Identification des unités de mesure dans les textes scientifiques

Soumia Lilia Berrahou[1, 2] Patrice Buche[1, 2] Juliette Dibie[3] Mathieu Roche[1, 4]
(1) LIRMM, 161 rue Ada, Montpellier, France
(2) IATE, 2 place Viala, Montpellier, France
(3) MIA, 16 rue Claude Bernard, Paris, France
(4) TETIS, 500 rue Jean-François Breton, Montpellier, France
berrahou@lirmm.fr, Patrice.Buche@supagro.inra.fr
Juliette.Dibie@agroparistech.fr, mathieu.roche@cirad.fr

Résumé. Le travail présenté dans cet article se situe dans le cadre de l'identification de termes spécialisés (unités de mesure) à partir de données textuelles pour enrichir une Ressource Termino-Ontologique (RTO). La première étape de notre méthode consiste à prédire la localisation des variants d'unités de mesure dans les documents. Nous avons utilisé une méthode reposant sur l'apprentissage supervisé. Cette méthode permet de réduire sensiblement l'espace de recherche des variants tout en restant dans un contexte optimal de recherche (réduction de 86% de l'espace de recherché sur le corpus étudié). La deuxième étape du processus, une fois l'espace de recherche réduit aux variants d'unités, utilise une nouvelle mesure de similarité permettant d'identifier automatiquement les variants découverts par rapport à un terme d'unité déjà référencé dans la RTO avec un taux de précision de 82% pour un seuil au dessus de 0.6 sur le corpus étudié.

Abstract.
Identification of units of measures in scientific texts.
The work presented in this paper consists in identifying specialized terms (units of measures) in textual documents in order to enrich a onto-terminological resource (OTR). The first step permits to predict the localization of unit of measure variants in the documents. We have used a method based on supervised learning. This method permits to reduce significantly the variant search space staying in an optimal search context (reduction of 86% of the search space on the studied set of documents). The second step uses a new similarity measure identifying automatically variants associated with term denoting a unit of measure already present in the OTR with a precision rate of 82% for a threshold above 0.6 on the studied corpus .

Mots-clés : ressource termino-ontologique, apprentissage, similarité.

Keywords: onto-terminological resource, learning, similarity.

1 Introduction

Le travail présenté dans cet article se situe dans le cadre de l'identification de termes spécialisés à partir de données textuelles pour enrichir une Ressource Termino-Ontologique (RTO). Les travaux de (McCrae *et al.*, 2011; Cimiano *et al.*, 2011) proposent d'associer une partie terminologique et/ou linguistique aux ontologies afin d'établir une distinction claire entre la manifestation linguistique (le terme) et la notion qu'elle dénote (le concept). Nous nous intéressons à l'enrichissement d'une RTO permettant de modéliser des relations n-aires entre des données quantitatives expérimentales (Touhami *et al.*, 2011), où les arguments peuvent être des concepts symboliques ou des quantités caractérisées par des unités de mesure. En effet, l'extraction des données quantitatives est un enjeu majeur pour de nombreux domaines scientifiques dont l'objectif concerne la capitalisation et la pérennisation des connaissances du domaine. Cependant, la forte variation d'écriture des unités de mesure dans les documents engendre des problèmes d'identification des instances numériques dans les textes.

Dans une démarche consensuelle, le Systeme International (SI) (Thompson & Taylor, 2008) organise, en posant plusieurs définitions formelles, le système des quantités et des unités de mesure. Il définit ainsi des unités de base, i.e. unités simples comme *kilogram*, et des unités dérivées, i.e. unités plus complexes comme $kg.m^{-1}$. Ce standard pose les règles d'écriture de l'ensemble des unités de mesure mais n'intègre pas la notion de variants d'unités. Ces principes sont repris dans des

