MuEVo, un vocabulaire multi-expertise (patient/médecin) dédié au cancer du sein

Solène Eholié, Mike Donald Tapi Nzali, Sandra Bringay, Clement Jonquet

MuEVo, un vocabulaire multi-expertise (patient/médecin) dédié au cancer du sein

Solène Eholié[1], Mike Donald Tapi Nzali[1,2],
Sandra Bringay[1], Clement Jonquet[1,3]

[1] LABORATOIRE D'INFORMATIQUE, DE ROBOTIQUE ET DE MICROÉLECTRONIQUE DE MONTPELLIER (LIRMM), Université de Montpellier, France
prenom.nom@lirmm.fr

[2] INSTITUT MONTPELLIÉRAIN ALEXANDER GROTHENDIECK (IMAG), Université de Montpellier, France
mike-donald.tapi-nzali@univ-montp2.fr

[3] CENTER FOR BIOMEDICAL INFORMATICS RESEARCH (BMIR), Stanford University, USA

Abstract : Il existe un écart notable à la fois d'ordre lexical et sémantique entre le vocabulaire des professionnels de la santé et celui des patients. À notre connaissance, il n'existe pas de ressource formalisée pour le français liant ces deux niveaux de vocabulaire. Nous présentons dans ce travail, une formalisation en SKOS d'un vocabulaire reliant ces deux niveaux d'expertise dans le cadre de la thématique du cancer du sein ainsi qu'une méthode d'alignement de la terminologie résultante, MuEVo, à des terminologies biomédicales de référence à savoir MeSH, SNOMED et MedDRA.

Mots-clés : Système d'organisation des connaissances (SOC), terminologies biomédicales, vocabulaire patient

1 Introduction et motivations

Selon une étude de TNS Sofres[1] réalisée en 2013, un Français sur deux a déjà recherché ou échangé des informations sur sa santé via le Web. De même, on trouve en ligne de très nombreuses publications scientifiques produites par les professionnels de santé comme dans la base bibliographique PubMed[2]. Des données en très grande quantité sont donc disponibles sur des sujets médicaux décrits selon deux niveaux d'expertise : patient et médecin.

Il existe de nombreuses formalisations du vocabulaire médecin en français, sous la forme de terminologies comme MeSH, SNOMED ou MedDRA[3]. Cependant, McCray *et al.* (1999) ont montré qu'il existe un écart notable entre le vocabulaire scientifique et technique utilisé par les médecins et celui vulgarisé des patients. Cet écart lexical et/ou sémantique handicape par exemple les patients dans leur recherche d'informations médicales (Kogan *et al.*, 2001). Certains travaux (Zeng & Tse, 2006; Jiang *et al.*, 2013) se sont donc intéressés à la création de CHV[4] (Consumer Health Vocabulary).

Le travail présenté dans cet article fait suite aux travaux de Tapi Nzali *et al.* (2015) qui

[1] http://www.patientsandweb.com/wp-content/uploads/2013/04/A-la-recherche-du-ePatient-externe.pdf

[2] http://www.ncbi.nlm.nih.gov/pubmed

[3] http://mesh.inserm.fr/mesh/, http://www.meddra.org

[4] vocabulaire composé d'un ensemble de termes utilisés par les non-experts (patients, leurs familles, etc.) pour exprimer des concepts médicaux

ont proposé une méthode originale de construction *semi-automatique* d'un vocabulaire patient/médecin à partir des médias sociaux. Ce vocabulaire que nous formalisons dans cet article et appelons MuEVo, est spécifique à la thématique du cancer du sein. Notre objectif est maintenant de construire une ressource structurée, exploitable par une machine et conforme aux standards du Web sémantique, qui permettra de faire le pont entre ces niveaux d'expertise, afin de pouvoir effectuer des traitements automatiques (e.g., recherche d'information (Zarro & Lin, 2011), classification automatique).

Nous proposons dans cet article une formalisation en SKOS[5] du vocabulaire MuEVo (section 2) puis une méthodologie pour l'aligner avec les terminologies présentes dans le serveur de terminologies francophones, SIFR BioPortal (section 3) (Jonquet *et al.*, 2016).

