


Université de Montpellier - Faculté des sciences

Navigation dans les règles d'implication multidimensionnelles sur l'agroécologie pour l'aide à la décision en santé animale et végétale

Master 2 Informatique, Parcours ICo

Stage effectué au LIRMM

par Musslin Lola

du 15/02/23 au 28/08/23

Tuteurs organisation
Marianne Huchard, Arnaud Sallaberry, Vincent Raveneau,
Pierre Martin, Alexandre Bazin, Pascal Poncelet

Tuteur université
Lucas Isenman

Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre de France 2030 portant la référence ANR-16-CONV-0004.

Table des matières

1 Introduction

1.1 Contexte du stage

Ce rapport relate le stage que j'ai effectué dans le cadre de mes études à l'Université de Montpellier en Master Intégration de Compétence en informatique (ICo), suite à la licence en biologie des écosystèmes. Il se situe donc naturellement à l'interface entre ces deux disciplines. J'ai choisi ce stage car, en étudiant l'informatique, mon ambition est de pouvoir développer des outils utiles au domaine de la biologie.

Dans le contexte actuel de changement climatique et de prise de conscience de l'impact de l'Homme sur l'environnement, la possibilité de pouvoir utiliser des plantes pour remplacer des pesticides de synthèse représente une solution plus durable. Ces plantes peuvent être utilisées de différentes manières, et ont généralement une action répulsive de plus longue durée, là où antibiotiques et pesticides ont une action immédiate sur les bioagresseurs. Le choix d'une plante pesticide doit cependant prendre en compte de nombreux facteurs tels que l'espèce de bioagresseur à cibler, ainsi que celle(s) à protéger, tout en considérant le fait qu'une plante bénéfique pour un organisme peut être toxique pour un autre (e.g. toxicité pour le technicien qui pulvérise l'extrait de plante pesticide sur une culture agricole).

Knomana [1] (Knowledge management on pesticides plants in Africa) est une base de connaissances qui regroupe des descriptions d'usage de ces plantes pesticides. L'objectif de ce stage est de se concentrer sur la manière de visualiser les données contenues dans cette base de connaissances. Il existe déjà des travaux portant sur la navigation et l'exploration de la base de Knomana [KOH+19]. La spécificité du travail, présenté dans ce mémoire, est que les données de Knomana sont d'abord traduites sous forme de règles d'implication avant d'être visualisées. Les règles d'implication ont l'avantage d'être textuelles, et sont par conséquent plus directement intelligibles pour un biologiste. Ce choix implique cependant le besoin de disposer d'une application complémentaire à celles qui visualisent les données sous forme de classifications conceptuelles.

1.2 Présentation de la structure d'accueil

Ce stage a été effectué au LIRMM (Laboratoire d'Informatique, de Robotique et de Micro-électronique de Montpellier), et a été encadré par les équipes Advanse [2], MaREL [3], et WEB3 [4] du LIRMM, ainsi que par un chercheur du CIRAD. L'équipe Advanse est spécialisée dans la fouille de données, l'apprentissage automatique, et la visualisation analytique de données complexes, qui est la thématique principale de ce stage. L'application développée s'inscrit dans le but général de visualiser et d'exploiter des données exprimées sous forme de règles d'implication, et en particulier celles de Knomana développé par l'unité de recherche AIDA du CIRAD dans laquelle travaille Pierre Martin [5].

Ce stage est financé par l'Institut de Convergence en Agriculture Digitale #Digitag [6]. Il fait suite au stage réalisé par Guilhèm Blanchard en 2022, portant sur la première version de RCAvizIR [GHSM22], et celui réalisé par Emile Muller en 2021, intitulé RCAViz : Visualizing and Exploring Relational Conceptual Structures [MHM+22] dont l'application est accessible en ligne [7].

1. https://agents.cirad.fr/Pierre+Martin/Knomana#
2. https://www.lirmm.fr/teams-en/advanse/
3. https://www.lirmm.fr/teams-en/marel/
4. https://www.lirmm.fr/teams-en/web3/
5. https://agents.cirad.fr/Pierre+Martin/Knomana
6. https://www.hdigitag.fr/fr/
7. https://rcaviz.lirmm.fr/

1.3 Présentation globale du stage

L'objectif plus précis du stage est de concevoir et de développer une application web permettant à un utilisateur de visualiser des règles d'implication multidimensionnelles produites à partir de données extraites de Knomana, afin d'identifier des connaissances d'intérêt (exprimées sous forme de règle).

1.4 Organisation du mémoire

Dans la section 2 de ce rapport, je présente le contexte du projet, en particulier, la provenance des données et la création des règles d'implication à visualiser. Dans la section 3, je présente les besoins de l'utilisateur, ainsi que les choix de conception y répondant. Dans la section 4, je présente la réalisation concrète de l'application. La section 5 revient sur la manière dont ce projet a été mené, en particulier sur l'organisation du travail, le planning et la communication autour du projet. Enfin, je conclus ce mémoire dans la section 6 et indique des éléments à ajouter dans l'application.

2 Contexte du projet

Dans cette section, je présente les données à visualiser par l'application développée durant ce stage. Je commencerai par parler de l'origine des données brutes en section 2.1 avant de décrire en section 2.2 les différents traitements qui leur sont appliqués pour obtenir les règles d'implication (section 2.3) que l'on cherche à visualiser.

2.1 Origine des données

Les règles d'implication, que nous utiliserons en entrée de l'application développée, sont produites à partir de données agroécologiques extraites de Knomana [MST+18, SMH+21]. Knomana recense des informations sur des utilisations de plantes à effet pesticide pour lutter contre des organismes bioagresseurs et protéger des cultures agricoles. Ces données se présentent sous la forme d'un tableau Microsoft Excel®. Chaque ligne du tableau (voir Figure 2) correspond à la description de l'utilisation d'une plante dans un cas donné (e.g. un pays voire une localité, une espèce de bioagresseur, une culture protégée, etc.). Les colonnes représentent les 35 attributs qui décrivent le cas d'utilisation, e.g. la plante à effet pesticide utilisée, son mode de préparation, les cultures qu'elle protège, les organismes qu'elle impacte et sa méthode d'utilisation. La figure 1 présente un extrait d'une ligne de données de Knomana.

		Plante	Informations plante (caractéristiques, localité)				Ravageurs, vecteurs, maladies ou auxiliaires		Mode d'emploi de l'extrait dans les essais biologiques **(LABO)**		
ID	N° espèce botanique	*Nom latin* plante	Nom (vernaculaire, français, autre)	Famille botanique	Forme d'utilisation (poudre, huile = huile essentielle)	Organisme à protéger (espèce d'animal)	Stade visé ou type d'organisme cible	*Nom latin*	Extrait ou HE employé seul ou/et associé à d'autres composantes; Fractions éprouvées	Modalité d'application vis à vis de l'organisme-cible	Dose appliquée: (µl/ml; µl/cm²; ml/ml; ppm)
35465		*Azadirachta indica*	Neem	Meliaceae	Huile	Maïs	Insectes	*Sitophilus zeamais*		Contact-Ingestion	3ml/kg

FIGURE 1 – Aperçu de quelques colonnes descriptives de l'utilisation d'une plante dans Knomana