travaux récents (Rijgersberg *et al.*, 2013) afin de modéliser formellement cette connaissance dans une ontologie dédiée à la représentation des données quantitatives et des unités de mesure. Les auteurs ont ainsi modélisé *OM* (Ontology of Units of Measure and Related Concepts). Les travaux de (Van Assem *et al.*, 2010) posent la problématique d'identification des données quantitatives présentes dans les cellules des tableaux représentés dans les documents. La localisation des variants d'unités n'est pas problématique dans ces travaux car la méthode repose sur le format structuré des tableaux. Les travaux de (Grau *et al.*, 2009) proposent des méthodes d'extraction des données expérimentales dans le domaine biomédical. L'identification des unités de mesure repose sur les unités référencées dans le Systeme International (Thompson & Taylor, 2008), la problématique de l'identification des variants d'unité référencée n'y est pas abordée.

Ainsi, à notre connaissance, les méthodes de l'état de l'art partageant l'objectif d'extraction de données quantitatives, ne permettent pas de résoudre la problématique d'extraction et d'identification des variants d'unités de mesure dispersés dans les documents scientifiques au format textuel non structuré. Dans cet article, nous présentons notre proposition qui tente de répondre à deux questions concernant l'identification des variants d'unités de mesure dans les documents textuels non structurés :

– La question concernant la localisation des variants dans le document. Sachant que nous travaillons sur l'intégralité des documents, nous préférons l'apprentissage afin de prédire la localisation des variants sans poser d'hypothèses préalables.

– La question de l'identification du variant une fois qu'il est localisé. A quel autre terme d'unité de mesure référencée dans la RTO peut-on le rapprocher, en sachant que les termes d'unités répondent à leurs propres règles syntaxiques ? Les méthodes existantes doivent être adaptées à ces nouvelles règles.

Les deux questions de recherche précédentes sont respectivement traitées en sections 2 et 3. Ces propositions sont alors expérimentées en section 4, avant la conclusion et les perspectives décrites en section 5.

2 Localisation d'unité de mesure dans les textes

La première étape de notre méthode illustrée dans la figure 1 consiste à prédire la localisation des variants d'unités de mesure dans les documents. Nous avons utilisé une méthode reposant sur l'apprentissage supervisé. Les pré-traitements choisis pour préparer nos documents reposent sur les étapes ci-contre : segmentation des documents en phrases, tokenisation des phrases, suppression des mots vides de la phrase, annotation automatique des phrases. Cette dernière étape, illustrée sur les figures 2 et 3, consiste à identifier les phrases contenant des concepts d'unités de mesure présents dans une RTO pour constituer un corpus d'apprentissage avec un ensemble d'exemples positifs (cf. phase 1 de la figure 1). Les figures montrent deux cas d'annotation automatique selon que l'unité est considérée comme un ou plusieurs tokens. Pour isoler la fraction du variant Ã identifier dans la deuxiéme étape, décrite dans la section 3, nous balayons la phrase de part et d'autre de l'unité identifiée jusqu'à trouver un terme du dictionnaire. Dans nos travaux, nous définissons une nouvelle représentation des documents sous forme de fen êtres textuelles, représentant un contexte phrastique précis (en considérant une à deux phrases autour de chaque phrase courante – cf. phase 2 de la figure 1). Dans notre contexte d'étude, pour chacune des fen êtres, nous avons sélectionné uniquement les termes apparaissant plus d'une fois dans le corpus dédié à l'apprentissage pour constituer un *sac de mots* (descripteurs). Cela réduit sensiblement l'espace de représentation des textes sans avoir pour autant un impact sur les résultats de la classification. Les termes ainsi pré-sélectionnés sont projetés sur chaque document. Hormis la représentation booléenne des descripteurs (présence/absence), les descripteurs propres à la représentation vectorielle peuvent être pondérés selon différentes approches statistiques comme TF, TF.IDF et OKAPI qui sont expérimentés en section 4.1 (cf. phase 3 de la figure 1). L'ensemble des vecteurs constitue la nouvelle représentation des documents pour les algorithmes d'apprentissage supervisé. Le but des modèles appris étant de prédire si une phrase d'un ensemble de test est susceptible de contenir une unité de mesure.

