Critères de comparaison des plateformes d'ontologies biomédicales NCBO BioPortal et CISMeF

Khedidja Bouarech

Université Montpellier II
Faculté des Sciences

MASTER 2 INFORMATIQUE

Spécialité DECOL

RAPPORT DE STAGE

Critères de comparaison des plateformes d'ontologies biomédicales

NCBO BioPortal et CISMeF

effectué à

Le Laboratoire d'Informatique de Robotique
et de Microélectronique de Montpellier
(LIRMM)

du 11 Mars au 31 Aout 2013

Par

BOUARECH Khedidja

Encadrant

M. Clément JONQUET

Remerciments

Je tiens, en premier lieu, à énormément remercier mon encadrant Dr. Clément Jonquet pour son encadrement et ses conseils tout au long du stage et lors de la rédaction du rapport.

Je tiens également à remercier Suzanne Perreira de la société VIDAL pour sa précieuse collaboration.

Je remercie les membres de l'équipe CISMeF pour m'avoir fait découvrir leur thèmes de recherche.

Je remercie aussi mes tuteurs à l'Université Montpellier II Michel Leclère et Konstantin Todorov.

Table des matières

Chapitre 1

Introduction

1.1 Contexte

Du 01 Mars 2011 au 31 août 2011, j'ai effectué mon stage de Master 2 DECOL au sein du LIRMM [1], à Montpellier. Ce stage de Master 2 fut l'occasion de mettre en application les nombreuses connaissances enseigneées en Master DECOL. Ce stage s'inscrit dans le cadre du projet SIFR [2] financé par l'Agence Nationale de la Recherche (ANR) [3].

Dans ce chapitre nous allons présenter le projet dans lequel s'inscrit ce stage et les différents partenaires du stage. Nous détaillerons également les objectifs et les technologies utilisées.

1.2 Lirmm

Le Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM) est une unité mixte de recherche, dépendant conjointement de l'Université Montpellier II [4] et du Centre National de la Recherche Scientifique (CNRS) [5]. Le laboratoire est organisé en trois départements scientifiques (informatique, robotique et microélectronique) et comprend 19 équipes de recherche.

1. http ://www.lirmm.fr/
2. http ://www.agence-nationale-recherche.fr/projet-anr/ ?tx_lwmsuivibilan_pi2%5BCODE%5D=ANR-12-JS02-0010
3. http ://www.agence-nationale-recherche.fr/
4. http ://www.univ-montp2.fr/
5. http ://www.cnrs.fr/

Les activités de recherche du LIRMM concernent : la conception et la vérification de systèmes intégrés, mobiles, communicants, la modélisation de systèmes complexes à base d'agents, les études en algorithmique, bioinformatique, interactions homme-machine, robotique,etc.
Outre les domaines propre au laboratoire, les recherches menées par LIRMM trouvent des finalités dans d'autres domaines applicatifs comme la chimie, la biologie, la santé, l'environnement,etc.

1.3 Projet SIFR

Le projet 'Semantic Indexing of French Biomedical Data Resources' (SIFR) a pour objectif de résoudre les défis scientifiques et techniques soulevés pour exploiter les ontologies dans la construction de services d'indexation, de fouille, et de recherche de données pour les ressources biomédicales françaises. Le but est de construire un workflow d'indexation basé sur les ontologies (i.e., French Annotator) similaire à celui qui existe pour les ressources en anglais, mais spécialisé pour le Français. Ce sera le premier jalon de la création (dans de futurs projets) d'un index de données qui permettra la recherche et la fouille sémantique et multilingue.
Le projet SIFR rassemble plusieurs chercheurs du LIRMM pour réaliser cet objectif. Des partenaires de très grande qualité sont également associés au projet : Stanford BMIR [6], un leader mondial en outils et services (anglais) basés sur les ontologies pour aider la construction de systèmes à base de connaissances biomédicales et le groupe CISMeF [7], leader national en services de terminologies pour la santé en France.
En outre, d'autres partenaires académiques et industriels ont également été identifiés et collaboreront à la valorisation concrète des résultats du projet en termes d'impact scientifique et économique (e.g., Ontologos Corp, CNRS-INIST).
SIFR permettra l'implantation d'une nouvelle thématique de recherche au LIRMM et offrira à la communauté biomédicale (e.g., cliniciens, professionnels de santé, chercheurs) des services d'indexation hautement performants basés sur les ontologies leur permettant d'améliorer leur processus de production et de consommation de données.

6. http ://bmir.stanford.edu/
7. http ://www.chu-rouen.fr/cismef/

1.4 Objectifs

L'objectif de ce stage consiste à faire une comparaison exhaustive des fonctionnalités de deux portails d'ontologies/terminologies biomédicales : le NCBO Bioportal[20, 32] et le portail CISMeF (accompagnée de tests fonctionnels et d'évaluation) et s'intéresser à l'interopérabilité des services qu'ils fournissent. En particulier, nous ferons une comparaison des outils d'annotation sémantique que proposent les deux portails.

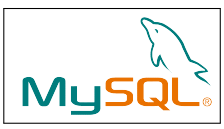
Une première partie de cet objectif était réalisée lors du TER Master 2 et qui consistait à effectuer une comparaison générale des différents services et outils proposés par les deux portails. Les détails de cette comparaison ainsi que les résultats sont présentés dans l'annexe de ce rapport.

La deuxième partie consiste à tester et évaluer les performances des outils d'annotation sémantique proposés par chaque portail en leur soumettant des corpus de données et en comparant les résultats avec un corpus de référence.

1.5 Technologies utilisées

Les différents traitements de données ainsi que les appels aux services d'annotation ont été réalisé avec le langage de programation JAVA sous l'IDE Eclipse [8] qui offre une possibilité d'ajouter de nombreux plugins.

L'agent HTTPClient est utilisé pour paramétrer la méthode d'appel au service d'annotation.



8. http ://www.eclipse.org/

Pour faciliter l'exploiatation des résultats obtenus des services d'annotation, nous les avons stockés dans une base de données MySQL.

1.6 Planning prévisionnel

Cette section présente le planning des tâches effectuées durant tout le déroulement de ce stage. La figure 1.1 montre le diagramme de Gantt qui décrit la division des tâches liées aux différentes étapes du stage. Description des couleurs :

Bleu clair Les phases de réalisation de la comparaison générale (TER)

Vert Les phases de traitement des données

Rouge Les phases d'annotation

Bleu foncé Les phases d'écriture

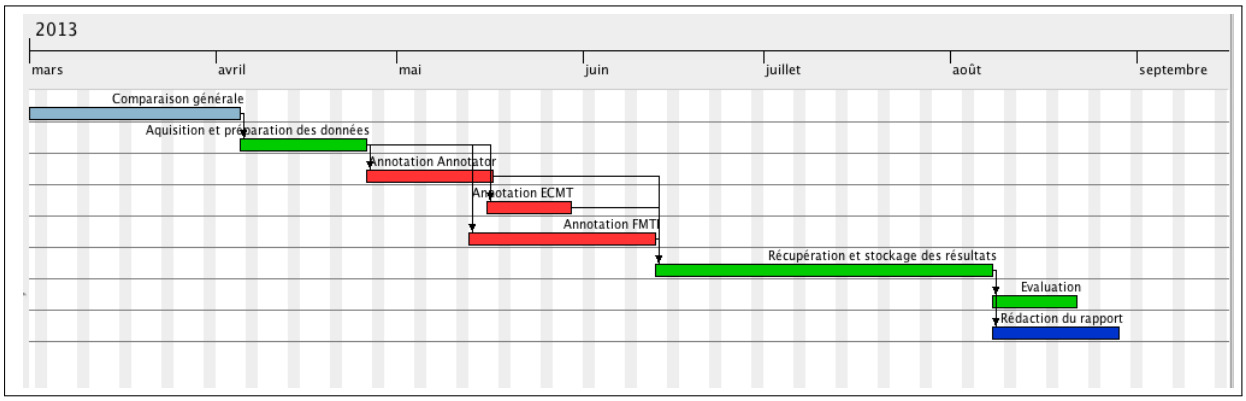


FIGURE 1.1 – Diagramme de Gantt pour le déroulement du stage

1.7 Plan du rapport

Nous commencerons par décrire dans le chapitre 2 les deux portails en détail en précisant pour chacun la philosophie adoptée, l'architecture et technologies utilisées et les différents services proposés. Dans le chapitre 3 nous détaillerons notre démarche de comparaison des workflow d'annotation sémantique pour les deux portails. Nous décrirons les jeux de données utilisés,le paramétrage et l'appel des services. Dans le chapitre 4 nous présenterons les résultats obtenus et leur évaluation. Pour chaque service, nous détaillerons le résultat d'évaluation et nous discuterons ce résultat. Puis nous concluons ce rapport par les perspectives dans le chapitre 5.

Chapitre 2

Présentation des portails

2.1 Introduction

Le volume des données dans le domaine biomédical est très grand et s'étend très vite. L'intégration et l'interopérabilité des données biomédicales est nécessaire pour permettre la recherche et l'interrogation de ces données ainsi de nouvelles découvertes scientifiques [19]. Ces données sont souvent mal structurées et sont disponibles dans plusieurs formats (bases de données, documents,etc.) empêchant des découvertes qui peuvent être faites en les fusionnant.

Pour faire face à ce problème, la communauté biomédicale s'est tournée vers les ontologies et les terminologies pour décrire ses données et les transformer en connaissances structurées et bien formées [6]. Les ontologies décrivent les connaissances d'un domaine à travers des concepts et des relations et les terminologies listent les termes d'un domaine, et pour chacun d'eux, proposent une fiche qui en décrit les usages, la (ou les) signification(s), ainsi que les relations entretenues avec des termes sémantiquement et/ou syntaxiquement proches [1] [12, 24].

L'équipe **Stanford BioMedical Informatics Research (BMIR)** [2] de l'Université de Stanford et l'équipe **Catalogue et Indexe des Sites Médicaux de langue Française(CISMeF)** [3] du Centre Hospitalier Universitaire de Rouen ont investi beaucoup d'efforts dans le développement d'outils et services à base d'ontologies pour assister

1. Avec la mise sur support informatique et la diversification des usages des terminologies dans les années 90, les terminologues se sont interrogés sur les notions de termes et de concepts, et sur leur articulation. Dans le reste du document nous utiliserons souvent le terme "ontologie" pour désigner ontologie et terminologie bien que nous soyons d'accord sur la différence entre ces deux structures détaillée dans [12, 24]
2. http ://bmir.stanford.edu
3. http ://www.cismef.org

les professionnels de santé dans leur recherche d'information disponible sur le Web et dans l'utilisation des ontologies. Les deux groupes ont développé des portails Web respectivement le NCBO BioPortal [20, 32] et le CISMeF Health Multi-Terminology Portal [14] qui offrent diverse services pour rechercher, indexer, explorer, annoter et visualiser les ontologies standards disponibles. En dépit de l'aspect spécifique de chacun, les deux portails partagent un objectif commun qui consiste à fournir à la communauté biomédicale une interface commune pour explorer l'information liée à la santé grâce aux connaissances représentées par des ontologies. Toutefois, les deux portails ne sont pas interopérables et sont parfois répétitifs. Une personne qui veut par exemple utiliser des bio-ontologies en anglais ou en français, doit faire face à un choix cornélien pour choisir un portail pour cette tâche. Plus loin encore, parfois, il faut utiliser les deux portails ensemble pour obtenir un résultat satisfaisant (utilisation des termes en anglais pour chercher des données francophones et vice-versa). Cette situation nécessite un effort supplémentaire et cause une perte de temps.

Afin de pouvoir dresser une étude comparative des portails NCBO et CISMeF et discuter leur possible convergence, nous allons d'abord présenter les deux portails et décrire leurs approches ainsi que les principaux services et les technologies sur lesquels ils reposent.

2.2 National Center for Biomedical Ontologies

Le Centre National pour les Ontologies Biomédicales (NCBO) [4] est un consortium d'informaticiens, biologistes, cliniciens et ontologistes qui développent des technologies pour permettre aux scientifiques de créer, gérer et partager les connaissances biomédicales en utilisant les ontologies comme fondement pour cette tâche[25]. La mission du centre comprend trois objectifs principaux[20] :

- Création et maintenance d'un entrepôt d'ontologies et de terminologies biomédicales.
- Développement d'outils et de services Web pour permettre l'utilisation des ontologies et terminologies de manière automatique.
- Mise en oeuvre de programmes d'éducation et de programmes de sensibilisation autour des ontologies biomédicales et des technologies du NCBO.

