Étude et définition d'une ontologie pour l'interprétation des exomes cancéreux

Laurence Gine I Cortiella

MEMOIRE

présenté en vue d'obtenir

le DIPLOME d'INGENIEUR CNAM :

SPECIALITE : INFORMATIQUE

OPTION : INGENIERIE DES SYSTEMES D'INFORMATION

Sujet :

Etude et définition d'une ontologie pour l'interprétation des exomes cancéreux

PAR : GINE I CORTIELLA Laurence

Soutenu le : 15 Janvier 2016

JURY

PRESIDENT : Pr. WEI Anne (Professeur au CNAM Paris)

MEMBRES :

M. MILLAN Thierry (Maitre de conférences - responsable de la filière informatique du CNAM Midi-Pyrénées)

M. DAYRE Pascal (Ingénieur de recherche CNRS - Tuteur)

INVITES :

Mme PAVARD Lise (Responsable du Plateau Recherche et Développement de Sogeti High Tech)

M YEKPE Gérard : Responsable du Bundle Training eLearning

REMERCIEMENTS

J'adresse mes sincères remerciements à toutes les personnes qui ont contribué à la rédaction de ce mémoire.

Mes remerciements vont en premier lieu à M. Pascal DAYRE, mon tuteur de stage dont les remarques et les commentaires m'ont permis de dégager la problématique, construire ma solution et écrire le mémoire.

Je remercie Mme Lise PAVARD qui a rendu possible mon détachement sur le plateau de Recherche et Développement de Sogeti High Tech pour la réalisation de mon stage. Je remercie aussi les membres du projet ICE qui m'ont accueillie durant le projet en particulier M. Arthur JUIN mon maître de Stage pour ses conseils et son soutien.

Enfin je remercie mes collègues de Sogeti High Tech qui ont relu mon mémoire et dont les remarques et conseils m'ont été très précieux.

RESUME : ETUDE ET DEFINITION D'UNE ONTOLOGIE POUR L'INTERPRETATION DES EXOMES CANCEREUX

MEMOIRE D'INGENIEUR DES SYSTEMES D'INFORMATION I.P.S.T C.N.A.M., TOULOUSE 2016

Le projet ICE (Interpretation of Clinical Exom) est un projet ambitieux qui vise à implémenter un outil innovant pour une approche prédictive de la lutte contre le cancer.

Ce projet est mené par un consortium regroupant des partenaires de différents domaines d'expertises dont Sogeti High Tech mon employeur, une société de service informatique qui est en charge de développer l'outil.

C'est un défi qui nous est lancé et qui commande que nous nous appuyons sur les technologies Big Data en raison du volume des données issues du séquençage des gènes et de leurs exploitations dans la lutte personnalisée du cancer. Un ensemble de ressources sont déjà disponibles décrivant les gènes, leurs mutations, les cancers, les molécules actives médicamenteuses, les essais cliniques et leurs publications scientifiques. Ces différents domaines de connaissance sont modélisés sous forme d'ontologies.

L'objectif de ce travail est d'étudier et de concevoir une ontologie pivot alignant les ontologies existantes et les particularités des patients pour permettre l'interprétation clinique des mutations des gènes et une lutte personnalisée du cancer la mieux adaptée pour chaque patient par l'émergence de nouvelles connaissances contextuelles à chaque individu.

Pendant six mois j'ai donc cherché à créer cette ontologie dont la particularité consiste à trouver les solutions pour faire le mapping.

Le mémoire que je vous présente est le résultat de six mois passés au sein de l'Equipe de recherche ICE de Sogeti High Tech pour concevoir cette ontologie pivot et la mise en place des technologies afférentes, fondement du prototype du futur système d'information à l'usage des cliniciens en cancérologie.

MOTS CLES :

Exome, ontologie, Web sématique , RDF, RDFS, OWL, OBO, SPARQL, Mapping

SUMMARY: STUDY AND DEFINITION OF ANTOLOGY FOR CLINICAL EXOM INTERPRETATION

THESIS FOR INFORMATION SYSTEM ENGINEER I.P.S.T C.N.A.M., TOULOUSE 2016

The ICE project (Interpretation of Clinical Exom) is a challenging project that aims at implementing an innovative tool for a predictive approach in cancer medication.

The project is led by a consortium of partners from different fields of expertise among which Sogeti High Tech my employer, an Information Technology service provider in charge of developing the tool.

The challenge resides in the use of Big Data Technologies due to the high volume of data from the sequencing of genes and their usage in gene based personalised medicine in cancer treatment. Some resources are already available other the internet that describe gene, gene variations, cancers, drugs, clinical trials and related publications. These resources are formalised through ontologies.

The task entrusted to me was to study and design an ontology as a backbone to align existing ontologies and patients' specifity for clinical interpretation of gene mutation. This shall therefore enable personalised cancer treatment that suits best a patient thanks to emerging contextual knowledge of each individual.

For six months I have therefore sought to create such ontology whose particularity resides in the search of solutions to map the ICE ontology with existing resources.

The thesis that I present you is the result of six months spent in the Sogeti ICE research team to design such backbone ontology and implementing related technologies serving as the basis of the system under construction for use by cancer doctors.

KEY WORDS:

Exom, ontology, semantic Web, RDF, RDFS, OWL, OBO, SPARQL, Mapping

TABLE DES MATIERES

1 INTRODUCTION

1.1 CONTEXTE

1.1.1 HISTORIQUE DU PROJET ICE

Le projet ICE (Interpretation of Clinical Exom) est un projet de grande envergure né de la conviction d'Integragen qu'il était possible aujourd'hui de passer d'une annotation descriptive des séquences génomiques à un véritable outil d'interprétation clinique en croisant les faits génomiques observés chez les patients avec les données pharmaco-génomiques disponibles via le World Wide Web .

Fort de cette conviction, Integragen approcha Sogeti pour étudier l'opportunité d'une collaboration avec Integragen et l'adéquation des technologies Big Data comme solution aux problématiques exposées par Integragen. Un Proof of Concept est envisagé. Vu le type de données gérées par Integragen, le choix pour la technologie Big Data est fait.

D'autres entités vont bientôt rejoindre SOGETI HIGH TECH et Integragen : Gustave Roussy, l'INSERM (Institut National de la Santé et de la Recherche Médicale) et CARPEM (Cancer Research and PErsonalized Medicine).

Ces organismes vont constituer un consortium qui va adresser pendant un quinquennat la conception de l'outil qui a pour but une approche prédictive de la lutte contre le cancer. En effet ICE devra permettre de guider le praticien dans son diagnostic au travers de la variation des gènes et lui proposer les options thérapeutiques les plus appropriées en relation avec les publications existantes.

Au sein de SOGETI HIGH Tech une équipe dédiée est créée pour mener le projet. Elle est connue sous l'appellation d'incubateur. La direction du projet est localisée à Paris pour la proximité avec les partenaires et les utilisateurs finaux tandis que l'équipe de développeurs est-elle basée à Toulouse.

1.1.2 OBJET DU MEMOIRE

L'outil ICE a pour objet d'assister le biologiste dans l'élaboration de son rapport suite aux résultats du séquençage génomique et d'aider le clinicien à déduire des éléments cliniques pertinents en partant du rapport du biologiste (ces deux utilisateurs ayant des interprétations de différents niveaux).

Ces interprétations sont basées sur le « patrimoine culturel » des experts ainsi que leur domaine de connaissance. Le but donc de cet outil est de formaliser ces domaines de connaissances non pas en offrant une aide trop automatisée qui serait rejetée par les experts, mais de trouver un compromis entre les experts et les utilisateurs « moins-sachant » qui y trouveraient un moyen d'élargir leur niveau de connaissance.

Le World Wide Web héberge des ontologies spécifiques et variées sur des domaines de la génomique, des maladies - notamment des cancers - et les essais cliniques avec des qualités disparates. Tous ces concepts font objets de publications.

La difficulté consiste à rassembler en un seul endroit tous ces concepts afin de :

1. Définir les interactions entre gènes
2. Analyser les impacts entre les variations du gène et les maladies susceptibles d'en découler

Lorsque la maladie est identifiée comme résultante de cette mutation du gène, il faut déterminer les pistes thérapeutiques possibles que ce soit :

1. En termes de traitements déjà sur le marché
2. Ou d'essais cliniques en cours

Pour chacun des points ci-dessus, lorsqu'il existe une documentation sur les sites de référence (PubMed, UMLS, MeSH, etc.), il convient de rattacher cette publication à son objet pour que les utilisateurs puissent avoir une connaissance la plus complète possible du concept ainsi étudié.

Notre mission sur le projet ICE a pour objet de concevoir une ontologie devant permettre une extraction d'informations (texte mining). L'ontologie fonctionnera comme un middleware qui interagira avec les ontologies biomédicales cartographiées alimentant ICE en données et l'application ICE qui doit permettre aux utilisateurs de visualiser les contenus des bases de connaissance existantes.

1.1.3 PROBLEMATIQUE

L'interprétation clinique des **exomes** constituera une avancée majeure dans la médecine prédictive pour les patients atteints de cancers.

- **Définition de l'exome** : L'exome est la partie du génome formée par les exons, c'est à dire les gènes codants pour les protéines. (Bahareh Rabbani, Mustafa Tekin, Nejat Mahdieh, 2014). C'est la partie du génome la plus directement liée au phénotype de l'organisme, à ses qualités structurelles et fonctionnelles. En effet l'exome petite partie du génome (< 2%)

a un intérêt crucial dans la recherche sur les maladies génétiques car elle regroupe 85 % des mutations. Aussi le séquençage de gènes dans le projet ICE cible spécifiquement cette portion du génome.

En effet être en mesure de prédire que telle mutation de gènes, donnera à coup sûr tel type de cancer, et pouvoir proposer une stratégie de soin qui s'adapte au mieux au cas particulier d'un patient donné en s'appuyant sur les dernières avancées connues en matière de soins, est une étape importante dans la lutte contre le cancer.

Les acteurs des industries pharmaceutiques, les cliniciens, les biologistes et la communauté scientifique fondent de grands espoirs sur cette approche innovante.

Les besoins des différents acteurs au regard du projet ICE sont intrinsèquement liés à leurs domaines d'expertises et à leurs activités.

Ainsi un biologiste, pour produire un rapport pertinent, aura besoin de consulter les données du patient. Il fera une revue générale des données biologiques issues du séquençage avant de s'attarder sur l'analyse détaillée du gène mutant. Il s'appuie alors sur la recherche et la consultation de la documentation et des publications disponibles pour produire un rapport expurgé ne contenant que des informations pertinentes pour le clinicien.

Le clinicien souhaite, quant à lui pouvoir visualiser les altérations du gène ; être en mesure de déterminer si cette mutation est associée ou non à un cancer afin d'établir un diagnostic. Il consultera les données sur différents sites de référence tels Clinical Trials (ClinicalTrials.gov) pour les essais cliniques en cours.

Ces cas d'utilisation montrent la nécessité d'intégrer une ontologie dans l'outil ICE afin de :

1. décrire de manière exhaustive :
    - le domaine de connaissance autour des gènes
    - les variations de ces gènes
    - les traitements et leur interaction avec ce qu'il est convenu d'appeler les voies d'interaction (Pathways) par lesquels ils interagissent avec l'organisme.
2. recenser toutes les publications disponibles susceptibles d'intéresser l'utilisateur quel qu'il soit.

C'est donc un vaste chantier que déterminent les spécifications suivantes :

- Un système pertinent représentant le domaine de connaissances tel que perçu par les utilisateurs
- Un système qui fasse le pont entres les ontologies existantes et l'application ICE
- Un système qui agrège les données éparses sur internet pour le présenter aux utilisateurs

Le but ici n'est pas de construire un système supplémentaire mais justement de réutiliser autant que possible l'existant et permettre de construire des requêtes complexes pour interroger ces systèmes.

Les questions suivantes se posent alors :

- Comment faire le maillage et la correspondance entre les classes de notre ontologie et celles des ontologies existantes pour en tirer le maximum de bénéfices ?
- Comment gérer les synonymes car les ontologies existantes ont été créées selon les usages des commanditaires avec des termes différents qui renvoient souvent aux mêmes concepts ?
- Comment s'affranchir de la nécessité d'avoir à télécharger les ontologies existantes de manière systématique ou de permettre une interrogation à distance par ICE des données des ontologies du web ?

Notre travail devra aboutir à cette vision du système en construction :

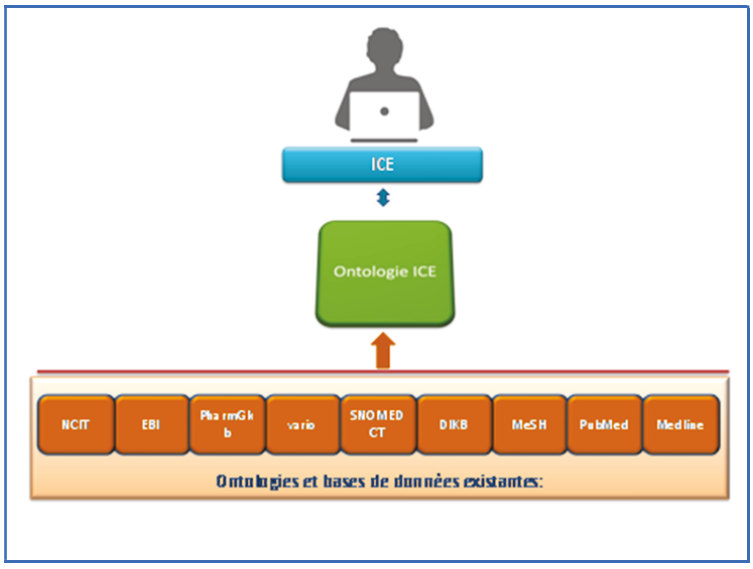


Figure 1 : ICE et les ontologies candidates au mapping

1.1.4 ANALYSE DES RISQUES

ICE est un projet à caractère totalement innovant pour SOGETI High Tech. En effet ce n'est pas un projet classique, en ce sens qu'il n'y pas de client avec un cahier de charges et des spécifications bien définies sur des technologies maîtrisées, mais plutôt un ensemble de partenaires avançant ensemble vers un objectif commun, innovant et ambitieux. Par ailleurs la nouveauté des sujets abordés comporte un certain nombre de risques qu'il convient d'analyser dans leur globalité. Nous distinguons ainsi des risques à la fois techniques, scientifiques, réglementaires et juridiques. De plus alors que la médecine personnalisée commence à entrer dans les mœurs il n'est pas certain que ICE rencontre un marché.

Le diagramme ci-dessous résume les risques pouvant compromettre la réussite du projet à moyen et long terme, et pour lesquels il faudra définir un plan d'action et de suivi.

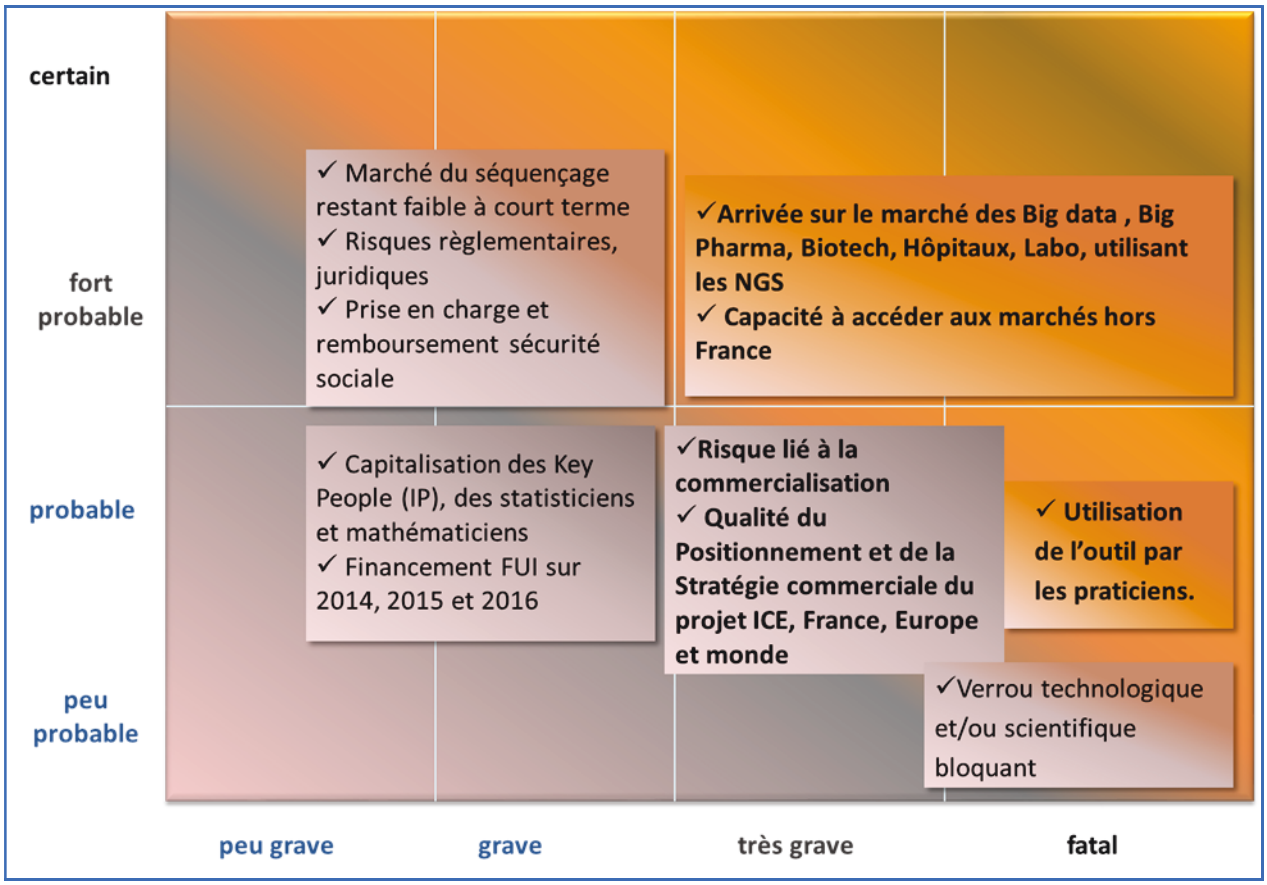


Figure 2 : Matrice des risques du projet ICE

Ces risques vont de ceux qui ont peu de chance de se produire à ceux dont la survenue est certaine. De même leur gravité varie. Ce sont :

1.1.4.1 LES RISQUES PEU OU PROBABLES DE GRAVITE FAIBLE:

- La capitalisation par l'équipe projet sur des profils clés que sont les mathématiciens et les statisticiens
- Financier : ICE est un projet mené par un consortium. Il n'y a pas de recherche de profit. Cependant le financement de cette activité de recherche est réparti entre les différents partenaires

1.1.4.2 LES RISQUES TRES PROBABLES DE GRAVITE FAIBLE

- Un risque juridique. La loi sur la bioéthique n'adresse pas encore le séquençage du génome humain. Des réflexions sont en cours. Aussi le cadre règlementaire reste inconnu à ce stade.
- Un marché restreint : le marché du séquençage génomique est naissant et peu développé à court terme
- Non remboursement des frais par la sécurité sociale : la sécurité sociale ne rembourse pas le coût du séquençage

1.1.4.3 LES RISQUES PEU OU PROBABLES DE GRAVITE FORTE

- Risque lié à la commercialisation : il n'est pas certain que ICE trouve un marché. Auquel cas l'outil ne sera pas viable
- Risque lié à la qualité du positionnement. ICE vise un marché potentiellement mondial. Il s'agit de ne pas se tromper sur la stratégie de commercialisation sur tous les continents

1.1.4.4 LES RISQUES TRES PROBABLES ET DE GRAVITE FORTE

- La concurrence : il y a une vive compétition dans la mise en œuvre de la médecine ciblée génétiquement. De nombreux acteurs sont donc entrés dans la course dont aussi bien des hôpitaux que des laboratoires ou encore des centres de recherche.
- La capacité d'accéder aux marchés extérieurs : le risque encouru étant que ICE ne soit utilisé qu'en France

1.1.4.5 LES RISQUES CERTAINS ET FATALS

En dépit des progrès considérables accomplis au cours de la dernière décennie dans les méthodes d'analyse « à haut débit » (High Throughput Analysis), l'analyse de l'exome à des fins de diagnostic et de thérapie doit faire actuellement face à des verrous scientifiques et technologiques majeurs qui sont listés ci-dessous :

1.1.4.5.1 VERROUS TECHNIQUES

- **Stockage, indexation temps réel et accès aux données** (en particulier dans le Cloud) : Définir un mode de stockage adapté des données telles les séquences brutes, variants, fichiers « full exome » de qualité de séquence
- **Extraction de connaissances, apprentissage et la visualisation de grandes masses de données** : Aider à la mise en place d'un outil de visualisation convivial et simple d'utilisation pour le clinicien qui ne possède pas forcément de fortes connaissances en génétique.
- **Qualité des données, confidentialité et sécurité des données :** Assurer la qualité des données par une méthode d'auto-apprentissage dans un contexte clinique et non de Recherche et Développement.Problèmes de propriété, de droit d'usage, droit à l'oubli : Prédéfinir les contraintes règlementaires de confidentialité, d'accès aux données par les utilisateurs, que faire des données sensibles inutiles**.**

1.1.4.5.2 VERROUS SCIENTIFIQUES

- **Méthode d'analyses avancées** : L'enjeu consiste à analyser les données volumineuses issues du séquençage génomique pour détecter, voir de prévenir une maladie chez un groupe de patients observés. Aussi est-il nécessaire de proposer des méthodes innovantes d'analyse de l'information fournie par ces séquences.
- **Identification de séquences temporelles** dans l'expression des différents gènes qui sont caractéristiques d'une pathologie. Pour identifier ces signatures, les solutions les plus prometteuses reposent sur la mise en œuvre de méthodes d'analyse et de décomposition des signaux (la méthode de statistiques bayésiennes et de simulation stochastique en sont des exemples).
- **Méthode d'annotation avancée des variants** identifiés par ajout de données médicales, génétiques issues de données externes publiques.
- **Interprétation** : Mettre en place un module d'interprétation clinique de la tumeur séquencée du patient (analyse sémantique sur des documents scientifiques,

méthodes mathématiques avancées de corrélation de données algorithme d'interprétation clinique). Ici une ontologie alliée aux méthodes mathématiques trouve son sens.

- **Contrôle Qualité** : Définir un ensemble de contrôle qualités à différents niveaux : qualité des échantillons, séquençage, variant détectés, gestion des données manquantes.

1.1.4.6 ADHESION DES UTILISATEURS FINAUX :

- Les utilisateurs cibles de ICE sont les praticiens et dans une moindre mesure les biologistes. Si l'outil n'obtient pas l'adhésion de ces utilisateurs la pérennité de l'outil est compromise. Il s'agit ici de s'assurer à chaque étape de l'avancement que les fonctionnalités attendues par ces utilisateurs sont bien couvertes par l'application.

Si ces verrous ne sont pas levés le projet échouera.

CONCLUSION

Nous venons de décrire les risques qui menacent le projet ICE dans sa globalité. Il nous faut donc garder en mémoire cette matrice pour qu'à terme elle soit convertie en matrice de succès.

Le chapitre qui suit aborde la méthodologie que j'ai suivie pour mener à bien l'implémentation de l'ontologie comme l'un des composants du projet ICE.

1.2 METHODOLOGIE

La réalisation de la mission qui nous a été confiée s'est faite en trois étapes :

1. Établissement d'un état de l'art
2. Modélisation de l'ontologie
3. Réalisation de différents prototypes pour confronter les résultats aux objectifs fonctionnels et techniques fixés initialement et d'ajuster les développements en fonction de ces résultats.

Notre travail s'est inscrit dans le cadre d'une méthode semi AGILE itérative décrite ci-après :

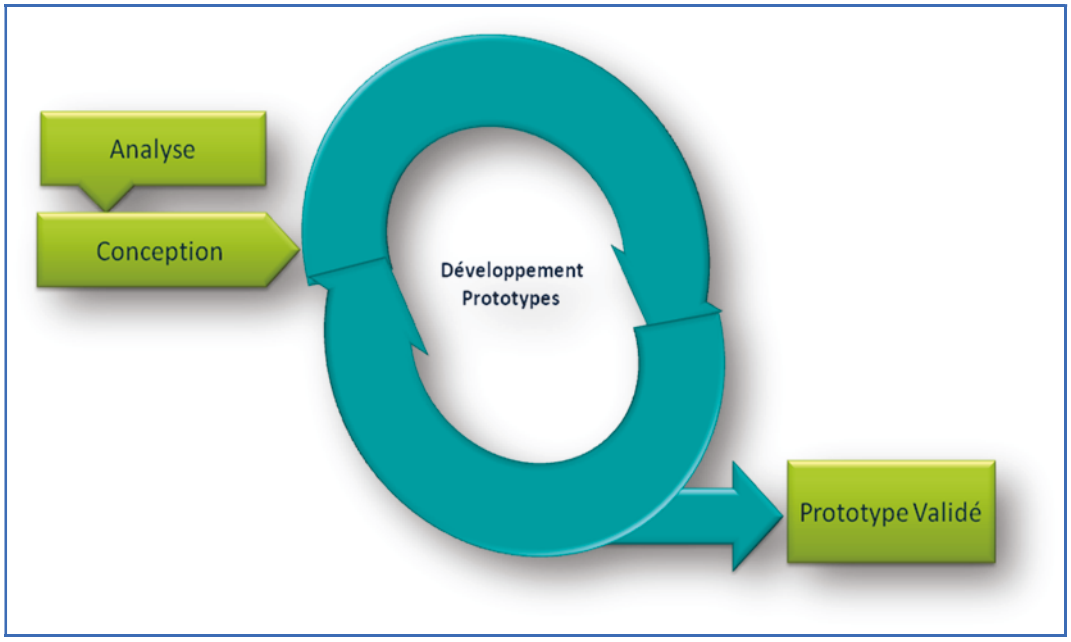


Figure 3 : Méthodologie Agile semi-itérative

1.2.1 ETUDE DE L'ETAT DE L'ART ET ANALYSE DU BESOIN

Le consortium du projet ICE souhaite développer un « outil d'assistance aux choix thérapeutiques basé sur la connaissance des tumeurs cancéreuses en les croisant avec l'ensemble des informations sur les offres pharmaceutiques, … établissant les liens entre leur efficacité et les caractéristiques génomiques du patient ».

L'état de l'art que nous avons effectué et dont nous parlerons plus en détail dans le prochain chapitre nous a permis d'étudier le problème pour comprendre les enjeux, avoir des pistes à explorer pour mettre en œuvre des solutions. Nous avons ainsi étudié :

- les technologies utilisées pour construire les ontologies (langages et outils)
- les ontologies existantes disponibles en rapport avec le domaine fonctionnel (oncologie)

1.2.2 CONCEPTION

Durant cette phase, nous avons modélisé l'ontologie. Il en a découlé un metamodèle que nous présenterons par la suite. Ce modèle a permis de représenter toutes les entités de l'ontologie et de déterminer les relations entre elles.

1.2.3 PROTOTYPAGE

Nous avons implémenté l'ontologie d'abord avec l'outil OBO-EDIT puis nous avons dû basculer sur l'éditeur Protégé du fait de difficultés rencontrées en cours de route.

1.2.4 DEVELOPPEMENT

Cette étape a vu la construction de l'ontologie au travers de différents prototypes itératifs qui ont été confrontés au fur et à mesure au besoin du partenaire et ajustés techniquement et fonctionnellement selon les retours.

1.2.5 VERIFICATION

Le modèle a été soumis à validation aux experts des domaines avant implémentation. Par la suite les différents prototypes ont été soumis à un expert technique qui a validé la véracité du modèle et vérifié que les inférences étaient cohérentes.

1.2.6 VALIDATION :

La validation est faite en interne par l'équipe dédiée SOGETI HIGH TECH travaillant dans l'incubateur[1]

[1] Equipe SOGETI High dédiée aux projets de Recherche et Développement

2 ETAT DE L'ART

Afin de mieux cerner la problématique de la création d'une ontologie biomédicale, il nous fallait dans un premier temps prendre la température de l'existant. Pour ce faire nous nous sommes attelés à rechercher quels sont les formalismes, les technologies disponibles, les langages de description des ontologies et les outils de conception.

Lorsque l'on parle d'ontologie, la notion qui vient en premier est le web sémantique. Nous avons donc cherché à comprendre ce qu'est ce concept et par quel organisme il est régit.

2.1 LE WEB SEMANTIQUE (LINKED DATA, 2014)

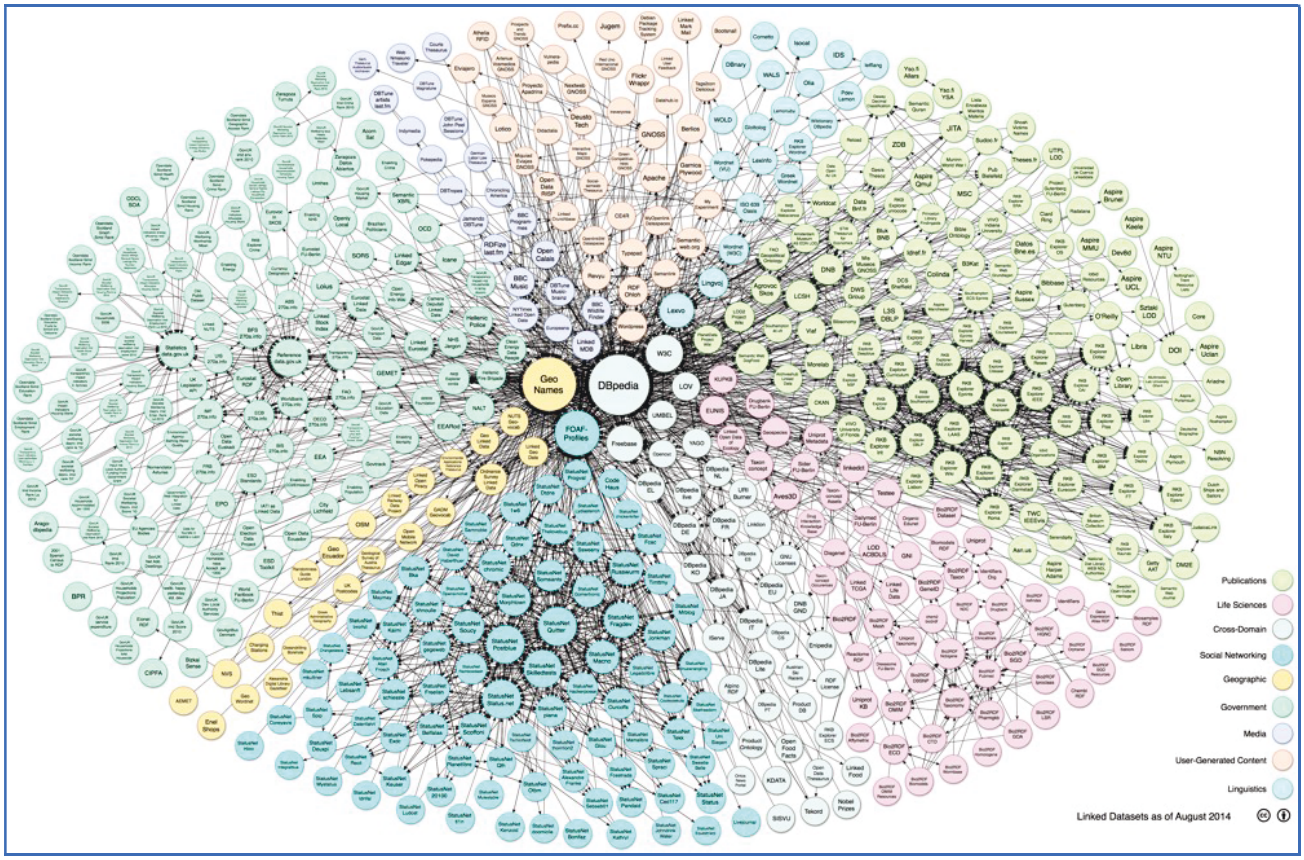


Figure 4 : Linked data/State of the LOD Cloud sept 2011

2.1.1 DEFINITION

Le web sémantique est un mouvement collaboratif qui vise à faire émerger de nouvelles connaissances en s'appuyant sur les données déjà présentes sur internet pour rendre le web intelligent. Selon Tim Berners-Lee, l'inventeur d'internet, il s'agit de « rendre le contenu

sémantique des ressources du web interprétable non seulement par les hommes mais aussi par les machines ».(Berners-Lee, 2001)

Toute cette acquisition de la connaissance par les machines est orchestrée par le World Wide Web Consortium (W3C).

2.1.2 LE WORLD WIDE WEB CONSORTIUM : W3C

Organisme à but non lucratif fondé en octobre 1994, le W3C définit les standards et les normes de compatibilité des technologies d'internet. Cet organisme entend mener le web à son plein potentiel en mettant en œuvre des directives et des protocoles pour assurer la pérennité du web.

Deux principes clés régissent le W3C :

1. Le Web Pour Tous : il s'agit de permettre à chaque être humain de bénéficier pleinement des avantages du web indépendamment des technologies, de la localisation et des capacités mentales ou physiques.
2. Le web en chaque chose : le web doit pouvoir être accédé quel que soit le support (mobile, domotique, télévision connectée …)

2.1.3 VISION

Le W3C souhaite créer un Web participatif, qui promeut le partage des connaissances et entend bâtir la confiance dans son utilisation à l'échelle mondiale.

Par ailleurs le W3C supervise le développement des standards pour le Web. Tous ces standards contribuent à rendre le web plus convivial. Cependant internet possède encore ses limites.

2.1.4 LE WEB AUJOURD'HUI

Bien souvent encore les sites internet sont conçus par des hommes pour des hommes. En effet les machines ne peuvent ni comprendre ni interpréter les contenus du net du fait d'une approche très souvent syntaxique.

L'adressage des sites web se fait au travers d'URLs : Uniform Resource Locators. Ce sont des adresses virtuelles qui permettent d'identifier les ressources du net de manière unique. Cependant ces hyperliens qui permettent de relier les ressources du net ne permettent pas aux machines de donner du sens aux contenus et de réaliser des traitements plus élaborés.

2.1.5 URL, URN, URI

En 1994, Tim Berners-Lee propose la mise en œuvre d'hypertexte (une chaîne de caractères qui renvoie vers un autre texte grâce à un hyperlien). Les technologies HTML (HyperText MarkUp Language) et HTTP (HyperText Transfer Protocol) voient alors le jour. Le premier permet de localiser une ressource, le second permet aux machines de dialoguer pour ainsi afficher la ressource à la demande de l'utilisateur. Les notions d'URL et URN sont ainsi nées. Alors que l'URL va permettre de localiser une ressource et de l'afficher, l'URN permet d'identifier une ressource par son nom dans un espace de nom donné. En creusant un peu plus il apparaît évident, pour leW3C, que les deux notions renvoient à une même problématique à savoir l'identification des ressources sur le net. La publication en juin 1994 de la RFC 1630 de Tim Berners-Lee définissant la syntaxe des URI est admise. Ainsi L'URI (Uniform Resource Identifier) permet d'identifier de manière unique la ressource sur le réseau. L'URI ne sert pas uniquement sur le web mais aussi dans le monde réel comme les codes ISBN ou les codes-barres. De fait le W3C ne fait pas de distinction entre ces trois notions, l'URI englobant URL et URN.

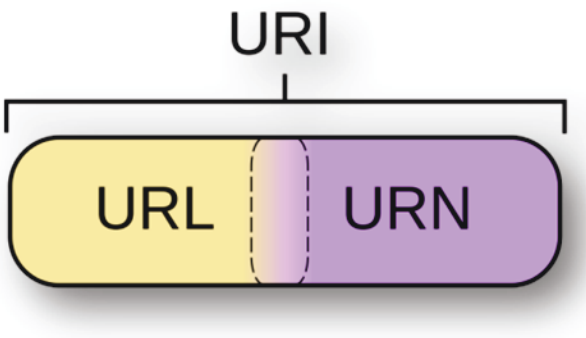


Figure 5 : URI soit une URL, une URN ou les deux

2.1.6 L'ARCHITECTURE DU WEB SEMANTIQUE

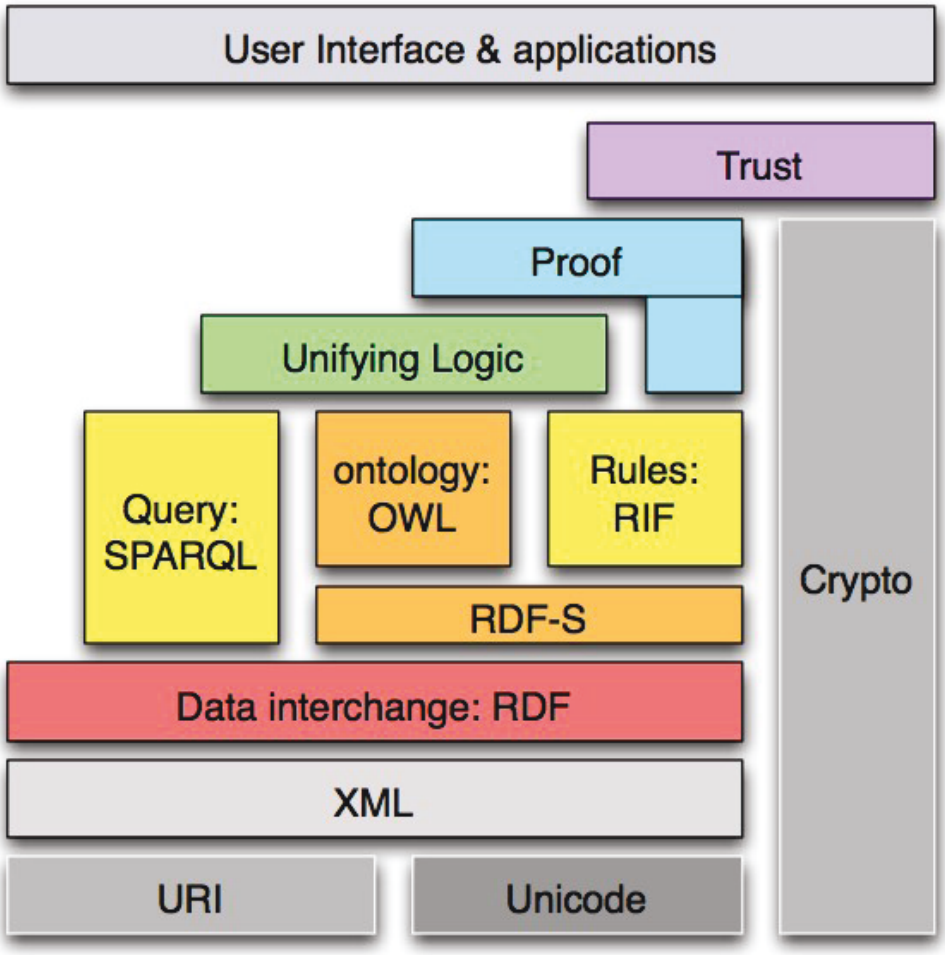


Figure 6 : Piles du Web sémantique (Semantic Web Stack)

L'architecture du Web sémantique est une architecture en couche faisant apparaître les niveaux suivants :

- ***Niveau 1*** : URI et Unicode : cette couche permet d'identifier lisiblement les ressources indépendamment du matériel et des technologies utilisées.
- ***Niveau 2*** : permet d'organiser et d'étiqueter les ressources afin qu'elles soient transmises et traitées en assurant une interopérabilité des systèmes.
- ***Niveau 3*** : permet de décrire les ressources pour leur associer des métadonnées.
- ***Niveaux 4 et 5***: permettent de construire un modèle de connaissances qui soient interprétables par la machine
- ***niveaux 6, 7 et 8***: permettent de définir les règles afin que la machine puisse déduire des raisonnements logiques et pour aboutir à des résultats par inférence.

Le but de cette architecture est de répondre pleinement à la vision du W3C c'est-à-dire instiller de la confiance en assurant la fiabilité des informations à chaque niveau de la couche.

2.2 ET LES ONTOLOGIES DANS TOUT ÇA

Une ontologie est un ensemble structuré de termes et concepts représentant le sens d'un champ d'informations, que ce soit par les métadonnées d'un espace de noms, ou les éléments d'un domaine de connaissances. Selon Gruber, c'est « la spécification explicite d'une conceptualisation », c'est-à-dire qui permet de spécifier dans un langage formel (compréhensible par la machine) les informations d'un domaine et leurs relations. A ce titre, les ontologies intéressent fortement le W3C et les communautés de pratique car pour le premier elles vont permettre de donner du sens au contenu, et pour les seconds de retrouver des contenus selon leur contexte : une profession, une science, un domaine artistique, etc.