La section suivante décrit la deuxième étape du processus : une fois l'espace de recherche réduit aux variants d'unités, elle propose une nouvelle mesure de similarité permettant d'identifier automatiquement les variants découverts par rapport à un terme d'unité déjà référencé dans la RTO.

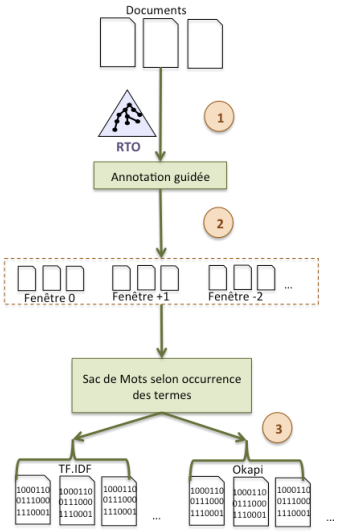


FIGURE 1: Représentation textuelle adaptée au contexte

gluten film ranged 88 amol/m.s.Pa 5 0 <um>%</um> relative humidity RH

(a) Unité de mesure référencée dans la RTO

gluten film ranged 88 amol/m.s.Pa 5 0 <um>%</um> relative humidity RH

(b) Un premier terme reconnu du dictionnaire

gluten film ranged 88 amol/m.s.Pa 5 0 <um>%</um> relative humidity RH

(c) Un second terme reconnu du dictionnaire

Unit fragment: 88 amol/m.s.Pa 5 0 %

(d) Le variant à identifier

FIGURE 2: Isoler le variant considéré comme un token

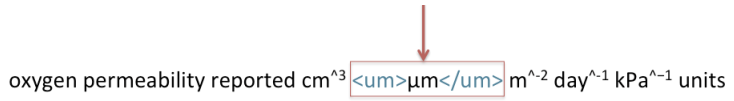
3 Identification de nouvelles unités de mesures

Les unités de mesures subissent de fortes variations terminologiques. Un m ême concept d'unités de l'ontologie peut donc être représenté par des termes d'unités très différents dans les documents, nous les nommons les variants d'unités. Contrairement aux variations terminologiques considérées pour évaluer la similarité entre deux cha înes de caractères, les unités de mesure possèdent leurs propres règles d'écriture établies librement par l'auteur du document.

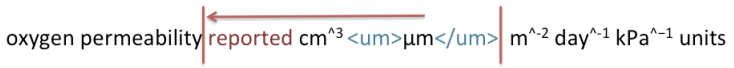
Par exemple, l'unité de mesure *amol/(m.s.Pa)* définie dans la RTO peut être écrite à l'aide de différents variants dans les documents scientifiques selon les cas :

- d'insertions de caractères comme dans *amol/m.sec.Pa* ou *amol.m-1.s-1.Pa-1*
- de suppressions de caractères comme dans *mol/m.s.Pa*
- d'inversions de certains blocs dans l'unité comme dans *amol.s-1.m-1.Pa-1*
- d'écriture non plus ponctuée mais comme un ensemble composé de blocs indépendants comme dans *amol m-1 s-1 Pa-1*

Nos travaux proposent d'extraire de telles variations terminologiques dans les documents afin d'enrichir la RTO de ces

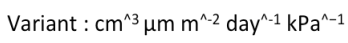


(a) Unité de mesure référencée dans la RTO



(b) Un premier terme reconnu du dictionnaire

(c) Un second terme reconnu du dictionnaire



(d) Fraction contenant le variant à identifier

FIGURE 3: Isoler le variant considéré comme plusieurs tokens

variants d'unités de mesure. Une telle extraction ne peut pas reposer sur des méthodes utilisant des expressions régulières car il n'existe aucune règle précise pour les établir du fait de la grande variation typographique des unités de mesure. Nous proposons d'utiliser et adapter les mesures de proximité aux spécificités des unités de mesure afin de répondre à la problématique d'identification des variants d'unités.