Créé en 2005, le Centre est organisé en six département (cores) dont le 1er (core1) est celui de la recherche en bioinformatique qui implique l'Université de Stanford, La Clinique Mayo, l'Universté de Victoria et l'Université de Buffalo. Le projet est financé par le National Institute of Health (NIH) [5] et fait partie du réseau des Centres nationaux

4. http ://www.bioontology.org/
5. http ://www.nih.gov

(américains) d'informatique biomédicale [6] [3].

2.2.1 BioPortal

BioPortal [7] est un entrepôt d'ontologies biomédicales qui héberge plus de 350 ontologies dans différents formats. Ces ontologies sont régulièrement mises à jour et accessibles via des navigateurs web et via des services web. En effet, BioPortal fournit une palette de services web comme la catégorisation, la recherche par terme, la visualisation graphique d'ontologies ou l'historique des versions d'une ontologie[32, 20, 31]. Nous détaillerons précisément ces fonctionnalités par la suite.

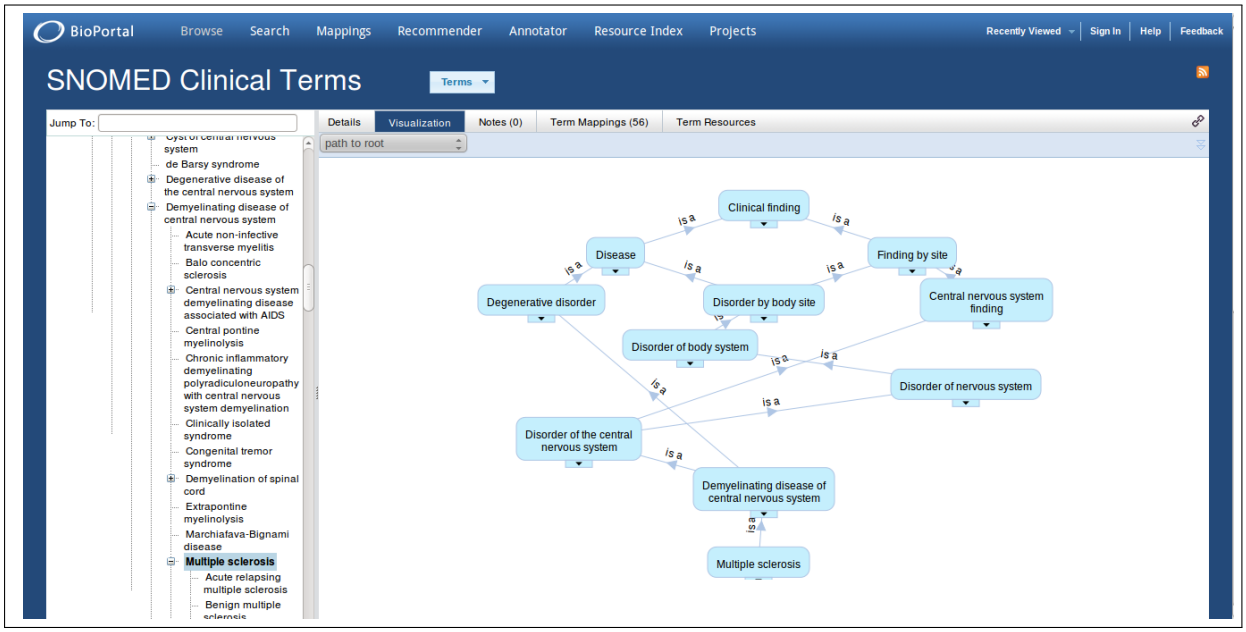


FIGURE 2.1 – Interface graphique de BioPortal, concept "Multiple sclerosis" de SNOMED-CT : à gauche un affichage en arborescence et à droite une visualisation graphique du concept

Architecture : BioPortal permet d'importer des ontologies dans différents formats tels que OWL, OBO [8], RDF ou bien Protégé Frame Language. BioPortal utilise LexGrid [9] pour stocker les ontologies en format OBO et RRF, et Protégé [10] pour les autres formats. BioPortal adopte une architecture multi-niveaux comme le montre la figure 2.1 [32]

6. http ://www.ncbcs.org
7. http ://bioportal.bioontology.org
8. http ://www.obofoundry.org
9. http ://informatics.mayo.edu/LexGrid
10. http ://protege.stanford.edu

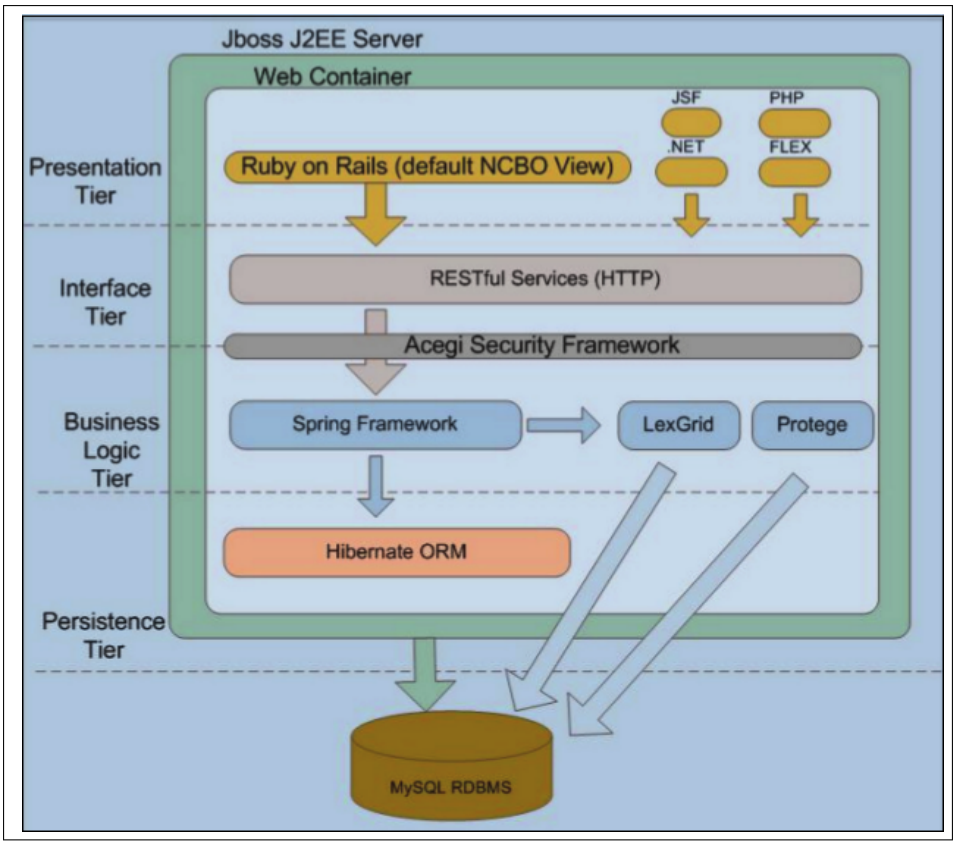


FIGURE 2.2 – Architecture multi-niveaux de BioPortal : La couche Présentation fournit l'interface utilisateur, la couche Interface contient les services web nécessaires pour les fonctionalités de BioPortal, la couche logique permet l'accès des API aux ontologies et au Resource Index [32]

A noter que cette architecture est entrain de changer pour passer au modèle RDF triple store avec 4stores[28], permettant l'accès aux ontologies du portail à travers des requêtes SPARQL. Une version Bêta est dors et déjà disponible [11].

Alignement

Les ontologies dans BioPortal comme dans la plupart des entrepôts d'ontologies se chevauchent en couverture [12]. Créer des alignements (mappings) parmi les ontologies en identifiant les concepts dont le sens est similaire est une étape critique dans l'intégration de données. BioPortal permet de créer manuellement des alignements pair à pair entre concepts d'ontologies différentes, soumettre au portail des alignements créés par ailleurs, chercher et télécharger des alignements existants sur le portail et les commenter. En plus de la contribution des utilisateurs, le NCBO utilise ses propres algorithmes lexicaux

11. http ://sparql.bioontology.org
12. 'overlap in coverage', dans le document original.

comme LOOM [13] pour créer les alignements [21, 10, 11].

Aspect communautaire

: BioPortal est une bibliothèque d'ontologies [7] communautaire. Les utilisateurs peuvent avoir accès au contenu du portail sans restriction, [14] et peuvent accéder aux opérations suivantes [22] :

- **Publication** : Les développeurs d'ontologies relevant du domaine biomédical peuvent les publier sur le portail sous condition de fournir les métadonnées essentielles (nom, auteur, licence, version,etc.) [29] à propos de leurs ontologies.
- **Commentaires** : Les utilisateurs peuvent laisser des notes discutant la modélisation d'une ontologie, pointant les problèmes liés aux définitions ou bien demandant un changement de la part de l'auteur. Ces commentaires sont inclus dans les métadonnées.
- **Évaluation** : BioPortal permet aux utilisateurs de participer à l'évaluation des ontologies en leur permettant de décrire leur projets utilisant des ontologies du portail. En présentant une revue des ontologies exploitées dans leurs projets, on peut distinguer quel type d'ontologies est plus adapté pour quel type de projets.
- **Création des alignements** : BioPortal permet à ses utilisateurs de créer, télécharger, exploiter, commenter et discuter des alignements. Malgré la performance des outils d'alignement automatique, BioPortal tient à impliquer la communauté d'experts dans cette tâche pour disposer d'un ensemble le plus riche possible d'alignements.

En outre, un utilisateur peut s'inscrire à un système de notifications RSS pour être informé de n'importe quelle contribution de la communauté sur une ontologie donnée.

2.2.2 Outils et services web

En plus de la plateforme d'hébergement d'ontologies biomédicales, le NCBO développe aussi des outils et services pour utiliser et exploiter ces ontologies et venir en aide aux chercheurs dans leur travail. Ces services sont disponibles à travers une interface de navigateur web[31] et sont accessibles via BioPortal REST API qui fournit les services d'accès, recherche, visualisation, etc. Elle permet deux formats de sorties XML et JSON[1].

13. 'Lexical OWL Ontology Matcher'

14. même si le portail offre un moyen pour restreindre l'accès à des ontologies qui ne sont pas en libre accès.

NCBO Annotator

Cet outil réalise ce qu'on appelle l'annotation sémantique de données biomédicales à partir d'ontologies, ce qui consiste à annoter des descriptions textuelles avec des concepts d'ontologies en faisant la relation entre la description et les termes d'ontologies qui lui correspondent. Le service d'annotation se compose de deux parties : la partie annotation directe qui consiste en la détection de la présence du concept dans le texte et la partie expansion sémantique qui crée de nouvelles annotations avec des concepts similaires. Ce services est directement accessible dans BioPortal. Lorsqu'un utilisateur visite un concept, il a accès (lien web) à l'ensemble des éléments annotés avec ce concept.[17, 5] .

NCBO Resource Index

Les annotations des ressources à partir des termes d'ontologies peuvent être utilisées pour indexer les ressources de données biomédicales hétérogènes. Le résultat obtenu est stocké dans une base de données qu'on appelle Resource Index. Un service web et une interface utilisateur sont disponibles pour accéder à cette base[15] .

