



**HAL**  
open science

# Improving Attribute Exploration for the Detection and Correction of Anomalies in an Agroecological Knowledge Base

Nassif Saab, Marianne Huchard, Pierre Martin

► **To cite this version:**

Nassif Saab, Marianne Huchard, Pierre Martin. Improving Attribute Exploration for the Detection and Correction of Anomalies in an Agroecological Knowledge Base. JOBIM 2022 - 22es Journées Ouvertes en Biologie, Informatique et Mathématiques, Jul 2022, Rennes, France. hal-03725155

**HAL Id: hal-03725155**

**<https://hal.science/hal-03725155v1>**

Submitted on 15 Jul 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



# Improving Attribute Exploration for the Detection and Correction of Anomalies in an Agroecological Knowledge Base

Nassif Saab<sup>1</sup>, Marianne Huchard<sup>1</sup> and Pierre Martin<sup>2</sup>

<sup>1</sup>LIRMM, Univ Montpellier, CNRS <sup>2</sup>CIRAD, UPR AIDA, Montpellier

## Attribute Exploration (AE)

- ▶ an interactive knowledge acquisition method of Formal Concept Analysis [1] aimed at finding dependencies between attributes [2].
- ▶ accepts data tables describing binary relationships between objects and a fixed set of attributes.

Attribute numbering: 6 5 4 3 2 1

	6	5	4	3	2	1
Objects	WdH	Fly	SH	WtH	Mg	RB
silver-gull		X	X	X		X
little-tern		X	X	X	X	
woodpecker	X	X				
giant-otter			X	X		
arctic-tern		X	X	X	X	X

**Attributes**

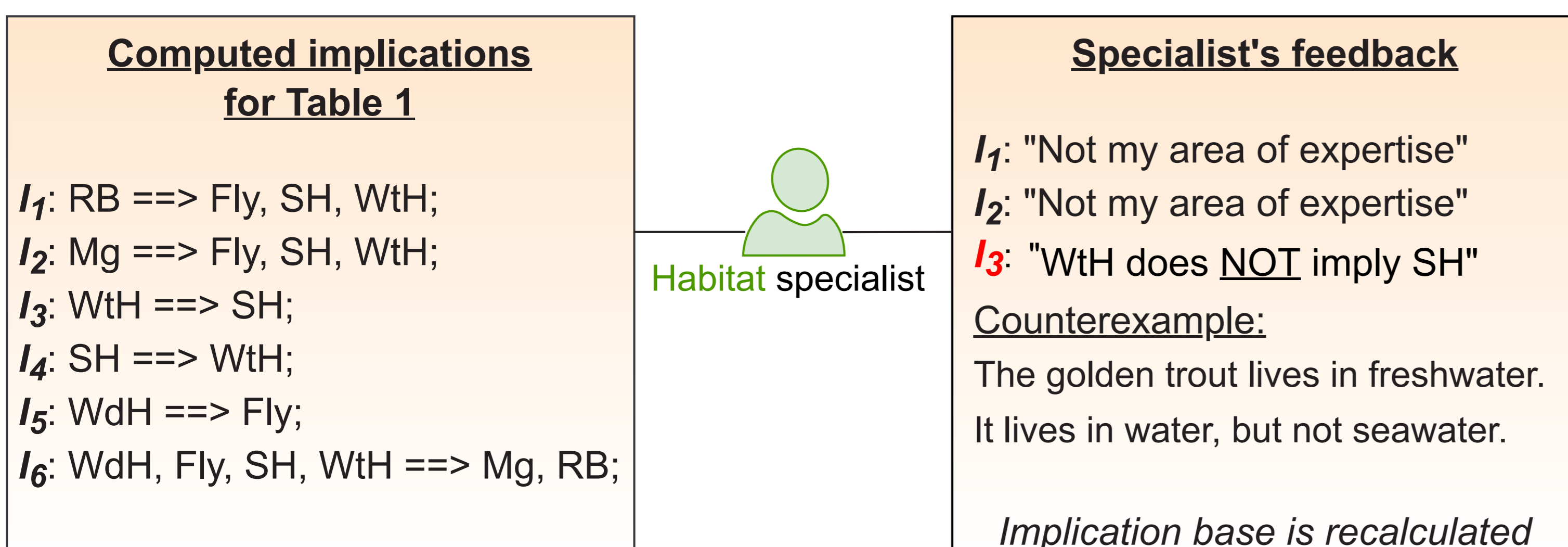
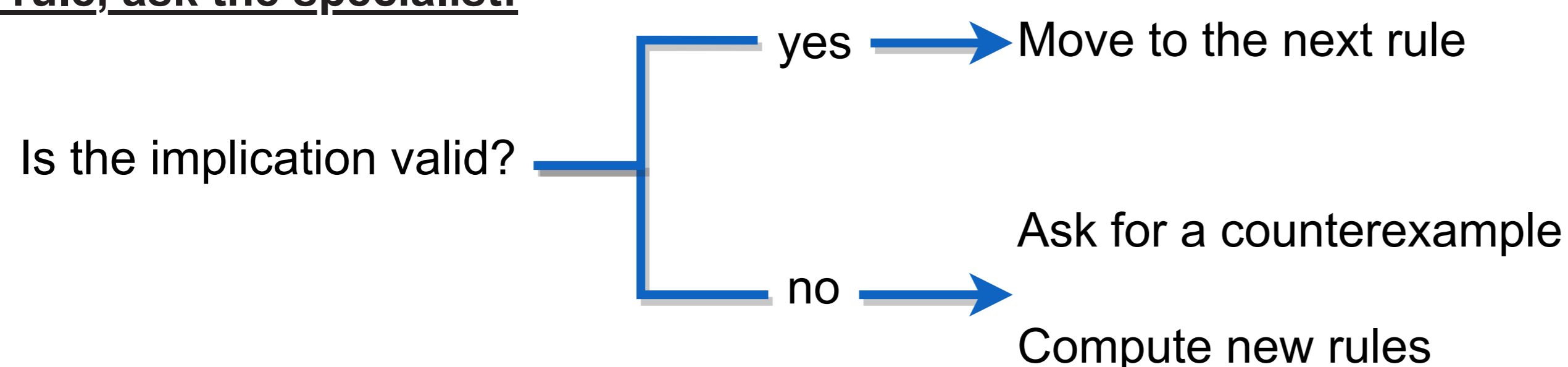
RB: red-bill  
Mg: migratory  
WtH: water-habitat  
SH: sea-habitat  
Fly: flies  
WdH: wood-habitat

