Modélisation de connaissances médicales pour améliorer le descriptif des maladies humaines avec leurs plus pertinents signes caractéristiques

Adama Sow, Abdoulaye Guissé, Oumar Niang

Laboratoire Traitement de l'Information et Systèmes Intelligents (LTISI)
Département du Génie Informatique et Télécommunications
Ecole Polytechnique Thiès (EPT), THIES, Sénégal
asow@ept.sn, aguisse@ept.sn, oniang@ept.sn

Résumé. Guérir un patient malade nécessite un diagnostic médical avant de proposer un traitement approprié. Avec l'explosion des connaissances médicales, nous nous intéressons à leur exploitation pour aider les médecins dans la collecte d'informations lors du processus de diagnostic. Le présent article porte sur la mise en place d'un modèle de données ciblant des connaissances disponibles dans des ressources aussi bien formelles que non formelles. L'objectif est de fusionner les forces de toutes ces ressources pour fournir l'accès à une variété de connaissances partagées facilitant l'identification et l'association des maladies humaines et à l'ensemble de leurs signes caractéristiques pertinents disponibles tels que les symptômes et les signes cliniques.
D'un côté, nous proposons une ontologie produite à partir d'une intégration de plusieurs ontologies et terminologies médicales existantes et ouvertes. D'un autre côté, nous exploitons des cas réels de patients dont le diagnostic aura déjà été confirmé par des médecins. Ils sont transcrits dans des rapports textuels en langue naturelle, et nous démontrons ici que leur analyse permet d'enrichir la liste des signes de chaque maladie. Ce travail aboutit alors à une base de connaissances chargée à partir des ontologies cibles connues sur la plate-forme de bioportail telles que DOID, MESH et SNOMED pour la sélection des maladies, SYMP, et CSSO pour tous les signes existants. L'échantillon de cas textuels choisis porte sur des maladies tropicales.

1 Introduction

Le diagnostic médical, tel que décrit dans le livre de Balogh et al. (2015) est une activité cognitive centrée sur le patient dont la compétence quintessentielle appartient au médecin. C'est un procédé qui consiste en une collecte continue des informations médicales qu'effectue le médecin avant de les intégrer et de les interpréter pour la gestion des problèmes de santé de son patient. Le diagnostic inclut généralement quatre étapes itératives : i) l'acquisition des informations contextuelles qui prend en compte les antécédents, les examens physiques premiers, les examens approfondis ou analyses cliniques avancées, ii) la formulation d'hypothèses

de diagnostics potentiels sous forme d'une liste d'une ou de plusieurs maladies, iii) la mise en cohérence des informations collectées avec chacune des hypothèses posées, iv) et enfin l'évaluation de chaque hypothèse pour identifier et confirmer le diagnostic le plus certain, sinon tout le processus doit être repris en élargissant la collecte.

Cette première étape de collecte est aussi capitale que complexe pour le médecin surtout lorsque cela nécessite de recourir rapidement, en un temps réduit, à des masses de connaissances médicales qui ne cessent d'exploser à l'échelle internationale. C'est dans l'optique d'assister les médecins dans l'exploitation de ces connaissances, que se situe notre recherche. Notre objectif consiste plus globalement à développer un moteur de recherche (Figure 1) qui guide l'accès aux informations médicales pertinentes à chacune des étapes du processus de diagnostic. Ce moteur permettrait de naviguer dans une base de connaissances constituées d'une ontologie médicale et d'une base de cas de diagnostics cliniques ayant déjà été validés par des médecins.

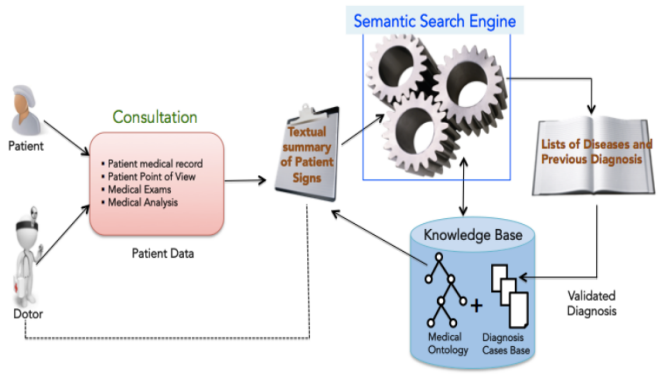

FIG. 1 – *Description du processus d'aide au diagnostic médical*

Le présent article met le focus sur la modélisation de cette base de connaissances (BC). Il s'agit de mettre en place un modèle de données ciblant des connaissances disponibles dans des ressources aussi bien formelles que non formelles. Le but est de fusionner les forces de toutes ces ressources pour fournir l'accès à une variété de connaissances partagées facilitant l'identification et l'association des maladies humaines et à l'ensemble de leurs signes caractéristiques pertinents disponibles tels que les symptômes et les signes cliniques.

Le noyau de cette BC est une ontologie produite à partir d'une fédération de plusieurs ontologies et terminologies médicales existantes et ouvertes. L'ontologie obtenue couvre une multitude d'informations descriptives des maladies mais aussi décrit la typologie et la sémantique des informations à collecter un patient. En effet, dans les ontologies existantes nous trouvons d'une part des ontologies de maladies associées à des symptômes dont l'exhaustivité est à éclaircir, et d'autre part des ontologies qui conceptualisent tous les signes susceptibles d'être identifiés chez un malade mais aucun lien avec les maladies concernées n'est identifié. Il s'agit notamment des signes cliniques dont les valeurs sont obtenues à partir des examens approfondis. Les ontologies de maladies visant une conception générique ne prennent pas en compte ces signes qui sont tout de même connus et formalisés dans les ontologies de signes.

Nous nous proposons alors d'enrichir notre ontologie en associant aussi les maladies à leurs signes cliniques. Cela est rendu possible en exploitant des cas réels de patients. Ces cas sont généralement transcrits soit dans des formulaires soit dans des rapports textuels en langue naturelle par des médecins. Ces derniers archivent systématiquement toutes les données de tout patient pris en charge dans son dossier médical (DM). Mêmes si les DM sont confidentiels, nous avons pu nous procurer, en collaboration avec des hôpitaux locaux, des rapports anonymes descriptifs de cas. L'analyse de leur contenu permet d'identifier l'ensemble des symptômes observés sur un patient ainsi les signes cliniques ayant permis de confirmer un diagnostic précis. Toutefois, ces derniers signes étant spécifiques à un patient donné, ils sont associés à une maladie via le cas qui les porte. Les cas sont ainsi stockés dans la base de connaissances. Chaque maladie de notre ontologie est décrite par l'ensemble des signes observés et vérifiés chez tous les patients portant cette même maladie. L'association maladie et ses signes est donc alimentée continuellement au fur à mesure qu'il y a des nouveaux cas de diagnostic.

