# Muskca : un système de fusion d'ontologies fondé sur le consensus et l'estimation de la confiance

**Fabien Amarger** [1], **Catherine Roussey** [2], **Ollivier Haemmerlé** [1], **Nathalie Hernandez** [1], **Romain Guillaume** [1]

1. *IRIT, UMR 5505 Université de Toulouse, UT2J 5 allées Antonio Machado F-31058 Toulouse Cedex, France*
   *prenom.nom@univ-tlse2.fr*
2. *UR TSCF, Irstea, 9 av. Blaise Pascal CS 20085, 63172 Aubière, France*
   *prenom.nom@irstea.fr*

RÉSUMÉ. *Aujourd'hui de nombreux jeux de données sont disponibles pour un même domaine d'intérêt sur le web de données liées. Ces jeux peuvent être de qualité variable ce qui rend difficile leur réutilisation. Dans cet article, nous présentons une approche permettant d'identifier les connaissances partagées par différents jeux en tenant compte de leur qualité. L'approche repose sur l'utilisation de métriques permettant d'évaluer la confiance des éléments communs extraits de jeux de données. Nous proposons plusieurs métriques dont une fondée sur l'intégrale de Choquet. Ces métriques ont été évaluées sur trois jeux de données réels du domaine agricole.*

ABSTRACT. *Today many datasets related to the same domain of interest are available on the web of Linked Data. These datasets can have variable quality, which makes them difficult to reuse. In this article, we present a novel approach for identifying knowledge shared by different datasets taking into account their quality. This approach is based on metrics used to evaluate the trust score of common elements extracted from various datasets. In this article we propose several metrics, one of them is based on the integral of Choquet. These metrics have been evaluated on a real case study from the agriculture domain.*

MOTS-CLÉS : *ontologies, bases de connaissances, consensus, fusion, fonction de confiance, intégrale de Choquet.*

KEYWORDS: *ontology development, trust, non-ontological sources, ontology design pattern, ontology merging.*

DOI:10.3166/RIA.32.313-344 

## 1. Introduction

Les technologies du web sémantique sont maintenant suffisamment matures pour permettre la publication de données structurées sur le web, contribuant ainsi au web de données liées. Le web de données liées doit actuellement faire face à un défi de taille car de plus en plus de données y sont publiées sans indication de leur qualité. Il devient donc difficile de réutiliser ces données. De plus, de nombreux jeux de données sont publiés sur un même domaine. Ces jeux de données mis en ligne par des organismes différents ont souvent été constitués pour répondre à un ou des usages spécifiques. La FAO [1] propose par exemple sur le web de données liées le thésaurus Agrovoc. Ce thésaurus est utilisé pour cataloguer toute ressource documentaire en lien avec l'agriculture. Les instituts de recherche français comme l'INRA [2] ou l'IRSTEA [3] ont également développé leur propre thésaurus pour cataloguer les articles scientifiques dans le domaine de l'agriculture. Parallèlement, le projet Agronomic Linked Data propose lui aussi plusieurs ontologies pour faciliter l'intégration de données hétérogènes dans le domaine de la biologie des plantes. Exploiter ces jeux de données pour un nouvel usage implique une analyse approfondie des éléments qui les composent ainsi qu'une évaluation de la qualité de leurs données.

Cet article présente une méthode de construction de bases de connaissances (ontologies avec ou sans individus) qui réutilise simultanément plusieurs bases de connaissances sources (BCS) de qualité variable. Notre objectif est d'exploiter les éléments communs aux sources ainsi que les élements complémentaires, tout en tenant compte des spécificités de chacune d'elles. Pour ce faire, nous attribuons à chaque élément extrait des différentes sources un score de confiance. Nos travaux reposent sur l'hypothèse suivante :

> La confiance d'un élément extrait des sources est fonction de deux critères : (1) le nombre de sources dans lesquelles il apparaît et (2) la qualité de ces sources.

Nos travaux s'intéressent à l'extraction de deux types d'éléments des bases de connaissances :

– des éléments de type sommet représentant des individus, des classes ou des litéraux,

– des éléments de type arc représentant des relations entre des sommets : des attributs (datatype properties), des relations entre instances (object properties), des relations de typage (rdf:type), des relations de subsomption (rdfs:subclassOf), etc.

Des travaux préliminaires ont déja été publiés sur notre approche. (Amarger *et al.*, 2015) présente l'approche générale de transformation de sources non ontologiques en bases de connaissances, ainsi que la fusion de bases de connaissances. Ces travaux considèrent les deux types d'éléments (sommet et arc) et présentent deux métriques

1. Food and Agriculture Organization of the United Nations.
2. Institut National de la Recherche Agronomique.
3. Institut national de Recherche en Sciences et Technologies pour l'Environnement et l'Agriculture.

d'évaluation de la confiance d'un élément extrait. Une troisième métrique fondée sur l'intégrale de Choquet a été proposée dans (Amarger *et al*., 2016). Cette métrique a été appliquée uniquement sur des sommets. Dans cet article, nous présentons notre approche de fusion de bases de connaissances prenant en compte les deux types d'éléments (sommet et arc). Nous proposons pour cela de nouvelles métriques d'évaluation de la confiance des arcs.

L'article est organisé de la façon suivante. Dans un premier temps, nous dressons un panorama des travaux portant sur la fusion de bases de connaissances. Nous présentons ensuite notre approche de fusion puis nous détaillons nos propositions pour calculer le score de confiance d'un élément. Finalement, nous présentons une évaluation de notre approche dans le domaine de l'agriculture.

Dans la suite de cet article, les exemples illustrant l'approche sont issus de trois sources disponibles sur le web :

**Agrovoc** Agrovoc est un thésaurus multilingue consacré à l'agriculture, utilisé par de nombreux acteurs dans différents pays. Agrovoc est une source intéressante car elle contient de nombreux labels dans différentes langues et elle a été réutilisée dans de nombreux projets (Caracciolo *et al*., 2013). La FAO gère cette source et la tient à jour régulièrement à l'aide d'un groupe d'experts.

**TaxRef** La taxonomie TaxRef du Muséum d'Histoire Naturelle est la taxonomie de référence en France pour les organismes vivants. Ce statut de référence pousse ses concepteurs à garantir la qualité des éléments de cette source. Elle est mise à jour régulierement à partir de sources existantes (Gargominy *et al*., 2016).

**NCBI** La taxonomie du National Center for Biotechnology Information (NCBI) n'a pas vocation à être une référence mais à être exhaustive. Pour cela, ses concepteurs s'autorisent des redondances. Elle bénéficie d'un processus de validation des taxons qui lui permet d'avoir un grand nombre de taxons issus de la littérature (Federhen, 2012). NCBI est particulièrement fournie en labels bien que la langue utilisée soit toujours l'anglais. Elle référence des labels anciens qui ont pu été remplacés depuis. Cette source est intéressante car elle est plus complète que les deux précédentes.

## 2. État de l'art sur la fusion de bases de connaissances

Construire une base de connaissances à partir de plusieurs BCS existantes est équivalent à un processus de fusion. Nous considérons dans cet article la définition de *fusion* de modèles telle qu'elle est proposée dans les travaux de (Pottinger, Bernstein, 2003) :

> En considérant deux modèles A et B et un ensemble de correspondances $Map_{AB}$ établies entre ces deux modèles, le processus de fusion génère un troisième modèle représentant l'union sans doublon des modèles de A et B conformément aux correspondances de $Map_{AB}$.

Cette définition est suffisamment générique pour considérer comme modèles plusieurs types de sources, dont les ontologies ou les bases de connaissances. La notion d' "union sans doublon" est particulièrement intéressante car elle impose de mettre en place un traitement particulier pour les éléments communs aux deux modèles. Nous nous intéressons aux travaux de fusion capables de générer automatiquement une nouvelle base de connaissances contenant les parties communes des sources. Comme souligné dans (X. L. Dong *et al.*, 2014), la fusion de connaissances est une tâche encore plus délicate que celle de la fusion de données. En effet, il ne s'agit pas uniquement de choisir dans les sources la bonne valeur à associer à une donnée mais il s'agit aussi d'identifier les éléments pouvant être mis en correspondance entre les différentes sources, or ce processus d'alignement peut lui aussi introduire des biais. Certains travaux incluent le calcul de l'alignement dans la fusion (X. Dong *et al.*, 2014) alors que d'autres considèrent l'alignement comme une entrée du processus (Raunich, Rahm, 2014). Comme il existe de nombreux systèmes d'alignement (Shvaiko, Euzenat, 2013), nous considérons qu'il n'est pas nécessaire de proposer un nouveau système mais plutôt de réutiliser des travaux existants et éprouvés.

Pour comparer les travaux traitant de fusion de bases de connaissances, deux autres caractéristiques sont d'importance :

**symétrique :** la notion de fusion symétrique implique que les deux modèles à fusionner aient la même importance. Il est aussi possible d'utiliser une technique de fusion asymétrique pour privilégier un modèle plutôt qu'un autre ;

**confiance :** suivant le processus de fusion appliqué, une confiance peut être associée aux éléments du modèle résultat de la fusion.

Certaines approches se sont portées sur la fusion de deux sources. L'approche présentée dans (Raunich, Rahm, 2014) propose un processus asymétrique de fusion. Ce processus dédié à la fusion de deux taxonomies donne une priorité à un des modèles pour lever les ambiguïtés. Les travaux de (Guzmàn-Arenas, Cuevas, 2010) proposent une fonction de confusion évaluant la disparité entre deux contraintes de domaine incompatibles. La problématique de fusion étant dans ces cas-là considérée entre deux modèles, la prise en compte de la confiance à accorder à un élément est simplifiée puisqu'il n'y a que deux possibilités et que la notion d'asymétrie permet de faire un choix. Néanmoins, puisque nous considérons que la fusion porte sur plus de deux sources (ce qui est plus réaliste à l'échelle du web de données liées), nous souhaitons généraliser la notion d'asymétrie en considérant l'importance relative de chaque source dans ce processus.

Des travaux de la littérature considèrent également plusieurs sources dans leur processus de fusion. L'approche présentée dans (X. Dong *et al.*, 2014) propose de fusionner à l'échelle du web des triplets extraits de 4 sources non ontologiques (textes, structure de documents HTML, tableaux et annotations posées manuellement) et d'une base de connaissances. Comme dans nos travaux, cette approche probabiliste prend en compte le nombre de sources dans lesquelles un triplet est présent. Cependant pour compenser les erreurs d'extraction dans les sources non ontologiques, une priorité est donnée aux triplets présents dans la base de connaissance. Ces triplets sont en effet

utilisés pour prédire la probabilité qu'un triplet candidat soit valide. Une approche similaire est présentée dans (X. L. Dong *et al.*, 2014), où trois techniques de fusion de données sont optimisées pour identifier la ou les valeurs correctes correspondant à l'objet de predicats extraits d'une base de connaissances. Dans notre cas, nous considérons que l'ensemble des sources correspondent à des bases de connaissances. De plus, nous ne souhaitons pas favoriser *a priori* une source par rapport aux autres, mais prendre en compte la qualité de chacune d'entre elles.