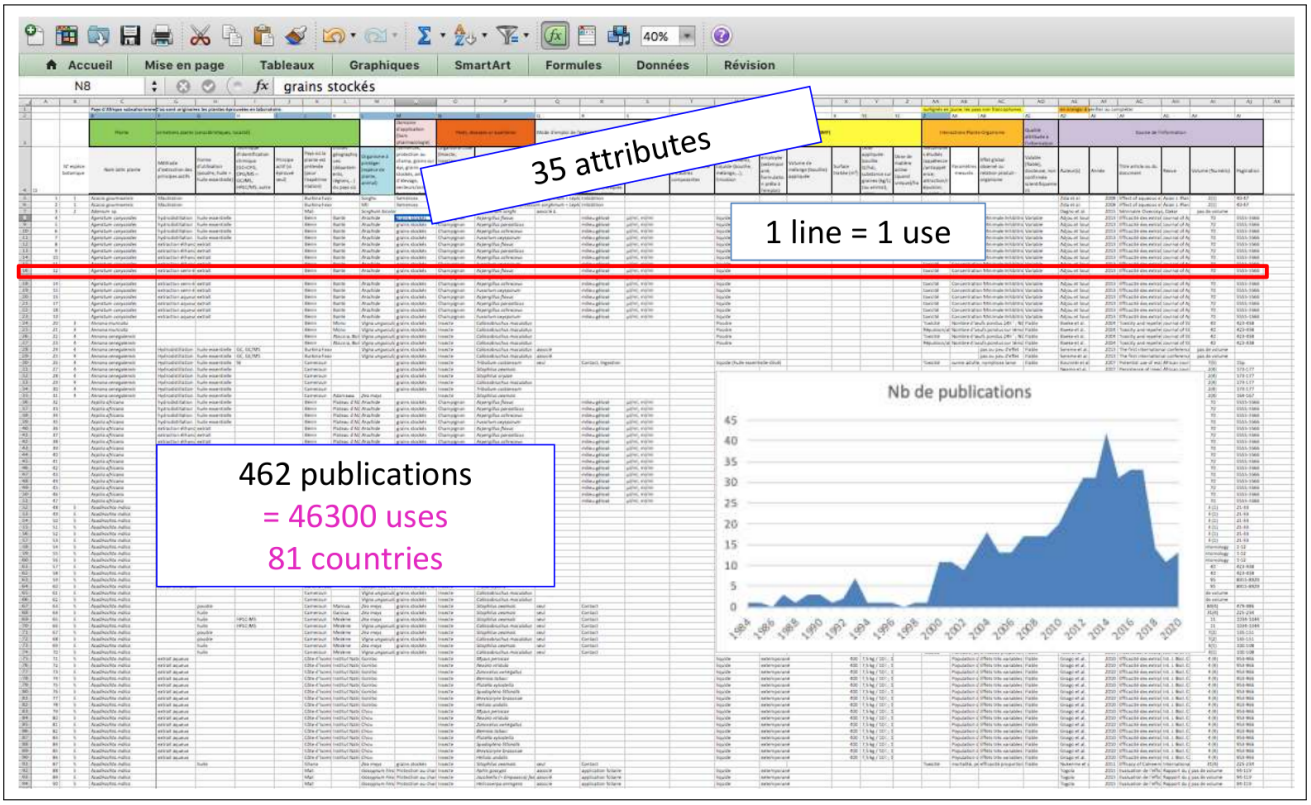

FIGURE 2 – Aperçu du tableau Excel® Knomana. Extrait de [MSH22]

2.2 Traitements appliqués aux données

Les données contenues dans le tableau Excel® peuvent être modélisées de différentes façons. Ce peut être, par exemple, selon un schéma entité-relation représentant un système constitué des trois éléments de description d'une utilisation : une plante utilisée comme pesticide, une culture protégée et un bioagresseur repoussé. Chacun de ces trois éléments étant une espèce biologique, il est possible de compléter la description de chacun par sa taxonomie.

Pour obtenir les règles d'implication à partir des données, on exporte tout d'abord les données d'intérêt depuis Excel®. Par exemple, on peut s'intéresser uniquement à l'espèce de la plante protégée, de la plante pesticide et du bioagresseur (voir Figure 3).

	A	B	C
1	ProtectedPlant	PesticidalPlant	Pest
2	Abelmoschus esculentus	Azadirachta indica	Spodoptera littoralis
3	Abelmoschus esculentus	Carica papaya	Spodoptera littoralis
4	Brassica oleracea	Azadirachta indica	Spodoptera littoralis
5	Brassica oleracea	Carica papaya	Spodoptera littoralis
6	Gossypium hirsutum	Dioscorea dumetorum	Spodoptera littoralis
7	Solanum lycopersicum	Vincetoxicum canescens	Spodoptera littoralis
8	Solanum lycopersicum	Vincetoxicum fuscatum	Spodoptera littoralis
9	Solanum lycopersicum	Vincetoxicum parviflorum	Spodoptera littoralis
10	Ricinus communis	Wedelia prostrata	Spodoptera litura
11	Zea mays	Azadirachta indica	Spodoptera spp.
12	Zea mays	Capsicum spp.	Spodoptera spp.
13	Zea mays	Chenopodium opulifolium	Spodoptera spp.

FIGURE 3 – Tableau résultant d'une extraction ciblée de la page Excel® de Knomana

Enfin, on traduit ce jeux de données sous la forme d'une famille de contextes relationnels, de laquelle les règles d'implication seront extraites par la librairie FCA4J [8] (Formal Concept Analysis for Java) [GHM22]. L'étape de traduction du jeu de données consiste à extraire les relations binaires des données et les représenter sous forme de tables à deux entrées. Ainsi, pour le jeu de données de la figure 3 qui indique qu'on utilise une plante pesticide (colonne B) pour protéger une plante (colonne A) contre un bioagresseur (colonne C), trois contextes formels sont construits, un pour chacune des colonnes (c.f. les figures 4 et 5 pour respectivement les colonnes B et C) et trois contextes relationnels (i.e. entre A et B, entre A et C, et entre B et C) pour exprimer la relation entre chaque paire de contextes formels. La figure 6 est un exemple de contexte relationnel qui représente la relation de contrôle d'un bioagresseur (en ligne sur la figure) par une plante pesticide (en colonne sur la figure). La mise en relation des contextes formels des figures 4 et 5 par le contexte relationnel de la figure 6 permet de représenter le contrôle d'un bioagresseur par une plante pesticide.

FormalContext PesticidalPlant					
	Azadirachta indica	Carica papaya	Dioscorea dumetorum	Wedelia prostrata	Vincetoxicum fuscatum
PesticidalPlant_AzadirachtaIndica	X				
PesticidalPlant_CaricaPapaya		X			
PesticidalPlant_DioscoreaDumetorum			X		
PesticidalPlant_Wedeliaprostrata				X	
PesticidalPlant_VincetoxicumFuscatum					X

FIGURE 4 – Contexte formel des plantes pesticides

FormalContext Pest			
	Pest_SpodopteraLittoralis	Pest_SpodopteraLitura	Pest_SpodopteraSpp.
Spodoptera littoralis	X		
Spodoptera litura		X	
Spodoptera spp.			X

FIGURE 5 – Contexte formel des pestes

RelationalContext controls			
source PesticidalPlant			
target Pest			
scaling exist			
	Spodoptera littoralis	Spodoptera litura	Spodoptera spp.
Azadirachta indica	X		X
Carica papaya	X		
Dioscorea dumetorum	X		
Wedelia prostrata		X	
Vincetoxicum fuscatum	X		

FIGURE 6 – Contexte relationnel reliant une plante pesticide à la peste qu'elle permet de contrôler

8. https://www.lirmm.fr/FCA4J

2.3 Règles d'implication

Une règle d'implication, comme celle présentée sur la figure 7, est composée de trois parties principales :

— une partie *métrique* entre chevrons (encadré bleu sur la figure)

— une partie *prémisse* après le chevron fermant la partie métrique (encadré rouge)

— une partie *conclusion* après la flèche ⇒ (encadré vert).

La partie métrique contient une ou plusieurs métriques (par exemple score de support) relatives à la règle. La partie prémisse est composée de plusieurs éléments séparés par des virgules, eux-mêmes composés d'une partie *relation* (avant la parenthèse) et d'une partie *objet* (entre parenthèses). De la même manière, la partie conclusion est composée de plusieurs éléments composés eux-mêmes d'une paire relation-objet.

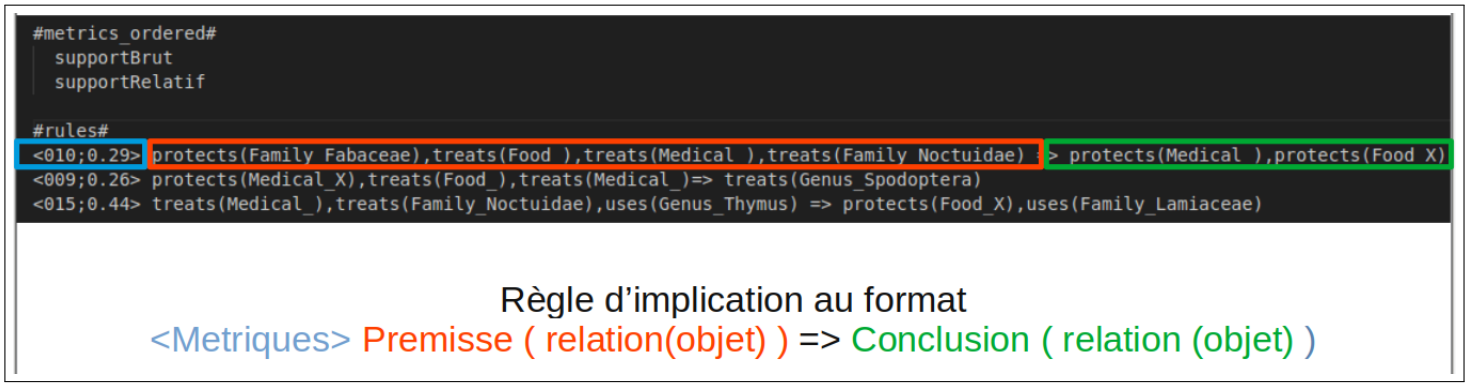

FIGURE 7 – Exemple de règle d'implication. En bleu, les valeurs de métriques (ici support brut et support relatif), en rouge la prémisse composée de 4 paires relation-objet et en vert la conclusion, composée de deux paires relation-objet.

La règle mise en évidence dans la figure 7 exprime le fait que « dans un système de protection, par une plante pesticide, dans lequel une plante (une culture agricole), de la famille fabaceae, est protégée et un bioagresseur, de la famille des Noctuidae qui n'a ni un usage alimentaire ni médical, est contrôlé, alors la plante protégée (la culture agricole) est utilisée pour l'alimentation humaine mais pas en soin médical ». Cette règle a un score de support brut de 10 (1er argument des métriques) et un score de support relatif de 0,29 (2e argument des métriques). Le support brut est le nombre d'objets portant la règle et le support relatif est le nombre d'objets portant la règle divisé par le nombre d'objets totaux.