Les ontologies décrivent généralement des:

- ***Individus*** : les objets de base
- ***Classes*** : ensembles, collections, ou types d'objets
- ***Attributs*** : propriétés, fonctionnalités, caractéristiques ou paramètres que les objets peuvent posséder et partager
- ***Relations*** : les liens que les objets peuvent avoir entre eux
- ***Événements*** : changements subis par des attributs ou des relations.

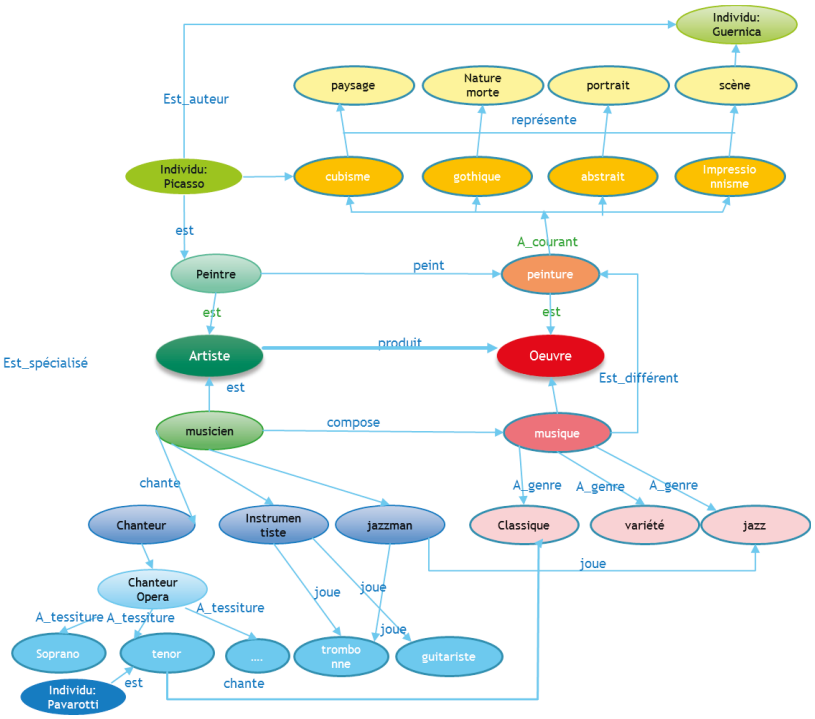


Figure 7 : Illustration d'une ontologie

L'ontologie appelle tous les composants de l'architecture ci-dessus représentée.

2.2.1 RESOURCE DESCRIPTION FRAMEWORK: RDF

Le W3C a édité le standard RDF afin d'offrir un formalisme pour la description des ressources. Ce formalisme permet de décrire les ressources selon un triplet Sujet, Prédicat et Objet

- ***Sujet***: la ressource que l'on décrit
- ***Prédicat***: relation existant entre une ressource et une autre ou une donnée
- ***Objet***: peut être une ressource ou une donnée

Figure 8 : Triplet RDF

Le RDF fait appel à divers langages :

- ***XML*** : il s'agit du format de sérialisation usuel pour le RDF.
- ***N-Triples*** : il s'agit simplement de la liste des triplets (chaque triplet occupe une ligne). Ce format est facile à analyser et à générer mais difficile à interpréter pour les humains.
- ***Turtle*** (Terse RDF Triple Language) : il s'agit d'une forme de sérialisation plus compacte et plus facile à lire que le XML.
- ***Notation3 (ou N3)***: ce format est très proche du Turtle mais il permet également d'exprimer des éléments supplémentaires (non RDF) permettant d'augmenter l'expressivité du document.

2.2.2 RDF SCHEMA: RDFS

Le RDFS est une extension du RDF qui fournit un vocabulaire de modélisation des données RDF. Ce vocabulaire permet de définir les types de ressources (personne, livre, etc.) et leurs propriétés (diplôme, titre, auteur, etc.). Il s'apparente au paradigme objet. Cependant il diffère de ces langages en ce sens qu'il décrit les propriétés sous formes de classes auxquelles elles s'appliquent.

2.2.3 OWL: ONTOLOGY WEB LANGUAGE

Le W3C a mis en œuvre un standard qui est le langage OWL pour modéliser les ontologies. Il permet de décrire des domaines de connaissances complexes : des entités, des groupes d'entités et des relations entre ces entités qui doivent être traités par la machine et non consultés par les humains. Ainsi OWL permet de décrire :

- Les relations entre classes
- La cardinalité
- L'égalité
- Une typographie plus riche des propriétés
- Des caractéristiques de propriétés
- Des classes énumératives.

De nombreuses ontologies existent sous OWL dont le célèbre FOAF qui permet de décrire les relations entre des personnes.

Exemple pour décrire M. Dan Brickley, son adresse mail, son site web ainsi qu'une de ses connaissances :

```
<rdf:RDF

 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"

 xmlns:foaf="http://xmlns.com/foaf/0.1/"

 xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">

<foaf:Person>

<foaf:name>Dan Brickley</foaf:name>

<foaf:title>Mr.</foaf:title>

<foaf:givenName>Dan</foaf:givenName>

<foaf:familyName>Brickley</foaf:familyName>

<foaf:mbox rdf:resource="mailto:webmaster@foaf-project.org"/>
```

<foaf:homepage rdf:resource=" http://danbri.org/"/>

<foaf:knows>

<foaf:Person">

<foaf:name> Libby Miller</foaf:name>

</foaf:Person>

</foaf:knows>

</foaf:Person>

</rdf:RDF>

2.2.4 LES ONTOLOGIES BIOMEDICALES

Depuis quelques années le monde de la biomédecine est en effervescence en ce qui concerne le séquençage du génome humain et les différentes applications thérapeutiques qui peuvent en découler. Parallèlement une course est ouverte quant au laboratoire qui proposera une approche prédictive et ciblée des soins se basant sur la détection de gènes dont la mutation est source de maladies graves et invalidantes comme le cancer. Le 30 janvier 2015, le président des Etats Unis d'Amérique, Barak Obama, annonçait un plan d'envergure dit « Programme de Médecine-Précision Génétique » avec une dotation de 215 millions de dollars qui a pour but de soutenir : « la recherche pour le traitement et la prévention des maladies en prenant en compte les variabilités des gènes de chaque individu, ses facteurs environnementaux et son mode de vie » (Warshaw, 2015). Toute cette science ne peut s'exprimer que si la machine «comprend » et guide les praticiens dans le choix de la stratégie de soins la plus efficiente. Aussi diverses ontologies ont vu le jour. Ces ontologies couvrent différents aspects de la biomédecine chacune se spécialisant dans un domaine particulier comme :

- ***Gene Ontology*** : ontologie décrivant les produits génétiques dans une base de données grâce à un vocabulaire contrôlé
- ***MeSH :*** (Medical Subject Heading) : thésaurus utilisant un vocabulaire contrôlé pour indexer et cataloguer les informations et les publications biomédicales.
- ***PharmGKB*** : base de connaissances sur l'impact de la variation génétique sur la réponse aux médicaments pour les cliniciens et les chercheurs.

- ***ClinVar*** : recueil des données sur les variantes trouvées dans des prélèvements de patients, les assertions faites sur leurs significations cliniques ainsi que les données sur les contributeurs et toutes autres informations pertinentes
- Et bien d'autres encore…

2.2.5 OBO FOUNDRY

Cette initiative a pour but d'aider la communauté scientifique à établir des ontologies dans le domaine biomédical. C'est un groupe de travail collaboratif qui établit les principes pour le développement des ontologies afin de créer un référentiel d'ontologies orthogonales et interopérables.

Pour ce faire OBO Foundry a défini un langage de modélisation des ontologies biomédicales, OBO, et mis en œuvre un outil de création des ontologies : OBO-Edit.

2.2.5.1 OBO VERSUS OWL

OBO-Edit est un logiciel promu par « la communauté biomédicale » alors qu'OWL est un standard du W3C. Tous deux permettent de définir les ressources en utilisant des concepts différents. Ainsi alors que dans OWL on parle de classe, dans OBO il s'agit de type. La notion de sous-classe n'existe pas dans OBO où la relation « is_a » détermine la spécialisation d'une classe OWL. La quantification est obtenue par la relation « part_of ». En exemple nous avons la notation suivante pour décrire un véhicule, une voiture, un habitacle et un volant :

Tableau i : tableau comparatif des langages OBO versus OWL

OBO	OWL
Type : Name : véhicule Definition : Moyen de transport Relation : has_a moteur Lorsqu'on intègre la voiture on a : Name : voiture Relation : is_a véhicule (voiture étant une sous classe de Véhicule) Relation : has_a habitacle (Habitacle étant une propriété de la classe voiture) de même que volant Relation : has_a volant Relation: part_of habitacle	Classe Superclass : véhicule SubClassOF : Voiture Habitacle SubClassOFpart_ofsome Voiture Volant SubClassOF_Part_ofsome habitacle Ici la classe volant est sous-classe de toutes les classes habitacle
OBO s'édite avec OBO-Edit	OWL offre plusieurs choix d'éditeurs dont Protégé est le plus connu

A ce jour OBO Foundry gère les ontologies suivantes :

OBO Foundry ontologies

Title	Domain	Prefix	File
Biological process	biological process	GO	go.obo
Cellular component	anatomy	GO	go.obo
Chemical entities of biological interest	biochemistry	CHEBI	chebi.obo
Molecular function	biological function	GO	go.obo
Ontology for biomedical investigations	experiments	OBI	obi.owl
Phenotypic quality	phenotype	PATO	pato.obo
Plant Ontology	anatomy and development	PO	plant_ontology.obo?view=co
PRotein Ontology (PRO)	proteins	PR	pro.obo
Xenopus anatomy and development	anatomy	XAO	xenopus_anatomy.obo
Zebrafish anatomy and development	anatomy	ZFA	zfa.obo

Figure 10 : Liste des ontologies OBO Foundry

En outre OBO Foundry a recensé plusieurs ontologies qu'elle juge susceptible d'intégrer la liste des ontologies OBO Foundry ou du moins d'intérêt manifeste pour la communauté biomédicale.

OBO Foundry candidate ontologies and other ontologies of interest

Title	Domain	Prefix	File	Last changed
Adverse Event Reporting Ontology	health	AERO	aero.owl	
Anatomical Entity Ontology	anatomy	AEO	aeo.obo	2012/06/01
Ascomycete phenotype ontology	phenotype	APO	ascomycete_phenotype.obo	2014/06/30
Basic Formal Ontology	upper	BFO	1.1	
Beta Cell Genomics Ontology	experiments	BCGO	bcgo.owl	
Biological Collections Ontology		BCO	bco.owl	
Biological imaging methods	experiments	FBbi	image.obo	2013/11/05
Biological Spatial Ontology	anatomy	BSPO	bspo.obo	
BRENDA tissue / enzyme source	anatomy	BTO	BrendaTissueOBO	
C. elegans development	anatomy	WBls	worm_development.obo	
C. elegans gross anatomy	anatomy	WBbt	wbbt.obo	
C. elegans phenotype	phenotype	WBPhenotype	wbphenotype.obo	
Cardiovascular Disease Ontology	health	CVDO	cvdo.owl	
Cell Line Ontology		CLO	clo.owl	
Cell type	anatomy	CL	cl.owl	
Chemical Information Ontology	biochemistry	CHEMINF	cheminf.owl	
Chemical Methods Ontology	health	CHMO	chmo.owl	
Common Anatomy Reference Ontology	anatomy	CARO	caro.obo	2011/12/14
Comparative Data Analysis Ontology		CDAO	cdao.owl	
CranioMaxilloFacial ontology	health	CMF	cmf.owl	
Ctenophore Ontology	anatomy	CTENO	cteno.owl	
Dictyostelium discoideum anatomy	anatomy	DDANAT	dicty_anatomy.obo	2013/01/19
Dictyostelium discoideum phenotype	anatomy	DDPHENO	dicty_phenotypes.obo	
Drosophila development	anatomy	FBdv	fly_development.obo	2010/07/01
Drosophila gross anatomy	anatomy	FBbt	fbbt.obo	
Drosophila Phenotype Ontology		FBcv	dpo.owl	
eagle-i resource ontology	resources	ERO	ero.owl	
Emotion Ontology	health	MFOEM	mfoem.owl	
Environment Ontology	environment	ENVO	envo-basic.obo	
Epidemiology Ontology		EPO	epo.owl	

Figure 11 : OBO Foundry - Ontologies Candidates ou d'intérêt

2.3 STOCKAGE DES TRIPLETS RDF (SEQUEDA, 2013)

Figure 12 : Stockages de RDF - Triplestore

Lorsqu'il s'agit de stocker les triplets RDF, on utilise des systèmes de gestion de bases de données spécifiques dits triplestores. Ils se définissent comme des systèmes de gestion de base de données dédiés au stockage des RDF et permettant de faire des requêtes sur ces triplets grâce au langage SPARQL (Introduction to Triplestores, 2013).

La caractéristique principale des triplestores est la capacité d'inférer (Wikipedia, Triplestore, 2015), c'est-à-dire l'opération par laquelle le système passe d'une assertion considérée comme vraie à une autre assertion au moyen d'un système de règles qui rend cette deuxième assertion également vraie, en un mot : tirer des « conclusions » de manière déductive[2], inductive[3] ou abductive[4] . Ces systèmes permettent (dans leur grande majorité) outre le chargement et le stockage des données, les accès concurrents, la sécurité et les contrôle d'accès ainsi que la récupération et les mises à jour des données.

On distingue trois catégories de Triplestores :

---

2 Type de relation que l'on rencontre en logique mathématique. Elle relie des propositions dites prémisses à une proposition dite conclusion et préserve la vérité.

3 Genre de raisonnement qui se propose de chercher des lois générales à partir de l'observation de faits particuliers, sur une base probabiliste.

4 Raisonnement par lequel on restreint dès le départ le nombre des hypothèses susceptibles d'expliquer un phénomène donné.

- ***Les triplestores natifs*** : ils sont construits spécifiquement pour gérer les triplestores. On compte parmi eux les systèmes suivants : 4Store, AllegroGraph, BigData, Jena TDB, Sesame, Stardog, OWLIM and uRiKa.
- ***Les triplestores adossés***: qui constituent une surcouche RDF au système de gestion de bases de données existantes. Ce sont : Jena SDB, IBM DB2 et Virtuoso.
- ***Les triplestores NoSQL***: les volumes générés par les triplets RDF pouvant rapidement atteindre des niveaux astronomiques, différentes solutions voient actuellement le jour pour stocker du RDF dans des systèmes Big Data NoSQL. Ainsi HBase tout comme CumulusRDF ou encore CouchBase sont apparus sur le marché.

Lorsque nous étudierons le mode de stockage de l'ontologie ICE, nous nous appuierons sur une étude comparative basée sur les systèmes les plus représentatifs de ces trois classes.

2.4 BASE DE DONNEES ORIENTEE GRAPHE

2.4.1 BASE DE DONNEES ORIENTEES GRAPHES : DEFINITION

Une base de données orientée graphe est une base de données orientée objet basée sur la théorie des graphes, donc avec des nœuds et des arcs, permettant de représenter et stocker les données.

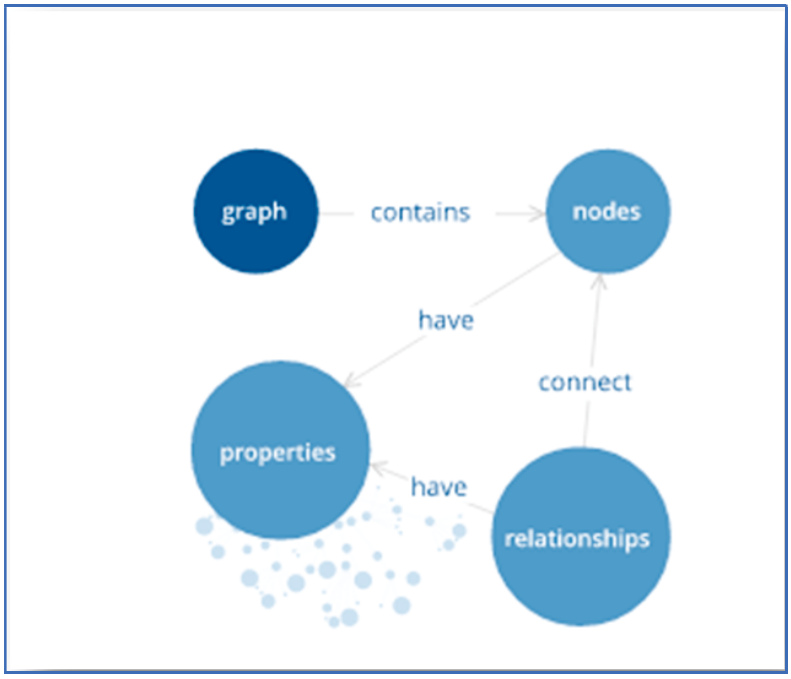


Figure 13 : Base de Données orientée Graphe

Trois notions clés sont à retenir :

- **Le graphe** : il est composé d'un nœud et de relation(s)
- **Le nœud** : chaque entité du graphe avec un identifiant unique
- **La relation** : unidirectionnelle qui lie les nœuds entre eux pour former le graphe.

Ainsi on peut réaliser le graphe ci-après pour une entreprise de revente de voitures comprenant un propriétaire, deux vendeurs et des clients :

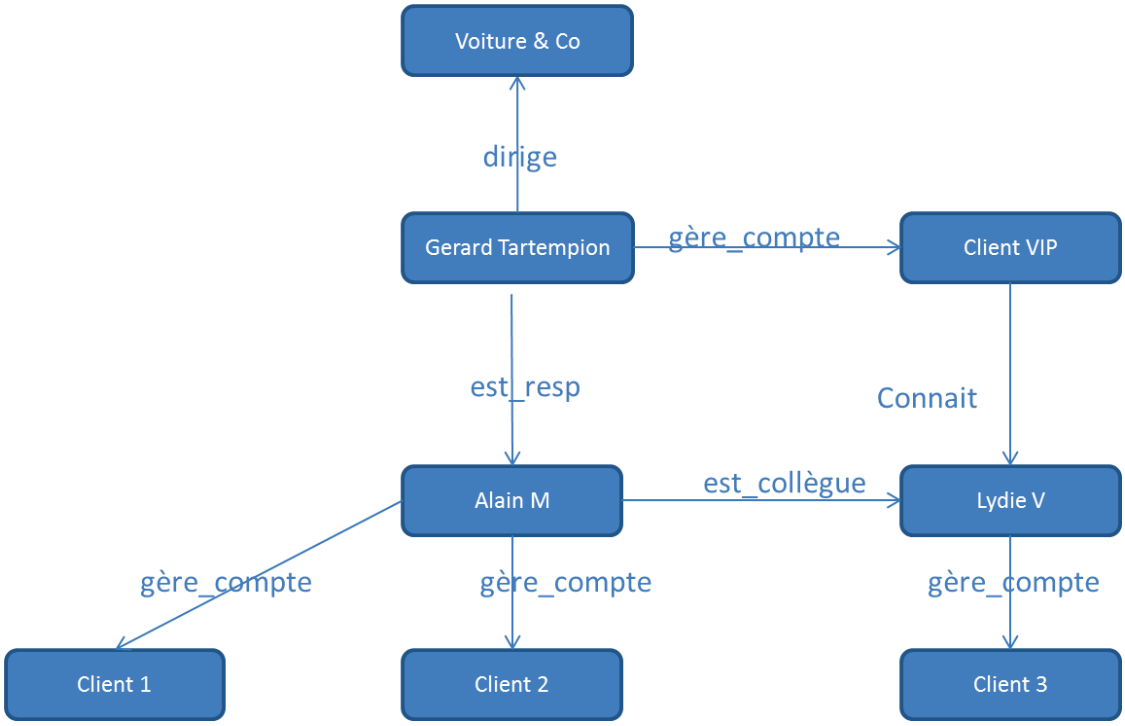


Figure 14 : Illustration base de données orientées graphe

Ces bases de données permettent d'accéder aux éléments voisins par le biais de points physiques dits d'adjacence. Les avantages de ces systèmes sont de grandes performances en ce qui concerne :

- Le traitement des données fortement connectées en évitant les jointures multiples très coûteuses
- La rapidité de temps de réponses pour les lectures en local par le parcours des graphes
- Le développement simple du fait des langages de requêtes dédiés au traitement de données connectées

La modélisation est facilitée du fait de :

- La schématisation non rigide

- La modélisation en langage naturel possible
- La découverte de nouveaux cas d'utilisation par la représentation naturelle des données.

Différents systèmes vont permettre d'implémenter les bases de données. Ainsi nous avons les exemples suivants :

- ***NEO4J***: c'est le plus connu des logiciels de bases de données orientées graphe. Il utilise le langage Cypher pour les requêtes
- ***Infinite Graph***: suite applicative permettant de construire des bases de données NoSQL orientées graphe distribuées
- ***InfoGrid*** : base de données web orientée graphe offrant de nombreux composants logiciels supplémentaires. Il est développé en Java.

2.5 BASE DE DONNEES ORIENTEES GRAPHE OU RDF QUELLE SOLUTION POUR ICE ?

Nous venons de voir deux types de bases de données permettant de stocker de très grands volumes de données. Ces deux systèmes de base de données permettent de gérer des informations connectées. Ils se basent principalement sur les relations que les entités ou classes peuvent avoir entre-elles.

Cependant, ils diffèrent principalement par le mode de requête. Ainsi RDF n'utilise que SPARQL pour faire des requêtes. Les bases de données orientées graphe sont plus polyvalentes et utilisent aussi bien SPARQL que Cypher pour Neo4J par exemple, ou encore Graphlog ou autre BON.

Les bases de données orientées graphe sont optimisées pour parcourir des graphes avec des performances supérieures aux bases de données relationnelles, comme nous l'avons vu plus haut.

Cependant seuls les Triplestores permettent de faire des inférences (tirer des conclusions par déduction, induction ou abduction).

Cette dernière différence nous fait opter pour une solution de création d'une ontologie nous servant de pierre angulaire pour récolter des données présentes sur le web en vue de les faire coïncider à notre modèle pour générer des résultats RDF qui sera chargé dans un système permettant de gérer le format RDF. En clair notre solution est un alignement d'ontologies sur le modèle ICE. Pour ce faire le processus ETL (Extract-Transform-Load) est envisagé.

2.6 EXTRAIRE, TRANSFORMER CHARGER (ETL), UN MOYEN POUR ICE

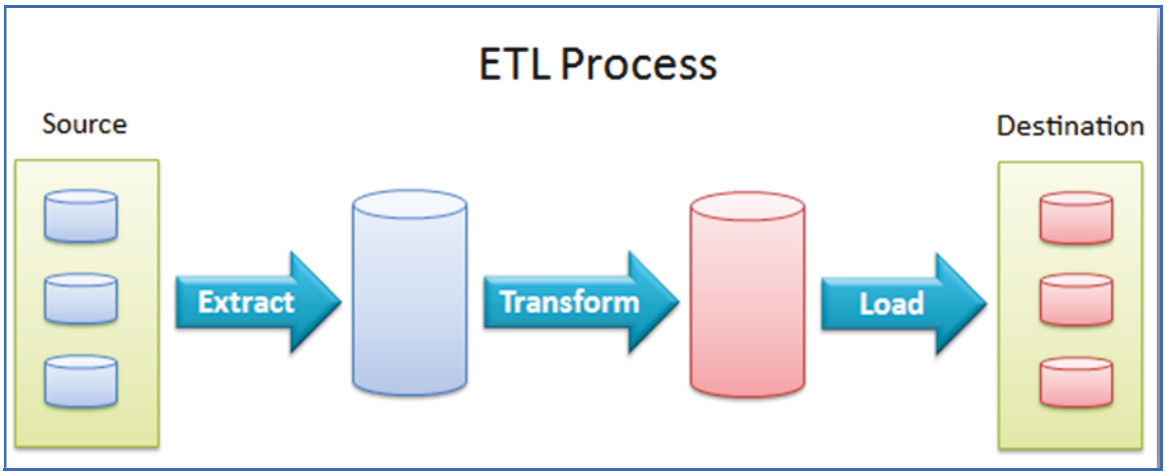


Figure 15 : Le Processus ETL

Le processus Extraire Transformer Charger est une technologie dans la manipulation des bases de données (en trois étapes) qui consiste en l'extraction massive de données en provenance de sources hétérogènes ou homogènes (de manière synchrone) pour les transformer en vue de les intégrer dans une base ou en entrepôt de données cible.

2.6.1 EXTRACTION DE DONNEES

C'est la première étape du processus. Il s'agit d'extraire des données de sources souvent hétérogènes. A cette étape on prend soin de ne pas altérer les systèmes sources. Les données peuvent être consolidées puis préparées pour la prochaine étape.

2.6.2 TRANSFORMATION DES DONNEES

Ici il s'agit de préparer les données pour l'intégration dans le système cible. Les données sont nettoyées, puis filtrées. Elles seront mises au format requis pour enfin appliquer des règles métiers. En sortie les données validées sont alors prêtes à être chargées dans la base ou l'entrepôt de données cible.

2.6.3 CHARGEMENT DES DONNEES

Le but ici est d'intégrer les données extraites puis transformer dans le système cible. Une attention particulière est portée à la qualité et à la cohérence des données. De même les données étant souvent volumineuse, elles sont stockées dans des entrepôts.

Le processus ETL supporte généralement l'ingénierie décisionnelle. Aussi les données du système source peuvent serviront:

- A l'analyse des données
- Au reporting
- À la fouille des données.

Le troisième point est celui qui revêt le plus d'importance pour ICE. En effet les données recueillies sont interrogées pour extraire du sens et établir des correspondances.

2.7 LE RAISONNEMENT EN ONTOLOGIE : PRINCIPES ET OUTILS

La force des ontologies réside dans le fait qu'elles permettent de faire des raisonnements par inférences. Différents outils sur le marché permettent de remplir cette fonction.

2.7.1 INFERER : COMMENT ?

L'inférence se classifie en quatre catégories :

- **La subsomption** : relation hiérarchique entre des concepts. Elle établit donc des liens explicites taxonomiques. Il s'agit d'une relation binaire ***transitive*** faisant appel à la notion de hyponymie–hyperonymie
- **La récupération** : le recensement de toutes les instances d'une classe
- **La réalisation** : recherche de toutes les classes auxquelles appartient une instance
- **La vérification de cohérence** : consistant à s'assurer que l'ontologie n'a pas d'incohérence

Le raisonnement appelle donc deux types de requêtes : les requêtes extensionnelles découlant de la subsomption et de la cohérence alors que les requêtes intentionnelles induites par la récupération et la réalisation requièrent une réflexion en amont sur les données.

Le fonctionnement même des raisonneurs consiste à vérifier la véracité ou la fausseté d'une formule booléenne à partir de variables d'entrée : on parle de Satisfiabilité. Il s'agit de vérifier qu'il n'y a pas de contradiction dans le concept. Le traitement qui en découle se fait donc en deux étapes : une étape de prétraitement ou le raisonneur réécrit l'ontologie dans une optique d'optimisation. Puis vient la phase de classification par subsomption et de vérification de la satisfiabilité.

Deux notions clés sont à l'œuvre lorsqu'on parle de raisonnement : Open World Assumption et le Closed World Assumption.

- L'Open World Assumption: toute information contenue dans l'ontologie est considérée comme incomplète et l'absence d'information ne conduit pas à sa négation. Ainsi le raisonneur répondra à une question dont il n'a pas trouvé de réponse qu'il ne sait. Drummond et Shearer en donne la définition suivante : *« si une proposition ne peut pas*

être prouvée comme vraie avec la connaissance disponible, le système ne peut pour autant conclure que cette proposition est fausse » (Drummond-and-Shearer, 2006)

- Closed World Assumption : dans les systèmes fermés, l'absence d'information est considérée comme information fausse. Ainsi lorsque le raisonneur ne trouve pas une réponse il répondra non.

2.7.2 LES RAISONNEURS DU MARCHE

Il existe une vingtaine de raisonneurs sur le marché: (WIKIPEDIA, 2015). Cependant les plus connus sont :

Tableau ii : Listes des principaux raisonneurs

Raisonneur	Editeur	Technologies	Outil
FaCTT++	The University of Manchester	C++ Open Source	Protege OWL-API
HermiT	University of Oxford		
Pellet	Clark & Parsia, LLC	Java open source et sous licence	Protege
Racer	Concordia University, Montreal, Canada; University of Lübeck, Germany	C++, FACT+ Libre	
Jena	APACHE	Java	OWL API

2.7.3 LA DECIDABILITE

Un problème de décision est dit décidable s'il existe un algorithme, une procédure mécanique qui termine en un nombre fini d'étapes, qui le décide, c'est-à-dire qui réponde par oui ou par non à la question posée par le problème. Ramené aux ontologies « La décidabilité est la capacité à réaliser un raisonnement dans un temps fini et limité. Tout raisonneur doit donc pouvoir faire ces opérations en un temps fini et répondre par oui ou non à une question posée. Cependant seul Protege est complètement décidable comparé aux autres raisonneurs (Fourtineau, 2013)

CONCLUSION DE L'ETAT DE L'ART:

Cette étude nous a permis de balayer toutes les notions autour des ontologies et de voir les standards du W3C ainsi que les outils développés par la communauté Biomédicale.

Nous avons également étudié le mode de stockage des données en analysant les bases de données RDF ou orientés Graphe.

Notre objectif est de construire un système « intelligent » permettant de retrouver une maladie depuis la mutation d'un gène mais aussi de parcourir le chemin inverse en associant par raisonnement un cancer à une variation génomique ou un traitement.

Le chapitre suivant nous permettra de décrire notre travail au cours de cette période de stage.

3 TRAVAUX EFFECTUES

3.1 ANALYSES DES BESOINS

L'outil ICE sera utilisé par deux acteurs:

- Les biologistes
- Les cliniciens

Ces deux utilisateurs ont des besoins différents. Ainsi le rapport biologique est l'élément central de l'outil ICE. Il est le produit de l'interprétation pour le biologiste alors qu'il représente une donnée d'entrée pour l'interprétation du clinicien : Le sens même de la notion d'interprétation est différent selon la catégorie d'utilisateur. Elle porte aussi sur des données différentes : les éléments du rapport pour le clinicien et les résultats de séquençage génomique pour le biologiste.

Le Clinicien quant à lui dispose des données du rapport certifié par le biologiste. Il peut y reconnaître des données attendues (soit par expérience, soit par connaissance du dossier médical) mais il se retrouve fréquemment submergé par une grande quantité d'informations qu'il ne sait pas toujours exploiter efficacement. Paradoxalement , il oscille entre la volonté de visualiser rapidement l'essentiel et celle de pouvoir disposer de tout le savoir connu concernant l'analyse y compris les paramètres d'expérimentation du séquençage qui lui permettront de juger et de comparer des rapports de chronologie et d'origines diverses.

Dans ce contexte, l'outil a pour vocation d'assister le biologiste dans l'élaboration de son rapport alors qu'il doit aider le clinicien à déduire des éléments cliniques pertinents du rapport dans le cas du clinicien. L'analyse de ces besoins nous a permis de modéliser le processus métiers et définir les différents cas d'utilisation

3.2 METHODOLOGIE

Comme évoqué dans l'introduction, notre démarche a été construite selon la méthode AGILE semi-itérative. Nous avons parcouru les étapes suivantes :

1. Analyse : Cette phase nous a amené à étudier le problème. Nous avons ainsi décrit les processus métiers qui sous-tendent l'utilisation de ICE. Nous avons également déterminé les fonctionnalités attendues par chaque acteur au travers de cas d'utilisation

2. Conception : cette phase a été l'occasion de deux activités qui ont structuré la base de connaissance que nous avons établie : il s'agit dans un premier temps de faire une carte mentale. Cette activité a permis de recenser toutes les notions relatives aux :
   - Gènes et leurs mutations
   - Aux maladies et leurs spécialisations
   - Les traitements qui comprennent les essais cliniques et les médicaments mais aussi la prise en charges administrative du patient
   - Le patient et ses données démographiques

3. Développement : cette activité a donné lieu à trois itérations : la première a permis de réaliser un premier prototype sous OBO-Edit. Le deuxième prototype a été réalisé en transférant les données OBO dans l'outil Protégé. Enfin la troisième itération de l'ontologie a été réalisée en la construisant complètement sous Protégé.

4. Validation des prototypes : chaque prototype a été vérifié et validé en interne.

Le schéma ci-dessous décrit cette démarche.

Figure 16 : Méthodologie du Projet

Le paragraphe qui va suivre décrit chaque étape de ce processus d'implémentation de l'outil ICE supportée par cette méthodologie.

3.3 MODELISATION DES PROCESSUS METIERS

Le modèle décrit d'une part les activités du biologiste dans le processus d'interprétation des données du séquençage, et d'autre part pour le clinicien les activités d'analyse des conclusions des biologistes au travers de leur rapport et la recherche d'informations pour poser un diagnostic et proposer une stratégie de soin individualisée.

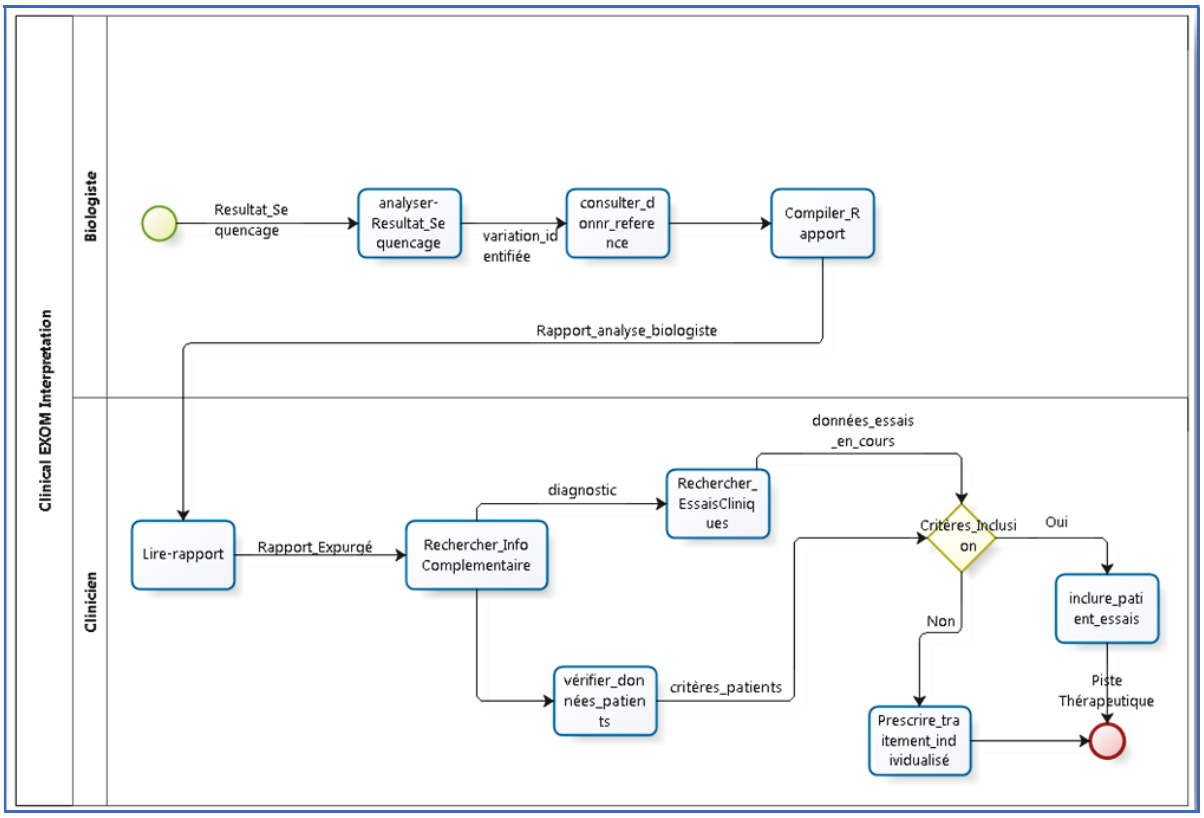


Figure 17 : Processus métier: Interpréter les exomes cliniques

3.3.1.1 PROCESSUS INTERPRETATION DES RESULTATS DE SEQUENÇAGE.

Ce processus nous décrit les activités qu'effectue le biologiste.

Tableau iii : Processus interpréter_résultats_sequençage

Activités	Description	Sortie
Analyser_resultat_sequencage	Le biologiste récupère en entrée les données brutes de séquençage qu'il analyse	Variation_identifié: le biologiste détecte s'il y a des mutations délétères
Consulter_donner_reference	Le biologiste est amené à rechercher des données pour étayer son analyse	Information_complementaires
Compiler_rapport	Rédaction de rapport à l'attention du clinicien	Rapport du biologiste

3.3.1.2 PROCESSUS INTERPRETATION DU RAPPORT D'ANALYSE DE BIOLOGIE

Le clinicien est l'acteur de ce processus. A la réception du rapport du biologiste, il mène ces activités pour pouvoir établir un diagnostic, affiner ces recherches et proposer une piste thérapeutique au patient.

Les activités de ce processus sont décrites dans le tableau ci-dessous :

Tableau iv : processus d'interprétation du rapport du biologiste

Activités	Description	Sortie
Lire_rapport	A la réception du rapport du biologiste, le clinicien se doit de le lire pour en extraire les informations pertinentes	Rapport_Expurgé
Rechercher_Info Complementaire	Le clinicien peut rechercher des informations supplémentaires sur les gènes délétères ou les maladies résultantes	Diagnostic
Rechercher_EssaisCliniques	Le clinicien peut rechercher les données sur les essais cliniques sur le cas qui l'occupe.	données_essais_en_cours
vérifier_données_patients	L'état du patient ainsi que les données démographiques (comme l'âge ou le sexe) sont pris en compte pour orienter la décision de traitement	critères_patients
Prescrire_traitement_individualisé	Si le patient ne remplit pas les conditions, le clinicien propose un traitement individualisé	traitement_personnalisé
inclure_patient_essais	Le patient qui remplit les conditions est inclus dans un essai clinique	piste_thérapeutique

Les activités ainsi décrites donnent lieu au cas d'utilisation ci-dessous.

3.3.2 LES CAS D'UTILISATIONS DES ACTEURS

Nous décrivons pour chaque acteur les cas d'utilisations. Le chapitre ci-dessous nous permet de décrire les fonctionnalités de l'outil ICE par le biologiste

3.3.2.1 CAS D'UTILISATION DU BIOLOGISTE

Le diagramme ci-dessous montre les cas d'utilisation du biologiste :

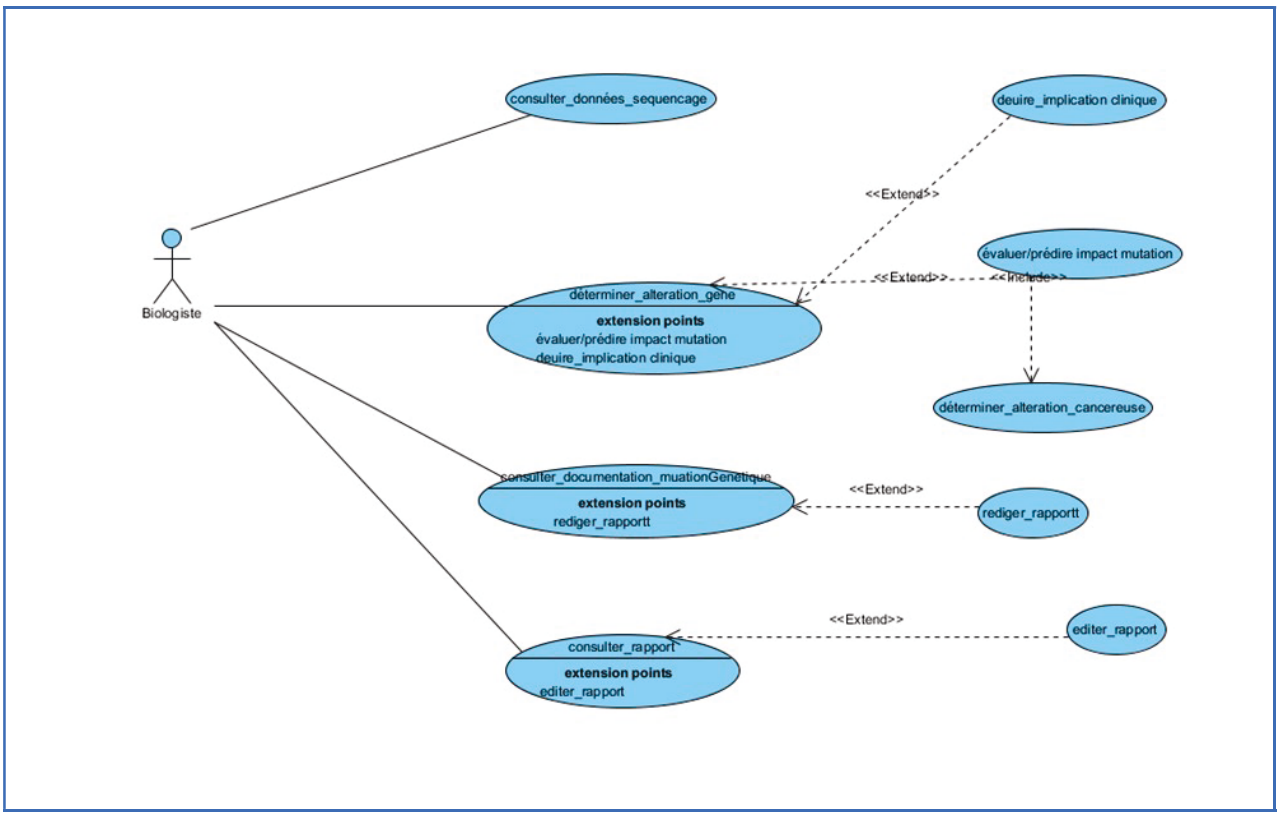


Figure 18 : Diagramme des cas d'utilisation du biologiste

Le tableau ci-dessous détaille chaque cas d'utilisation du biologiste.