Ainsi, dans la *section 2* , nous dressons une étude de l'état de l'art se rapportant aux ontologies médicales et de leur utilisation dans les systèmes de diagnostic. Nous présentons ensuite, dans la *section 3*, notre sélection d'ontologies médicales en vue de constituer une ontologie plus adéquate pour le diagnostic. Enfin, dans la *section 4*, avant la conclusion, nous décrivons notre méthode d'intégration d'ontologies cibles, en montrant le modèle de la structure de données utilisée, la récupération des informations utiles à charger dans l'ontologie, l'analyse textuelle de cas cliniques réels pour l'association des maladies à leurs signes pertinents, et la description de l'ontologie résultante en chiffres.

2 Ontologies médicales

Pour établir le diagnostic (Balogh et al. (2015)), il est important pour le médecin de recouper toutes les informations sur l'état de santé de son patient. Il s'agit précisément (Bringay et al. (2005)) de l'avis du patient sur son état pour identifier ses douleurs, des examens physiques d'usages effectués par un médecin à l'occasion de consultations, et des examens approfondis (cliniques et paracliniques), qui à la demande du médecin, permettent l'identification des signes les plus complexes et implicites.

Dans les systèmes d'aide au diagnostic médical (S-Ortiz et al. (2013)), cette phase de collecte est une activité cognitive où la sémantique des informations est contrôlée au travers des connaissances connues dans le jargon médical. C'est dans ce but précis que les ontologies médicales ont été conçues (Hoehndorf et al. (2015); Anbarasi et al. (2013)). Elles mettent en place des vocabulaires médicaux communs reposant sur des concepts partagés qui facilitent l'interopérabilité des documents entre les acteurs du domaine et surtout l'élaboration des connaissances. Les ontologies médicales représentent une évolution des thésaurus médicaux ; elles ne se limitent pas à définir une terminologie mais elles vont plus loin en modélisant clairement les entités médicales telles que les maladies, leurs signes caractéristiques, leurs traitements connus, jusqu'aux processus hospitaliers de prise en charge des patients.

Nous nous intéressons ici aux ontologies médicales des maladies humaines. La liste est longue et chaque ontologie présente ses propres spécificités. Mais globalement la plupart des maladies connues sont couvertes et renvoient chacune à un concept regroupant ses divers termes nominatifs et leurs synonymes, ses différentes définitions et axiomes textuelles et ses signes caractéritiques. Ces derniers indiquent entre autres des signes cliniques et des symp-

tômes (Cox et al. (2014)), mais aussi éventuellement l'agent en cause de la maladie, le mode de transmission, et la localisation dans l'anatomie humaine. Aussi, des liens taxonomiques (ou hiérarchiques) sont définis d'entre les concepts de maladies afin de les classer dans des catégories de maladies. Cela est facilité par le fait que ces ontologies sont implémentées dans des langages formels, comme le OWL (Ontology Web Language), basés sur le principe des graphes conceptuels, du concept orienté objet et de la logique de description.

Les supports d'aide au diagnostic médical sont des systèmes experts (Wang et Tansel (2013)) où les ontologies médicales sont utilisées comme une base de connaissances (S-Ortiz et al. (2013); Mohammed et al. (2012)). Elles sont exploitées globalement pour l'aide à la prise de décision (Khoo et al. (2011)) : soit pour faciliter la compréhension des termes présents dans les documents et les rapports médicaux, soit pour permettre le raisonnement et la recherche d'informations notamment, dans le processus de diagnostic, lors qu'il s'agit d'identifier les maladies associées à un symptôme donné ou alors les symptômes caractéristiques d'une maladie précise. Elles ont aussi été utilisées pour alerter les médecins sur les effets des substances chimiques sur le traitement de certaines maladies.

Ainsi, le processus de diagnostic étant basé sur le raisonnement autour des maladies et de leurs signes caractéristiques, la difficulté actuelle, au regard des ontologies de maladies existantes, réside sur le fait que ces signes sont listés de façon peu exhaustive (Mohammed et al. (2012)) et peu formelle (Oberkampf et al. (2012)). Seules les symptômes les plus communs sont énoncés dans ces ontologies mais d'un patient à un autre il existe des différences. D'ailleurs, les signes cliniques qui prennent des valeurs chez un patient ne sont pas même pris en compte. Il y a toutefois des ontologies spécifiques à la conceptualisation des signes (Mohammed et al. (2012); Anbarasi et al. (2013)) mais ils ne sont pas associés à des maladies.

Nous ne visons pas ici la construction d'une ontologie à partir de ressources non formelles (Charlet et al. (2012)) mais notre objectif est de fédérer les forces de plusieurs ontologies existantes afin de disposer d'une ontologie suffisamment fournie en terme de maladies et d'associer à chacune d'elles l'ensemble de ses signes pertinents et apparaissant dans la plus part des patients qui ont été affectés par les mêmes maladies. Cette association a déjà fait l'objet de travaux de recherches. En effet, Oberkampf et al. (2012), proposent dans leur projet d'ontologie Disy, de laisser la latitude aux médecins de remplir pour chaque maladie tous ces signes. Les travaux de Mohammed et al. (2012) quant à eux proposent une intégration d'ontologies afin de regrouper pour chaque maladie l'ensemble de ses signes présents dans ces ressources cibles. De notre côté, notre proposition est similaire à celle de ces derniers auteurs dans la mesure où nous cherchons aussi une fédération d'ontologies de maladies humaines et de signes afin de constituer un nouvelle qui soit adaptée pour le diagnostic médical. Cependant, malgré cette volonté de fédération, les ontologies actuelles ne sont toujours pas de taille pour décrire dans les moindres détails les maladies avec l'ensemble de leurs signes caractéristiques. Pour pallier à cela, nous tentons de nous intéresser à l'analyse de cas réels de patients ayant déjà été diagnostiqués et dont les médecins ont transcrits tout le processus dans des rapports textuels. Cette analyse permet alors de répertorier des signes nouveaux, jusqu'ici non encore pris en compte.

