HAL
open science

# I$^2$DEE : intégrer et visualiser des données biologiques pour concevoir une ressource termino-ontologique

Fabien Jalabert, Sylvie Ranwez, Michel Crampes, Vincent Derozier

**To cite this version:**

Fabien Jalabert, Sylvie Ranwez, Michel Crampes, Vincent Derozier. I$^2$DEE : intégrer et visualiser des données biologiques pour concevoir une ressource termino-ontologique. IC - 17èmes Journées francophones d'Ingénierie des Connaissances, Jun 2006, Nantes, France. pp.141-150. hal-01026267

**HAL Id: hal-01026267**

**https://inria.hal.science/hal-01026267**

Submitted on 21 Jul 2014

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# I²DEE : intégrer et visualiser des données biologiques pour concevoir une ressource termino-ontologique

Fabien Jalabert, Sylvie Ranwez, Michel Crampes et Vincent Derozier

Centre de Recherche LGI2P - Ecole des Mînes d'Alès,
Parc Scientifique Georges Besse, F – 30 035, Nîmes Cedex 1,
http://www.lgi2p.ema.fr, {Prénom.Nom}@ema.fr

**Résumé**

Dans différents domaines, le besoin d'organiser et de structurer les données pour améliorer leur exploitation et leur diffusion monopolise de nombreuses équipes de recherche. Au cœur de ces travaux on trouve souvent une ressource terminologique ou ontologique (RTO) spécifique à une application dans le domaine considéré. Cependant la conception de cette RTO ignore trop souvent des données hétérogènes provenant de ressources spécifiques. Dans le domaine biomédical, il peut s'agir de rapports d'actes médicaux, de ressources bibliographiques, mais également de données biologiques issues de bases de données telles que *GOA*, *Gene Ontology* ou encore *KEGG*.

Cet article présente un environnement intégré d'ingénierie ontologique expérimenté dans le domaine de la biologie. Son objectif principal est l'intégration de données hétérogènes dans le processus de conception d'une RTO spécifique à une application donnée. Cet environnement permet, grâce à une chaîne d'analyse et de traitement de données, de filtrer les concepts et relations pertinents pour cette application et de les présenter à l'utilisateur au travers d'une carte de connaissances avec laquelle il peut interagir via une interface personnalisable.

**Mots clés :** Ingénierie des connaissances ; ressources terminologiques et ontologiques ; cartographie de connaissances ; intégration de données hétérogènes ; bioinformatique.

## 1 Introduction

L'expertise, i.e. le savoir et le savoir faire dans un domaine particulier, constitue ce que nous appelons la ***connaissance du domaine***. Cette connaissance désigne les "notions acquises dans ce domaine" (*Petit Larousse*) et peut être stockée informatiquement. Les bases de données remplissent ce rôle de stockage et de partage des connaissances au sein d'une communauté. Les données stockées peuvent être de types très différents. De nombreuses équipes de recherche essaient de satisfaire un besoin d'organisation et de structuration de ces données pour améliorer leur exploitation et leur diffusion. De façon générale, les ressources terminologiques et ontologiques (RTO) sont souvent présentées comme le pilier incontournable de cette structuration, si bien qu'aujourd'hui les notions telles que ontologies, thésaurus et modélisation des connaissances, ainsi que leur représentation visuelle interrogent des communautés scientifiques aussi diverses que les sciences humaines et sociales, la biologie, l'économie ou la productique.

L'ingénierie des connaissances s'intéresse depuis longtemps à la conception de méthodes et d'outils pour assister l'expert dans sa tâche de conception d'une RTO. Une grande part des travaux concerne l'extraction automatique de termes candidats et de relations sémantiques à partir de corpus textuels, laissant à la charge de l'opérateur humain l'affinage des relations et l'organisation de tous ces termes. Le choix du corpus documentaire est déterminant car il doit refléter au mieux la connaissance du domaine. Or ce corpus se limite souvent aux données textuelles. Il serait profitable d'intégrer dans l'ingénierie ontologique des ressources plus diversifiées, issues de données expérimentales, afin que la RTO reflète au mieux la connaissance du domaine.

Dans le domaine biomédical, par exemple, les RTO sont souvent utilisées dans des processus d'analyse (analyse de données d'expression de puces ADN, analyse de pathologies, etc.). Pourtant la réciproque n'est pas vraie : les données expérimentales ne sont pas prises en compte dans la conception de RTO spécifiques. En plus des données textuelles provenant de rapports médicaux et des ressources bibliographiques, souvent tirées de *PubMed/Medline* [16], il serait souhaitable d'intégrer des données complémentaires issues de *KEGG* [10], *GO* [7] ou *GOA* [8], par exemple. Cependant leur hétérogénéité et la difficulté d'en avoir une représentation synthétique expliquent souvent leur absence du corpus.

C'est pour répondre à ce manque que nous avons conçu et développé un environnement intégré d'ingénierie ontologique, I²DEE – *Interactive and Integrated Data Exploration Environment*, que nous avons appliqué à l'ingénierie ontologique. Notre objectif est d'**intégrer des connaissances hétérogènes lors de la conception de RTO pour une application spécifique et de les manipuler graphique-**

**ment.** Notre approche est générique et peut s'appliquer à différents domaines. Cependant nous avons choisi de l'expérimenter dans le domaine des sciences du vivant, ce qui explique certains choix dans la suite de cet article, notamment concernant les types de données intégrés. Nous avons mis en place une chaîne complète d'extraction et de fusion de concepts à partir de différentes bases de données biologiques. Elle permet d'extraire un ensemble d'éléments de connaissance pertinents pour une tâche précise. Pour cela on associe au corpus textuel des données disponibles dans d'autres bases biologiques. Cet ensemble de concepts est ensuite visualisé sur une carte des connaissances personnalisable avec laquelle l'utilisateur peut interagir pour une tâche spécifique.

La section 2 expose notre problématique et l'état de l'art. Nous y détaillons différentes approches et notre positionnement. La section 3 décrit la chaîne de traitements qui permet d'extraire de différentes sources un ensemble de concepts et de relations pertinents. La section 4 montre les résultats obtenus en matière de visualisation et présente l'utilité de l'approche. Nous terminons par un bilan soulevant les limites et les perspectives de nos travaux.

# 2 Problématique et état de l'art

La conception d'une RTO est une tâche qui repose essentiellement sur un expert ou un groupe d'experts et requiert de leur part, à la fois une expertise du domaine et une bonne maîtrise de l'utilisation qui va être faite de la RTO résultante. Ce processus est réalisé en deux temps : tout d'abord la classification ou le regroupement automatique de termes et l'extraction de relations sémantiques (hyperonymie, méronymie, etc.) à partir de méthodes statistiques et linguistiques, ensuite, l'organisation et la structuration des éléments de connaissances extraits. Ces deux étapes peuvent s'appuyer sur des méthodes et outils informatiques présentés dans cet état de l'art.