Tableau v : Description des cas d'utilisation du biologiste

Cas d'utilisation	Description
Consulter_données_sequencage	Le biologiste reçoit et analyse les données de séquençage génomique
Déterminer_altération_gene	À l'analyse le biologiste peut déterminer s'il existe des mutations génétiques ou non
Deduire_implication_clinique	Si le biologiste rencontre des variations, il peut en déduire les implications cliniques
Evaluer/predire_impact_mutation	Lorsque le biologiste a déterminé des variations génétiques, il peut évaluer ou prédire l'impact de la mutation
Determiner_alteration_cancereuse	Après avoir évalué les impacts de la variation génétique, le biologiste est à même de prédire les cancers causés par les altérations.
rediger_rapport	Le biologiste rédige son rapport
Consulter_documentation_mutationGénétique	Le biologiste pour étayer son rapport consulte la documentation sur les mutations qu'il a identifiées
Consulter_rapport	Le biologiste peut consulter les rapports que lui ou ses collègues ont rédigés par le passé
Editer_rapport	L'action d'éditer le rapport implique de le consulter

Figure 19 : Cas d'utilisation du biologiste

3.3.2.2 CAS D'UTILISATION DU CLINICIEN

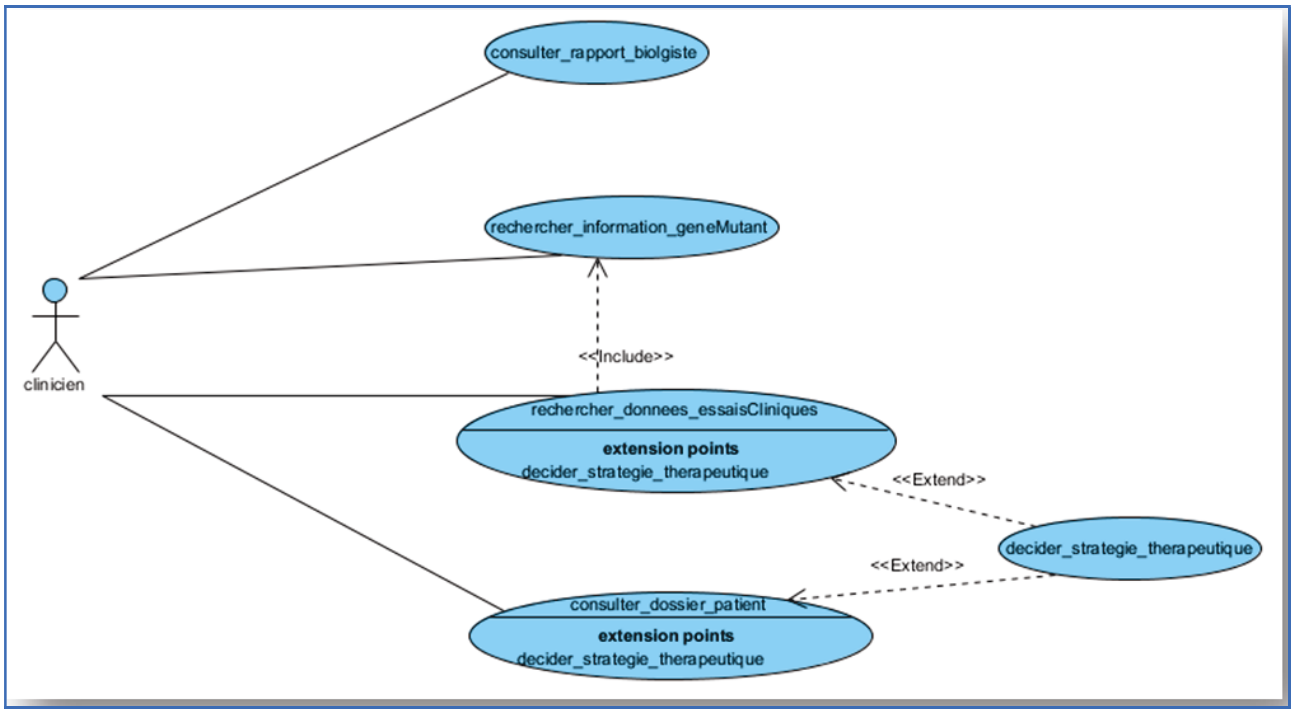


Figure 20 : cas d'utilisation du clinicien

Le tableau ci-dessous nous explique les différents cas d'utilisation du clinicien :

Tableau vi : description des Cas d'utilisation du clinicien

Cas d'Utilisation	Description
consulter_rapport_biologiste	A la réception du rapport du biologiste, le clinicien se doit de le lire pour en extraire les informations pertinentes
rechercher_information_geneMutant	Le clinicien, pour compléter les remarques du biologiste, recherche des informations sur les mutations génétiques identifiées dans diverses sources telles ClinVar, ou encore GeneOntology
rechercher_donnees_essaisCliniques	Selon la maladie induite par la mutation génétique, le médecin recherche des informations sur les essais cliniques en cours pour voir dans quelle mesure il peut demander à inclure son patient. Cette recherche entraîne systématiquement une recherche sur les gènes délétères ou les voies génétiques d'interactions des médicaments (Pathways). Les informations ainsi recueillies permettent au clinicien de décider s'il inclut ou non son patient dans l'essai.
consulter_dossier_patient	Le clinicien, pour arrêter son diagnostic et affiner sa stratégie thérapeutique, consulte le dossier du patient. Cela déclenchera à un moment ou à un autre un choix de traitement
décider_stratégie_thérapeutique	Cette action découle à la fois des conditions du patient et des conditions de l'essai cliniques : deux pistes seront envisagées : inclure le patient dans un essai clinique ou proposer un parcours de soins basé sur des traitements déjà éprouvés en tenant compte de la spécificité du patient.

Dans ce chapitre nous avons étudié les processus métiers du biologiste et du clinicien. Cela a conduit à leur modélisation. Nous avons ensuite déduit les cas d'utilisation de ces deux acteurs. Au travers de ces cas d'utilisation, nous avons appris les fonctionnalités attendues par les acteurs

ainsi que les données qu'ils vont manipuler. Cela décrit donc leur domaine de connaissance mais aussi les relations entre les classes que notre système devra implémenter.

Le chapitre qui va suivre nous mène à réfléchir sur le domaine de connaissance. Cela s'est fait au travers d'une carte mentale.

3.4 MODELISATION DE L'ONTOLOGIE ICE

La conception de l'ontologie a commencé par une réflexion autour du patient et des différents concepts que nous serons amenés à implémenter dans l'ontologie. Une carte mentale a été créée pour nous permettre d'aller au bout de notre raisonnement.

3.4.1 CARTE MENTALE

Nous avons alors créé une carte mentale (mind map) pour décrire de manière exhaustive les concepts. La réflexion a porté sur le patient, ses gènes et leurs mutations, les maladies et leurs traitements. On obtient la carte ci-dessous :

3.4.1.1 LA BRANCHE DEMOGRAPHIE

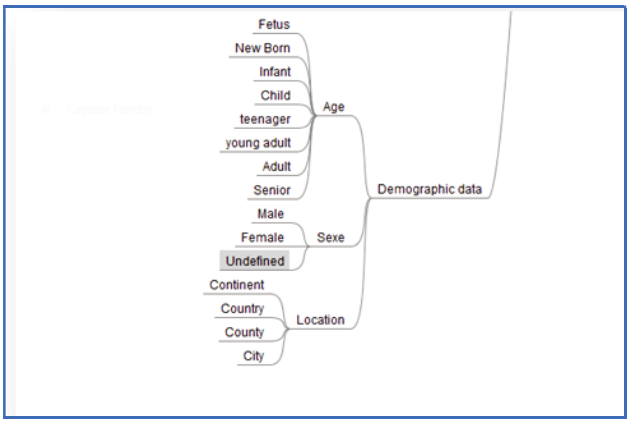


Figure 21 : Carte Mentale - branche démographie

Cette branche permet de décrire les informations du patient à savoir son âge, son sexe et sa localisation géographique.

3.4.1.2 LA BRANCHE GENOME

Cette branche permet de décrire l'un des piliers de l'ontologie ICE à savoir la description des gènes et de leurs variations. J'ai ainsi décrit les variations relatives à l'ADN, à l'ARN, ainsi qu'aux protéines. Sur chaque composant j'ai décrit selon leur origine les mutations concernant :

3.4.1.2.1 LE CHROMOSOME

La structure : ces variations peuvent concerner la chaîne de l'ADN. Dans ce cas on peut observer des suppressions, des insertions, de la translocation ou encore de la substitution dans la chaîne de l'ADN. Il peut aussi s'agir d'amplification

Le nombre de chromosomes : ici, soit le nombre de chromosomes est anormal, il y en a en trop (comme dans la trisomie 21) ou pas assez, soit des copies multiples d'un même chromosome sont observées (on parle de ploïdie).

3.4.1.2.2 LA FONCTION DU GENE

On constate ici un défaut au niveau des fonctions des gènes :

- **Transfert**
- **Réparation**
- **Réplication**
- **Régulation**
- **Transcription**

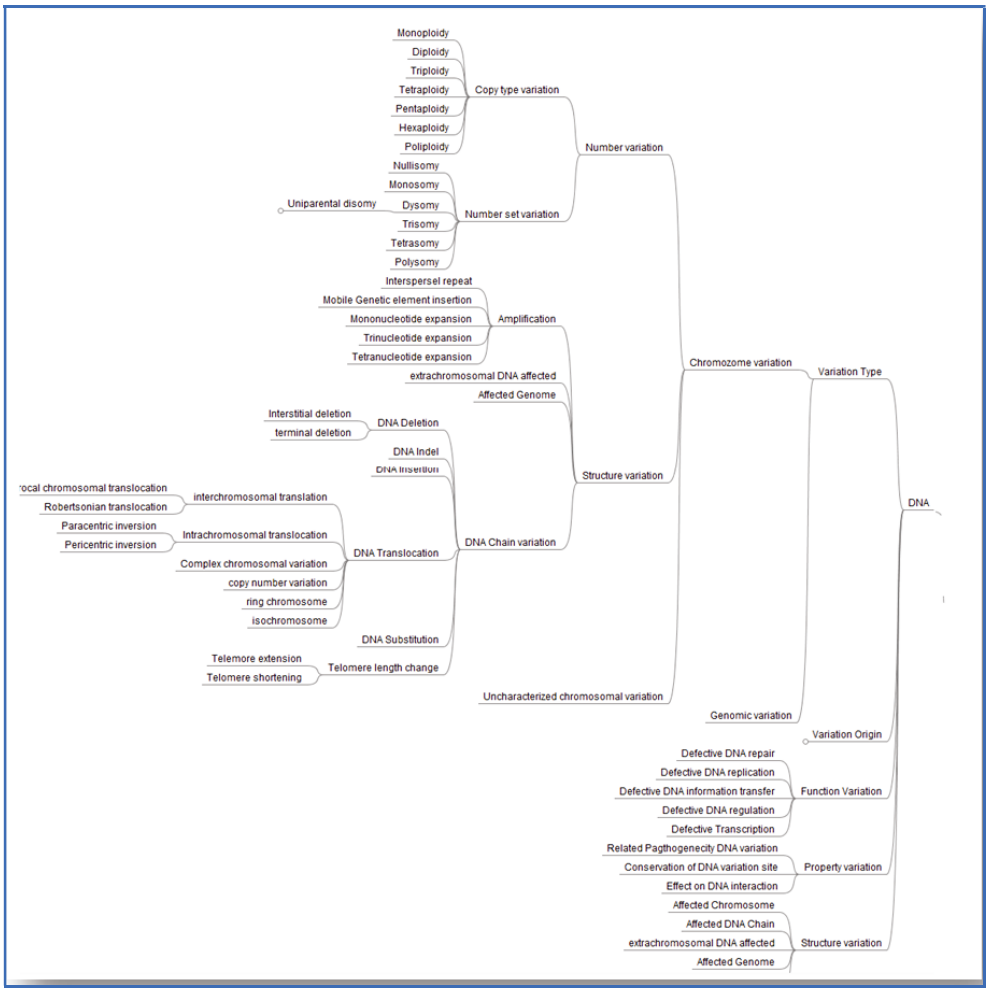


Figure 22 : Carte Mentale -Variation du Gène

3.4.1.3 LA BRANCHE MALADIE

Sur cette branche je me suis attachée à décrire les maladies résultant de prolifération cellulaires qui occasionnent les cancers. Les cancers que nous avons recensés sont de divers types :

3.4.1.3.1 CANCERS DES CELLULES

On y retrouve :

- ***Les blastomes***: tumeurs malignes développées à partir d'un type cellulaire embryonnaire
- ***les carcinomes*** : tumeurs développées à partir des cellules d'un épithélium. Un carcinome est qualifié de tumeur solide puisqu'il forme un bloc de cellules plus ou moins soudées entre elles.

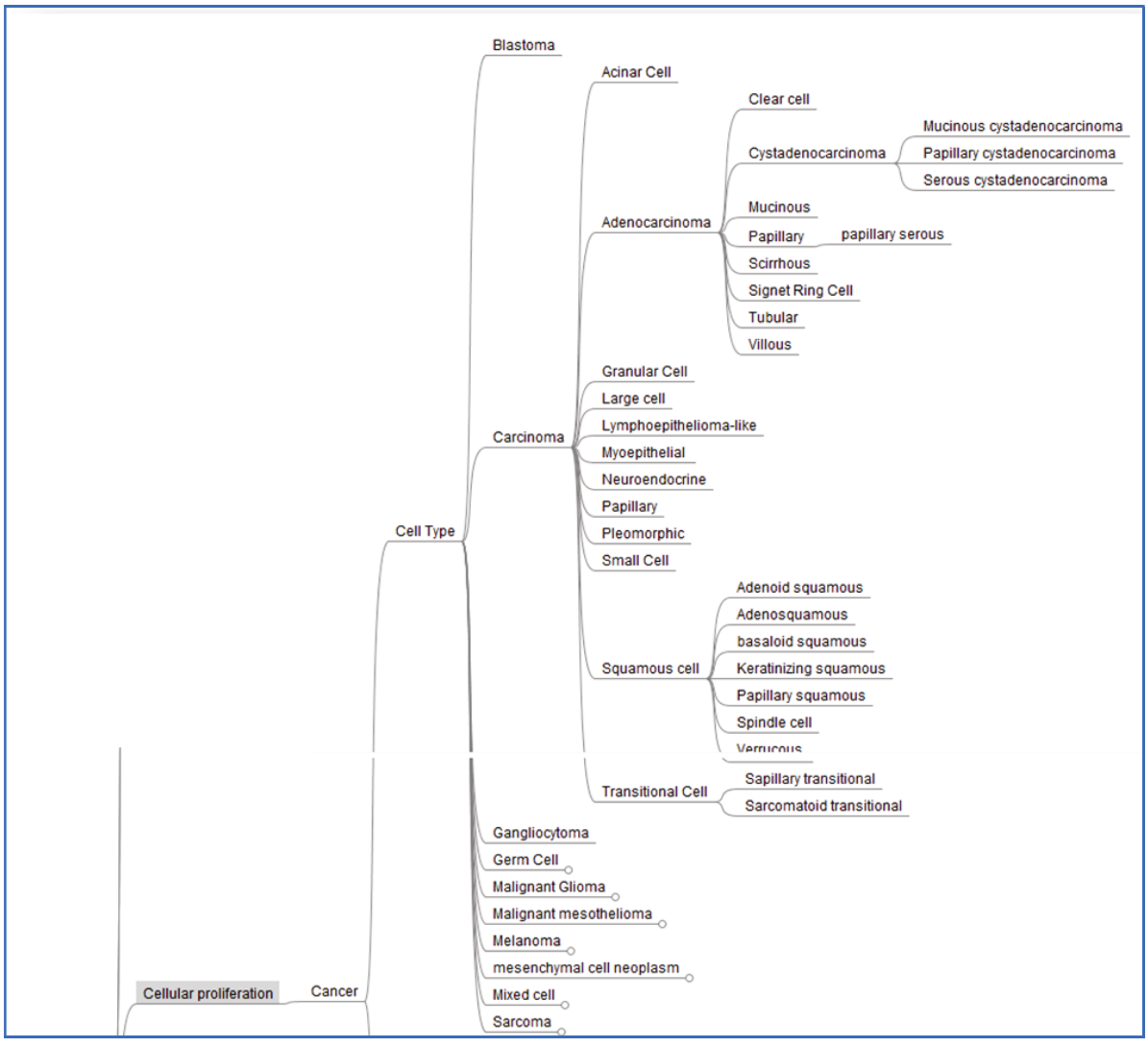


Figure 23 : Carte mentale - Cancer

Dans un souci d'exhaustivité, nous avons décrit toutes les autres formes de cancers mais notre étude se focalisera sur les carcinomes, les autres seront intégrés au fur et à mesure.

3.4.1.4 LA BRANCHE TRAITEMENT

Cette branche m'a permis de décrire tous les concepts autour des médicaments. En effet un traitement s'inscrit sur trois niveaux : actes médicaux, traitement médicamenteux et essais cliniques :

3.4.1.4.1 LES ACTES MEDICAUX

Ce sont tous les actes que le corps médical sera amené à opérer sur un patient. Ce peut être un acte purement administratif comme l'admission du patient pour des soins, ou son transfèrement vers d'autres services. Il peut s'agir aussi de tous les actes d'investigation pour poser un diagnostic comme les examens de laboratoires.

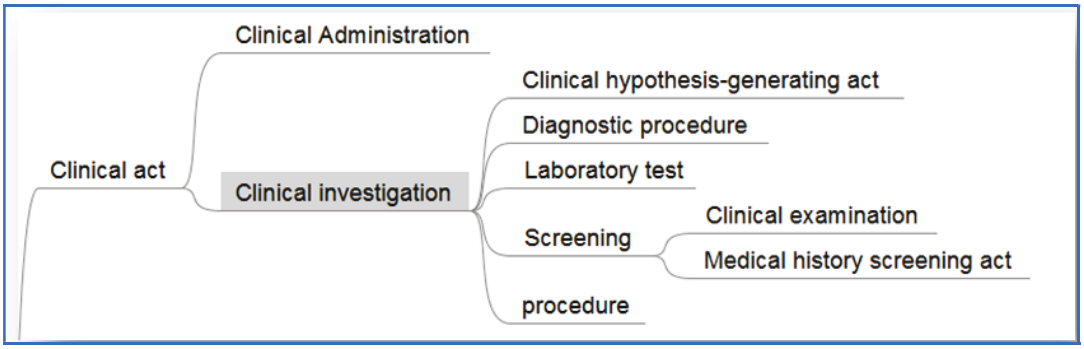


Figure 24 : Carte Mentale - Actes Cliniques

3.4.1.4.2 LES TRAITEMENTS MEDICAMENTEUX

Ici nous avons décrit tous les concepts autour des médicaments. Ainsi un médicament se caractérise par :

- ***Une autorité de régulation*** et de mise sur le marché comme la Federal Drug Administration (FDA) aux Etats-Unis.
- ***Un fabricant***: le laboratoire qui a mis au point la molécule et qui la commercialise
- ***La molécule***
- ***Le mode*** par lequel le médicament interagit avec l'organisme
- ***Les interactions*** avec d'autres médicaments
- ***Les preuves*** de son activité Pharmacocinétique (absorption, distribution, métabolisme et élimination).
- Enfin toutes les ***publications*** relatives à ce médicament

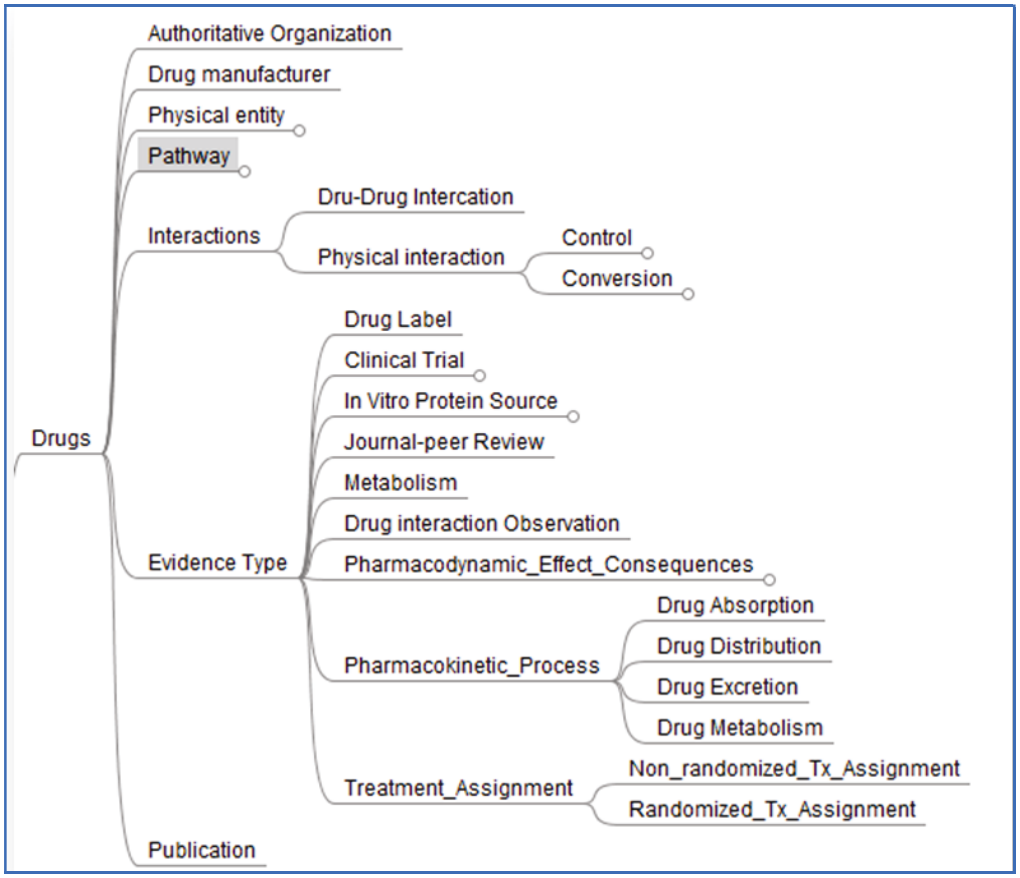


Figure 25 : Carte Mentale - Médicaments

3.4.1.5 LES ESSAIS CLINIQUES

L'exploration du concept essai clinique m'a permis de référencer tous les types d'essais cliniques. Il en ressort que les essais cliniques sont classifiés en trois catégories :

- ***Etudes analytiques***
- ***Etudes descriptives***
- ***Etudes longitudinales***
- ***Les études analytiques***

Ce sont des études quantitatives qui peuvent revêtir les formes suivantes :

3.4.1.5.1 ETUDE A ACCES ELARGI

Ce type d'étude « permet à des patients atteints de maladies graves et qui ne peuvent pas participer à des essais cliniques contrôlés d'avoir accès au médicament à condition qu'ils ne soient pas exposés à des risques déraisonnables » (CHEBI, C. E. (s.d.). Chemical Entities of Biological Interest Ontology. Récupéré sur Bioportal: http://purl.obolibrary.org/obo/CHEBI_23888)

3.4.1.5.2 ETUDES EXPERIMENTALES

Ce sont des études au cours desquelles un traitement est administré pour en observer les résultats.

3.4.1.5.3 ETUDES CONTROLEES

Un groupe-contrôle est indispensable pour valider l'efficacité d'une procédure afin d'aboutir à des conclusions non biaisées et hâtives

Dans les essais quantitatifs, il y aussi le mode de conception des essais pour déterminer l'essai qui correspond au mieux à l'étude envisagée.

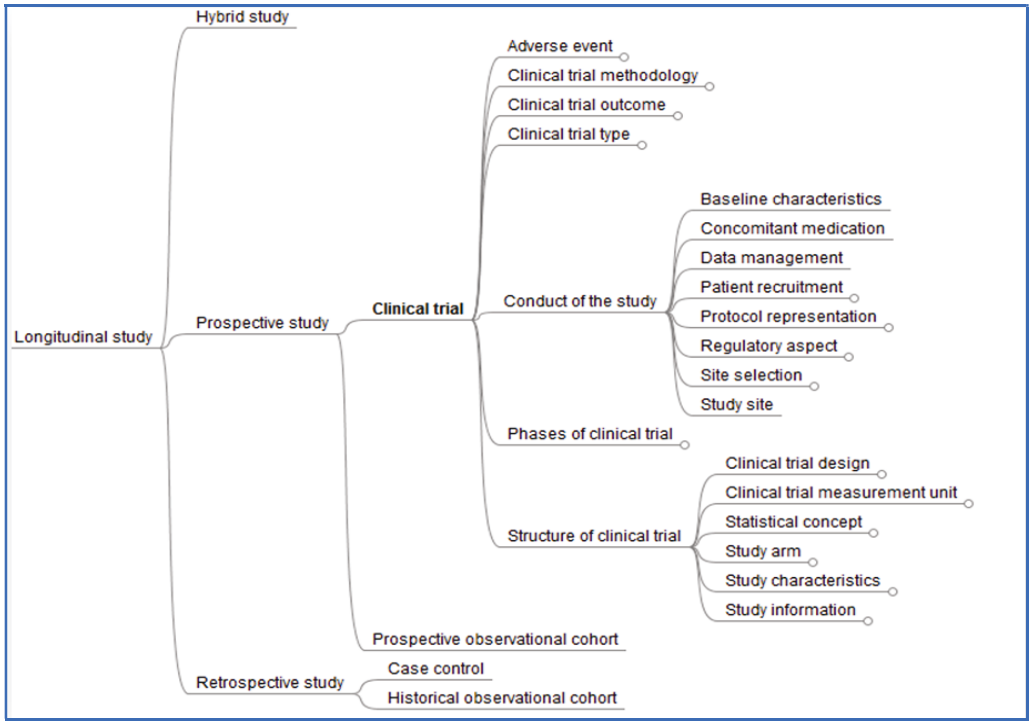


Figure 26 : Carte Mentale - Essais Cliniques

3.4.1.5.4 LES ETUDES DESCRIPTIVES

Ce sont des études conçues uniquement pour décrire la distribution de certaines variables. Elles ne recherchent pas les liens de causalité entre ces variables. (CTO)

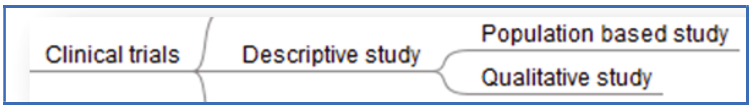


Figure 27 : Carte Mentale - Etudes Descriptives

Il y a deux catégories d'essais descriptifs : les essais basés sur des populations et les essais quantitatifs.

3.4.1.5.5 LES ETUDES LONGITUDINALES

Dans cette catégorie nous rencontrons les études prospectives, rétrospectives et des études hybrides.

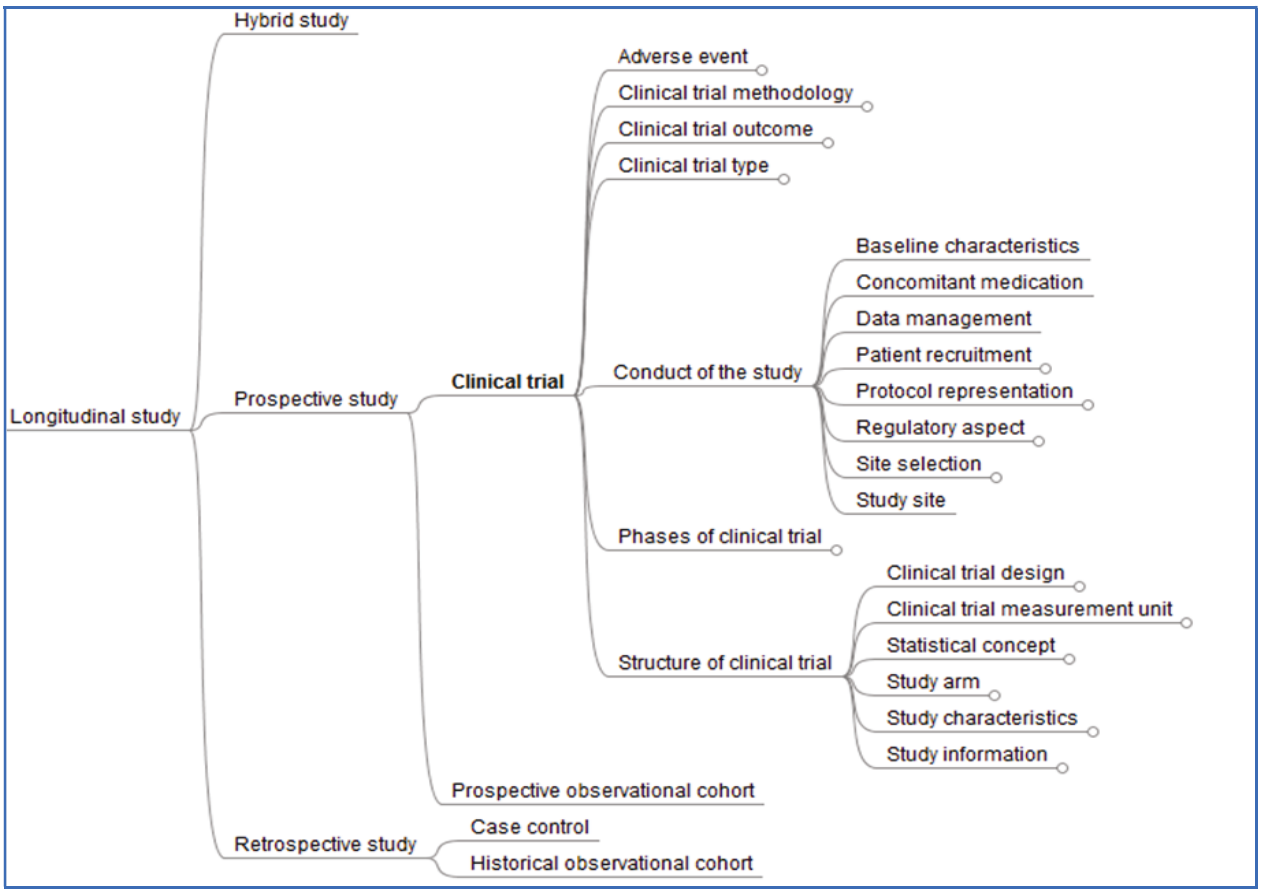


Figure 28 : Carte Mentale - Etudes Longitudinales

3.4.1.5.6 LES ETUDES PROSPECTIVES

Ces études ont pour but de déterminer les facteurs causant des troubles en comparant un panel d'individus qui n'ont pas la manifestation d'un résultat mais sont exposés à une cause potentielle comparativement à un panel concurrent qui lui n'a pas été exposé à la cause (effet placebo). On pourra ainsi étudier :

- Les effets (bénéfiques ou adverses)
- La méthodologie de l'essai
- Les résultats

- Le type d'essais
- Les différentes phases de l'essai
- Sa structure
- Les études rétrospectives

Il s'agit d'études pour évaluer les causes possibles d'une maladie au sein d'un panel de personnes ayant cette maladie ou ce trouble comparé aux personnes d'un autre panel qui ne l'ont pas eu égard à une exposition précédente à une cause potentielle.

Les essais hybrides : c'est une combinaison des deux types d'essais précédents.

Au terme de cette réflexion autour du gène, de ses variations, des maladies notamment les cancers de type carcinomes pouvant en résulter ainsi que des traitements j'ai pu concevoir le modèle ICE que nous avons ensuite implémenté dans l'outil OBO-EDIT

3.4.2 LE MODELE DE L'ONTOLOGIE ICE

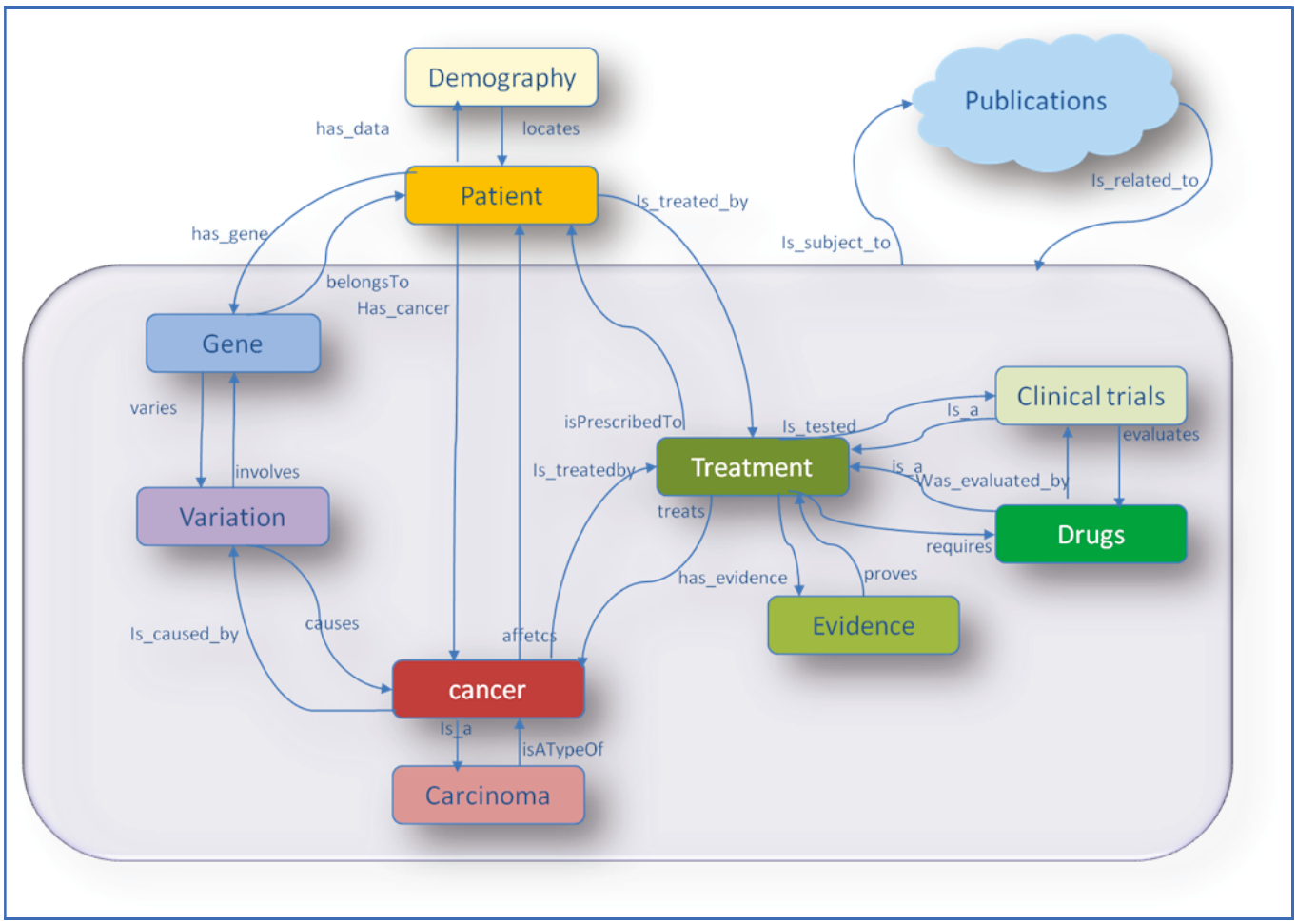


Figure 29 : Metamodèle ICE

Le metamodèle de l'ontologie ICE fait apparaître les classes (types selon la notation OBO) suivantes :

3.4.2.1 LES CLASSES DU MODELE

Tableau vii : Classes de l'ontologie ICE

Classes	Définition
Patient	Personnes recevant un traitement médical.
Demography	Données statistiques permettant de catégoriser un patient : âge, sexe, localisation, adresse
Gene	Unité d'hérédité contrôlant un caractère particulier. Cet élément génétique correspondant à un segment d'ADN ou d'ARN, situé à un endroit bien précis (locus) sur un chromosome
Variation	Altération dans l'ADN, l'ARN ou les protéines : Une modification ou la différence d'une norme ou standard. (NCIT)
Cancer	Maladie de la prolifération cellulaire incontrôlée qui est maligne et primaire, caractérisée, l'invasion des cellules locales et les métastases (DOID, 2015) Elle peut être causée par une variation de gène.
Carcinoma	Cancer développé à partir d'un tissu épithélial (peau, muqueuse) (DOID, 2015)
Treatment	Procédé délivré par un professionnel de la santé dans le but de soulager les symptômes associés à un trouble. (OBI, 2011)
Clinical Trials	Etude contrôlée pour évaluer l'efficacité et l'innocuité de nouveaux médicaments, dispositifs, traitements en comparant deux ou plusieurs groupes.
Drugs	Toute substance qui, lorsqu'elle est absorbée par un organisme vivant peut modifier une ou plusieurs de ses fonctions. Le terme est généralement admis pour une substance prise dans un but thérapeutique, mais est aussi couramment utilisé pour les substances consommées. (CHEBI)
Evidence	Elément permettant de soutenir une assertion particulière telle que l'existence d'une interaction, d'une voie clinique. La création de l'instance de preuve se fera par l'instanciation d'un code de confiance, une evidence-code ou un formulaire d'expérimentation (DIKB)
Publication	Des copies d'une œuvre ou d'un document distribué au public par la vente, la location, ou de prêt. (ALA Glossaire de Bibliothèque et sciences de l'information, 1983, p 181). Dans notre cas, la littérature biomédicale portant sur les entités de notre ontologie.

3.4.2.2 LES RELATIONS ENTRE LES RESSOURCES DU MODELE

Table viii : Relations du modèle ICE

Relation	Description
Has_data	Cette relation relie le patient à la classe démographie
Locates	Relation entre la Classe Demography et la classe Patient. Elle précise que la démographie permet de localiser le patient
Has_gene	Lie le patient à la classe Gene
belongsTo	Lie la classe Gene à la classe Patient
Varies	Lie la classe Gène à la classe Variation
Involves	Lie la classe Variation à la classe Gène
varies	Lie la classe Variation à la classe Gène
Causes	Lie la classe Variation à la classe Cancer
Is_caused	Lie la classe Cancer à la classe Variation
Affects	Lie la classe Cancer à la classe Patient
has_cancer	Lie la classe Patient avec la classe Cancer
Is_a	Lie la Classe Carcinome à la classe Cancer
isATypeOf	Lie la classe Carcinome à la classe Cancer
Is_treated_by	Lie la classe Treatment à la classe Patient
hasPrescription	Lie la classe Patient à la classe Treatment
isPrescribedTo	Lie la classe Treatment à la classe Patient
Is_subject_to	Lie les classes Gene, Variation, Treatment, Drug, Cancer, Clinical Trials à la classe Publication
isRelatedTo	Lie la Classe Publication aux Classes Gene, Variation, Treatment, Drug, Cancer, Clinical Trials
Tests	Lie la classe Clinical Trial à la classe Treatment
isTestedBy	Lie la Classe Treatment à la classe Clinical Trials
isRequired	Lie la classe Drugs à la classe Treatment

Requires	Lie la classe Treatment à la classe drugs
Tested	Lie la classe Clinical Trial à la classe Drugs
wastestedBy	Lie la classe Drugs à la classe Clinical Trials
has_evidence	Lie la classe Treatment à la classe Evidence
Proves	Lie la classe Evidence à la classe Treatment
Is_treatedby	Lie la classe Cancer à la classe Treatment
Treats	Lie la classe Treatment à la classe Cancer

Au sortir de cette modélisation, nous avons implémenté l'ontologie ICE avec l'outil OBO-Edit.