De nombreuses mesures ont été proposées pour analyser la qualité d'une base de connaissances sur le web de données liées (Zaveri *et al.*, 2016). Dans notre approche, nous proposons d'évaluer la qualité d'un élement d'une base de connaissances en considérant à la fois la présence de cet élement dans les différentes sources et la qualité de la source elle-même.

D'autres approches proposent de prendre en compte la cohérence logique au moment de la fusion. Les travaux présentés dans (Lin, Mendelzon, 1999) ont pour objectif de générer une base de connaissances cohérente, contenant l'ensemble des axiomes non contradictoires de plusieurs bases de connaissances sources. Dans le cas ou deux BCS contiennent des axiomes contradictoires, ($A$ et non $A$), l'axiome qui apparaît le plus souvent dans les différentes BCS est conservé. Si le même nombre de BCS contiennent $A$ et non $A$, les deux axiomes sont éliminés de la BC finale. Pour ce faire, un ordre partiel sur les ensembles d'axiomes non contradictoires issus de $N$ BCS a été défini. Il est à noter qu'il n'est pas possible de calculer automatiquement l'opérateur de fusion car le problème est NP-complet. Dans notre cas, nous modélisons ce principe de majorité par des scores de confiance qui tiennent compte de la fiabilité des relations d'équivalence. Nous étendons également ce calcul de score en tenant compte de la qualité de la BCS.

Nous proposons plus précisément de quantifier la confiance à accorder à un élément extrait d'une source en fonction du nombre de sources dans lesquelles il est présent, ainsi que la qualité de chacune d'elles.

## 3. Processus général

En raison de sa généricité, nous avons choisi de travailler avec la méthodologie NeOn. NeOn se décline en neuf scénarios, chacun proposant une méthode de construction d'ontologies différente. Chaque scénario implique différents processus pour la construction collaborative d'ontologies et de réseaux d'ontologies. Nous avons tout d'abord étudié le scénario 2 de Neon "réutilisation et réingénierie de sources non ontologiques" (Villazon-Terrazas *et al.*, 2010). Ce scénario génère des ontologies à partir de sources non ontologiques comme des systèmes d'organisation des connaissances, des bases de données ou d'autres types de sources. Cette méthode prend en compte les choix de modélisation et d'implémentation et applique le même ensemble de règles de transformation sur la source. Chaque ensemble de règles de transformation constitue un patron de transformation. Comme la transformation d'un thésaurus doit être guidée, nous avons également étudié le scénario 7 "réutilisation de patrons de conception

d'ontologies". Un patron de conception d'ontologies (ou ODP, pour Ontology Design Pattern) (Gangemi, Presutti, 2009) est défini en tant que solution de modélisation pour un problème récurrent de conception d'ontologies (Gangemi, Presutti, 2009). Les ODP sont normalement générés par l'expérience des ontologues, qui les soumettent en ligne dans des répertoires de patrons de conception [4]. Ces patrons sont évalués par la communauté des ontologues et sont en général adoptés en tant que bonne pratique. Le scénario 7 sélectionne des patrons pour construire des ontologies.

Comme vu précédemment, la méthodologie NeOn (Suárez-Figueroa *et al.*, 2012) propose plusieurs méthodes de création de bases de connaissances exploitant des sources existantes, ontologiques ou non. Ces méthodes utilisent plusieurs sources séparément afin d'enrichir la base de connaissances de manière séquentielle. À notre connaissance, aucune méthode n'exploite simultanément plusieurs sources existantes. Or l'intérêt d'une telle approche est de comparer les éléments présents dans les différentes sources, afin de statuer sur leur réutilisation dans la base de connaissances finale. Nous faisons l'hypothèse que des éléments communs à plusieurs sources font partie d'un consensus sur le domaine et que leur présence dans la base de connaissances finale est souhaitable.

Nous situons notre proposition dans le cadre de la méthodologie NeOn avec les scénarios 2 et 7.

Ce processus général, présenté sur la figure 1, est composé de trois étapes (Amarger *et al.*, 2015) :

– **1 Analyse de sources :** pendant ce processus, l'expert du domaine sélectionne les sources les plus appropriées pour la construction de la base de connaissances. Il inspecte chaque source potentielle afin d'évaluer sa couverture, mais également afin d'évaluer la faisabilité de la transformation automatique en base de connaissances. Cette étape correspond aux activités 1, 2 et 3 du scénario 2 de NeOn.

– **2 Transformation de sources :** ce processus spécifie la transformation à appliquer sur chaque source pour obtenir une base de connaissances au format OWL. Les spécifications de la base de connaissances finale sont tout d'abord définies. Ces spécifications sont utilisées pour générer un module ontologique. Ce module correspond à la fusion de plusieurs patrons de conception. Cette étape correspond au scénario 7. Une transformation utilisant en entrée des patrons de transformation et le module est appliquée sur la source. Le résultat obtenu est appelé une base de connaissances source (BCS). Cette base correspond au module enrichi avec des éléments extraits de la source. Cette étape correspond aux activités 4 et 5 du scénario 2 de NeOn.

– **3 Fusion de bases de connaissances :** ce processus construit la base de connaissances finale en se fondant sur toutes les BCS extraites à partir des sources. Il est à noter que toutes ces BCS sont un enrichissement du même module. L'originalité de notre approche vient de ce processus de fusion. Il utilise plusieurs bases de connais-

4. Par exemple sur le site web http://ontologydesignpatterns.org

sances simultanément, afin d'extraire des connaissances communes aux différentes sources. Cette étape correspond à l'activité 6 du scénario 2 de NeOn.

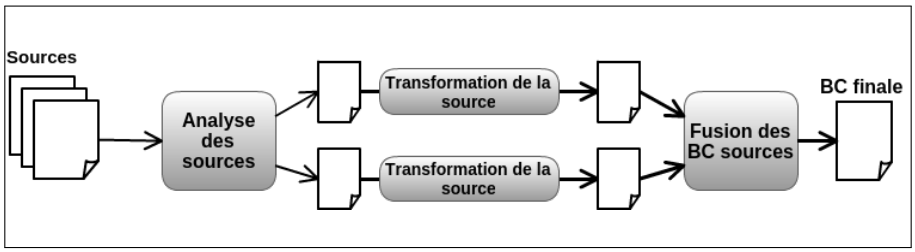


*Figure 1. Processus général*

## 4. Processus de fusion de bases de connaissances

Le processus de fusion que nous proposons est décomposé en plusieurs étapes présentées dans la figure 2.

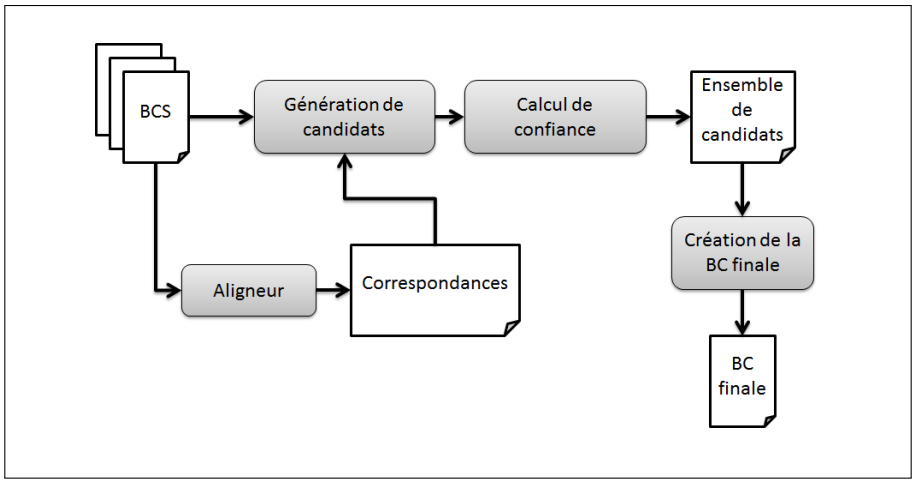


*Figure 2. Processus de fusion des bases de connaissances*

Ce processus prend en entrée les différentes BCS à fusionner. Il est à noter que toutes les BCS partagent la même modélisation des connaissances car elles sont toutes un enrichissement du même module. Il fournit en sortie une liste d'éléments pondérés, intitulés candidats, éléments potentiels de la base de connaissances finale. Ces candidats sont sélectionnés à partir d'un seuil pour être intégrés à la base de connaissances finale.

Quatre étapes sont présentes dans ce processus :

**Alignement des BCS :** cette étape établit des alignements entre tous les couples de BCS considérés ;

**Génération de candidats :** à partir des alignements, des candidats sont générés ;

**Calcul de la confiance :** un score de confiance est calculé et associé à chaque candidat ;

**Construction de la BC :** une sélection des candidats est effectuée à partir de leur score de confiance pour déterminer ceux qui appartiendront à la base de connaissances finale. Un filtre automatique peut être complété par une validation manuelle. Ensuite, une fois les candidats sélectionnés, il faut construire pour chacun d'eux l'élément le représentant dans la BC finale (choix de l'URI, choix des métadonnées, choix des labels, etc.).

### *4.1. Alignement des BCS*

Nous définissons une BCS comme un graphe $S$ orienté composé d'un ensemble de sommets et un ensemble d'arcs $S = (V_S, E_S)$ tels que :

– $V_S$ est l'ensemble des sommets de $S$. Les sommets sont les classes, les individus et les littéraux de la BCS ;

– $E_S$ est l'ensemble des arcs de $S$. Les arcs sont toutes les propriétés utilisées pour lier les individus, les classes et les littéraux.

La première étape du processus de fusion est l'alignement entre les $N$ différentes BCS. Conformément au fonctionnement des aligneurs, nous effectuons cet alignement entre chaque paire de BCS. Pour chaque paire, nous obtenons un alignement qui est un ensemble de correspondances. Soit $T$ le nombre d'alignements obtenus entre les $N$ BCS. Dans cet article, nous ne considérons comme correspondances que les relations d'équivalence stricte ($\equiv$) entre deux sommets appartenant respectivement à chacune des deux sources alignées, c'est-à-dire deux BCS $Si = (V_{Si}, E_{Si})$ et $Sj = (V_{Sj}, E_{Sj})$. Cette relation est pondérée par un degré de fiabilité (fourni par l'aligneur) représentant la probabilité que cette équivalence soit correcte.

Une correspondance se définit comme une arête entre deux sommets $\{oe_i \in V_{Si}$ ; $oe_j \in V_{Sj}\}$ pondérée par $valueE(oe_i, oe_j)$. Elle satisfait les contraintes suivantes :

– $V_{Si} \neq V_{Sj}$ car une correspondance est toujours établie entre deux sommets appartenant à des ensembles de sommets de BCS différentes ($V_{Si}$ et $V_{Sj}$) ;

– une correspondance est toujours établie entre deux sommets de même nature (soit des individus, soit des classes) ;

– $valueE()$ est une application qui, à toute arête définie comme correspondance, associe un unique degré de fiabilité compris entre 0 et 1 tel que $valueE(oe_i, oe_j) = valueE(oe_j, oe_i)$.