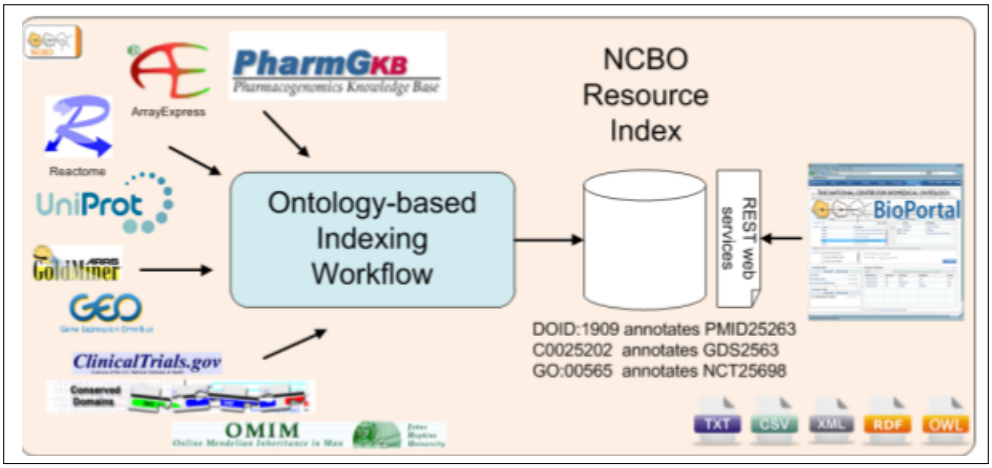


FIGURE 2.3 – NCBO Resource Index : Les ressources annotées sont stockées dans la base de données RI qui est accessible via un service web[15]

NCBO Ontology Recommender Service

Les chercheurs en bioinformatique utilisent les ontologies pour annoter leurs données afin de faciliter l'intégration translationnelle des données. Avec la croissance du nombres d'ontologies, il devient problématique pour un chercheur de choisir l'ontologie la mieux adaptée pour annoter ses données. Ce service utilise en entrée des données textuelles ou un ensemble de mots clés décrivant un domaine d'intérêt et propose en résultat l'ontologie appropriée pour l'annotation ou la représentation de ces données. Ce résultat

est basé sur tois critères. Le premier est la couverture, ou quelle ontologie contient les plus de termes couvrant l'ensemble du texte. Le second est la connectivité, ou les ontologies qui sont le plus alignées. Le dernier critère est le nombre de concepts dans l'ontologie. Les ontologies utilisées proviennent de BioPortal et UMLS.[16]

2.3 Catalogue et Index des Sites Médicaux de langue Française

Le **Catalogue et Index des Sites Médicaux de langue Française (CISMeF)** [15] est un projet initié par le CHU de Rouen. Le but de CISMeF est d'assister les professionnels de santé dans leurs quêtes d'informations et de connaissances médicales électroniques disponibles sur le web. Le cadre de CISMeF est centré sur la santé et les sciences médicales, dépassant la médecine proprement dite. Trois axes prioritaires ont été définis : les ressources concernant l'enseignement, la médecine factuelle (recommandations pour les bonnes pratiques cliniques et conférences de consensus) et les documents spécialement destinés aux patients et au grand public, dans le but de participer à l'amélioration de l'éducation sanitaire dans le monde francophone.[9] .

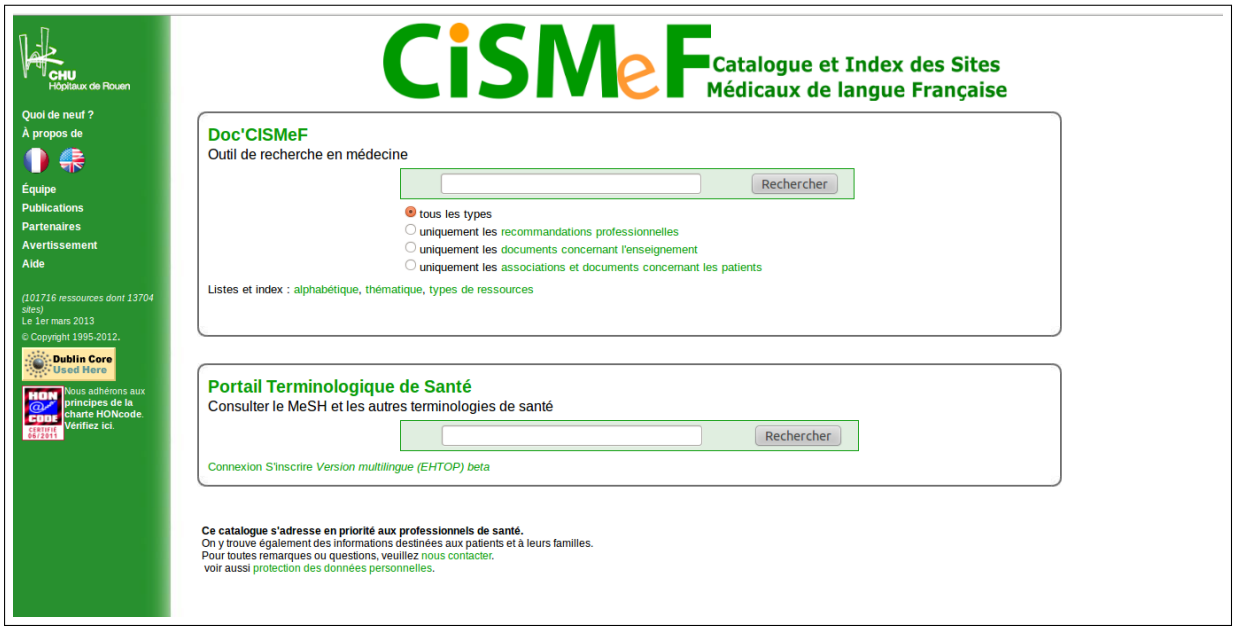


FIGURE 2.4 – Le portail CISMeF

15. http ://www.cismef.org/

2.3.1 Approche et technologies

CISMeF a une approche "plutôt terminologique". Il s'appuie sur l'ontologie MeSH [16] (Medical Subjects Headings), produite par la National Library of Medicine américaine (NLM), traduite en français par l'INSERM [17] et utilisée entre autre pour le catalogage d'ouvrages et la description d'articles de la base de données Medline [18].[27, 18]. Le CISMeF avait pour mission d'origine d'indexer manuellement des ressources de données avec un petit groupe de d'experts médicaux en utilisant MeSH, mais l'hétérogénité des ressources médicales sur le web ont conduit l'équipe du CISMeF à améliorer et enrichir MeSH.[27]. Cet enrichissement a donné lieu à de nombreux services, d'abord en recherche d'information (DocCISMeF et InfoRoute) puis en indexation automatique (ECMT et FMTI). Plus récemment, le catalogue est devenu multi-terminologique en introduisant dans son système d'information d'autres ontologies telles que (FMA, SNOMED, GO,etc.), cela a donné lieu à une autre gamme de services (HMTP).

Architecture

Avant 2005, CISMeF utilisait deux outils pour l'indexation des ressources : MeSH et le standard de métadonnées Dublin Core [19], les données sont stockées dans la bases de données CISMeF (Oracle 11.1g database). A partir de 2005, le catalogue est devenu multi-terminologique et il fallait donc revoir le modèle d'architecture pour s'adapter à ce changement et intégrer les nouvelles ontologies dans la base de données. Pour cela, il adapte une architecture combinant le modèle EAV (Entity-Attribute-Value) pour la base de donnée et le modèle UMV pour représenter les terminologies[14]

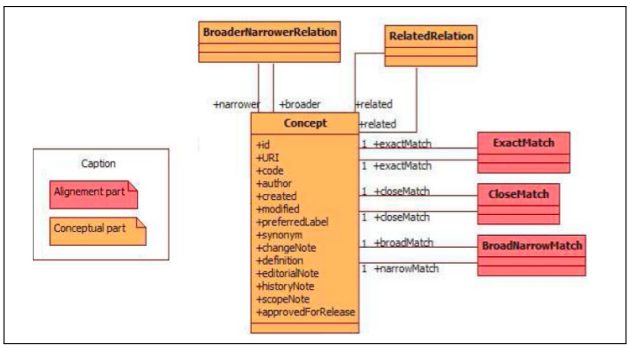


FIGURE 2.5 – Modèle EAV de la base de données [14]

16. http ://www.ncbi.nlm.nih.gov/mesh
17. http ://www.inserm.fr
18. http ://www.nlm.nih.gov/pubs/factsheets/jsel.html
19. http ://dublincore.org

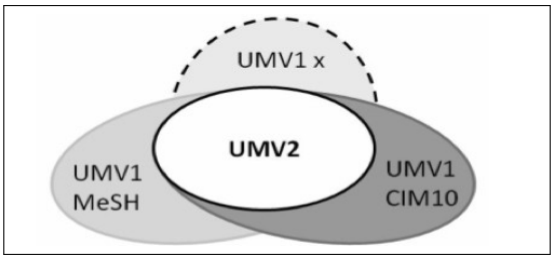


FIGURE 2.6 – Représentation du méta-modèle UMV2 et son extension UMV1 X où X représente une ontologie[14]

Alignement

L'équipe du CISMeF a développé un ensemble de méthodes d'alignement automatique telles que la méthode d'alignement conceptuel basée sur UMLS et la méthode lexicale. L'ajout d'alignements est effectué par les experts/curators.

Aspect communautaire

Les utilisateurs peuvent accéder aux services de recherche sans restriction [20]. CISMeF possède certains droits de propriété sur le contenu des ontologies hébergées sur son portail, ce contenu peut présenter un intérêt financier pour CISMeF. Ces droits se présentent comme ceci

- **Ontologie** : CISMeF n'a pas d'autorité sur le contenu de l'ontologie (concepts, hiérarchie, synonymes,etc.) tel qu'il était quand l'ontologie était importée sur le portail.
- **Intra-ontologie** : CISMeF possède des droits de propriété sur les ajouts/-modifications effectués par son équipe d'experts sur le contenu de l'ontologie (nouveaux synonymes, nouveaux attributs, traduction,etc.).
- **Inter-ontologies** : CISMeF possède aussi des droits de propriété sur les connaissances ajoutées par ses curators entre les ontologies (alignements, alignements multilingues).

20. Cependant, le nombre d'ontologies exploitables est limité, pour avoir la liste complète d'ontologies, une inscription sur le site est obligatoire.

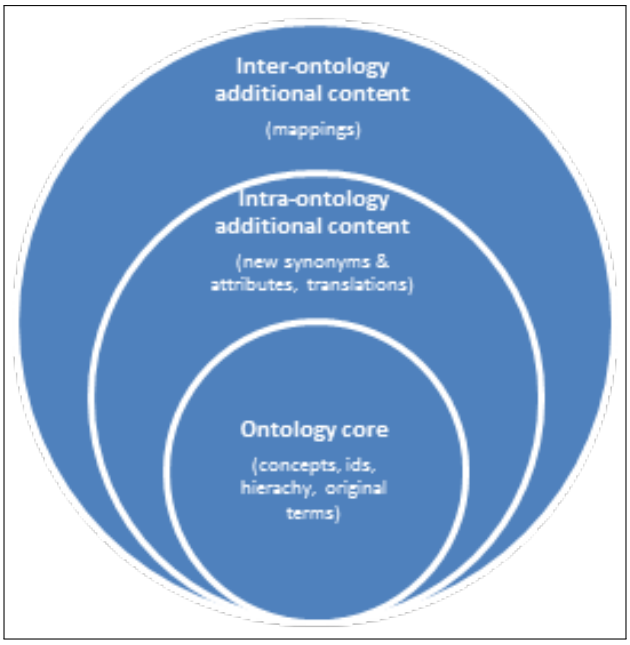


FIGURE 2.7 – Niveau de connaissances dans les ontologies du CISMeF : L'autorité de CISMeF porte sur le niveau intra-ontologique et le niveau inter-ontologique

2.3.2 Doc'CISMeF

Doc'CISMeF est le moteur de recherche qui permet de chercher parmi les ressources indéxées par CISMeF. Il exploite les synonymes et la hiérarchie des concepts. Grâce à la gestion des synonymes, une recherche sur 'tumeurs rénales' apporte les mêmes résultats qu'une recherche sur 'cancer du rein' car ces deux termes sont définis dans MeSH comme synonymes. Avec la plupart des moteurs de recherche généralistes, ces deux requêtes obtiennent des résultats très différents. Grâce à la hiérarchisation des concepts de MeSH, tous les termes spécifiques d'un descripteur sont pris en compte dans une requête. Une recherche sur les tumeurs rénales apporte donc toutes les ressources concernant les formes particulières de ce cancer (néphrocarcinome, tumeur de Wilms,etc.). C'est ce qu'on appelle l'"explosion sémantique" ou l'expansion sémantique de requête. Lors d'une recherche avec Doc'CISMeF cette explosion s'effectue par défaut, multipliant ainsi le nombre de réponses pertinentes. Doc'CISMeF exploite également le caractère international de MeSH en proposant une fonctionnalité appelée "Recherche Complémentaire" grâce aux traductions disponibles de MeSH et des autres ontologies utilisées. Pour chaque requête effectuée avec Doc'CISMeF des recherches vers des sources anglophones sont automatiquement proposées sous forme de liens hypertextes. Ainsi, après une requête portant sur "SIDA", une requête "Acquired Immunodeficiency Syndrome" est accessible d'un simple clic sur plusieurs bases de données (figure 2.9)[18].