3.4.2.3 LE CHOIX DES OUTILS

Lorsqu'il s'est agi de construire l'ontologie avec un outil, nous avons été confrontés au choix de l'outil le mieux adapté. Après avoir parcouru les différentes offres du marché, deux outils sortaient du lot, Protégé et OBO-Edit.

3.4.2.3.1 QU'EST-CE QUE PROTEGE ? 

Protégé est édité par la Stanford University. C'est un logiciel libre, d'édition d'ontologie, basé sur le langage OWL. Il offre une interface graphique qui permet décrire les ressources et les propriétés de ces ressources selon le format RDF. Il offre les formalismes suivants : RDF/XML, OWL/XML, OBO.

Il est basé sur le langage java et permet d'utiliser des raisonneurs tels Pellet pour vérifier les inférences de raisonnement à l'intérieur des ontologies.

Protégé utilise également des API comme Jena Fuseki pour le requêtage et la manipulation des données de l'ontologie. Enfin SPARQL est le langage de requêtage utilisé pour interroger la base de connaissance.

L'interface de Protégé se présente comme suit :

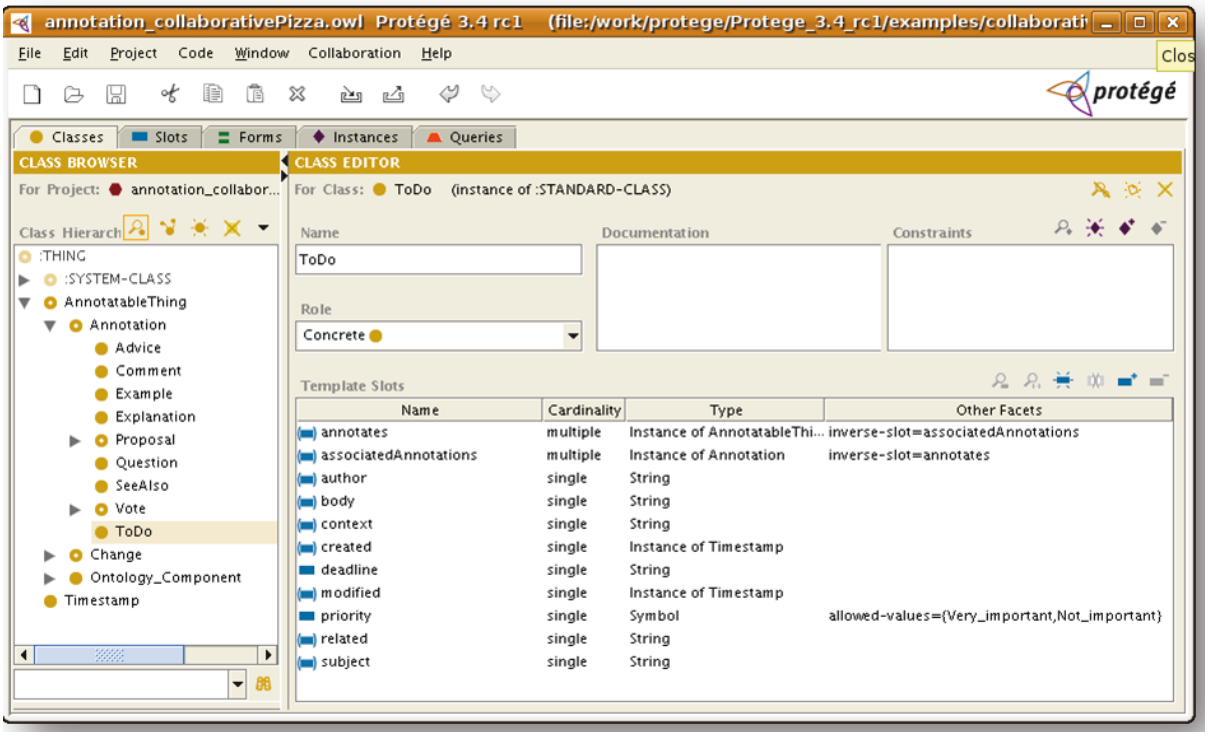


Figure 30 : Interface Protegé

3.4.2.3.2 OBO-EDIT

OBO-Edit est l'émanation d'OBO-Foundry. Comme nous l'avons cité plus haut c'est un organisme collaboratif qui a pour but de développer un ensemble de principes pour la création d'ontologies biomédicales. Il en a résulté le Consortium Gene Ontology qui regroupe un ensemble d'ontologies et d'activités de recherche dans le cadre du projet Gene Ontology. OBO-Edit est donc essentiellement un logiciel de développement d'ontologies biomédicales.

Comme Protégé il est libre et développé en Java. Il va générer des données en Format OBO, mais permet aussi de gérer des formats OWL. Comme Protégé il embarque un raisonneur qui permet de vérifier à chaque sauvegarde la cohérence de l'ontologie et de détecter les erreurs.

OBO-Edit offre l'interface suivante :

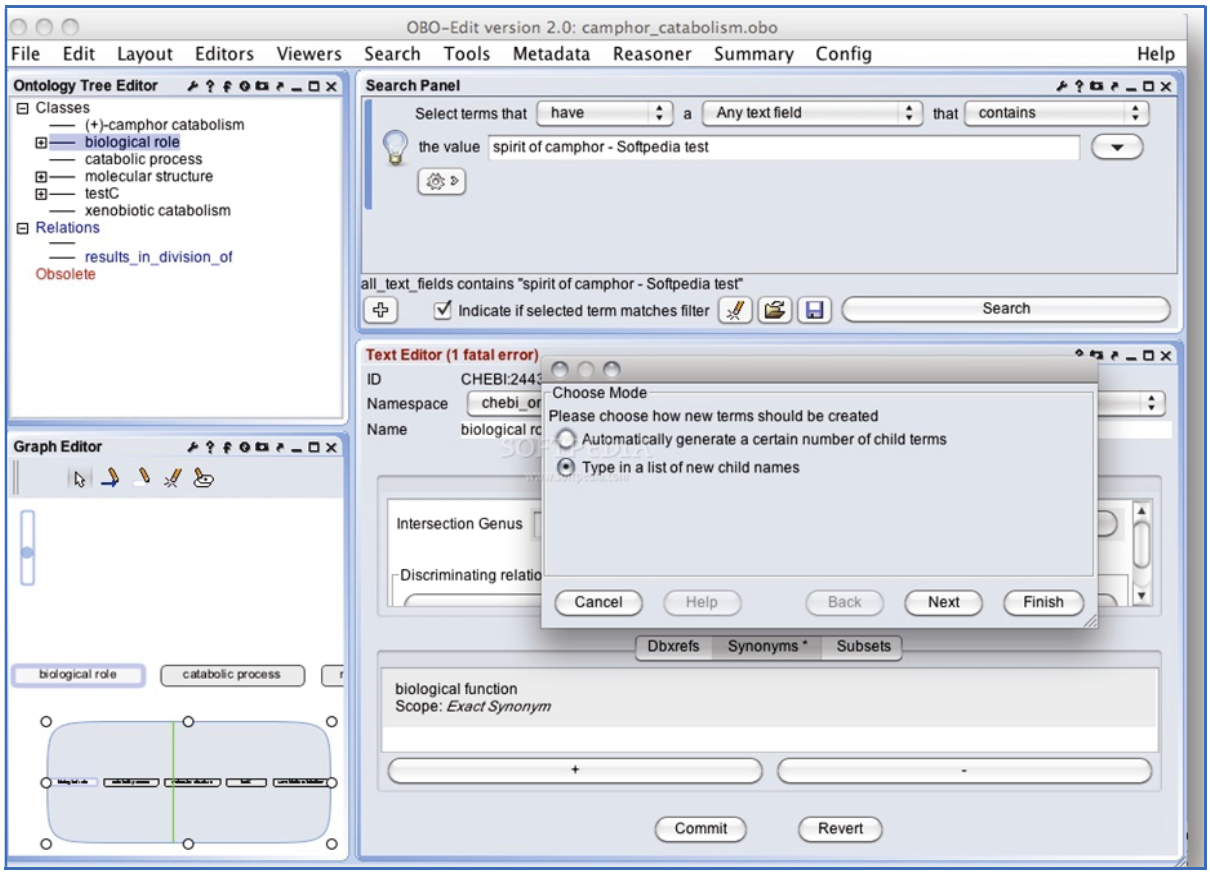


Figure 31 : Interface OBO-Edit

Dans le tableau qui suit nous comparons les deux outils.

Tableau ix : Tableau Comparatif OBO-Edit et Protegé

Editeur	Description	Techno	Format	Facilité d'usage	Requêtage	Raisonneur	API
OBO-Edit	Permet de créer et de manipuler les ontologies biomédicales notamment	JAVA	OBO, OWL	☺☺	SPARQL	Raisonneur intégré	BioJava OWL API
Protégé	Permet des gérer des ontologies	JAVA	OWL, RDF/XML	☺	SPARQL	Pellet HermiT, Fact++ …	Jena, Jena Fuseki, OWL API

Au sortir de cette analyse notre choix en premier s'est porté sur OBO-Edit comme outil de développement de l'ontologie ICE car c'est un projet qui s'inscrit dans la recherche biomédicale. En outre l'import d'ontologies biomédicales existantes dans Protégé a échoué du fait de leur volume trop important. Enfin La majorité des ontologies que nous comptons intégrer dans l'outil sont référencées sur le site Bioportal ontology et font partie de OBO-Foundry sont soit des ontologies candidates ou notées comme d'intérêt.

3.5 PROTOTYPAGE

En accord avec la méthodologie du projet, nous avons implémenté trois prototypes durant le projet. Ses ontologies prototypes ont été conçues respectivement sous OBO-Edit puis Protégé. A chaque itération nous avons amélioré, adapté l'ontologie en fonction des contraintes mais aussi objectifs que nous souhaitions atteindre.

Le chapitre qui suit, décrit quels ont été les prototypes et quel a été le moteur pour le passage au prototype suivant.

3.5.1 PROTOTYPE 1 : ICE ONTOLOGY FORMAT OBO

Dans cette partie, nous abordons la création de l'ontologie dans OBO-Edit qui se présente comme suit :

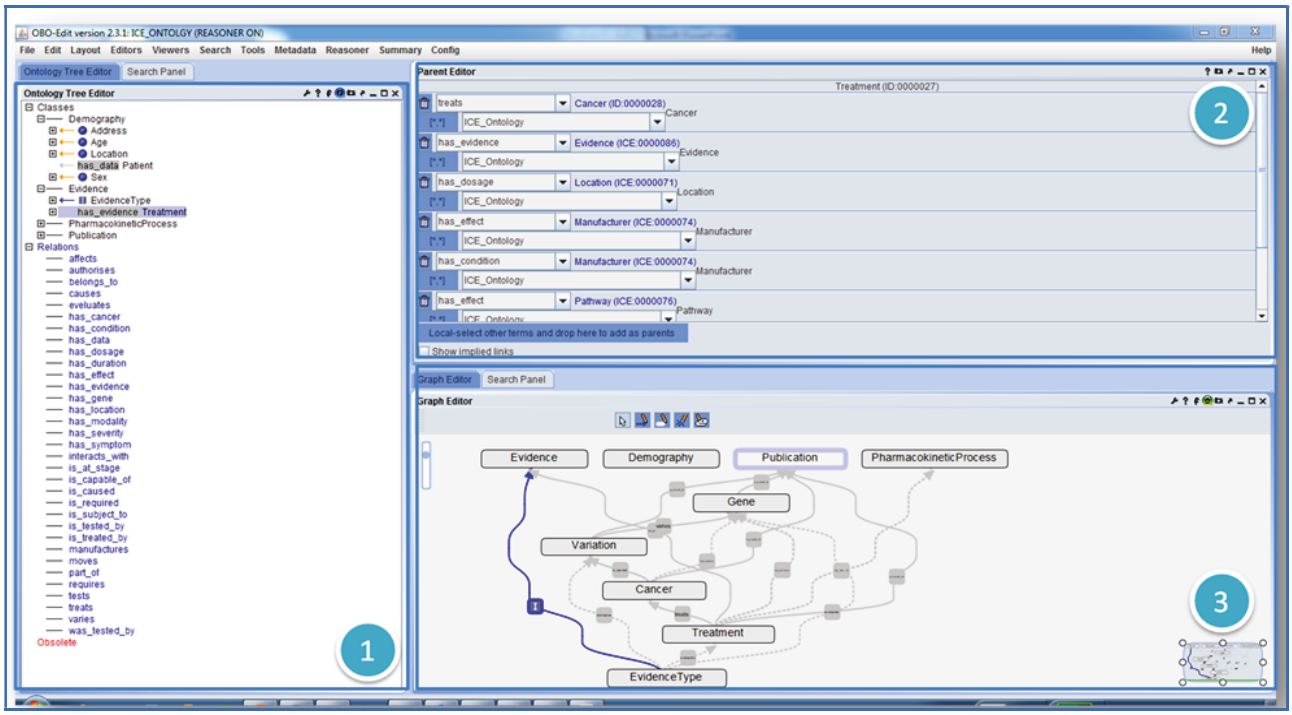


Figure 32: Vue Générale de l'ontologie ICE

Cet écran nous montre trois fenêtres d'édition de l'ontologie :

1. ***L'Ontology Tree Editor***: Cette fenêtre montre l'arborescence de l'ontologie. Il permet de voir pour chaque type d'objet les enfants et tous les types avec qui il est en relation.
2. ***Le Parent Editor*** : Lorsqu'on sélectionne un type dans l'éditeur, le Parent Editor affiche toutes les relations avec les autres types. On peut ainsi éditer ces relations.

3. ***Le Graph Editor :*** Il permet d'afficher le graphe des classes parentes et filles. On peut également gérer les relations entre elles.

Ce sont les vues les plus importantes qui nous ont le plus servi dans l'application. Nous avons pu ainsi implémenter les classes comme suit :

3.5.1.1 DESCRIPTION DES COMPOSANTS DE L'ONTOLOGIE

Remarque : du fait de la redondance par inférence détectée par le raisonneur, nous n'avons intégré qu'une relation par classe du metamodèle présenté plus haut.

3.5.1.1.1 LA CLASSE DEMOGRAPHY

Les types d'objet et les relations avec les autres sont les suivantes :

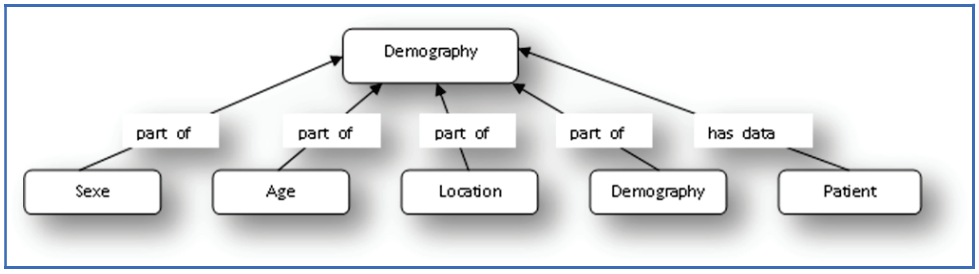


Figure 33 : la classe Demography, ses sous-classes et les relations entre elles

Cela donne dans OBO-Edit le modèle suivant :

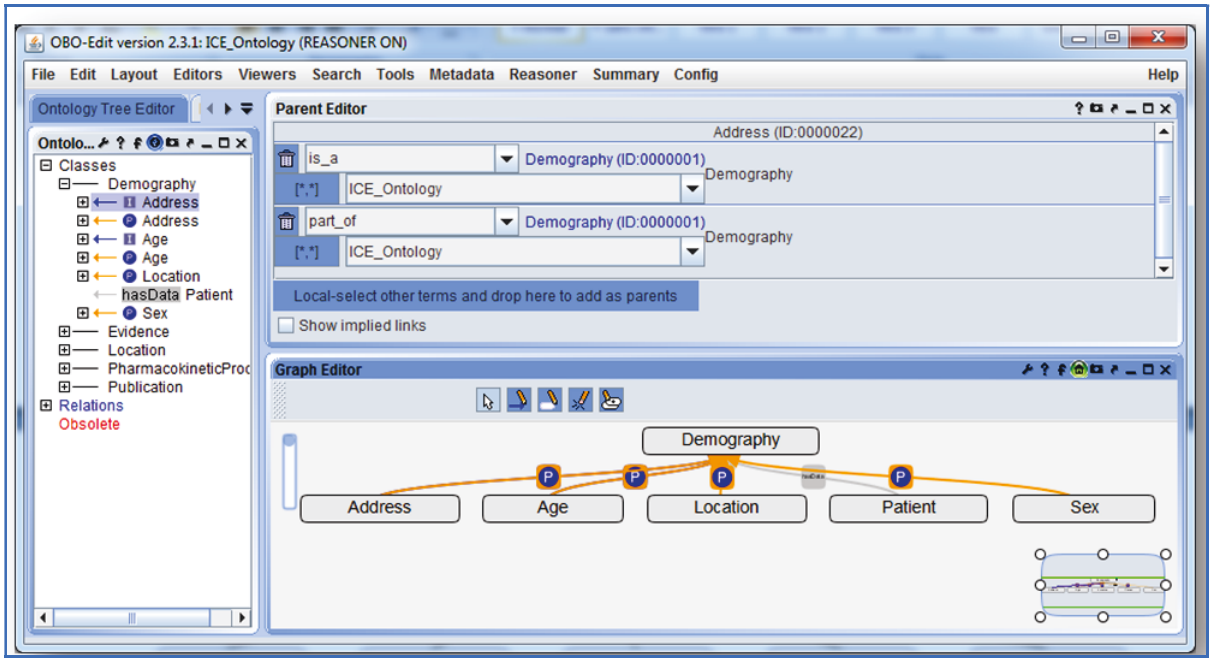


Figure 34 : Classe Demagraphy sous OBO edit

La classe Patient a été implémentée comme suit :

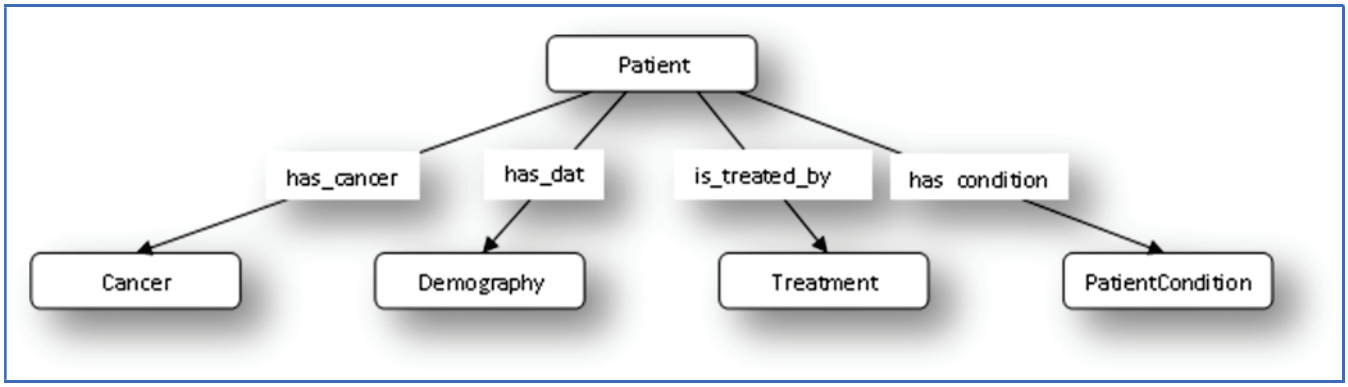


Figure 35 : classe Patient

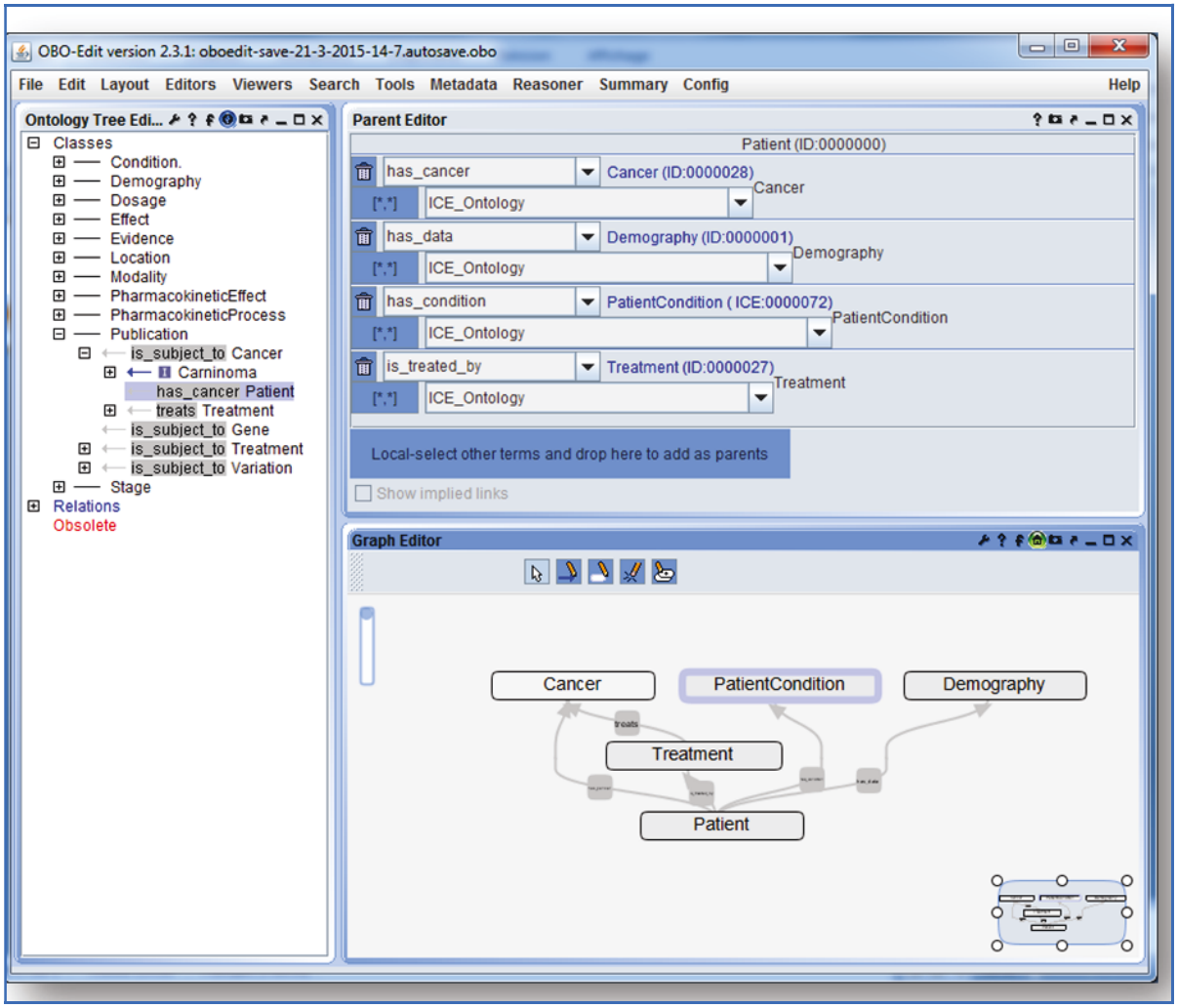


Figure 36 : Vue de la Classe Patient dans OBO-Edit

3.5.1.1.2 LA CLASSE TREATMENT.

Le traitement a été implémenté dans OBO-Edit de la manière suivante :

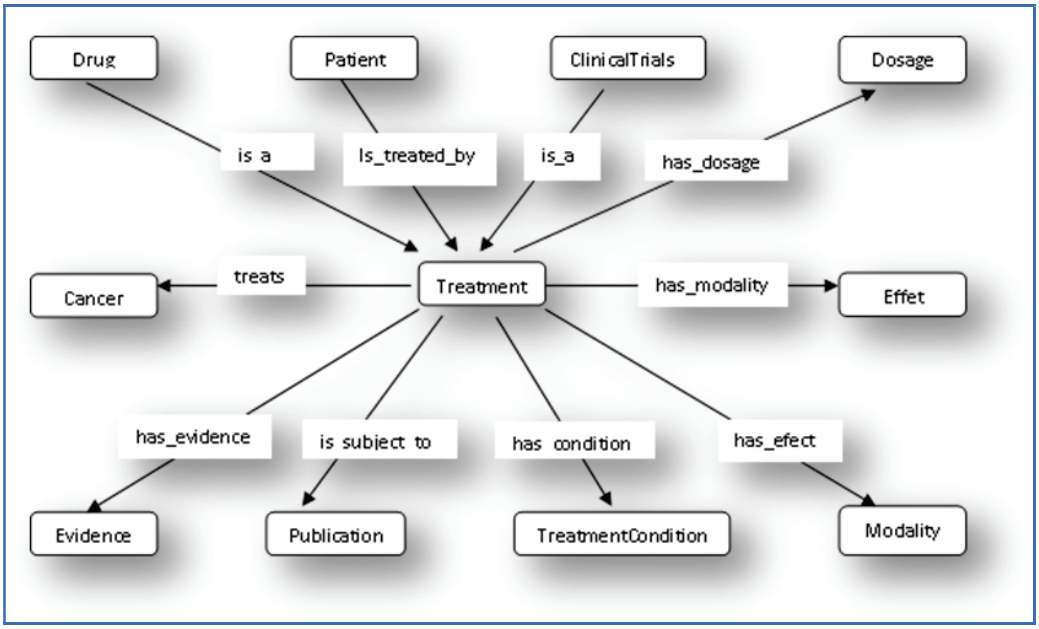


Figure 37 : Classe Treatment

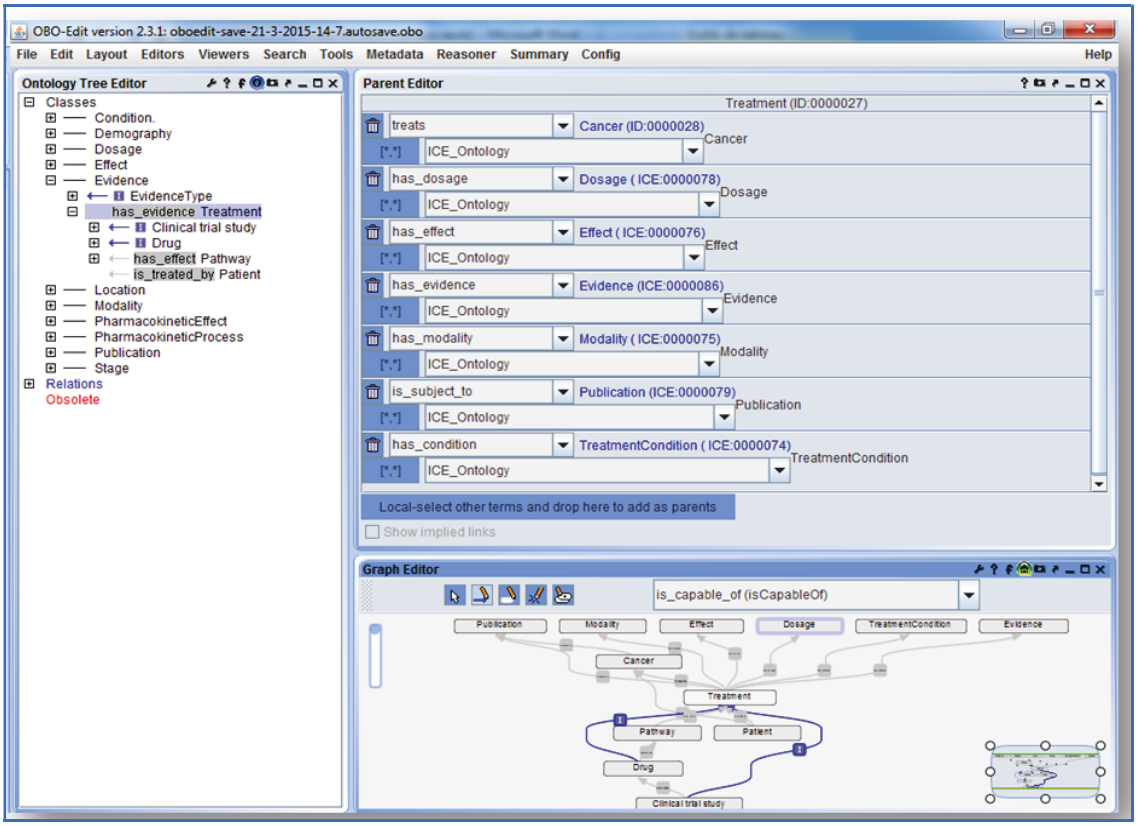


Figure 38 : Vue de la classe Treatment dans OBO-Edit

3.5.1.1.3 LA CLASSE CANCER

Elle est décrite de la manière suivante :

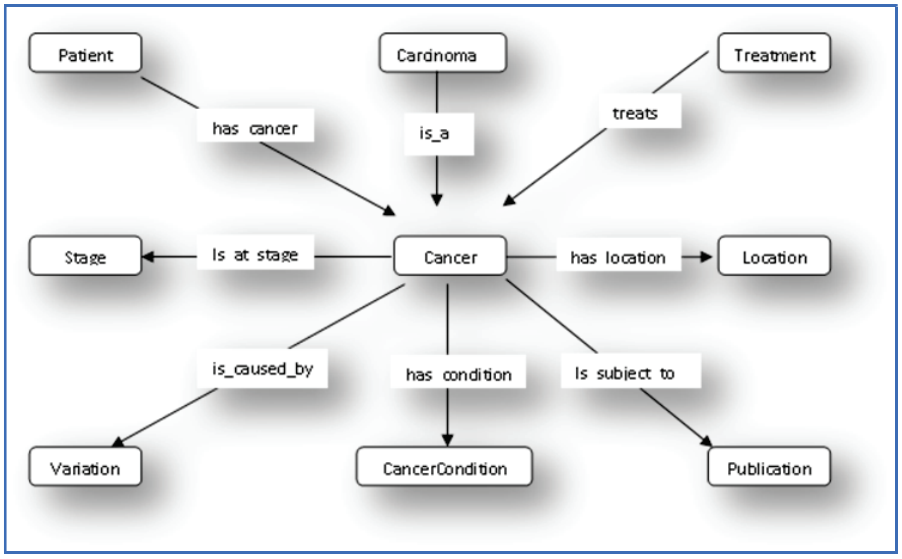


Figure 39 : Classe Cancer

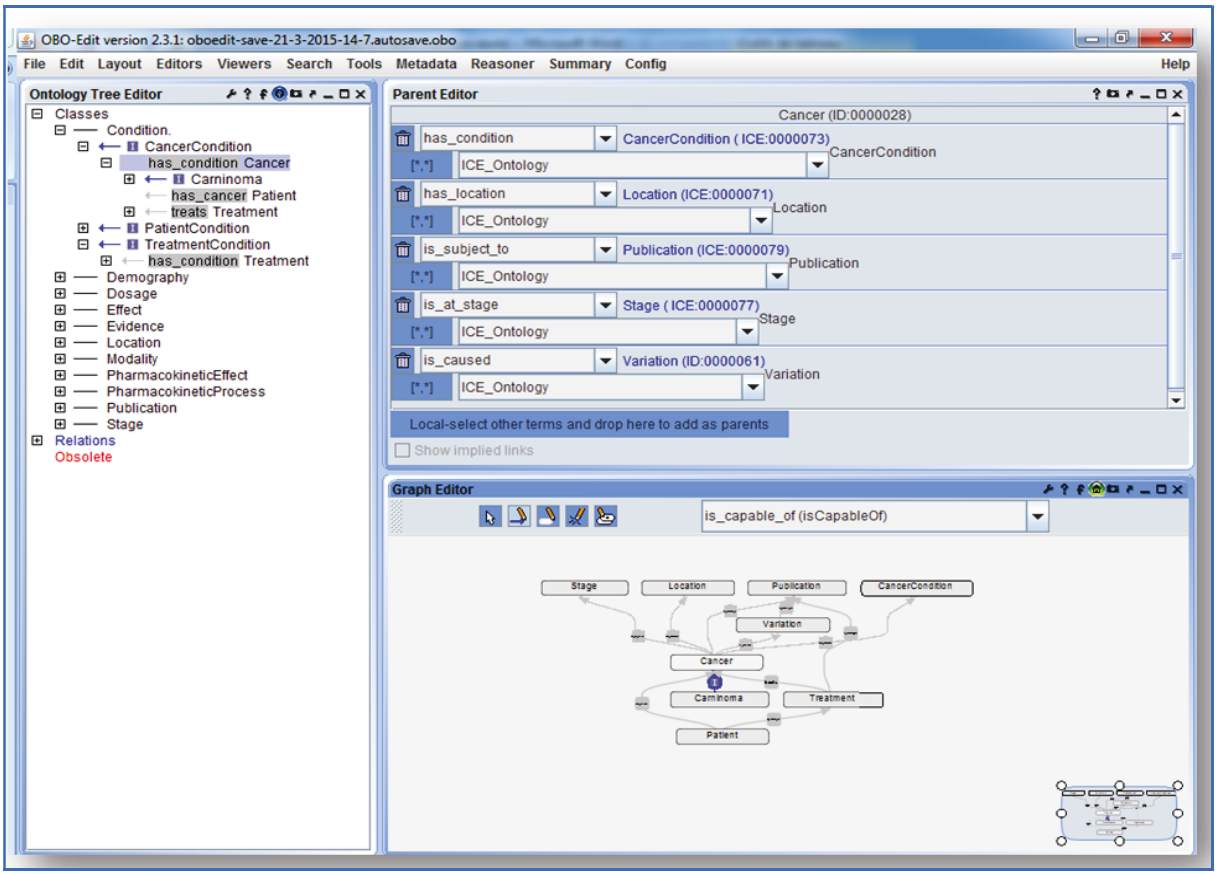


Figure 40 : Vue de la classe Cancer dans OBO Edit

3.5.1.1.4 LA CLASSE VARIATION

Les variations ont été décrites de la manière suivante :

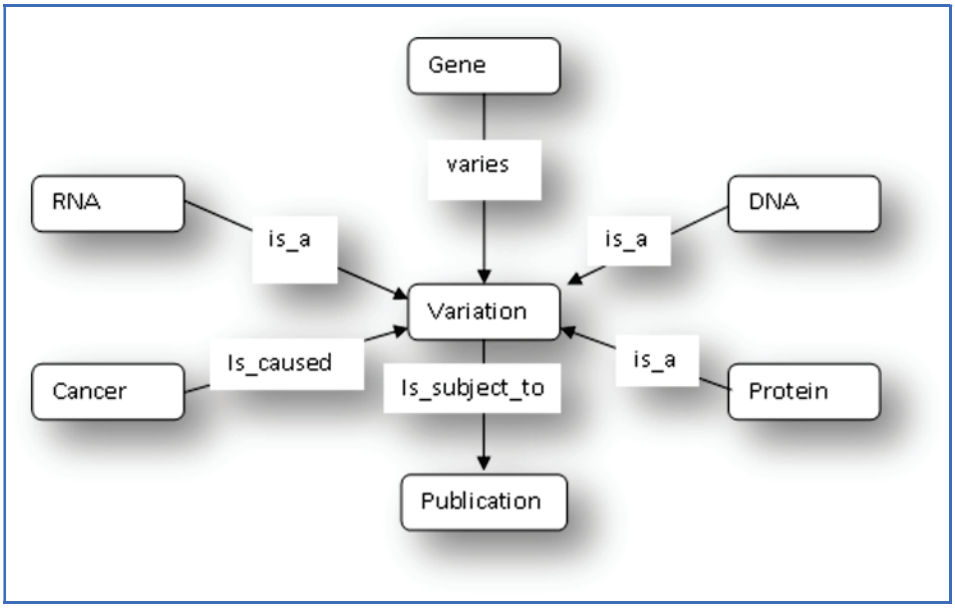


Figure 41 : Classe Variation

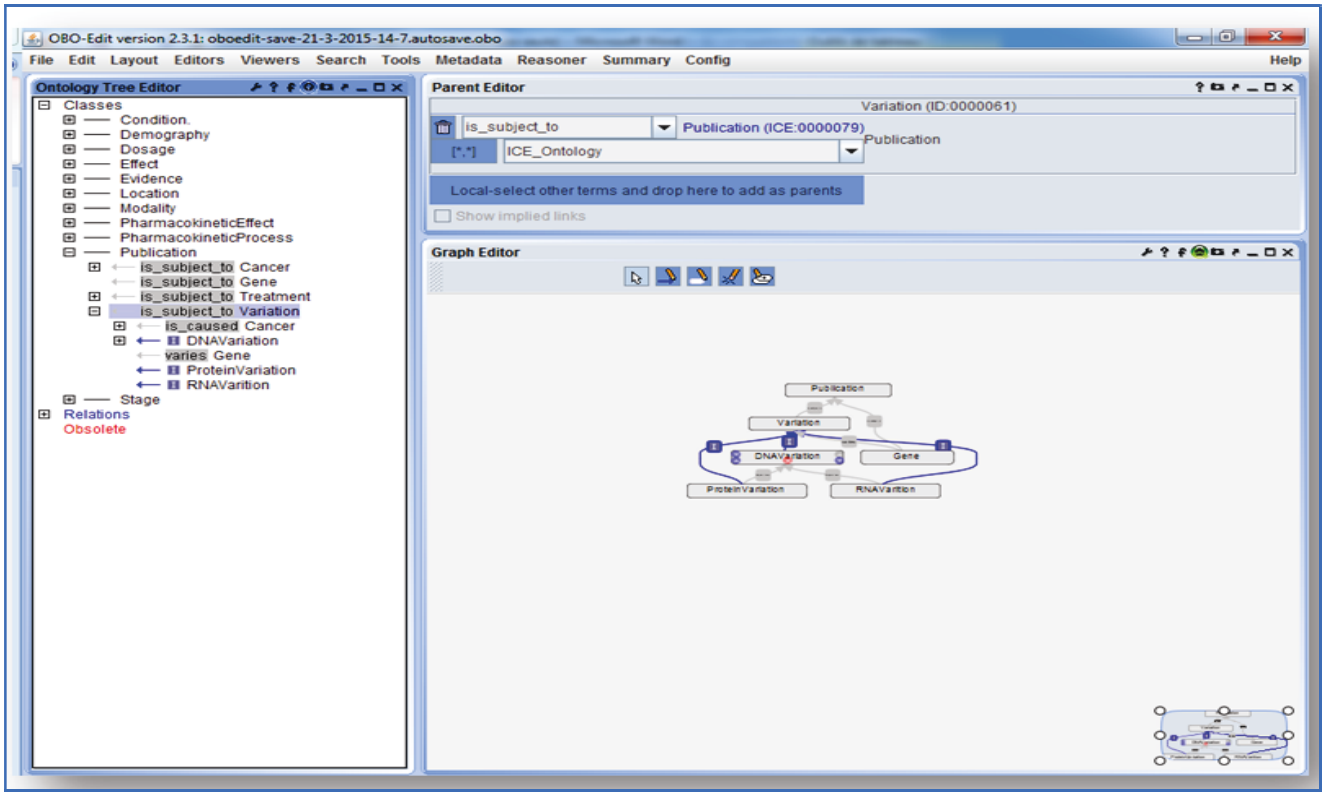


Figure 42 : Vue de la classe Variation dans OBO-Edit

3.5.1.1.5 LA CLASSE DRUG

La classe des médicaments a été implémentée comme suit :