X indicates incorrect information

Table 1: An example of formal context on the characteristics of five animals.

- ▶ relies on the computation of the Duquenne-Guigues basis [3], a minimal set of implication rules from which all implications can be inferred.
- ▶ implemented by tools like ConExp [4] as follows:

For each rule, ask the specialist:



## The problem with the rule display order

- ▶ ConExp's AE displays rules consecutively in the lexic order of attribute sets. Accordingly, set A is presented before set B if:

$$\min((A \cup B) \setminus (A \cap B)) \in B$$

In the premise of  $I_5$ , {WdH} is represented as  $A = \{6\}$ .

In the premise of  $I_6$ , {WdH, Fly, SH, WtH} is represented as  $B = \{3, 4, 5, 6\}$ .

$I_5$  is presented before  $I_6$ , because the smallest differing element  $3 \in B$ .

- ▶ The lexic order is not meaningful to specialists because it does not regard their interests, e.g. **habitat**. Consequently, specialists may have to endure questions outside their areas of expertise, e.g.  $I_1, I_2$ . With large numbers of attributes, the AE process becomes unproductive and time-consuming.

## A solution to the current problem

- ▶ To display the rules in an order that considers the nature of the data, we propose rearranging attributes in conformity with the definition of the lexic order and the numbering of attributes.
- ▶ Shifting attributes to the right in Table 1 gives them display priority during ConExp's AE. This allows us to present the computed implication rules in a new descending order of relevance for each specialist.

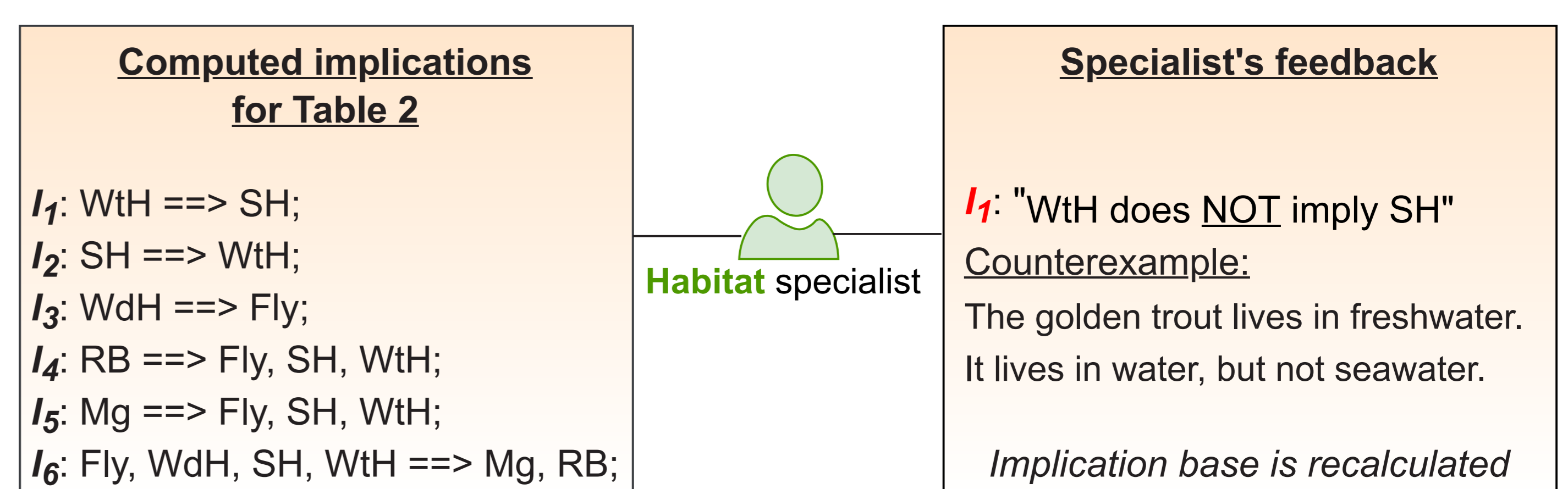
## Sorting the data & Rearranging the attributes

To present all the implication rules within the specialist's area of expertise first, attributes are grouped into categories, e.g. **habitat**, and reordered by their relevance to the specialist.