FIGURE 2.8 – Doc'CISMef : Exemple de résultat de recherche sur "sclérose en plaques" : A droite se trouvent les résultats de recherche parmi les ressources indéxés et à gauche les différentes options pour affiner la recherche

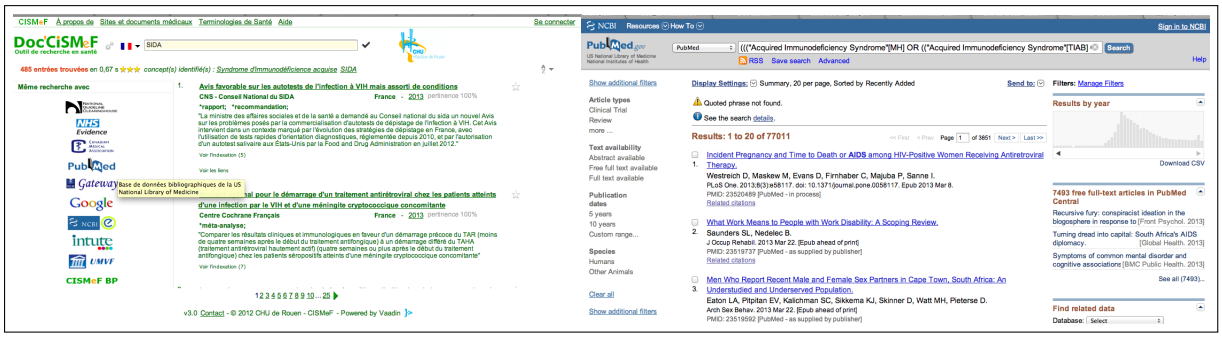


FIGURE 2.9 – Les résultats de recherche pour "SIDA" avec des liens de recherche sur sites anglophones (PubMed, Gateway, etc.), à gauche et le résultat pour "Acquired Immunodeficiency Syndrome" en choisissant PubMed par exemple, à droite

Doc'CISMeF recense et indexe plus de 100000 ressources médicales locales et internationales en accès libre et gratuit de sources institutionnelles et professionnelles (CNRS, Institut Pasteur, Santé Canada,etc.). Ces ressources se présentent sous forme de documents (articles, thèses, recommandations,etc.) qui représentent 85% des ressources indéxées et de sites internet qui représentent 15% de ces ressources. [2]

2.3.3 HMTP

Health Multi-Terminological Portal (HMTP [21]**)**, en français **Portail Terminologique de Santé (PTS)**, est proposé par l'équipe CISMeF depuis Mars 2005. C'est un espace consacré à sa terminologie qui permet la recherche et la consultation des termes médicaux utilisés dans le catalogue médical des ressources. Le changement apporté par HMTP concerne toute la stratégie de CISMeF qui gâce à cela est passée à l'univers multi-terminologique en introduisant de nouvelles ontologies et de nouvelles méthodes comme l'alignement et la méta-modélisation pour assurer l'intégration des ces nouvelles ontologies. HMTP offre un accès aux professionnels de santé pour manipuler des termes et peut être incorporé dans des applications dans le milieu médical. En outre, il offre ces fonctionalités sous forme de service web [14].
En Janvier 2013, HMTP compte 45 ontologies avec plus de 1600000 concepts[4]

Plus récemment, **The European Health Terminology/Ontology Portal (EHTOP** [22]**)** représente une évolution de HMTP. Ce portail offre les services d'accès aux ontologies en français et en anglais, mais également en italien, espagnol et allemand. Il compte plus de 32 ontologies. Depuis Janvier 2010, il est utilisé par les documentalistes de CISMeF pour l'indexation des ressources médicales en mode multi-terminologique. [13].

2.3.4 ECMT

L'Extracteur de Concepts Multi-Terminologiques (ECMT) [23] est un service permettant à partir d'une requête textuelle d'annoter les mots composant cette requête avec des termes de la base terminologique qui leur corespondent. Pour réaliser cette tâche, ECMT comporte deux modules : un module basé sur l'algorithme de sac de mots et un module d'indexation textuelle. ECMT s'est largement inspiré de l'algorithme de recherche d'information de Doc'CISMeF et de **F-MTI** un indexeur automatique multi-terminologique développé par CISMeF en collaboration avec la société VIDAL[26, 23].

2.3.5 InfoRoute

InfoRoute [24] est un outil permettant d'effectuer de la recherche d'information contextuelle en santé. Il a pour objectif de répertorier les moteurs de recherche spécialisés dans le domaine de la santé et de proposer une recherche simple via des "infobuttons" en

21. http ://pts.chu-rouen.fr/index.html
22. http ://www.ehtop.eu/
23. http ://doccismef.chu-rouen.fr/Interpreteur.html
24. http ://inforoute.chu-rouen.fr/irsite/

tirant partie des possibilités de recherche des moteurs de recherche et de la base terminologique. Ainsi, une requête en InfoRoute est automatiquement éxpansée sémantiquement et redirigée vers tous les moteurs de recherche disponibles dans InfoRoute (Google, Doc'CISMeF, Gateway,etc.). Cette application se distingue de Doc'CISMeF car il n'y a pas d'indexation du contenu des ressources recherchées. La résolution de la requête est confiée directement au moteur de recherche de la source originale. Cette application est similaire à ENTREZ [25] [8].

2.4 Conclusion

Dans cette section, nous avons présenté les deux portails d'ontologies biomédicales NCBO BioPortal et le portail CISMeF. Nous avons décrit leurs architectures, approches et les différents services qu'ils proposent. Comme nous avons pu voir, les deux portails présentent quelques similitudes dans leurs services. Dans la section suivante, nous allons nous intéresser aux services d'annotation sémantique proposés par ces portails.

25. http ://www.ncbi.nlm.nih.gov/sites/gquery

Chapitre 3

Comparaison des workflows d'annotation sémantique

3.1 Introduction

Dans cette partie, nous nous intéressons aux outils d'annotation sémantique proposés par les deux portails et détaillés précédemment à savoir ECMT, FMTI et le NCBO Annotator. Cette comparaison portera sur les performances de ces trois outils pour des données anglophones et francophones.

3.2 Jeu de données et corpus de référence

Pour effectuer cette comparaison fonctionnelle, nous avons choisi un ensemble de corpus de données textuelles pour nous servir de jeu de test comme le montre le tableau suivant :

Table 3.1 – Corpus choisis pour l'annotation

Corpus	Nature	Nombre documents
Citations Pubmed	Titres d'articles	2000 fr et 2000 en
Labtestsonline	Données de tests en biologie	140 fr et 280 en

Notant que le corpus PubMed est préalablement annoté par MeSH. Il servira donc comme corpus de référence pour évaluer les résultats de chaque outil. Nous considérons aussi MeSH comme seul critère d'évaluation pour les trois outils.

3.3 Démarche de comparaison

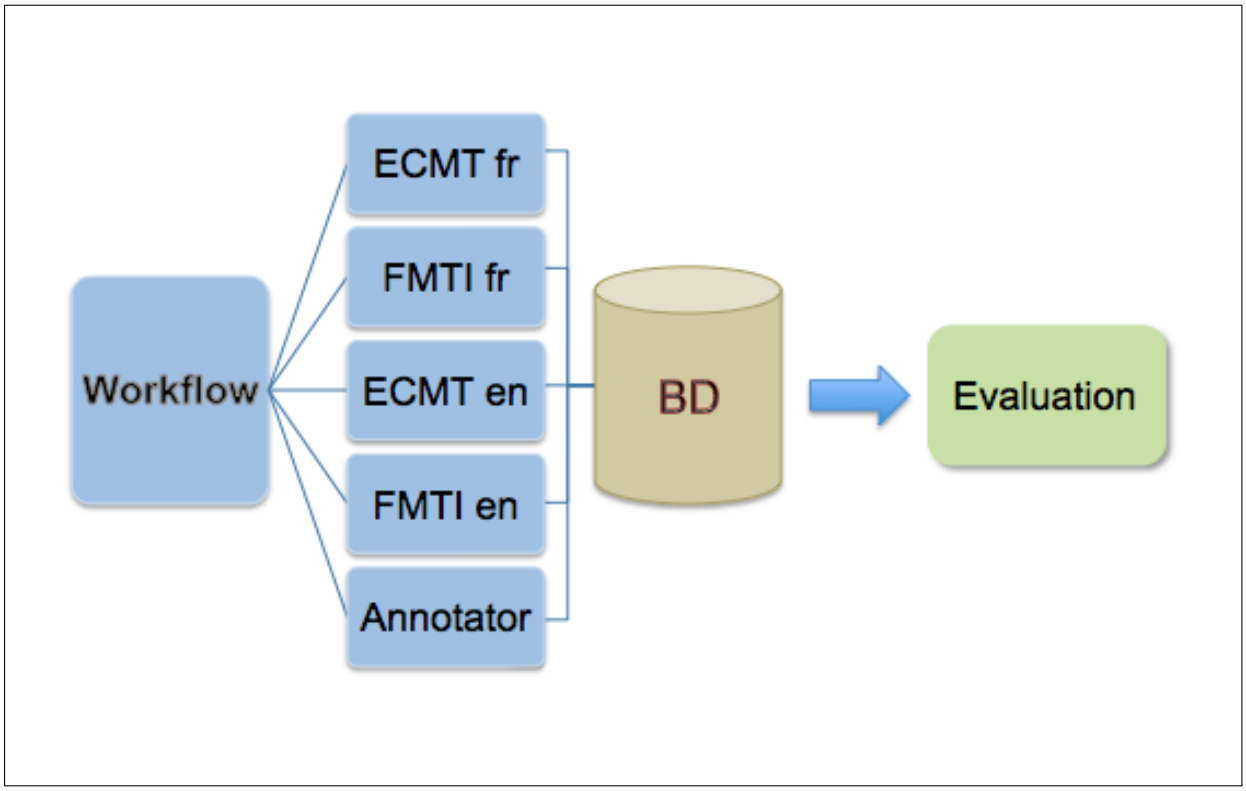


FIGURE 3.1 – Démarche de comparaison : Annotation des corpus textuels par les outils puis stockage des résultats dans une base de données ensuite évaluation

Notre démarche consiste à comparer les perfomances d'annotation des outils de la façon suivante :

Données anglophones Pour les citations PubMed, nous comparons les résultats obtenus par l'Annotator, ECMT et FMTI avec le corpus de référence. Pour le corpus Labtestsonline nous comparons les résulats obtenus par chaque outil par rapport aux autres.

Données francophones Comme nous nous disposons pas de l'équivalent de l'Annotator pour le français, la comparaison s'effectuera donc pour ECMT et FMTI avec le corpus de référence pour les citations PubMed et entre les deux outils pour les données Labtestsonline.