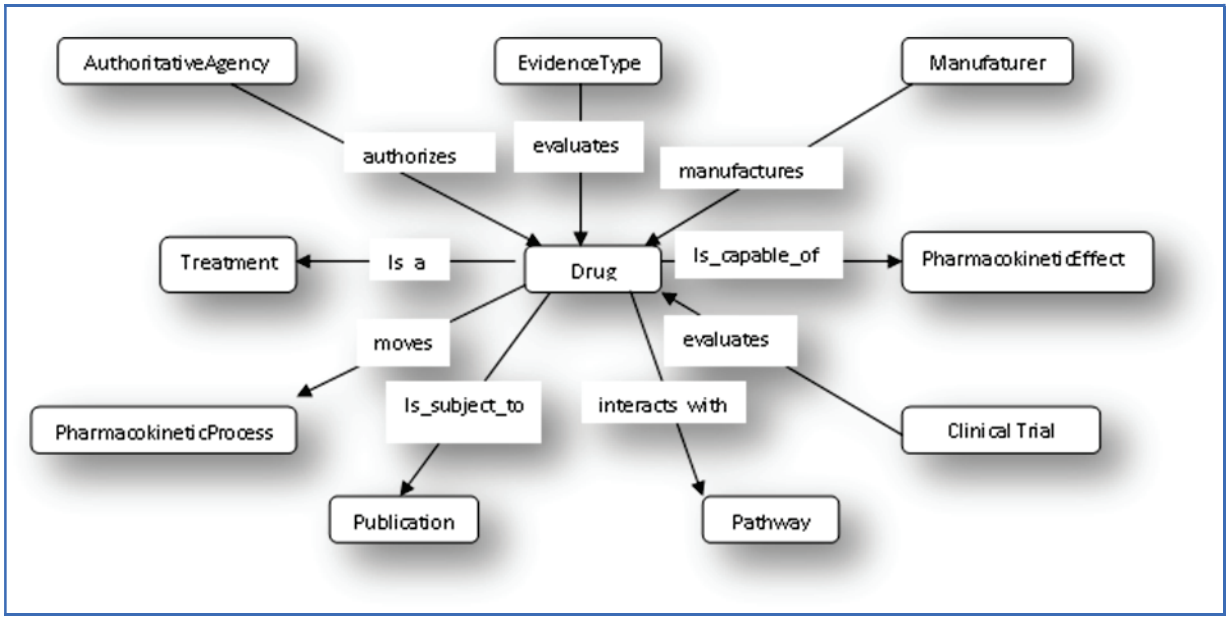


Figure 43 : Classe Drug

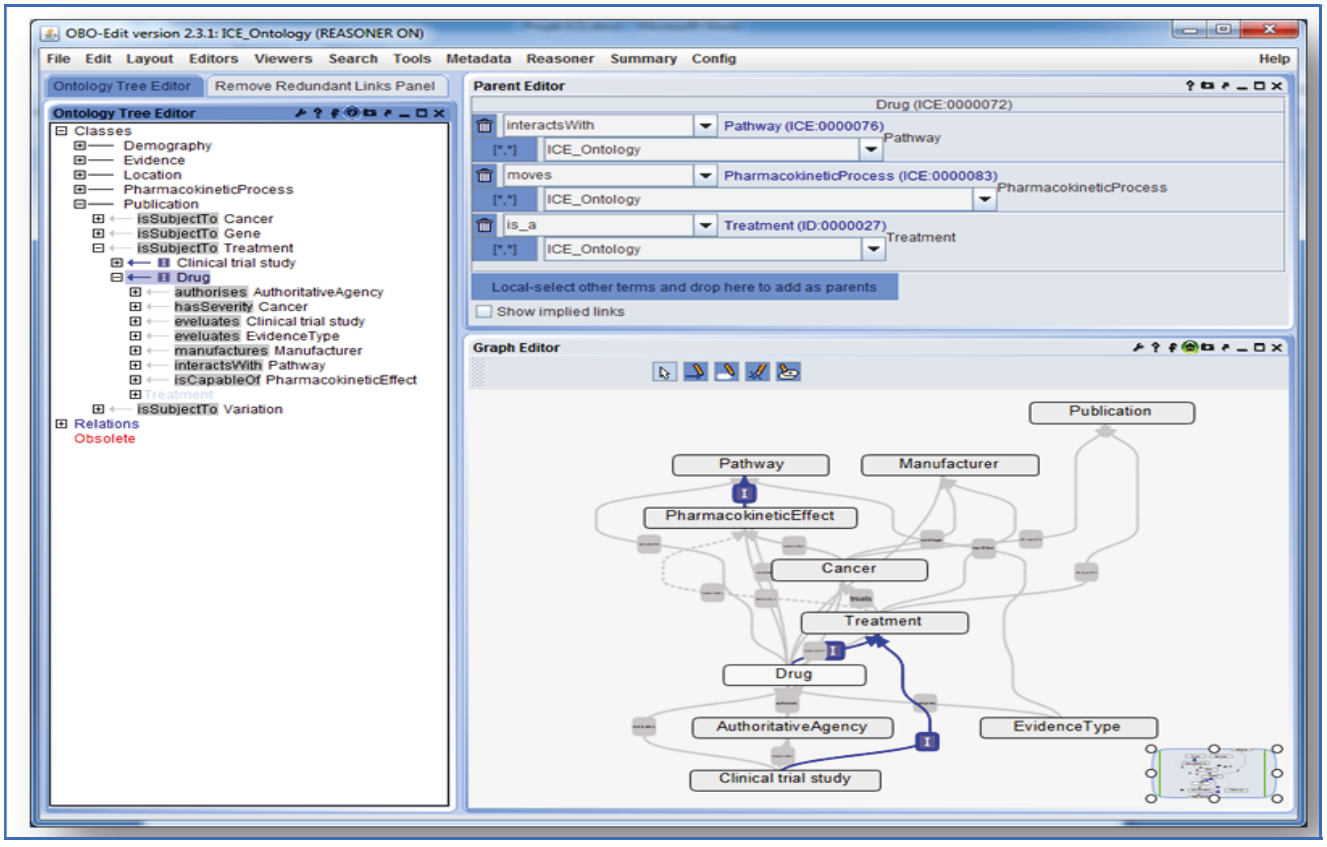


Figure 44 : Vue de la classe Drug dans OBO-Edit

3.5.1.1.6 CLASSE CLINICAL TRIAL STUDY

Cette classe se présente comme suit :

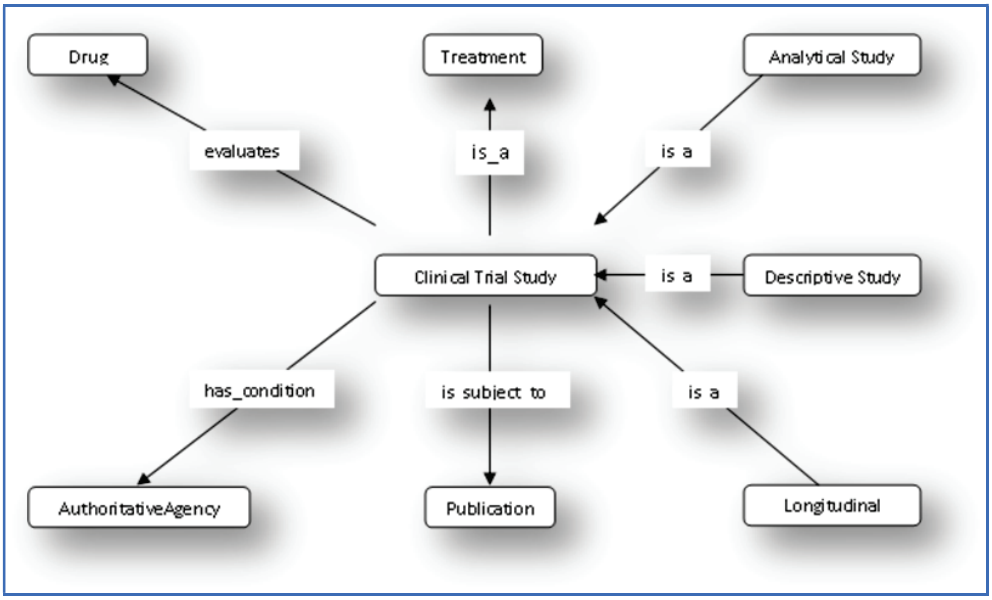


Figure 45 : Classe Clinical Trials

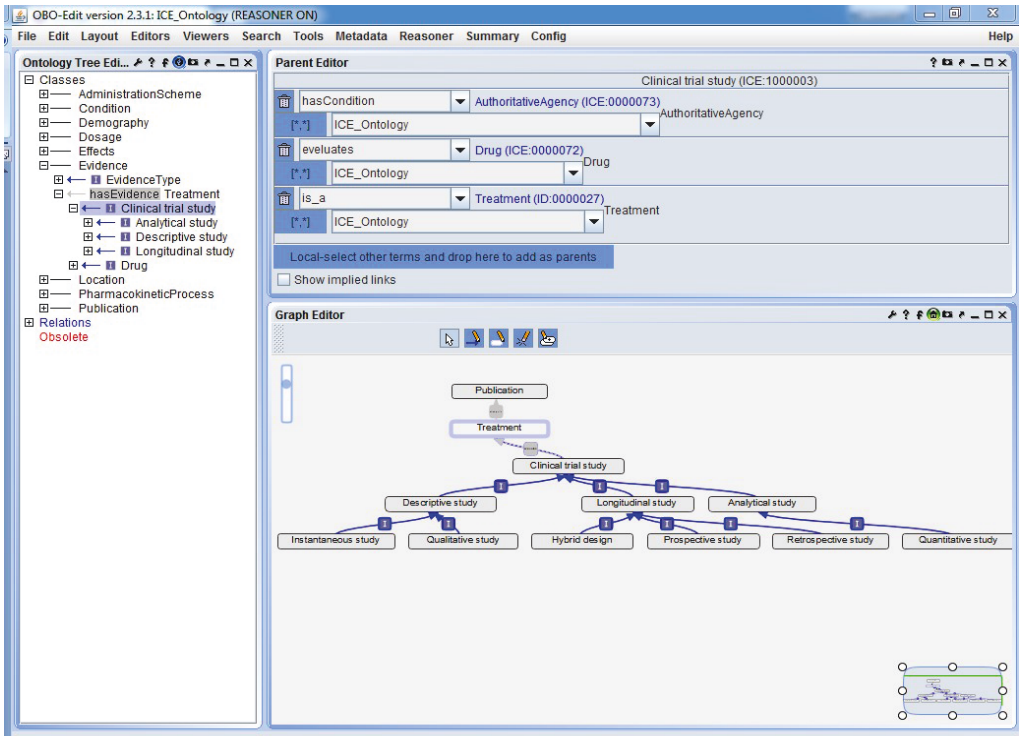


Figure 46 : Vue de la classe Clinical Trial Study dans OBO-Edit

CONCLUSION :

Dans ce chapitre nous avons abordé la transposition du modèle établi au chapitre précédent en ontologie format OBO. Nous avons, pour ce faire, utilisé l'éditeur OBO dont nous avons décrit les fonctionnalités. En y regardant de plus près, des différences apparaissent. Cependant le modèle décrit de manière satisfaisante le domaine de connaissance autour des mutants de gènes, leurs conséquences cliniques et les traitements liés.

Le but de l'ontologie étant de servir de pivot pour les ontologies existantes sur internet, le pas suivant est de faire un mapping (établir une correspondance) entre les ressources de notre ontologie et celles des ontologies déjà existantes pour commencer à peupler ICE.

3.5.2 MAPPING AVEC LES ONTOLOGIES EXISTANTES

Au sortir de l'implémentation de l'ontologie avec l'outil OBO-Edit, nous avons souhaité entreprendre le peuplement de l'ontologie : Trois solutions étaient alors envisagées :

1. Le téléchargement des ontologies existantes en local puis leur chargement dans l'ontologie ICE
2. L'importation des ontologies en utilisant les adresses URLs
3. L'alignement ou mapping des ontologies avec la nôtre aux travers d'outils comme Karma Intégration Tool ou MogMap

Etant donné la quantité de données à brasser, le téléchargement des données des données s'est révélé peu efficace. En effet les données génomiques, les variations ou encore les traitements en passant par les essais cliniques sont très volumineuses. Et ce sans compter les sites de publications comme Pubmed ou MeSH qui comportent des millions de références. Ce volume de données fait planter Protégé qui ne peut charger toutes les bases de connaissances requises.

Aussi avons-nous opté pour la troisième solution à savoir faire un mapping hors Protégé et générer un fichier RDF directement utilisable dans la base de données HBase.

3.5.2.1 CONVERSION DU FORMAT OB EN FORMAT OWL

Le format OBO-Edit est assez confidentiel et utilisé uniquement par l'initiative OBO-Foundry. Aussi afin de pouvoir travailler avec d'autres applications nous avons été amenés à convertir le format OBO en OWL (Web Ontology Language) qui est le standard du W3C comme évoqué au chapitre 3.2.3

Cette conversion se fait à la sauvegarde du fichier en changeant d'adaptateur. Les images ci-dessous décrivent le processus de cette conversion :

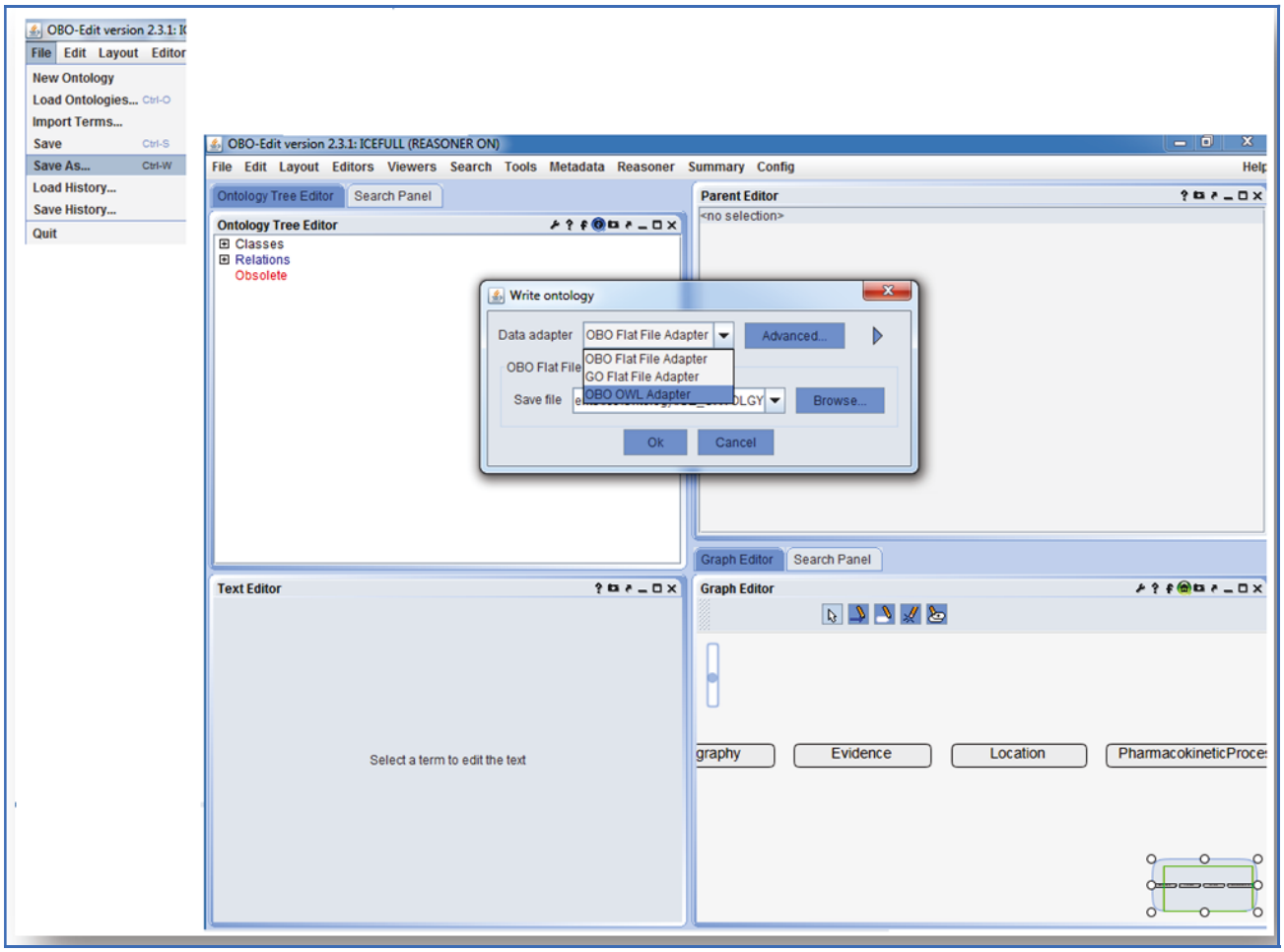


Figure 47 : Sauvegarde d'un fichier OBO sous format OWL depuis OBO-Edit

A la sauvegarde, OBO-Edit demande de choisir un adaptateur de format de fichier. Sur la sélection de l'adaptateur OBO OWL, l'écran suivant s'affiche pour confirmation et nommage du fichier.

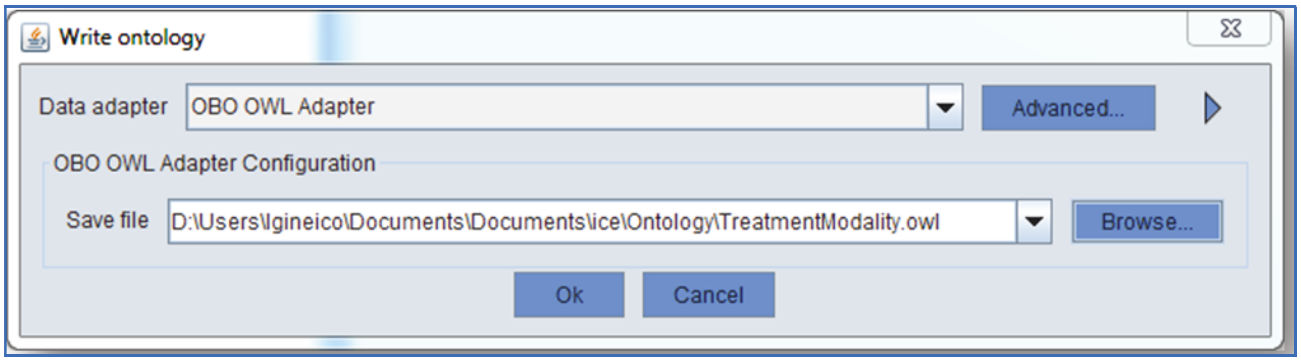


Figure 48 : Sélection de l'emplacement du fichier OWL

La fenêtre qui suit permet de donner un donner un nom au fichier OWL dans le répertoire sélection à la figure précédente.

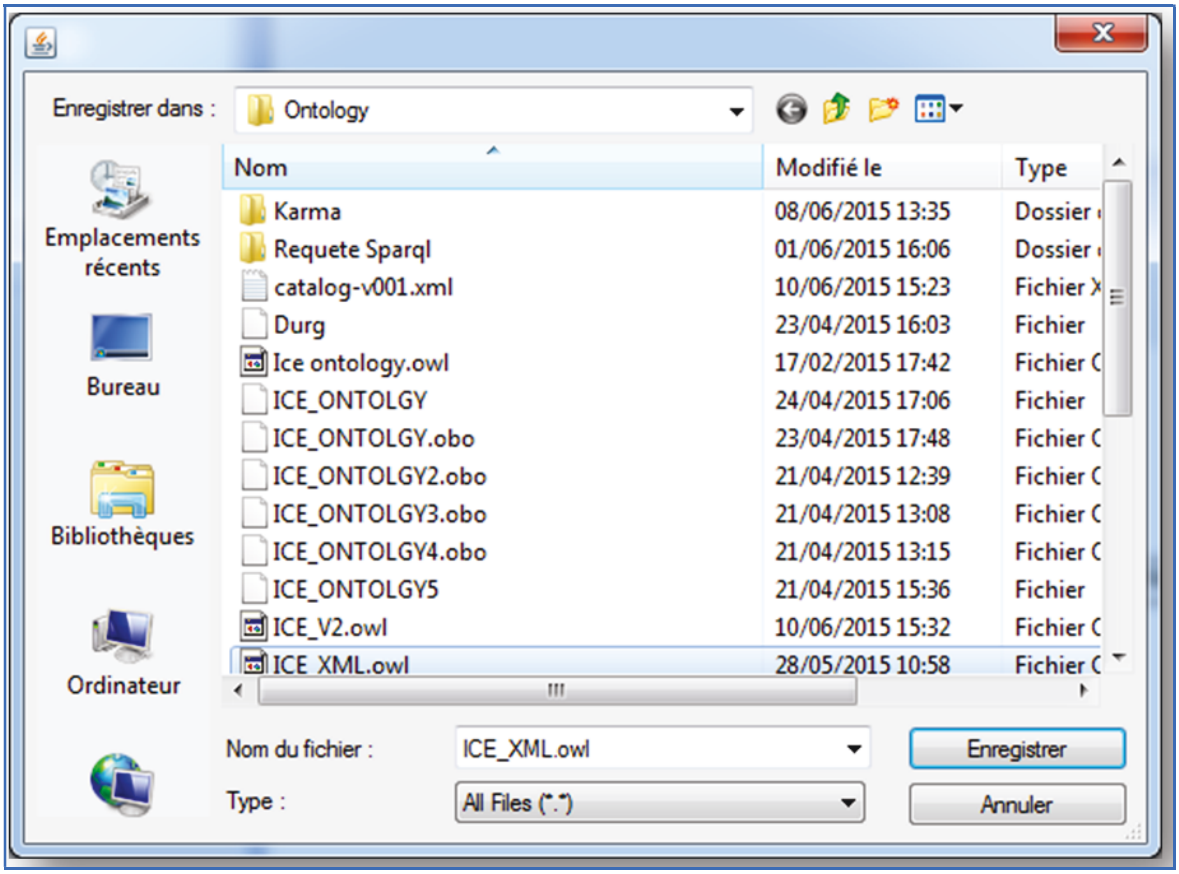


Figure 49 : Enregistrement de l'ontologie au format OWL

Une fois le fichier converti en format OWL nous envisageons alors de faire le mapping avec l'outil Karma

3.5.3 PRESENTATION DU LOGICIEL KARMA (KARMA, 2013)

Karma se présente comme un outil d'intégration d'informations qui permet aux utilisateurs d'intégrer très facilement et très rapidement des informations de sources variées et de différents formats. Ces informations peuvent être issues de bases de données, de fichiers CSV, XML JSON et bien d'autres formats encore. L'intégration des données se fait par la modélisation selon le modèle d'une ontologie de son choix au travers d'une interface graphique. Karma apprend à reconnaître le mapping des données sur les classes et utilise l'ontologie pour proposer un modèle qui lie les classes entre elles. Il est possible pendant ce processus de transformer les données selon ses besoins pour normaliser les différents formats. Une fois le modèle finalisé, l'outil permet de générer un fichier au Format RDF ou un autre format, ou encore de le stocker dans une base de données.

La force de Karma réside dans l'utilisation de la programmation par l'exemple, des techniques d'apprentissage et de l'algorithme d'optimisation de l'arbre de Steiner pour faire coïncider les modèles avec l'ontologie cible en ajustant automatiquement le modèle généré grâce

à son interface graphique sans affronter les règles très compliquées de mapping des autres systèmes.

A la suite de l'installation nous entreprenons de travailler avec Karma. Pour ce faire nous allons vous présenter ses fonctionnalités.

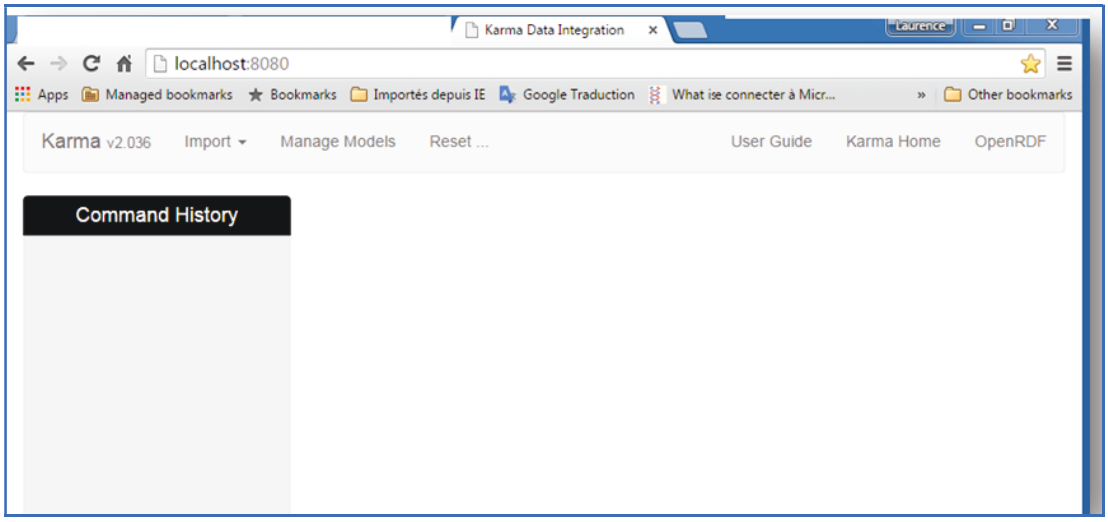


Figure 50 : Page d'accueil de Karma

3.5.3.1 IMPORTATION DE FICHIERS

Karma offre une interface qui permet d'importer des fichiers différents sources:

- Base de données
- SQL
- De service
- De fichier

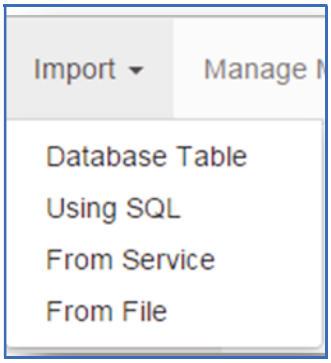


Figure 51: Formats de fichiers pour importation

3.5.3.2 GESTION DES MODELES DE FICHIERS

Karma permet de charger, supprimer ou rafraîchir des modèles et de savoir l'heure de la dernière publication.

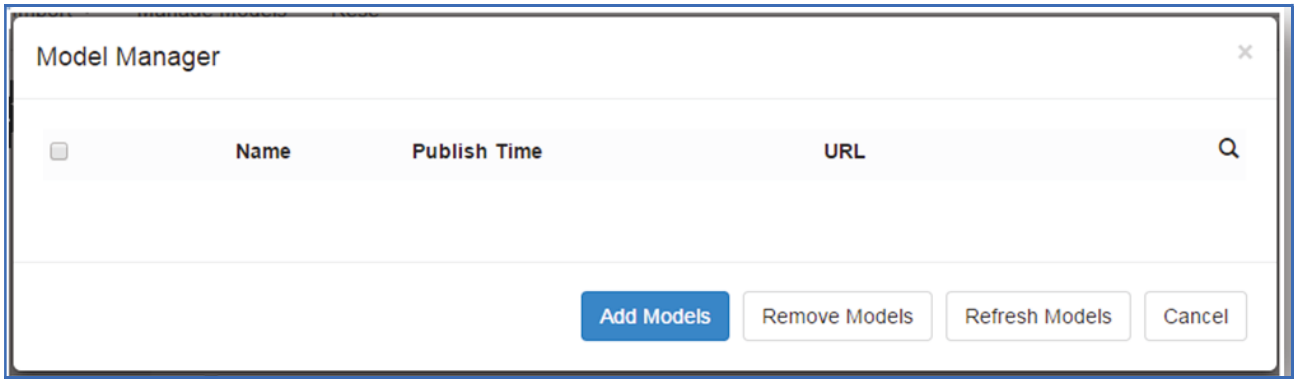


Figure 52 : Gestion des modèles dans Karma

3.5.3.2.1 LE MENU RESET

Ce menu permet de faire un RAZ en effaçant tous les types sémantiques « appris » par Karma lors de précédentes opérations de mapping.

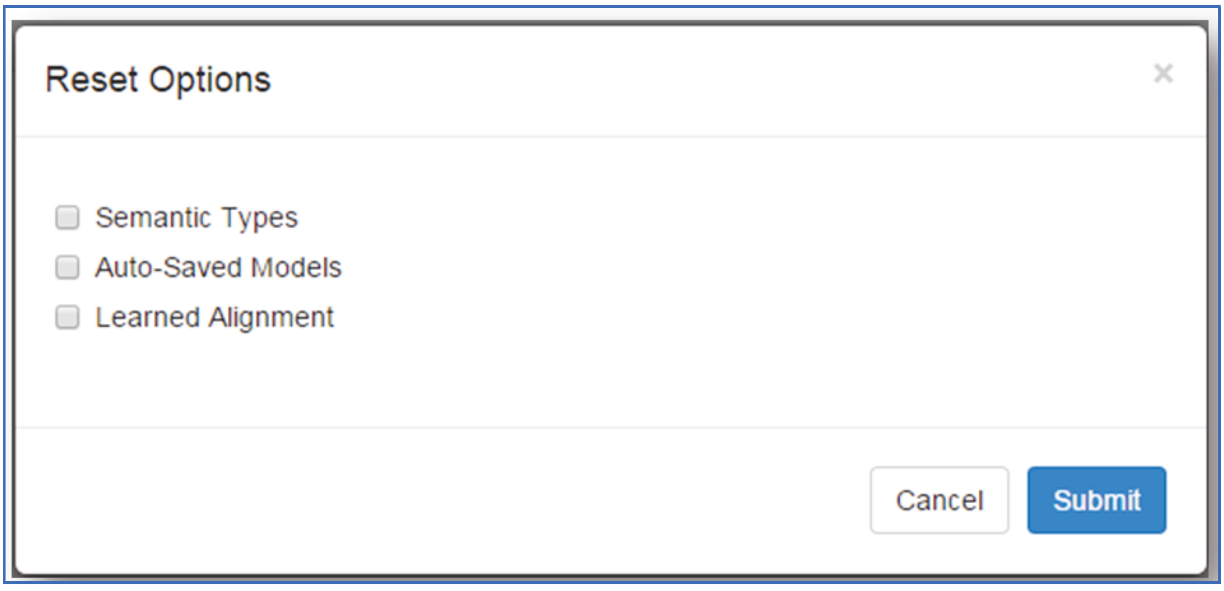


Figure 53 : Menu Reset

Après l'installation et le lancement de l'application, le chapitre suivant traitera du mapping avec l'ontologie comme modèle.

3.5.3.3 TENTATIVE DE MAPPING AVEC LE FICHIER FORMAT OBO CONVERTI EN FORMAT OWL

Karma n'est pas en mesure de traiter des fichiers OBO. Aussi comme évoqué au chapitre Conversion du Format OB en format OWL, nous avons dû convertir notre fichier OBO en OWL pour faire le mapping.

Le mapping se fait en trois étapes :

1. Chargement du modèle et des fichiers,
2. Transformation si nécessaire du fichier cible
3. Publication d'un nouveau fichier RDF ou d'un nouveau modèle.

3.5.3.3.1 IMPORTATION DES MODELES

L'importation des fichiers se fait au travers du menu import :

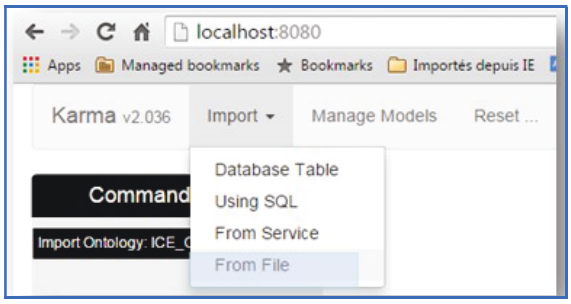


Figure 54 : Etape 1 choix de la source

Nous choisissons l'option « From File ». Ensuite nous sélectionnons le fichier OWL correspondant

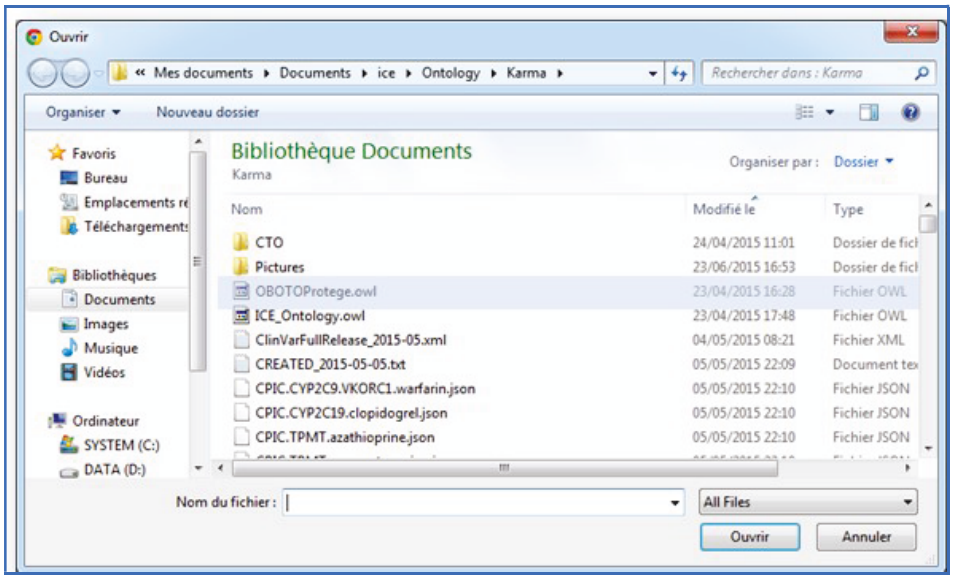


Figure 55 : Etape 2 Sélection de fichier

A la suite il faut préciser le format du fichier à sélectionner

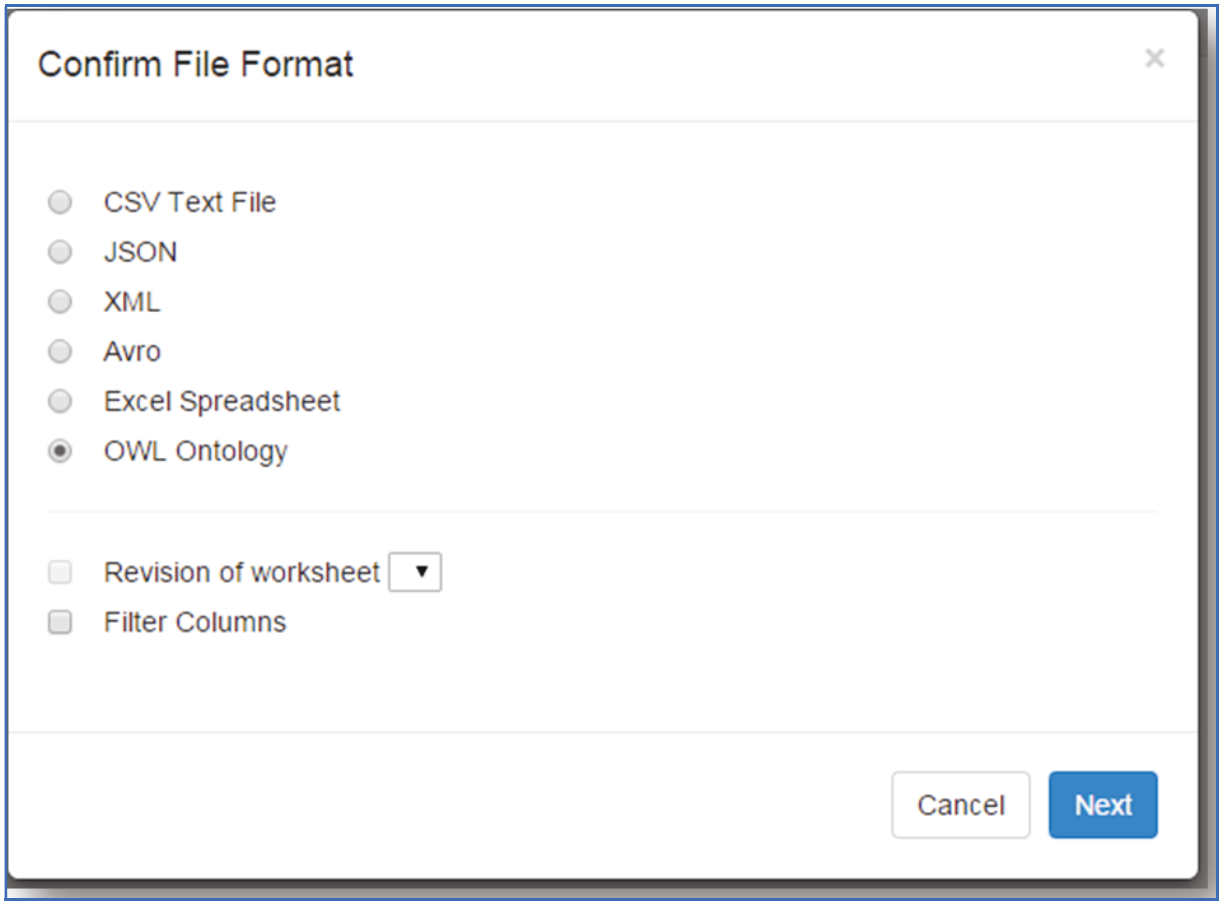


Figure 56 : Etape 3 Confirmer le format d'importation

Une fois le format de fichier sélectionné et le bouton « Next » cliqué, il faut valider les options de l'import.

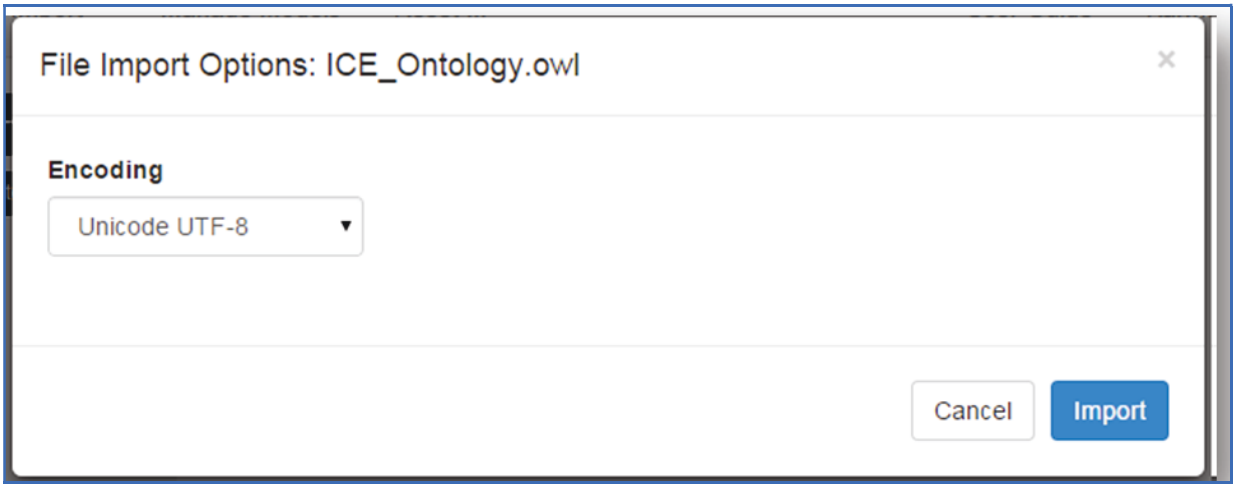


Figure 57 : Etape 4 Confirmer l'import du modèle

Le modèle importé s'affiche et le bouton « Import » permet de finaliser l'action. On obtient la page ci-dessous :

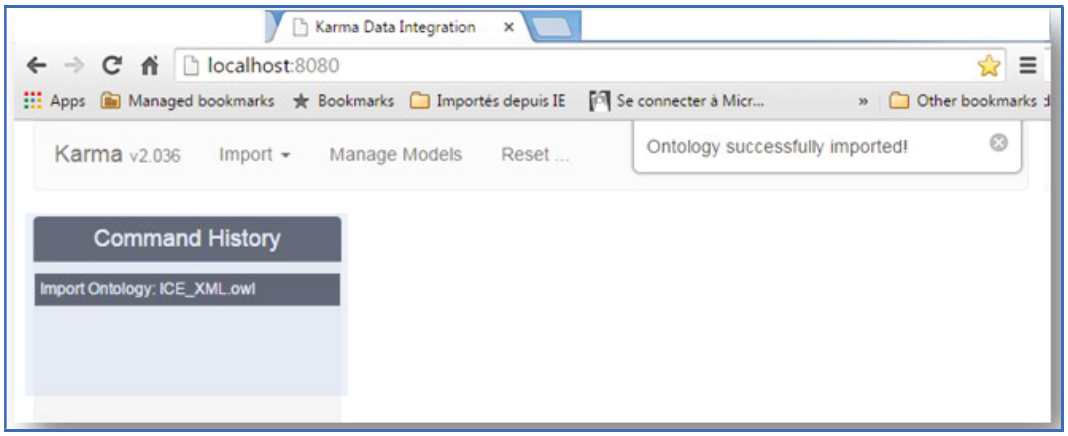


Figure 58 : Etape 5 Modèle importé avec succès

Une fois le modèle d'ontologie importé, il faut procéder de la même manière pour importer dans karma le(s) fichier(s) à mapper sur le modèle.