## 2.1 Extraction de termes candidats

Les travaux d'extraction de termes candidats trouvent leurs origines dans la construction automatique de thésaurus pour la recherche d'information. Les techniques employées alors s'appuient essentiellement sur des mesures statistiques s'intéressant à la distribution des termes dans un corpus. [41] et [30] proposent un état de l'art récent et approfondi de ces techniques. Les mesures les plus courantes sont basées sur la fréquence d'emploi d'un terme dans un document. [45] propose une pondération, TF.IDF (*Term Frequency x Inverse Document Frequency*) permettant de spécifier l'intérêt d'un terme simple au sein d'un corpus.

Suivant l'hypothèse que l'association récurrente de deux termes n'est pas le fruit du hasard, d'autres mesures sont utilisées pour fouiller les associations entre termes ou les regrouper en classes. On parle de **cooccurrence** ou d'**information mutuelle** [26]. Ces mesures déterminent des similarités thématiques entre termes et construisent des réseaux, des thésaurus ou encore effectuent des regroupements par catégories. Des mesures plus spécifiques recherchent des associations de termes contigus et ordonnés afin d'extraire des **unités polylexicales** [32], appelées aussi **segments répétés** [40]. On parle alors de **collocation** dans le cas de deux termes simples [25][38], et de **n-grammes** dans le cas d'une vision prédictive [46].

Ces approches purement statistiques sont justifiées et correspondent à une réalité observée à travers les corpus. Cependant certains travaux cherchent à améliorer la qualité des résultats en s'appuyant sur une analyse linguistique. Un état de l'art est donné dans [24]. Z. Harris propose d'appliquer cette analyse distributionnelle à des unités syntaxiques [36], travaux repris par la suite par [35] et [23]. Différents outils reconnus et utilisés par la communauté scientifique résultent de ces approches, dont *NOMINO* [31], *ACABIT* [29] et *FASTR* [39].

Signalons enfin des approches visant à visualiser des réseaux lexicaux. HyperLex [48] se base sur des cooccurrences pour visualiser les différents contextes d'usage d'un terme polysémique. [44] propose la notion de *contexonymes* pour trouver le terme substituable le mieux adapté à un contexte (dans la cadre de la traduction automatique). Enfin, [34] étudie plus en profondeur la structure et la topologie des graphes suivant une similarité, la *proxémie*.

En biologie l'intégration de différentes ressources est parfois nécessaire, par exemple pour comparer des séquences de gènes ou prédire certaines caractéristiques de gènes ou de protéines. C'est ce qui est à l'origine *d'UMLS,* qui est un médiateur entre différentes ontologies existantes plus ou moins spécifiques et souvent complémentaires [19]. Cependant, à notre connaissance, aucun système d'extraction de connaissances terminologiques n'intègre conjointement des ressources bibliographiques et conceptuelles et des connaissances biologiques comme les données génomiques et protéomiques, les annotations, les voies métaboliques, etc. C'est pourquoi nous avons voulu intégrer ces différentes données dès le début du processus d'extraction. Nous détaillerons dans la section 3 les techniques mises en place pour chaque type de données.

## 2.2 Environnement de structuration et d'édition de RTO

L'ingénieur des connaissances dispose de nombreux outils pour manipuler, organiser, mettre en relation des concepts et exploiter les ontologies résultantes. Pour un inventaire de ces environnements et leurs caractéristiques, on peut consulter [42] ou [27]. Les principales fonctionnalités répertoriées peuvent être regroupées en fonction de différentes motivations :

- respecter des formalismes et / ou méthodologies ;
- proposer des outils d'inférence et de vérification de consistance ;

- tenir compte d'une architecture distribuée et du travail collaboratif ;
- doter l'outil d'une interface conviviale et expressive avec laquelle l'interaction est facilitée.

Dans notre approche ce dernier point est capital. À l'heure actuelle la plupart des environnements proposent une structure arborescente dans laquelle il est difficile de naviguer (souvent, seule la relation *"est un"* est explicitée).

## 2.3 Contexte et positionnement

Notre objectif est clairement identifié : nous souhaitons assister l'expert (un biologiste dans notre cas) dans sa tâche de conception d'une RTO, par l'intermédiaire d'un environnement intégré et interactif d'exploration de données hétérogènes. Les interactions sont réalisées au travers d'une carte des connaissances. Pour cela, il nous faut répondre à plusieurs interrogations qui dirigent l'ensemble de la chaîne de traitements. Ces interrogations concernent soit la nature des données traitées, soit leur visualisation sur la carte.

Concernant leur ***nature***, les questions sont les suivantes. Quelles données sont pertinentes dans le processus d'ingénierie ontologique ? Comment intégrer des données hétérogènes ? Et comment généraliser notre approche à différents domaines ? Concernant la ***visualisation*** des données, rapidement on est confronté à des problèmes de surcharge. Il faut donc répondre aux questions suivantes. Comment filtrer au mieux les informations pertinentes ? Quelles fonctionnalités de l'interface peuvent permettre une manipulation simple et efficace (ajout de ressources pour un contexte précis, etc.) ? Quels outils facilitent la perception des informations (multi-échelle, zoom, distorsion optique et sémantique, etc.) ? Comment exploiter au mieux les différentes relations afin de faciliter la navigation sur la carte ? Enfin, comment prendre en compte la personnalisation de la carte ?

Ces questions sont au cœur de notre démarche. Nous proposons des éléments de réponse pour aller au delà des limites soulevées.

Deux projets de collaboration avec des spécialistes des sciences du vivant, nous ont confirmé l'intérêt que peut représenter une carte des connaissances pour les chercheurs de ce domaine. Le premier concerne l'analyse des données d'expression de puces ADN. Le second concerne la conception d'une RTO servant de support à la recherche d'information et à la veille technologique sur un domaine précis. C'est cette deuxième application qui est détaillée dans cet article même si certains choix au cours de l'implémentation ont été dirigés par la première application. La RTO produite concerne le *Plasmodium Falciparum*, un parasite à l'origine du paludisme (ou *malaria*). Notre objectif est de réunir dans un environnement unique des connaissances hétérogènes afin de suggérer des réponses à un problème donné, soit par association de certaines connaissances, soit par leur proximité sémantique.

# 3 Un environnement intégré d'extraction et fusion de concepts

I$^2$DEE prend en charge l'ensemble de la chaîne de traitements des données, depuis l'extraction à partir de corpus textuels ou de bases de connaissances, jusqu'à la visualisation de la connaissance du domaine. Nous présentons dans cette section une vue générale, puis nous détaillons les différentes données qui sont intégrées et les différents processus qui mènent vers la carte des connaissances.

## 3.1 Présentation générale de I$^2$DEE

Les principales étapes de la chaîne d'analyse de données sont présentées dans FIG. 1. À la suite d'un ensemble de mots-clés sélectionnés par un utilisateur, un moteur de recherche interroge différentes bases de données pour ne retenir dans la masse d'informations que celles qui semblent en rapport avec la requête. Suivent une série d'étapes qui vont permettre d'affiner ce filtrage. En tout premier lieu une *lemmatisation*, utilisant des dictionnaires spécialisés permet d'extraire une série de lemmes qui, après un traitement statistique, vont constituer une *stop-list*. En parallèle une analyse de *cooccurrence* permet de proposer un ensemble de relations. Ces relations sont comparées avec les relations sémantiques présentes dans *UMLS*. On dispose alors d'un ensemble de concepts et de relations avec ces concepts, qui forment un graphe. C'est ce graphe qui est visualisé à l'aide d'une carte. L'utilisateur peut interagir directement avec cette carte en fonction de la tâche qu'il a à accomplir. Par exemple, l'intégration dans cette carte de données de natures différentes lui permet de visualiser des associations, des proximités sémantiques. Il est possible de mettre en relief, par ce biais, une relation forte entre certains gènes et certaines protéines, par exemple. On pourra également identifier les résultats de recherche qui semblent proches de certains thèmes cibles, etc.