3 Ontologies cibles à intégrer

La constitution d'une ontologie de maladies et de signes consiste à une fédération d'un ensemble d'ontologies autour d'une structure unifiant toutes les maladies humaines ainsi que leurs signes caractéristiques. Les maladies correspondent aux diagnostics possibles. Les signes sont ceux susceptibles d'être identifiés sur un patient afin de conclure sur un diagnostic précis qui lui peut renvoyer à une ou plusieurs maladies.

Les maladies sont organisées de façon hiérarchisée ; elles et leurs formes dérivées sont regroupées par catégories, qui peuvent elles-mêmes être composées de sous-catégories de maladies. Les maladies sont lexicalisées afin d'avoir pour chaque maladie l'ensemble des termes nominatifs les plus connus et leurs synonymes. Pour chaque maladie, il sera important de conserver les définitions afin de contrôler la sémantique la mieux partagée. La plupart des signes connus de chaque maladie sont listés formellement à partir de ceux disponibles dans les ontologies médicales cibles.

Nous analysons ici des ontologies médicales mises à la disposition du public via la plateforme BioPortal. Notre choix s'est porté sur la DOID [1], la MESH [2], la SNOMED [3] comme ontologies de maladies, ainsi que la SYMP [4], et la CSSO [5] comme ontologies de signes.

L'ontologie DOID (Disease Ontology) nous sert d'ontologie de référence. Elle propose une hiérarchie de *10389* maladies humaines et de catégories de maladies. Avec la figure 2, nous pouvons voir que chaque maladie a un identifiant unique (*rdf :about*), et est classées dans une ou plusieurs catégories (*rdfs :subClassOf*). La maladie de l'*Hépatite A* appartient à la catégorie "*DOID_37*" des maladies de la peau ("*skin disease*") et à la catégorie "*DOID_934*" des maladies infectieuses virales ("*viral infectious disease*"). Toutefois, d'un identifiant à un autre, il n'y a aucune description permettant de dire qu'un identifiant donné renvoie à une maladie ou à une catégorie de maladies. Mais, en considérant le graphe hiérarchique, tous les concepts feuilles correspondent aux maladies et ceux qui ont des fils constituent des catégories.

Chaque maladie dans DOID fait référence (*oboInOwl :hasDbXref*) à la même maladie dans d'autres bases ontologiques comme celle de la ressource terminologique MESH (Medical Subject Headings). Elle est l'un des thésaurus de référence dans le domaine biomédical. Elle est connue pour les multitudes de termes synonymes proposés comme dénominations d'une maladie. Les termes sont en anglais et sont produits par la NLM [6], mais la traduction dans d'autres langues est aussi assurée dans plusieurs pays, notamment en français avec les travaux de l'Inserm [7]. Chacune des maladies dispose d'un terme préférentiel (*prefLabel : hepatitis A*) qui est la dénomination la plus utilisée, mais aussi de plusieurs termes synonymes (*altLabel : Viral hepatitis A, Viral hepatitis type A, Hepatitis Infectious, Hepatitides Infectious, Infectious Hepatitis, Infectious Hepatitides*). Ces termes correspondent aux différentes appélations de l'Hépatite A dans le monde.

1. http ://purl.bioontology.org/ontology/DOID
2. https ://www.nlm.nih.gov/mesh/
3. http ://purl.bioontology.org/ontology/SNOMEDCT
4. http ://purl.bioontology.org/ontology/SYMP
5. http ://purl.bioontology.org/ontology/CSSO
6. U.S. National Library of Medicine
7. http ://mesh.inserm.fr/mesh/

```
<owl:Class rdf:about="http://purl.obolibrary.org/obo/DOID_12549">
    <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">hepatitis A</rdfs:label>
    <rdfs:subClassOf rdf:resource="http://purl.obolibrary.org/obo/DOID_37"/>
    <rdfs:subClassOf rdf:resource="http://purl.obolibrary.org/obo/DOID_934"/>
    <obo:IAO_0000115 rdf:datatype="http://www.w3.org/2001/XMLSchema#string">A viral infectious disease that
    results_in inflammation located_in liver, has_material_basis_in Hepatitis A virus,
    which is transmitted_by ingestion of contaminated food or water,
    or transmitted_by direct contact with an infected person.
    The infection has_symptom fever, has_symptom fatigue, has_symptom loss of appetite, has_symptom nausea,
    has_symptom vomiting, has_symptom abdominal pain, has_symptom clay-colored bowel movements,
    has_symptom joint pain, and has_symptom jaundice.</obo:IAO_0000115>
    <oboInOwl:hasAlternativeId rdf:datatype="http://www.w3.org/2001/XMLSchema#string">DOID:12547</oboInOwl:hasAlternativeId>
    <oboInOwl:id rdf:datatype="http://www.w3.org/2001/XMLSchema#string">DOID:12549</oboInOwl:id>
    <oboInOwl:hasDbXref rdf:datatype="http://www.w3.org/2001/XMLSchema#string">MESH:D006506</oboInOwl:hasDbXref>
    <oboInOwl:hasDbXref rdf:datatype="http://www.w3.org/2001/XMLSchema#string">NCI:C3096</oboInOwl:hasDbXref>
    <oboInOwl:hasDbXref rdf:datatype="http://www.w3.org/2001/XMLSchema#string">SNOMEDCT_US_2016_03_01:154347003</oboInOwl:hasDbXref>
    <oboInOwl:hasDbXref rdf:datatype="http://www.w3.org/2001/XMLSchema#string">SNOMEDCT_US_2016_03_01:40468003</oboInOwl:hasDbXref>
    <oboInOwl:hasDbXref rdf:datatype="http://www.w3.org/2001/XMLSchema#string">UMLS_CUI:C0019159</oboInOwl:hasDbXref>
    <oboInOwl:hasRelatedSynonym rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Viral hepatitis A</oboInOwl:hasRelatedSynonym>
    <oboInOwl:hasExactSynonym rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Viral hepatitis, type A (disorder)</oboInOwl:hasExactSynonym>
    <oboInOwl:hasOBONamespace rdf:datatype="http://www.w3.org/2001/XMLSchema#string">disease_ontology</oboInOwl:hasOBONamespace>
</owl:Class>
```