2 Formalisation SKOS

2.1 Présentation des données

Pour valider ce travail préliminaire, nous avons utilisé 173 relations entre termes patient/médecin, spécialisées sur le cancer du sein, issues de (Tapi Nzali *et al.*, 2015). Ces relations ont été obtenues via un alignement entre un corpus patient constitué d'un ensemble de messages extraits de médias sociaux (forums[6] et groupes Facebook publics[7]) et un vocabulaire médecin cible à savoir la liste de termes de référence proposée par l'INCa[8] (Delavigne, 2012). À chaque relation, sont associés un type (*abréviation*, *erreur d'orthographe* ou *association*), la méthode utilisée pour détecter la relation et un poids assigné par la méthode. Le tableau 1 présente quelques exemples de relations.

Terme patient	Terme médecin	Type de relation
nez	pharynx	association
abaltion	ablation	erreur d'orthographe
onco	oncologue	abréviation
traitement hormonal	hormonothérapie	association

Tableau 1 : Exemples de relations patient/médecin extraits de Tapi Nzali *et al.* (2015)

2.2 Spécification du modèle

SKOS est une recommandation du W3C pour représenter des vocabulaires controlés (Miles *et al.*, 2005). C'est un standard très utilisé dans la communauté du Web sémantique. Le thésaurus AGROVOC[9] par exemple est formalisé en SKOS. L'unité de connaissance en SKOS est le *skos:Concept*. Un *skos:Concept* est une ressource RDF qui formalise une idée, une réalité.

[5] Simple Knowledge Organization System - https://www.w3.org/2004/02/skos/

[6] *cancerdusein.org*

[7] *Cancer du sein*, *Octobre rose 2014*, *Cancer du sein - breast cancer, brustkrebs*

[8] Institut National de Cancer - http://www.e-cancer.fr/Dictionnaire/

[9] http://aims.fao.org/fr/agrovoc

On peut lui associer au plus un label préféré (*skos:prefLabel*), c'est-à-dire la dénomination privilégiée du concept. D'autres termes peuvent être associés au concept comme variantes valides (*skos:altLabel*) ou variantes existantes mais déconseillées (*skos:hiddenLabel*).

Ce modèle initial ne suffit pas pour conserver les méta-données relatives au processus d'extraction de chaque relation patient/médecin, à savoir le poids de la relation, la méthode ayant généré la relation et enfin son type. Nous avons donc étendu notre usage de SKOS pour intégrer la provenance de la relation, en particulier à l'aide du vocabulaire PROV[10] qui est une recommandation du W3C pour représenter les informations de provenance. Le modèle final obtenu se présente comme décrit sur la figure 1.

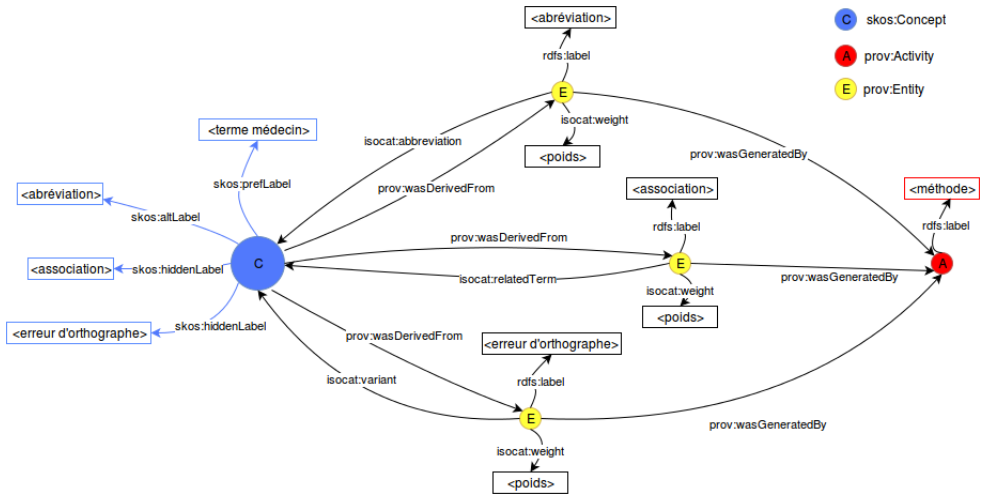

Figure 1: Modèle de representation des relations patient/médecin en SKOS+PROV dans MuEVo