Dans notre approche, il est fondamental de pouvoir prendre en considération les particularités d'écriture des unités de mesure, en notant que chaque bloc est indépendant dans l'écriture de l'unité. De ce fait, l'ordre des blocs n'est pas important à prendre en compte, en revanche, la comparaison des blocs entre eux nous semble plus pertinente et plus adaptée dans le calcul de la similarité. Il est alors intéressant de proposer une mesure qui calcule la similarité en deux temps, qui s'appuie à la fois sur les unités déjà référencées dans la RTO et sur les caractères spécifiques (/, (,), ., ×, ^...) utilisés comme séparateurs de blocs.

Dans un premier temps, **les candidats sont présectionnés** selon la mesure de Jaccard. Le principe ci-dessous est alors mis en œuvre :

– Soit un couple composé du variant candidat u_i et d'une unité référence dans la RTO u_j. $J(u_i, u_j)$ (cf. formule (1)) calcule dans un premier temps le score de similarité entre l'ensemble u_i et l'ensemble u_j par rapport aux blocs communs sans tenir compte de leur ordre.

$$J(u_i, u_j) = \frac{|u_i \cap u_j|}{|u_i \cup u_j|} \tag{1}$$

– On sélectionne le couple (u_i, u_j) comme étant pertinent à être comparé si $J(u_i, u_j) > K'$, K' étant le seuil minimal défini préalablement par l'utilisateur.

Prenons l'exemple du couple composé d'un variant candidat localisé et extrait à partir d'un document $kg\ \ Pa\hat{}^{-1}s\hat{}^{-1}m\hat{}^{-2}$ et son référent dans la RTO $lb.m.m^{-2}.s^{-1}.Pa^{-1}$. Dans ce contexte, le calcul de la mesure de Jacccard donne le résultat suivant :

$$J(kg\ m\ Pa\hat{}^{-1}s\hat{}^{-1}m\hat{}^{-2}, lb.m.m^{-2}.s^{-1}.Pa^{-1}) = \frac{4}{6} = 0.7$$

Dans un deuxième temps, après cette phase de présélection, **les candidats sont sélectionnés** selon une mesure étendue de Damereau-Levenshtein.

La mesure de similarité de Levenshtein (Levenshtein, 1966) calcule le co ût minimal pour transformer une première cha îne de caractères en une deuxième cha îne de caractères en considérant les opérations de remplacement de caractères entre les deux cha înes, d'ajout d'un caractère ou et de suppression d'un caractère. Le co ût est ensuite normalisé pour obtenir une valeur de la distance entre les deux cha înes entre 0 et 1. Cette mesure est étendue par Damerau (Damerau, 1964) qui

inclut dans celle de Levenshtein la notion de transposition de caractères d'une cha îne à une autre, i.e. dans *litre* et *liter*, il y a transposition entre les caractères "e" et "r".

La mesure adaptée à notre contexte (cf. formule (2)) ne considère plus la comparaison des caractères mais des blocs de caractères, correspondant à des unités simples. Le variant candidat et l'unité de référence, composant le couple présélectionné lors de la première phase, sont, dans cette seconde phase, comparés bloc à bloc pour déterminer leur similarité finale.

$$SM_{D_b}(u_i, u_j) = max[0; \frac{\min(|u_i|, |u_j|) - D_b(u_i, u_j)}{min(|u_i|, |u_j|)}]; SM_{D_b}(u_i, u_j) \in [0; 1] \quad (2)$$

- (u_i, u_j) représente le couple sélectionné à partir de la mesure de Jaccard ;
- Chaque bloc de u_i est comparé aux blocs de u_j pour calculer la nouvelle distance D_b ;
- u_i est validée comme un variant de l'unité u_j si $SM_{D_b} > K$, avec K un seuil de similarité défini préalablement.