Dans nos travaux, nous utilisons l'aligneur LogMap [5] car ce système a obtenu de bons résultats lors de l'évaluation OAEI 2014 (Dragisic *et al.*, 2014). De plus, cet

5. http://www.cs.ox.ac.uk/isg/projects/LogMap/

aligneur met en correspondance des individus et pas seulement des classes (Jiménez-Ruiz, Grau, 2011). Il n'existe pas à l'heure actuelle d'aligneur capable de générer des correspondances entre propriétés.

### *4.2. Génération de candidats*

Deux types de candidats sont générés : les candidats sommets et les candidats arcs.

#### *4.2.1. Candidat sommet*

Les candidats sommets sont générés en exploitant les correspondances établies entre les sommets des graphes $Si$ représentant les différentes BCS.

Un candidat $CandS = (V_{CandS}, E_{CandS}, valueE)$ est un graphe non-orienté connexe dont les sommets sont des sommets provenant de BCS différentes et les arêtes sont les correspondances issues des $T$ alignements entre les $N$ BCS. Les composants d'un candidat respectent les contraintes suivantes :

– $V_{CandS}$ : $\forall v \in V_{CandS}$ avec $v \in V_{Si}$ $\nexists v' \in V_{CandS}$ tel que $v' \in V_{Si}$ et $v \neq v'$. Tous les sommets d'un candidat appartiennent à des BCS différentes. Par conséquent $|V_{CandS}| \leq N$ ;

– $E_{CandS}$ : l'ensemble des arêtes d'un candidat est inclus dans l'ensemble des arêtes des $T$ alignements. Les arêtes de $CandS$ sont des correspondances ;

– un candidat est un graphe connexe. $\forall v_1, v_2 \in V_{CandS}$, il existe forcément un chemin $path = \{e_j, ..., e_k\}$ avec $\forall e_i, e_i \in path, e_i \in E_{CandS}$ reliant $v_1$ à $v_2$. Par conséquent, tous les sommets de $CandS$ sont liés à au moins un autre sommet de $CandS$ par une correspondance.

La figure 3 présente deux candidats liant des individus issus de trois BCS. Les deux candidats représentent donc des éléments potentiels de la base de connaissances finale, ici "Triticum" et "Triticum Durum".

La génération des candidats équivaut à chercher les composantes connexes dans le graphe global constitué des $N$ BCS alignées. Nous recherchons les composantes de taille inférieure ou égale à $N$. Nous vérifions que chaque sommet de la composante appartienne à des BCS différentes. Nous effectuons un parcours en profondeur du graphe global en testant les contraintes précédentes. Nous étiquetons les sommets avec l'identifiant du candidat pour éviter les boucles infinies (Amarger, 2015).

#### *4.2.2. Candidat arc*

Un candidat arc $CandA = (V_{CandA}, E_{CandA})$ est un graphe biparti orienté étiqueté. Un candidat arc se construit à partir d'au moins un candidat sommet. Un candidat arc peut lier deux candidats sommets, comme le candidat arc2 de la figure 3. Un candidat arc peut aussi lier un candidat sommet avec un sommet du module. Ainsi, l'ensemble de sommets $V_{CandA}$ se compose soit de deux ensembles de sommets de candidats sommets distincts, soit d'un ensemble de sommets d'un candidat sommet

et d'un singleton composé d'un sommet du module. Tous les arcs de $E_{CandA}$ sont étiquetés par la même étiquette d'arc.

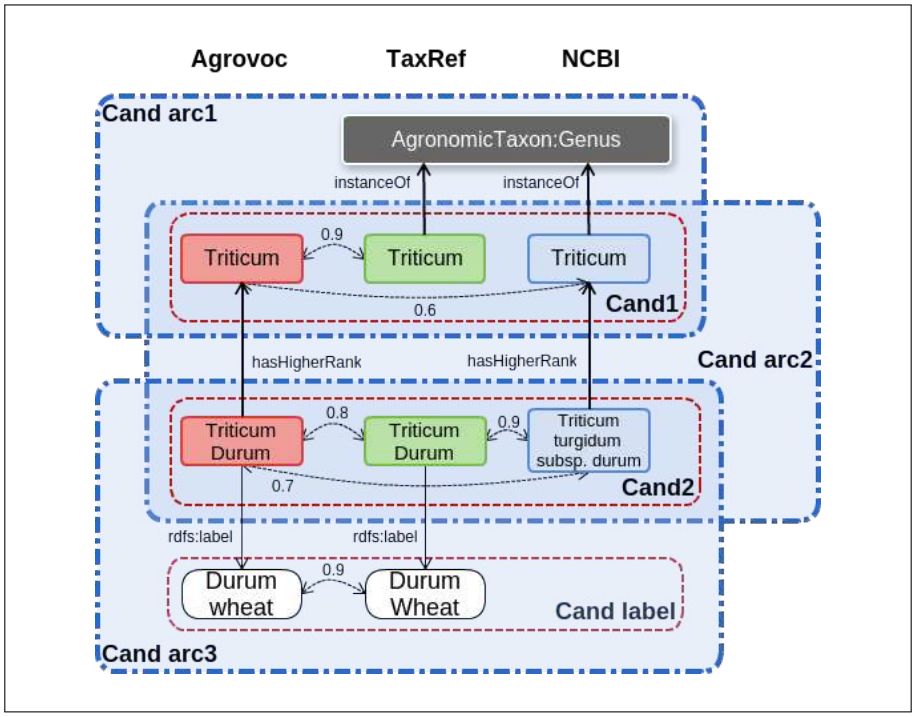


*Figure 3. Exemple de candidats arcs*

## 5. Calcul de la confiance d'un candidat

Une fois les candidats générés, nous leur affectons un score de confiance. Le premier critère de notre hypothèse cherche à favoriser les candidats contenant le plus grand nombre de sommets issus des différentes BCS et identifiés par l'aligneur comme étant équivalents. Nous définissons une première fonction $trust_{likelihood}$ qui prend en compte le nombre de sources impliquées dans le candidat (Amarger *et al.*, 2014). Nous définissons une autre fonction intitulée $trust_{degree}$ qui intègre les degrés de fiabilité des correspondances (Amarger *et al.*, 2014). Le deuxième critère de notre hypothèse consiste à tenir compte de la qualité des BCS dans le calcul de la confiance d'un candidat. Nous proposons, par la fonction $trust_{choquet}$, de prendre en compte l'implication relative de chaque source pour un candidat donné (Amarger *et al.*, 2016).

### *5.1. Fonction Trust Likelihood*

Nous définissons une première façon de calculer le niveau de consensus avec à la fonction $trust_{likelihood}$ issue des probabilités (fonction de vraisemblance en français).

Nous évaluons la confiance d'un candidat en calculant le ratio entre le nombre de sources impliquées dans le candidat et le nombre total des sources considérées dans la fusion.

Si le candidat est un candidat sommet, alors ses composantes sont $CandS = (V_{CandS}, E_{CandS}, valueE)$. Le calcul de cette fonction est présenté dans l'équation 1. Cette fonction ne prend en compte que l'ensemble des sommets du candidat $V_{CandS}$ et le nombre de sources alignées $N$.

$$trust_{likelihood}(CandS) = \frac{|V_{CandS}|}{N} \tag{1}$$

Les deux candidats sommets présentés sur la figure 4 obtiennent le même score de confiance à l'aide de la fonction $trust_{likelihood}$.

Si le candidat est un candidat arc, alors ses composantes sont :
$CandA = (V_{CandA}, E_{CandA})$. Le calcul de cette fonction est présenté dans l'équation 2. Cette fonction ne prend en compte que l'ensemble des arcs du candidat $E_{CandA}$ et le nombre de sources alignées $N$.

$$trust_{likelihood}(CandA) = \frac{|E_{CandA}|}{N} \tag{2}$$

### *5.2. Fonction Trust Degree*

Les candidats sont générés à partir de plus ou moins de correspondances. La fonction $trust_{degree}$ évalue la confiance proportionnellement au nombre de correspondances et à leurs degrés de fiabilité.

Si le candidat est un candidat sommet, alors ses composantes sont $CandS = (V_{CandS}, E_{CandS}, valueE)$. Le calcul de cette fonction est présenté dans l'équation 3.

$$trust_{degree}(CandS) = \frac{\sum_{e_i \in E_{CandS}} valueE(e_i)}{\frac{N(N-1)}{2}} \tag{3}$$

Cette fonction fait la somme de tous les degrés de fiabilité des correspondances utilisées pour générer le candidat sommet. Cette somme est normalisée en divisant le résultat par le nombre maximum de correspondances possibles, c'est-à-dire le nombre de paires possibles entre toutes les sources considérées.

Les correspondances étant utilisées dans le processus de génération des candidats, cette fonction prend en compte indirectement le nombre d'éléments présents dans le candidat. Cette fonction est proportionnelle au nombre d'arêtes : plus le graphe du candidat sommet contiendra d'arêtes, plus il contiendra de sommets.

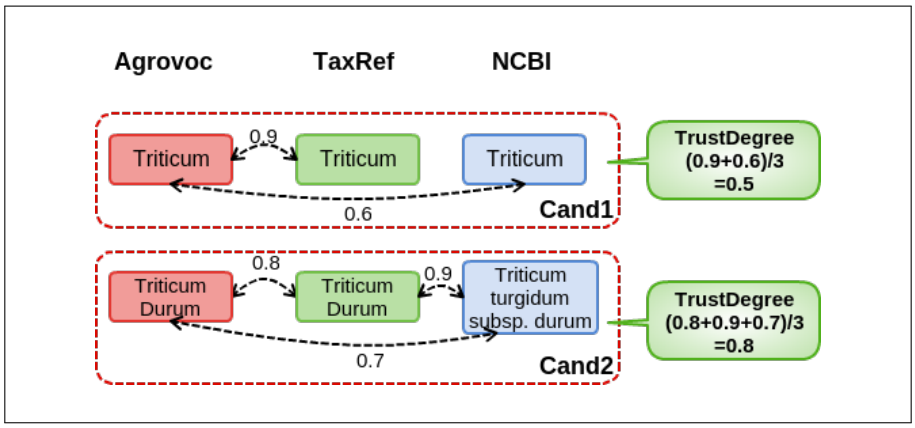


*Figure 4. Calcul des scores de confiance avec trustDegree*

La figure 4 présente les confiances associées aux deux candidats à l'aide de la fonction $trust_{degree}$. Nous observons que $trust_{degree}(Cand2) > trust_{degree}(Cand1)$. La fonction $trust_{degree}$ ordonne les candidats qui impliquent le même nombre de sources.