FIG. 2 – *Descriptif de la maladie de l'Hépatite A dans DOID*

MESH est aussi connue pour son rôle de dictionnaire médical de référence à travers ces définitions explicites en langue humaine. La DOID quant à elle propose une définition (*obo :IAO_*) dans un langage sémi-formalisé qui va un peu plus loin dans la description de la maladie. Il est d'ailleurs aisé de décomposer cette description à partir des groupes de mots verbaux tels que *results_in, located_in, caused_by (ou has_material_basis_in), transmitted_by ou has_symptom* qui renvoient aux signes caractéristiques d'une maladie, et correspondant respectivement à la manifestation de la maladie, à sa localisation dans l'anatomie humaine, à l'agent à l'origine de la maladie, à ses modes de transmissions, et à ses symptômes. Cette liste de caractéristiques est très variante d'une maladie à une autre dans la DOID, elles ne sont pas toujours toutes prises en compte.

Pour donc pallier ce manque d'informations, nous utilisons SNOMED (aussi référenciée avec *oboInOwl :hasDbXref*) qui est une des ontologies les plus abouties dans le domaine médical. Elle a été conçue à partir du méta-thésaurus UMLS (Unified Medical Language System). Elle est très référenciée d'autant qu'elle couvre tous les champs de la médecine. SNOMED propose une catégorisation des différentes caractéristiques d'une maladie. Elle propose un panorama riche et variée de catégories de termes comme décrit dans le tableau 1 suivant.

L'ensemble de ces caractéristiques existe actuellement sous forme de liste de termes dans SNOMED. Elles sont, sauf naturellement les catégories "Organismes vivants" et "Eléments topographiques", des signes qui peuvent être descellés chez un patient malade. Mais pour un médecin la priorité réside dans l'observation des symptômes ("Eléments morphologiques") et l'identification et la mesure des signes cliniques ("Fonctions biologiques"). Toutes les autres catégories d'informations sont complémentaires pour faciliter la prise de décision.

C'est dans cette optique que nous avons considéré les ontologies SYPM et CSSO. La première est développée dans le même projet que la DOID. De la même façon que celle-ci pour

Catégorie	Description
Agents physiques	Antécédents du patient, en terme d'organismes artificiels présents dans son corps (*prothèse, implant*)
Organismes vivants	Virus, de la bactérie, des microbes, etc. à l'origine de la maladie (*bacille, absence de micro-organismes*)
Eléments morphologiques	Symptômes présents sur le patient malade (*lésion, gonflement, inflammation, infection*)
Fonctions biologiques	Signes cliniques présents sur le patient malade (*odeur d'urine, température cutanée*)
Composées chimiques	Produits chimiques pouvant être exposés au patient malade (*eau de Javel , drogue illicite*)
Conditions sociales	Caractère social du patient malade (*résident tropical, refus de nourriture*)
Eléments topographiques	Localisation de la maladie à un endroit précis du corps (*tissu osseux, épiderme, oreille interne*)
Procédures médicales	Antécédents du patient en terme d'examens approfondis effectués pour les besoins d'un diagnostic (*interventions chirurgicales, thérapie, orthopédie, analyse laboratoire*)

TAB. 1 – *Tableau des caractéristiques des maladies dans SNOMED*

les maladies, SYMP propose une structure hiérarchique complète de tous les signes cliniques et symptômes, qui sont aussi classées dans des catégories de signes. SYMP appose à chaque signe une définition renvoyant à la façon dont celui-ci se manifeste chez le patient. La seconde brandit aussi le même objectif que la SYMP mais elle est un peu moins aboutie. Seul le tiers des signes de SYMP sont prises en compte dans CSSO. Toutefois, cette dernière apporte un plus, une terminologie pour chaque signe. Par exemple, pour le signe *Fatigue* de l'Hépatite A (Figure 2), nous avons les termes synonymes suivants : *Lassitude, Tiredness, Weariness*. Cependant, aucune de ces deux ontologies ne fait la différence entre un signe clinique et un symptôme, il faudra alors faire la correspondance (*mapping*) avec la catégorisation des signes de la SNOMED.

4 Méthodologie d'intégration

4.1 Structure de données

Les différents formats de données des ontologies que nous avons sélectionnées sont implémentés avec les standards W3C du Web sémantique autour des langages RDF, RDFs et OWL. Donc pour faciliter la récupération des données ciblées sur chacune de ces ressources, nous proposons une structure (Figure 3) utilisant les mêmes technologies et qui hérite de celles-ci les mêmes formalismes conceptuels.

La structure est centrée sur la maladie (*Disease Class*) à laquelle sont associées toutes les classes d'informations nécessaires à la compréhension de la maladie ainsi qu'à la recommen-