En posant $K = 0.5$, le couple $kg\ m\ Pa^{\hat{}-1}s^{\hat{}-1}m^{\hat{}-2}, lb.m.m^{-2}.s^{-1}.Pa^{-1}$, SM_{D_b} calcule la similarité du couple en comparant chaque bloc dans les unités :

$$SM_{D_b}(kg\ m\ Pa^{\hat{}-1}s^{\hat{}-1}m^{\hat{}-2}, lb.m.m^{-2}.s^{-1}.Pa^{-1}) = max[0; \frac{5-1}{5}] = 0.8$$

Un tel processus fondé sur ces deux phases consécutives de présélection et de sélection finale permet la découverte de nouveaux variants à intégrer comme détaillé dans la section suivante.

4 Expérimentations

Les expérimentations ont été menées à partir d'un corpus de 115 articles scientifiques en anglais issus du domaine des emballages alimentaires. Elles s'appuient également sur une liste de 211 termes dénotant les différents concepts d'unités de mesure pour le domaine des emballages alimentaires. Ces différentes fen êtres correspondent à des sous-ensembles du corpus représentant 5000 phrases (e.g. f_0) à 15000 phrases (e.g. f_{+2}). Le corpus complet comportant plus de 35000 phrases. Le sac de mots représente un ensemble de 3000 à 4800 descripteurs selon les différentes représentations.

4.1 Évaluation de la méthode de localisation des unités de mesure

Notre objectif, au cours de cette première étape (cf. section 2), est de produire un modèle d'apprentissage appris à partir des données représentées sous forme de fen êtres textuelles, qui permette de réduire l'espace de recherche des variants d'unités. Nous avons testé plusieurs fen êtres textuelles. Nous restituons en résultats des expérimentations uniquement ceux révélant les fen êtres d'étude les plus pertinentes dans le tableau 1. Par souci de lisibilité, les fen êtres textuelles sont exprimées de la manière suivante :

- f_0 : représente la fen être comportant la phrase où au moins un terme d'unité dénotant un concept de la RTO est identifié.
- f_{+2} : représente la fen être comportant la phrase où au moins un terme d'unité dénotant un concept de la RTO est identifié ainsi que les deux phrases suivantes.
- f_{-2} : représente la fen être comportant la phrase où au moins un terme d'unité dénotant un concept de la RTO est identifié ainsi que les deux phrases précédentes.

Les tableaux 1 et 2 restituent les résultats obtenus sur le corpus des emballages réalisé avec quatre algorithmes d'apprentissage (Naives Bayes, C4.5, DMNB (Discriminative Multinominal Naive Bayes), SMO (Sequential minimal optimization qui est une variante de SVM)) et une 10-validation croisée. Le tableaun1 restitue les résultats, toutes mesures confondues. L'analyse des résultats montrent que Naives Bayes produit une F-mesure allant de 0.85 à 0.88, l'arbre de décision établi sur l'algorithme C4.5 (J48) produit de meilleurs résultats autour de 0.93 à 0.96. DMNB et SMO produisent les meilleurs

résultats, conformément à ce qui est souligné dans la littérature du domaine (0.95 à 0.99). Outre ces résultats analytiques, nous remarquons qu'un plus large contexte, à partir des fen êtres f_{+2} et f_{-2}, n'améliorent pas les résultats d'apprentissage. Nous pouvons donc en déduire que la plus petite fen être textuelle, c'est-à-dire celle où au moins un terme d'unité référencé dans la RTO apparait, est le contexte le plus favorable à la découverte de variants d'unités. Cette conclusion permet de réduire sensiblement l'espace de recherche des variants, i.e. 5000 phrases à considérer plut ôt que 35000 initialement dénombrées, tout en restant dans un contexte optimal de recherche (réduction de 86% de l'espace de recherche).

