Détection de consensus entre sources et calcul de confiance fondé sur l'intégrale de Choquet

Fabien Amarger[1,2], Jean-Pierre Chanet[1], Romain Guillaume[2], Ollivier Haemmerlé[2], Nathalie Hernandez[2], Catherine Roussey[1]

[1] UR TSCF, Irstea, 9 av. Blaise Pascal CS 20085, 63172 Aubière,France
prénom.nom@irstea.fr

[2] IRIT, UMR 5505, Université de Toulouse, UT2J 5,allées Antonio Machado, F-31058 Toulouse
prénom.nom@univ-tlse2.fr

Résumé : Aujourd'hui de nombreux entrepôts sont disponibles sur le Web de données liées pour un même domaine d'intérêt. Ces entrepôts peuvent être de qualité variable ce qui rend difficile leur réutilisation. Dans cet article, nous présentons une approche permettant d'identifier la connaissance partagée par différents entrepôts en favorisant la connaissance issue de sources de qualité. L'approche repose sur l'utilisation de l'intégrale de Choquet. Notre approche a été évaluée dans le domaine de l'agriculture.

Mots-clés : ontologies, bases de connaissances, consensus, fusion, fonction de confiance, intégrale de Choquet

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 que de déterminer la qualité des 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. De cette manière, il devient possible d'exploiter les éléments communs ainsi que la complémentarité des sources tout en tenant compte des spécificités de chacune d'elles. Chaque élément extrait des différentes sources se voit attribuer un score de confiance. Nos travaux reposent sur l'hypothèse suivante :

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

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.

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 visant à 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 : le thésaurus Agrovoc de la FAO, la taxonomie des organismes vivants du NCBI[4] et la taxonomie française de référence TaxRef du Muséum d'Histoire Naturelle.

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* telle qu'elle est proposée dans les travaux 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 modèle 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. Les travaux les plus anciens et les plus emblématiques sont ceux du projet Prompt Noy & Musen (2003). Nous sommes intéressés par les travaux de fusion capables de générer automatiquement une nouvelle base de connaissances contenant les parties communes des sources. Pour comparer les travaux traitant de fusion de bases de connaissances, nous avons défini trois critères :

symétrique : la notion de fusion symétrique implique que les deux modèles à fusionner ont la même importance. Il est aussi possible d'utiliser une technique de fusion asymétrique pour privilégier un modèle plutôt qu'un autre. Dans ce cas, le résultat de la fusion suivra l'organisation du modèle privilégié ; dans notre cas, les sources ont été restructurées en fonction d'un modèle commun Amarger *et al.* (2015).

align : un processus d'alignement génère les correspondances utilisées dans la fusion des modèles. Certains travaux incluent le calcul de l'alignement dans la fusion (inclus) alors que d'autres considèrent l'alignement comme une entrée du processus (entrée) ; il existe de nombreux systèmes d'alignement. Il n'est ici pas nécessaire de proposer un nouveau système mais plutôt de réutiliser des travaux existants et éprouvés. Par conséquent, nous sommes intéressés par les travaux qui dissocient la fusion du calcul d'alignements.

confiance : suivant le processus de fusion appliqué, une confiance peut être associée aux éléments du modèle résultat de la fusion. Nous sommes intéressés par les systèmes de fusion capables de calculer des degrés de consensus entre sources.

4. National Center for Biotechnology Information

Approche	Symétrique	Align.	Confiance
Curé (2009)	sym	entrée	non
Guzmán-Arenas & Cuevas (2010)	sym	inclus	non
Raunich & Rahm (2014)	asym	entrée	non