Nous allons maintenant détailler ce processus en présentant les différentes données qui sont intégrées dans notre environnement, puis les différentes techniques mises en œuvre pour traiter ces données.

## 3.2 Données à intégrer

Les données partagées par les biologistes sont nombreuses et peuvent être regroupées suivant quatre classes : **ressources bibliographiques**, **ressources conceptuelles**, **bases de données biologiques** et **données expérimentales**. Les outils actuels exploitent majoritairement les ressources bibliographiques et intègrent parfois des ressources conceptuelles. Il nous a semblé pertinent de tenir compte également des connaissances expérimentales relatives au vivant car elles sont plus représentatives de l'étendue des connaissances biologiques.

**Ressources bibliographiques**. Il s'agit ici d'entrepôts d'ouvrages, d'articles scientifiques ou de résumés. Le plus

répandu est *PubMed* qui est partagé par toute la communauté biomédicale [16]. Il propose plus de 14 millions d'articles dont plus de la moitié sont décrits par leurs résumés. D'autres ressources plus spécifiques existent, comme *CISMEF* [4] qui est un portail français, *OMIM* [13] qui synthétise une connaissance biologique axée autour des gènes et des désordres génétiques chez l'homme, ou *Biomed Central* [2] et *PubMed Central* [17] qui proposent le contenu intégral d'articles en HTML et PDF. Ces ressources contiennent non seulement une connaissance bibliographique, mais aussi une grande part de la connaissance biologique. Cette conviction motive de nombreux travaux qui ont déjà aboutis à de multiples résultats [47]. Dans notre chaîne, nous avons privilégié *PubMed*.

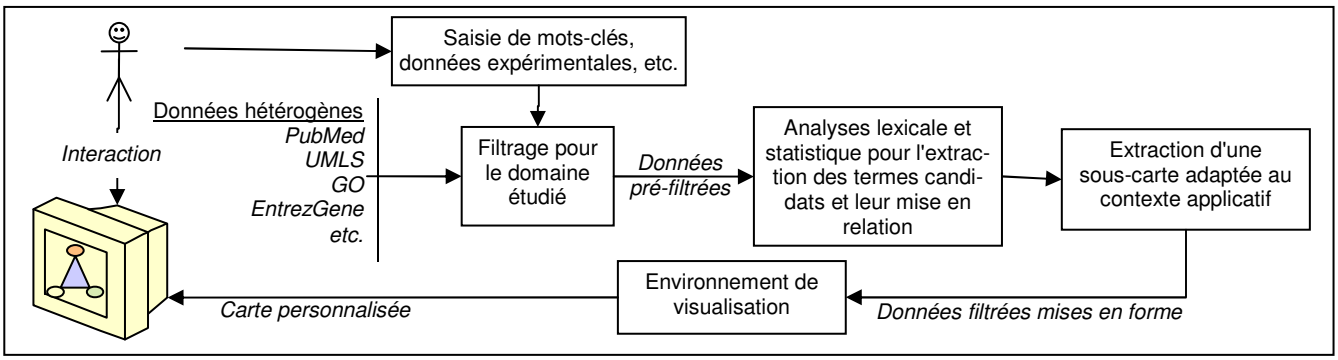



**FIG. 1 –** *Architecture générale de* I²DEE

**Ressources conceptuelles**. La masse croissante de données, particulièrement en biologie avec, par exemple, le décryptage du génome humain, pose le problème de l'accès à ces données. Les chercheurs ont besoin d'outils pour uniformiser des requêtes, indexer des données ou partager des connaissances au sein d'équipes de recherche. De nombreuses "ontologies", plus ou moins spécifiques et plus ou moins formelles ont été conçues. Les principales sont les suivantes. La *GeneOntology* (*GO* – [7]) est utilisée par tous les biologistes pour annoter des gènes et des protéines. Le *MeSH* [12] est employé par *PubMed* et sa traduction française par *CISMEF*. *UMLS* [19] est un entrepôt qui réunit plus d'une centaine d'ontologies de différentes langues dans un *metathesaurus* et facilite ainsi leur interopérabilité, leur alignement ou plus simplement leur réutilisation. Une initiative similaire existe, à une moindre échelle, avec le projet français *TermSciences* [18] qui utilise la traduction du *MeSH*. Cependant, alors que *UMLS* cible les terminologies du biomédical, *TermSciences* s'intéresse à la recherche documentaire dans des thématiques scientifiques plus générales. Dans notre expérience, nous nous somme limités à l'intégration de *UMLS* (qui inclut *GO* et *MeSH* notamment), et de *GODatabase* [1] qui fournit des liens avec des composés chimiques et des gènes, une distance ultramétrique (longueur du plus court chemin entre deux nœuds du graphe), des liens bibliographiques et des relations de synonymie.

**Bases de données biologiques : connaissance partagée du vivant.** Ces ressources sont celles qui posent le plus de problèmes d'hétérogénéité. On peut distinguer des ressources partagées par une grande partie de la communauté des sciences du vivant et d'autres plus spécifiques. Dans la première catégorie, on trouve *UniProt* [20] et *GeneBank* [6], qui répertorient les protéines et les gènes, *KEGG* [10] qui entrepose les voies métaboliques pour différentes espèces et *GOA* [8] qui propose des annotations. Parmi les ressources plus spécifiques, on peut citer, par exemple, *PDB* [14] qui contient les informations structurales sur les protéines. On trouve également des ressources qui normalisent la connaissance partagée, comme *EntrezGene* [5] qui normalise et centralise les gènes, leurs alias, leurs nomenclatures et leurs annotations ainsi que des références bibliographiques vers *PubMed*. D'autres ressources, très spécifiques, centralisent la connaissance d'un sous-domaine, comme *PlasmoDB* [15] pour le *Plasmodium Falciparum*.

I²DEE intègre *EntrezGene*, *KEGG* et *PlasmoDB*.

**Données expérimentales.** La sélection du corpus est une étape décisive. Or, même si l'hétérogénéité des données expérimentales d'une communauté pose un frein, elles sont un excellent point de départ pour constituer un corpus reflétant au mieux la connaissance du domaine de cette communauté. Les travaux du groupe MAGE en témoignent en ce qui concerne l'analyse de données d'expression de gènes [11].

Pour I²DEE nous avons utilisé comme point de départ de nos expérimentations, la liste des gènes présents sur une puce ADN du génome du *Plasmodium Falciparum*.

## 3.3 Résultat attendu

La plupart des travaux à notre connaissance intègrent les données au travers d'un schéma consensus. L'objectif est d'avoir un modèle pivot du domaine permettant de conserver l'expressivité du schéma de chaque ressource. L'extensibilité d'une telle approche reste limitée et l'ingénierie lourde à mettre en place. Par exemple le schéma génomique unifié *GUS* [9], exploité notamment dans *PlasmoDB*, contient plus de 300 tables.

