# Construction d'un vocabulaire patient/médecin

Mike Donald TAPI NZALI[1,2] **S. Bringay**[2] C. Lavergne[1]
T. Opitz[1] J. Azé[2] C. Mollevi[3]

[1]IMAG, Université de Montpellier, France

[2]LIRMM, Université de Montpellier, France

[3]ICM, Institut du Cancer de Montpellier, France

1er juillet 2015





# Plan

# Financements

## 3 projets

- Patients Mind : Réseau National des MSH (Maisons des Sciences de l'Homme) 2013-2014 et 2014-2015
- SFIR : Projet ANR 2012-2015
- Qualité de vie : Contrat Iresp 2012-2015

# Qualité de vie


Média sociaux


QLQ BR-23

Formulaire sur la qualité de vie

# Qualité de vie

## Travaux réalisés : extraction supervisée et non supervisée

- des thèmes d'intérêt des patients
- de la temporalité (avant, après la chirurgie...)
- des sentiments
- des locuteurs

Ma sœur a très peur de ne pas supporter sa chimiothérapie du 8 janvier 2015.

# Difficulté : traiter les textes mal rédigés des patients

[redacted] [163904] posté le 05/01/2012 09:45:00 par [redacted]

Je m'interroge aussi. J'ai 31 ans et on m'a annoncé clairement que jamais je ne pourrais avoir d'enfants. Maintenant pkoi ne pas bloquer les ovaires...si reeellement je ne donnerai jms la vie, ces trucs ne servent plus qu'a nourir mon cancer.

Je vois mon gygy ce mois ci je vais bien lui en parler

# Acquisition du vocabulaire Patient/Médecin

## Motivations : Faciliter les traitements des textes issus des réseaux sociaux

- Rechercher plus efficacement des contenus dans ces messages (mots-clés patients)
- Classifier

## Objectif

- Étendre une ontologie médicale existante avec le vocabulaire fréquemment utilisé par les patients (INCA)

## Intuition

- Utiliser les Patient Author Text pour extraire le vocabulaire semi-automatiquement