TABLE 1 – Travaux sur la fusion

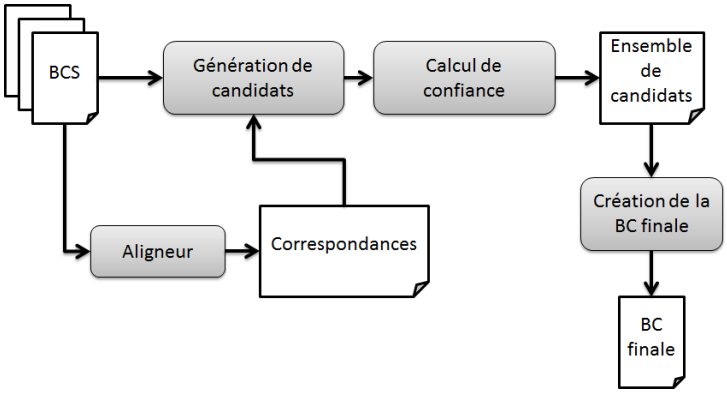

FIGURE 1 – Processus de fusion des bases de connaissances

Nous pouvons remarquer sur le tableau 1 que seule l'approche présentée dans (Raunich & Rahm, 2014) propose un processus asymétrique de fusion. Ce processus asymétrique définissant un modèle prioritaire par rapport à l'autre, certaines ambiguïtés susceptibles d'apparaître lors de la fusion peuvent être résolues automatiquement. Nous remarquons surtout que les processus de fusion présentés ici ne traitent pas de la notion de confiance. On note que les travaux de Guzmán-Arenas & Cuevas (2010) proposent une fonction de confusion pour évaluer la disparité entre deux contraintes de domaine incompatibles. La problématique de fusion étant toujours 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 dans ce cas-là. 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. 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 de fusion de bases de connaissances

Le processus de fusion que nous proposons est décomposé en plusieurs étapes présentées dans la figure 1. Ce processus prend en entrée les différentes BCS à fusionner et 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 connaissance finale.

Quatre étapes sont présentes dans ce processus :

alignement des bases de connaissances sources : 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 connaissance 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.).

3.1 Génération des correspondances

Nous définissons une base de connaissances source comme un graphe S 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 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. 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 remplit 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 (Si et 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 aligneur permet de mettre 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. En conséquence, dans la suite de cet article, par soucis de simplification, nous ne travaillerons que sur la fusion des sommets des graphes représentant les BCS.

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

3.2 Candidat

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

Un candidat $C = (V_C, E_C, valueE_C)$ 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. Un candidat est un sous-graphe du multigraphe construit à partir des N bases de connaissances sources alignées. Les composants d'un candidat respectent les contraintes suivantes :

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

— E_C : l'ensemble des arêtes d'un candidat est inclus dans l'ensemble des arêtes des T alignements. Les arêtes de C sont des correspondances ;

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

La figure 2 présente deux candidats liant des individus issus de 3 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 Calcul de 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 avons défini dans Amarger *et al.* (2014) une première fonction $trust_{simple}$ qui prend en compte le nombre de sources impliquées dans le candidat. Cette fonction n'intègre pas le degré de fiabilité accordé par l'aligneur pour la correspondance. Nous définissons dans cet article une autre fonction intitulée $trust_{degree}$ décrite ci-dessous.

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.

4.1 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 leur degré de fiabilité. Le calcul de cette fonction est présenté dans l'équation 1. $C = (V_C, E_C, valueE)$

est le candidat étudié. E_C est l'ensemble des arêtes du candidat, N le nombre de sources alignées et $valueE$ l'application qui, à chaque arête de E_C, associe son degré de fiabilité.

$$trust_{degree}(C) = \frac{\sum_{e_i \in E_C} valueE(e_i)}{\frac{N(N-1)}{2}} \quad (1)$$

Cette fonction fait la somme de tous les degrés de fiabilité des correspondances utilisées pour générer le candidat. 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 permet de prendre en compte indirectement le nombre d'éléments présents dans le candidat. Cette fonction est proportionnelle au nombre d'arêtes : plus le graphe du candidat contiendra d'arêtes, plus il contiendra de sommets.