Concernant les candidats arcs, nous définissons la fonction de calcul de confiance $trust_{degree}$ de la même manière que $trust_{likelihood}$.

$$trust_{degree}(CandA) = trust_{likelihood}(CandA) = \frac{|E_{CandA}|}{N} \quad (4)$$

### *5.3. Trust degree avec propagation*

La confiance d'un candidat arc peut être influencée par la confiance des candidats sommets reliés par ce candidat arc. Nous considérons ici que la confiance d'un candidat arc est d'autant plus grande que les candidats sommets qu'il relie ont un score de confiance élevé. Nous définissons une confiance $trust_{degree_p}$ pour les candidats arcs prenant en compte ce phénomène de propagation de la confiance.

Tout d'abord, si le candidat arc relie deux candidats sommets, alors nous pouvons définir la confiance $trust_{degree_p}$ pour calculer la confiance de ce candidat arc.

$$trust_{degree_p}(CandA) = \frac{|E_{CandA}| + \frac{trust_{degree}(CandS1)+trust_{degree}(CandS2)}{2}}{N+1} \quad (5)$$

tel que $CandS1$ et $CandS2$ sont liés par $CandA$

Cette formule prend en compte $|E_{CandA}|$ (le nombre d'arcs présents dans le candidat arc) et la moyenne des scores de confiance des candidats sommets liés par le can-

didat arc. Nous avons normalisé ce résultat avec $N$ (le nombre d'arcs maximal) plus 1 (le maximum de la somme de deux scores de candidat sommet divisé par deux). Si un candidat arc relie un candidat sommet à un sommet du module, alors le $trust_{degree}$ du sommet du module est égal à 1.

### *5.4. Fonction Trust Choquet*

L'intégrale de Choquet (équation 10) est utilisée pour la prise de décision sur un ensemble de critères (Grabisch, Roubens, 2000). Elle permet de pondérer l'intérêt de chaque sous-ensemble de critères au lieu de pondérer chaque critère indépendamment des autres, comme le ferait une somme pondérée.

Dans notre cas, chaque source a un intérêt variable pour un candidat donné qui dépend du nombre de sources avec lesquelles elle est en accord et de la qualité de ces sources. Par exemple, considérer une nouvelle source de qualité pour un candidat impliquant déjà un grand nombre de sources de bonne qualité aura moins d'importance que considérer cette source pour un candidat impliquant des sources de mauvaise qualité. Dans notre cas, l'implication d'une source $S_i$ dans le candidat $Cand$ est considérée comme un critère. Cette implication est évaluée par la fonction $implic(S_i, Cand)$. L'implication ordonne les sources localement en fonction d'un candidat donné.

#### *5.4.1. Implication des sources pour un candidat donné*

La fonction $implic(S_i, Cand)$ détermine la force de l'implication de la source $S_i$ dans la construction du candidat $Cand$. Cette force est fonction des correspondances associées au sommet de la source $S_i$ dans le candidat $Cand$. En effet, nous considérons que plus un sommet de la source $S_i$ est lié avec des correspondances fortes vers les autres sommets du candidat $Cand$, plus la source est impliquée dans la construction de ce candidat.

Prenons l'exemple présenté sur la figure 4. Nous remarquons dans le candidat "Cand1" que le sommet "Triticum" provenant de la source Agrovoc a une implication plus forte que les autres sommets. En effet, deux correspondances lient le sommet provenant d'Agrovoc vers les deux autres sommets du candidat. Les sommets de sources TaxRef et NCBI n'étant liés qu'à un seul sommet, l'implication de ces sources dans le candidat "Cand1" est moindre. Pour représenter formellement cette implication, nous sommons les degrés de fiabilité des correspondances liant le sommet provenant de la source et appartenant au candidat. Afin de normaliser cette valeur, nous divisons cette somme par l'implication maximale possible, c'est-à-dire les correspondances avec un degré de fiabilité de 1 vers tous les sommets du candidat. En d'autres termes, nous la divisons par le nombre de sources considérées moins un. Nous calculons l'implication d'une source $S_i$ par rapport à un candidat sommet $Cand$ en utilisant l'équation 6. Rappelons qu'une source $S_i$ est un graphe ayant comme ensemble de sommet $V_{S_i}$. Un candidat sommet $Cand$ est un graphe non-orienté $Cand = (V_{Cand}, E_{Cand}, valueE)$.

La fonction $implic(S_i, Cand)$ utilisée pour évaluer la source $S_i$ est définie par l'équation suivante :

$$implic(S_i, Cand) = \frac{\sum_{\substack{e \in E_{Cand} \\ e=(oe_i, oe_j) \ avec \ oe_i \in V_{S_i}}} valueE(e)}{N-1} \quad (6)$$

Si nous reprenons l'exemple du candidat "Cand1" de la figure 4, l'implication de la source "Agrovoc" dans ce candidat peut être définie de la manière suivante :

$$implic(Agrovoc, Cand1) = \frac{0,9 + 0,6}{3-1} = \frac{1,5}{2} = 0,75 \quad (7)$$

Alors que l'implication de la source TaxRef dans ce même candidat peut être évaluée de la manière suivante :

$$implic(TaxRef, Cand1) = \frac{0,9}{3-1} = \frac{0,9}{2} = 0,45 \quad (8)$$

Nous ne considérons ici que la correspondance qui a un degré de fiabilité de 0,9 puisque c'est la seule qui implique le sommet provenant de la source TaxRef. De la même manière, nous définissons l'implication de la source NCBI dans le candidat "Cand1" de la manière suivante :

$$implic(NCBI, Cand1) = \frac{0,6}{3-1} = \frac{0,6}{2} = 0,3 \quad (9)$$

Cette notion d'implication est particulièrement pertinente dans cet exemple puisque nous observons que le sommet provenant d'Agrovoc est central dans la construction de ce candidat. Les deux autres sommets n'ont pas de correspondance entre eux. Si ce sommet n'était pas présent, alors le candidat n'existerait tout simplement pas. Il est donc cohérent que l'implication de la source Agrovoc soit bien plus grande que l'implication des deux autres sources pour ce candidat.

Si un candidat $Cand$ n'a qu'un seul sommet, et donc aucune correspondance à utiliser, alors nous définissons l'implication de la source $S_i$ de la manière suivante : $implic(S_i, Cand) = \frac{1}{N-1}$. Rappelons que $N$ est le nombre de sources alignées dans le processus de fusion.

### *5.4.2. Agrégation de l'implication des sources*

Cette fonction d'implication ordonne localement les sources pour construire des sous-ensembles de sources. Soit $N$ le nombre de sources considérés. $S_1$ est la source qui a la plus faible implication et $S_N$ celle qui a la plus forte implication dans le candidat $Cand$. Nous pouvons construire $N$ sous-ensembles.

– Le sous-ensemble constitué des $N$ sources. Soit $L_1$ cet ensemble de sources. Nous voulons agréger toutes les valeurs d'implication des sources de $L_1$. Ainsi $\forall S_j \in L_1$ nous agrégeons les valeurs d'implication en dessous de la borne $B_1 = implic(S_1, Cand)$.

– Le sous-ensemble constitué des $N-1$ sources ayant l'implication la plus élevée. Soit $L_2$ ce sous-ensemble. Nous voulons agréger toutes les valeurs d'implication des

sources de $L_2$. $\forall S_j \in L_2$ nous agrégeons les valeurs en dessous de la borne $B_2 = implic(S_2, Cand)$.

– ...

– Le sous-ensemble constitué des deux sources ayant l'implication la plus élevée pour le candidat donné. Soit $L_{N-1}$ ce sous-ensemble. $\forall S_j \in L_{N-1}$ nous agrégeons les valeurs en dessous de la borne $B_{N-1} = implic(S_{N-1}, Cand)$.

– Le sous-ensemble constitué de la source ayant l'implication la plus élevée pour le candidat donné. Soit $L_N$ ce singleton. $\forall S_j \in L_N$ nous agrégeons les valeurs en dessous de la borne $B_N = implic(S_N, Cand)$ .

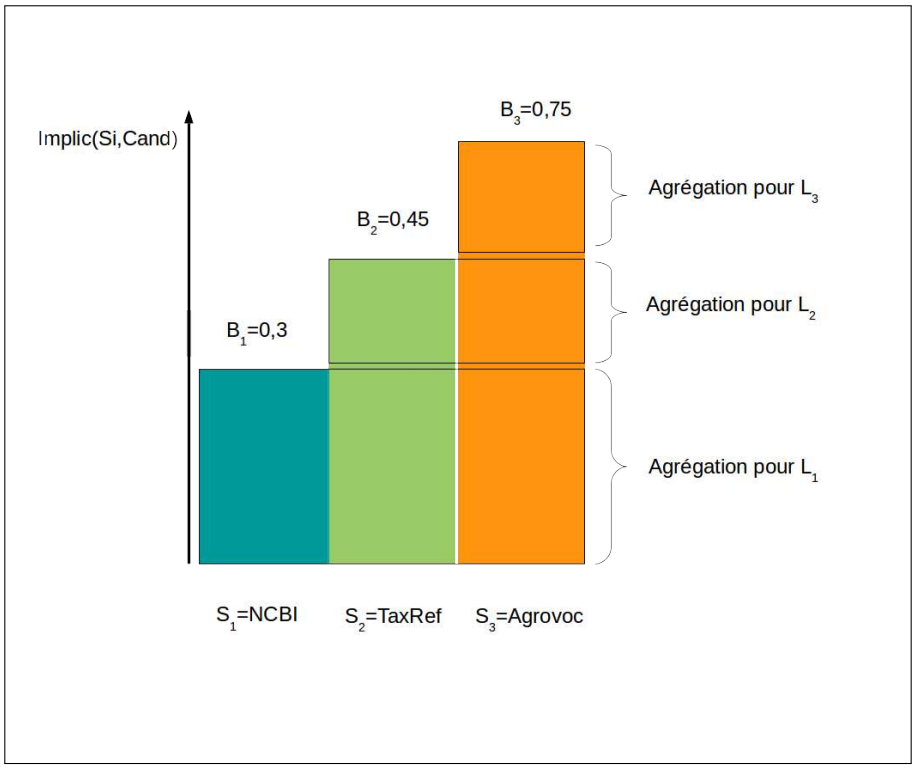


*Figure 5. Illustration de l'agrégation de l'implication des sources*

L'objectif est d'agréger l'implication des sous-ensembles de sources. La figure 5 illustre cette agrégation pour le candidat $Cand1$. L'agrégation calcule la somme des surfaces présentées dans la figure 5 pour trois sous-ensembles de sources : $L_1 = \{Agrovoc, TaxRef, NCBI\}$, $L_2 = \{Agrovoc, TaxRef\}$ et $L_3 = \{Agrovoc\}$.