Le tableau 2 synthétise les résultats selon les différentes mesures de pondération et la matrice booléenne pour la fen être optimale f_0. Notre objectif étant d'évaluer quel algorithme produit le modèle restituant des valeurs de F-mesure stables sur les différentes mesures de pondération et la matrice booléenne. La F-mesure, ainsi que les valeurs de précision et de rappel restent stables et élevées avec le modèle DMNB, en restituant une valeur constante autour de 0.95.

	Dec. Tree J48			**Naive Bayes**			**DMNB**			**SMO**		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
f_0	0.99	0.87	0.93	0.83	0.93	0.88	0.95	0.96	**0.95**	0.99	0.99	**0.99**
f_{+2}	0.99	0.92	0.96	0.95	0.77	0.85	0.93	0.96	**0.95**	0.99	0.97	**0.99**
f_{-2}	0.99	0.92	0.95	0.77	0.98	0.86	0.94	0.96	**0.95**	0.99	0.97	**0.98**

TABLE 1: Résultats des instances de la classe "Unit" : Précision (P), Rappel (R), F-mesure (F) restitués pour chaque fen être textuelle.

	Dec. Tree J48			**Naive Bayes**			**DMNB**			**SMO**		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
Boolean	0.99	0.87	0.93	0.83	0.93	0.88	**0.95**	**0.96**	**0.95**	0.99	0.99	0.99
TF	0.99	0.86	0.92	0.69	0.85	0.76	**0.95**	**0.96**	**0.95**	0.84	0.90	0.87
TF.IDF	0.99	0.86	0.92	0.69	0.85	0.76	**0.95**	**0.96**	**0.95**	0.84	0.90	0.87
Okapi	0.99	0.86	0.92	0.69	0.86	0.76	**0.95**	**0.96**	**0.95**	0.77	0.88	0.82

TABLE 2: Résultats des instances de la classe "Unit" sur f_0 : Précision (P), Rappel (R), F-mesure (F) restitués pour chaque mesure de pondération et le modèle booléen.

4.2 Évaluation de la méthode d'identification des unités de mesure

Dans la deuxième étape de notre processus (cf. section 3), nous nous appuyons sur les résultats obtenus précédemment afin d'identifier les variants d'unités. Notons que dans notre contexte, nos mesures doivent sélectionner les variants les plus pertinents à présenter aux experts, ce qui revient à minimiser le bruit. Ainsi, nous privilégions la mesure de précision pour évaluer nos propositions.

Le tableau 3 restitue les résultats obtenus avec la nouvelle mesure et permet de les comparer par seuil de similarité :

1. La précision est globalement plus élevée avec le processus complet comparativement à l'application de la seule mesure de présélection (Jaccard), ceci confirme donc l'intér êt d'utiliser notre mesure SMD_b.
2. Les seuils les plus intéressants à exploiter sont au-dessus de 0.6 avec un taux de précision supérieur à 82% après application des deux étapes successives ; l'essentiel des variants sont identifiés.
3. En dessous de 0.5, les résultats se tassent largement. En choisissant de ne considérer que les seuils au-delà de 0.5, nous créons forcément du silence mais un silence "contr ôlé". En effet, le processus d'extraction et d'identification des variants étant un processus itératif, les nouvelles unités intégrées dans la RTO favorisent la découverte d'autres variants qui s'expriment dans cette plage de silence.

Le tableau 3 montre la validation des couples variants et unités de référence pertinents (avec K et K' ayant les valeurs 0.5). Un m ême variant peut former un couple avec plusieurs unités de référence dans la RTO. En effet, prenons l'exemple du variant *amol / m s Pa*, sa comparaison avec les unités de référence *amol/m/s/Pa, amol/(m.s.Pa), amol/(m s Pa)...* est considérée comme pertinente. Pour tous les couples pertinents validés, nous n'intégrons qu'une seule fois le variant *amol / m s Pa*. Considérant cette remarque, sur les 267 couples cumulés dans le tableau 3, validés par SMD_b (260 couples sélectionnés), nous obtenons 121 variants d'unités uniques à intégrer dans notre RTO.