3 Conception d'une solution d'exploration visuelle des règles d'association

Dans cette section, je vais d'abord présenter les besoins utilisateurs en section 3.1, puis exprimer les les choix de conception globaux qui ont été adoptés (section 3.2).

3.1 Expression des besoins des utilisateurs

Lorsque l'on passe des données de la base de connaissances Knomana aux règles d'implication générées par FCA4J, la quantité de règles générées peut être très importante. Pour réduire ce nombre, les experts travaillent généralement avec une partie des données de Knomana. Malgré cela, on peut encore facilement atteindre plusieurs milliers de règles, dont les prémisses et conclusions peuvent chacune contenir plusieurs centaines d'éléments. Aussi, rechercher une information en particulier au milieu d'un grand nombre de règles réunies dans une liste est un exercice complexe.

Les problématiques à résoudre par le logiciel à développer sont les suivantes :

— **P1**
Pouvoir regrouper les règles par éléments de prémisse ou de conclusion en commun (par exemple, regrouper les règles donnant des informations sur la protection d'une espèce, ou sur le contrôle d'un bioagresseur).

— **P2**
Pouvoir estimer la quantité de règles se trouvant dans chacun des groupes formés (par exemple, 50% des règles comprennent un élément protect(espèce A) alors que seulement 5% des règles comprennent un élément protect(espèce B).

— **P3**
Pouvoir naviguer dans ces règles et sélectionner des groupes possédant des caractéristiques communes d'intérêt en affinant la sélection de proche en proche.

3.2 Choix de conception

Pour pouvoir s'y retrouver dans cette grande quantité d'informations, nous allons avoir besoin d'utiliser la technique de visualisation de données. Elle permet de représenter une grande quantité de données de manière visuelle et d'y naviguer facilement.

L'application développée est prévue de façon à fonctionner sur un navigateur web (par exemple Mozilla Firefox ou Opera). En effet, proposer une application en ligne offre l'avantage, pour l'utilisateur, d'être fonctionnel, quelque soit le système d'exploitation sur lequel le navigateur fonctionne, de ne pas avoir d'installation logicielle à se préoccuper, et de n'avoir qu'à ouvrir ses fichiers pour l'appliquer sur ses données. Cette application doit permettre de charger des fichiers contenant une liste de règles d'implication (dont le format a été explicité dans la section 2.3). Une fois ces règles chargées, nous devons être en mesure de les séparer en différents groupes selon les paires relation-objet contenues dans la partie prémisse ou conclusion. Nous devons être en mesure de quantifier le nombre de règles contenues dans les différents groupes ainsi créés (nombre d'implications) et d'afficher les métriques présentées en première partie des règles.

Compte tenu de la grande variété d'éléments de prémisse et de conclusion possibles, le tri des règles d'implication se fera donc en plusieurs étapes successives, permettant d'affiner la sélection à chaque clic, et de naviguer dans ces règles en les triant par relation ou par objet. Au terme de ce processus, l'utilisateur pourra sélectionner les éléments de prémisse et de conclusion de manière unitaire. Nous expliquerons plus en détail ces étapes dans les sections suivantes.

4 Implémentation de cette solution : l'outil RCAVizIR

L'application développée est accessible sur une page web, à l'exemple de RCAViz. Elle respecte également la charte graphique de RCAviz de façon à montrer la complémentarité des deux applications. Les visualisations de données ont été implémentées à l'aide de la bibliothèque *D3.js*[9].

4.1 Vue d'ensemble

L'application RCAVizIR est en ligne, dont un prototype est temporairement testable à l'adresse `https://rcavizir.lirmm.net/projetstage/`. Une vue d'ensemble de son interface est présentée à la figure 8.

Pour effectuer ces étapes successives de sélection des règles d'implication, nous avons divisé l'interface en trois colonnes, l'affinage de la sélection s'effectuant de la gauche vers la droite. A gauche de ces trois colonnes, un panneau latéral permet de visualiser la légende. Il est repliable pour permettre à l'utilisateur d'avoir plus d'espace pour afficher les trois colonnes principales. En effet, l'utilisateur doit pouvoir se concentrer tour à tour sur chaque étape du processus. Aussi, de façon à donner le plus de confort possible, la taille de chaque colonne peut être redimensionnée par l'utilisateur.

Le première colonne permet d'effectuer une première sélection groupée des règles selon la partie relation de leurs prémisses et conclusions. La seconde colonne permet de réaliser une sélection sur la partie objet de ces prémisses et conclusion, contenant un attribut (comme par exemple ici, le nom d'une espèce), d'où son nom de matrice « Attributs ». La troisième permet de réaliser la sélection des règles elles-mêmes.

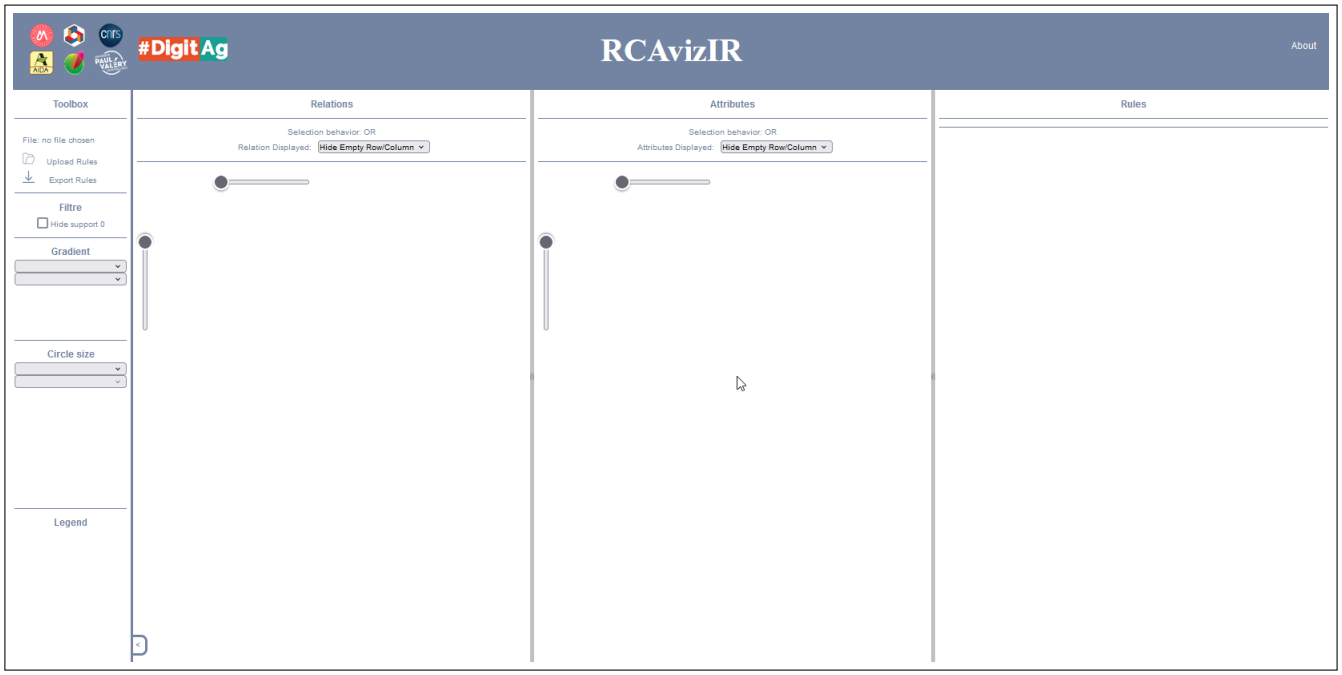

FIGURE 8 – Vue d'ensemble de l'application avant chargement d'un fichier de règles. De la gauche vers la droite : le panneau de légende, puis les trois colonnes pour affiner la sélection.

9. `https://d3js.org/`

4.2 Parseur

Le parseur est le point d'entrée de l'application. Il permet de passer d'un fichier texte contenant les règles d'implication, à un objet javascript dont les différents éléments pourront être exploités pour être visualisés.

Au début de la réalisation de cette étape, le fichier texte contenant les règles d'implication n'était pas encore pourvu de métadonnées, car les règles d'implication ne contenaient alors qu'une seule métrique entre chevrons, en l'occurrence le support brut. Dans la mesure où il est possible que d'autres métriques soient associées aux règles (e.g. le lift pour une règle d'association), nous avons prévus l'ajout d'autres métriques dans la partie située entre les premiers chevrons, où chacune est séparée de la suivante par un point virgule, comme par exemple <0,1 ;0,2>. Pour savoir quelle métrique correspond à quelle information, il a fallu ajouter des métadonnées qui informent de leur ordre d'apparition. Pour réaliser les tests, nous avons ajouté une métrique supplémentaire, i.e. le *support relatif*, afin de tester l'affichage de différentes métriques dans RCAvizIR. De cette manière, il est possible, pour l'utilisateur, de définir lui-même les métriques apparaissant dans les règles d'implication, et le nom à associer à chacune de ces métriques.