L'intégrale de Choquet agrège l'implication des sous-ensembles de sources (notée $L_i$) en suivant l'équation :

$$trust_{choquet}(Cand) = B_1\mu(L_1) + \sum_{i=2}^{N}(B_i - B_{i-1})\mu(L_i) \quad (10)$$

avec

– $N$ le nombre de sources impliquées dans le candidat $Cand$,

– $S_i$ les sources ordonnées par ordre croissant en fonction de leur implication dans le candidat $Cand$.

– $B_i = implic(S_i, Cand)$ les seuils nécessaires pour calculer l'agrégation. $B_i < B_{i+1}$,

– $L_i = \{S_j | j \in \{i, ..., N\}$ les sous-ensembles de sources. $\forall S_j \in L_i$ nous avons $implic(S_j, Cand) \geq B_i\}$.

– L'intégrale de Choquet utilise la fonction $\mu$ pour pondérer les sous-ensembles de sources en fonction de leur intérêt.

Cette intégrale calcule le score de confiance du candidat $Cand$ à partir des sous-ensembles de sources $S_1, ..., S_N$ impliqués dans le candidat.

### *5.4.3. Intérêt des sources*

La dernière fonction à définir pour utiliser l'intégrale de Choquet est la fonction $\mu$ qui représente l'intérêt d'un ensemble de sources dans la prise de décision. Des priorités entre les sources sont ainsi définies. Nous favorisons par exemple les candidats impliquant la source "TaxRef" plutôt que ceux impliquant "Agrovoc". Pour ce faire, nous définissons une fonction $Q(S_i)$ retournant une valeur, comprise entre 0 et 1, représentant la qualité de la source $S_i$. L'intérêt d'une source sera fonction de sa qualité.

Dans notre exemple, nous considérons trois sources de qualité différente. Pour évaluer $Q(S_i)$, nous utilisons les scores de qualité définis avec nos experts lors de la construction de notre référence sur la taxonomie des blés (voir section Expérimentation). La source TaxRef, qui est une référence nationale dans ce domaine, a un score de qualité fixé à 0,9. Du fait de son processus de validation manuel et de sa mise à jour régulière, la source NCBI a également un score relativement élevé fixé à 0,8. La source Agrovoc, quant à elle, a un score de qualité de 0,6 puisque des travaux (Soergel *et al.*, 2004) ont montré qu'elle contient un certain nombre d'erreurs. Elle reste néanmoins une source intéressante.

Nous devons définir la fonction $\mu(L_i)$ caractérisant l'intérêt d'un sous-ensemble de sources ($L_i = \{S_i, ..., S_N\}$) en fonction de leur qualité. De cette façon, nous prennons en compte la diversité et la multiplicité des sources. Un candidat impliquant un grand nombre de sources de mauvaise qualité pourra être considéré aussi pertinent qu'un candidat impliquant peu de sources de très bonne qualité. Nous considérons non seulement que chaque source a un intérêt variable mais aussi que l'évolution de

l'intérêt des sources n'est pas linéaire. Cette non linéarité prend en compte une évolution variable de l'intérêt des sources. Nous considérons, par exemple, que si un candidat implique déjà un grand nombre de sources de bonne qualité, alors l'ajout d'une nouvelle source de bonne qualité dans la définition du candidat ne va pas augmenter significativement sa confiance. Notre intuition sur cette répartition non linéaire est qu'il existe un point représentatif à partir duquel l'intérêt des sources va croître significativement. Ce point d'explosion [6] est spécifique au problème étudié. Il dépend non seulement du nombre de sources considérées mais aussi de la nécessité de favoriser la qualité ou non des sources. De plus, l'intensité de l'explosion est aussi spécifique au problème étudié.

Nous définissons la fonction $\mu(L_i)$ tel que $L_i$ est un sous-ensemble des sources :

$$\mu(L_i) = \frac{\lambda(\sum_{k=i}^{N} Q(S_k)) - \lambda(0)}{\lambda(\sum_{k=1}^{N} Q(S_k)) - \lambda(0)} \tag{11}$$

$$\lambda(x) = arctan(\frac{x - x_0}{\gamma}) \tag{12}$$

La fonction $Q(S_k)$ retourne la qualité de la source $S_k$. L'équation $\sum_{k=1}^{N} Q(S_k)$ calcule la somme des scores de qualité de toutes les sources considérées dans le processus. Ce qui donne pour notre exemple la somme suivante :

$$\sum_{k=1}^{N} Q(S_k) = 0,9 + 0,8 + 0,6 = 2,3 \tag{13}$$

De la même façon, l'équation $\sum_{k=i}^{N} Q(S_k)$ calcule la somme des scores de qualité des sources du sous-ensemble $L_i$.

Cette fonction $\mu(L_i)$ représente l'intérêt du sous-ensemble de sources $L_i$. La fonction $\lambda(x)$ est inspirée de la fonction de répartition de la loi gamma. Ainsi elle suit une répartition qui respecte notre intuition sur l'évolution de l'intérêt des sources.

Dans la fonction $\lambda(x)$, deux paramètres sont utilisés. Le premier, $x_0$, définit le point d'explosion, point à partir duquel l'intérêt des sources augmente particulièrement. Le deuxième paramètre est la valeur $\gamma$ qui définit l'indice de linéarité de la courbe. Plus la valeur de $\gamma$ tend vers 0, plus l'intérêt des sources est crénelé, c'est-à-dire qu'il est très proche de 0 en dessous de $x_0$ et très proche de 1 au dessus. À l'inverse, plus $\gamma$ s'approche de $\sum_{k=i}^{N} Q(S_k)$ (somme des scores de qualité de toutes les sources considérées) plus la courbe est linéaire.

Pour notre cas d'étude, nous devons définir les deux paramètres $x_0$ et $\gamma$. Nous définissons arbitrairement le point d'explosion à 50 % de la qualité disponible. Soit $x_0 = 2,3/2 = 1,15$.

6. Point auquel la dérivée $\mu'$ est à son maximum

Toujours arbitrairement, nous définissons un taux de linéarité de la répartition à 20 %. Nous définissons $\gamma = 2,3 * 0,20 = 0,46$.

La figure 6 présente la répartition de la fonction $\mu(x)$ en fonction de nos paramètres.

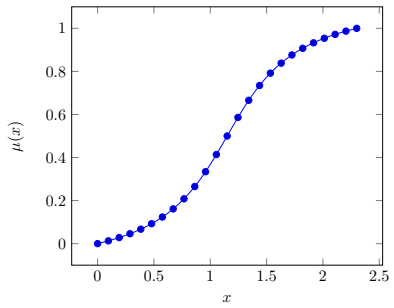


*Figure 6. Répartition de $\mu(x)$ avec les paramètres $\sum_{k=i}^{N} Q(S_k) = 2,3$, $x_0 = 1,15$ et $\gamma = 0,46$*

### *5.4.4. Application de l'intégrale de Choquet pour le calcul de la confiance des candidats*

Le calcul de la confiance du candidat "Cand1" en utilisant l'intégrale de Choquet est :

$$\begin{aligned} trust_{choquet}(Cand1) =& implic(NCBI, Cand1) * \mu(Agrovoc, TaxRef, NCBI) \\ &+ [implic(TaxRef, Cand1) - implic(NCBI, Cand1)] \\ &* \mu(Agrovoc, TaxRef) \\ &+ [implic(Agrovoc, Cand1) - implic(TaxRef, Cand1)] \\ &* \mu(Agrovoc) \\ =& 0,3 * \mu(Agrovoc, TaxRef, NCBI) \\ &+ (0,45 - 0,3) * \mu(Agrovoc, TaxRef) \\ &+ (0,75 - 0,45) * \mu(Agrovoc) \\ =& 0,3 * 1 + 0,15 * 0,77 + 0,3 * 0,14 = 0,46 \end{aligned} \tag{14}$$

De la même manière, nous appliquons l'intégrale de Choquet pour calculer la confiance du candidat "Cand2" présenté sur la figure 4. Nous obtenons les implications suivantes :

- $implic(Agrovoc, Cand2) = (0,7 + 0,8)/2 = 0,75$
- $implic(TaxRef, Cand2) = (0,8 + 0,9)/2 = 0,85$

– $implic(NCBI, Cand2) = (0,9 + 0,7)/2 = 0,8$

Nous obtenons alors le calcul suivant :

$$\begin{aligned} trust_{choquet}(Cand2) =& 0,75 * \mu(Agrovoc, TaxRef, NCBI) \\ &+ (0,8 - 0,75) * \mu(TaxRef, NCBI) \\ &+ (0,85 - 0,8) * \mu(TaxRef) \\ =& 0,75 * 1 + 0,05 * 0,86 + 0,05 * 0,29 \\ =& 0,81 \end{aligned} \quad (15)$$

L'intégrale de Choquet est applicable de la même manière sur un candidat arc, mais cela pose un problème : il n'existe pas d'implication d'une source dans la constitution d'un candidat arc car ces candidats ne sont pas générés à partir de correspondances. Nous considérons donc que chaque source est impliquée à un niveau maximum si elle est présente dans le candidat arc. Le calcul de la confiance d'un candidat arc en utilisant l'intégrale de Choquet telle que définie précédemment revient à calculer la fonction $\mu(L_i)$ en considérant $L_i$ comme étant le sous-ensemble des sources impliquées dans le candidat arc. Par exemple nous considérons le candidat "Cand arc2" dans lequel apparaît la relation "hasHigherRank" entre les candidat sommets "Cand2" et "Cand1". Ce candidat arc implique les sources Agrovoc et NCBI. Si nous appliquons la démarche du calcul de l'intégrale de Choquet concernant ce candidat, nous obtenons les implications suivantes :

– $implic(Agrovoc, Candarc2) = 1$
– $implic(NCBI, Candarc2) = 1$

Par conséquent le calcul de l'intégrale de Choquet est le suivant :

$$\begin{aligned} trust_{choquet}(Candarc2) =& 1 * \mu(Agrovoc, NCBI) \\ =& 1 * 0,70 \\ =& 0,70 \end{aligned} \quad (16)$$

## 6. Évaluation

Nous présentons dans cette section les expérimentations que nous avons menées afin d'évaluer notre approche et notamment son implémentation avec l'outil Muskca.

Nous souhaitons tout d'abord évaluer la capacité de notre approche à générer une base de connaissances à partir de plusieurs sources non ontologiques.

D'une part, nous cherchons à valider notre hypothèse sur la pertinence d'un élément partagé par plusieurs sources. Pour cela, nous devons demander à des experts d'évaluer les candidats générés. D'autre part, nous voulons évaluer l'impact des scores de confiance pour aider un expert du domaine à sélectionner des candidats.