Le tableau suivant résume cette démarche de comparaison

TABLE 3.2 – Etape de comparaison des différents outils

	Français	Anglais
Citations PubMed	ECMT/Référence FMTI/Référence	ECMT/Référence FMTI/Référence Annotator/Référence
Labtestsonline	ECMT/FMTI	ECMT/FMTI ECMT/Annotator FMTI/Annotator

3.4 Condition d'appel et paramétrage des services

3.4.1 Le NCBO Annotator

Le NCBO Annotator est appelé via un service web Annotator REST service. Pour effectuer l'appel nous avons utilisé l'exemple du client Java fourni par le NCBO [1]

```
try {
    HttpClient client = new HttpClient();
    client.getParams().setParameter(
    HttpMethodParams.USER_AGENT,"Annotator Client Example - Annotator");  //Set this string for your application

    String text = "Melanoma is a malignant tumor of melanocytes which are found predominantly in skin but also in the bowel and the eye";

    PostMethod method = new PostMethod(annotatorUrl);

    // Configure the form parameters
    method.addParameter("longestOnly","true");
    method.addParameter("wholeWordOnly","true");
    method.addParameter("filterNumber", "true");
    //method.addParameter("stopWords","I,a,above,after,against,all,alone,always,am,amount,an,and,any,are,around,as,at,back,be,before,behind,below,between,
    method.addParameter("withDefaultStopWords","true");
    method.addParameter("isTopWordsCaseSensitive","false");
    method.addParameter("mintermSize","3");
    method.addParameter("scored", "true");
    method.addParameter("withSynonyms","true");
    method.addParameter("ontologiesToExpand", "");
    method.addParameter("ontologiesToKeepInResult", "");
    method.addParameter("isVirtualOntologyId", "true");
    method.addParameter("semanticTypes", "");
    method.addParameter("levelMax", "0");
    method.addParameter("mappingTypes", ""); //null, Automatic, Manual
    method.addParameter("textToAnnotate", text);  //"Melanoma is a malignant tumor of melanocytes which are found predominantly in skin but also in the bc
    method.addParameter("format", "xml"); //Options are 'text', 'xml', 'tabDelimited'
    method.addParameter("apikey", "YourAPIKey");
```

FIGURE 3.2 – Exemple du client java pour l'Annotator avec les paramétrages par défaut

Nous avons paramétré le client Java de la façon suivante :

1. http ://www.bioontology.org/wiki/index.php/Annotator_Client_Examples

TABLE 3.3 – Corpus choisis pour l'annotation

Paramètre	Valeur choisie	Signification
longestOnly	False	Prendre toutes les phrases en considération et non seulement la plus longue
wholeWordOnly	True	Prendre en considération uniquement les mots entiers et non tronqués
filterNumber	True	Pour filtrer les chiffres
withDefaultStopWords	True	Pour utiliser le filtre 'mots vides' fourni par défaut
mintermSize	3	Taille minimale d'un terme pour qu'il soit pris dans l'annotation
scored	True	classer les résultats selon leur score
withSynonyms	True	Pour garder les annotations effectuées avec des termes synonymes
ontologiesToExpand	Tous	Pour choisir quelle ontologie à expanser (le paramètre est laissé par défaut ici vu que nous utilisons une seule ontologie)
ontologiesToKeepInResult	1351	Pour choisir l'ontologie utilisée pour l'annotation. Dans notre cas nous avons choisi MeSH dont le virtual ID est 1351
isVirtualOntologyId	True	Pour dire que c'est le virtual ID et non le local ID (c'est l'ID de l'ontologie indépendemment de la version)
semanticTypes	Tous	Pour choisir les types sémantiques UMLS à inclure
levelMax	5	Profondeur maximale pour la closure
mappingTypes	Tous	Pour définir les types de mapping à inclure
textToAnnotate	text	'text' faisant référence à la variable qui contient le texte à envoyer au service
format	XML	Pour choisir le format de sortie des résultats
apikey	3bbe5511-42e2-4387-b411-0d9677f24ab4	un apikey est fournit pour chaque utilisateur enregistré sur BioPortal. Il est nécessaire pour exploiter les NCBO REST services

Comme l'annotator peut traiter un texte assez large (e.g. un chunk de 500 mots) [2], l'annotation de notre corpus est effectuée par l'envoi de chaque document à la fois. Le résultat est récupéré dans un fichier xml correspondant au document envoyé.

2. Voir l'annexe

```
<annotatorResultBean>
  <resultID>OBA_RESULT_68e0</resultID>
  <statistics>
    <statisticsBean>
      <contextName>MGREP</contextName>
      <nbAnnotation>3</nbAnnotation>
    </statisticsBean>
    <statisticsBean>
      <contextName>MAPPING</contextName>
      <nbAnnotation>0</nbAnnotation>
    </statisticsBean>
    <statisticsBean>
      <contextName>CLOSURE</contextName>
      <nbAnnotation>0</nbAnnotation>
    </statisticsBean>
  </statistics>
  <parameters>
    <longestOnly>false</longestOnly>
    <wholeWordOnly>true</wholeWordOnly>
    <filterNumber>true</filterNumber>
    <withDefaultStopWords>true</withDefaultStopWords>
    <stopWords/>
    <minTermSize>3</minTermSize>
    <withContext>true</withContext>
    <withSynonyms>true</withSynonyms>
    <ontologiesToExpand/>
    <levelMax>5</levelMax>
    <mappingTypes/>
    <semanticTypes/>
    <isStopWordsCaseSenstive>false</isStopWordsCaseSenstive>
    <ontologiesToKeepInResult>
      <string>1351</string>
    </ontologiesToKeepInResult>
    <isVirtualOntologyId>true</isVirtualOntologyId>
    <textToAnnotate>Lung carcinosarcoma A case report</textToAnnotate>
    <outputFormat>xml</outputFormat>
    <apiKey>3bbe5511-42e2-4387-b411-0d9677f24ab4</apiKey>
  </parameters>
  <annotations>
    <annotationBean>
      <score>10</score>
      <concept>
        <id>33735824</id>
        <localConceptId>46836/D002296</localConceptId>
        <localOntologyId>46836</localOntologyId>
        <isTopLevel>0</isTopLevel>
        <fullId>http://purl.bioontology.org/ontology/MSH/D002296</fullId>
        <preferredName>Carcinosarcoma</preferredName>
```


FIGURE 3.3 – Exemple de résultat obtenu pour une citation PubMed : L'entête du fichier contient les paramètres d'annotation définis auparavant, ensuite les concepts trouvés sont affichés par ordre décroisaant selon le score

3.4.2 ECMT

ECMT peut être appelé via une requête formulée en url sur la base fixe "http://ecmt.chu-rouen.fr/servlets/Interpreteur?Mot="
Pour effectuer l'appel, la phrase à annoter est concaténée à cette url. La phrase est préalablement encodée pour le format url (remplacement des espaces et autres caractères spéciaux par leur encodage url).

```
donnees=ligne;
URL urlEcmt;
urlEcmt=new URL("http://ecmt.chu-rouen.fr/servlets/Interpreteur?Mot="+donnees);
ecmt = urlEcmt.openConnection();
ecmt.setDoInput(true);
ecmt.setDoOutput(true);
ecmt.setUseCaches(false);
ecmt.setDefaultUseCaches(false);
ecmt.setRequestProperty("","");
DataOutputStream out = new DataOutputStream(ecmt.getOutputStream());
out.writeBytes(donnees);
out.flush();
out.close();
BufferedReader in=new BufferedReader(new InputStreamReader(ecmt.getInputStream()));
```

FIGURE 3.4 – Exemple du programme utilisé pour l'appel du service ECMT : l'appel est effectué via une requête sur la base fixe avec la phrase à annoter (contenu de la variable donnees)

ECMT n'offre pas d'options de paramétrage de l'appel ou de choix d'ontologies, néanmoins il permet de choisir le type d'annotation souhaitée : directe ou expansée.
Pour obtenir l'expansion sémantique, il faut rajouter "&Exp" à la fin de la phrase à envoyer.
La taille du texte géré par ECMT est assez limitée (e.g. phrase de 50 mots) [3], nous avons découpé chaque document de notre corpus en phrases avant de les envoyer une à une. Le résultat est récupéré sous forme de fichier xml.

3.4.3 FMTI

FMTI est un outil appartenant à la société VIDAL [4], nous avons pu compter sur les responsables de l'outil dans la société VIDAL pour réaliser l'annotation de nos corpus anglais et français avec cet outil. L'annotation est effectuée avec MeSH. FMTI réalise l'annotation par fichier, il fallait donc découper notre corpus en fichiers coresspondant à chaque documet. Le résultat obtenu est un fichier xml.

3. Voir annexe
4. http ://www.vidal.fr/

```xml
<?xml version="1.0" encoding="ISO-8859-1"?>
<INDEXINGFMTI>
<DOCUMENT ID="">
<TEXT> Métastase splénique isolée révélant un cancer du colon.</TEXT>
<SENTENCE ID="1">
<TEXTSENTENCE>Métastase splénique isolée révélant un cancer du colon.</TEXTSENTENCE>
<SENTENCEINDEXING>
<TERM CODE="D009369">
<TERMLABEL>tumeurs</TERMLABEL>
<TERMTERMINO>MSH</TERMTERMINO>
</SENTENCEINDEXING>
</SENTENCE>
<INDEXINGLANGUAGE>FR</INDEXINGLANGUAGE>
<TEXTLANGUAGE>FR</TEXTLANGUAGE>
<INDEXINGTERMINOS>MSH-FR</INDEXINGTERMINOS>
<FINALINDEXING>
<TERM CODE="D009369">
<TERMLABEL>tumeurs</TERMLABEL>
<TERMTERMINO>MSH</TERMTERMINO>
<TERMOCC>1</TERMOCC>
</TERM>
</FINALINDEXING>
</DOCUMENT>
</INDEXINGFMTI>
```

FIGURE 3.5 – Exemple de résultat obtenu

3.5 Traitement des résultats

Les résultats obtenus sous forme de fichiers xml représentent les annotations et les concepts relatifs trouvés par les outils. Pour faciliter leur exploitation plus tard, nous les avons stockés dans une base de données selon le schéma suivant :

Concept Cette table contient tout les concepts de MeSH avec leurs identifiants et noms.

Ressources Chaque document de nos corpus est représenté avec son numéro et la ressource à laquelle il appartient (PubMed, Labtestsonline).

Pour chaque outil, les annotations sont représentées dans une table contenant l'annotation, le concept associé, la ressource à laquelle elle appartient et si elle est expansée ou non.

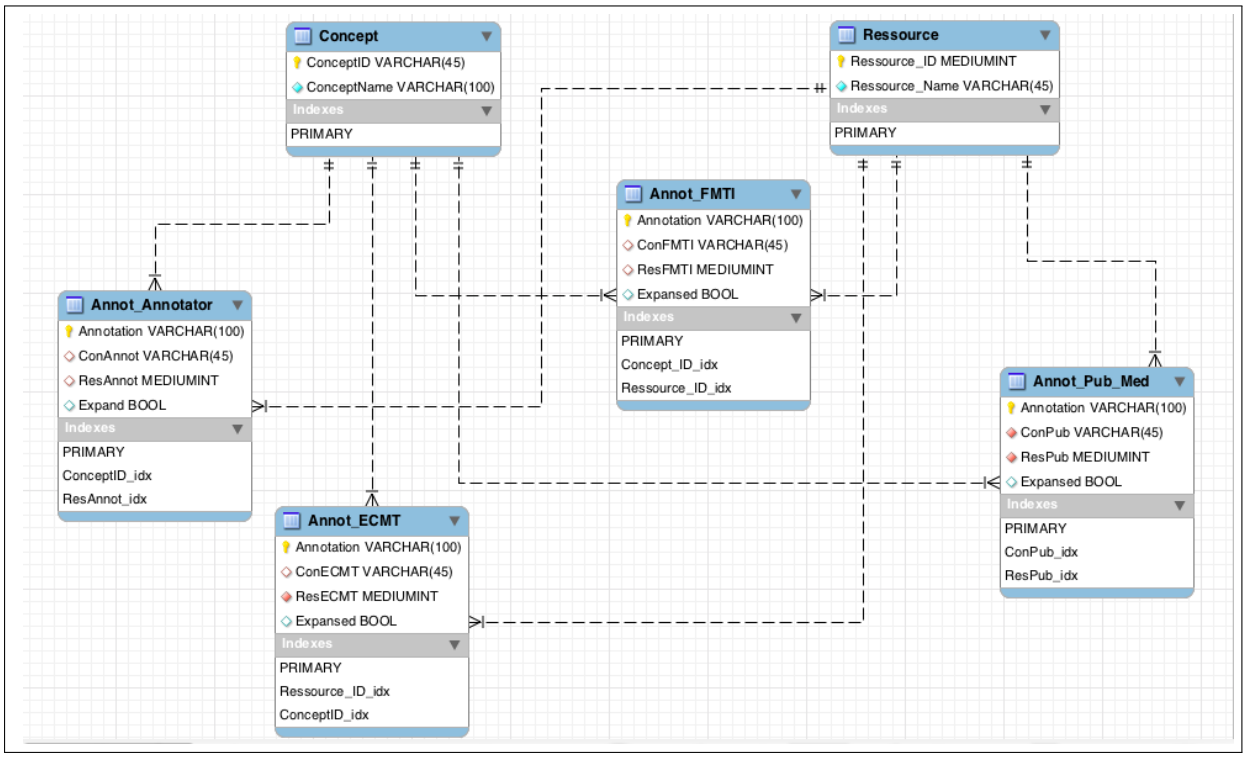


FIGURE 3.6 – Schéma de la base de données pour stocker les résultats d'annotation : chaque outil est représenté par une table spécifique

3.6 Conclusion

Dans cette section nous avons vu les différents paramétrages pour effectuer l'appel aux trois services d'annotation. Les annotations obtenues par chaque outil ont été stockées dans la base de données. Dans la section suivante nous allons évaluer ces résultats et ainsi la performance de ces services.