Les figures 9 et 10 présentent un fichier de règles d'implication utilisé pour tester RCAVizIR avant, et après ajout des métadonnées.

```
<010> protects(Family_Fabaceae),treats(Food_),treats(Medical_),treats(Family_Noctuidae) => protects(Medical_),protects(Food_X)
<009> protects(Medical_X),treats(Food_),treats(Medical_)=> treats(Genus_Spodoptera)
<015> treats(Medical_),treats(Family_Noctuidae),uses(Genus_Thymus) => protects(Food_X),uses(Family_Lamiaceae)
```

FIGURE 9 – Fichier d'entrée de RCAVizIR tel qu'il était produit par FCA4J au début du stage.

```
#metrics_ordered#
  supportBrut
  supportRelatif

#rules#
<010;0.29> protects(Family_Fabaceae),treats(Food_),treats(Medical_),treats(Family_Noctuidae) => protects(Medical_),protects(Food_X)
<009;0.26> protects(Medical_X),treats(Food_),treats(Medical_)=> treats(Genus_Spodoptera)
<015;0.44> treats(Medical_),treats(Family_Noctuidae),uses(Genus_Thymus) => protects(Food_X),uses(Family_Lamiaceae)
```

FIGURE 10 – Fichier d'entrée de RCAVizIR après ajout d'une métrique et de métadonnée pour faciliter son traitement par le parseur javascript, et pour tester l'affichage de différentes métriques. Dans cet exemple, les métriques figurant en première partie d'une règle représentent, dans l'ordre, le support Brut et le support Relatif, comme spécifié dans les métadonnées en début de fichier.

4.3 Construction de la première matrice

La première matrice est la matrice *Relation*. Elle permet dans un premier temps de filtrer les règles d'implication selon la/les relation(s) contenue(s) dans leur prémisse et leur conclusion. Chaque disque, contenu dans la matrice, représente un ensemble de règles comportant une certaine relation dans la prémisse (en ligne) et une relation dans la conclusion (en colonne). Par exemple, un disque situé à l'intersection de la relation `protects` en ligne, et de la relation `treats` en colonne signifie qu'il contient les règles possédant au moins une relation `protects` en prémisse, et au moins une relation `treats` en conclusion. Le fait de demander à l'utilisateur de choisir la relation entre les données (plutôt que les attributs) repose sur le fait qu'il y aura dans la plupart des cas moins de types de relations différents que d'attributs différents. Dans notre cas par exemple, les seules relations possibles sont `treats`, `protects` et `uses`, alors que les attributs correspondent à des noms d'espèces (il y en a donc un très grand nombre).

Les métriques ont un intérêt notoire pour l'utilisateur, car elles permettent de donner de précieuses informations sur les règles. Par exemple, s'il y a beaucoup de règles contenues dans un disque, mais qu'elles ont toutes un score de support très faible, elles peuvent avoir moins d'intérêt qu'un disque contenant de nombreuses règles au score de support élevé. Dans la mesure où chaque disque peut contenir de nombreuses règles, nous ne pouvons donc pas afficher le détail de ces métriques. RCAVizIR permet donc à l'utilisateur de sélectionner l'agrégation de la métrique qu'il souhaite afficher (min, max, moyenne et médiane). Le panneau légende permet donc à l'utilisateur de choisir deux des métriques à afficher ainsi que l'agrégation voulue. Elles seront respectivement représentées par le rayon des disques de la matrice, ou par un gradient allant du noir au blanc. Il est possible de sélectionner jusqu'à trois disques dans cette matrice pour créer la matrice de la colonne centrale. Les figures 11 et 12 présentent respectivement une vue d'ensemble, et une vue rapprochée de cette première étape de sélection.

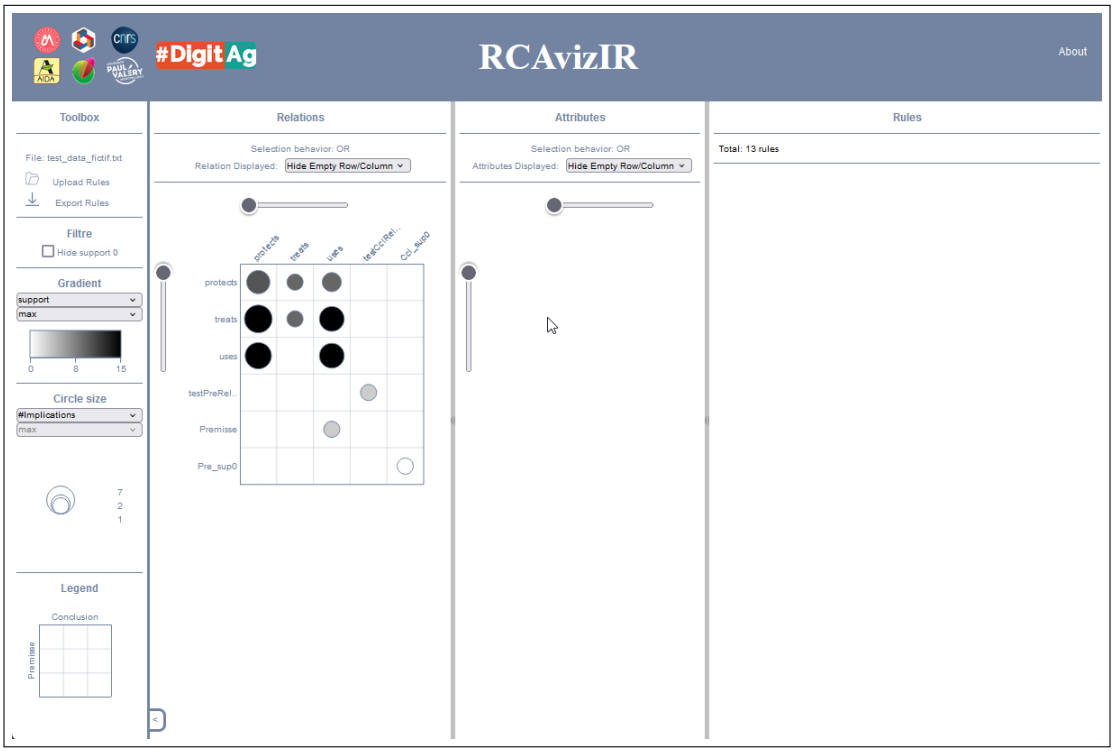

FIGURE 11 – Vue d'ensemble de l'application après chargement d'un fichier de règles, et application des options sélectionnées par défaut dans la légende. Il s'agit de la première étape de sélection : l'étape matrice *Relation* qui regroupe les différentes règles en disques selon la partie relation de leurs prémisses et conclusions

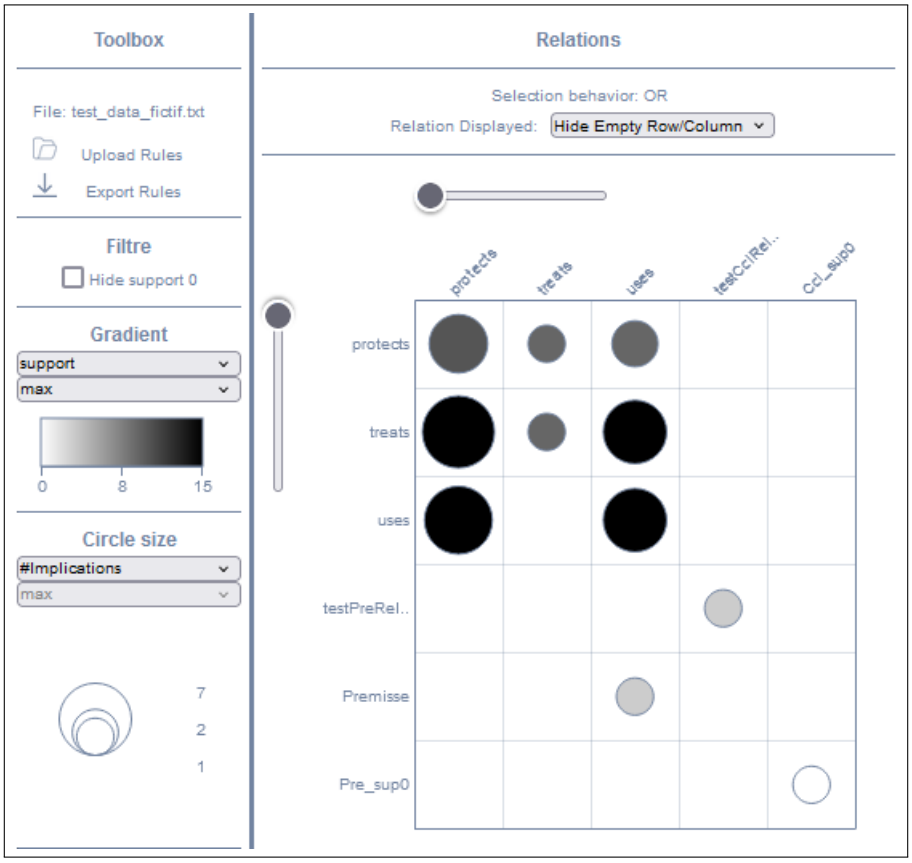

FIGURE 12 – Vue rapprochée de l'étape "Matrice Relation"

4.4 Construction de la seconde matrice

La seconde matrice est la matrice *Attribut*. Elle commence par récupérer le sous-ensemble de règles inclues dans le ou les disques sélectionnés dans la première matrice. De la même manière que dans la matrice *Relation*, elle sépare les règles en différents groupes selon les attributs contenus dans les prémisses (en colonne) et les attributs contenus dans les conclusions (en ligne). Généralement, cette matrice est bien plus grande que la matrice de la première colonne car un jeu de données contient toujours plus d'attributs que de relations, une liste d'espèces dans notre exemple. C'est pour cela que des sliders verticaux et horizontaux ont été ajoutés pour permettre de faire dérouler la matrice dans le sens vertical et horizaontal.