3.5.3.4 IMPORTATION DE FICHIER(S) POUR MAPPING

Pour ce faire les trois premières étapes de l'action précédente sont identiques. Depuis le menu Import cliquer sur « import from file » avec le modèle dans l'historique des commandes dans la barre à gauche :

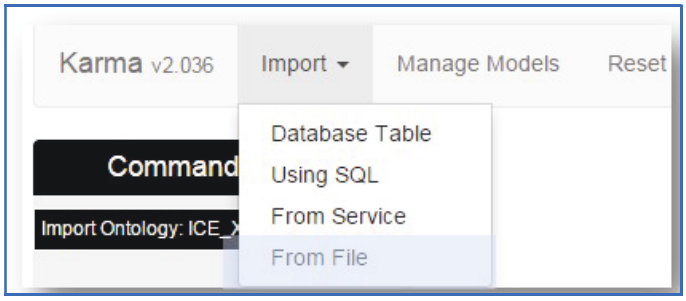


Figure 59 : Etape 1 Import de fichier

A la suite, il faut sélectionner le fichier à importer ; dans notre cas nous choisissons d'importer un fichier au format TSV

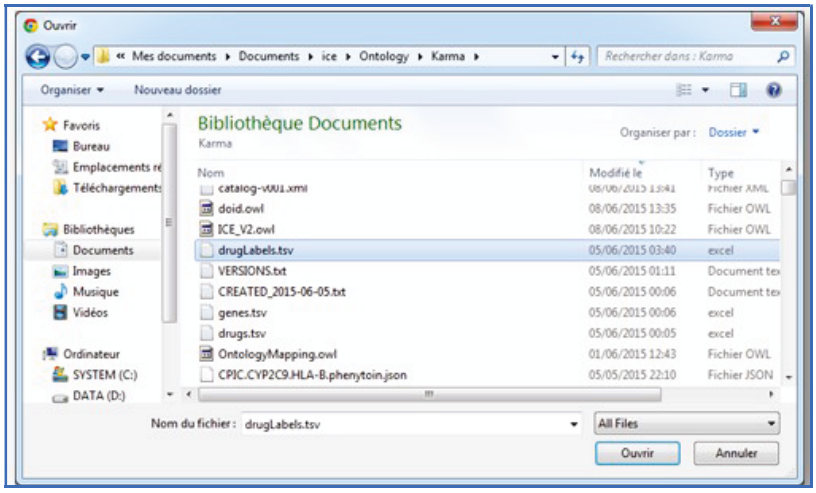


Figure 60 : Etape 2 sélection du fichier TSV

Lorsque le fichier est sélectionné, il faut alors indiquer le format d'importation

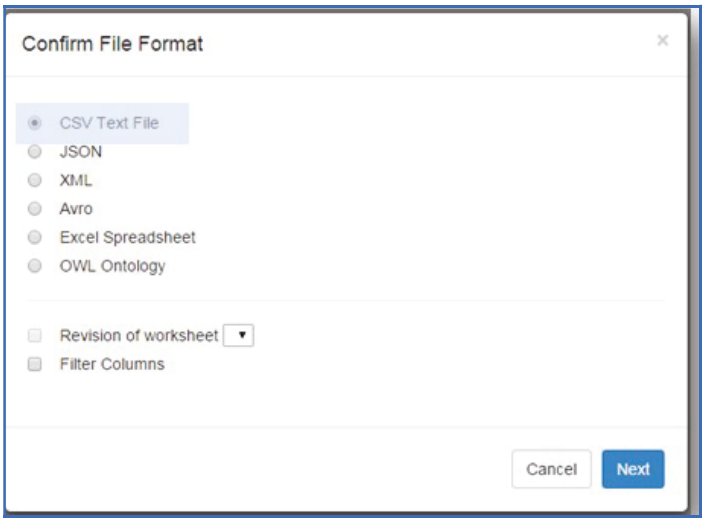


Figure 61 : Etape 3 confirmation du format d'import

Ensuite, il convient d'indiquer le délimiteur de colonne. Nous sélectionnons Tab et la fenêtre nous présente les informations en prévisualisation.

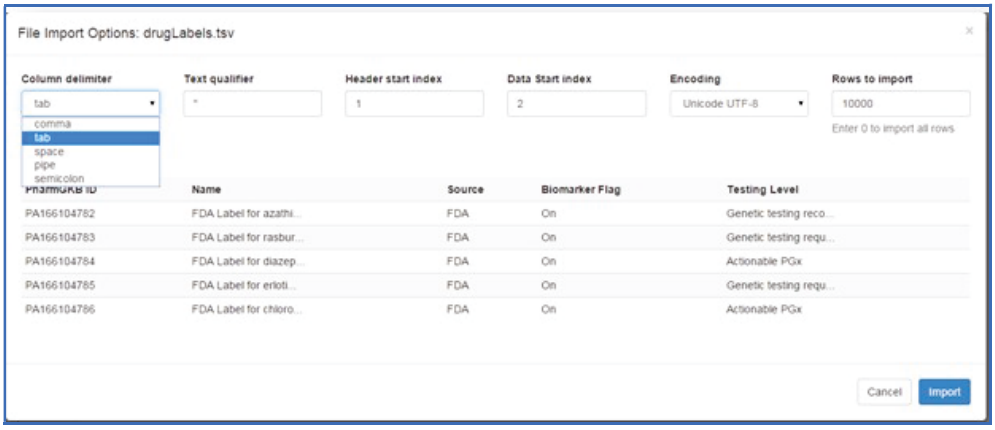


Figure 62 : Etape 4 sélection du délimiteur de colonne

Cette figure nous donne un aperçu de la disposition des données une fois le délimiteur de colonnes sélectionné.

En cliquant sur le menu import, nous importons le fichier dans Karma

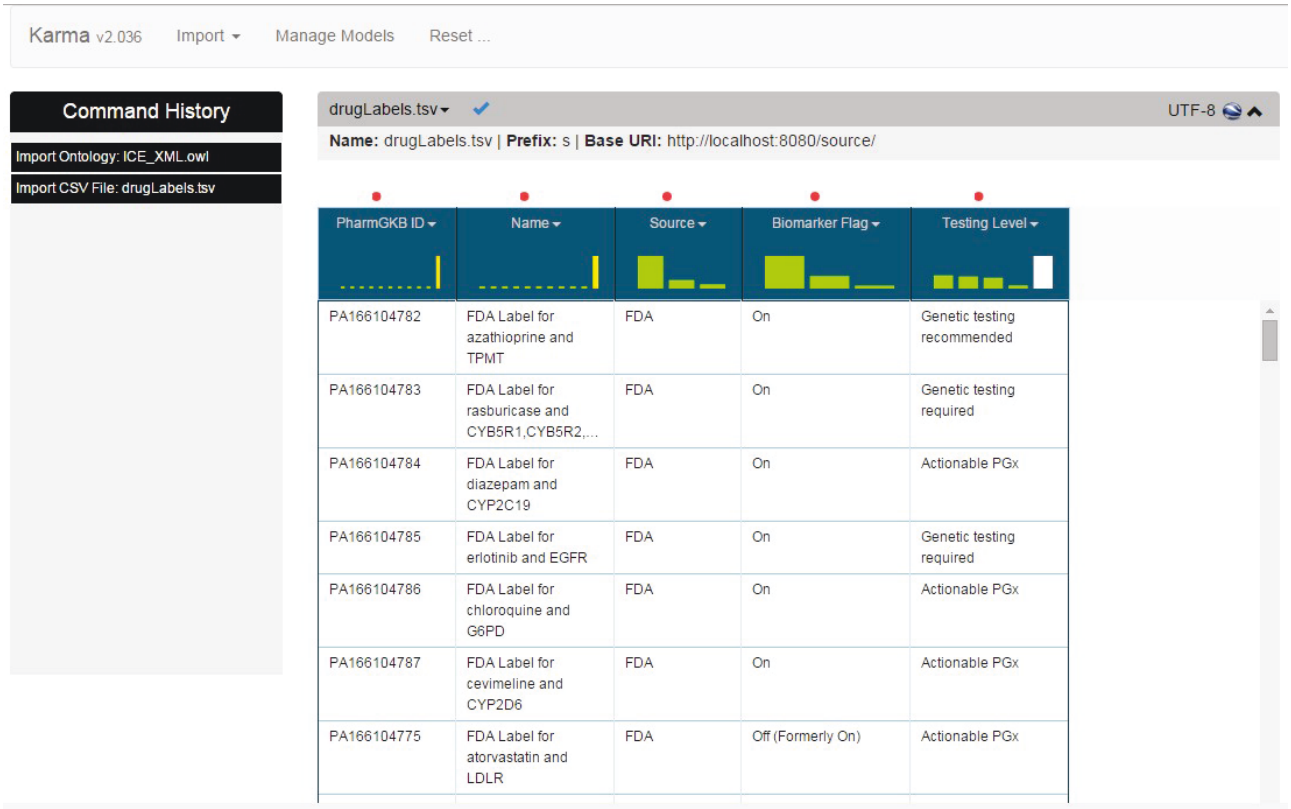


Figure 63 : Fichier Drug importé du site PharmGKB

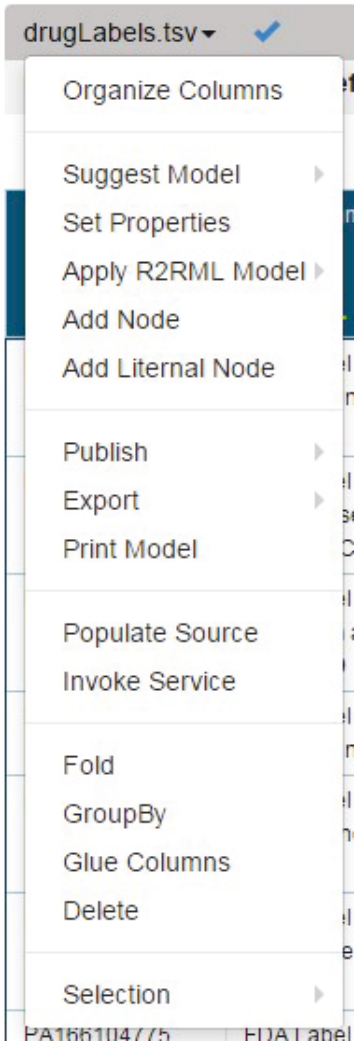
Il s'agit d'un fichier téléchargé depuis le site PharmGKB listant tous les médicaments et leurs voies d'interaction (Pathway) dans l'organisme. Dans ce fichier tous les points rouges représentent des classes dans l'ontologie d'où a été exporté ce fichier. Le bandeau à gauche de l'écran retrace tout l'historique des actions de l'instance Karma en cours. Dans ce cas nous avons importé successivement une ontologie et un fichier au format TSV. Le mapping consistera à faire coïncider les colonnes de ce fichier à notre ontologie modèle.

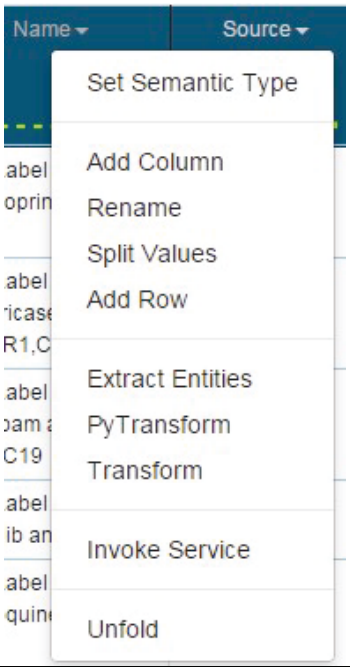
Le chapitre qui suit indique comment nous avons mené cette action.

3.5.3.5 MAPPING DE FICHIER SUR NOTRE MODELE D'ONTOLOGIE

Une fois le fichier chargé, deux menus sont alors accessibles permettant de gérer les fichiers.

1. Le menu niveau fichier : c'est un menu déroulant qui s'affiche en cliquant sur l'icône au niveau du nom de fichier il permet de :
   - D'organiser les colonnes
   - Définir des propriétés ou de gérer des nœuds dans le modèles,
   - De publier, exporter ou imprimer un modèle
   - Peupler la source



1. Au niveau des colonnes, le menu permet de : - transformer des colonnes avant manipulation, - leur attribuer des types sémantiques - de rajouter ou supprimer des colonnes - etc.	

La première opération consiste à transformer les colonnes notamment à définir des URIs puis à leur attribuer des types sémantiques pour faire coïncider les colonnes et les propriétés de notre modèle.

3.5.3.5.1 TRANSFORMATION AVEC PYTRANSFORM POUR CREER DES URIS

Nous utilisons le menu PyTransform pour définir des URI afin d'avoir des identifiants uniques que nous pourrons utiliser dans le fichier RDF qui sera publié à la fin du processus.

Pour ce faire au niveau de la colonne sur laquelle nous souhaitons définir l'URI, nous cliquons sur le menu déroulant pour sélection l'action « ***PyTransform*** ». Nous créons une deuxième colonne pour d'URI. L'exemple ci-dessous nous montre comment nous avons procédé pour la colonne « ***Name*** ».

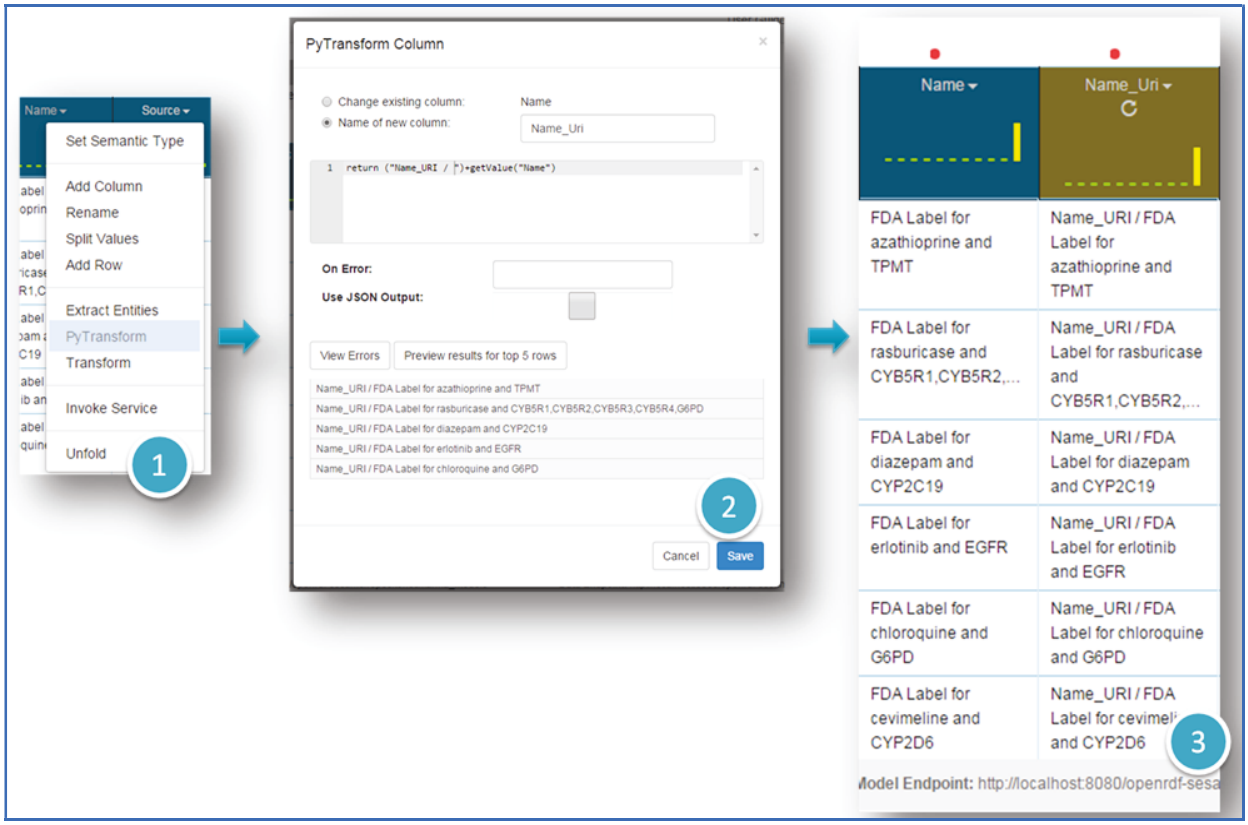


Figure 64 : PyTransform pour créer des colonnes URI

Cette action sera répétée sur chaque colonne du fichier et au final nous obtiendrons la transformation ci-dessous :

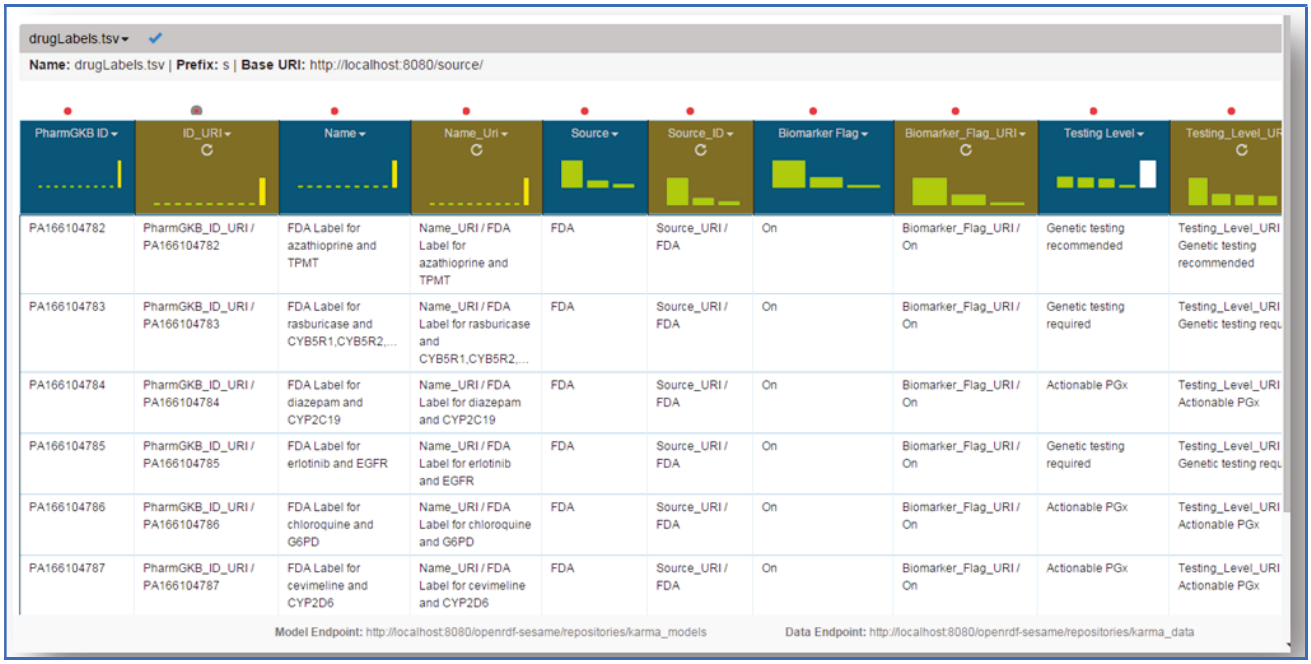


Figure 65 : Colonnes d'URI insérées après transformation

Cette figure nous montre que pour chaque colonne d'information du fichier originel, nous avons inséré une colonne où nous avons récupéré la valeur du champ en y accolant devant une chaîne de caractères rappelant l'information de la colonne.

Toutes les actions sont ainsi récapitulées dans le bandeau « Command History». Si l'on souhaite annuler une action, il suffit de cliquer sur la flèche pour éliminer cette action.

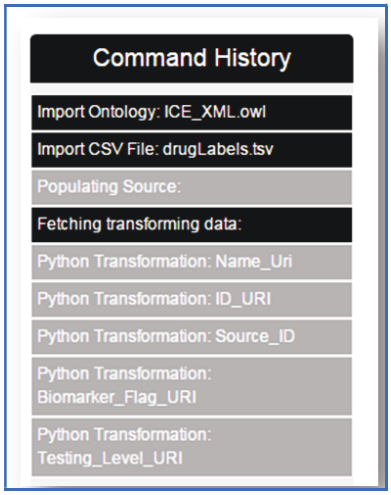


Figure 66 : Karma action log

Une fois toutes les transformations faites, il faut indiquer un « type sémantique » pour chaque colonne. C'est-à-dire faire correspondre la colonne à une propriété d'une classe de l'ontologie modèle.

Le chapitre suivant nous montre le mapping à proprement parler.

3.5.3.6 MAPPING FICHIER/MODELE

Nous essayons ici de faire matcher les colonnes du fichier avec les propriétés des classes. Pour ce faire le menu « Set Semantic Type » est utilisé.

Dans un premier temps, nous essayons de définir des URIs avec les colonnes d'URI que nous avons créées précédemment :

3.5.3.6.1 DEFINITION D'URIS

Nous procédons comme suit :

1. Cliquer sur « Set Semantic Type » dans le menu déroulant
2. Cocher l'option « URI of Class »

3. Cliquer sur le bouton “Edit” pour sélectionner la classe dont on souhaite définir une nouvelle URI

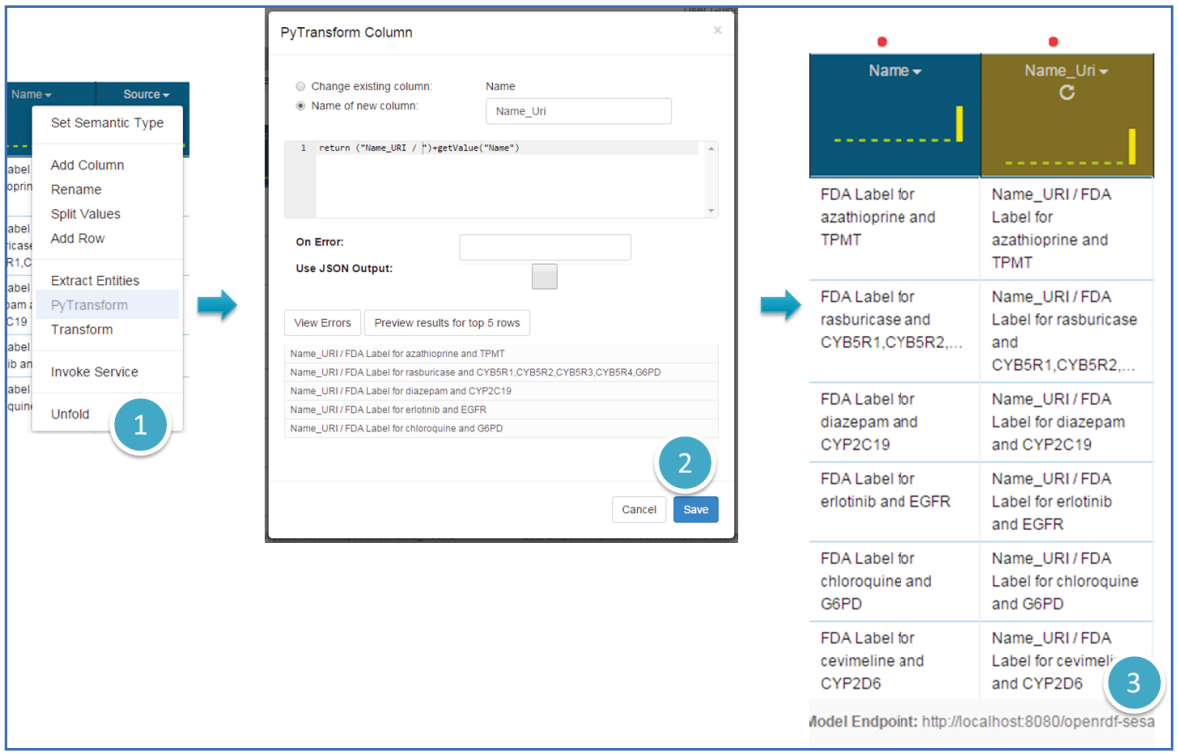


Figure 67 : Semantic Type URI

A notre surprise Karma n’affiche pas les intitulés des classes. Nous retentons l’expérience avec les propriétés :

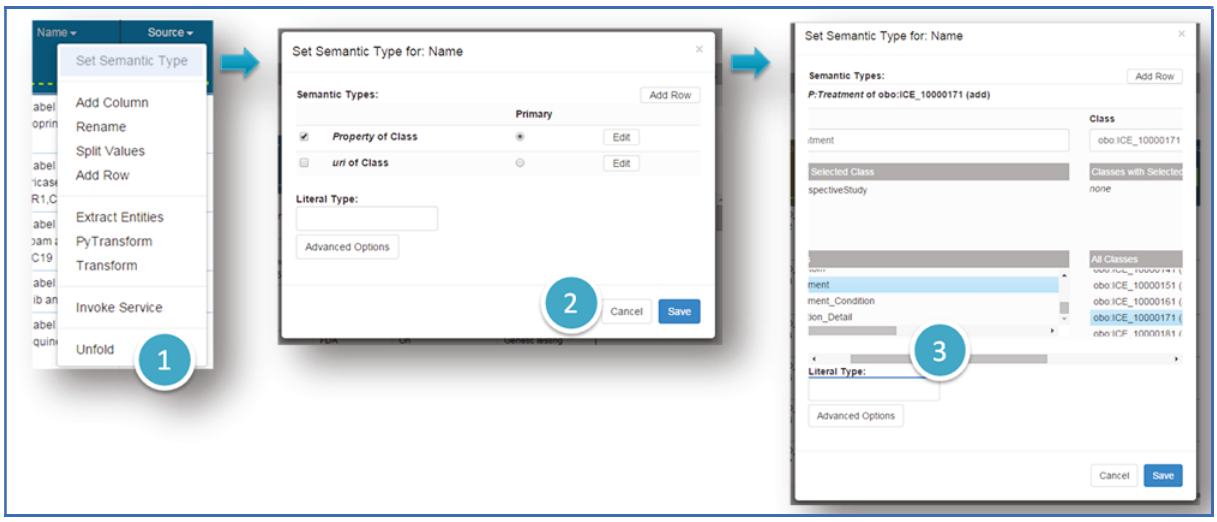


Figure 68 : Semantic Type Class

Nous revenons au même constat : Karma affiche des identifiants qui ne donnent aucune indication sur les classes auxquelles ils sont censés se référer. Après plusieurs tentatives soldées

par un même échec, nous décidons de transférer notre modèle sous l'outil Protégé et voir ce qu'il se passe. D'autant que nous avons fait une autre tentative avec un autre logiciel dit Logmap[5] dont le résultat n'a pas été concluant.

5 LogMap est un système de mapping d'ontologie hautement évolutif avec un raisonneur integer et des capacities de corrections de coherence.

3.5.4 PROTOTYPE 2 : CONVERSION OBO EN OWL SOUS PROTEGE

La tentative de mapping sous Karma ayant échoué avec le modèle directement converti en OWL depuis OBO, nous avons transcrit le modèle d'ontologie sous Protégé afin de voir si le problème ne venait pas d'une incompatibilité avec le modèle généré depuis OBO-Edit.

L'ontologie transférée sous Protégé se présente comme suit :

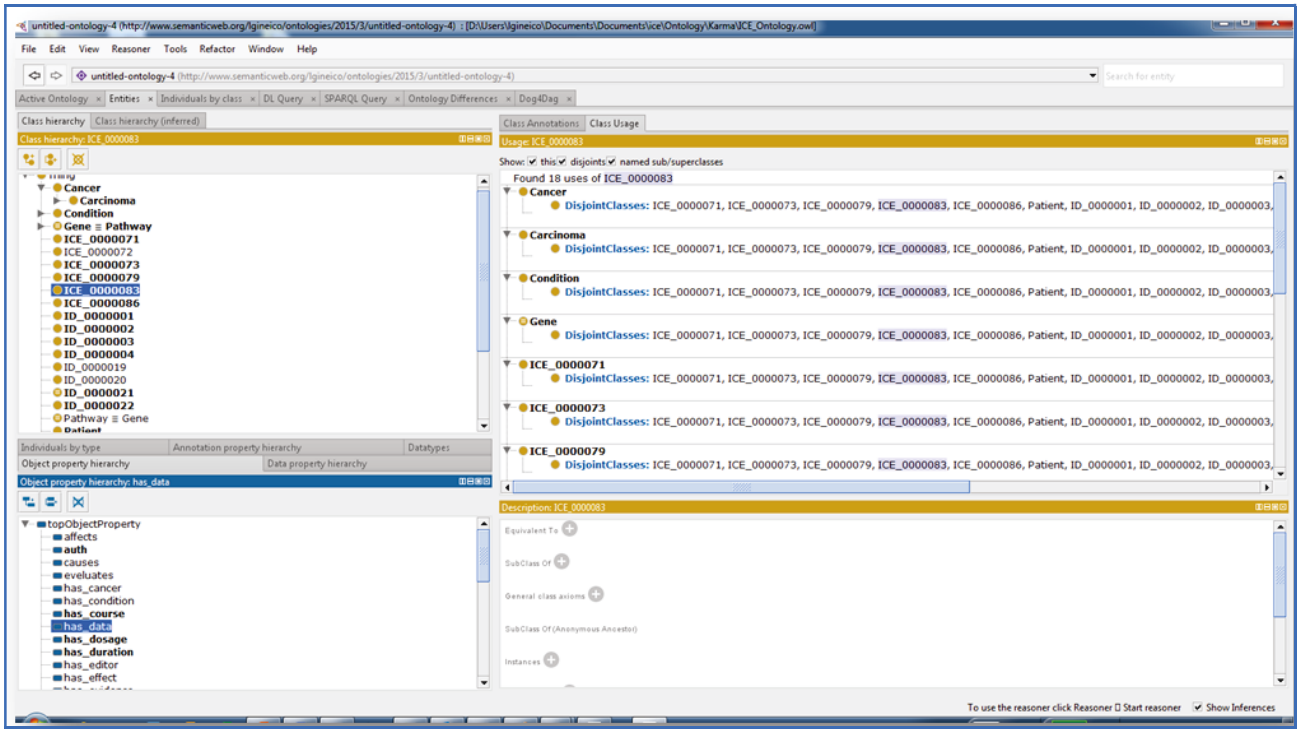


Figure 69 : Ontologie transférée sous Protégé

Sur cette figure nous notons la présence d'identifiants comme sous-classes de l'entité Pathway. Ces mêmes identifiants que nous retrouvons sous Karma sans pouvoir déterminer à quelles classes ils se réfèrent. Nous décidons de supprimer ces classes fantômes pour ne faire apparaître que les classes que nous avons effectivement créées sous OBO-Edit.

Nous tentons à nouveau de faire un mapping avec karma mais obtenons l'image ci-dessous :

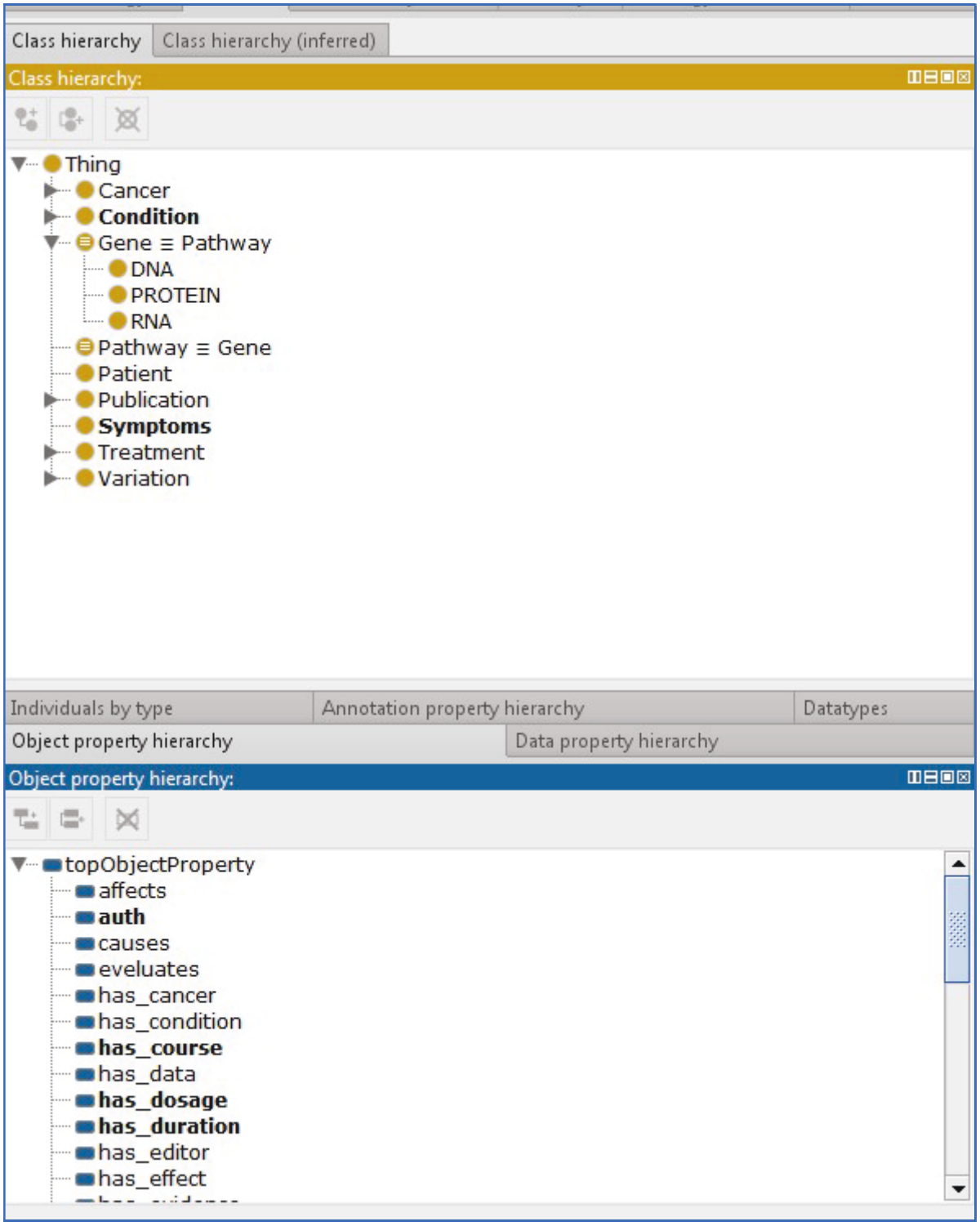
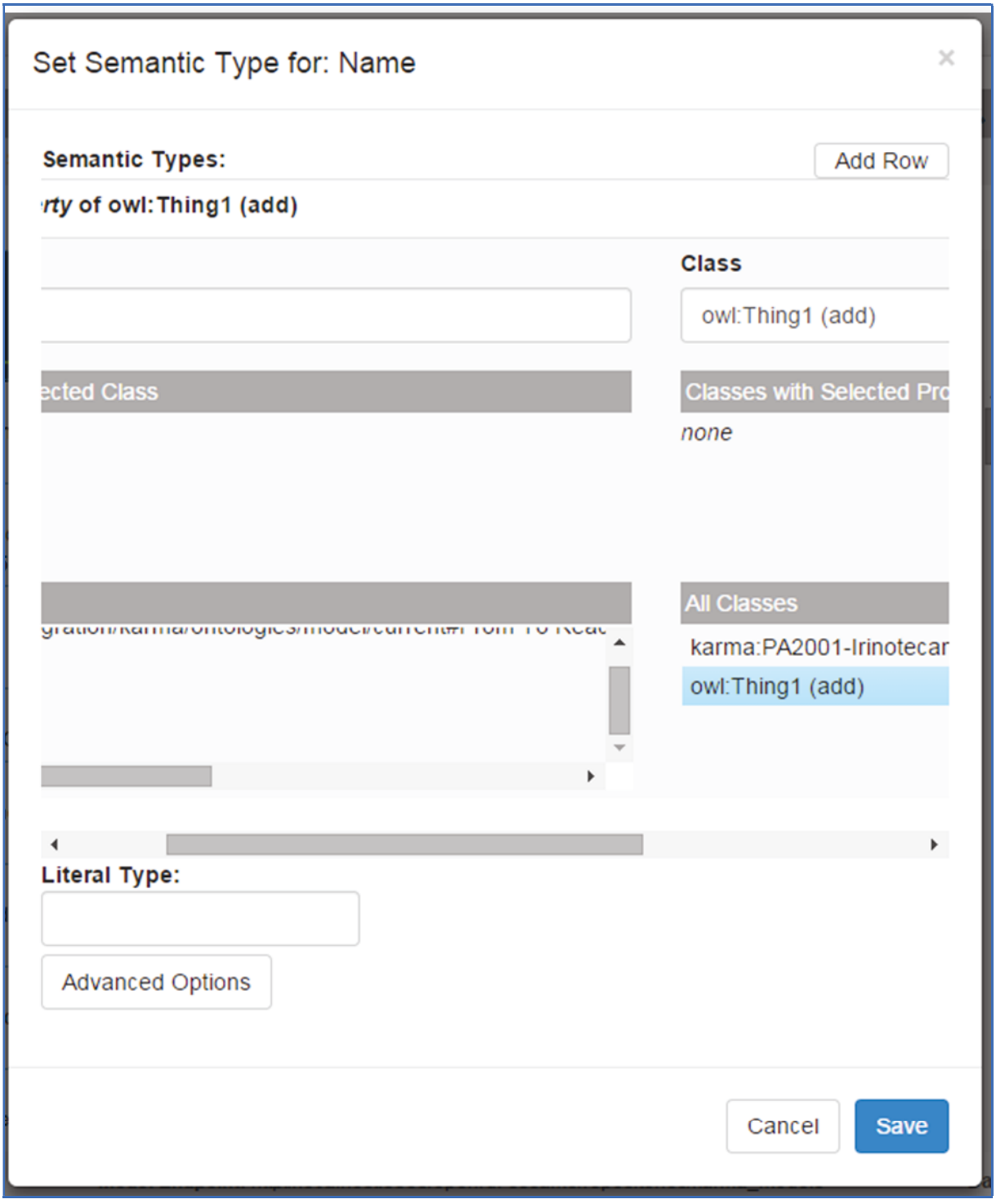


Figure 70 : Classes et Object Properties dans Protegé

Ici toutes les entités et leurs propriétés ont disparu. Nous constatons que ce modèle-ci ne fonctionne pas.



Il nous faut envisager d'autres solutions. Nous entendons parler d'un outil dit LogMap qui ferait la même chose que Karma. Nous décidons de faire le mapping avec cet outil.