Chapitre 4

Evaluation des outils

4.1 Introduction

Dans ce chapitre nous allons procédé à l'évaluation des résultats d'annotation obtenus par chaque outil. Pour ce faire nous allons calculer la précision et le rappel pour chacun des outils d'annotation, par rapport au corpus de référence. Pour les données Labtestsonline, nous comparons les résultats d'annotation entre les outils comme cité dans le chapitre précédent.

4.2 Corpus PubMed

Le corpus de citations PubMed utilisé pour le test des outils est est déjà annoté manuellement. Pour évaluer les trois services d'annotations, nous calculons les valeurs de précision et de rappel de chaque outil par rapport à l'annotation de référence.

Définition 1 Le rappel désigne le nombre de documents pertinants retrouvés par rapport au nombre total des documents pertinents de référence. Dans notre cas, il désigne le nombre d'annotations de référence retrouvées au regard du nombre total d'annotations de référence.

$$\boxed{Rappel = \frac{Annotations\ du\ corpus\ de\ référence\ retrouvées}{Total\ des\ annotations\ du\ corpus\ de\ reférence}}$$

Définition 2 La précision désigne le nombre de documents pertienets retrouvé par rapport au nombre de documents retrouvés. Dans notre cas, cela indique le nombre d'annotations de référence retrouvées au regard du nombre d'annotation retrouvées par chaque outil.

$$Précision = \frac{Annotations\ du\ corpus\ de\ référence\ retrouvées}{Total\ des\ annotations\ retrouvées\ par\ l'outil}$$

Nous désignons ici par annotation le concept MeSH retrouvé puisque nous pouvons retrouver le même concept dans l'annotation de plusieurs ressources.
Le corpus d'annotation est manuellement annoté avec 3098 concepts répartis sur 19003 annotations pour 2000 documents. Le tableau suivant montre les résultats et observations obtenus

TABLE 4.1 – Résultat de l'évaluation des annotations du corpus PubMed par les trois outils pour le français et l'anglais

	Annotator	**ECMT**		**FMTI**	
Langue	Anglais	Anglais	Français	Anglais	Français
Nombre annotations	1205	2075	1190	1215	1104
Nombre annotations exactes	900	907	558	818	492
Rappel	0,2905	0,2927	0,1801	0,2640	0,1588
Précision	0,7468	0,4371	0,4689	0,6732	0,4456

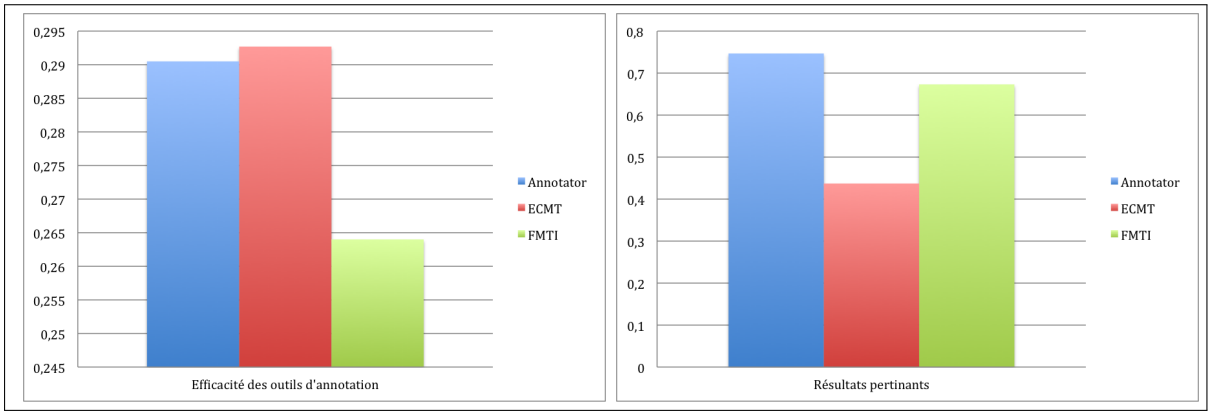


FIGURE 4.1 – Résultats des données en anglais retrouvés et leur précision

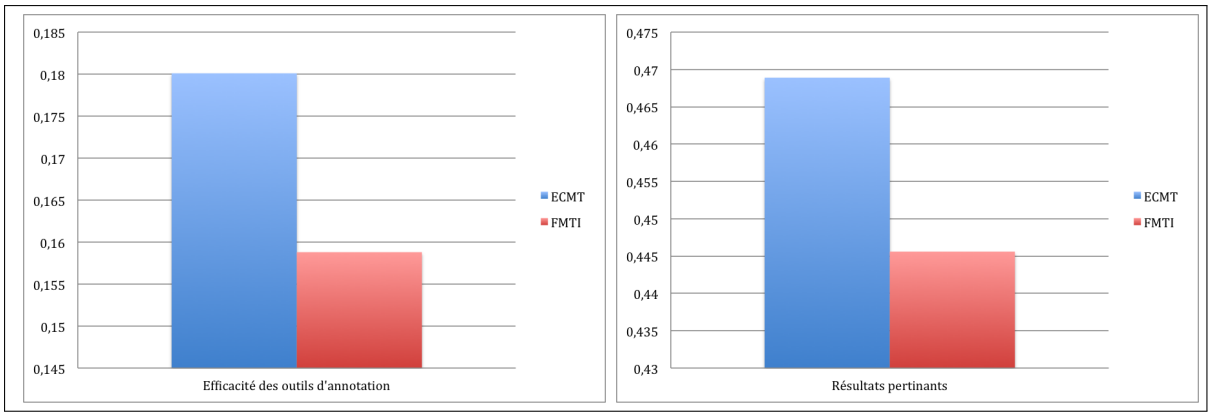


FIGURE 4.2 – Résultats des données en français retrouvés et leur précision

D'après les résultats du tableau, nous pouvons remarquer que les trois outils d'annotation affichent un taux de rappel similaire et faible par rapport à l'annotation manuelle de référence.

Pour les données anglophones, le NCBO Annotator donne de meilleurs résultats que les deux autres (0,7468), ECMT étant le moins précis. Pour les données francophones, nous constatons que ECMT est plus précis que FMTI.

Ces valeurs restent néanmoins relatives vu qu'elles ne considèrent que MeSH comme ontologie pour l'annotation et qu'un outil comme ECMT ne permet pas de choisir l'ontologie et/ou la relation utilisée pour l'expansion, certains résultats peuvent être pénalisés.

4.3 Corpus Labtestsonline

Pour ce corpus nous ne possédons pas de référence, nous allons donc comparer les résultats de chaque outils par rapport aux deux autres.

TABLE 4.2 – Annotations trouvées pae chaque outil

	Annotator	**ECMT**	**FMTI**
Nombre annotations trouvées	4477	1560 fr et 5749 en	et 1230 fr 3681 en

Table 4.3 – Comparaison des résultats pour le NCBO Annotatore/ECMT et le NCBO Annotator/FMTI

Annotations communes	1047
Annotations Annotator	3430
Annotations ECMT	4702

Annotations communes	1230
Annotations Annotator	3247
Annotations FMTI	2451

Table 4.4 – Comparaison des résultats entre ECMT et FMTI

	Français	Anglais
Annotations communes	920	1097
Annotations ECMT	640	4652
Annotations FMTI	133	2584

d'après les résultats de ces tableaux, nous pouvons remarquer que les annotations communes pour les trois outils sont assez faibles et que beaucoup présentes dans l'un ne le sont pas dans l'autre. Toutefois, nous n'avons aucun moyen de déterminer la qualité de ces annotations comme dans le corpus précédent en l'absence d'un corpus de référence.

4.4 Conclusion

Dans ce chapitre, nous avons évalué la qualité d'annotation des trois outils d'annotation au regard du corpus de référence et ceci pour les données en français et en anglais.

Chapitre 5

Conclusion

5.1 Bilan

Durant ces cinq mois de stage, nous nous sommes intéressés aux caractéristiques de deux portails de ressources biomédicales.

Chaque portail était présenté en précisant l'approche, les fonctionnalités et les services qu'il propose. Nous avons réalisé une synthèse générale démontrant les similarités et les différences entre les deux portails.

Nous avons mis l'accent sur la comparaison fonctionnelle des workflows d'annotation sémantique proposés par les deux portails. Nous avons détaillé notre démarche de comparaison basée sur des tests de performance des outils d'annotation. Nous avons évalué les résultats de ces outils et démontré les limites de chaque service.

5.2 Perspectives

Le résultat du travail proposé amène les perspectives suivantes.

La première concerne l'évaluation des outils qui était réalisée par rapport à l'ontologie MeSH. Ces tests peuvent être réalisés pour d'autres ontologies et en impliquant un jeu de test plus riche pour avoir un meilleur feedback.

Suite à cette comparaison, un modèle commun qui permet la réutilisation des propriétés des deux portails (alignement, visualisation,etc.) pourrait être envisagé.

5.3 Difficultés rencontrées

Les difficultés rencontrées durant ce stage concernaient particulièrement le processus de traitement des résultats pour extraire les données, ce qui a retardé la rédaction de ce rapport.

La partie comparaison générale fût un peu laborieuse vu qu'il fallait choisir les crtières de comparaison pertinents à partir de l'état de l'art. Etant nouvelle dans ce domaine, un temps d'adaptation fût nécessaire.

5.4 Apport

La comparaison réalisée lors de ce stage permettra d'étudier plus tard une possibilité de convergence des deux portails. La base de données contenant les résultats d'annotation pourrait être réutilisée pour d'autres travaux concernant le projet SIFR.

D'un point de vue plus personnel, ce stage passé au Lirmm fût une occasion pour moi de découvrir le domaine de la recherche, collaborer avec des experts et approfondir mes connaissances dans le domaine du web sémantique et traitement de données.

Annexe A

Comparaison des deux portails

Le portail du NCBO et celui du CISMeF possèdent plusieurs caractéristiques communes mais chacun d'entre eux a ses propres spécificités. Nous pouvons d'ores et déjà regrouper les fonctionnalités des deux portails dans 5 grands groupes tels que :

- Hébergement d'ontologies.
- Workflow d'annotation sémantique.
- Indexation des ressources.
- Services propres à chaque portail.
- API et web services.

Les caractéristiques de chaque groupe sont détaillées dans les tableaux suivants :

Hébergement d'ontologies

Dans cette partie nous présentons les caractéristiques liées à l'hébergement des ontologies pour chaque portail. Ces caractéristiques regroupent le contenu, les fonctionnalités d'accès, l'alignement des ontologies, l'aspect communautaire et autre.

Caractéristique	NCBO	CISMeF
Contenu		
Nom du service	BioPortal [1]	Health Multi-Terminological Portal (HMTP) ou Portail Terminologique de Santé (PTS) en français [2]. Et plus récemment, European Health Terminology/Ontology Portal (EHTOP) [3]
Provenance des ontologies	– Importées d'UMLS. – Importées d'OBO Foundry. – Ajoutées directements par les utilisateurs.	Ajoutées par les experts/curators du CISMeF au cas par cas.
Format accepté	OWL, OBO, RDF, RRF, Protege Framework Language.	Pas d'import automatique.
Politique d'hébergement	BioPortal accepte tout, en supposant que c'est biomédical et que c'est "bien formé".	CISMeF procède au cas par cas et se réserve le choix d'importer ou non une ontologie.

1. http ://bioportal.bioontology.org/
2. http ://pts.chu-rouen.fr/
3. http ://www.ehtop.eu/

Métadonnées des ontologies	Formalisées sous forme d'ontologie "BioPortal Metadata Ontology"(http://bioportal.bioontology.org/ontologies/42948?p=summary), par exemple (nom, ID, format, description, URI, catégorie, groupe, revues, versions, URL, métriques, vues, alignements, notes et projets liés, utilisateurs).	Pas de spécification particulière pour les métadonnées. On peut avoir accès directement à certaines métadonnées directement dans l'onglet "Terminologies" : Nom, description, éditeur, URL, version, langue, hiérarchie.
Gestion des versions d'ontologies	Pour chaque ontologie, chaque version est disponible et accessible grâce au versionID qui identifie la version "virtuelle" de l'ontologie. C'est un identifiant unique quelle qu'elle soit la version de l'ontologie dans le portail et qui pointe vers la dernière version, et le versionID qui identifie une version spécifique de l'ontologie. [31]	Pas de gestion de versions.
Condition d'accès aux ontologies	Tout est en libre accès. En outre, le portail propose l'accès sécurisé au cas par cas	Accès libre pour MeSH et CISMeF. Le reste est en accès restreint (inscription).
Fonctionnalités d'accès		
Inscription	Requise uniquement pour ajouter des ontologies mais non nécessaire pour y accéder.	Requise pour avoir accès aux ontologies non libre d'accès.

