathematical Foundations. Springer 1999, ISBN 978-3-540-62771-5, pp. I-X, 1-284