

# Pharmacovigilance du Web Social par une approche fondée sur les bases de connaissances du Web Sémantique

Damien Leprovost, Marie-Christine Jaulent

Laboratoire Limics, INSERM U-1142, Univ. UPMC Paris 6, Sorbonne Universités  
Université Paris 13, Sorbonne Paris Cité, 75006 Paris, France – {prenom.nom}@inserm.fr

## 1 Introduction

La pharmacovigilance s'intéresse à la surveillance des effets indésirables des médicaments autorisés sur le marché. On estime que moins de 5% de ces événements sont effectivement déclarés, ce qui entraîne un coût économique et humain très important (Hazell et Shakir, 2006).

Nous présentons ici nos premiers travaux visant à améliorer tant la détection de ces effets que celle des signaux. Nous utilisons la structure des données du Web Sémantique pour réaliser une pré-évaluation automatique des données du Web Social, afin de fournir un ensemble limité de signaux potentiels. La figure 1 présente l'architecture du framework LINPHAS développé.

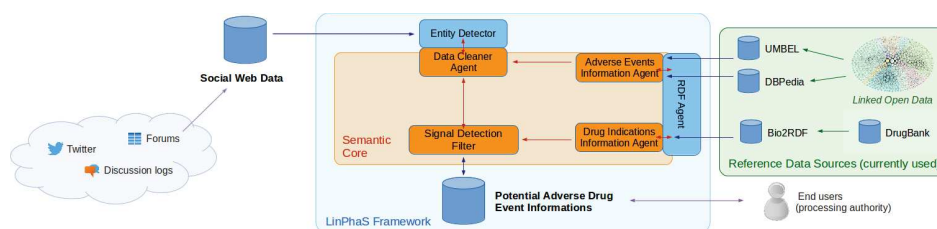


FIG. 1 – Architecture fonctionnelle du framework LINPHAS

## 2 Données utilisées et processus

Les données sociales peuvent être n'importe quel contenu Web généré par des utilisateurs. Dans notre étude, 82 924 tweets sont extraits entre septembre 2014 et mars 2015 à partir de 34 médicaments ciblés<sup>1</sup>. Les données ouvertes et liées utilisées proviennent de trois jeux de données : *UMBEL*, pour extraire et naviguer entre les concepts ; *DBpedia*, pour déterminer la terminologie et *DurgBank via Bio2RDF* pour accéder aux informations des médicaments.

Après nettoyage et normalisation, les messages sont mis en correspondance avec les données extraites du Web Sémantique (états de santé et médicaments). Ne sont alors conservées

1. Le détail des termes retenus est disponible à <https://goo.gl/pKXUkX>.

que les contributions avec un potentiel lien causal. L'étape de filtrage marque ensuite les relations préalablement connues par référence d'indication ou d'effet. Un message conservé non marqué est alors enregistré comme « signal potentiel », nécessitant une attention particulière.

Dans cette étude, LINPHAS met en évidence 31 661 des 60 224 tweets en anglais comme contenant une relation médicament–effet (53 %), dont 5 463 (17 %) comme signaux potentiels. La table 1 montre les états pour lesquels une relation non connue est supposée. LINPHAS retrouvant les relations préalablement connues, la table 2 en liste les principaux signaux.

état de santé	médicament	qté	suggestion
weight loss	orlistat	1558	obesity
bleeding	rivaroxaban	1260	anticoagulant
noma	ipilimumab	1041	melanoma
alcoholism	baclofen	560	<i>no match</i>
noma	vemurafenib	436	melanoma
internal bleeding	rivaroxaban	377	anticoagulant

TAB. 1 – Principaux éléments identifiés

état de santé	médicament	qté
kyste	ivacaftor	1477
fibrose	ivacaftor	1371
obésité	orlistat	1092
mélanome	ipilimumab	1032
psoriasis	ustekinumab	1002
saignement	dabigatran	965
cancer	eribulin	818

TAB. 2 – Indications détectés

### 3 Travaux futurs

La continuité de nos travaux portera sur l'amélioration de la prise en compte de la disparité entre terminologie médicale et langage courant, et sur le raisonnement sur les indications et les effets pour l'évaluation de la mesure dans laquelle les relations trouvées sont nouvelles.

### 4 Travaux en relation

De nombreux outils statistiques ont été proposés, mais la diversité des sources et le manque de références standards de taille suffisante pour évaluer et comparer les méthodes amplifient le problème de détection du signal (Harpaz et al., 2012). Les données sur le Web Social ont également été considérées (Lamb et al., 2013). Mais la pauvreté sémantique relative d'un tweet est souvent pointée comme une limite à l'exploitation de ce type de données (Saif et al., 2012).

## Références

- Harpaz, DuMouchel, Shah, Madigan, Ryan, et Friedman (2012). Novel data-mining methodologies for adverse drug event discovery and analysis. *Clin. Pharma. Ther.* 91(6), 1010–1021.
- Hazell, L. et S. A. Shakir (2006). Under-reporting of adverse drug reactions. *Drug Safety* 29(5), 385–396.
- Lamb, A., M. J. Paul, et M. Dredze (2013). Separating fact from fear : Tracking flu infections on twitter. In *HLT-NAACL*, pp. 789–795.
- Saif, H., Y. He, et H. Alani (2012). Alleviating data sparsity for twitter sentiment analysis. In *CEUR Workshop Proceedings (CEUR-WS. org)*.