Chaque *skos:Concept* (représenté en bleu sur la figure) est une représentation formelle de toutes les relations trouvées pour un "terme médecin" donné. Pour un *skos:Concept* donné, l'identifiant implicite est le "terme médecin" qui le décrit. Il doit donc être unique. On l'assigne alors au champ *skos:prefLabel*. Chaque mesure mise en jeu est représentée par une *prov:Activity* (en rouge sur la figure). Par souci de lisibilité, une seule méthode est mentionnée sur la figure mais plusieurs peuvent être utilisées. Chaque relation reliant le "terme médecin" du concept à un "terme patient" est représentée via les labels standards SKOS (*skos:altLabel* ou *skos:hiddenLabel*). En complément, nous conservons les informations de provenance à l'aide d'une entité RDF de type *prov:Entity* reliée au concept par un label ISOcat qui sert à préciser le type de la relation. La fonction de détermination des labels SKOS et ISOcat est donnée par le tableau 2. Le poids de la relation est également stocké dans la *prov:Entity* correspondante à l'aide du label *isocat:weight*. Chaque entité modélisant une relation avec le "terme médecin" du concept est stockée dans le *skos:Concept* associé au "terme médecin" à l'aide d'une information de provenance *prov:wasDerivedFrom*.

Après formalisation des relations patient/médecin en SKOS, nous souhaitons aligner *MuEVo* à des terminologies de référence.

[10]https://www.w3.org/TR/prov-dm/

Type de la relation	Label SKOS	Label ISOcat
abréviation	*skos:altLabel*	*isocat:abbreviation*
erreur d'orthographe	*skos:hiddenLabel*	*isocat:variant*
association	*skos:hiddenLabel*	*isocat:relatedTerm*

Tableau 2 : Fonction d'attribution des labels SKOS et ISOcat

3 Alignement du vocabulaire médecin

BioPortal (Noy *et al.*, 2009) est un serveur de terminologies biomédicales. Dans le cadre du projet SIFR[11], une instance de BioPortal[12] donne accès à une version *en français* des principales terminologies du domaine biomédical (Jonquet *et al.*, 2016). Via ce portail Web, un utilisateur peut partager une terminologie sur le serveur et la relier à celles déjà disponibles via des mappings étiquetés là encore à l'aide de SKOS. Nous avons donc chargé MuEVo dans SIFR BioPortal après l'avoir formalisé précédent et souhaitons maintenant relier nos concepts à ceux des terminologies biomédicales standards disponibles.

Le vocabulaire médecin initial, la liste de l'INCa, est une liste plate et de taille réduite. La création de ces liens (mappings) nous permettra de bénéficier de la connaissance plus large et structurée offerte par ces terminologies lors de l'usage explicite du vocabulaire patient/médecin pour indexer sémantiquement le contenu de forums par exemple.

Nous visons uniquement l'établissement de liens d'équivalence *skos:exactMatch* et de liens hiérarchiques : hyperonymie ou généralisation (*skos:broadMatch*) et hyponymie ou spécialisation (*skos:narrowMatch*). Pour nos premières expérimentations, nous nous sommes limités à trois terminologies cibles en français : MeSH, SNOMED et MedDRA. L'approche d'alignement adoptée s'articule en deux phases : un alignement direct et indirect.

3.1 Alignement direct

La phase d'alignement direct consiste à rechercher à l'aide de l'API REST[13] de BioPortal chaque terme de l'INCa dans notre vocabulaire. Si l'on retrouve exactement le même terme comme appellation préférée ou variante d'un concept d'une terminologie cible, alors on établit un lien d'équivalence, *skos:exactMatch*, entre le concept étudié et celui de la terminologie cible. Sur la figure 2, le terme *abdomen* est l'appellation préférée d'un concept d'une terminologie standard. Le concept *Abdomen* est alors relié. Le terme *cancer* apparaît comme variante du concept standard *Tumeurs* donc un lien *skos:exactMatch* est créé.

Pour les termes n'apparaissant comme label d'aucun concept des terminologies cibles, nous recherchons un alignement indirect.

[11] Semantic Indexing of French Biomedical Data Resources - http://www.lirmm.fr/sifr/