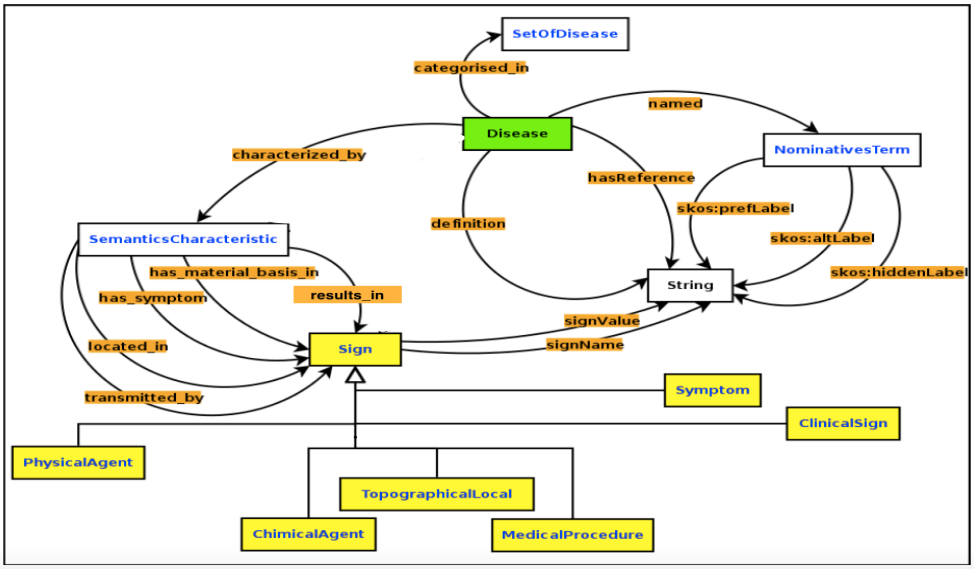

FIG. 3 – *Structure de données ontologiques combinant maladies et signes*

dation de diagnostics potentiels. Chaque maladie est identifiée (*categorized_in*) dans une catégorie ou plusieurs catégories (*SetOfDiseases Class*). Chaque maladie est associée (*named*) à ensemble de termes nominatifs (*NominativeTerms*) synonymes, du terme préféré (skos :prefLabel), aux termes alternatifs (skos :altLabel, skos :hiddenLabel). Chaque maladie regroupe (characterized_by) un ensemble de caractéristiques sémantiques (SemanticCharacteristics Class) et à travers les relation *has_symptom, transmitted_by, located_in, caused_by, results_in* renvoient respectivement aux différents types de signes (Sign Class) tels que Symptom ou ClinicalSign (symptômes ou signes cliniques), PhysicalAgent (c-a-d transmise par action d'un organisme vivant), TopographicalLocate (éléments topographiques), PhysicalAgent (organismes vivants), ChimicalAgent (composées chimiques) ou Symptom (éléments morphologiques) ou Medical-Procedure (Procédure médicale).

Chaque signe a un nom et éventuellement une valeur, surtout dans le cas des signes cliniques mesurables. Chacune des classes de signes, identifiables dans SNOMED (Tableau 1), regroupent et répertorient tous les signes possibles mais une maladie donnée n'est associée qu'aux signes les plus communs, les autres signes sont rattachés sur un cas de patient, et varie d'un cas à un autre. D'ailleurs dans la structure de données d'ensemble (figure 4) pour le moteur de recherche de diagnostics, nous pouvons voir que le patient est matérialisé par le descriptif textuel de son état de santé (*SourceTextForPatientState*), et il est associé à un cas de diagnostic médical (*MedicalDiagnosisCase*). Ce dernier étant relié (*associatedDisease*) à une maladie sur la base d'un ensemble de signes (*hasSign*).

4.2 Extraction de données à partir des ontologies cibles

Notre ontologie (Figure 3) est chargée par interrogation des différentes ressources ontologiques cibles avec le langage de requête SPARQL. Ces dernières sont directement exécutées

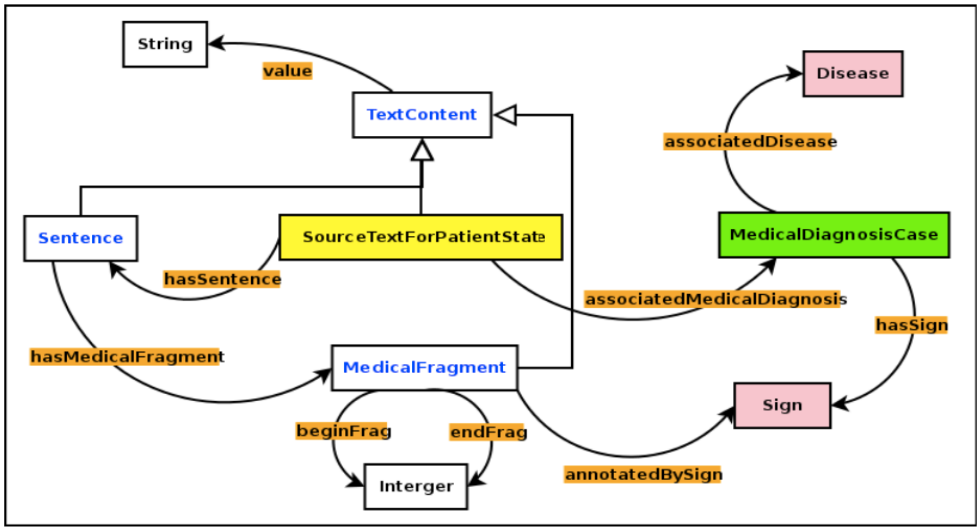

FIG. 4 – *Structure de données d'ensemble pour le moteur de recommandation de diagnostics*

sur des SPARQL EndPoint, des interfaces de requêtes ouvertes pour la navigation dans des graphes RDF. Nous utilisons ici celui du BioPortal [8]. Au total nous avons cinq (5) patrons de requêtes SPARQL qui permettent de récupérer :

- toutes les maladies (Figure 5) qui constituent les feuilles des classes à partir de la DOID, ainsi que leurs définitions à partir de MESH ;
- toutes les catégories de maladies (Figure 6) à partir de la DOID où nous sélectionnons leur nom, leur description, et leurs catégories mères ;
- tous les termes nominatifs synonymes des maladies (Figure 7) à partir de la DOID, mais surtout à partir de MESH, soient le label préférentiel, ainsi que les labels alternatifs pour chaque maladie ;
- tous les signes caractéristiques de base (Figure 8) pour chaque maladie à partir des descriptions sémi-formalisées de la DOID ;
- tous les termes nominatifs synonymes des signes : les labels préférentiels sont extraits de SYMP, les labels alternatifs sont quant à eux extraits des ontologies CSSO, et SNOMED.

Ainsi, dans le tableau 2, nous présentons le nombre de maladies et de catégories extraites de la DOID. Le nombre de termes nominatifs de maladies sont ceux obtenus pour le moment de la DOID, le travail d'extraction continue afin de les compléter avec les termes disponibles dans MESH. Enfin, les signes ainsi que leurs termes nominatifs sont tous tirés de la SYMP, ils sont à compléter avec la SNOMED et l'évolution de l'ontologie CSSO.