La légende de cette seconde matrice diffère de la première. En effet, si la métrique représentée par la taille des disques est conservée, le gradient de gris n'est plus affiché, et est remplacé par un camembert de couleur. Lors de la sélection d'un disque dans la première matrice, une couleur spécifique lui est attribuée. Nous pouvons ainsi avoir au maximum trois couleurs distinctes. Le camembert couleur de la matrice *Attribut* permet d'identifier la provenance des règles par rapport à la sélection faite dans la matrice *Relation*. Dans certains cas, une règle peut faire partie de deux ou plus, des disques sélectionnés en matrice *Relation*. Dans ce cas, la portion correspondant à ces règles est colorée en gris dans les camemberts. Dans cette matrice, l'utilisateur peut également sélectionner de 1 à 3 camemberts, et une couleur spécifique sera attribuée à chacun. Les figures 13 et 14 représentent respectivement une vue d'ensemble, et une vue rapprochée de cette seconde étape de sélection. Les disques représentés sur cette matrice contiennent uniquement les règles présentes dans la sélection faite sur la Matrice Relation

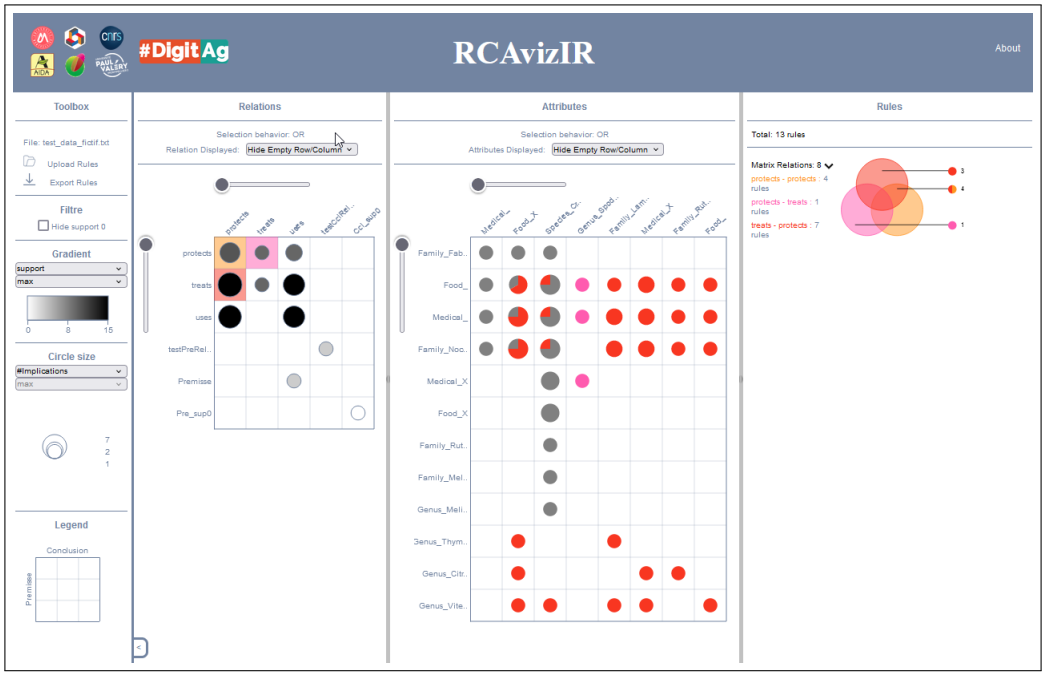

FIGURE 13 – Vue d'ensemble de l'étape « Matrice Attribut ».

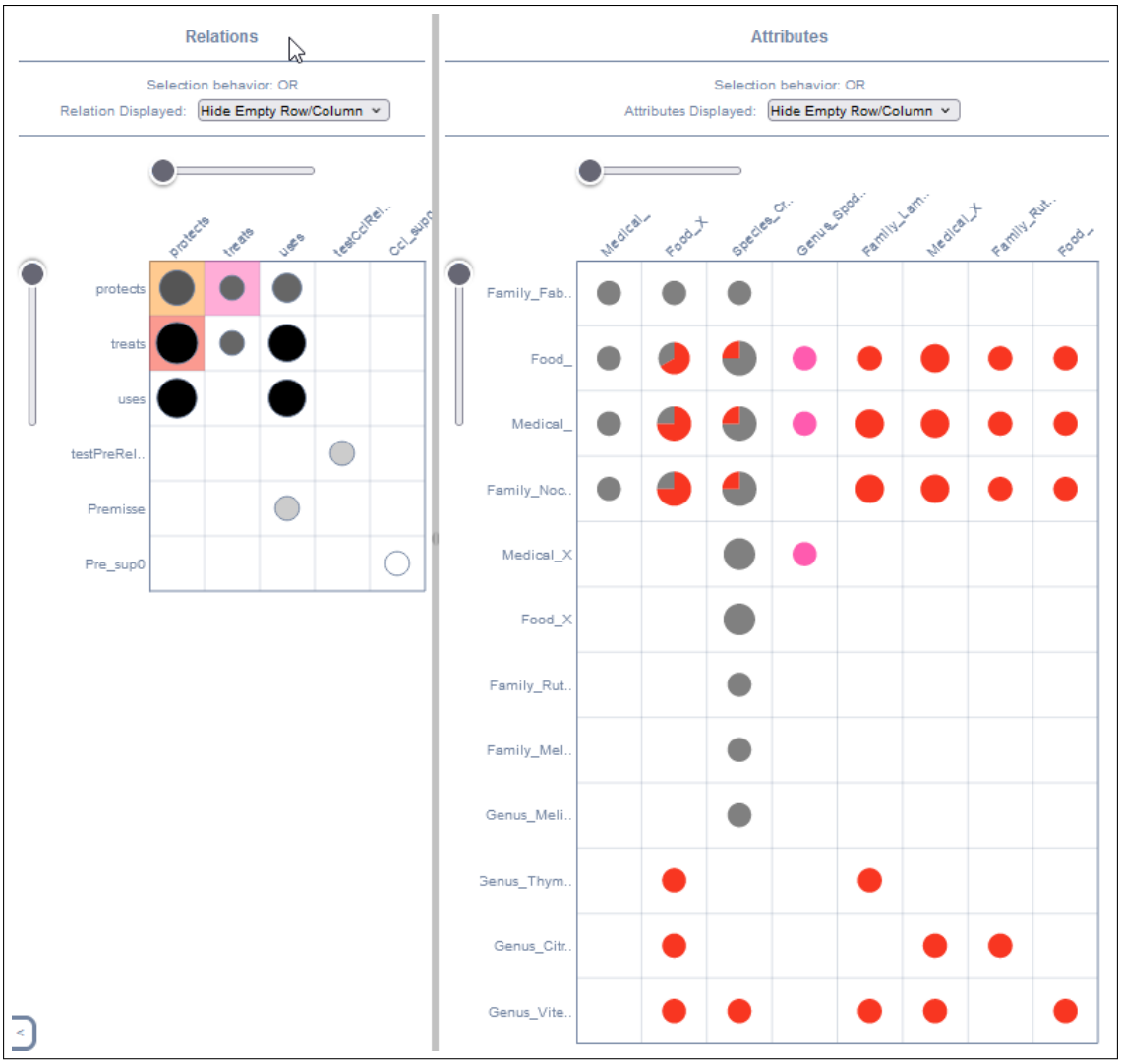

FIGURE 14 – Vue rapprochée de l'étape « Matrice Attribut ».

4.5 Affichage et sélection des règles à exporter

La dernière colonne de RCAVizIR permet tout d'abord de visualiser le détail des sélections faites dans les deux matrices. Cela permet de répondre à la problématique P2 et d'avoir une vue d'ensemble de la répartition de chacune de ces règles, et le nombre de chacun de ces types de règles. Comme dans la plupart des cas des jeux de données, des règles seront communes à plusieurs sélections, nous avons choisi de représenter ces sélections par un diagramme de Venn, qui est optimal pour visualiser et séparer les règles appartenant à une seule sélection, à deux sélections en particulier, ou aux trois sélections. Cela nous permet d'avoir plus d'informations sur la partie grise des camemberts affichés dans la matrice attribut.

La partie *Rules* de cette troisième colonne permet également d'accéder à la sélection des disques choisis dans la matrice attribut. Cette fois-ci, chacun des éléments de prémisse et de conclusion sont représentés dans leur entièreté pour permettre de réaliser une dernière sélection.