3.5.4.1.1 MAPPING AVEC LOGMAP

Nous décidons d'utiliser l'interface Web de cet outil pour faire le mapping. Cette opération se fait en trois étapes :

1. Remplir le formulaire en ligne avec les modèles et fichiers

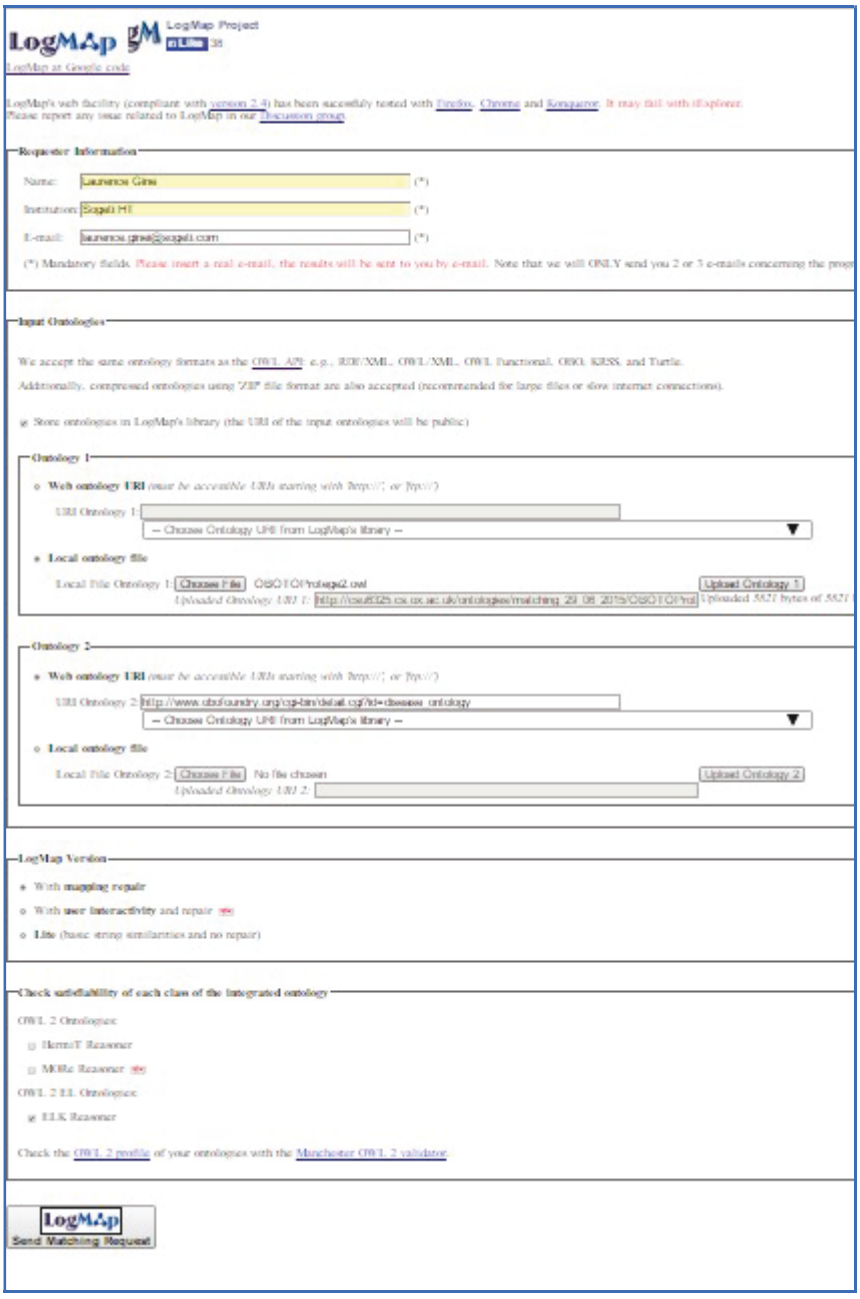


Figure 71: LogMap - Formulaire de demande de mapping

2. Envoyer à LogMap pour matching

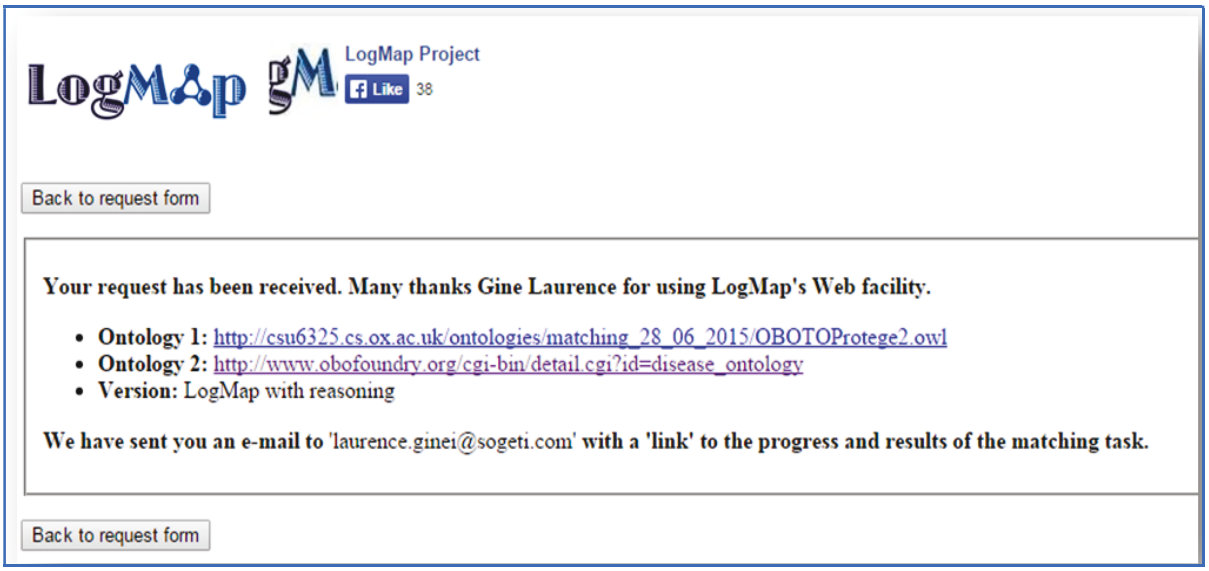


Figure 72 : LogMap - Confirmation de transfert

3. Recevoir le fichier RDF du nouveau modèle.

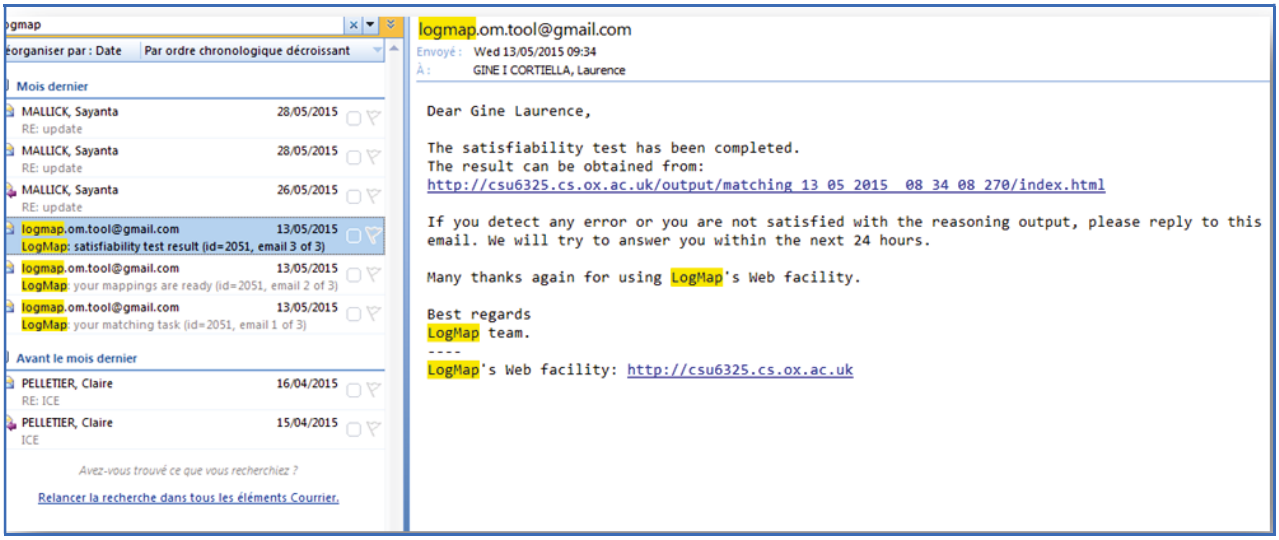


Figure 73 : LogMap - Courriel de confirmation de réception et de traitement de la demande

Lorsque nous cliquons sur le lien dans le mail ci-dessus, nous constatons l'échec de notre démarche

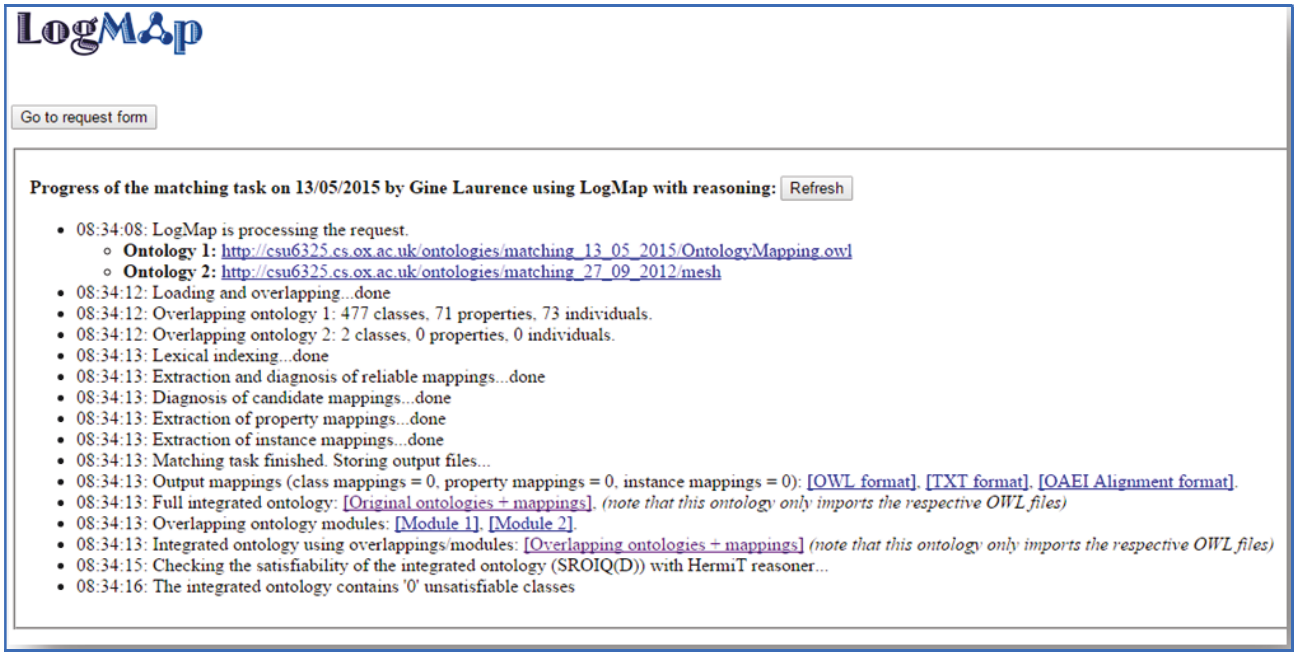


Figure 74 : LogMap - Echec du mapping

Ce deuxième échec nous convainc que Karma n'est pas à incriminer mais que c'est le modèle qui ne fonctionne pas. Il s'avère qu'OBO diverge d'OWL car le metamodèle OBO n'est pas représenté par un triple mais les entités et leur propriété sont représentées comme des sous-classes des unes des autres ce qui casse le modèle OWL. Par conséquent les outils se basant sur le métamodèle OWL comme Karma ne peuvent gérer des fichiers OBO.

Par ailleurs aux détours d'autres essais avec Karma, nous constatons sur le modèle précédent, des objets que nous avons créés dans Protégé remontent sans problème dans Karma. Nous décidons alors de recréer le modèle de zéro dans Protégé.

Le prochain chapitre nous indique comment nous avons construit ce troisième prototype.

3.5.5 PROTOTYPE 3 : OWL FROM SCRATCH

Ayant constaté une incompatibilité entre OBO et Karma, nous avons décidé de changer notre fusil d'épaule et de recréer l'ontologie en partant de zéro sous Protégé. Cette ontologie se présente comme suit :

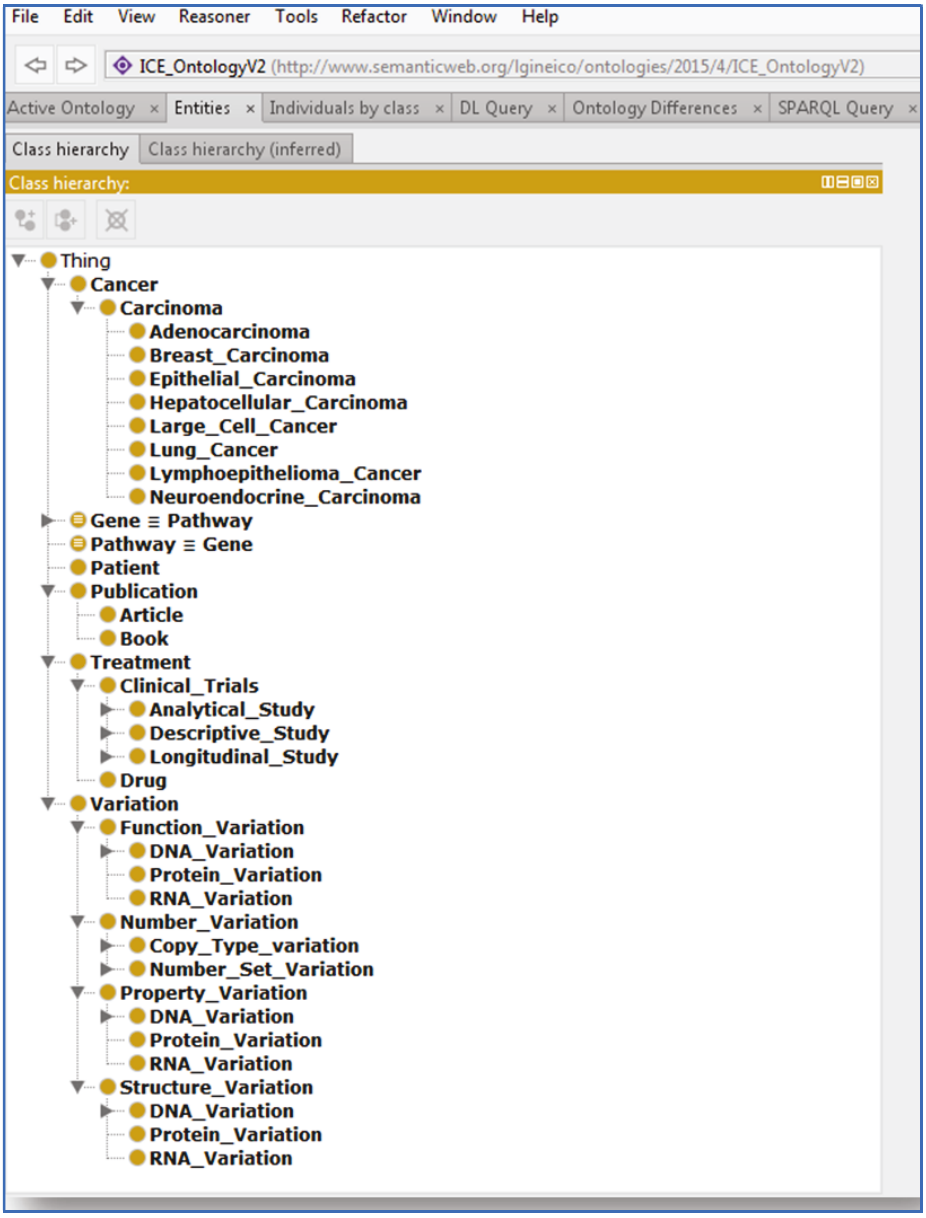


Figure 75 : Classe de l'ontologie ICE sous Protege

Cette ontologie reprend la majorité des classes de l'ontologie sous OBO. Cependant étant donné que sous OBO, les propriétés étaient apparentées à des sous-classes, nous avons dû réorganiser les classes pour rendre le modèle plus conforme au RDF. Cela donne les ressources suivantes :

3.5.5.1.1 LES CLASSES DE L'ONTOLOGIE

Tableau x : Les classes de l'ontologie

Classes	Sous-classes	Sous classe
Cancer	Carcinoma	Adenocarcinoma Breast_carcinoma Epithelial_carcinoma Hepatocellular_Carcinoma Lung_Cancer Lymphoethelioma_Cancer Neuroendocrine_Carcinoma
Gene	DNA	
	Protein	
	ARN	
Pathway		
Variation	Function variation	DNA_Variation Protein_Variation RNA_Variation
	Property Variation	DNA Variation Protein_Variation RNA_Variation
	Structure Variation	
	Number variation	Copy_Type_Variation Number_Set_Variation
Treatment	Clinical Trials	Analytical Study
		Descriptive Study
		Longitudinal study

	Drugs	
Publication	Article	
	Book	

3.5.5.1.2 LES PREDICATS OU « OBJECTS PROPERTIES

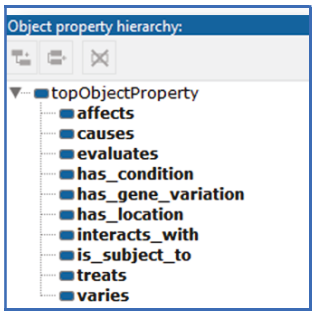


Figure 76 : Protégé List des relations (object properties)

Ces prédicats sont recensés dans le tableau ci-dessous

Tableau xi : Object Properties de l'ontologie V2

Data Properties	Description
Affects	Lie la classe Cancer à la classe Patient
Causes	Lie la Variation à la classe Cancer
Evaluates	Lie la classe Clinical Trial à la classe Treatment
has_gene_variation	
has_location	Lie la classe cancer à la propriété location
Varies	Lie la classe Gène à la classe Variation
Interacts_with	Lie la classe Pathway à la class Drug
Treats	Lie la classe Treatment à la classe cancer
Involves	Lie la classe Variation à la classe Gène
varies	Lie la classe Variation à la classe Gène
Is_subject_to	Lie les classes Gene, Variation, Treatment, Drug, Cancer, Clinical Trials à la classe Publication

3.5.5.1.3 LES DATA PROPERTIES:

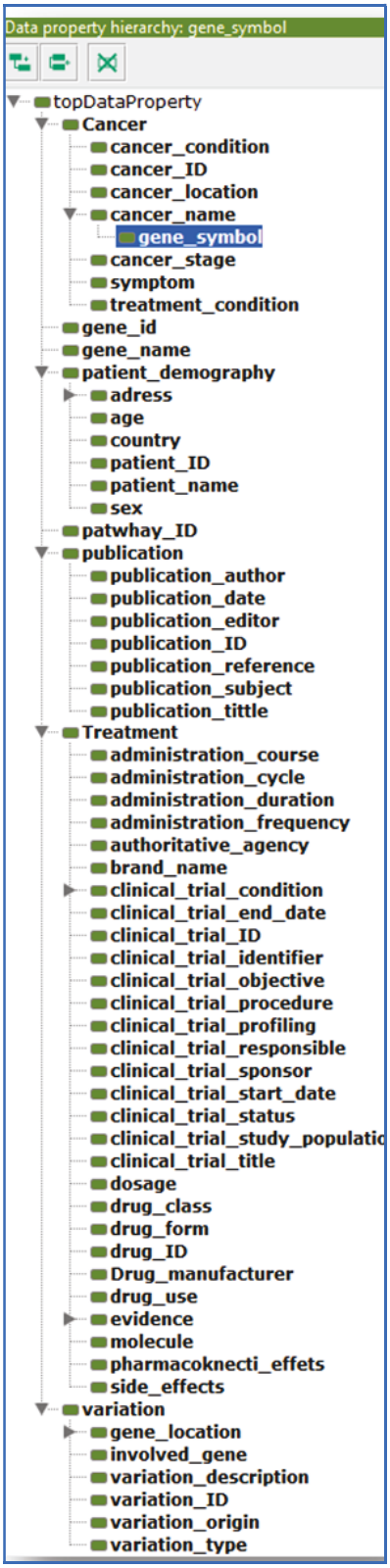


Figure 77 : Protege Data Property

Tableau xii : Data Properties (propriétés des classes)

Classes	Data Properties	Description	
Cancer	Cancer ID	Numéro identifiant cancer	
	Cancer_condition	Condition de prescription d'un médicament à un pour un cancer	
	Cancer_name	Nom du cancer	
	Cancer_stage	Stade du cancer	
	Symptom	Symptômes du cancer	
	Treatment_condition	Condition du traitement	
Gene	gene_id	Identifiant du gène	
	gene_name	Nom du gène	
	gene_symbol	Symbole du gène	
	gene_location	Localisation du gène	
Patient_demography	Adress	Adresse postale et email	
	Age	Age du patient	
	Country	Pays du patient	
	patient_name	Nom du patient	
	Sex	Sexe du patient	
Publication	publication_author	Auteur de la publication	
	publication_date	Date de la publication	
	publication_editor	Editeur de la publication	
	publication_id	Identifiant de la publication	
	publication_reference	Référence de la publication	
	publication_author	Auteur de la publication	
	publication_tittle	Titre de la publication	
Treatment	Drug	Administration_course	Cours du traitement

		Administration_cycle	Cycle du traitement
		Administration_duration	Durée de traitement
		Administration_frequency	Fréquence de traitement
		Authoritative_agency	Autorité de régulation
		Brand_name	Nom commercial du médicament
		Dosage	Dosage
		drug_class	Classe du médicament
		drug_form	Forme du médicament : comprimé, injection etc.
		drug_id	Identifiant médicament
		drug_manufacturer	Fabricant du médicament
		drug_use	Indication thérapeutique
		Evidence	Preuve (référence de l'essai clinique)
		Molecule	Molécule
		pharmacoknecti-effects	Effets pharmacocinétiques.
		side_effect	Effets adverses
	Clinical_trials	clinical_trials_condition	Conditions de l'essai clinique

		clinical_trials_end_date	Date de fin de l'essai clinique
		clinical_trials_id	Identifiant de l'essai clinique
		clinical_trials_idenditifier	Référence de l'essai clinique
		clinical_trials_objective	Objectifs de l'essai clinique
		clinical_trials_procedure	Procédure pour l'essai clinique
		clinical_trials_profiling	Profilage des participants à l'essai clinique
		clinical_trials_ responsible	Responsable de l'essai clinique
		clinical_trials_sponsor	Sponsor de l'essai clinique
		clinical_trials_start_date	Daté de début de l'essai clinique
		clinical_trials_status	Etat de l'essai clinique
		clinical_trials_study_populatio n	Population cible de l'essai clinique
		clinical_trials_title	Titre de l'essai clinique
Variation	gene_location	Localisation du gène mutant	
	Involved_gene	Gène mutant	
	variation_description	Description de la variation	
	variation_id	Identifiant de la variation	
	variation_origin	Origine de la variation	
	Variation_type	Type de variation	

Nous venons de lister toutes les ressources de notre ontologie. Nous allons voir quelques exemples de ressources décrites dans l'ontologie avec le raisonneur activé.

Vue d'une instance de gène :

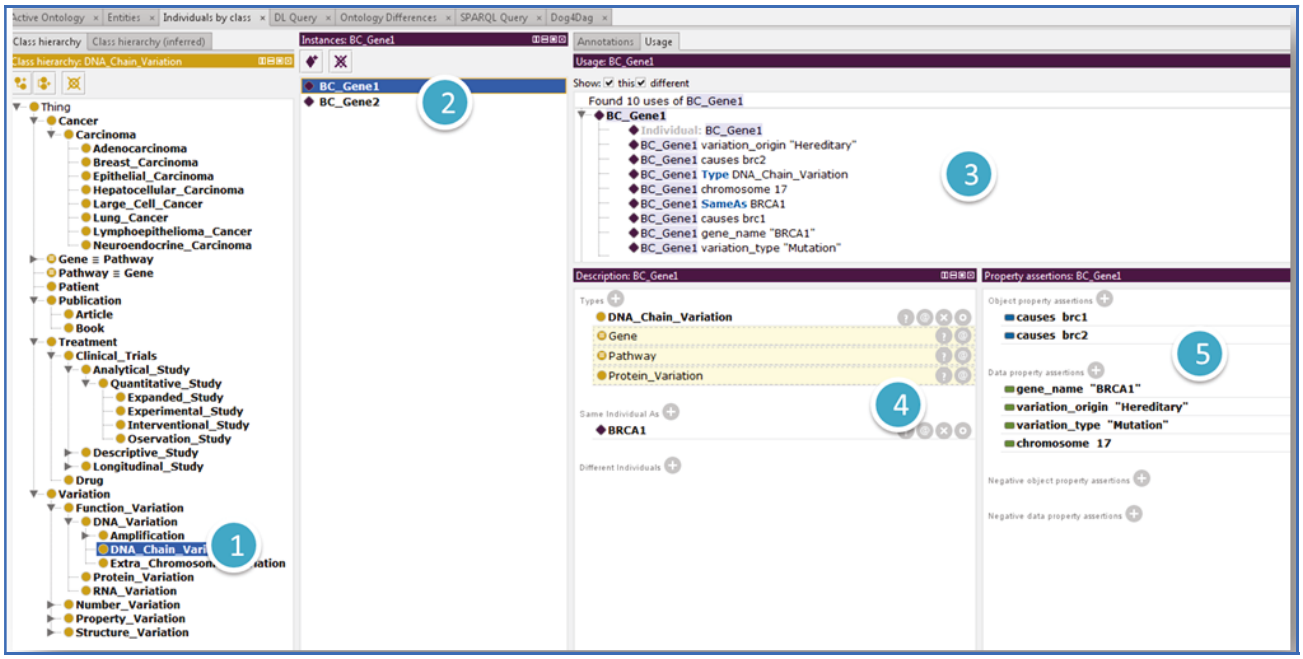


Figure 78 : Instance de Gène

Nous notons ici une instance de variation de gène incriminé dans le cancer du sein. On distingue :

- Le panneau 1 nous identifie BC_Gene1 (panneau 2) comme une variation.
- Cette variation est décrite dans le panneau 3.
- Le « ***BC_Gene »*** est un type de « ***DNA_Chain_Variation*** » avec toutes les inférences remontées par le raisonneur
- Le panneau 5 nous montre toutes les valeurs des propriétés de l'individu « ***BC_Gene1*** »
- Vue d'une instance de médicament « ***Drug*** »

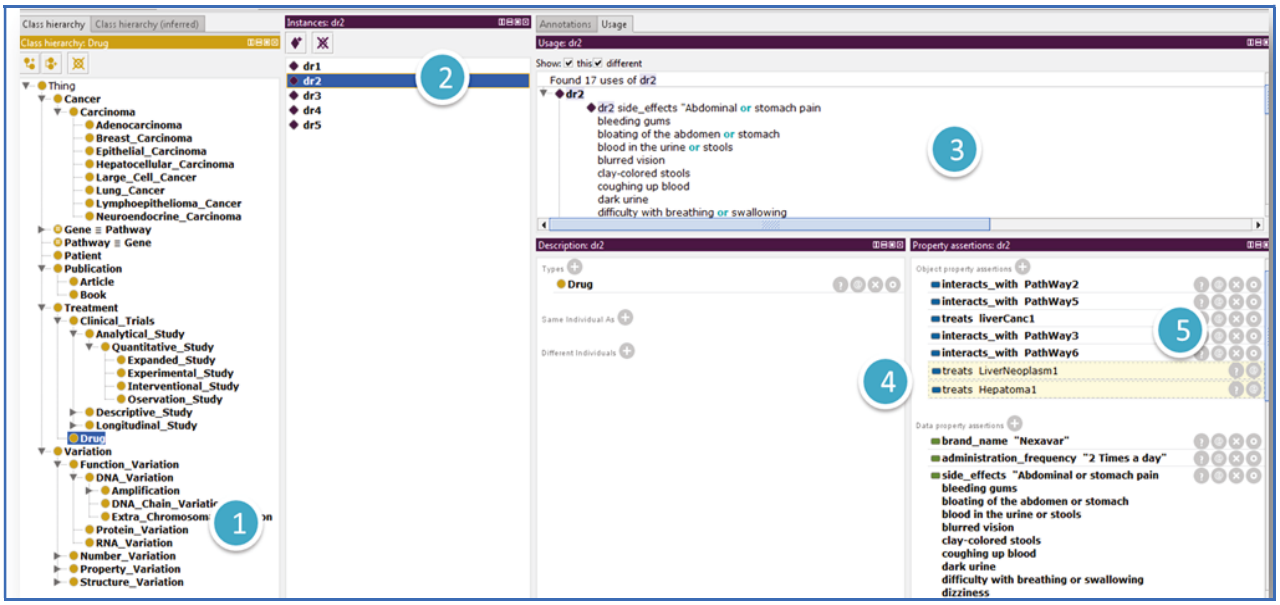


Figure 79 : Instance de Drug

Nous notons ici une instance de variation de médicament dénommé « ***Nexavar »***, utilisé pour le traitement du cancer du foie. Sur cette instance on voit :

- Le panneau 1 nous identifie la classe d'appartenance du médicament.
- Le panneau 2nous donne l'individu « ***dr2*** » comme une variation.
- Cet individu est décrit dans le panneau 3.
- Le panneau 5 nous montre toutes les valeurs des propriétés de l'individu « ***dr2 »*** avec en jaune les inférences remontées par le raisonneur.

Vue d'une instance d'essai clinique

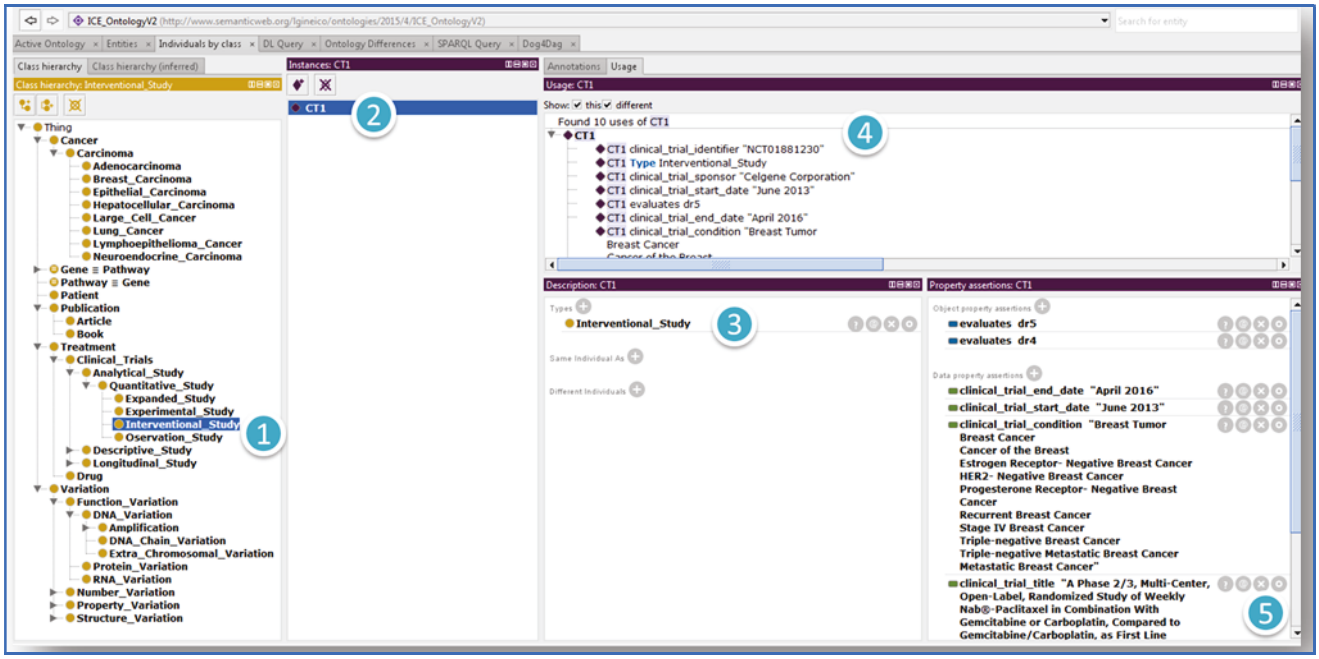


Figure 80 : instance d'essai clinique

Cette image nous montre un exemple d'essai clinique. Ainsi nous avons l'individu « ***CT1 »*** décrit comme suit :

Une étude interventionnelle, dont l'instance « ***CT1 »*** a pour type « ***Interventional study »,*** *a*vec les différents usages cités dans le cadre 3. Les valeurs de propriétés sont indiquées dans la zone « ***property Assertion*** ».

3.5.6 REQUETES SPARQL DANS L'ONTOLOGIE

Pour vérifier que notre prototype fonctionne correctement, nous avons développé quelques requêtes dans le langage dédié au format RDF. Ce langage est appelé SPARQL.

3.5.6.1 DESCRIPTION DU LANGAGE SPARQL

Le langage SPARQL est le standard du W3C pour les requêtes de bases données triple stores ou RDF. Le SPARQL a un protocole dit SPROT (SPARQL Protocole) qui vise à transporter les requêtes SPARQL clients vers les processors de requêtes. Le protocole SPARQL a deux fonctions :

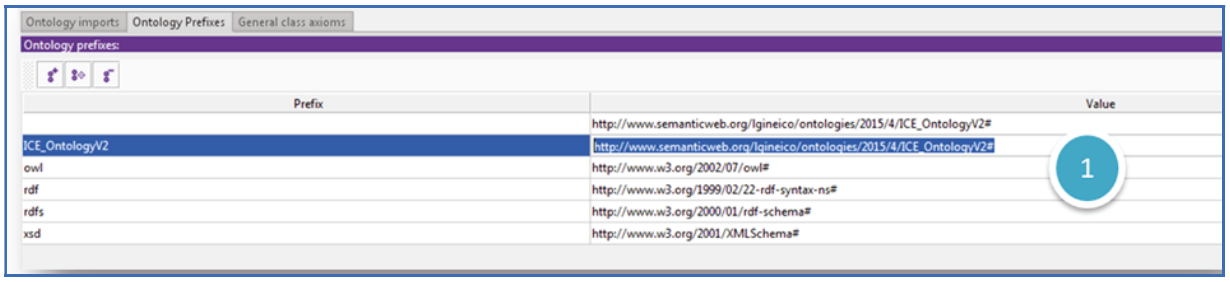
- Comme une interface abstraite indépendante de toute réalisation concrète, de mise en œuvre, ou de liaison à un autre protocole
- Comme point de fixation des protocoles HTTP et SOAP a cette même interface

Pour notre prototype, n'ayant pas encore recueilli, ni intégré les données de séquençage d'Integragen et n'ayant testé le mapping avec Karma qu'à titre expérimental, nous avons saisi quelques données à la main directement dans l'ontologie. Ces données nous ont permis d'implémenter deux types de requêtes directement dans Protégé :

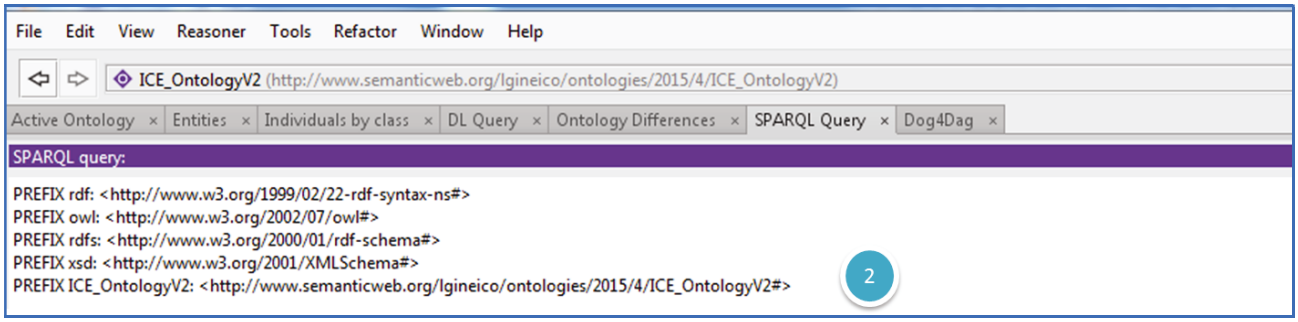
- Des requêtes simples et
- Des requêtes à résultats multiples

Dans Protégé avant de construire une requête, il faut d'abord préfixer son l'URI de l'ontologie (c'est un raccourci très utile pour ne pas avoir à retaper à chaque fois l'URI). Pour ce faire nous avons procédé comme suit :

Retrouver et recopier l'URI de l'ontologie, en l'occurrence ICE_OntologieV2



Le retranscrire dans l'onglet « Sparql Query »



Une fois le préfixe établi, il pourra se substituer au long URI. Les requêtes que nous avons par la suite développées sont :

3.5.6.2 TROUVER LE SYNONYME DU CANCER HEPATOMA1

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

PREFIX owl: <http://www.w3.org/2002/07/owl#>

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

PREFIX ICE_OntologyV2: <http://www.semanticweb.org/lgineico/ontologies/2015/4/ICE_OntologyV2#>

SELECT ?Synonym

WHERE

{

ICE_OntologyV2:Hepatoma1 owl:sameAs ?Synonym

}

Cette requête donne le résultat ci-dessous :

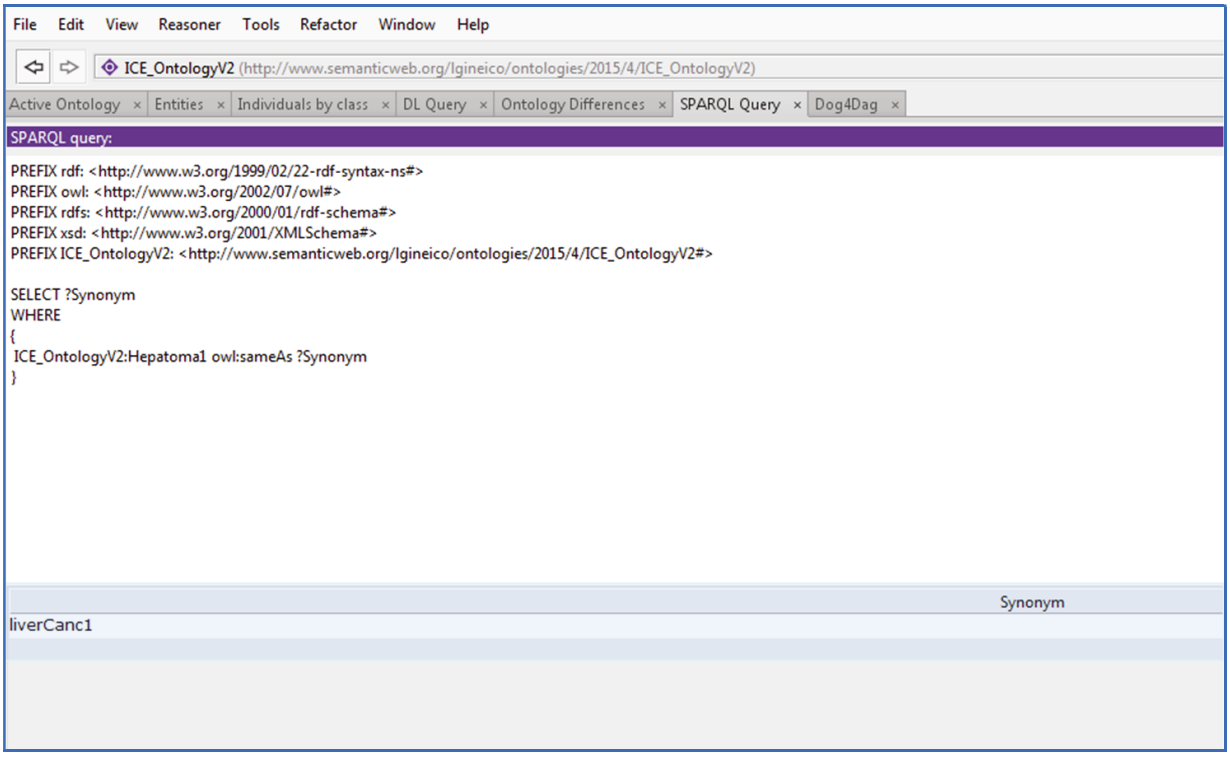


Figure 81: Requête 1 - Recherche de synonyme

3.5.6.3 RECHERCHE DES MALADIES TRAITEES PAR UN MEDICAMENT DONNE EN L'OCCURRENCE LE « DR1 »

*SELECT **

WHERE

{

ICE_OntologyV2:dr1 ICE_OntologyV2:treats ?Cancer.