8. http ://sparql.bioontology.org/

Association des maladies humaines à leurs plus pertinents signes caractéristiques

Types Eléments	**Objet ontologique**	**Ontologies d'origines**	**Nbr. d'individus**
Diagnostics			
Maladies	Disease Class	DOID	6442
Catégories	SetOfDiseases Class	DOID	3947
Termes Synonymes	AnnotationProperty (prefLabel, altLabel, hiddenLabel)	DOID, MESH	27586
Signes			
Symptômes et Signes Cliniques	Symptom Class et CinicalSign Class - subClassOf Sign Class	SYMP	942
Autres signes	PhysicalAgent Class, ChimicalAgent Class, TopographicalLocate Class, MedicalProcedure Class : subClassOf Sign Class	DOID, SNOMED	6020
Termes Synonymes	AnnotationProperty (prefLabel, altLabel)	CSSO, SNOMED	1346

TAB. 2 – *Description de l'état actuel de notre ontologie de maladies et de signes*

```
<Disease rdf:about="&ourOnto;Hepatitis_A">
   <categorized_in rdf:resource="&ourOnto;SET_37"/>
   <categorized_in rdf:resource="&ourOnto;SET_934"/>
   <definition xml:lang="en">
        Inflammation of the liver in humans caused by a
        member of the hepatovirus genus, human hepatitis A virus.
        It can be transmitted through fecal contamination of food or water.
   </definition>
   <named rdf:resource="&ourOnto;Terms_12549"></named>
   <characterized_by rdf:resource="&ourOnto;SC_12549"></characterized_by>
   <hasReference>http://purl.obolibrary.org/obo/DOID_12549</hasReference>
   <hasReference>http://purl.bioontology.org/ontology/MESH/D006506</hasReference>
</Disease>
```

FIG. 5 – *Récupération des maladies*

```
<SetOfDiseases rdf:about="&ourOnto;SET_37">
   <rdfs:label xml:lang="en">skin disease</rdfs:label>
   <rdfs:subClassOf rdf:resource="&ourOnto;SET_16"/>
   <description xml:lang="en">An integumentary system disease that is located in the skin.</description>
   <hasReference>http://purl.obolibrary.org/obo/DOID_12549</hasReference>
</SetOfDiseases>
```

FIG. 6 – *Récupération des catégories de maladies*

4.3 Extraction de données à partir de cas réels

Après avoir constitué notre fédération d'ontologies, nous allons découvrir jusqu'à quelle mesure nous pouvons enrichir cette ontologie à partir de l'analyse de rapports médicaux de cas de diagnostics ayant déjà été validés par des médecins. A l'image, de nos exemples précédents,

```
<NominativesTerm rdf:about="&ourOnto;Terms_12549">
   <skos:prefLabel xml:lang="en">Hepatitis A</skos:prefLabel>
   <skos:altLabel xml:lang="en">Viral hepatitis A</skos:altLabel>
   <skos:altLabel xml:lang="en">Viral hepatitis type A</skos:altLabel>
   <skos:altLabel xml:lang="en">Hepatitis Infectious</skos:altLabel>
   <skos:altLabel xml:lang="en">Hepatitides Infectious</skos:altLabel>
   <skos:altLabel xml:lang="en">Infectious Hepatitis</skos:altLabel>
   <skos:altLabel xml:lang="en">Infectious Hepatitides</skos:altLabel>
   <hasReference>http://purl.obolibrary.org/obo/DOID_12549</hasReference>
   <hasReference>http://purl.bioontology.org/ontology/MESH/D006506</hasReference>
</NominativesTerm>
```

FIG. 7 – *Récupération des termes nominatifs de maladies*

```
<SemanticsCharacteric rdf:about="&ourOnto;SC_12549">
   <results_in rdf:resource="&ourOnto;Symp_inflammation"/><!--Symptôme le plus général-->
   <located_in rdf:resource="&ourOnto;TLSign_liver"/>
   <caused_by rdf:resource="&ourOnto;PASign_Hepatitis-A-virus"/>
   <transmitted_by>direct contact with an infected person</transmitted_by>
   <transmitted_by>ingestion of contaminated food or water</transmitted_by>
   <has_symptom rdf:resource="&ourOnto;Symp_fever"/>
   <has_symptom rdf:resource="&ourOnto;Symp_fatigue"/>
   <has_symptom rdf:resource="&ourOnto;Symp_loss-of-appetite"/>
   <has_symptom rdf:resource="&ourOnto;Symp_nausea"/>
   <has_symptom rdf:resource="&ourOnto;Symp_vomiting"/>
   <has_symptom rdf:resource="&ourOnto;Symp_abdominal-pain"/>
   <has_symptom rdf:resource="&ourOnto;Symp_clay-colored-bowel-movements"/>
   <has_symptom rdf:resource="&ourOnto;Symp_joint-pain"/>
   <has_symptom rdf:resource="&ourOnto;Symp_jaundice"/>
   <hasReference>http://purl.obolibrary.org/obo/DOID_12549</hasReference>
</SemanticsCharacteric>
```

FIG. 8 – *Récupération des signes caractéristiques de maladies*

```
<SYMPTOM rdf:about="&ourOnto;Symp_fatigue">
   <skos:prefLabel xml:lang="en">Fatigue</skos:prefLabel>
   <skos:altLabel xml:lang="en">Lassitude</skos:altLabel>
   <skos:altLabel xml:lang="en">Tiredness</skos:altLabel>
   <skos:altLabel xml:lang="en">Weariness</skos:altLabel>
   <hasReference>http://purl.jp/bio/11/csso/CSSO_000119</hasReference>
   <hasReference>http://purl.obolibrary.org/obo/SYMP_0019177</hasReference>
   <hasReference>http://purl.bioontology.org/ontology/SNOMEDCT/84229001</hasReference>
</SYMPTOM>
```

FIG. 9 – *Récupération des termes nominatifs des signes de maladies*

nous considérons ici un cas de patient diagnostiqué Hépatite A [9]. Cette maladie est d'origine virale désignant une inflammation du foie. Elle est répertoriée parmi les infections sexuellement transmissibles et est avec le Sida, dans le top dix (10) des maladies les plus dangereuses

9. Exemple traduit en français à partir du site http ://www.immunologyclinic.com/

au Sénégal [10].