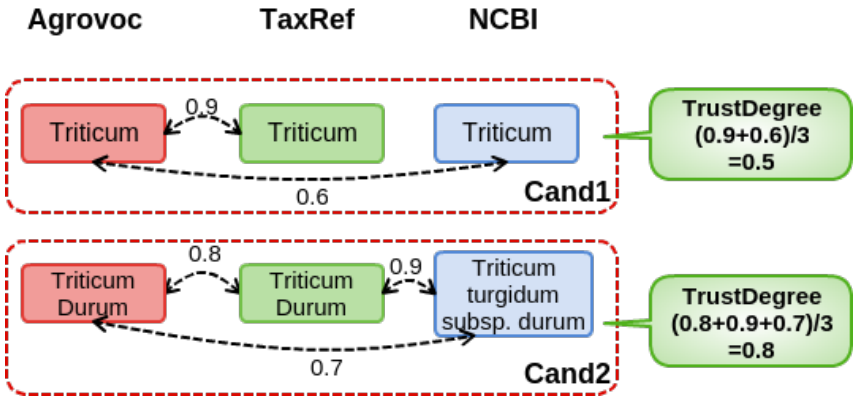

FIGURE 2 – Calcul des scores de confiance avec trustDegree

La figure 2 présente les confiances associées aux deux candidats à l'aide de la fonction $trust_{degree}$. Nous observons que $trust_{degree}(Cand2) > trust_{degree}(Cand1)$. Nous pouvons donc ordonner les candidats par leur confiance, évaluée à l'aide de la fonction $trust_{degree}$ [6].

4.2 Fonction Trust Choquet

L'intégrale de Choquet (eq.2) est utilisée pour la prise de décision sur un ensemble de critères (Grabisch & Roubens (2000)). Elle permet de pondérer l'intérêt de sous-ensembles de critères au lieu de pondérer chaque critère indépendamment des autres, comme le ferait une somme pondérée des critères. Elle permet ainsi de modéliser des conjonctions et des disjonctions sur des sous-ensembles de critères. Pour illustrer l'intérêt de l'intégrale de Choquet, nous allons présenter deux exemples de prises de décision impliquant deux critères x_1 et x_2. La fonction $\mu\{x\}$ représente l'intérêt du critère x dans la prise de décision. Considérons deux cas :

— le cas de la conjonction de critères : le décideur n'est satisfait que si les deux critères x_1 et x_2 sont réalisés simultanément et il n'est pas satisfait si l'un des critères est réalisé sans l'autre. Dans ce cas $\mu(\{x_1\}) = \mu(\{x_2\}) = 0$ mais $\mu(\{x_1, x_2\}) = 1$;

6. Il est à noter que si la confiance était évaluée uniquement par le nombre de sources impliquées, les deux candidats obtiendraient le même score

— le cas de la disjonction : le décideur est satisfait si l'un des deux critères est réalisé sans l'autre. Il n'est pas plus satisfait si les deux critères sont réalisés simultanément. Dans ce cas, $\mu(\{x_1\}) = \mu(\{x_2\}) = 1$ et $\mu(\{x_1, x_2\}) = 1$.

Dans notre cas, les sources S_i impliquées dans le candidat C sont considérées comme les critères de la prise de décision. Une source a un intérêt variable 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 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, la fonction $\mu()$ prendra ses valeurs dans l'intervalle $[0..1]$.

La fonction $f(S)$ retourne une évaluation de la source S.

$$trust_{choquet}(C) = \sum_{i=1}^{n} [f(S_{(i)}) - f(S_{(i-1)})]\mu(A_i) \tag{2}$$

avec $A_i = \{S_{(i)}, ..., S_{(n)}\}$ et $S_{(i)}$ est la permutation des sources S_i tel que $f(S_{(0)}) = 0$ et $0 \leq f(S_{(1)}) \leq f(S_{(2)}) \leq ... \leq f(S_{(n)})$. Cette intégrale calcule le score de confiance du candidat C à partir des sous-ensembles de sources S_i impliquées dans le candidat.

4.2.1 Implication des sources dans un candidat

La fonction f détermine la force de l'implication de la source S dans la construction du candidat C. Cette force est fonction des correspondances associées au sommet de la source S et impliquées dans le candidat C. En effet, nous considérons que plus un sommet de la source S est lié avec des correspondances fortes vers les autres sommets du candidat C, plus la source est impliquée dans la construction de ce candidat.

Prenons l'exemple présenté sur la figure 2 à la page 6. Nous pouvons remarquer pour 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 leur source 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 C en utilisant l'équation 3. Rappelons qu'une source S est un multigraphe orienté étiqueté S ayant comme ensemble de sommet V_S. Un candidat sommet C est un graphe non-orienté $C = (V_C, E_C, valueE)$.