Seuil de similarité	**Présélection par Jaccard** (étape 1)	**Précision** (étape 1)	**Sélection par** SMD_b (étape 2)	**Précision** (étape 2)
[0.9-1]	64	**0.87**	54	0.84
[0.8-1]	121	0.79	102	**0.84**
[0.7-1]	238	0.73	209	**0.88**
[0.6-1]	266	0.75	220	**0.82**
[0.5-1]	317	0.77	249	**0.78**
[0.4-1]	479	0.50	267	**0.56**

TABLE 3: Résultats obtenus avec la nouvelle mesure combinée

5 Conclusion

La méthode proposée, guidée par la connaissance de la RTO, permet à partir d'un processus complet de localisation et d'identification des variants d'unités de mesure, d'enrichir la RTO de nouveaux termes d'unités. Cet enrichissement est une étape fondamentale car les problématiques d'identification des unités de mesure complexes sont une des causes des problématiques d'identification et d'extraction des instances d'arguments quantitatifs.

Nous avons montré que la première étape de notre méthode, s'appuyant sur l'apprentissage supervisé, permet de localiser automatiquement les variants d'unité dans un contexte phrastique optimal de recherche. La méthode repose sur une nouvelle représentation des données, sous forme de fen êtres textuelles d'étude, que nous obtenons en étant guidé par la RTO. Par la suite, nous avons proposé une nouvelle mesure de similarité adaptée aux spécificités des unités de mesure. Le choix de la mesure de Damereau-Levenshtein est appropriée à notre contexte car elle prend en charge toutes les variations constatées pour les unités de mesure. De plus, associée à l'indice de Jaccard, la nouvelle mesure permet de rapprocher les couples variant-unité référencée de manière plus pertinente en octroyant un premier score global de similarité qui ne tient pas compte de l'ordre des blocs dans la construction de l'unité complexe. Dans un second temps, la nouvelle mesure SMD_b affine ce rapprochement en comparant chaque bloc du variant et de l'unité référencée sélectionnée. Le processus d'enrichissement de la RTO étant un processus itératif, une nouvelle phase d'extraction et d'identification permettrait alors de comparer d'autres variants avec ces nouvelles unités intégrées dans la RTO, qui deviennent des référents.

Références

CIMIANO P., BUITELAAR P., MCCRAE J. & SINTEK M. (2011). LexInfo : A declarative model for the lexicon-ontology interface. *J. Web Sem.*, **9**(1), 29–51.

DAMERAU F. (1964). A technique for computer detection and correction of spelling errors. *Commun. ACM*, **7**(3), 171–176.

GRAU B., LIGOZAT A.-L. & MINARD A.-L. (2009). Corpus study of kidney-related experimental data in scientific papers. In *Proceedings of the Workshop on Biomedical Information Extraction*, p. 21–26.

LEVENSHTEIN V. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, **10**, 707.

MCCRAE J., SPOHR D. & CIMIANO P. (2011). Linking lexical resources and ontologies on the semantic web with lemon. In *Proceedings of the 8th extended semantic web conference on The semantic web : research and applications - Volume Part I*, ESWC'11, p. 245–259, Berlin, Heidelberg : Springer-Verlag.

RIJGERSBERG H., VAN ASSEM M. & TOP J. (2013). Ontology of units of measure and related concepts. *Semantic Web*.

THOMPSON A. & TAYLOR B. N. (2008). Guide for the use of the international system of units (SI).

TOUHAMI R., BUCHE P., DIBIE-BARTHÉLEMY J. & IBANESCU L. (2011). An Ontological and Terminological Resource for n-ary Relation Annotation in Web Data Tables. In *International Conference ODBASE, OTM Workshops 2011*, volume 7045, p. 662–679 : Lecture Notes in Computer Science series.

VAN ASSEM M., RIJGERSBERG H., WIGHAM M. & TOP J. (2010). Converting and annotating quantitative data tables. *The Semantic Web–ISWC 2010*, p. 16–31.