Notre approche se focalise au contraire sur deux points : la souplesse d'extension et une structure adaptée à l'interaction utilisateur (visualiser et manipuler la connaissance en ayant une compréhension de sa structure). Dans ce but, la carte que nous proposons adopte une approche *navigationnelle* sur un graphe typé et valué. L'ajout de nouvelles ressources et les interactions pour la recherche d'information, par exemple, sont simplifiés. Les algorithmes utilisés dans cette approche proviennent de différentes communautés : recherche d'information, analyse de liens et visualisation de grands graphes.

## 3.4 Processus d'intégration

Nous avons choisi d'intégrer les différentes données de façon séquentielle, suivant un ordre déterminé par des dépendances fonctionnelles. *UMLS* est intégré en premier avec *PubMed*. Des bases plus spécifiques sont ensuite intégrées telles que *GODatabase*, *Gene*, *GOA*, *KEGG* et *PlasmoDB*. Les données d'origines sont téléchargées puis intégrées en fonction des formats donnés. La persistance est assurée par un serveur *MySQL*.

La suite de cette section décrit pour chacune de ces ressources le processus d'intégration qui a été mis en place.

**UMLS** est proposé avec l'outil *MetamorphoSys* qui permet de formater un sous-ensemble *d'UMLS* dans un fichier optimisé pour le chargement dans un SGBDR (Système de gestion de bases de données relationnelles). *UMLS* est composé de quatre niveaux de base. À l'aide de cet outil nous avons extrait seulement le niveau le plus haut (les concepts) et le niveau le plus bas (les occurrences – atomes). Nous avons également conservé les relations sémantiques, les types sémantiques et les sources de différents atomes concernés.

*UMLS, chargé au complet, occupe près de 20 Go. Le nombre de concepts obtenus est de l'ordre du million et le nombre d'atomes de l'ordre de 4 à 5 millions.*

**PubMed** est fourni dans un format XML compressé. Nous avons utilisé la librairie *BioText* [3] qui permet d'intégrer les données dans un schéma relationnel. De cette façon nous extrayons de *PubMed* un sous-ensemble de documents qui contiennent un ou plusieurs mots-clés d'une liste donnée ("*microarray*", "*gene expression*", "*plasmodium*", "*malaria*", etc.).

*Ce filtrage permet d'identifier 1% des documents initiaux (qui représentaient 50 Go), soit 120 000 documents et les informations associées : auteurs, mots-clés, composés chimiques, gènes et journaux.*

**GoDatabase** est directement téléchargeable dans des fichiers compatibles avec un SGBDR. L'ontologie étant déjà présente dans *UMLS*, une simple mise à jour des concepts inexistants est réalisée. D'autre part, cette base propose une version précalculée des distances ultramétriques entre concepts. Enfin des définitions, références croisées et relations de synonymie sont présentes.

**Entrez Gene** est proposé dans un format XML. Il est possible d'en télécharger un sous-ensemble concernant un organisme précis, c'est le cas pour *Plamodium Falciparum*. De cette base nous retirons une gestion des alias (différents noms et numéros d'accession d'un même gène), des références bibliographiques (documents *PubMed* décrivant un gène) ainsi que des commentaires divers.

**PlasmoDB** produit les informations les plus à jour concernant les gènes du *Plasmodium Falciparum*. Nous l'utilisons principalement pour consulter les annotations.

**GOA** est, dans notre cas, redondant avec *GO* et PlasmoDB qui fournit une liste à jour de toutes les annotations de gènes. Dans un cas plus général, il permet de consulter les annotations de gènes basées sur *GO* soit dans une approche pluri-espèce, soit quand des portails ne sont pas disponibles dans le domaine étudié.

Enfin, nous intégrons les données expérimentales, provenant ici de fichiers textes comportant les données d'expressions pour l'ORF (Open Reading Frame) de chaque gène. La correspondance entre ORF et gène est obtenue grâce à *Entrez Gene*

## 3.5 Analyse lexicale et distributionnelle

Une fois toutes les données intégrées, nous effectuons une analyse lexicale dont le but est de lister les concepts et entités nommées (gènes, protéines, etc.) présents dans les documents textuels. Ne disposant pas d'un analyseur librement accessible, suffisamment rapide et robuste pour des termes qui ne sont pas uniquement alphabétiques (un nom de molécule peut contenir des virgules, des parenthèses, des chiffres, etc.), nous avons développé un outil permettant de découper le texte en recherchant les segments les plus longs. Il utilise un dictionnaire de lemmes et la liste des concepts, leurs orthographes et prend aussi en compte différentes variations (flexions, présence ou non de virgule, etc.). Le fonctionnement est basé sur un parcours d'arbre à lemmes. Analyser près de 120 000 résumés d'articles ne prend que quelques minutes. Nous n'avons pas encore effectué d'évaluation en profondeur de la précision, mais les premiers tests nous ont montré rapidement qu'on arrivait à retrouver des termes complexes.

Après cette étape on possède une liste ordonnée d'éléments de connaissance pour chaque ressource textuelle utilisée par la suite dans une analyse distributionnelle produisant fréquence, cooccurrence et collocations.

## 3.6 Extraction d'une sous-carte

Les données obtenues après la chaîne de traitements que nous venons de décrire sont encore très volumineuses. Près d'un million de nœuds proviennent *d'UMLS*. *PubMed* représente plusieurs centaines de milliers de nœud supplémentaire, auxquelles peuvent se rajouter encore plusieurs dizaines de millions de nœuds s'il s'agit d'un domaine ou la

littérature est particulièrement fournie. Les gènes, annotations et autres informations provenant de bases de données plus spécifiques sont finalement en nombre plus raisonnable. Si une telle masse de données peut être fouillée automatiquement, il est impensable de les manipuler directement manuellement compte tenu des limites posées par les postes de travail et les outils de visualisation. Il est donc indispensable d'extraire une sous-carte de cet ensemble de données. Cette étape est en grande partie dirigée par l'usage prévu de la carte.

Dans notre cas, nous avons pris comme donnée initiale la liste des 500 gènes présents sur une puce du génome de *Plasmodium Falciparum*. Les nœuds situés à une distance maximale de 2 de ces gènes sont ajoutés. Nous ne prenons pas en compte les relations de cooccurrence, ni les distances ultra métriques qui engendrent une très forte connexité. Le graphe résultant est composé de 500 000 nœuds car la présence des types sémantiques dans UMLS fait que certains concepts sont reliés directement à plusieurs dizaines de milliers de nœuds.

Pour réduire le nombre d'éléments, nous appliquons un algorithme de pondération des nœuds inspirés des travaux décrits dans [43][22]. Les gènes centraux sont initialisés avec un poids de 1, les nœuds à une distance de 1 de ces gènes avec un poids de 0,5, les autres avec un poids nul. En procédant itérativement à partir du centre du graphe, chaque sommet propage sa pondération à ses voisins. Soit $n_i$ un sommet et $n_j$ un de ses voisins :

$$rank(n_j) \leftarrow rank(n_j) + \frac{rank(n_i)}{\deg ree(n_i)}$$

Cette méthode permet de limiter rapidement l'expansion en largeur du graphe. Les valeurs obtenues permettent de limiter le nombre de sommets souhaités par seuillage.