# CHV : Consumer Health Vocabulary

## 2 CHV en anglais

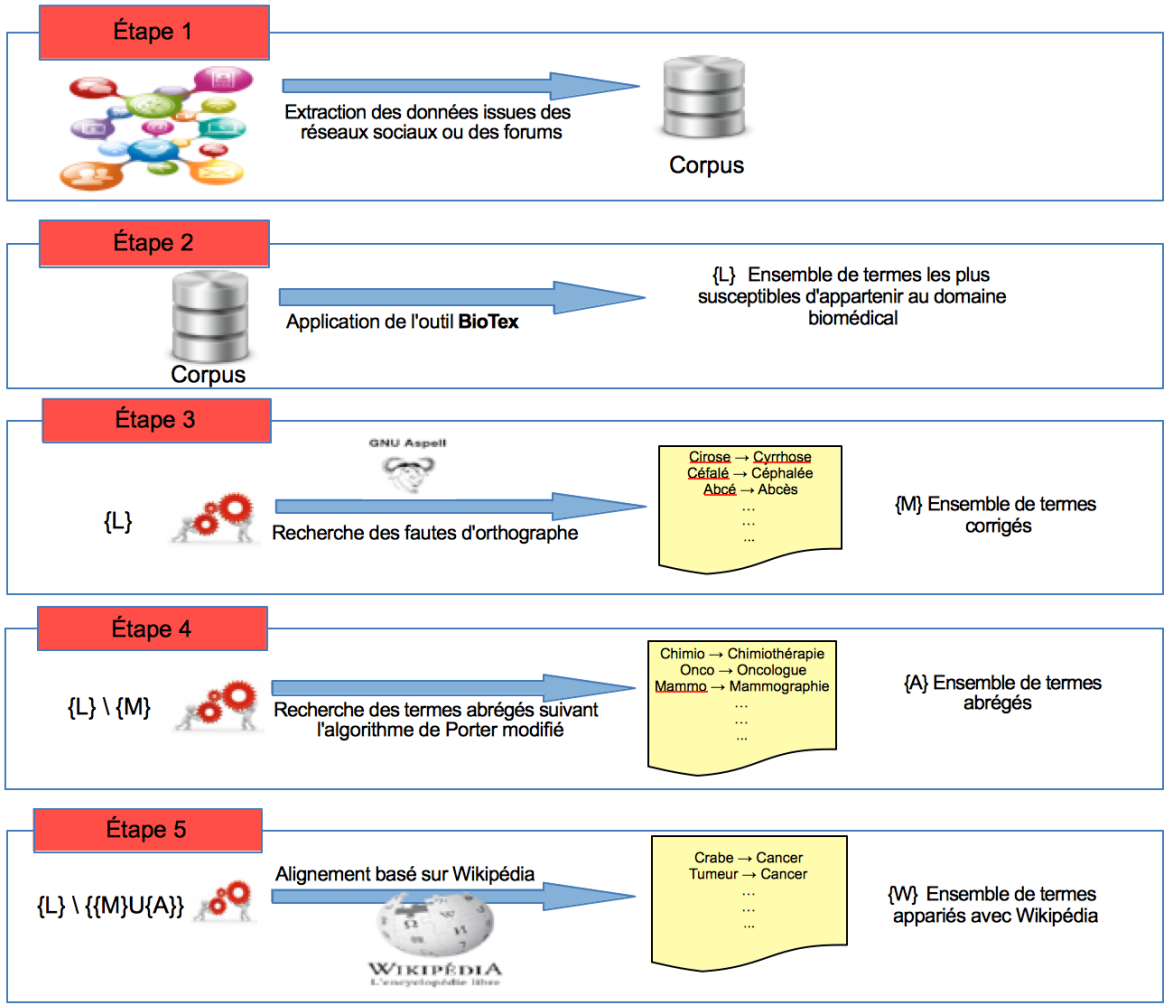
- *MedlinePlus* produit par la National Library of Medecine
- *Open and Collaborative Consumer Health Vocabulary* inclus dans l'UMLS

## Utilisation

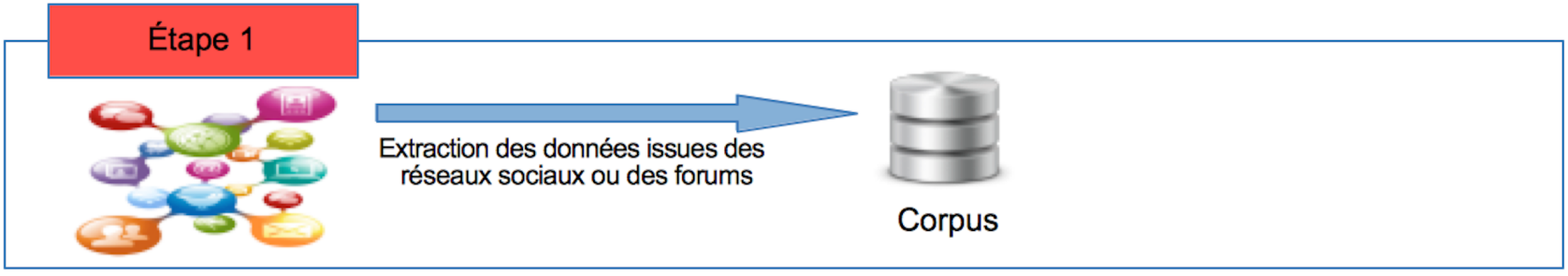
- Réduire les écarts de connaissances patient/médecin
- Améliorer la lisibilité des documents médicaux
- Augmenter la confiance que l'on a dans le discours du médecin (mieux compris)