Notre expérimentation porte sur la construction d'une taxonomie agronomique des blés. Les taxonomies regroupent les espèces en fonction de leurs noms de taxons scientifiques organisés sous la forme d'une hiérarchie (espèce, genre, famille, ordre, classe, embranchement, règne). Il existe différentes taxonomies organisant les espèces vivantes en fonction de critères génétiques ou morphologiques. De plus les taxonomies évoluent en fonction de nouvelles découvertes. Pendant 10 000 ans de culture, de nombreuses formes de blé ont évolué sous l'effet de la sélection par l'homme. Cette diversité s'est traduite par une grande confusion dans la dénomination des espèces de blés. Notre cas d'usage porte donc sur la génération d'une base de connaissances représentant une taxonomie des espèces de blés. Cette taxonomie doit représenter un consensus en France. En effet, nous souhaitons que cette taxonomie puisse être utilisée à la fois par des agriculteurs et des agronomes.

Nous présentons dans la section suivante le cadre expérimental mis en place, puis les résultats de l'évaluation.

### *6.1. Mise en place du cadre expérimental*

Mettre en place notre approche sur un cas réel implique de définir le module ontologique et d'identifier les sources et les patrons d'extraction exploitant ces sources. Nous avons également défini un processus d'interaction avec les experts pour valider notre approche.

#### *6.1.1. Module ontologique : AgronomicTaxon*

Conformément au processus défini dans notre méthodologie, nous avons construit un module ontologique (Roussey *et al.*, 2013). Ce module, intitulé AgronomicTaxon, définit des bornes du domaine d'étude. Il définit aussi les entités de haut niveau nécessaires à la représentation d'une taxonomie scientifique [7].

AgronomicTaxon réutilise un certain nombre de patrons de conception provenant d'autres ontologies ou du portail de patrons de conception ontologique ODP [8]. Nous retrouvons par exemple le patron de conception consacré à la représentation de taxonomie : LinnaeanTaxonomy [9].

Trois sources ont été sélectionnées comme étant pertinentes pour enrichir le module : Agrovoc, TaxRef et NCBI.

Des patrons de transformation sont appliqués sur chacune des sources pour obtenir une BCS. Les détails et les implémentations de ces transformation sont disponibles sous la forme de projet Java accessible aux adresses :

– https://github.com/Murloc6/T2RKB pour Agrovoc.

7. Une description détaillée du module est disponible sur le site web https://sites.google.com/site/agriontology/home/irstea/agronomictaxon

8. Ontology Design Patterns - http://ontologydesignpatterns.org/

9. http://ontologydesignpatterns.org/wiki/Submissions:LinnaeanTaxonomy

– https://github.com/Murloc6/TaxRef2RKB/ pour TaxRef.
– https://github.com/Murloc6/NCBI2RKB/ pour NCBI.

Le tableau 1 résume les éléments extraits de chaque source.

*Tableau 1. Quantités d'éléments extraits des sources pour la taxonomie des blés*

| Source | Sommets Classes | Sommets Individus | Sommets Labels | Arc hasHigherRank | Arc Type |
|---|---|---|---|---|---|
| Agrovoc | 5 | 21 | 108 | 27 | 16 |
| TaxRef | 6 | 17 | 89 | 16 | 17 |
| NCBI | 15 | 99 | 157 | 97 | 91 |

Nous observons dans ce tableau que la source NCBI est bien plus fournie que les deux autres sources avec 15 classes extraites et 99 individus représentant des taxons. Ces individus sont typés par la classe $Taxon$. Nous avons extrait 108 labels pour 21 individus de type $Taxon$ et 5 classes provenant d'Agrovoc. Nous avons restreint cette extraction aux labels français et anglais, puisque ce sont les seules langues disponibles dans les autres sources.

### *6.1.2. Données de référence*

Une fois les BCS générées, nous avons demandé aux experts d'analyser leurs éléments, afin de déterminer s'ils devaient être représentés dans la base de connaissances finale. Ainsi nous évaluons si les éléments validés par les experts apparaissent bien dans au moins un candidat de la base de connaissances finale. Un autre intérêt de cette approche est de pouvoir déterminer s'il existe un consensus au sein des sources et si les candidats générés contiennent bien les éléments communs aux trois sources.

Nous avons implémenté une interface web pour demander aux experts l'intérêt des éléments de chaque source. En d'autres termes, les experts valident les individus de type $Taxon$ issus des trois BCS. Pour chacun des taxons nous leur avons posé les questions suivantes :

1. **Est-ce que le taxon appartient au domaine ?** Nous demandons dans un premier temps si l'élément présenté est vraiment un taxon (un élément de la taxonomie). Nous voulons également savoir s'il est dans le périmètre de la base de connaissances, c'est à dire Triticum. Cette question nous permettra de savoir si cet élément doit être retrouvé dans la base de connaissances finale.

2. **Est-ce que le taxon est plus spécifique qu'un autre taxon ?**

Nous voulons valider les arcs "hasHigherRank" entre deux taxons de la source. Nous demandons donc si le premier taxon est un bien un fils du second en considérant une hiérarchie taxonomique.

3. **Est-ce que ce taxon appartient au rang donné ?** Il y a parfois des informations concernant le rang du taxon dans la source. Nous souhaitons donc valider cette extraction et définir le type du taxon : est-ce une espèce, un genre, etc. Cette question

permet de valider la relation d'instanciation (arc $rdf : type$) entre l'individu de type $Taxon$ et une classe du module (sous classe de la classe $Taxon$).

L'interface web propose trois possibilités de réponse pour chacune des questions : Valide, Non valide, Ne sait pas. Les individus ont été validés par la première question, les arcs "hasHigherRank" par la deuxième et les arcs "rdf:type" par la dernière. À la fin de la validation, nous avons une liste d'éléments ontologiques validés par les experts. Nous les comparons aux candidats générés par Muskca.

Pour analyser les candidats générés par notre approche, nous souhaitons analyser dans quelle mesure nos candidats s'appuient sur des éléments des sources validés par les experts. Tout d'abord, nous avons vérifié que les experts étaient d'accord sur les éléments valides des sources. Nous considérons qu'il existe un accord entre les experts quand au moins deux experts ont validé le même élément et que le troisième a sélectionné l'option "Ne sait pas". Nous avons donc calculé le rapport entre le nombre d'experts et leur nombre de validations et nous obtenons une valeur de **0,82**. Nous avons également calculé le score de Fleiss Kappa (Fleiss, Cohen, 1973) sur les validations des experts. Nous obtenons alors un score de Fleiss Kappa de **0,69**. Ces deux valeurs montrent que les experts s'accordent la plupart du temps sur la validation des éléments. Nous utilisons donc les retours des experts comme des données de référence permettant d'évaluer les candidats.

Les éléments validés par les experts composent un graphe intitulé $Gold$ constitué d'une sous-partie des graphes représentant les BCS. Le graphe $Gold$ se défini tel que :

$$\begin{aligned} Gold &= (V_{Gold} \subseteq (V_{TaxRef} \cup V_{NCBI} \cup V_{Agrovoc} \\ &, \quad E_{Gold} \subseteq (E_{TaxRef} \cup E_{NCBI} \cup E_{Agrovoc}) \end{aligned}$$

### *6.2. Évaluation de la qualité des candidats générés*

Cette section décrit l'évaluation de la qualité de la base de connaissances finale générée par notre approche. Pour cela nous allons utiliser le jeu de données de référence généré avec l'aide de trois experts, présenté dans le paragraphe précédent. Nous comparerons les éléments ontologiques impliqués dans les candidats et les éléments validés par les experts. Cette comparaison repose sur les calculs de précision, rappel et F-mesure. Dans cette section nous présentons dans un premier temps notre stratégie d'évaluation. Ensuite, les résultats sont présentés et analysés.

#### *6.2.1. Stratégie de validation*

Notre objectif ici est de vérifier la qualité des candidats générés et l'impact de la fonction de confiance utilisée pour les ordonner. Notre approche de fusion des 3 BCS présentées précédemment génère un ensemble de candidats sommets et de candidats arcs.

Soit $BC_{final}$ l'ensemble des candidats générés par le processus de fusion.
$BC_{final} = (CANDS_{final}, CANDAfinal)$.

Rappelons que

– un candidat sommet est défini comme un graphe ayant des étiquettes d'arc valuées : $\forall CandS_i \in CANDS_{final}, CandS_i = (V_{CandS_i}, E_{CandS_i}, valueE)$

– et un candidat arc se définit comme un graphe biparti orienté :
$\forall CandA_j \in CANDAfinal, CandA_j = (V_{CandA_j}, E_{CandA_j})$

Nous allons maintenant définir les candidats valides.
Soit $BC_{valid} = (CANDS_{valid}, CANDAvalid)$ l'ensemble de ces candidats. Nous considérons un candidat sommet $CandS_i$ valide si tous ses sommets ont été validés par les experts :
$CandS_i \in CANDS_{valid}$ ssi $\forall v \in V_{CandS_i}, v \in V_{Gold}$.

Nous considérons un candidat arc $CandA_j$ valide si tous ses sommets et tous ses arcs ont été validés par les experts :
$CandA_j \in CANDA_{valid}$ ssi $\forall v \in V_{CandA_j}, v \in V_{Gold}$ et $\forall e \in E_{CandA_j}, e \in E_{Gold}$

A partir des candidats validés et des candidats générés, nous calculons les valeurs de précisions, rappel et f-mesure. Nous définissons la mesure de précision de la manière suivante :

$$\begin{aligned} Precision(BC_{final}) \quad = \quad & \\ & \frac{|CANDS_{final} \cap CANDS_{valid}|}{|CANDS_{final}| + |CANDA_{final}|} \\ + \quad & \frac{|CANDA_{final} \cap CANDA_{valid}|}{|CANDS_{final}| + |CANDA_{final}|} \end{aligned} \tag{17}$$

Nous considérons dans cette formule que $BC_{final}$ est l'ensemble de tous les candidats générés. $BC_{valid}$ est l'ensemble des candidats validés. Le numérateur de cette précision est le nombre de candidats générés et valides. Le dénominateur est le nombre de candidats générés. Plus la précision est élevée et plus les candidats générés sont considérés comme valides par les experts.

Nous calculons aussi le rappel qui représente l'exhaustivité des résultats. Ce rappel évalue la proportion d'éléments validés par les experts apparaissant dans les candidats générés. Nous calculons le rappel de la manière suivante :

$$Rappel(BC_{final}) = \frac{\left|V_{Gold} \cap \left(\bigcup_{\forall CS_i \in CANDS_{final}} V_{CS_i}\right)\right|}{|V_{Gold}| + |E_{Gold}|} + \frac{\left|E_{Gold} \cap \left(\bigcup_{\forall CA_j \in CANDA_{final}} E_{CA_j}\right)\right|}{|V_{Gold}| + |E_{Gold}|} \quad (18)$$

Le numérateur compte le nombre d'éléments (sommets ou arcs) présents dans le jeu de données de référence $Gold$ et qui apparaissent aussi dans un candidat de la base de connaissances finale $BC_{final}$. Le dénominateur est le nombre d'éléments dans le jeu de données de référence. Plus ce rappel sera élevé et plus les éléments validés par les experts seront présents dans des candidats.