[12] http://bioportal.lirmm.fr/

[13] http://data.bioportal.lirmm.fr/documentation

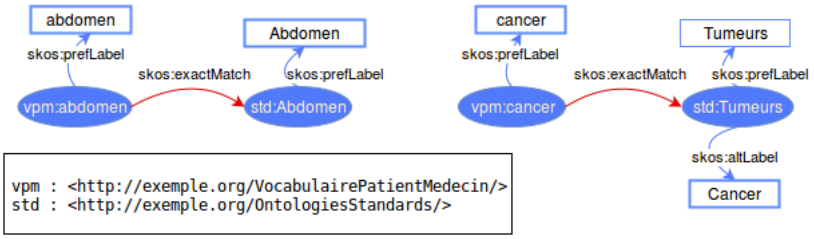

Figure 2: Exemples d'alignements directs

3.2 Alignement indirect

Nous faisons ici l'hypothèse qu'il existe des ressources plus généralistes intermédiaires entre la liste de l'INCa et les entrées des terminologies standards cibles. Ainsi, pour un "terme médecin" t_m donné de MuEVo, il s'agit d'utiliser des ressources externes, Wiktionary[14] (Meyer & Gurevych, 2012) dans notre cas, pour trouver des termes en lien avec t_m par une relation sémantique de type *synonyme, hyperonyme, hyponyme* et qui apparaissent eux comme labels dans les terminologies cibles. Le protocole adopté se décrit comme suit :

1. On recherche[15] le terme médecin dans Wiktionary. Si l'entrée existe alors on récupère l'ensemble des synonymes, hyperonymes et hyponymes

2. Pour chaque terme *t* de la liste ainsi constituée, une recherche parmi les labels des terminologies cibles à l'aide l'API de BioPortal est effectuée

3. En cas de succès, on définit les mappings suivants entre notre concept initial $C_{initial}$ et le concept C_{cible} de la terminologie cible retourné par l'API de recherche : si *t* était un synonyme : $C_{initial}$ *skos:exactMatch* C_{cible}; si *t* était un hyperonyme : $C_{initial}$ *skos:broadMatch* C_{cible}; si *t* était un hyponyme : $C_{initial}$ *skos:narrowMatch* C_{cible}.

Par exemple (voir figure 3), pour le terme *cure*, un synonyme est *traitement*; *oncologue* a pour hyperonyme *médecin spécialiste* et un hyponyme de *atome* est *ion*.

4 Résultats

Le vocabulaire MuEVo est consultable[16] sur SIFR BioPortal. Les résultats de l'alignement des 64 termes médecin du vocabulaire étudié sont résumés dans le tableau 3. Ces alignements ont été réalisés via un programme que nous avons écrit pour automatiser le processus. Les trois terminologies cibles choisies couvrent 84,38% du vocabulaire médecin : MeSH (70,31%), SNOMED (51,56%) et MedDRA (37,5%). Parmi les 10 termes manquants, 3 ont pu être alignés avec succès grâce aux hyponymes extraits de Wiktionary. Pour les 7 restants, l'alignement est manuel.

[14]http://www.wiktionary.org

[15]Nous automatisons la recherche en utilisant l'API JWKTL (Zesch *et al.*, 2008) après l'avoir adapté pour le français

[16]MuEVo - http://bioportal.lirmm.fr/ontologies/MUEVO

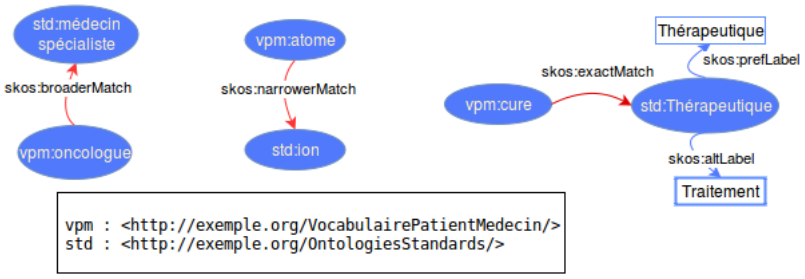

Figure 3: Exemple d'alignement indirect pour les termes oncologue, atome et cure

	Nombre	**Exemples**
1A : Singulier	51	abdomen -> Abdomen (MeSH)
1B : Pluriel	17	glucide -> Glucides (MeSH)
1A+1B	54	
2 : Hyponymes	3	atome -> ion (SNOMED)

Tableau 3 : Résultats obtenus automatiquement pour 64 termes en entrée de la phase d'alignement direct (1A, 1B) et 10 termes à la phase d'alignement indirect 2

5 Conclusions et perspectives

Dans cet article, nous avons proposé une méthode de formalisation d'un vocabulaire patient/médecin en terminologie au format SKOS ainsi que des pistes pour aligner le vocabulaire médecin correspondant aux terminologies de référence existantes. Une telle ressource peut être utilisée pour rendre des productions médicales (dossiers médicaux par exemple) plus compréhensibles aux patients (Zeng & Tse, 2006) ou pour de l'indexation multi-expertise (Soualmia *et al.*, 2003). Dans la suite de nos travaux, nous envisageons d'explorer trois points : la structuration interne de MuEVo à l'aide des relations sémantiques extraites de définitions (Medelyan *et al.*, 2009), l'acquisition de nouvelles relations patient/médecin en utilisant le métathésaurus UMLS[17] (Keselman *et al.*, 2008), plus large que celui de l'INCa et enfin l'exploitation de la ressource pour des tâches de classification supervisées et non supervisées exploitant la hiérarchie des terminologies auxquelles MuEVo est aligné (Wijewickrema *et al.*, 2015).