La fonction f utilisée pour évaluer la source S lors du calcul du score de confiance du candidat C avec l'intégrale de Choquet est définie par la fonction $implication$ comme présenté dans l'équation suivante :

$$f(S) = implication(S, C) = \frac{\sum_{\substack{e \in E_C \\ e=(oe_i, oe_j) \text{ avec } oe_i \in V_S}} valueE(e)}{N - 1} \tag{3}$$

Si nous prenons l'exemple du candidat "Cand1" de la figure 2, l'implication de la source "Agrovoc" dans ce candidat peut être définie de la manière suivante :

$$f(Agrovoc) = implication(Agrovoc, Cand1) = \frac{0,9 + 0,6}{3-1} = \frac{1,5}{2} = 0,75 \quad (4)$$

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

$$f(TaxRef) = implication(TaxRef, Cand1) = \frac{0,9}{3-1} = \frac{0,9}{2} = 0,45 \quad (5)$$

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 :

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

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.

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

4.2.2 Intérêt des sources

La deuxième fonction à définir pour utiliser l'intégrale de Choquet est la fonction μ qui représente l'intérêt des sources dans la prise de décision. Cela permet de définir des priorités entre les sources. Nous pouvons par exemple favoriser les candidats impliquant la source "TaxRef" plutôt que ceux impliquant "Agrovoc". Pour ce faire, nous définissons une fonction $Q(S)$ retournant une valeur, comprise entre 0 et 1, représentant la qualité de la source S. L'intérêt d'une source sera fonction de sa qualité.

Dans notre exemple, nous considérons trois sources de qualité différente. Pour évaluer $Q(S)$, 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 $\mu(L_S)$ caractérisant l'intérêt d'un sous-ensemble de sources ($L_S = \{S_j, ..., S_k\}$) en fonction de leur qualité. De cette façon, nous pouvons prendre 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 permet de prendre en compte une évolution variable de l'intérêt des sources. Nous pouvons, par exemple, considérer 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 [7] 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 $\mu(L_S)$ suivante, en considérant L_S comme étant un sous-ensemble des sources considérées :

$$\mu(L_S) = \frac{\lambda(\sum_{i=1}^{|L_S|} Q(S_i)) - \lambda(0)}{\lambda(\sum_{i=1}^{N} Q(S_i)) - \lambda(0)} \tag{7}$$

$$\lambda(x) = arctan(\frac{x - x_0}{\gamma}) \tag{8}$$

La fonction $Q(S_i)$ permet de récupérer la qualité de la source S_i définie précédemment. L'équation $\sum_{i=1}^{N} Q(S_i)$ permet d'obtenir la somme des scores de qualité de toutes les sources considérées dans le processus. Dans l'exemple, nous pouvons définir :

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

De la même façon, nous utilisons l'équation $\sum_{i=1}^{|L_S|} Q(S_i)$ qui permet d'obtenir la somme des scores de qualité des sources présentes dans le sous-ensemble L_S.

Cette fonction $\mu(L_S)$ permet de représenter l'intérêt du sous-ensemble de sources L_S. Nous utilisons la fonction $\lambda(x)$ qui est inspirée de la fonction de répartition de la loi gamma et qui permet d'avoir 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, permet de définir le point d'explosion, point à partir duquel l'intérêt des sources augmente particulièrement. Le deuxième paramètre est la valeur γ qui définit l'indice de linéarité de la courbe. Plus la valeur de γ tend vers 0 et 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 γ s'approche de $\sum_{i=1}^{N} Q(S_i)$ (somme des scores de qualité de toutes les sources considérées) et plus la courbe est linéaire.

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

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

La figure 3 présente la répartition de la fonction $\mu(x)$ en fonction de nos paramètres.