Pour calculer une valeur unifiée, nous calculons la F-Mesure. Cette fonction agrège les valeurs de précision et de rappel.

$$F - Mesure(BC_{final}) = \frac{2 * Precision(BC_{final}) * Rappel(BC_{final})}{Precision(BC_{final}) + Rappel(BC_{final})} \quad (19)$$

Nous avons mené plusieurs expérimentations afin d'évaluer les candidats que nous générons. Chacune de ces expérimentations utilise l'une des fonctions de calcul du score de confiance. Pour chaque fonction, nous calculons les mesures de précision, rappel et F-mesure.

### *6.2.2. Résultats*

Nous commençons par calculer la précision, le rappel et la F-mesure pour tous les candidats générés.

*Tableau 2. Résultats sur la taxonomie des blés sans filtrage*

| | **Précision** | **Rappel** | **F-Mesure** |
|---|---|---|---|
| **Individus** | 0,87 | 1 | 0,93 |
| **HasHigherRank** | 0,74 | 0,99 | 0,85 |
| **Types** | 0,45 | 1 | 0,62 |

Nous menons ensuite une expérimentation pour évaluer l'impact de chacune des fonctions de confiance sur le classement des candidats. Nous considérons d'abord $trust_{likelihood}$ puis $trust_{degree}$ et enfin $trust_{choquet}$. Pour chacune de ces expérimentations, nous appliquons plusieurs filtres sur le score de confiance. Nous appliquons un filtre sur les candidats par pas de 0,1 jusqu'à la valeur 0,9. Nous calculons les 3 mesures pour chacun des pas. De cette manière, il est possible de comparer l'intérêt

des fonctions de confiance. La fonction de confiance qui aura une meilleure représentation du consensus verra sa précision être d'autant plus élevé que le filtre sera élevé lui aussi.

Dans le cadre de chacune des expérimentations, nous avons également évalué notre approche sur les différents types de candidats générés :

– les "candidats sommets individus" qui représentent les individus de type $Taxon$ extraits des différentes sources ;

– les "candidats arcs $hasHigherRank$" qui sont les relations "hasHigherRank" entre deux taxons ;

– les "candidats arcs $type$" qui sont les relations d'instanciation des individus de type $Taxon$. Ces relations identifient leur rang taxonomique.

Nous ne considérons pas ici les "candidats de type classe" car ils n'ont pas été validés par les experts.

#### 6.2.2.1. Analyse de l'ensemble des candidats générés

Le premier fait notable sur ces résultats est constitué par les scores encourageants sur le tableau 2. Par exemple, la F-mesure des candidats individus est à 0,93. Ce score signifie que notre proposition génére des candidats de bonne qualité, même sans filtrage. Le score le moins élevé ici concerne les candidats $type$ avec une précision de 0,45. Ceci s'explique par de nombreuses erreurs dans les sources concernant la catégorisation des taxons. Les taxons et leurs relations $hasHigherRank$ entre taxons sont corrects et consensuels mais le rang taxonomique des taxons n'est pas consensuel.

#### 6.2.2.2. Fonction de confiance : $trust_{likelihood}$

Nous présentons dans cette section les résultats obtenus en utilisant la fonction de confiance $trust_{likelihood}$. Les tableaux 3, 4, 5, présentent ces résultat en fonction du seuil utilisé pour le filtrage. Ce seuil considére uniquement les candidats ayant un score supérieur ou égal à ce seuil.

*Tableau 3. Expérimentations $trust_{likelihood}$ : individus*

| Seuil | Précision | Rappel | F-Mesure |
|---|---|---|---|
| **0,1** | 0,87 | 1 | 0,93 |
| **0,2** | 0,87 | 1 | 0,93 |
| **0,3** | 0,87 | 1 | 0,93 |
| **0,4** | **0,99** | 0,63 | 0,77 |
| **0,5** | 0,99 | 0,63 | 0,77 |
| **0,6** | 0,99 | 0,63 | 0,77 |
| **0,7** | 1 | 0,60 | 0,75 |
| **0,8** | 1 | 0,60 | 0,75 |
| **0,9** | 1 | 0,60 | 0,75 |

*Tableau 4. Expérimentations $trust_{likelihood}$ : $hasHigherRank$*

| Seuil | Précision | Rappel | F-Mesure |
|---|---|---|---|
| **0,1** | 0,74 | 0,99 | 0,84 |
| **0,2** | 0,74 | 0,99 | 0,84 |
| **0,3** | **0,74** | 0,99 | 0,84 |
| **0,4** | 0,58 | 0,43 | 0,50 |
| **0,5** | 0,58 | 0,43 | 0,50 |
| **0,6** | **0,58** | 0,43 | 0,50 |
| **0,7** | **1** | 0,21 | 0,35 |
| **0,8** | 1 | 0,21 | 0,35 |
| **0,9** | 1 | 0,21 | 0,35 |

*Tableau 5. Expérimentations $trust_{likelihood}$ : $type$*

| Seuil | Précision | Rappel | F-Mesure |
|---|---|---|---|
| **0,1** | 0,45 | 1 | 0,62 |
| **0,2** | 0,45 | 1 | 0,62 |
| **0,3** | **0,45** | 1 | 0,62 |
| **0,4** | 0,48 | 0,48 | 0,46 |
| **0,5** | 0,48 | 0,45 | 0,46 |
| **0,6** | 0,48 | 0,45 | 0,46 |
| **0,7** | 0,55 | 0,37 | 0,44 |
| **0,8** | 0,55 | 0,37 | 0,44 |
| **0,9** | 0,55 | 0,37 | 0,44 |

Les résultats obtenus grâce à la fonction $trust_{likelihood}$ sont intéressants, notamment ceux présentés dans le tableau concernant les individus (tableau 3). Nous voyons qu'ici la précision est à 0,99 à partir du seuil 0,4. Ceci implique que si un candidat sommet individu est composé d'éléments provenant d'au moins deux sources, il est très probablement validé par les experts. Ce score est plus mitigé pour les candidats arcs $type$ (tableau 5) puisque la précision est à 0,45. Ceci s'explique par le fait que les relations d'instanciation ("instanceOf") sont souvent spécifiques à une source et n'apparaissent que rarement dans une autre. Ceci s'observe d'autant plus que la précision augmente peu par rapport à l'augmentation du seuil. Un point important dans ces résultats apparaît dans le tableau des candidats arcs $hasHigherRank$ (tableau 4) représentant la relation "hasHigherRank". La précision est de 0,74 pour les seuils de 0,1 à 0,3, puis de 0,58 jusqu'au seuil 0,6 et passe directement à 1 ensuite. Cette progression s'explique par le fait que beaucoup de candidats $hasHigherRank$ considérés comme valides n'apparaissent que dans une seule source. Ils sont donc rejetés à un niveau de seuil impliquant qu'il y ait au moins 2 sources pour considérer le candidat, ce qui explique la chute. Néanmoins la remontée à 1 à la fin signifie que tous les candidats $hasHigherRank$ qui apparaissent dans 3 sources sont validés par les experts.

Globalement nous observons une augmentation de la précision avec l'augmentation du seuil dans les différents tableaux. Ainsi, notre hypothèse de départ est vérifiée. Si un élément est présent dans plusieurs sources, alors il est de meilleure qualité.

6.2.2.3. Fonction de confiance : $trust_{degree}$

*Tableau 6. Expérimentation $trust_{degree}$ : individus*

| Seuil | Précision | Rappel | F-Mesure |
|---|---|---|---|
| **0,1** | 0,99 | 0,6 | 0,77 |
| **0,2** | 0,99 | 0,63 | 0,77 |
| **0,3** | **1** | 0,62 | 0,77 |
| **0,4** | 1 | 0,60 | 0,75 |
| **0,5** | 1 | 0,60 | 0,75 |
| **0,6** | 1 | 0,60 | 0,75 |
| **0,7** | 1 | 0,26 | 0,41 |
| **0,8** | 1 | 0,26 | 0,41 |
| **0,9** | 1 | 0,26 | 0,41 |

La fonction de confiance $trust_{degree}$ pour les candidats arcs étant la même que $trust_{likelihood}$, nous ne présentons ici que les résultats concernant les candidats individus.

En observant les résultats de cette fonction $trust_{degree}$ présentés dans le tableau 6 et en les comparant avec ceux de la fonction $trust_{likelihood}$ (tableau 3), nous observons plusieurs phénomènes. Nous remarquons tout d'abord une augmentation plus significative de la précision dès le seuil de 0,1. Néanmoins, le rappel est lui fortement impacté par ce seuil. Ceci s'explique par l'effet discriminant des correspondances. Les candidats ayant un score de confiance élevé sont ceux qui impliquent beaucoup de correspondances. Les candidats impliquant trois sources n'utilisent pas forcément beaucoup de correspondances. Les éléments ontologiques ne sont alors pas fortement connectés entre eux par des correspondances, ce qui explique la diminution significative du rappel pour des seuils hauts. Les candidats impliquant moins de trois sources sont ici rejetés dès les seuils assez bas, bien que certains soient valides. Ceci explique la diminution rapide du rappel.

Nous en déduisons que l'utilisation des correspondances dans la fonction $trust_{degree}$ est discriminante. Les candidats n'impliquant pas beaucoup de correspondances (ou ayant des degrés de fiabilités faibles) ont un score de confiance bas. Néanmoins, nous continuons à vérifier notre hypothèse : plus un candidat utilise de correspondances (et donc plus son score $trust_{degree}$ est élevé), plus sa qualité est assurée. Nous l'observons avec la précision à 1 dés le seuil 0,3.

6.2.2.4. Fonction de confiance : $trust_{degree_p}$ avec propagation

Le tableau 7 présente les résultats obtenus en utilisant la fonction $trust_{degree_p}$ sur les candidats arcs. Cette fonction prend en compte la propagation de la confiance des

*Tableau 7. Expérimentations $trust_{degree_p}$ : sur les candidats arcs*

| | Seuils | Precision | Rappel | F-Measure |
|---|---|---|---|---|
| **Relations** | 0,6 | 0,68 | 0,51 | 0,58 |
| | 0,9 | 0,93 | 0,32 | 0,47 |
| **Types** | 0,6 | 0,77 | 0,39 | 0,52 |
| | 0,9 | 0,81 | 0,39 | 0,53 |

candidats sommets associés au candidat arc. Pour ces expérimentations, nous avons sélectionné uniquement deux seuils représentatifs. Si nous comparons ces résultats avec ceux obtenus pour $trust_{likelihood}$ (tableaux 4, 5) nous observons une nette amélioration pour les candidats arcs $type$. Nous remarquons aussi une amélioration des résultats pour les candidats arcs $hasHigherRank$, même si cette amélioration est moindre. Nous déduisons alors que la propagation de confiance d'un candidat sommet sur la confiance des candidats arcs associés améliore les résultats.

6.2.2.5. Fonction de confiance : $trust_{choquet}$

Nous utilisons maintenant la fonction de confiance $trust_{choquet}$. Cette fonction nécessite la définition de plusieurs paramètres. Nous reprenons les paramètres présentés dans la section 5.4.