Le cas, pris en exemple (Figure 10), transcrit dans un résumé textuel le descriptif symptomatique de l'état d'un patient dont le diagnostic est confirmé à la suite d'un ensemble d'examens approfondis. Ce descriptif recoupe tous les signes caractéristiques permettant de conclure sur la maladie de l'Hépatite A, et son analyse soulève un certain nombre de questionnements pratiques pouvant permettre d'améliorer et d'étendre notre ontologie :

Typologies des signes présents. Pour la maladie Hépatite A, seuls 9 symptômes généraux figurent dans l'ontologie. Il s'agit d'une fièvre (fever), d'une fatigue (fatigue), d'une perte d'appétit (loss of appetite), d'une nausée (nausea), d'un vomissement (vomiting), d'une douleur abdominale (abdominal pain), d'une présence de selles de couleur argile (clay colored bowel movements), d'une douleur articulaire (joint pain), et d'une jaunisse (jaundice). Seuls 7/9 sont identifiables dans ce présent cas et représentent les premiers signes observables chez le patient malade. Les autres signes, bien que existant dans l'ontologie et non associés à des maladies, correspondent aux signes généraux indiquant son sexe, son âge, et ses excès, mais aussi aux antécédents et aux signes cliniques. Ces derniers découlent d'examens approfondis et renvoient à des valeurs. Au final plus de 16 signes supplémentaires viennent s'ajouter à ceux qui décrivent l'Hépatite A dans l'ontologie : ce qui prouve que l'analyse des cas réels est primordiale à la confirmation d'un diagnostic.

Extraction des signes à partir du descriptif textuel. L'analyse du cas pose deux problèmes : l'identification des symptômes généraux connus de l'ontologie pour l'Hépatite A et l'extraction des signes, de type clé/valeur, spécifiques au patient et non répertoriés pour cette même maladie. Dans ces deux tâches, est posé un problème de traitement automatique de la langue (TAL) d'autant que ce type de descriptif est transcrit en langue naturelle. Si les termes renvoyant aux symptômes ne sont pas difficiles à détecter au regard du lexique détaillé (termes préférentiels et synonymes) apposé à chaque signe. Pour les autres signes, en plus d'être identifiés dans le texte, il est impératif d'extraire leurs valeurs. Elles renvoient à des entités nommées précises et l'affectation d'un de ces signes à une valeur devra aussi passer par l'identification de la relation (le plus souvent verbale) qui les associe dans le texte. Il est alors nécessaire d'utiliser des outils TAL pour identifier tous les fragments de texte décrivant un signe ou une valeur de signe. Cette partie n'est pas détaillée dans cet article mais fait l'objet de nos travaux de recherche en cours.

Prise en compte des nouveaux signes dans l'ontologie. Le modèle de données décrit plus haut (Figures 3 et 4) montre que notre ontologie, bien que répertoriant tous les signes susceptibles d'être présents chez un patient à partir des ontologies de signes cibles, n'associe que les signes les plus communs à une maladie donnée. Donc les signes spécifiques décrits dans le contenu d'un cas sont aussi stockés dans l'ontologie mais leurs valeurs ne peuvent être consignées que dans le cas de type *"MedicalDiagnosisCase"*, qui lui est associé à l'ensemble des signes présents dans son contenu ainsi que leurs valeurs, et à la maladies (ou aux maladies) à laquelle il correspond. Par conséquent, une maladie sera toujours en même temps reliée à l'ensemble de ces symptômes communs via l'ontologie de maladies et de signes généraux, et à un ensemble de signes spécifiques suivant le nombres de cas déjà diagnostiqués.

Apport de ce descriptif dans le processus de diagnostic. Nous pouvons enfin remarquer dans l'exemple que les signes associés à l'Hépatite A dans l'ontologie n'apparaissent pas

10. http ://www.who.int/countries/sen/fr/

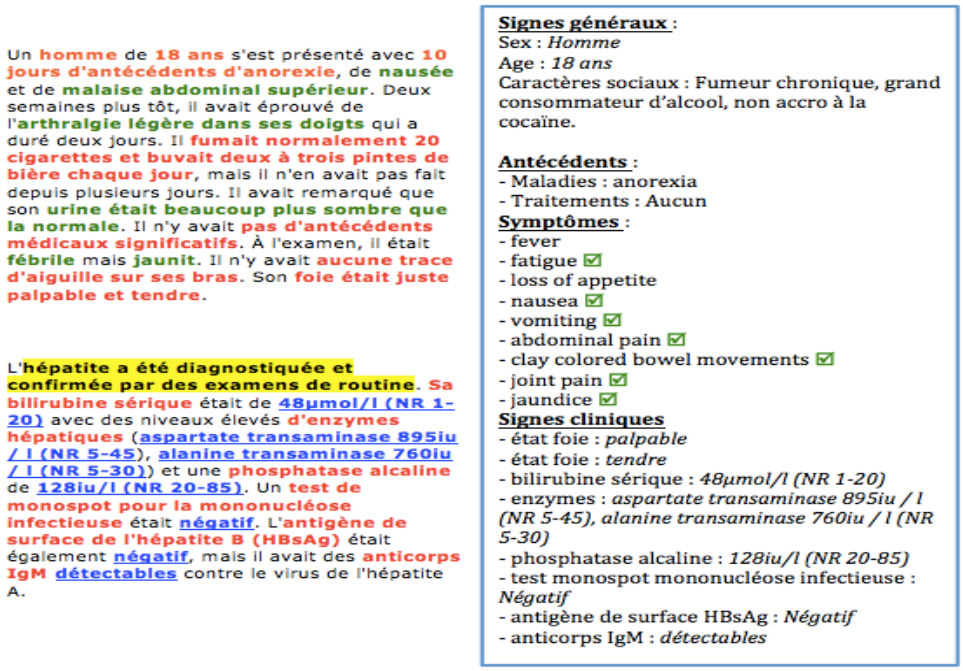

FIG. 10 – *Cas réél d'un patient diagnostiqué Hépatite A*

tous et leur présence variera certainement d'un patient à un autre, de même que les signes spécifiques. Un premier apport serait alors de pouvoir calculer, considérant tous les cas enregistrés, la fréquence d'apparition de chaque signe pour l'ensemble des patients diagnostiqués une même maladie. Le classement de ces fréquences nous amène à un second clair apport qui consistera à identifier les signes spécifiques nécessaires, l'un après l'autre, à la confirmation d'un diagnostic donnée.