# Plan

# Méthode proposée



# Étape 1 : Acquisition des données



# Forums de santé : de la création des sites jusqu'en octobre 2014

## http ://www.cancerdusein.org/le-forum

- 665 membres
- 1050 topics
- 17 000 messages

## http ://www.lesimpatientes.com

- 4 486 membres
- 93 topics
- 132 789 messages

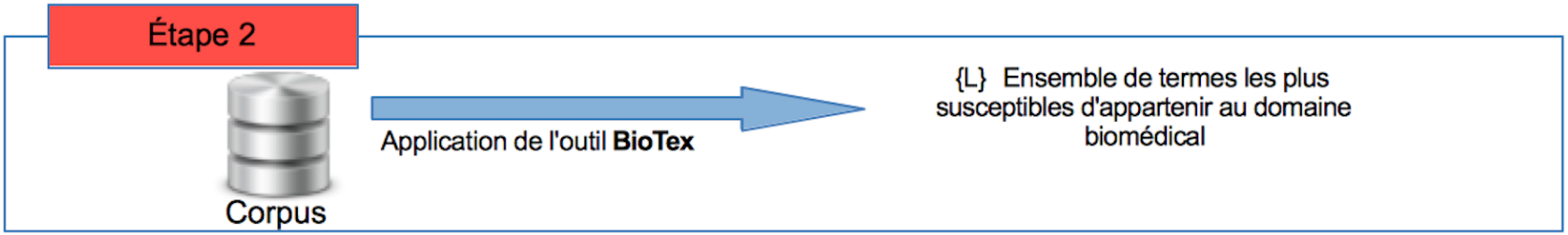
# Acquisition des données : Groupes du réseau social facebook

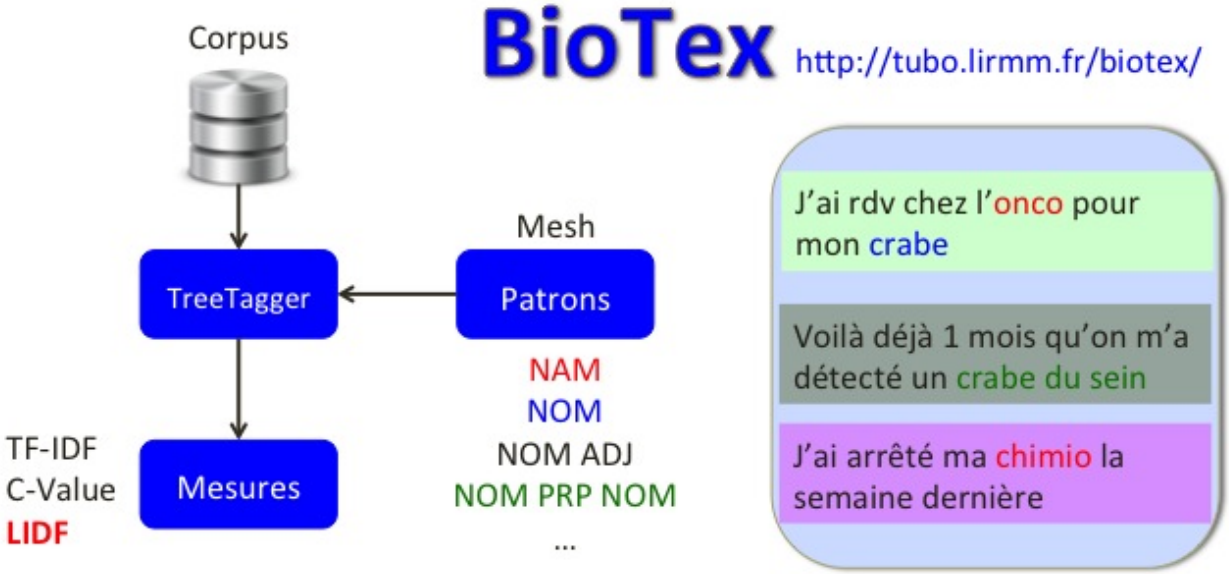
## Groupes facebook

- Post d'un utilisateur (ou membre)
- Date et heure du post
- Commentaires/"like" liés aux posts