6 Remerciements

Ce travail est réalisé au sein du projet SIFR financé par le programme JCJC ANR-12-JS02-01001 et le projet "Comparison of longitudinal analysis models of the health-related quality of life in oncology" financé par l'IRESP.

[17]Unified Medical Language System - https://www.nlm.nih.gov/research/umls/

References

DELAVIGNE V. (2012). Peut-on «traduire» les mots des experts? un dictionnaire pour les patients atteints de cancer. *Dictionnaires et traduction*, p. 233–263.

JIANG L., YANG C. C. & LI J. (2013). Discovering consumer health expressions from consumer-contributed content. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, p. 164–174: Springer.

JONQUET C., ANNANE A., BOUARECH K., EMONET V. & MELZI S. (2016). SIFR BioPortal : Un portail ouvert et générique d'ontologies et de terminologies biomédicales françaises au service de l'annotation sémantique. In *16th Journées Francophones d'Informatique Médicale, JFIM'16*.

KESELMAN A., SMITH C. A., DIVITA G., KIM H., BROWNE A. C., LEROY G. & ZENG-TREITLER Q. (2008). Consumer health concepts that do not map to the umls: where do they fit? In *Journal of the American Medical Informatics Association*, volume 15, p. 496–505: Elsevier.

KOGAN S., ZENG Q., ASH N. & GREENES R. A. (2001). Problems and challenges in patient information retrieval: a descriptive study. In *Proceedings of the AMIA Symposium*, p. 329: American Medical Informatics Association.

MCCRAY A. T., LOANE R. F., BROWNE A. C. & BANGALORE A. K. (1999). Terminology issues in user access to web-based medical information. In *Proceedings of the AMIA Symposium*, p. 107: American Medical Informatics Association.

MEDELYAN O., MILNE D., LEGG C. & WITTEN I. H. (2009). Mining meaning from wikipedia. *International Journal of Human-Computer Studies*, **67**(9), 716–754.

MEYER & GUREVYCH (2012). Wiktionary: A new rival for expert-built lexicons? exploring the possibilities of collaborative lexicography. In *Granger, S. and Paquot, M., International Conference on Dublin Core and Metadata Applications*.

MILES A., MATTHEWS B., WILSON M. & BRICKLEY D. (2005). Skos core: simple knowledge organisation for the web. In *International Conference on Dublin Core and Metadata Applications*, p. pp–3.

NOY N. F., SHAH N. H., WHETZEL P. L., DAI B., DORF M., GRIFFITH N., JONQUET C., RUBIN D. L., STOREY M.-A., CHUTE C. G. *et al.* (2009). Bioportal: ontologies and integrated data resources at the click of a mouse. In *Nucleic acids research*, p. gkp440: Oxford Univ Press.

SOUALMIA L., DARMONI S. J., DOUYÈRE M. & THIRION B. (2003). Modelisation of consumer health information in a quality-controlled gateway. In *Studies in health technology and informatics*, p. 701–706: IOS Press; 1999.

TAPI NZALI M. D., BRINGAY S., LAVERGNE C., OPITZ T., AZÉ J. & MOLLEVI C. (2015). Construction d'un vocabulaire patient/médecin dédié au cancer du sein à partir des médias sociaux. In *Ingénierie des Connaissances*, p. 9–20.

WIJEWICKREMA C. M. *et al.* (2015). Impact of an ontology for automatic text classification. *Annals of Library and Information Studies (ALIS)*, **61**(4), 263–272.

ZARRO M. & LIN X. (2011). Using social tags and controlled vocabularies as filters for searching and browsing: A health science experiment. *Mountain View, CA*.

ZENG Q. T. & TSE T. (2006). Exploring and developing consumer health vocabularies. In *Journal of the American Medical Informatics Association*, volume 13, p. 24–29: Elsevier.

ZESCH T., MÜLLER C. & GUREVYCH I. (2008). Extracting lexical semantic knowledge from wikipedia and wiktionary. In *LREC*, volume 8, p. 1646–1652.