Nous avons ainsi fixé le nombre de nœuds entre 5 000 et 15 000. Ce qui représente un facteur multiplicatif de 10 par rapport au nombre de gènes initial. Ce facteur semble cohérent, un gène possédant généralement autour de 6 annotations et 2 à 5 liens bibliographiques. Enfin, une fois cet ensemble de nœuds déterminé, nous remontons les relations hiérarchiques ("est un", "partie de", etc.) puis nous ajoutons toutes les relations de tout type dans le graphe.

La sous-carte que nous utilisons dans la suite possède plus de 6000 nœuds, dont près de 2000 sont des concepts.

# 4 Interface utilisateur adaptable

Comme nous l'avons dit en introduction, l'objectif est de fournir à différents utilisateurs, pour une tâche particulière (la conception d'une ontologie, par exemple), une carte des connaissances avec laquelle il peuvent interagir de façon conviviale et efficace. Pour cela une adaptation de la carte à un utilisateur et à son contexte est nécessaire. Cette adaptation peut être de deux types : adaptation du contenu de la carte et adaptation des techniques de visualisation. Les données hétérogènes intégrées dans I$^2$DEE ne pouvant être manipulées simultanément, il faut déterminer lesquelles sont utiles et doivent être représentées. Par exemple, dans le cas de l'utilisation par un biologiste, celui-ci ne souhaite pas forcément visualiser les gènes et leurs regroupements. Ceux-ci peuvent ne pas figurer sur la carte. D'autre part, l'environnement doit être adaptable en termes de visualisation. L'utilisateur peut la paramétrer en fonction de ses besoins qui évoluent régulièrement au cours de sa tâche. Des développements ultérieurs permettront au système d'appliquer des fonctions d'apprentissage basées sur les interactions afin de filtrer les informations peu pertinentes.

## 4.1 Techniques de visualisation

La visualisation de la carte doit tenir compte de différentes contraintes : elle représente des connaissances hétérogènes et est employée dans plusieurs applications. Il est donc impossible d'anticiper la forme du graphe représenté et d'utiliser des méthodes spécifiques à certaines topologies. Il faut au contraire une méthode robuste et adaptable. Nous avons opté pour une visualisation dynamique. Pour cela nous utilisons la solution proposée dans [33], qui exploite la notion physique de forces. Le principe est de disposer sur un plan les nœuds d'un graphe et de les relier par des forces (souvent appelées ressorts). La disposition des nœuds et leur répartition dans le plan s'effectuent automatiquement. Certaines forces peuvent être négatives pour repousser certains types de nœuds ; on évite ainsi les chevauchements. D'autres sont positives et vont favoriser les attractions. La sémantique, extraite des bases de données, qui est appliquée à ces forces, permet d'émettre des hypothèses sur les relations entre certains nœuds en fonction de leur proximité sur la carte. De plus, il est possible d'activer ou non, de façon interactive, certains types de forces. La disposition du graphe dépend des relations et peut donc être modifiée au cours de l'exécution. Dans les expériences suivantes, par exemple, nous organiserons le graphe autour de la relation *"partie-de"* puis de la relation *"est-un"*.

Nous avons développé une première version du prototype à partir de la librairie *Prefuse* [37]. Cette dernière permet la visualisation simultanée de plusieurs milliers de nœuds. Elle ne devient saccadée qu'à partir de quelques dizaines de milliers de relations (Celeron M 1.4 GHz). Son architecture permet d'utiliser un filtre de graphe en continu. Le rendu à l'écran est accéléré par la gestion du *double buffering*. Enfin, elle dispose d'une gestion des forces efficace et de multiples fonctionnalités sont présentes : zooms, déplacements, captures d'écran haute définition, affichage des statistiques de rafraîchissements, etc.

Nous avons modifié l'architecture afin d'intégrer plus de souplesse et de performance dans la gestion des propriétés des nœuds et des arêtes du graphe (types, valeurs, etc.). L'inertie présente dans l'intégrateur de forces a été supprimée. Enfin, nous avons introduit un descripteur de type

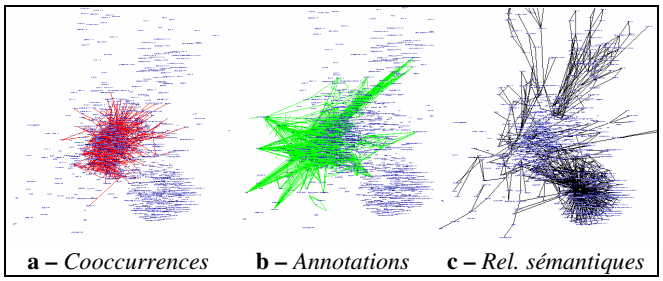
permettant de sélectionner un sous-ensemble de nœuds en fonction d'un critère commun (type de relation, valeur, etc.), ainsi qu'une couche supplémentaire de description des types indépendante des données et de l'application.

Dans la section suivante nous présentons nos résultats et discutons les différentes vues obtenues.

NOTE – *Les captures d'écran présentées tiennent compte du fait que le lecteur peut n'avoir qu'une impression noir et blanc. Cependant le manque de couleurs ainsi que l'impossibilité de faire apparaître l'aspect dynamique et interactif de la carte peuvent altérer la perception des résultats.*

## 4.2 Résultats de la visualisation

Les premières captures d'écran que nous proposons dans FIG. 2 représentent uniquement des concepts (près de 2000). Ici, nous ne visualisons pas les gènes, ni les documents et nous ne prenons pas en compte les relations d'ordre bibliographique. Les 6000 relations en lien avec ces concepts se répartissent en trois types : les **cooccurrences**, les **annotations** et les **relations sémantiques** *("est un"*, "*partie de", etc.).* Les trois captures montrent que la typologie du graphe répartit ces types de relations dans des parties connexes mais relativement distinctes de la carte.



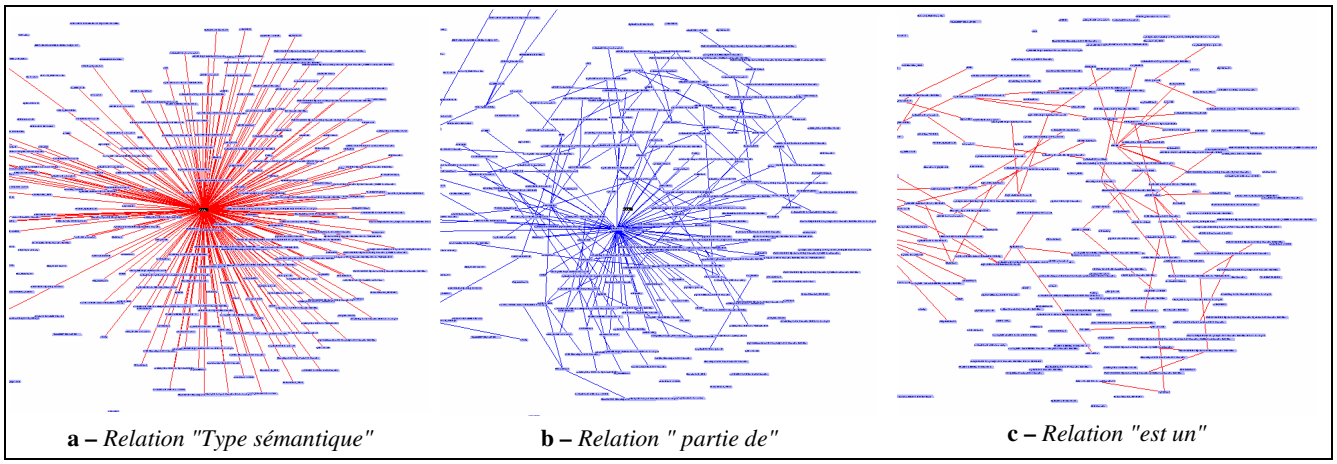
**a –** *Cooccurrences* **b –** *Annotations* **c –** *Rel. sémantiques*