*Tableau 8. Expérimentations $trust_{choquet}$ : individus*

| Seuil | Précision | Rappel | F-Mesure |
|---|---|---|---|
| **0,1** | 0,87 | 0,98 | 0,92 |
| **0,2** | 0,99 | 0,63 | 0,77 |
| **0,3** | 0,99 | 0,63 | 0,77 |
| **0,4** | 1 | 0,62 | 0,77 |
| **0,5** | 1 | 0,60 | 0,75 |
| **0,6** | 1 | 0,60 | 0,75 |
| **0,7** | 1 | 0,26 | 0,41 |
| **0,8** | 1 | 0,26 | 0,41 |
| **0,9** | 1 | 0,26 | 0,41 |

Les résultats obtenus avec la fonction $trust_{choquet}$ montrent que cette fonction intègre les avantages des deux autres fonctions de confiance. En analysant les résultats présentés dans le tableau 8, nous observons une augmentation de la précision sur les différents types de candidats. La diminution du rappel s'explique par l'effet discriminant des correspondances. Néanmoins, cet effet est diminué ici puisque compensé par une implication de l'intérêt des sources.

### *6.2.3. Analyse*

L'analyse des résultats obtenus en utilisant les différentes fonctions de confiance nous a permis d'observer majoritairement trois aspects du score de confiance.

*Tableau 9. Expérimentations $trust_{choquet}$ : $hasHigherRank$*

| Seuil | Précision | Rappel | F-Mesure |
|---|---|---|---|
| **0,1** | 0,74 | 0,99 | 0,85 |
| **0,2** | 0,70 | 0,96 | 0,81 |
| **0,3** | 0,583 | 0,43 | 0,50 |
| **0,4** | 0,58 | 0,43 | 0,50 |
| **0,5** | 0,58 | 0,43 | 0,50 |
| **0,6** | 0,58 | 0,43 | 0,50 |
| **0,7** | 0,58 | 0,43 | 0,50 |
| **0,8** | 0,30 | 0,42 | 0,351 |
| **0,9** | 1 | 0,21 | 0,35 |

*Tableau 10. Expérimentations $trust_{choquet}$ : $type$*

| Seuil | Précision | Rappel | F-Mesure |
|---|---|---|---|
| **0,1** | 0,45 | 1 | 0,62 |
| **0,2** | 0,39 | 0,98 | 0,56 |
| **0,3** | 0,48 | 0,45 | 0,46 |
| **0,4** | 0,48 | 0,45 | 0,46 |
| **0,5** | 0,48 | 0,45 | 0,46 |
| **0,6** | 0,48 | 0,45 | 0,46 |
| **0,7** | 0,48 | 0,45 | 0,46 |
| **0,8** | 0,23 | 0,43 | 0,30 |
| **0,9** | 0,56 | 0,37 | 0,44 |

Tout d'abord, l'évaluation de la fonction $trust_{likelihood}$ nous a permis de vérifier que notre hypothèse de départ, qui stipule que plus un candidat implique de sources, plus il est susceptible d'être validé, est une bonne intuition. Nous affirmons ici que le consensus est un critère de qualité. Le deuxième aspect important est que l'utilisation des correspondances dans la fonction de confiance a un effet discriminant assez fort sur le score des candidats. En effet les candidats, même s'ils impliquent plusieurs sources, n'impliquent pas forcément beaucoup de correspondances. Ces correspondances apportent en revanche un aspect de qualité supplémentaire au candidat. Nous observons ce phénomène dans les tableaux de résultats utilisant la fonction $trust_{degree}$. La fonction $trust_{choquet}$ considère les deux aspects présentés précédemment en donnant en plus la possibilité de paramétrer l'importance des sources. De cette manière, l'utilisation des correspondances est toujours forte mais elle est compensée par l'intérêt des sources impliquées dans le processus. L'utilisation de la fonction $trust_{degree_p}$, considérant le rayonnement, améliore les résultats par rapport à la fonction $trust_{degree}$. Ce phénomène n'ayant pas été pris en compte dans $trust_{choquet}$, il serait intéressant d'appliquer ce rayonnement sur cette fonction de confiance.

## 7. Conclusion et perspectives

Dans cet article, nous avons présenté notre méthode de fusion de plusieurs bases de connaissances. Cette méthode intègre deux scénarios de la méthodologie NeOn. Par rapport à l'existant, notre méthode est la première qui travaille avec plus de deux bases simultanément. En effet, nous souhaitons extraire de plusieurs sources les éléments consensuels, c'est-à-dire ceux qui sont communs à plusieurs sources. Ces éléments forment des candidats potentiels à intégrer dans la base de connaissances finale. Notre proposition évalue la confiance dans les éléments extraits des sources. Nous avons présenté plusieurs fonctions de confiance, tenant compte du nombre de sources considérées, des correspondances entre les éléments de ces sources et de la qualité des sources. Notre méthode de fusion travaille sur plusieurs formes de candidats représentant des classes, des individus et les propriétés liant deux individus ou un individu et une classe. Pour évaluer notre méthode nous avons travaillé avec trois bases de connaissances représentant les taxonomies des blés : le thésaurus Agrovoc, la taxonomie TaxRef et la taxonomie du NCBI. Notre évaluation a montré l'intérêt de la fonction $trust_{choquet}$, capable de tenir compte des correspondances entre les éléments et de la qualité des sources. Dans l'avenir nous devrons reconduire des expérimentations pour valider notre méthode de manière plus approfondie. Il serait aussi intéressant de détecter les incohérences entre candidats pour construire une base de connaissances cohérente.

*Remerciements*

*Nous tenons à remercier les trois experts mobilisés sur ce travail :*

- *Franck Jabot de l'Irstea Clermont-Ferrand*
- *Jacques Le Gouis de l'INRA Clermont-Ferrand*
- *Vincent Soulignac de l'IRSTEA Clermont-Ferrand*

## Bibliographie

Amarger F. (2015). *Vers un système intelligent de capitalisation de connaissances pour l'agriculture durable : construction d'ontologies agricoles par transformation de sources existantes*. Thèse de doctorat non publiée, Université de Toulouse 2 le Mirail.

Amarger F., Chanet J., Haemmerlé O., Hernandez N., Roussey C. (2014). SKOS Sources Transformations for Ontology Engineering: Agronomical Taxonomy Use Case. In *Metadata and Semantics Research: 8th Research Conference, MTSR 2014*, p. 314–328. Karlsruhe, Germany, Springer.

Amarger F., Chanet J., Haemmerlé O., Hernandez N., Roussey C. (2015). Construction d'une ontologie par transformation de systèmes d'organisation des connaissances et évaluation de la confiance. *Ingénierie des Systèmes d'Information*, vol. 20, n° 3, p. 37–61.

Amarger F., Chanet J.-P., Guillaume R., Haemmerlé O., Hernandez N., Roussey C. (2016). Détection de consensus entre sources et calcul de confiance fondé sur l'intégrale de choquet. In *27es journées francophones d'Ingénierie des Connaissances.* Montpellier, France, HAL.

Caracciolo C., Stellato A., Morshed A., Johannsen G., Rajbhandari S., Jaques Y. *et al*. (2013). The AGROVOC Linked Dataset. *Semantic Web*, n° 3, p. 341–348.

Dong X., Gabrilovich E., Heitz G., Horn W., Lao N., Murphy K. *et al*. (2014). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th acm sigkdd international conference on Knowledge Discovery and Data Mining*, p. 601–610. New York, USA.

Dong X. L., Gabrilovich E., Heitz G., Horn W., Murphy K., Sun S. *et al*. (2014). From data fusion to knowledge fusion. *Proceedings of the VLDB Endowment*, vol. 7, n° 10, p. 881–892.

Dragisic Z., Eckert K., Euzenat J., Faria D., Ferrara A., Granada R. *et al*. (2014). Results of the Ontology Alignment Evaluation Initiative 2014. In *9th ISWC workshop on ontology matching (OM)*, p. 61–104. Riva del Garda, Italy, HAL.

Federhen S. (2012, janvier). The NCBI Taxonomy database. *Nucleic Acids Research*, vol. 40, n° D1, p. D136–D143.

Fleiss J. L., Cohen J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*.

Gangemi A., Presutti V. (2009). Ontology design patterns. In *Handbook on ontologies*, p. 221–243. Springer.

Gargominy O., Tercerie S., Régnier C., Ramage T., Schoelinck C., Dupont P. *et al*. (2016). *TAXREF v10. 0, référentiel taxonomique pour la France: Méthodologie, mise en oeuvre et diffusion*. Rapport technique. Paris, Muséum National d'Histoire Naturelle.

Grabisch M., Roubens M. (2000). Application of the choquet integral in multicriteria decision making. *Fuzzy Measures and Integrals-Theory and Applications*, p. 348–374.

Guzmàn-Arenas A., Cuevas A.-D. (2010). Knowledge accumulation through automatic merging of ontologies. *Expert Systems with Applications*, vol. 37, n° 3, p. 1991–2005.

Jiménez-Ruiz E., Grau B. C. (2011). Logmap: Logic-based and scalable ontology matching. In *10th International Semantic Web Conference, proceedings, part i*, vol. 7031, p. 273–288. Bonn, Germany, Springer.

Lin J., Mendelzon A. O. (1999). Knowledge base merging by majority. In *Dynamic worlds*, p. 195–218. Springer.

Pottinger R. A., Bernstein P. A. (2003). Merging models based on given correspondences. In *29th international conference on Very Large Data Bases*, p. 862–873. Berlin, Germany.

Raunich S., Rahm E. (2014). Target-driven merging of taxonomies with Atom. *Information Systems*, vol. 42, p. 1–14.

Roussey C., Chanet J.-P., Cellier V., Amarger F. (2013). Agronomic taxon. In *Proceedings of the 2nd International Workshop on Open Data*, p. 5. Paris, France.

Shvaiko P., Euzenat J. (2013). Ontology matching: State of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, n° 1, p. 158–176.

Soergel D., Lauser B., Liang A., Fisseha F., Keizer J., Katz S. (2004). Reengineering thesauri for new applications: The AGROVOC example. *Journal of Digital Information*, vol. 4, n° 4.

Suárez-Figueroa M. C., Gómez-Pérez A., Motta E., Gangemi A. (2012). *Ontology engineering in a networked world*. Springer Science & Business Media.

Villazon-Terrazas B., Carmen Suarez-Figueroa M., Gomez-Perez A. (2010). A Pattern-Based Method for Re-Engineering Non-Ontological Resources into Ontologies. *International Journal on Semantic Web and Information Systems*, vol. 6, n° 4, p. 27–63.

Zaveri A., Rula A., Maurino A., Pietrobon R., Lehmann J., Auer S. (2016). Quality assessment for linked data: A survey. *Semantic Web*, vol. 7, n° 1, p. 63–93.