Quelques chiffres. Le tableau 3 prouve ce que nous venons de dire, et nous pouvons voir les symptômes et les signes cliniques trouvés sur une dizaine de cas réels de patients ayant déjà été diagnostiqués. Les exemples choisis portent sur des maladies tropicales [11] que nous trouvons au Sénégal. Nous pouvons alors remarquer que pour chaque maladie, il y a un nombre précis de symptômes généraux indiqués par l'ontologie de maladie mais la totalité d'entre eux ne sont pas présents chez les patients. Par ailleurs, de nouveaux symptômes non répertoriés dans l'ontologie font leur apparition ainsi que les signes cliniques spécifiques à chaque patient.

5 Conclusion

Dans cet article, la problématique a porté sur la mise en place d'un système d'aide au diagnostic médical à partir des ressources ontologies ouvertes et partagées. Il est question ici

11. Exemples tirés de http ://medecinetropicale.free.fr/

Association des maladies humaines à leurs plus pertinents signes caractéristiques

Maladie	**Symptômes indiqués par l'ontologie**	**Symptômes de l'ontologie présents sur le cas**	**Symptômes nouveaux**	**Signes cliniques**
Hepatitis A	9	7	7	9
Cholera	5	3	10	2
Rougeaole	6	4	11	3
Dengue	10	5	11	20
Tétanos	4	3	12	4
Paludisme	6	4	8	24
Syphilis	5	2	12	14
Chikungunya	9	5	7	29
FièvreTyphoide	8	5	8	3
Meningite	9	4	5	7

TAB. 3 – *Nombre de Simptômes et de Signes cliniques trouvés sur des cas réels*

de constitution d'une ontologie centrale fédérant un ensemble d'ontologies et terminologies médicales cibles, qui répondent au besoin en informations afin de faciliter la tâche du médecin dans l'identification des diagnostics potentiels, parmi lesquels il aura la latitude de choisir ou de valider le plus fiable en connaissance de cause. Ce type de système ne se substitue donc nullement au médecin.

Nous avons donc proposé une méthodologie d'intégration autour d'une structure de graphe RDF facilitant la récupération des maladies humaines et de leurs plus pertinents signes caractéristiques, à partir d'une analyse de cas réels de diagnostics confirmés par des médecins. Au final, nous disposons d'une ontologie de maladies et de signes qui devrait servir de base de connaissances dans le moteur de recherche que nous visons. Le travail en perspective serait de faire valider cette ontologie par les acteurs du domaine mais cela ne se fera que pour évaluer sa pertinence et sa consistance dans son rôle pour le moteur, qui est d'identifier et d'apposer une sémantique aux signes présents chez un patient, et de trouver des maladies pertinentes comme diagnostics.

Références

Anbarasi, M., P. Naveen, S. Selvaganapathi, et M. Nowsath (2013). Ontology based medical diagnosis decision support system. In *Actes de International Journal of Engineering Research and Technology (Traitement automatique des langues naturelles)*.

Balogh, E. P., B. T. Miller, et J. R. Ball (2015). Improving diagnosis in health care. In *Actes de National Academies of Sciences, Engineering, and Medicine. The National Academies Press, Washington, DC (Traitement automatique des langues naturelles)*.

Bringay, S., C. Barry, et J. Charlet (2005). Les documents et les annotations du dossier patient hospitalier. In *Actes de Information-Interaction-Intelligence, Volume 4, No. 1, 2005 (Traitement automatique des langues naturelles)*.

Charlet, J., G. Declerck, F. Dhombres, P. Gayet, p. Miroux, et P. Vandenbussche (2012). Construire une ontologie médicale pour la recherche d'information : problématiques terminologiques et de modélisation.

Cox, A. P., P. L. Ray, M. Jensen, et A. D. Diehl (2014). Defining 'sign' and 'symptom'. In *Actes de IWOOD Workshop, In ICBO, Houston, TX, USA, October 6-7, 2014 (Traitement automatique des langues naturelles)*, pp. 101–110.

Hoehndorf, R., P. Schofield, et G. Gkoutos (2015). The role of ontologies in biological and biomedical research : a functional perspective. In *Actes de Briefings in Bioinformatics Journal, 2015 (Traitement automatique des langues naturelles)*.

Khoo, C. S. G., J. cheon Na, V. W. Wang, et S. Chan (2011). Developing an ontology for encoding disease treatment information in medical abstracts. In *Actes de DESIDOC Journal of Library and Information Technology 31 (2) (Traitement automatique des langues naturelles)*, pp. 103–115.

Mohammed, O., R. Benlamri, et S. Fong (2012). Building a diseases symptoms ontology for medical diagnosis : An integrative approach. In *Actes de IEEE International Conference on Future Generation Commnication Technology (FGCT 2012), Décembre 2012 (Traitement automatique des langues naturelles)*.

Oberkampf, H., S. Zillner, et B. Bauer (2012). Interpreting patient data using medical background knowledge. In *Actes de 3rd International Conference on Biomedical Ontology (ICB0), Autria July 21-25, 2012 (Traitement automatique des langues naturelles)*.

S-Ortiz, J. A. R., A. L. Jimenez, J. Cater, et C. A. Malendés (2013). Ontology-based knowledge representation for supporting medical decisions. In *Actes de Recherche in Computer Science, 2013 (Traitement automatique des langues naturelles)*.

Wang, H.-T. et A. U. Tansel (2013). Composite ontology-based medical diagnosis decision support system framework. In *Actes de Communications of the IIMA : Vol. 13 (Traitement automatique des langues naturelles)*.

Summary

Healing a sick patient requires a medical diagnosis before proposing appropriate treatment. With the explosion of medical knowledges, we are interested in their exploitation to help clinicianc in collecting informations during diagnostic process. This article focuses on the development of a data model targeting knowledges available in both formal and non-formal resources. Our goal is to merge the strengths of all these resources to provide access to a variety of shared knowledges facilitating the identification and association of human diseases and to all of their available relevant characteristic signs such as symptoms and symptoms. clinical signs.
On one side, we propose an ontology produced from an integration of several existing and open medical ontologies and terminologies. On another side, we exploit real cases of patients whose diagnosis has already been confirmed by clinicians. They are transcribed in textual reports in natural language, and we show here that their analysis improves the list of signs of each disease. This work then results in a knowledges base loaded from the known target ontologies on the bioportal platform such as DOID, MESH and SNOMED for disease selection, SYMP, and CSSO for all existing signs. The sample of selected textual cases concerns tropical diseases.