- 5 groupes
- 1 389 utilisateurs
- 96 792 messages

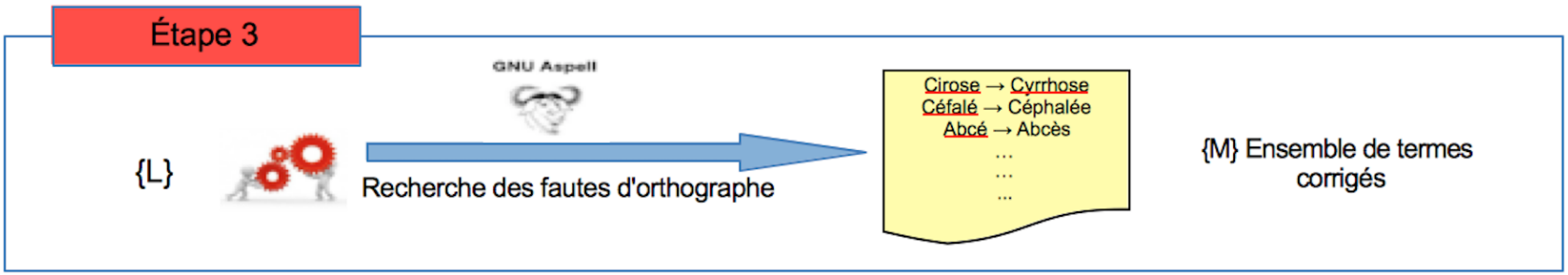
# Étape 2 : Extraction des termes candidats



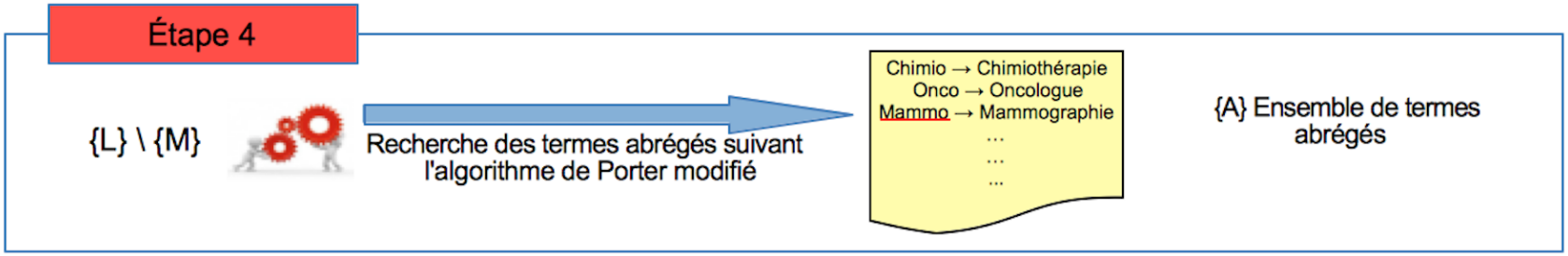

(Lossio-Ventura et al. 2014) BIOTEX : A system for Biomedical Terminology Extraction, Ranking, and Validation.

# Étape 3 : Erreurs d'orthographe fréquentes




- Modification de l'algorithme ASPELL : INCA
- Mot patient $\rightsquigarrow$ Mot médecin
  - Cirose $\rightsquigarrow$ Cirrhose
  - Céfalé $\rightsquigarrow$ Céphalée
  - Abcé $\rightsquigarrow$ Abcès