Recherche de concepts dans les ontologies	Les deux portails offrent des options de recherche similaires : – Onglet de recherche. – Suggestion de recherche. – Recherche avancée (terme avec définition, terme exact,.etc). – Choix des ontologies incluses dans la recherche.	
Recherche et organisation des ontologies	BioPortal permet la recherche parmi les ontologies. Les ontologies sont organisées par catégories (anatomie, santé, molécule, etc.) et par groupes : – CTSA Health Ontology Mapper (CTSA-HOM) – Cancer Biomedical Informatics Grid (caBIG®) – Clinical and Science Translational Science Awards (CTSA) – OBO Foundry – Proteomics Standards Initiative (PSI) – Unified Medical Language System (UMLS) – World Health Organization (WHO) Family of International Classifications (WHO-FIC) (http://bioportal.bioontology.org/ontologies)	Pas de spécification particulière et pas de regroupement d'ontologies.
Visualisation des ontologies	Interface de visualisation qui permet l'exploration des relations complexes, alignements, etc. Sous forme d'arbre de navigation standard et sous forme de graphe d'ontologie.	Arbre de navigation standard et liste mais pas de graphe d'ontologie.
Edition des ontologies	Pas d'environnement pour l'édition d'ontologies. Un futur avec WebProtege [4], un éditeur d'ontologies basé sur le Web qui sera intégré à BioPortal. [30]	Les experts/curators peuvent modifiér le contenu dans le backend à l'aide d'un accès spécifique.

4. http ://protegewiki.stanford.edu/wiki/WebProtege

Téléchargement des ontologies	Disponible pour chaque ontologie (exemple avec l'ontologie NCIt http://rest.bioontology.org/bioportal/ontologies/download/47638?apikey=4ea81d74-8960-4525-810b-fa1baab576ff&userapikey=)	Pas de possibilité de téléchargement mais il existe un travail en cours sur une fonctionnalité d'export.
Publication d'ontologies	Les utilisateurs inscrits peuvent publier leurs ontologies sur le portail.	Seuls les administrateurs et les curators de CISMeF ont le droit d'import et de publication des ontologies.
Modification des ontologies	Aucune.	Possible et effectuée par les experts de CISMeF.
Alignements		
Ajout des alignements	– Importation automatique (algorithmes et méthodes d'alignement). – Ajout manuel par les utilisateurs via l'interface utilisateur.	– Importation automatique (algorithmes et méthodes d'alignement). – Ajout manuel par les curators.

Visualisation des alignements	Possibilité de visualiser les alignements : – Pour une ontologie (liste des ontologies alignées avec cette ontologie), exemple SNOMED-CT (http://bioportal.bioontology.org/ontologies/1353/?p=mappings) – Par couple d'ontologies, en sélectionant une ontologie dans la liste des alignements d'une ontologie, on peut voir les alignements entre les deux. Exemple dans la liste d'alignements de SNOMED-CT en cliquant sur NCIt, on peut visualiser en détail les 32042 alignements entre les deux ontologies. – Par concept. Exemple du concept "AIDS" dans SNOMED-CT et la liste des alignements avec des concepts dans d'autres ontologies (http://bioportal.bioontology.org/ontologies/1353?p=terms&conceptid=http%3A%2F%2Fpurl.bioontology.org\%2Fontology%2FSNOMEDCT%2F62479008#mappings)	Possibilité de visualisation des alignements pour un concept via l'onglet "Relations", exemple avec le concept "AIDS" (http://pts.chu-rouen.fr/recherche.html).
Téléchargement des alignements	Les utilisateurs peuvent télécharger les alignements (sans pour autant télécharger des ontologies), exemple	Pas de spécification particuleère.
Aspect communautaire		
Commentaires des utilisateurs	Possibilité de laisser des commentaires sur les ontologies et/ou les alignements. Les commentaires sont rajoutés dans les métadonnées de l'ontologie, exemple avec la liste de commentaires sur NCIt (http://bioportal.bioontology.org/ontologies/1032/?p=notes)	Non disponibles.
Notification et flux RSS	Possibilité de s'inscrire au service de notification pour être informé des nouvelles ontologies ajoutées et des commentaires ajoutés sur les ontologies. (http://bioportal.bioontology.org/syndication/rss)	Non disponible.

Documentation et assistance	– Onglet "Aide" pour assister l'utilisateur dans l'exploration du portail (http://bioportal.bioontology.org/help). – NCBO Wiki qui est un support de documentation sur les logiciels, projets et technologies du NCBO (http://www.bioontology.org/wiki/index.php/Main_Page).	– Rubrique "A propos de" qui fournit des informations générales et plus techniques sur le portail (http://www.chu-rouen.fr/cismef/cismef.html). – Blogue CISMeF, un blogue qui aborde l'information médicale francophone, Internet et les questions de santé, l'informatique médicale, l'accès à la littérature scientifique et, bien sûr, CISMeF (http://www.cismef.org/cismef/blog/)

Autorité sur le contenu	Aucune, les ontologies hébergées sur BioPortal appartiennent à leurs créateurs	CISMeF possède un droit de propriété sur : – Les ontologies créées par son équipe (e.g. CISMeF). – les ontologies modifiées par ses experts (traductions, nouveaux synonymes,etc.) – les connaissances ajoutées sur les ontologies par ses experts (alignements, alignements multilingues, etc.)
Architecture et technologies utilisées		

Architecture	Multi-couches : – Couche Présentation qui fournit l'interface utilisateur. – Couche Interface qui contient les Web services nécessaires pour les fonctionalités de BioPortal. – Couche logique qui permet l'accès des API aux ontologies et au Resource Index. – Couche Persistance pour les données. Cette architecture tend à passer au modèle RDF triplestore avec l'utilisation de 4store.	Oracle database avant 2005. A partir de 2005, modèle EAV+UMV [14]. EAV (Entity-Attribute-Value) : Un méta-modèle est créé pour la base de données pour l'adapter aux nouvelles ontologies. Ce méta-modèle factorise les objets communs à toutes les ontologies (classes, attributs, relations). Le modèle de chaque ontologie est une spécialisation du méta-modèle. UMV(Unifying Model of Vocabulary) : Pour Représenter chaque ontologie.
Technologies utilisées	Protege, LexGrid, MySQL, Spring/JDBC, Hibernate, 4store	SKOS, Oracle
Autres critères		

Approche/ Philosophie	Approche ontologique	Approche terminologique au départ qui tend à devenir ontologique
Multilinguisme	– Interface utilisateur disponible uniquement en anglais. – Pas de relation entre les termes en anglais et ceux dans d'autres langues. – Les ontologies dans d'autres langues sont représentées comme des vues de l'ontologie anglaise.	– UI disponible en français et en anglais (et en d'autres langues pour EHTOP). – Possibilité de navigation et de recherche entre les termes en multilingue. – Le méta-modèle inclus la relation entre les termes français et anglais (et autres langues pour EHTOP).

Workflow d'annotation sémantique

Dans cette partie, nous présentons les caractéristiques liées au workflow d'annotation sémantique dans chaque portail.

Caractéristique	**NCBO**	**CISMeF**
Nom du service	NCBO Annotator [5]	Extracteur de Concepts Multi-Terminologiques (ECMT) [6]
Performance	Peut traiter un texte assez large (e.g. chunk de 500 mots).	Taille de texte traité assez limitée (e.g. phrase de 50 mots).
Fonctionalités	– Possibilité de choisir l'ontologie voulue pour l'annotation – Expansion sémantique avec las alignements et la relation "is-a". [17, 5]	– Pas de choix d'ontologies pour l'annotation. – Pas d'expansion sémantique.
Méthode de reconnaissance des concepts	Détection de la présence du mot dans le texte par correspondance syntaxique.	Méthodes NLP (algorithme de sac de mots).

5. http ://bioportal.bioontology.org/annotator
6. http ://ecmt.chu-rouen.fr/

Ressources indexées

Dans cette partie nous nous intéressons aux ressources de données indexées et leur représentation dans chaque portail.

Caractéristique	**NCBO**	**CISMeF**
Nom du service	NCBO Resource Index [7]	Doc'CISMef [8]
Type de ressources indexées	Documents (e.g. Articles, thèses, etc.), Base de données en ligne (e.g. BioModels http://www.ebi.ac.uk/biomodels-main/).	
Méthode d'indexation	A partir de l'Annotator	Manuellement par les curators et automatiquement avec (FMTI/ECMT)

Services propres à chaque portail

Cette partie détaille les services proposés par chaque portail et qui n'ont pas d'équivalent dans l'autre.

Services propres à chaque portail	**NCBO**	**CISMeF**
Inforoute [9]	Pas d'équivalent.	– Répertorier les moteurs de recherche dans le domaine biomédical. – Expansion sémantique des reqûetes. – Les ressources recherchées ne sont pas indexées.
NCBO Recommender [10]	– A partir d'un texte, décider quelle ontologie est la plus adéquate. – Les ontologies utilisées par ce service proviennent de BioPortal et UMLS.	Pas d'équivalent.

9. http ://inforoute.chu-rouen.fr/ir_site/xsl/index.jsp
10. http ://bioportal.bioontology.org/recommender

Gestion des projets liés	Pour chaque ontologie, la liste des projets liés est précisée (e.g. SNOMED-CT est utilisée dans le NCBO Resource Index [11] , OntoCAT [12] http://bioportal.bioontology.org/ontologies/1353/?p=summary). La liste complète des projets qui utilisent les ontologies de BioPortal est disponible dans l'onglet "Project" (http://bioportal.bioontology.org/projects).	Les organismes et projets impliqués sont listés sous l'onglet "Partenaires" (http://pts.chu-rouen.fr/).

API Web services et Interface utilisateur

Dans cette partie nous présentons les API fournies par chaque portail pour accéder aux différents services ainsi que les interfaces utilisateur. Nous présentons aussi une liste des métriques propres à chaque portail.

Fonctionnalité	**NCBO**	**CISMeF**

11. http ://bioportal.bioontology.org/resource_index
12. http ://www.ontocat.org/

API Web services	API's REST [13] pour : – Services d'accès aux ontologies et aux versions d'ontologies. Exemples : liste des groupes d'ontologies (http://rest.bioontology.org/bioportal/groups?apikey=YourAPIKey), liste des versions d'une ontologie (http://rest.bioontology.org/bioportal/ontologies/versions/1104?apikey=YourAPIKey) – Services d'accès aux vues des ontologies et aux vues des versions d'ontologies. Exemples : télécharger une vue (http://rest.bioontology.org/bioportal/ontologies/download/43072?apikey=YourAPIKey), avoir toutes les versions des vues d'une ontologie (http://rest.bioontology.org/bioportal/views/versions/1104?apikey=YourAPIKey) – Service de recherche. Exemple : recherche dans BioPortal (http://rest.bioontology.org/bioportal/search/?query=Gene&apikey=YourAPIKey) – Service de terme. Exemple : avoir tous les termes d'une version d'ontologie (http://rest.bioontology.org/bioportal/concepts/42431/all?pagesize=50&pagenum=500&apikey=YourAPIKey) – Service de propriété. Exemple : avoir les prpriétés d'une version d'ontologie(http://rest.bioontology.org/bioportal/ontologies/properties/38801?apikey=YourAPIKey) – Service de la hiérarchie. Exemple : avoir le chemin racine/feuilles d'un concept (http://rest.bioontology.org/bioportal/virtual/rootpath/1032/Melanoma?apikey=YourAPIKey) – Service Annotator [14] – Service Recommender [15] – Service Resource Index [16]	Web Service SOAP pour PTS http://pts.chu-rouen.fr/axis2/services/WSInterface?wsdl .