L'utilisateur peut sélectionner les éléments un par un. Lors de la sélection, seules les règles contenant l'élément de prémisse ou de conclusion choisi sont conservées. Et lorsque plusieurs éléments sont sélectionnés, les règles conservées contiennent tous les éléments sélectionnés.

Lorsque cette sélection est faite et qu'il ne reste plus que les règles d'intérêt pour l'utilisateur, ces règles peuvent être exportées grâce au bouton `export` situé dans la légende. Cette dernière étape de sélection est présentée en figures 15 et 16

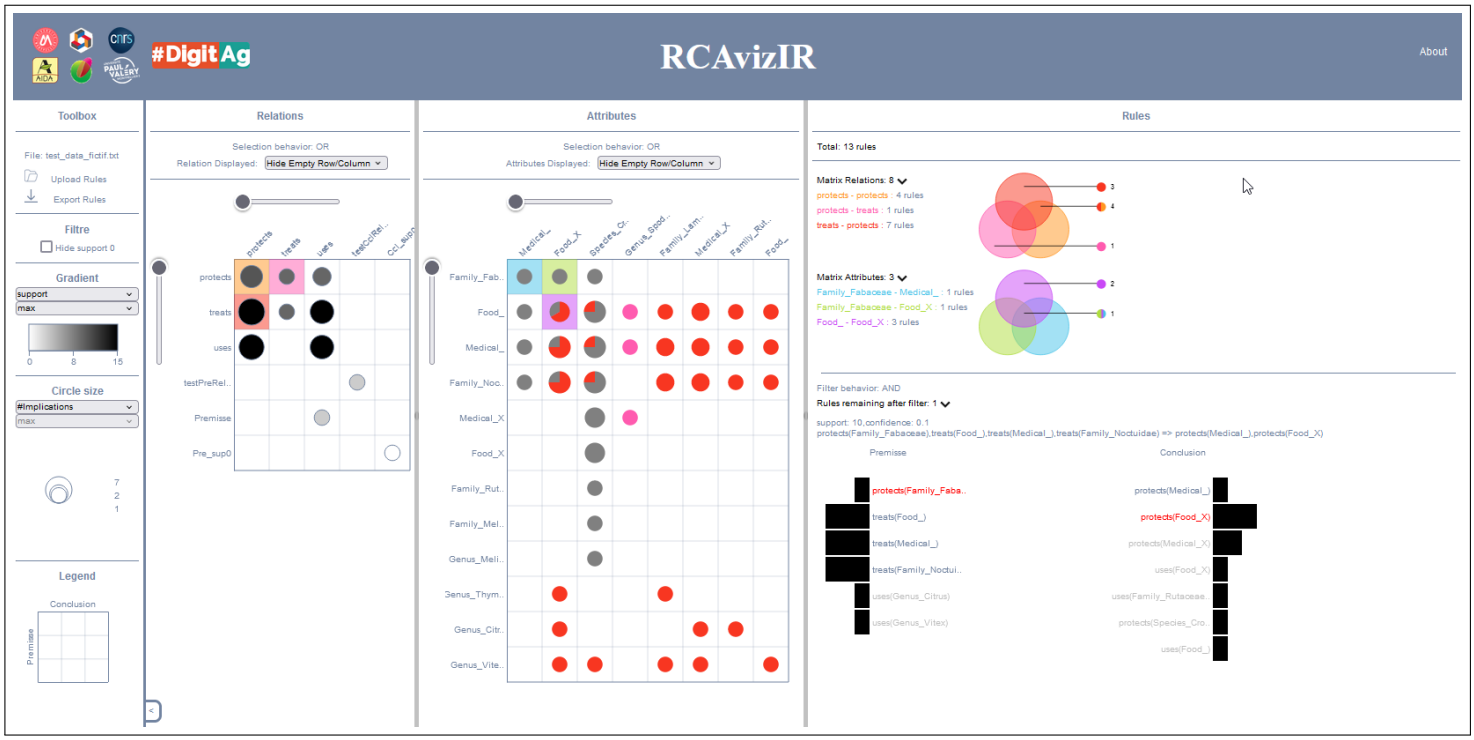

FIGURE 15 – Vue d'ensemble de l'application à l'étape « Rule »

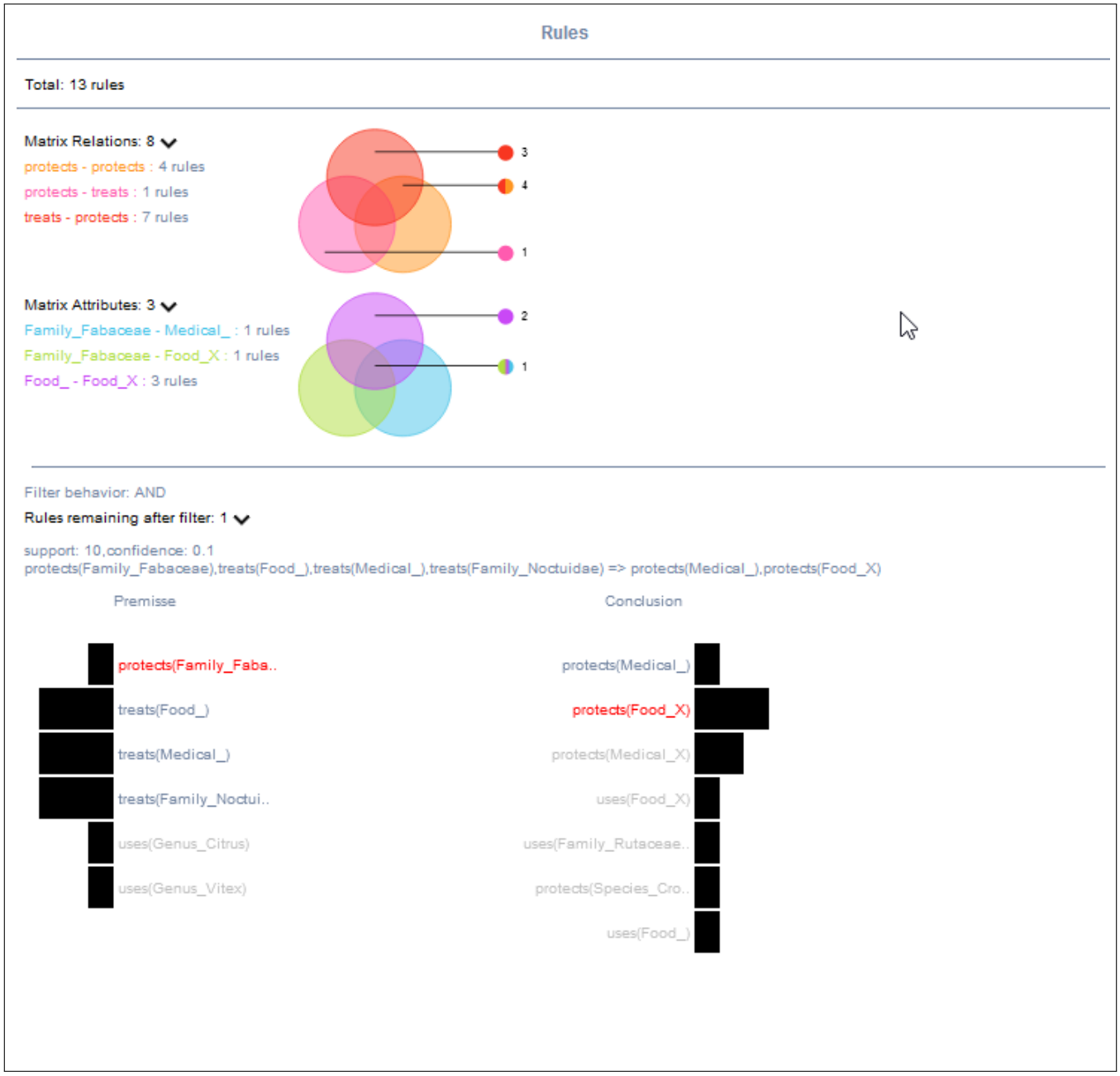

FIGURE 16 – Vue rapprochée de l'étape « Rule »

4.6 Option supplémentaire ou en développement

A la date de la soutenance, le développement de l'application n'est pas terminé. Aussi, une prolongation du stage est prévue. A cette occasion, certains éléments visuels seront changés. C'est le cas des diagrammes en barre de la colonne *Rules*. Ces barres indiquent, pour le moment, uniquement le nombre de fois où l'élément de prémisse ou de conclusion associé est présent dans les règles d'implication sélectionnées. L'objectif est de décomposer ces barres en fractions colorées selon la sélection dont elles sont issues dans la matrice *Attribut*. Cette coloration se fera de la même manière que les camemberts de la matrice *Relation*, le gris représentant l'appartenance d'une règle à plusieurs sélections.