Comme vu précédemment, si nous considérons le candidat "Cand1" de la figure 2, nous avons les implications des sources suivantes :

7. Point auquel la dérivée μ' est à son maximum

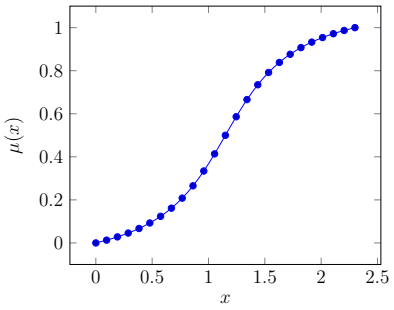

FIGURE 3 – Répartition de $\mu(x)$ avec les paramètres $\sum_{i=1}^{N} Q(S_i) = 2,3$, $x_0 = 1,15$ et $\gamma = 0,46$

— $implication(Agrovoc, Cand1) = 0,75$
— $implication(TaxRef, Cand1) = 0,45$
— $implication(NCBI, Cand1) = 0,3$

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

$$\begin{aligned} trust_{choquet}(Cand1) =& f(NCBI) * \mu(Agrovoc, TaxRef, NCBI) \\ &+ [f(TaxRef) - f(NCBI)] * \mu(Agrovoc, TaxRef) \\ &+ [f(Agrovoc) - f(TaxRef)] * \mu(Agrovoc) \\ =& 0,3 * \mu(Agrovoc, TaxRef, NCBI) \\ &+ (0,45 - 0,3) * \mu(Agrovoc, TaxRef) \\ &+ (0,75 - 0,45) * \mu(Agrovoc) \\ =& 0,3 * 1 + 0,15 * 0,77 + 0,3 * 0,14 = 0,46 \end{aligned} \tag{10}$$

5 Évaluation

Nous avons construit semi-automatiquement trois bases de connaissances en transformant les 3 sources suivantes : le thésaurus Agrovoc de la FAO, la taxonomie des organismes vivants du NCBI et la taxonomie française de référence TaxRef des organismes vivants du Muséum d'Histoire Naturelle. Le processus de transformation est présenté dans Amarger *et al.* (2014). Nous avons aussi construit une référence sur la taxonomie des blés avec l'aide de trois experts Amarger (2015). Cette référence contient l'union de tous les éléments des 3 BCS précédentes validés manuellement par nos experts.

La précision, le rappel et la $F - mesure$ sont calculés afin d'évaluer la qualité des candidats générés par notre processus de fusion. Nous considérons qu'un candidat généré est valide si la totalité de ses sommets appartiennent à la référence. Nous avons adapté ces mesures de la manière suivante :

— la précision est le rapport entre le nombre de sommets des candidats valides et le nombre total de sommets des candidats générés par la fusion ;
— le rappel est le rapport entre le nombre de sommets des candidats valides et le nombre total de sommets de la référence ;
— la $F - mesure$ est une combinaison des mesures de précision et rappel.

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

TABLE 2 – Expérimentations $trust_{degree}$ vs $trust_{choquet}$

Nous avons analysé l'impact de chacune des fonctions de confiance en sélectionnant les candidats en fonction de leur score de confiance. Un candidat est sélectionné si son score est supérieur à un seuil. Nous sélectionnons les candidats par pas de 0,1. Nous calculons les 3 mesures sur l'ensemble de candidats générés pour chacun des pas. La fonction de confiance représentant au mieux le consensus aura une mesure de précision d'autant plus élevée que le filtre sera élevé lui aussi. Nous ne présentons ici que les résultats concernant les candidats ne contenant que des sommets de nature individus.

En observant les résultats de la fonction $trust_{degree}$, présentés dans le tableau 2, nous observons plusieurs phénomènes. Tout d'abord, la précision est importante dès le seuil 0,1. Le rappel est moins satisfaisant. Ceci s'explique par le fait que la fonction de confiance $trust_{degree}$ discrimine l'ensemble des candidats en fonction des correspondances. Les candidats ayant un haut score de confiance sont ceux qui impliquent beaucoup de correspondances. Les candidats impliquant moins de 3 sources sont ici rejetés dès les seuils assez bas, bien que certains soient valides. Ceci explique la diminution rapide du rappel. Les candidats impliquant trois sources n'utilisent pas forcément beaucoup de correspondances ou des correspondances avec de faibles degrés de fiabilité. Les sommets des candidats ne sont alors pas fortement connectés, ce qui explique les faibles valeurs du rappel pour des seuils hauts.