13. http ://www.bioontology.org/wiki/index.php/BioPortal_REST_services
14. http ://www.bioontology.org/wiki/index.php/Annotator_Web_service

Interface utilisateur (UI)	– Partie ontologie : Pour la recherche de termes, l'interface utilisateur est similaire, avec des onglets différents pour afficher les informations liées aux termes (description, hiérarchie, alignements, etc.). La présentation de CISMeF est plus claire et synthétique, BioPortal offre en plus un onglet "visualisation" pour afficher ces informations sous forme de graphe et un onglet "notes" pour voir et/ou ajouter des commentaires. BioPortal offre aussi une page pour la recherche des alignements, projets liés et métadonnées. – Partie données : Les deux portails offrent une UI spécifique pour la recherche de ressources indexées (Doc'CISMeF et NCBO Resource Index) qui retourne les résultat de recherche sous forme de liste de ressources. Doc'CISMeF ne sélectionne pas le concept d'ontologie à utiliser dans la recherche comme le fait le Resource Index grâce à l'auto complétion. Pour une ressource trouvée, les deux services affichent les concepts par lesquels elle est indexée et le lien externe de cette ressource(e.g. "melanoma" http://bioportal.bioontology.org/resource_index/resources/PGDI?conceptids=1032/Melanoma http://doccismef.chu-rouen.fr/dc/#env=basic&n=20&f=1&s=&format=null&lang=fr&wt=true&res=DOC_70217&tab=0&filter=null&objti=DOC&ee=false&q=melanoma). BioPortal offre en plus la possibilité de voir l'annotation du texte original avec le concept et Doc'CISMeF offre une possibilité de recherche dans d'autres ressources non indexées (e.g. PubMed, Google, etc.). Les deux portails offre une UI pour l'annotation sémantique (NCBO Annotator et ECMT). La taille du texte à annoter est plus grande pour le NCBO Annotator qui offre également une page pour la présentation des résulats de l'annotation, ce qui n'est pas disponible pour ECMT.

15. http ://www.bioontology.org/wiki/index.php/Ontology_Recommender_Web_service
16. http ://www.bioontology.org/wiki/index.php/Resource_Index

Métriques [17]	– Ontologies 337 – Concepts 5 841 808 – Synonymes – Définitions – Relations et hiérarchie – Ressources 38 – Ressources indexées 5 125 438 – alignements 2 000 000 – curators – utilisateurs enregistrés –	– Ontologies 45 – Concepts 1 620 000 – Synonymes 3 700 000 – Définitions 220 000 – Relations et hiérarchie 5 400 000 – Ressources – Ressources Indexées 100 000 – alignements 1 600 000 – curators – utilisateurs enregistrés 1 100

17. Nous nous disposons pas de toutes les métriques.

Bibliographie

[1] Bioportal rest services. http://www.bioontology.org/wiki/index.php/BioPortal_REST_services.

[2] Catalogue cismef, contenu et innovation. http://www.chu-rouen.fr/cismef/actes_18_ans/2_pres_18ans_GK.pdf.

[3] Ncbo public wiki. http://www.bioontology.org/wiki/index.php/Main_Page.

[4] Pts : pourquoi ? principales applications. http://www.chu-rouen.fr/cismef/actes_18_ans/3_pres_18ans_PTS_JG.pdf.

[5] Nipun Bhatia, Nigam H. Shah, Daniel L. Rubin, Annie P. Chiang, and Mark A. Musen. Comparing Concept Recognizers for Ontology-Based Indexing : MGREP vs. MetaMap. Research report BMIR-2008-1332, Stanford University, CA, USA, March 2008.

[6] Olivier Bodenreider and Robert Stevens. Bio-ontologies : Current Trends and Future Directions. *Briefings in Bioinformatics*, 7(3) :256–274, August 2006.

[7] Mathieu d'Aquin and Natasha F. Noy. Where to Publish and Find Ontologies ? A Survey of Ontology Libraries. *journal of Web Semantics*, 11 :96–111, March 2012.

[8] SJ Darmoni, S Pereira, A Neveol, P Massari, B Dahamna, C Letord, G Kedelhue, J Piot, A Derville, and B Thirion. French infobutton : an academic and... business perspective. In *AMIA Symp.*, page 920. IOS Press, 2008.

[9] Stéfan Jacques Darmoni, Benoît Thirion, J. P. Leroy, Magaly Douyère, F. Baudic, and J. Piot. Cismef : a structured health resource guide for healthcare professionals and patients. In *RIAO*, pages 819–829, 2000.

[10] Amir Ghazvinian, Natalya F. Noy, and Mark A. Musen. Creating Mappings For Ontologies in Biomedicine : Simple Methods Work. In *American Medical Informatics Association Annual Symposium, AMIA'09*, pages 198–202, Washington DC, USA, November 2009.

[11] Amir Ghazvinian, Natasha F. Noy, Clement Jonquet, Nigam H. Shah, and Mark A. Musen. What Four Million Mappings Can Tell You about Two Hundred Ontologies. In A. Bernstein, D. R. Karger, T. Heath, L. Feigenbaum, D. Maynard, E. Motta, and K. Thirunarayan, editors, *8th International Semantic Web Conference, ISWC'09*,

volume 5823 of *Lecture Notes in Computer Science*, pages 229–242, Washington DC, USA, November 2009. Springer.

[12] Christine Golbreich, Grosjean Julien, and Stéfan Darmoni. FMA and HMTP Portal in OWL : Reconciling Ontology with Terminology in Life Sciences via Metamodeling. Technical report, June 2010.

[13] J Grosjean, T Merabti, N Griffon, B Dahamna, LF Soualmia, and SJ Darmoni. Multi-terminology cross-lingual model to create the health terminology/ontology portal. In *AMIA*, Chicago, 2012.

[14] Julien Grosjean, Tayeb Merabti, Badisse Dahamna, Ivan Kergourlay, Benoit Thirion, Lina F. Soualmia, and Stefan J. Darmoni. Health Multi-Terminology Portal : a semantics added-value for patient safety. In V. Koutkias, J. Nies, S. Jensen, N. Maglaveras, and R. Beuscart, editors, *Patient Safety Informatics - Adverse Drug Events, Human Factors and IT Tools for Patient Medication Safety*, volume 166 of *Studies in Health Technology and Informatics*, pages 129–138. IOS Press, 2011.

[15] Clement Jonquet, Paea LePendu, Sean Falconer, Adrien Coulet, Natalya F. Noy, Mark A. Musen, and Nigam H. Shah. NCBO Resource Index : Ontology-Based Search and Mining of Biomedical Resources. *Web Semantics*, 9(3) :316–324, September 2011. 1st prize of Semantic Web Challenge at the 9th International Semantic Web Conference, ISWC'10, Shanghai, China.

[16] Clement Jonquet, Mark A. Musen, and Nigam H. Shah. Building a Biomedical Ontology Recommender Web Service. *Biomedical Semantics*, 1(S1), June 2010. Selected in Pr. R. Altman's 2011 Year in Review at AMIA TBI.

[17] Clement Jonquet, Nigam H. Shah, and Mark A. Musen. Un service Web pour l'annotation semantique de donnees biomedicales avec des ontologies. In M. Fieschi, P. Staccini, O. Bouhaddou, and C. Lovis, editors, *13emes Journees Francophones d'Informatique Medicale, JFIM'09*, volume 17 of *Informatique et Sante*, Nice, France, April 2009.

[18] G. Kerdelhué. Utilisation du thésaurus MeSH dans le site CISMeF. *Documentaliste - Sciences de l'information*, 44(1), 2007.

[19] Maurizio Lenzerini. Data Integration : A Theoretical Perspective. In L. Popa, editor, *21st ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, PODS'02*, pages 233–246, Madison, WI, USA, June 2002.

[20] Mark A. Musen, Natalya Fridman Noy, Nigam H. Shah, Patricia L. Whetzel, Christopher G. Chute, Margaret-Anne D. Storey, and Barry Smith. The national center for biomedical ontology. *JAMIA*, 19(2) :190–195, 2012.

[21] Natalya F. Noy, Nicholas B. Griffith, and Mark A. Musen. Collecting Community-Based Mappings in an Ontology Repository. In A. P. Sheth, S. Staab, M. Dean,

M. Paolucci, D. Maynard, T. W. Finin, and K. Thirunarayan, editors, *7th International Semantic Web Conference, ISWC'08*, volume 5318 of *Lecture Notes in Computer Science*, pages 371–368, Karlsruhe, Germany, October 2008. Springer.

[22] Natasha F. Noy, Michael Dorf, Nicholas B. Griffith, Csongor Nyulas, and Mark A. Musen. Harnessing the Power of the Community in a Library of Biomedical Ontologies. In T. Clark, J. S. Luciano, M. S. Marshall, E. Prud'hommeaux, and S. Stephens, editors, *Workshop on Semantic Web Applications in Scientific Discourse, SWASD'09*, volume 523 of *CEUR Workshop Proceedings*, page 11, Washington DC, USA, November 2009.

[23] Suzanne Pereira, P. Massari, Antoine Buemi, Badisse Dahamna, E. Serrot, Michel Joubert, and Stéfan Darmoni. F-MTI : outil d'indexation multi-terminologique : application à l'indexation automatique de la SNOMED. In *Risques, technologies de l'information pour les pratiques médicales : comptes rendus des treizièmes journées francophones d'informatique médicale*, pages 57–67, France, 2009.

[24] Axel Reymonet, Jérôme Thomas, and Nathalie Aussenac-Gilles. Modélisation de Ressources Termino-Ontologiques en OWL. In F. Trichet, editor, *Actes des Journées Francophones d'Ingénierie des Connaissances (IC 2007)*, pages 169–180, Grenoble, France, 2007. Cépaduès Editions.

[25] Daniel L. Rubin, Suzanna E. Lewis, Chris J. Mungall, Sima Misra, Monte Westerfield, Michael Ashburner, Ida Sim, Christopher G. Chute, Margaret-Anne Storey, Barry Smith, John Day-Richter, Natalya F. Noy, and Mark A. Musen. National Center for Biomedical Ontology : Advancing Biomedicine through Structured Organization of Scientific Knowledge. *OMICS : A Journal of Integrative Biology*, 10(2) :185–198, June 2006.

[26] S Sakji, Q Gicquel, S Pereira, I Kergoulay, D Proux, Darmoni SJ, and MH Metzger. Evaluation of a french medical multi-terminology indexer for the manual annotation of natural language medical reports of healthcare-associated infections. In *MEDINFO 2010 - Proceedings of the 13th World Congress on Medical Informatics*, volume 160, pages 252–256, Cape Town, South Africa, 2010.

[27] Saoussen Sakji, Catherine Letord, Suzanne Pereira, Badisse Dahamna, Michel Joubert, and Stefan J. Darmoni. Drug Information Portal in Europe : information retrieval with multiple health terminologies. In K-P. Adlassnig, B. Blobel, J. Mantas, and I. Masic, editors, *22th International Conference of the European Federation for Medical Informatics, MIE'09*, volume 150 of *Studies in Health Technology and Informatics*, pages 497–501, Sarajevo, Bosnia, August 2009. IOS Press.

[28] Manuel Salvadores, Matthew Horridge, Paul R. Alexander, Ray W. Fergerson, Mark A. Musen, and Natalya F. Noy. Using sparql to query bioportal ontologies and metadata. In *Proceedings of the 11th international conference on The Semantic*

Web - Volume Part II, ISWC'12, pages 180–195, Berlin, Heidelberg, 2012. Springer-Verlag.

[29] J Shon and M A Musen. The low availability of metadata elements for evaluating the quality of medical information on the world wide web. *Proc AMIA Symp*, pages 945–9, 1999.

[30] Tania Tudorache, Csongor Nyulas, Natalya Fridman Noy, and Mark A. Musen. Web-protégé : A collaborative ontology editor and knowledge acquisition tool for the web. *Semantic Web*, 4(1) :89–99, 2013.

[31] Patricia L. Whetzel, Natalya F. Noy, Nigam H. Shah, Paul R. Alexander, Csongor Nyulas, Tania Tudorache, and Mark A. Musen. BioPortal : enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Research*, 39((web server)) :541–545, June 2011.

[32] Patricia L. Whetzel, Nigam H. Shah, Natalya F. Noy, Benjamin Dai, Michael Dorf, Nicholas B. Griffith, Clement Jonquet, Cherie H. Youn, Chris Callendar, Adrien Coulet, Daniel L. Rubin, Barry Smith, Margaret-Anne Storey, Christopher G. Chute, and Mark A. Musen. BioPortal : Ontologies and Integrated Data Resources at the Click of the Mouse. In B. Smith, editor, *International Conference on Biomedical Ontology, ICBO'09*, page 197, Buffalo, NY, USA, July 2009.

Liste des tableaux

Table des figures