Le deuxième objectif d'amélioration concerne l'utilisation de RCAVizIR. En effet, nous avons constaté que les règles d'implications importées dans l'application comportent souvent un grand nombre de règles au support faible ou égal à 0. Si l'option permettant de filtrer les règles de support zéro est d'ores et

déjà intégré, il serait pertinent d'améliorer ce filtre en le remplaçant par un slider pour choisir le seuil de support minimum au dessous duquel l'utilisateur ne souhaite pas visualiser les règles. Cela permettra de faire un premier filtrage des règles n'intéressant pas l'utilisateur.

Le troisième objectif d'amélioration, impliquant des changements plus en profondeur, concerne l'ajout de données supplémentaires relatives aux règles à visualiser. Les attributs présents dans les données test de Knomana étant des espèces d'organisme, nous avons déjà noté qu'il en existe un très grand nombre. Cela contribue à accroître sensiblement la taille de la matrice attribut. De plus, les informations relatives à leur classification dans la taxonomie n'est pas prise en compte dans la sélection. Pour répondre à ces deux problématiques conjointement, nous avons donc prévu de permettre à l'utilisateur d'importer un fichier de taxonomie pour regrouper les espèces appartenant par exemple à un même genre ou une même famille. Cette taxonomie pourrait être représentée dans notre matrice attribut par des colonnes dépliantes. Cette colonne permettrait d'afficher toutes les espèces d'un groupe de manière synthétique, et pourrait être dépliée, en se divisant en plusieurs colonnes, pour afficher le détail de chaque espèce présente dans ce groupe. Les attributs n'apparaissant pas dans cette taxonomie seraient laissés dans des colonnes non dépliantes. Cela permettrait également d'afficher les attributs ne correspondant pas à des noms d'espèces (par exemple, les espèces d'intérêt alimentaire et médical). Le chargement d'une taxonomie étant laissé au soin de l'utilisateur, cette fonctionnalité pourrait également s'appliquer aux taxinomies et donc permettre de regrouper relativement n'importe quel type d'attribut en différents groupes situés dans une arborescence (par exemple, une taxinomie concernant l'habitat des espèces avec des attributs comme forêt, spécialisé en forêt tropicale, et forêt équatoriale).

Le dernier objectif d'amélioration concerne la généralisation de l'application. Il s'agira d'offrir la possibilité de charger sur RCAVizIR des règles triadiques à la place des règles d'implication actuelles [BBN19] élaborées sur des données relationnelles à deux dimensions. Les règles triadiques sont actuellement calculées par une application différente de FCA4J. Leurs format final n'étant pas encore fixé, il est encore possible de les exporter sous un format proche de celui des règles d'implication actuellement visualisées (avec les mêmes séparateurs). Elles pourraient donc tout à fait être traitées par le parseur actuel et donc visualisées par RCAVizIR.

4.7 Exemple de scénario d'utilisation de l'application

Un agriculteur cultive une plante A, sur un terrain qu'il sait attaqué par deux espèces B et C, qui est déjà protégé par une plante pesticide D. Il aimerait cultiver une autre espèce de plante. Il pourra choisir dans la première matrice les relations "protect" en prémisse, et "protect" en conclusion, (car il cherche un système dans lequel une nouvelle espèce sera protégée en plus de celle qui l'est déjà) puis sélectionner l'espèce de plantes qu'il cultive actuellement, en premisse, et des espèces de plante qu'il aimerait éventuellement cultiver en conclusion. Dans la dernière colonne, il pourra visualiser uniquement les règles répondant à ces critères. Il pourra ensuite sélectionner, en élément de prémisse, la plante pesticide qu'il utilise actuellement : il pourra ainsi afficher les règles qui mettront en évidence les bioagresseurs contre lesquelles la plante pesticide actuellement utilisée permet de lutter. Il pourra également sélectionner, en élément de conclusion, les bioagresseurs contre lesquelles il aimerait protéger sa culture, pour voir si une autre plante pesticide permettrait de les contrôler. Si plusieurs règles d'implication d'intérêt se dégagent de sa sélection, le score de support permettra de les départager.

5 Gestion de Projet

5.1 Organisation

Pour organiser le développement de cette application, il a été nécessaire de faire des points réguliè-rement (hebdomadaire) avec quelques encadrants afin d'assurer le suivi de ce qui était développé et de fixer petit à petit de nouveaux objectifs. Au cours des réunions mensuelles réunissant l'ensemble des encadrants, nous avons évoqué et validé les choix de conception, et même testé l'application.

Réunion technique hebdomadaire Nous avons organisé une fois par semaine des réunions tech-niques. Ces réunions avaient pour but de discuter des problèmes pratiques liés au développement tels que faire des choix entre différentes manières de coder une fonctionnalité, ou, par exemple, discuter de l'optimisation de l'application. Nous avons aussi pu aborder des problèmes liés à la visualisation, par exemple, le choix de l'emplacement des menus interactifs ou des différents boutons, ainsi que le choix des couleurs utilisées. Au terme de chaque réunion, j'ai constitué une liste des objectifs à court, moyen et long terme que j'ai classé (voire reclasser) par ordre de priorité, cette liste étant revisitée à l'issue de chaque réunion. Cela me permettait d'avoir une vision des objectifs à accomplir pour la semaine (en réalisant au moins un prototype de chaque objectif prioritaire) et à plus long terme.

Réunion Mensuelle La réunion mensuelle, quant à elle, m'a permis de mieux comprendre les données agroécologiques utilisées, ainsi que le fonctionnement des concepts relationnels et comment obtenir et lire les règles d'implication obtenues à partir de ces données. Cette réunion était également l'occasion de re-situer la travail par rapport aux besoins des utilisateurs (par exemple ajouter des options nous permettant de mettre en valeur une information d'intérêt pour un biologiste).

5.2 Développement

La première semaine du stage m'a permis de découvrir et de prendre en main la bibliothèque *D3.js*, et de comprendre les données à visualiser en discutant avec les membres de l'équipe.

Les deux premiers mois ont été consacrés au développement de la base du système, en s'inspirant de l'aspect visuel du prototype réalisé en 2022 par Gulhèm Blanchard. Lors de cette étape, j'ai pu développer en javascript les trois colonnes destinées à recevoir les visualisations d3, ainsi que le volet coulissant de légende situé à gauche de l'application. J'ai aussi reproduit la bannière de RCAVizIR sur le modèle de celle de RCAviz. Un aperçu de ce à quoi ressemblait RCAVizIR pendant cette étape est visible en figures 17 et 18

Au cours des quatre mois suivants, le développement a été beaucoup moins linéaire car celui-ci s'ajustait en fonction des décisions de chaque réunion. En effet, comme il s'agit d'un projet de recherche, nous n'avions pas une idée précise de l'application que nous obtiendrions à la fin, mais une liste de problématiques à résoudre et d'informations d'intérêt à visualiser. Nous avons donc, pour cette partie, adopté une méthode de travail Agile, en revenant régulièrement sur des points précédemment développés pour y apporter des modifications. Nous avons aussi proposé des prototypes de certaines fonctionnalités, qui ont été conservés, ou non, suite aux retours des utilisateurs. Nous avons par exemple abandonné l'affichage des gradients en matrice Attribut, et changé de place plusieurs fois les boutons de sauvegarde, d'upload et d'export. Lorsqu'une partie me posait particulièrement problème, par exemple, à cause de mon manque de maîtrise d'une fonctionnalité d3, j'ai développé en parallèle des petits projets secondaires pour pouvoir extraire cette fonctionnalité du reste de l'application pendant son développement, et pour mieux la comprendre, avant de l'ajouter au projet principal. Cela a été le cas par exemple pour les camemberts réalisés dans la matrice Attribut.

Le projet est actuellement hébergé sur le GitLab du LIRMM. Celui-ci comprend une branche *main*, qui correspond à la version stable accessible à l'url `https://rcvizir.lirmm.net/projetstage/` . Il comprend aussi une branche de développement, elle aussi accessible en ligne à partir d'une url, et rendant possible son accès par les membres de l'équipe participant aux réunions techniques.