	6	5	4	3	2	1
	Fly	Mg	RB	WdH	SH	WtH
silver-gull	X		X		X	X
little-tern	X	X			X	X
woodpecker	X			X		
giant-otter					X	X
arctic-tern	X	X	X		X	X

Grouped attributes about the **habitat** (WdH, SH, WtH) are shifted to the right.

Table 2: The same formal context post data manipulation.

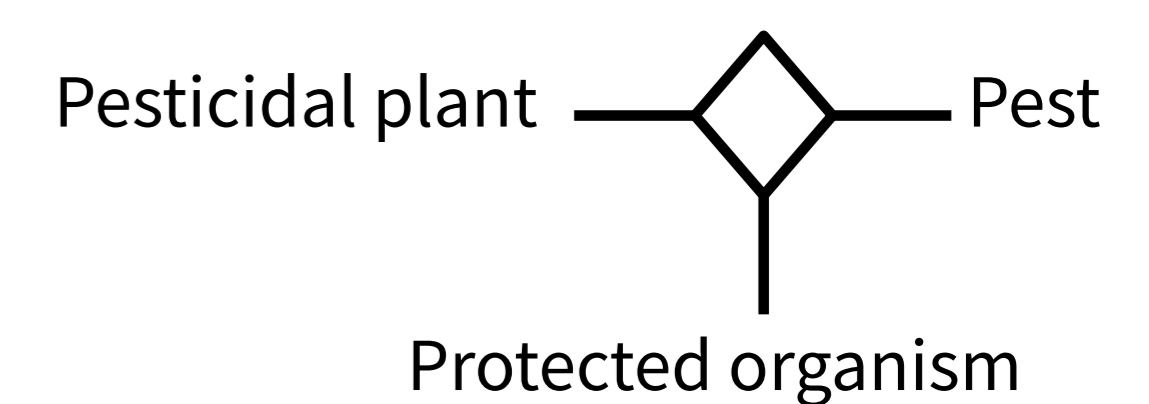


By accommodating AE to the interest of the specialist, the anomaly detection and correction process becomes more efficient.

## Applying AE to Knomana

### Knomania [5]

- an agroecological knowledge base
- a relational data model
- over 45,000 descriptions on plant use
- thousands of computed implication rules



### AE to detect and correct anomalies

Latin name (plant)	Botanic family	Used part of the plant
Salvia officinalis	Lamiaceae	leaf, flower
Laurus nobilis	ID 45444	leaf
Myrtus communis	Myrtaceae	leaf
Citrus aurantium	Rutaceae	

Annotations: Incorrect spelling (Lamiaceae), Incorrect value type (ID 45444), Missing information (empty cell).

Table 3: An example of three types of anomalies found in Knomana.

The detection and correction of anomalies can be done by:

1. Rejecting an incorrect implication rule during AE.
2. Identifying the culprit among the objects with all the attributes of the rule.
3. Providing a correction for this object.

**Requirements:** an extension of improved AE for Relational Concept Analysis [6] to handle Knomana's ternary relationships. RCA is an extension of FCA intended for entity-relationship data models.

## References

- [1] Bernhard Ganter and Rudolf Wille. Formal Concept Analysis. Springer Berlin Heidelberg, 1999.
- [2] Bernhard Ganter and Sergei Obiedkov. Conceptual Exploration. Springer Berlin Heidelberg, 2016.
- [3] J. L. Guigues and Vincent Duquenne. Famille minimale d'implications informatives résultant d'un tableau de données binaires. Math. et Sci. Hum., 24(95):5–18, 1986.
- [4] Serhiy A. Yevtushenko. Conexp, 2022.
- [5] Pierre J. Silvie, Pierre Martin, Marianne Huchard, Priscilla Keip, Alain Gutierrez, and Samira Sarter. Prototyping a knowledge-based system to identify botanical extracts for plant health in sub-saharan africa. Plants, 10(5), 2021.
- [6] Mohamed Rouane-Hacene, Marianne Huchard, Amedeo Napoli, and Petko Valtchev. Relational concept analysis: mining concept lattices from multi-relational data. Annals of Mathematics and Artificial Intelligence, 67(1):81–108, 2013.

## Acknowledgments

Funded by Key Initiatives MUSE DATA & LIFE SCIENCES through an interdisciplinary internship grant.