**FIG. 2 –** *Carte des concepts et principales relations présentes.*

La topologie du sous-graphe de cooccurrence (Fig. 2-a) est similaire à celle d'un réseau social, au centre du graphe. Bien qu'ayant une densité d'arêtes faible, ce sous-graphe est généralement difficile à visualiser [21]. Les annotations présentées dans Fig. 2-b occupent aussi une partie bien délimitée du graphe. Elles reflètent l'utilisation des termes dans le contexte expérimental. Enfin, les relations sémantiques de Fig. 2-c sont à la périphérie du graphe, organisées en branches, parmi lesquelles un *hub* est localisé en bas à droite de la carte. Le reste de la carte adopte une forme grillagée dû à la présence de plusieurs hiérarchies qui ne sont que partiellement en accord. Ce *hub* (détaillé dans Fig. 3) est structuré par trois relations. La relation *"type sémantique"*, provenant d'UMLS (Fig. 3-a), a une structure étoilée dont le centre est le concept *cell component*, facilement lisible grâce au zoom. La relation *"partie de"* (FIG. 3-b) est centrée sur *cytoplasma*. Elle est corrélée et décrite dans un arbre plus profond qui laisse supposer une modélisation plus détaillée. Les nombreuses arêtes enchevêtrées résultent de l'action simultanée des forces des relations *"est un"* (FIG. 3-c) et *"partie de"* qui sont transversales. On voit bien ici que la visualisation simultanée de plusieurs types de relation peut gêner la perception et le décodage de la carte.

En réponse, l'utilisateur peut organiser la disposition en fonction d'une relation spécifique. FIG. 4-a organise le *hub* avec la relation *"partie de"*. On peut alors visualiser séparément les trois types de relations sémantiques. FIG. 4-b procède de même pour une organisation avec la relation *"est un"*.

Cette expérience montre que la possibilité d'organiser la carte suivant différentes relations, de façon alternative, est essentielle pour sa bonne interprétation. Plus généralement dans la carte présentée dans FIG. 5, les structures multi-hiérarchiques périphériques du graphe sont le résultat de la relation *"est un"*, Ce qui n'est pas surprenant car il s'agit de la relation la plus utilisée dans les ontologies présentes dans *UMLS*. La relation *"partie de"* reste cloisonnée dans le *hub* comme le montre l'ellipse située en partie droite de FIG. 5. Après une observation microscopique, on constate que les nœuds les plus généraux sont au centre et les plus spécifiques sont à l'extérieur.



**a –** *Relation "Type sémantique"* **b –** *Relation " partie de"* **c –** *Relation "est un"*

**FIG. 3 –** *Agrandissement du hub et des trois relations qui le structurent.*

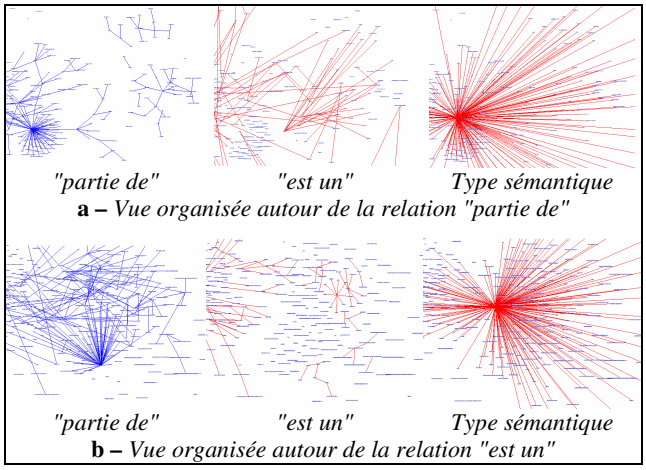



**FIG. 4 –** *Différentes vues de la même carte en fonction de types de relation différents.*

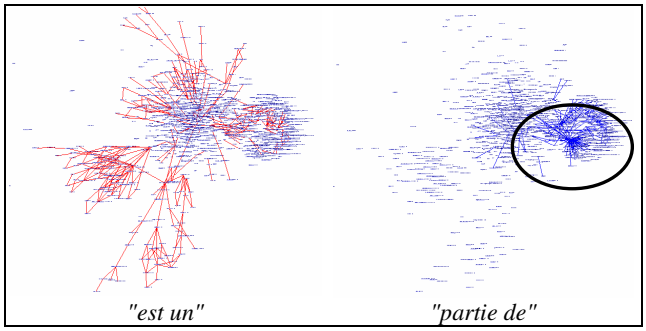



**FIG. 5 –** *Répartition des relations "est un" et "partie de" dans le graphe.*

FIG. 6 détaille une branche structurée par la relation *"est-un"*. Le résultat obtenu a la topologie d'un graphe orienté sans circuit (couramment appelé DAG). Deux raisons peuvent justifier un tel résultat : soit certaines ontologies comme *GO* préfèrent cette topologie à celle d'une arborescence, soit différentes ontologies arborescentes ont un choix différent quant aux propriétés permettant de considérer qu'un concept est hyponyme d'un autre. Dans l'exemple de FIG. 6, *adenosine catabolism* est hyponyme de *purine ribonucleoside catabolism* et de *adenosine metabolism* qui sont tous deux hyponymes de *purine ribonucleoside metabolism.* Pour adopter un des deux points de vue, l'utilisateur peut faire un choix en fonction de son objectif concernant la RTO. Il peut aussi conserver une topologie de DAG et considérer qu'il s'agit de concepts composés par d'autres relations sémantiques (c.f. FIG. 7).

## 4.3 Adaptabilité

Les résultats présentés précédemment montrent à la fois l'intérêt de représenter différents types de données sur une même carte et la nécessité de doter cette carte d'outils de personnalisation afin de faciliter son interprétation.

Dans le cas de la conception d'une RTO, différentes vues en fonction de différents types de relation, permettent de mieux comprendre certains phénomènes et aident l'utilisateur dans sa sélection. Cependant cela soulève deux problèmes :

- L'utilisateur doit constamment alterner les vues en fonction de son besoin. Cette manipulation nécessite une bonne maîtrise de l'environnement et de ses fonctionnalités.
- Même au terme de filtrages successifs, la quantité d'information reste trop grande pour l'utilisateur. Des techniques d'adaptation sont nécessaires pour tenir compte de l'utilisateur, de ses interactions, de ses choix et filtrer automatiquement et progressivement les informations pertinentes.