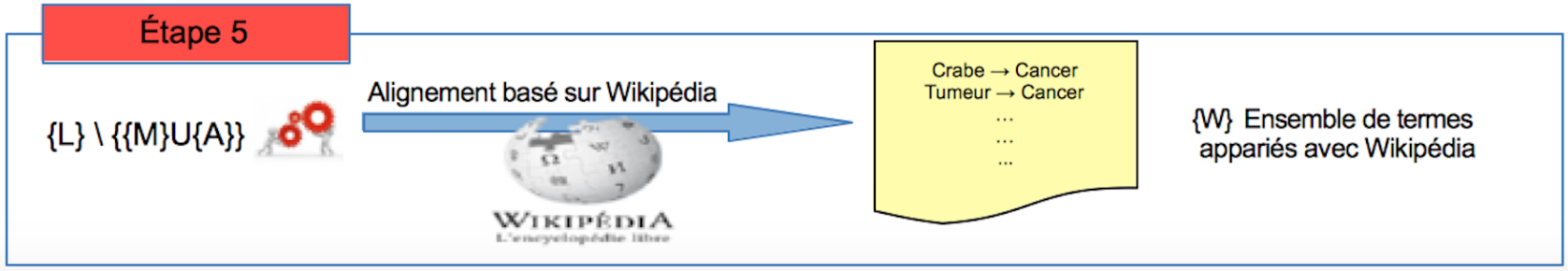
# Étape 4 : Détection des termes abrégés




- Liste de suffixes : *logue, logie, thérapie, thérapeute...*
- Mot patient $\rightsquigarrow$ Mot médecin
  - onco $\rightsquigarrow$oncologue
  - chimio $\rightsquigarrow$ chimiothérapie

(Paternostre et al. 2002) Carry, un algorithme de désuffixation pour le français.

# Étape 5 : Alignement basé sur Wikipédia





$$M(w_k, t_n) = \frac{N(w_k, W(t_n)) + N(t_n, W(w_k))}{2}$$

$$Poids(w_j, t_i) = \frac{M(w_j, t_i)}{\sum_{k=1}^{|W|} M(w_k, t_i)}$$

où $N(a, W(b))$ : fréquence de $a$ dans la page wikipédia de $b$.

# Plan

# Protocole de validation

## Connaissances extraites

- Relation *r*
- Mot patient *pat*
- Mot médecin *med*
- Etiquette $meth \in \{ortho, abb, asso\}$

## Validation automatique

- JeuxDeMot : réseau lexical issu d'un jeu contributif (http ://www.jeuxdemots.org/diko.php)

## Validation manuelle

- Relations non validées automatiquement
- Annotations manuelles : 4 annotateurs

# Résultats

## Validation automatique

- Sur 432 relations obtenues, 211 ont été validées automatiquement (48, 8%)
- 32 relations - erreurs orthographiques (7, 4%)
- 10 relations - abbréviations (2, 3%)
- 169 relations - associations wikipedia retrouvées dans JDM (39, 1%)

## Validation manuelle

- Sur 218 relations restantes, 93 ont été retenues après consencus (42, 6%)
- Kappa : 0,21 (faible)

# Plan

# Conclusions

## Des travaux préliminaires

+ Précision globale : 71% vs. 31% (Doing-Harris 2011) pour un CHV en anglais généraliste
+ Alignements d'expressions composées
+ Solicitation de l'expert limitée > Impact du poids
- Peu de relations > Spécificité de la ressource pour le cancer du sein (dictionnaire INCA - 1 227 termes)

(Doing-Harris 2011) Computer-assisted update of consumer health vocabulary through mining of social network data.

# Perspectives

## À court terme

- Applications à d'autres domaines que la cancérologie
- Rappel : combien de mots patients intéressants a t'on oublié ?
- Impact des données en entrée (forums » réseaux sociaux...)

## À plus long terme

- Choix Wikipedia (ni grand public, ni scientifique) > Autres mesures d'appariement « google, yahoo... »
- Étendre une ontologie au format « SKOS » > export dans BioPortal
- Évaluer l'effet sur la classification

# Ressources

## En français

- Vocabulaire Patient : http ://www.lirmm.fr/ tapinzali/Ressources/VocPatMed
- Vocabulaire Sentiment : https ://www.lirmm.fr/patient-mind/ > onglet Ressource

# Principal auteur de ces travaux

Mike-Donald.Tapi-Nzali@lirmm.fr