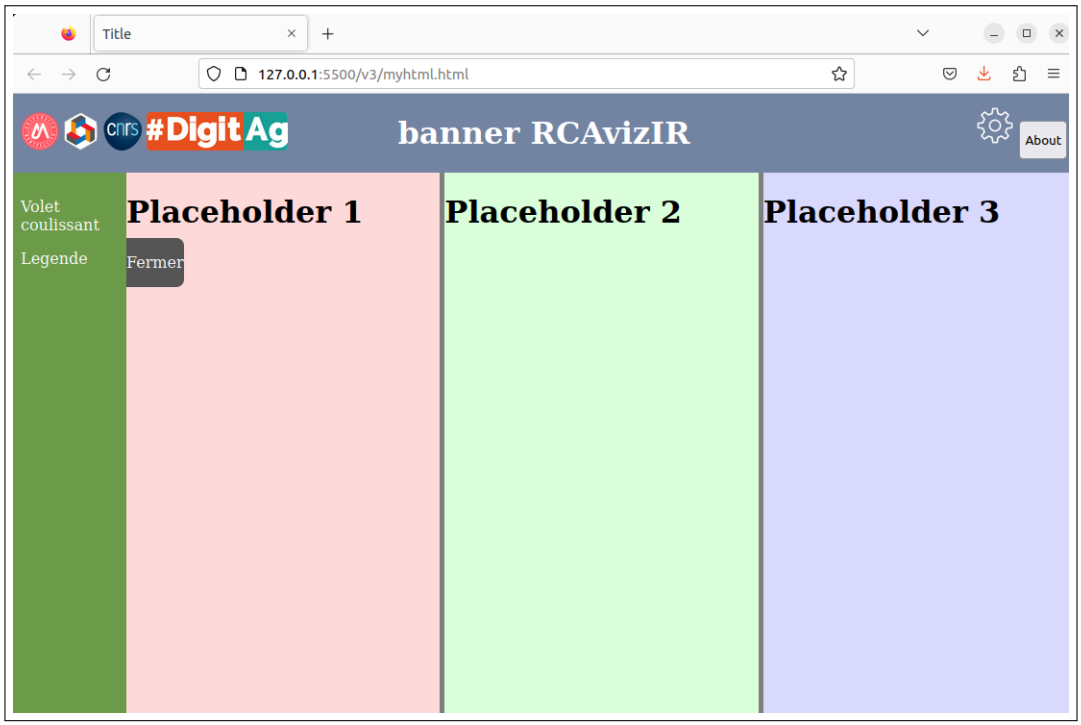

FIGURE 17 – Aperçu de l'application au 24/02/2023

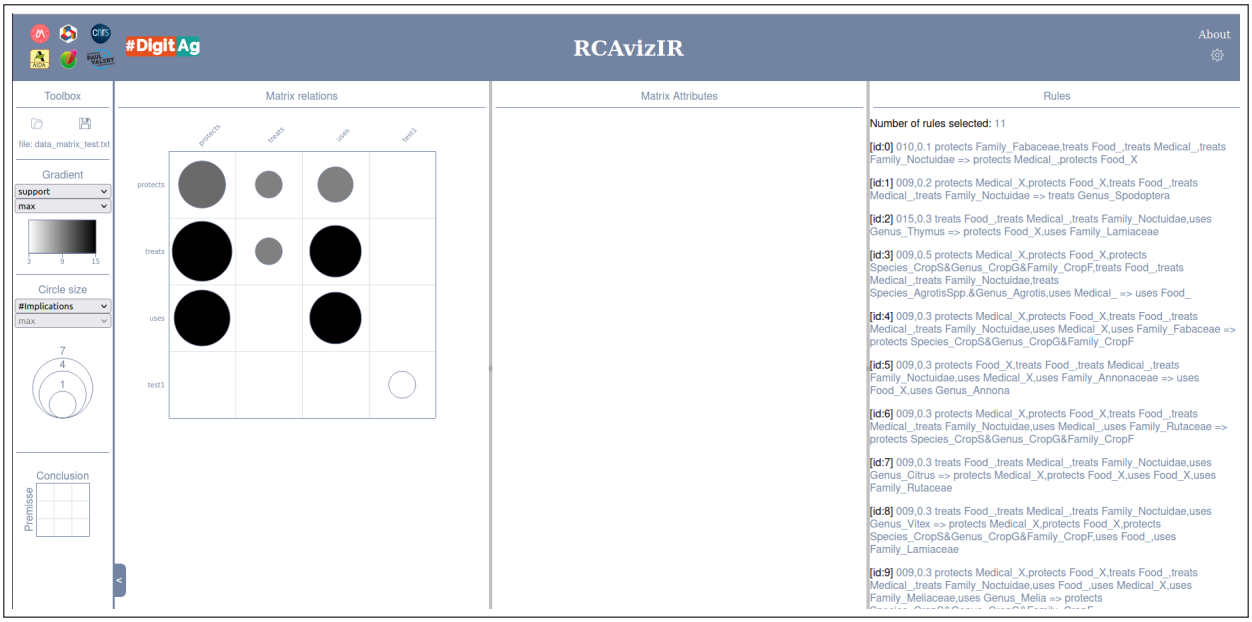

FIGURE 18 – Aperçu de l'application au 23/03/2023

6 Conclusion

L'objectif de ce stage était de réaliser une application permettant de visualiser des fichiers contenant une liste de règles d'implication, et de les regrouper par éléments de prémisse et de conclusion tout en quantifiant les règles appartenant aux différents groupes ainsi crées. Cela nous a permis de développer une application permettant de naviguer au sein de ces règles selon différents critères laissés au choix de l'utilisateur, tels qu'une métrique, un élément de prémisse ou une conclusion d'intérêt. Si l'application développée, RCAVizIR, est un premier prototype fonctionnel nous permettant de visualiser ces règles, il reste encore de nombreuses voies à explorer pour approfondir cette application. Comme nous en avons parlé en section 4.6, la généralisation de l'application aux règles triadiques, ainsi que l'ajout d'une taxonomie pour nous permettre d'organiser mieux la matrice Attribut seront les prochaines fonctionnalités à implémenter à la suite de ce stage. Compte tenu du fait que l'application RCAVizIR est disponible en ligne, dans le futur, on pourrait imaginer la possibilité de récupérer les données d'une base de connaissances traitées par FCA4J, directement depuis FAC4J sans avoir à gérer les fichiers. Ceci permettrait de ne pas se reposer sur les capacités de calcul et de stockage locales de la machine utilisée.

Ayant commencé l'informatique à partir de la quatrième année, ce stage m'a permis de pratiquer les connaissances acquises durant ces deux années, et d'apprendre plus en profondeur javascript, et la librairie d3 que je n'avais jamais utilisée. Plus personnellement, étant issue d'un parcours universitaire entre la biologie et l'informatique, ce stage m'a donné l'opportunité de pouvoir réaliser ce pourquoi j'ai entrepris ce master : développer des outils informatiques au service de la biologie.

7 Remerciements

Je tiens à remercier tous mes encadrants de stage, Marianne Huchard, Arnaud Sallaberry, Vincent Raveneau, Pierre Martin, Alexandre Bazin, Pascal Poncelet pour l'aide apportée, et pour leur disponibilité tout au long de ce projet. Plus particulièrement, Vincent Raveneau pour l'aide technique apportée très régulièrement durant le développement de l'application, Arnaud Sallaberry et Pascal Poncelet pour leur présence aux réunions hebdomadaires, Alexandre Bazin et Marianne Huchard pour les explications très complètes sur la création des règles d'implication, et Pierre Martin pour l'interprétation des données agroécologiques, et les explications sur la base de données Knomana.

Références

[BBN19] Alexandre Bazin, Aurélie Bertaux, and Christophe Nicolle. Représentation condensée de règles d'association multidimensionnelles. In Marie-Christine Rousset and Lydia Boudjeloud-Assala, editors, Extraction et Gestion des connaissances, EGC 2019, Metz, France, January 21-25, 2019, volume E-35 of RNTI, pages 225–236. Hermann-Éditions, 2019.

[GHM22] Alain Gutierrez, Marianne Huchard, and Pierre Martin. FCA4J : A Java Library for Relational Concept Analysis and Formal Concept Analysis. In Alexandre Bazin, Karell Bertet, Christophe Demko, Pierre Martin, and Ants Torim, editors, ETAFCA 2022 - ExistingTools and Applications for Formal Conceptual Analysis Workshop@CLA2 volume 3308 of CEUR Workshop Proceedings, pages 207–212, Tallinn, Estonia, June 2022. CLA Conference Series (cla.inf.upol.cz). in conjunction with CLA 2022. The 16th International Conference on Concept Lattices and Their Applications, Tallinn, Estonia, June 20 - 22, 2022.

[GHSM22] Blanchard Guilhèm, Marianne Huchard, Arnaud Sallaberry, and Pierre Martin. Navigation dans les règles d'implication extraites de connaissances agroécologiques en santé animale et végétale pour l'aide à la décision. 2022.

[KOH+19] Priscilla Keip, Amirouche Ouzerdine, Marianne Huchard, Pierre Silvie, and Pierre Martin. Navigation conceptuelle dans une base de connaissances sur l'usage des plantes en santé animale et végétale. 2019.

[MHM+22] Emile Muller, Marianne Huchard, Pierre Martin, Pascal Poncelet, and Arnaud Sallaberry. Rcaviz : Visualizing and exploring relational conceptual structures. In Pablo Cordero and Ondrej Krídlo, editors, Proceedings of the Sixteenth International Conference on Concept Lattices and Their Applications (CLA 2022) Tallinn, Estonia, June 20-22, 2022., Tallinn, Estonia, June 20-22, 2022, volume 3308 of CEUR Workshop Proceedings, pages 133–146. CEUR-WS.org, 2022.

[MSH22] P. Martin, P. J. Silvie, and M. Huchard. Conference, using a pesticidal plant requires managing knowledge-intensive inputs. Yamoussoukro, Côte d'Ivoire, July 2022. Conference, ICCP 3, Yamoussoukro, Côte d'Ivoire, July 25-29.

[MST+18] Pierre Martin, Samira Sarter, Appolinaire Tagne, Zakaria Ilboudo, Pascal Marnotte, and Pierre Silvie. Knowing the useful plants for organic agriculture according to literature : Building and exploring a knowledge base for plant and animal health. In African organic conference, pages 137–141, 2018.

[SMH+21] Pierre 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 :896, 04 2021.