Nous pouvons en déduire que l'utilisation des correspondances dans le calcul de la confiance discrimine rapidement les candidats. Les candidats n'impliquant pas beaucoup de sources auront un score de confiance assez bas. Néanmoins, nous vérifions la validité de notre hypothèse initiale. 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 dès le seuil 0,3.

Pour l'évaluation de la fonction de confiance $trust_{choquet}$, nous réutilisons les scores de qualité des sources utilisées. Les résultats de la fonction $trust_{choquet}$ sont meilleurs sur les seuils bas que ceux de la fonction $trust_{degree}$. On note un très fort rappel et $F - mesure$ pour le seuil 0,1 pour la fonction $trust_{choquet}$. En effet, l'implication des sources pour le candidat permet de contrebalancer l'aspect discriminant des correspondances. En revanche, les deux fonctions ont des résultats identiques sur les seuils hauts. Il est à noter que nos expérimentations n'ont porté que sur la fusion de 3 sources. Des expérimentations impliquant plusieurs sources de qualité variée pourront montrer tout l'impact de la fonction $trust_{choquet}$.

6 Conclusion et Perspectives

Dans cet article, nous avons présenté notre méthode de fusion de plusieurs bases de connaissances. Notre méthode est la première qui travaille avec plus de deux bases. En effet, nous souhaitons extraire de plusieurs sources les éléments consensuels, c'est-à-dire ceux qui sont communs à plusieurs sources. Notre proposition évalue la confiance dans les éléments extraits des sources. Nous avons présenté plusieurs fonctions de confiance. Une évaluation a montré l'intérêt de la fonction $trust_{choquet}$, capable de tenir compte de l'implication locale d'une source dans un élément et de la qualité de cette source. Notre méthode de fusion s'est focalisée uniquement sur la fusion des classes et des individus des BCS. Bien que les aligneurs ne soient pas encore capables de générer des correspondances entre propriétés, nous devons étendre notre méthode de fusion aux candidats représentant les liens entre les classes et les individus.

Références

AMARGER F. (2015). *Vers un système intelligent de capitalisation de connaissances pour l'agriculture durable : construction d'ontologies agricoles par transformation de sources existantes*. PhD thesis, 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 *8th Research Conference Metadata and Semantics Research MTSR Karlsruhe, Germany, November, 2014*, p. 314–328.

AMARGER F., CHANET J., HAEMMERLÉ O., HERNANDEZ N. & ROUSSEY C. (2015). Incompatiblity treatment of candidates from several knowledge bases alignments. In *26es Journées Francophones d'Ingénierie des Connaissances, Rennes, juin, 2015*, p. 203–208.

CURÉ O. (2009). Merging expressive ontologies using formal concept analysis. In *On the Move to Meaningful Internet Systems : OTM 2009 Workshops, Vilamoura, Portugal, November, 2009*, volume 5872, p. 49–58.

DRAGISIC Z., ECKERT K., EUZENAT J., FARIA D., FERRARA A., GRANADA R., IVANOVA V., JIMÉNEZ-RUIZ E., KEMPF A. O. & LAMBRIX P. E. A. (2014). Results of the Ontology Alignment Evaluation Initiative 2014. In *9th ISWC workshop on ontology matching (OM), Riva del Garda, Italy, October, 2014*, p. 61–104.

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

GUZMÁN-ARENAS A. & CUEVAS A.-D. (2010). Knowledge accumulation through automatic merging of ontologies. *Expert Systems with Applications*, **37**(3), 1991–2005.

JIMÉNEZ-RUIZ E. & GRAU B. C. (2011). Logmap : Logic-based and scalable ontology matching. In *10th International Semantic Web Conference, Bonn, Germany, October, 2011, Proceedings, Part I*, p. 273–288.

NOY N. F. & MUSEN M. A. (2003). The PROMPT suite : interactive tools for ontology merging and mapping. *International Journal of Human-Computer Studies*, **59**(6), 983–1024.

POTTINGER R. A. & BERNSTEIN P. A. (2003). Merging models based on given correspondences. In *29th International Conference on Very large data bases, Berlin, Germany, September, 2003*, p. 862–873.

RAUNICH S. & RAHM E. (2014). Target-driven merging of taxonomies with Atom. *Information Systems*, **42**, 1–14.

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*, **4**(4).