Dans I$^2$DEE, cette adaptabilité est permise par l'ajout d'un ensemble de critères de sélection, regroupés suivant plusieurs axes. Les critères **structuraux** sont fonction de la situation d'une relation ou d'un noeud dans le graphe (centralité, degré, *hub*, *ranking*, etc.). Les critères **distributionnels** correspondent aux analyses statistiques employées couramment (fréquence, cooccurrence, TF.IDF, approches Harrissiennes, etc.). Les critères **de pertinence** jugent d'une adéquation d'un élément avec le corpus initial (approches vectorielles, etc.). Enfin la **fiabilité** de la source semble importante : certaines ontologies sont considérées plus fiables que d'autres car reconnues par la communauté. Une extraction automatique de relations peut donc être parfois considérée moins fiable qu'une relation importée d'une ontologie conçue par un expert. Les valeurs de ces critères sont dépendantes de la donnée, du corpus, et ne varient pas au cours de l'utilisation.

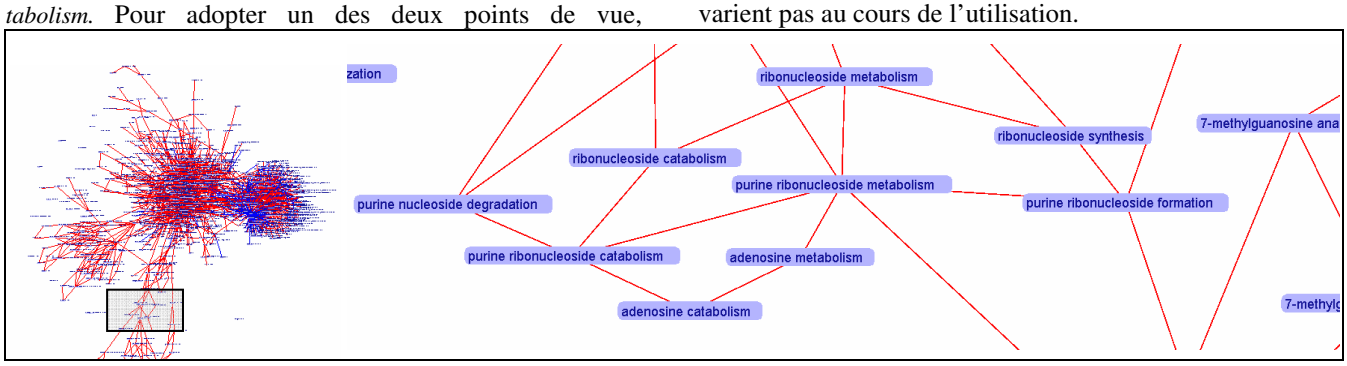



**FIG. 6 –** *Agrandissement d'une branche représentant la relation "est un" dans la carte.*

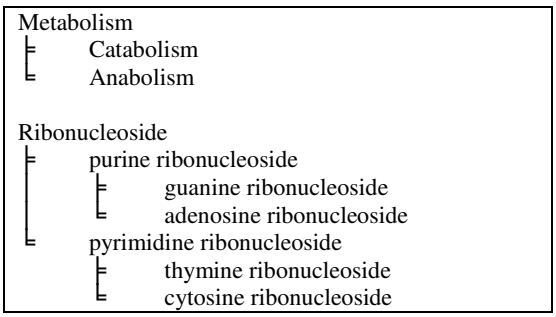



**FIG. 7 –** *Le concept adenosine ribonucleoside peut être considéré comme un concept composé de deux concepts primitifs qui appartiennent chacun à une hiérarchie.*

Pour permettre une adaptabilité de l'interface, nous avons ajouté un dernier critère, celui d'**utilité**. L'hypothèse sur laquelle nous nous basons est la suivante : si l'utilisateur s'intéresse à un élément, il va probablement considérer ses voisins. S'il décide de conserver ou supprimer un élément, et s'il n'effectue aucune action sur un voisin, alors la probabilité d'utilité de cet élément est d'autant plus grande. À partir de ces différents critères, l'utilisateur peut filtrer ou non la donnée à l'aide de deux seuils minimum et maximum et ainsi masquer des termes trop fréquents ou trop rares.

Nous disposons d'un environnement de visualisation de cartes de connaissances, appelé MolAge, doté de nombreuses fonctionnalités d'interaction et de visualisation (zoom, lentilles sémantiques, spectres sémantiques, etc.) [28]. La taille des données traitées dépassait les limites de cet environnement cependant de récentes mises à jour vont permettre prochainement de l'utiliser pour visualiser nos cartes de connaissance à la place de l'application *Prefuse*

### 4.4 Utilisation : limites et perspectives

Une évaluation auprès de biologistes est en cours. Elle révèle leur manque de familiarisation avec des outils d'ingénierie des connaissances. Ils se focalisent constamment sur la connaissance biologique, et non sur les concepts qui la structurent. Bien qu'expert du domaine, le biologiste peut rencontrer des difficultés à comprendre la structure de la RTO proposée et les liens entre concepts. Très souvent derrière une relation entre deux termes c'est la relation biologique qui est recherchée. Rapidement, la priorité est donnée à la proximité des concepts au détriment des relations qui peuvent les réunir. Cependant la cooccurrence, majoritairement représentée, ne fait pas forcément sens pour lui. Il faut y adjoindre des relations complémentaires (synonymie, définitions, contexte d'emploi, etc.) pour leur donner du sens.

De plus l'auto-organisation des concepts peut entraîner des superpositions de plusieurs thématiques et perturber ainsi le biologiste. C'est alors qu'émergent des besoins de filtrage, de coloration, de lentilles sémantiques ou de réorganisation du paysage pour aider à son interprétation. Toutes ces fonctionnalités étant présentes dans MolAge, nous pourrons répondre à ses attentes.

Les premiers retours sont cependant positifs. L'approche semble bonne, et après interprétation de la carte, les relations de cooccurrence s'avèrent refléter des réalités biologiques. Les principales limites concernent donc l'interaction avec la carte et non la pertinence des informations qui y sont représentées. En effet, concernant le *Plasmodium Falciparum*, les imperfections de la carte reflètent l'état des connaissances sur cet organisme : celles-ci sont pauvres et semblent se rapprocher de celles concernant les bactéries plutôt que de celles qui concernent les eucaryotes.

## 5 Conclusion

Cet article a présenté un environnement intégré d'ingénierie ontologique, dont l'originalité est d'intégrer différentes connaissances hétérogènes dans le processus de conception d'une RTO spécifique. Cette dernière est visualisée dans une carte de connaissance interactive qui permet à l'utilisateur d'extraire de nouvelles connaissances, d'effectuer des recoupements ou d'analyser l'information présente.

Une première évaluation a été organisée avec des biologistes. Ils ont soulignés l'intérêt de la prise en compte de différentes sources de connaissances de la conception de la RTO jusque dans sa visualisation. Les principales limites concernent l'interaction avec la carte. Nos travaux continuent donc en ce sens et devraient apporter rapidement de nouveaux résultats.

La principale limite que nous percevons concerne les capacités technologiques des environnements de travail. Les performances sont acceptables sur une station de travail récente mais le resteront-elles si l'on s'intéresse à un domaine où la connaissance est beaucoup plus prolifique que celle disponible pour notre étude (*Plasmodium Falciparum*) ? Nous envisageons donc de confronter $I^2$DEE à d'autres domaines et à d'autres types d'application comme, par exemple, l'analyse de données de puces à ADN.

## 6 Références

### 6.1 Références en ligne

[1] Amigo – Gene Ontology Software and Databases http://www.godatabase.org

[2] BioMed Central – http://www.biomedcentral.com/

[3] BioText - http://biotext.berkeley.edu/

[4] CISMEF, Catalogue et Index des Sites Médicaux Francophones – http://www.cismef.org/

[5] Entrez Gene http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene

[6] Genebank – http://www.ncbi.nlm.nih.gov/Genbank/

[7] GO – the Gene Ontology – http://www.geneontology.org/

[8] GOA – Gene Ontology Annotation http://www.ebi.ac.uk/GOA/index.html

[9] GUS, The Genomics Unified Schema http://www.gusdb.org/

[10] KEGG: Kyoto Encyclopedia of Genes and Genomes http://www.genome.jp/kegg/

[11] MAGE, MicroArray and Gene Expression http://www.mged.org/Workgroups/MAGE/mage.html

[12] MeSH, Medical Subject Headings http://www.nlm.nih.gov/mesh/meshhome.html

[13] OMIM, Online Mendelian Inheritance in Man http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM

[14] PDB, the RCSB Protein DataBank – http://www.rcsb.org/pdb/

[15] PlasmoDB, The Plasmodium Genome Resource http://plasmodb.org/

[16] PubMed – http://www.pubmed.gov

[17] PubMed Central - http://www.pubmedcentral.nih.gov/

[18] TermSciences - http://termsciences.inist.fr/

[19] UMLS, Unified Medical Language System http://www.nlm.nih.gov/research/umls/

[20] Uniprot, The Universal Protein Resource http://www.expasy.uniprot.org/

## 6.2 Références bibliographiques

[21] Bennouas T. Modélisation de Parcours du Web et Calcul de Communautés par Emergence, *Thèse de doctorat*, Université Montpellier 2, 2005.

[22] Borodin, A., Roberts, G. O., Rosenthal, J. S., et Tsaparas, P. Link analysis ranking: algorithms, theory, and experiments. *ACM Tranactions. On Internet Technology (TOIT),* vol. 5:1, pp. 231-297, 2005.

[23] Bourigault D. UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus, *TALN 2002*, Nancy, 2002

[24] Bourigault D. et Jacquemin C. Construction de ressources terminologiques, editor, *Ingénierie des langues*, Hermès, Paris, J-M. Pierrel, 2000.

[25] Choueka Y. Looking for needles in a haystack, *Conference on User-Oriented Context Based Text and Image Handling (RIAO 88)*. Cambridge, MA. 1988.

[26] Church K.W. et Hanks P. Word Association Norms, Mutual Information and Lexicography. *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pp. 76-83, Vancouver, 1989.

[27] Corcho O., Fernández-López M. et Gómez-Pérez A.. Methodologies, tools and languages for building ontologies: where is their meeting point ? *Data Knowledge Engineering*, vol. 46:1, pp. 41-64 : Elsevier Science Publishers B. V., 2003.

[28] Crampes M., Ranwez S., Velickovski F., Mooney C and Mille N. An Integrated Visual Approach for Music Indexing and Dynamic Playlist Composition, *MMCN2006, 13th Annual Multimedia Computing and Networking*, San Jose, California, January 18-19, 2006.

[29] Daille B. Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques. *Thèse de Doctorat en Informatique Fondamentale*. Université Paris 7, 1994.

[30] Daille B. Conceptual structuring through term variations. *ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pp. 9-16, F. Bond, A. Korhonen, D. MacCarthy and A. Villacicencio, 2003.

[31] David S. et Plante P. De la nécessité d'une approche morpho-syntaxique en analyse de textes. *ICO, Intelligence artificielle et sciences cognitives au Québec,* vol. 2:3, pp. 140-151, 1990.

[32] Dias G., Guilloré S. et Lopes J.G.P. Extraction automatique d'associations textuelles à partir de corpora non traités, *5èmes Journées Internationales d'Analyse Statistiques des Données Textuelles, JADT'00*, Lausanne, Suisse, 2000

[33] Eades, P. A heuristic for graph drawing. *Congressus Numerantium 42*, pp. 149-160, 1984.

[34] Gaume, B. Balades aléatoires dans les petits mondes lexicaux*, I3 : Interaction, Information, Intelligence*, vol. 4:2, 2005.

[35] Habert B., Nazarenko A. La syntaxe comme marche-pied de l'acquisition des connaissances : bilan critique d'une expérience. *Actes des Journées d'acquisition des connaissances, JAC 96*, pp. 137-149, 1996.

[36] Harris Z. *Mathematical Structures of Language*, New-York, John Wiley & Sons, 1968.

[37] Heer J. Prefuse: a software framework for interactive information visualization *Masters of Science, Computer Science Division,* University of California, Berkeley, 2004.

[38] Hindle D. et Rooth M. Structural ambiguity and lexical relations, *Computational Linguistics*, Special issue on using large corpora, Vol. 19:1, pp. 103-120, 1993.

[39] Jacquemin C. FASTR : A unification grammar and a parser for terminology extraction from large corpora. *Journées IA'94*, pp. 155-164, Paris, 1994.

[40] Lebart L. et Salem A. *Statistique textuelle*, Dunod, 1988.

[41] Malaisé V. Méthodologie linguistique et terminologique pour l'exploitation d'outils d'extraction terminologique et la constitution d'ontologies différentielles à partir de corpus textuels, *Thèse de doctorat*, Université Technologique de Compiègnes, 19 octobre 2005.

[42] Mizoguchi R. Ontology Engineering Environments, *Handbook on Ontologies*, pp. 175-298, S. Staab et R. Studer, 2004.

[43] Page L., Brin S., Motwani R. et Winograd T. The PageRank Citation Ranking: Bringing Order to the Web, Stanford Digital Library Technologies Project, 1998.

[44] Ploux, S., Ji, H., A model for matching semantic maps between languages (French/English, English/French) *Computational Linguistics*, vol. 29:2, pp. 155-178, 2003.

[45] Salton G. et McGill M.J. *Introduction to Modern Information Retrieval*, McGraw Hill, 1983

[46] Suen C. Y. N-Gram Statistics for Natural Language Understanding and Text Processing, *IEEE Trans. on Pattern Analysisand Machine Intelligence*, vol. PAMI-1:2, pp.164-172, 1979.

[47] Swanson D.R. Fish oil, Raynaud's syndrome, and undiscovered public knowledge, *Perspectives in Biology and-Medicine*, vol 30:1, pp. 7-18, 1986.

[48] Véronis, J. Hyperlex : cartographie lexicale pour la recherche d'informations. *TALN'2003,* pp. 265-274. 2003