}

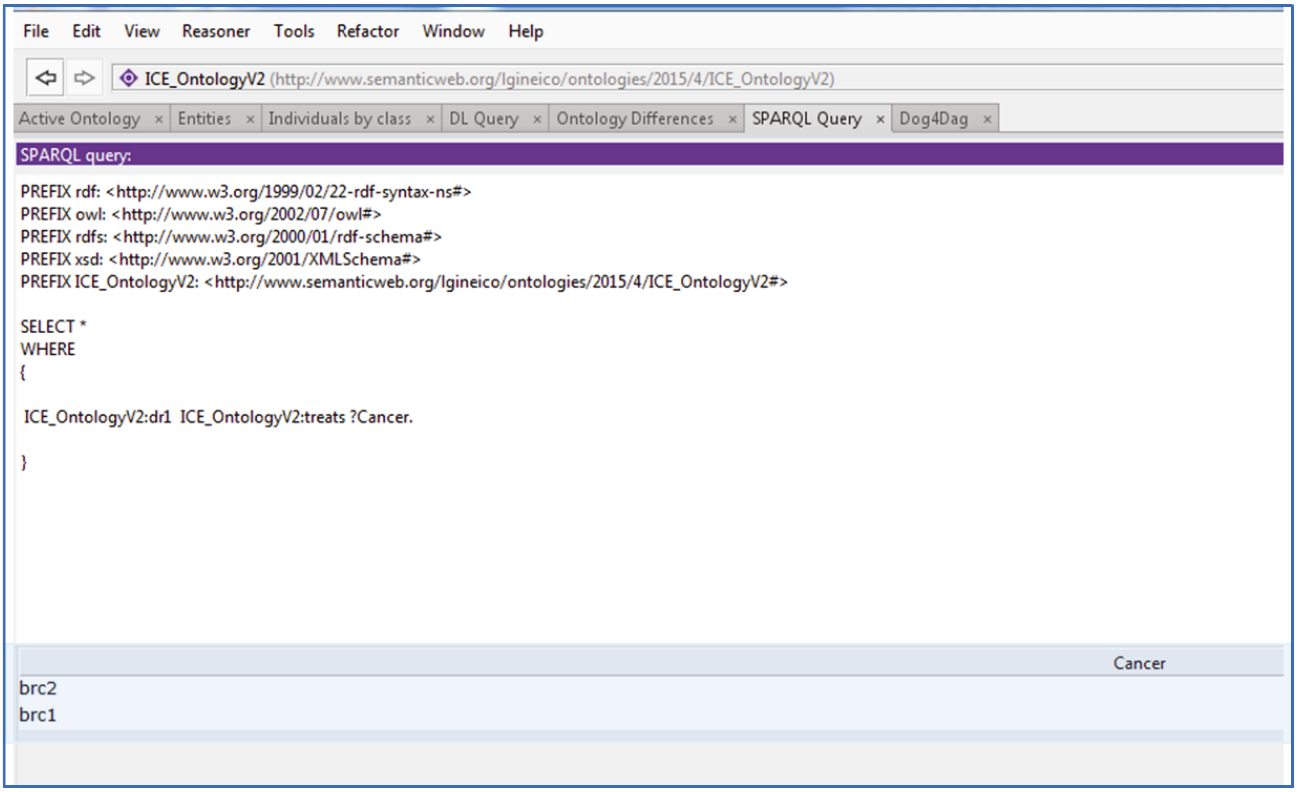


Figure 82 : Requête 2 Recherche de cancer traitée par un médicament

3.5.6.4 TROUVER LES CANCERS QUI ONT DONNE LIEU A PUBLICATION

*SELECT **

WHERE

{

?Cancer ICE_OntologyV2:is_subject_to ?Article.

}

Requêtes multiples

Nous recherchons tous les médicaments utilisés pour traiter tous les cancers de notre base ainsi que leur indication thérapeutique.

*SELECT **

WHERE

{

?Drug ICE_OntologyV2:brand_name ?DrugName.

?Drug ICE_OntologyV2:treats ?Cancer.

?Drug ICE_OntologyV2:drug_use ?Usage.

}

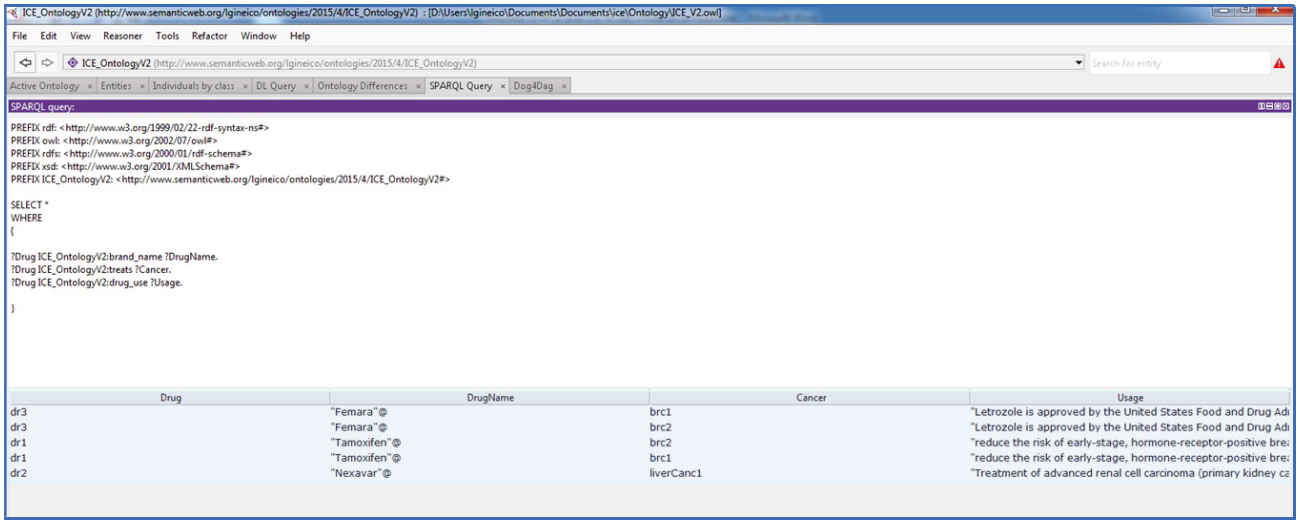


Figure 83 : Recherche de Médicament, Cancer et d'indication thérapeutique

Voilà les quelques requêtes que nous avons développées dans Protégé. Ces requêtes seront à étendre lorsque toutes les données auront été chargées dans la base HB.

Le chapitre suivant aborde le Mapping que nous avons réalisé sous KARMA avec le modèle d'ontologie du prototype 3.

3.5.7 MAPPING D'ONTOLOGIE AVEC L'OUTIL KARMA

Comme énoncé précédemment, nous avons fait un mapping avec des données de la base de données PharmGKB sur notre ontologie. Le but ici est de démontrer la faisabilité de faire correspondre des informations contenues dans des ontologies existantes sur notre modèle ICE.

Cela à terme doit constituer un moyen pratique de peupler l'ontologie et générer du RDF que l'on chargera dans la base de données HBase sans avoir à importer de gros volumes de données directement dans Protégé qui connaît de graves problèmes de performance lorsqu'il s'agit de données volumineuses.

3.5.7.1 LE MODELE DE MAPPING

Nous avons fait deux essais de mapping avec les données de médicaments et les données sur les gènes

1. Mapping des médicaments
2. Mapping des gènes

Pour faire cette correspondance, nous avons téléchargé depuis le site PharmGKB, le fichier des gènes en format TSV que nous avons chargé dans Kamara.

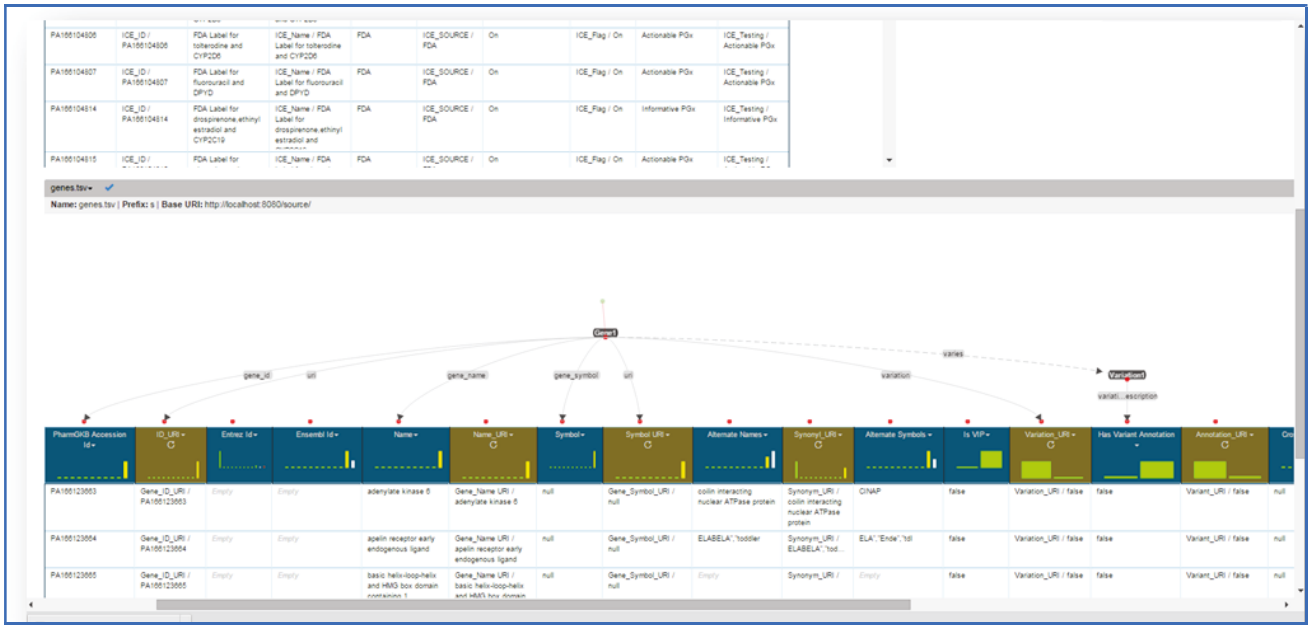


Figure 84 : modèle de mapping avec le fichier de médicaments PharmGKB

Ici nous avons établi les correspondances suivantes :

Tableau xiii : Tableau de correspondance modèle vs fichier

Classes Ontologie ICE	Propriétés de la classe	Colonnes de fichier Drugs correspondantes
Gènes	Gene_id	PharmGKB_accession_id
	Gene_name	name
	Gene_symbol	symbol
Variation	Variation_description	Has_variation_annotation

Le résultat du mapping des données

Le résultat de ce mapping est un fichier RDF qui sera chargé dans Jena.

A l'issue de cette expérience, nous avons démontré qu'il était possible de peupler l'ontologie en passant par Karma. Cette solution offre l'avantage d'exploiter les contenus d'ontologies existantes sans avoir à les importer dans Protéger. Cela correspond au cahier de charge que nous confié l'équipe ICE c'est-à-dire vérifier la réutilisabilité des ressources du net.

Cette étape marque la fin de notre travail. L'équipe ICE reprendra le flambeau pour l'extraction des données et les solutions de stockages.

CONCLUSION DES TRAVAUX EFFECTUES

Pendant cette période de 6 mois, nous avons étudié les ontologies et construit un modèle qui correspond au besoin exprimé par l'équipe ICE. Nous avons ainsi construit une ontologie modèle, qui servira de pivot pour les ressources du net. Nous avons démontré qu'il était possible de faire correspondre les ressources du net avec notre modèle en utilisant l'outil Karma.

Le relais sera pris par l'équipe ICE pour charger les données de séquençages produites par Integragen et recueillir les données des ontologies du web. Ils feront alors le mapping selon notre modèle pour générer du RDF à charger dans la base de données HBase.

Le chapitre qui suit décrit comment le projet exploitera l'ontologie

4 PERSPECTIVES

4.1 LES LIMITES DE L'ONTOLOGIE MODELE

Nous avons implémenté une ontologie qui doit servir de modèle pour faire les alignements nécessaires avec les ontologies existantes. L'alignement que nous effectuons avec l'ontologie ICE est manuel. Il va donc nécessiter une manipulation à intervalle régulier. La limite de notre solution est donc un manque d'automatisation. Cette piste reste à étudier car notre mission s'achève et nous n'aurons pas de temps pour investiguer dans ce sens.

4.2 LA GESTION DE DONNEES

Les données sont au cœur du système ICE. En effet le projet ICE tel que nous avons évoqué depuis le début intègrera des données issues du séquençage des gènes, (des données très volumineuses) mais aussi des données en provenance des ontologies déjà existantes. Cela confère une complexité au système qu'il convient d'adresser correctement. En effet la majorité des risques identifiés concernent la gestion des données que ce soit :

- Leur stockage
- Leur manipulation des données
- La qualité des données
- La règlementation notamment le droits à l'oubli des malades du cancer

Aussi convient-il de sélectionner un système de gestion de données qui répondent aux exigences ci-dessous :

- la montée en charge avec le temps,
- la sécurité des données (les données médicales nominatives doivent être tracées),
- la réplication des données,
- la consistance des transactions,
- la répartition de charge (load balancing),
- et le traitement par lot (batching).

HBase est le système qui a été sélectionné comme la solution de stockage des données sélectionnée. C'est une base de données NoSQL open source développée en Java. Elle est basée sur le modèle big table de Google. HBase fonctionne en surcouche de Hadoop Distributed File System (HDFS). Il offre ainsi des très grandes capacités de stockage et permet une tolérance aux

échecs. C'est-à-dire que ce type de système est capable de fonctionner correctement même en cas de problème majeur, là où les systèmes traditionnels s'arrêtent. Ses caractéristiques principales sont résumées dans le tableau ci-dessous :

- réplication : Hbase permet une réplication automatique avec une très haute disponibilité permettant au système de fonctionner en transparence en cas d'erreur ou d'échec des traitements
- distribution : Hbase permet une répartition automatique des traitements
- la consistance immédiate des données
- Indexation des données
- Nombre illimité de lignes

Cette solution répond à certains risques identifiés lors de l'analyse notamment en ce qui concerne le « *Stockage, indexation temps réel et accès aux données ».*

Par ailleurs le couplage avec l'écosystème Hadoop et Map Reduce permettra la manipulation et l'extraction des données. On obtient le schéma ci-dessous qui donne une vision du système en construction.

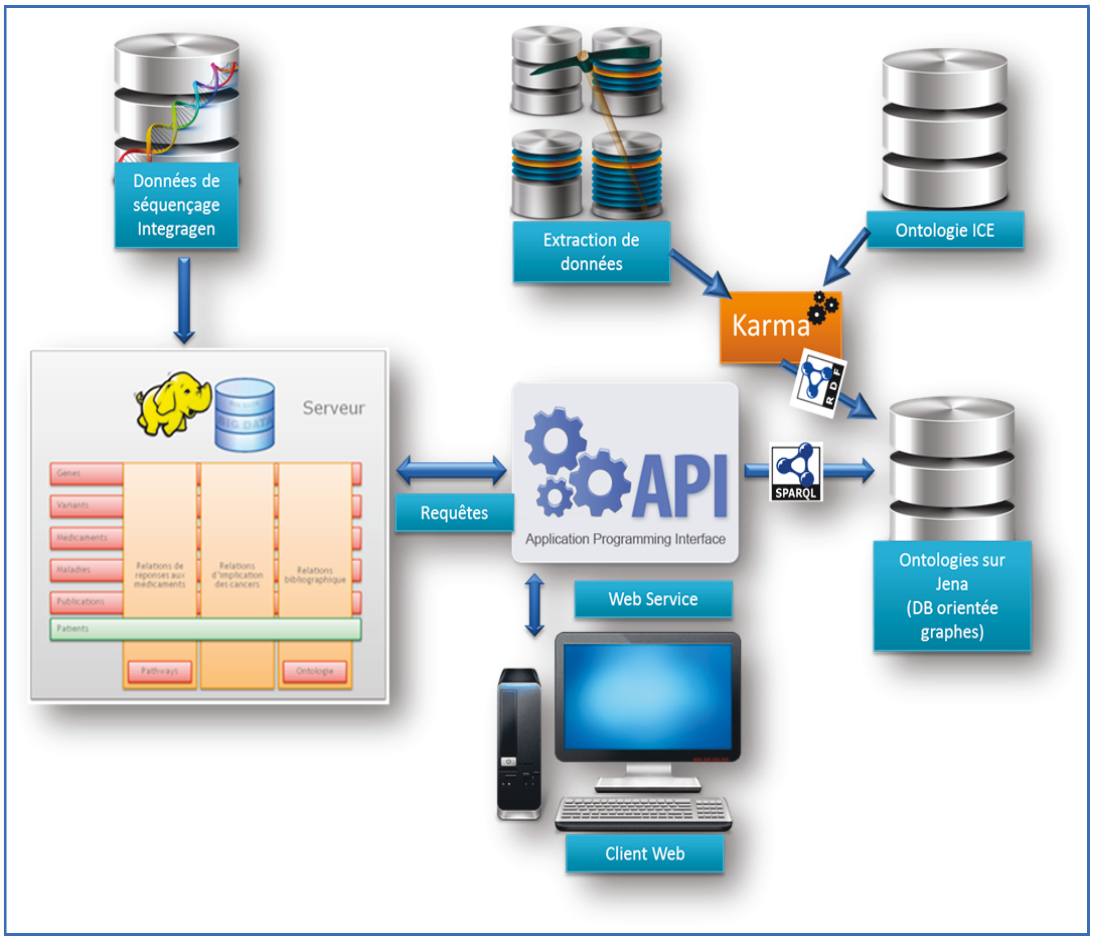


Figure 85 : Architecture ICE

4.3 LA FOUILLE DES DONNEES (TEXT-MINING)

La fouille de données, pour faire des recherches pertinentes en tenant compte du contexte et des inférences du raisonneur, sera l'autre chantier que le projet devra embrasser pour sa finalisation à échéance dernier trimestre 2015.

En effet des sites internet tel PubMed sont aujourd'hui un exemple de ce qui est possible de faire en termes d'extraction d'informations comme par exemple en recherchant les mots clés d'un concept donné, en retrouvant les auteurs d'une publication allant même jusqu'à retrouver des publications en fonction sujet donné.

L'outil ICE devra répondre à cette exigence. A ce jour une étude a été menée pour référencer les moteurs de recherche et les outils existants pour extraire les informations de diverses ontologies. Le pas suivant sera de choisir l'outil qui répondra au mieux aux exigences et mettre en œuvre les requêtes directement exploitables par l'application ICE.

Trois cas d'utilisation sont ainsi envisagés :

1. faire correspondre des données identiques (synonymes)
2. renvoyer les publications les plus adaptées par rapport à un patient (publications relatives à une ressource)
3. objectif plus lointain : créer de valeur, réinjecter de la valeur (ajout, modification de contenus)

4.3.1 GESTION DES SYNONYMES

Comme nous l'avons évoqué précédemment, il existe de nombreuses ontologies relatives aux gènes et au cancer. Ces ontologies n'utilisent pas la même nomenclature pour décrire les ressources. Ainsi un cancer du foie sera nommé hépato-carcinome, adénocarcinome, adénocancer hépatique, épithélioma ou carcinome hépatocellulaire selon les ontologies. Le but de ICE est donc de gérer ces synonymes et de pouvoir restituer tous ces noms au travers des requêtes souvent complexes.

4.3.2 EXTRACTIONS DES PUBLICATIONS

De même par rapport à un sujet donné, l'outil devra être capable de parcourir les différents nœuds de l'ontologie pour extraire les publications et les classifier par ordre de pertinence. Ainsi l'utilisateur pourra afficher une liste des documentations publiées sur le net relatives à sa recherche.

4.3.3 CREATION DE VALEUR

L'objet à moyen et long terme de ICE est de permettre à la communauté scientifique d'enrichir la base de connaissance. Les utilisateurs pourront donc injecter de nouveaux termes, corriger des termes erronés créant ainsi un cercle vertueux de la création et du partage de la connaissance

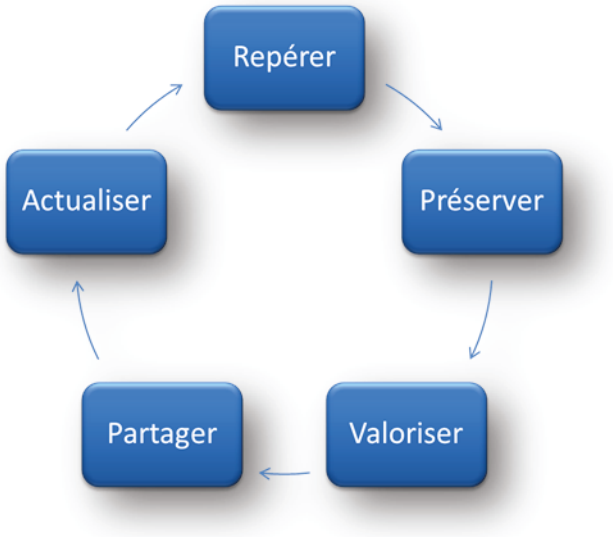


Figure 86 : Cercles vertueux de la connaissance

CONCLUSION

Comme nous venons de voir, il reste encore des chantiers importants pour le projet ICE qui est maintenant à la deuxième année du quinquennat. L'ontologie n'est qu'une brique à cette construction. Cependant elle servira de pierre angulaire pour toutes les ontologies qui devront interagir avec le système de connaissance en construction.

5 CONCLUSION GENERALE

NOUS ARRIVONS AU TERME DE NOTRE ETUDE. ELLE A DEBUTE PAR UN TOUR D'HORIZON DES TECHNOLOGIES GRAVITANT AUTOUR DES ONTOLOGIES. CETTE CARTOGRAPHIE NOUS A PERMIS DE CONNAITRE LES STANDARDS DU W3C. AINSI CE CONSORTIUM DEFINIT LES REGLES QUI VONT REGIR LE WEB DE DEMAIN. MAIS DEJA ILS SONT A L'ŒUVRE POUR DONNER DU SENS AU CONTENU DU WEB. LE BUT EST DE CONSTRUIRE DES MACHINES INTELLIGENTES PERMETTANT AUX SITES DE « DIALOGUER ». DEUX STANDARDS CLES SONT TRES IMPORTANTS DANS LE DOMAINE DES ONTOLOGIES :

- *LE RDF (RESOURCE DESCRIPTION FRAMEWORK) : CE STANDARD PERMET DE DECRIRE LES RESSOURCES SOUS FORME DE TRIPLET. LES ONTOLOGIES VONT UTILISER CE FORMALISME PERMETTANT DE DECRIRE LES RESSOURCES DE L'ONTOLOGIE. CE FAISANT LES RAISONNEURS VONT PERMETTRE DE « FAIRE LA RELATION ENTRE LES DIFFERENTES ENTITES » ET DE FAIRE DES INFERENCES.*
- *LES URI (UNIFORM RESOURCE IDENTIFIER) : CET IDENTIFIANT UNIQUE POUR TOUTE RESSOURCE, ELLE REGROUPE LES NOTIONS DE URL (UNIFORM RESOURCE LOCATOR) ET URN (UNIFORM RESOURCE NAME). CET IDENTIFIANT PERMET AUX MACHINES DE RECONNAITRE LES RESSOURCES ET DE POUVOIR LES INTERCONNECTER.*

LE W3C A EGALEMENT EMIS UNE RECOMMANDATION QUANT AU LANGAGE DE DEVELOPPEMENT DES ONTOLOGIES ; IL S'AGIT DU OWL (ONTOLOGIE WEB LANGUAGE).

DEUX OPTIONS SE PRESENTAIENT POUR MODELISER LE DOMAINE DE CONNAISSANCE : UNE BASE DE DONNEES ORIENTEE GRAPHE CLASSIQUE OU UNE ONTOLOGIE QUI EST UN TYPE DE BASE DE DONNEES ORIENTEES GRAPHE MAIS AVEC L'AVANTAGE DU RAISONNEMENT PAR INFERENCE. CETTE QUALITE DES ONTOLOGIES A FAIT PENCHER LA BALANCE EN FAVEUR DE CETTE SOLUTION. CEPENDANT POUR CONSTRUIRE CETTE ONTOLOGIE, NOUS AVONS DU PARCOURIR LES CONCEPTS DU DOMAINE DE CONNAISSANCES AU TRAVERS DE CARTE MENTALE POUR ETRE SURE QUE RIEN N'ETAIT OUBLIE.

LA MISE EN ŒUVRE S'EST FAITE PAR ITERATIONS SUCCESSIVES QUI A PERMIS DE CONSTRUIRE UN PROTOTYPE. NOUS ETIONS PARTIS SUR LA PISTE OBO-EDIT, UN OUTIL QUE PROMEUT LA COMMUNAUTE BIOMEDICALE MAIS SON MANQUE DE VERSATILITE NOUS A CONTRAINT A OPTER POUR PROTEGE UN OUTIL PLUS REPANDU.

L'OBJECTIF QUI NOUS A ETE ASSIGNE A SAVOIR REUTILISER AU MAXIMUM L'EXISTANT POUR LES FAIRE COÏNCIDER A ETE ATTEINT GRACE A L'OUTIL KARMA INTEGRATION SYSTEM QUI NOUS PERMET DE NOUS AFFRANCHIR DE TELECHARGER DIRECTEMENT DANS PROTEGE DES ONTOLGIES DU WEB. CET OUTIL EN EFFET NOUS PERMET D'ALIGNER LES RESSOURCES DES ONTOLOGIES EXISTANTES SUR NOTRE MODELE POUR GENERER DU RDF.

NOTRE ONTOLOGIE SERVIRA DE PIVOT DANS LE SYSTEME ICE. C'EST UN PROJET DE GRANDE ENVERGURE QUI A ENCORE PLUS DE TROIS ANNEES DEVANT LUI. LE CHARGEMENT DES DONNEES EN BASE ET LA FOUILLE DE TEXTE CONSTITUENT LE PAS SUIVANT. LES RISQUES IDENTIFIES EN DEBUT D'ETUDE SERONT GERES ET MITIGES AU MOMENT OPPORTUN.

NOUS AURIONS AIME POUVOIR CONTINUER SUR CETTE VOIE MAIS NOTRE MISSION S'ACHEVE. CE SEMESTRE A ETE UNE PERIODE DE DECOUVERTE ET D'ENRICHISSEMENT. NOUS AVONS EMBRASSE LES DOMAINES DES ONTOLOGIES ET DE LA MEDECINE GENOMIQUE, JUSQU'ALORS INCONNUS DE NOUS.

6 Annexe

6.1 DOSSIER DE LA PROPOSITION DU SUJET DE STAGE

PROPOSITION DE SUJET DE MEMOIRE

En vue d'obtenir

LE DIPLOME D'INGENIEUR CNAM

en

Informatique

par

Yao Affoue Laurence épouse Gine I Cortiella

Lieu du stage : Sogeti High Tech
Plateau Recherche & Développement – Projet ICE (Interpretation of Clinical Exoms)
Responsable en entreprise : Lise Pavard

Etude et définition d'une ontologie pour l'interprétation des exomes cancéreux

Date de soutenance envisagée : Juillet 2015

2 Passage Roquemaurel - B204 – 31300 /0630927502 / laurence_yao@yahoo.fr

Table des matières

1 PRESENTATION DU SUJET

1.1 Contexte

Dans le cadre du projet ICE – Interpretation of Clinical Exome- en partenariat avec Integragen, l'Inserm, et l'Institut Gustave Roussy, Sogeti High Tech a la charge du développement d'une infrastructure logicielle permettant de fournir des diagnostics différenciés pour chaque patient atteint d'un cancer.
Le projet est basé sur les technologies du BigData comme les ontologies, le text mining, les bases de données NoSQL, la représentation graphique de données.

1.2 But du stage

Il s'agit de participer au projet ICE et de contribuer à l'étude et la définition d'une ontologie dans le domaine médical.
Le principe des ontologies consiste à modéliser et conceptualiser les connaissances d'un domaine afin d'obtenir une base de connaissances la plus complète et exhaustive possible tout en étant dotée de dispositifs de recherches intuitifs et efficaces.

Le stage s'articulera en 3 parties :

- Etablir un état de l'art général des ontologies dans le milieu médical (outils et méthodes)
- Implémenter les ontologies pertinentes et les intégrer au projet ICE
- Coordonner les contributeurs du plateau R&D intervenant sur la partie ontologie du projet ICE

1.3 Contraintes

Le projet ICE est un projet de recherche dont les technologies ont une maturité faible et sont susceptibles de changer au cours de l'avancement.
Les objectifs initiaux sont ambitieux car le sujet est complexe, et la technologie des ontologies n'est pas totalement maîtrisée.
Le travail débutera par l'établissement d'un état de l'art préliminaire sur les outils et les méthodes des ontologies.

2 METHODOLOGIE

La méthodologie déployée sera celle d'un projet de recherche : établissement d'un état de l'art, puis conception de prototypes permettant de confronter les résultats aux objectifs fonctionnels et techniques fixés initialement et d'ajuster les développements en fonction de ces résultats.

La stagiaire pourra s'appuyer des ressources du plateau R&D pour le développement des prototypes et assurera la coordination de leurs interventions sur le sujet.

Notre travail s'inscrit dans le cadre d'une méthode agile semi itérative décrite ci-après :

2.1 Etude de l'art et analyse du besoin :

Notre partenaire souhaite développer un « outil d'assistance aux choix thérapeutiques basé sur la connaissance de tumeur cancéreuse en les croisant avec l'ensemble des informations sur les molécules pharmaceutiques, … établissant les liens entre leur efficacité et les caractéristiques génomiques du patient ».

Il s'agit d'étudier le problème pour comprendre les enjeux, et faire un état de l'art pour savoir d'où l'on part et quelles sont les solutions que nous pouvons envisager. Pour ce faire nous étudierons :
La connaissance actuelle sur les Ontologies,

- les technologies utilisées pour construire les ontologies (langages et outils)
- les ontologies existantes disponibles en rapport avec le domaine fonctionnel (oncologie)

Au sortir de cette phase nous aurons les éléments nécessaires pour concevoir et proposer une solution qui répondra à la demande initiale.

2.2 Conception :

Nous établirons une conception pour la mise en œuvre d'une ontologie adaptée au contexte du projet, en fonction des éléments retenus dans la phase d'étude.
Ce qui nous amènera à comparer les outils les mieux adaptés à la conception de l'ontologie et définir une architecture cible.

2.3 Prototypage :

a Développement :

Cette étape verra la construction de l'ontologie au travers de différents prototypes itératifs qui seront confrontés au fur et à mesure au besoin du partenaire et seront ajustés techniquement et fonctionnellement selon ses retours.
L'objectif est de pouvoir s'adapter aux attentes qui sont susceptibles d'évoluer au cours de projet, et de maintenir un échange régulier avec le partenaire.

b Vérification :

Chaque itération du prototype donnera lieu à des tests unitaires afin d'assurer que le prototype répond aux attentes d'une part et que d'une itération à l'autre il n'y a pas de régression d'autre part.

c Validation :

Les prototypes seront mis à la disposition d'un panel d'utilisateurs qui feront des retours en rapport avec leurs attentes fonctionnelles.

Toutes ces activités seront décrites dans un mémoire qui sera soutenu devant un jury au mois de décembre 2015.

3 TECHNIQUES

3.1. Outils

Les outils utilisés pour le stage ne sont pas encore définis et le seront à l'issue de la phase de conception.
Nous avons cependant une liste d'outils et de langages liés aux ontologies que nous avons présélectionnés pour l'étude préliminaire :

- Les outils de conceptions d'ontologie : (Protégé, KMGEN, SemanticWorks etc.)
- Langage de développement d'ontologie (OWL, Ontologie, SKOS ou N3)

3.1 Difficultés

Les difficultés du stage proviennent de 2 aspects du projet ICE.

Tout d'abord, les ontologies sont une technologie innovante et mal maîtrisée, implémentée pour la 1ère fois dans le domaine médical chez Sogeti High Tech. La stagiaire ne pourra pas s'appuyer sur des référents métiers dans ce domaine et devra faire preuve d'autonomie dans ses activités.

Par ailleurs, le projet ICE s'inscrit dans une démarche de recherche et développement dans le cadre d'un partenariat subventionné. Les besoins utilisateurs ne sont donc pas aussi bien définis que sur une prestation client classique, et pourront être amenés à évoluer en fonction des avancées technologiques du projet.
Ce contexte demande une constante adaptation et justifie l'adoption d'une méthodologie semi-itérative.

4 LIVRABLES

Les livrables attendus vont s'échelonner comme suit :

Figure 1: Planning des livrables

Les étapes en amont dites préparatoires ont eu pour but de définir le projet mettre en place l'organisation pour concilier mon activité professionnelle et la réalisation du stage. Cette étape a permis de rédiger le document de proposition de sujet soumis au CNAM.

Pour ce qui concerne le stage a proprement parlé, les livrables sont les suivants :

- Document1: Etat de l'art et recommandations d'utilisation des ontologies pour le projet ICE
- Prototype 1 : Première implémentation d'ontologies pour ICE
- Prototype 2 : Prise en compte des retours sur le 1er prototype
- Prototype 3 : Intégration des ontologies dans le projet ICE
- Document 2 : Capitalisation sur le stage (Spécification du prototype)
- Reporting : Etat d'avancement hebdomadaire du sujet au cours du stage.

5 PLANNING

Le stage démarrera à la mi-novembre 2014 et durera jusqu'au 31 aout 2015. Le planning ci-dessous décrit l'enchainement des activités.

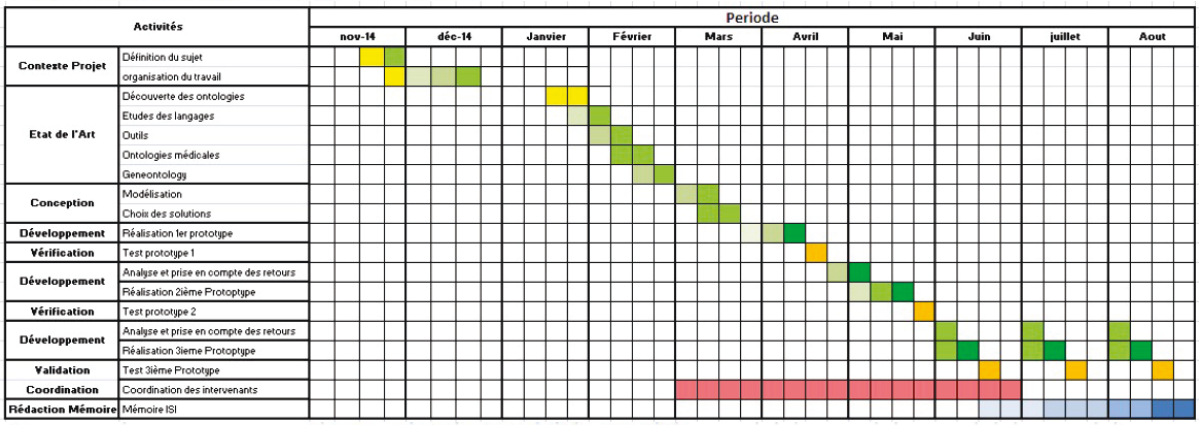


Figure 2 : Planning des activités

Notre planning fait apparaître sept types d'activités que sont :

5.1 Contexte Projet :

Cette activité a pour objectif de définir le cadre de du stage. Elle a été menée sur les deux derniers mois de l'année 2014. En effet, en accord avec les équipes du projet ICE et la responsable du plateau recherche et développement, nous avons défini le sujet du stage et établi l'organisation matérielle de son déroulement.

5.2 Etat de l'art :

Etude de la demande et recherche au sujet de :

- la connaissance actuelle sur les ontologies,
- les langages connus utilisés pour les construire
- les outils permettant de les construire
- l'étude de l'ontologie des gènes : Geneontology

5.3 Conception :

- Conceptualisation et modélisation de l'ontologie
- Sélection de la solution
- Initialisation de l'ontologie

5.4 Développement :

Réalisation de prototypes de manière itérative pour arriver à un résultat répondant à l'attente du partenaire

5.5 Vérification :

Tests unitaires de l'ontologie que nous mettons en place basé sur des scenarii de tests que nous écrirons

5.6 Validation :

Test de validation de l'opérabilité de l'ontologie :

- Recherche de termes dans l'ontologie
- Mise en relation d'un terme avec un exome
- Lien entre l'exome et les traitements proposés.

5.7 Coordination d'équipe :

Cette activité s'étale sur toute la durée du stage

5.8 Rédaction du mémoire :

Cette tâche débute au mois de mars et se prolonge jusqu'à la fin du mois d'août 2015.

6.2 EXEMPLE DE FICHE DE SUIVI HEBDOMADAIRE DU PROJET

Plateau R&D BIG DATA

Week n°15

POLE : ICE - Ontology

Groupe de travail : Laurence Gine l Cortiella - Claire Pelletier

Réalisé	En cours	A faire
Finalisation de la conception sous OBO-Edit	Importation manuelle de 'ontologie pour peupler ICE Ontologie	Installer Karma ou un autre logiciel de mapping Voir comment faire le mapping entre différentes ontologies
Début de mise en place de manuel d'installation de maven	Recherche sur les méthodes d'évaluations des ontologies	Faire valider par Sayanta la cohérence du model

Contraintes/difficultés: nécessité d'un avis métier pour valider l'exactitude, la cohérence des données.

Impossible d'installer Karma : impossible d'installer karma : problème de compiler

Propositions :

Objectif Semaine +2 : construire une première requête avec l'ontologie et s'assurer de la cohérence du résultat

6.3 LISTES DES STANDARDS DU WORLD WIDE WEB CONSORTIUM

- *ARIA (Accessible Rich Internet Applications)*
- *ATAG (Authoring Tool Accessibility Guidelines)*
- *AWWW (Architecture of the World Wide Web)*
- *CC/PP (Composite Capabilities/Preferences Profiles)*
- *CSS (Cascading Style Sheet / Feuilles de style en cascade)*
- *DOM (Document Object Model)*
- *EXI (en) (Efficient XML Interchange)*
- *GRDDL (Gleaning Resource Descriptions from Dialects of Languages)*
- *HTML (HyperText Markup Language)*
- *InkML (en) (Ink Markup Language)*
- *MathML (Mathematics Markup Language)*
- *OWL (Web Ontology Language)*
- *PICS (en) (Platform for Internet Content Selection)*
- *PNG (Portable Network Graphics)*
- *POWDER (en) (Protocol for Web Description Resources)*
- *RDF (Resource Description Framework)*
- *RDFa (Resource Description Framework for HTML)*
- *SKOS (Simple Knowledge Organization System)*
- *SMIL (Synchronized Multimedia Integration Language)*
- *SML (Service Modeling Language)*
- *SOAP (Simple Object Access Protocol)*
- *SPARQL (langage de requête et un protocole pour RDF)*
- *SVG (Scalable Vector Graphics)*
- *UAAG (User Agent Accessibility Guidelines)*
- *WCAG (Web Content Accessibility Guidelines)*
- *WSDL (Web Service Definition Language)*
- *XForms*
- *XHTML (eXtensible HyperText Markup Language)*
- *XML (Extensible Markup Language)*
- *XML Encryption*
- *XML Schema*
- *XML Signature*
- *XPath*
- *XPointer (XML Pointer)*
- *XProc (XML Pipeline Language)*
- *XQuery*

- *XSLT (Extensible Stylesheet Language Transformations)*
- *…*

6.4 LISTE DES ONTOLOGIES CANDIDATES A L'ALIGNEMENT

Ontologie	Description	URL
National Cancer Institute Thesaurus (NCIT)	Le Centre NCI est un organisme américain qui œuvre pour l'informatique biomédicale et de la technologie de l'information (CBIIT). Il fournit et plaide pour l'utilisation appropriée de la science de données, l'informatique et la technologie de l'information (TI) pour soutenir et accélérer la Mission NCI qui est : prévenir et soigner le cancer.	http://ncicb.nci.nih.gov/core/EVS
Expirmental Factor Ontology : EFO	Cette ontologie fournit une description systématique des nombreuses variables expérimentales disponibles dans les bases de données EBI, ainsi que pour des projets externes tels que le catalogue NHGRI GWAS. EFO permet de soutenir l'annotation, l'analyse et la visualisation des données traitées par de nombreux groupes liés à l'EBI et constitue une base ontologique pour le Centre de la Validation	http://www.ebi.ac.uk/efo/index.html

	thérapeutique.	
PharmGKB	PharmGKB est une base de connaissance de la pharmaco génomique qui regroupe les données cliniques, y compris les indications posologiques et l'étiquetage des médicaments. PharmGKB recueille, purge et diffuse des connaissances sur l'impact de la variation génétique humaine sur les réponses aux médicaments	https://www.pharmgkb.org/index.jsp
VariO: Variation Ontology	Vario est une ontologie pour la description standardisée et systématique des effets, des conséquences et des mécanismes des variations des gènes	
SNOMEDCT : Systematized Nomenclature of Medicine -	SNOMED CT est le plus exhaustif et précis des référentiels en terminologie médicale. Il est mis	http://ihtsdo.org

Clinical Terms	en œuvre et détenu par le International Health Terminology Standards Development Organisation (IHTSDO).	
DIKB: Drung Interaction knwoledge base Ontology	Une taxonomie de preuves pour les essais cliniques qui combinés avec un ensemble de critères d'inclusion, permettent aux experts de la pharmacologie de préciser le degré de leur confiance dans les visées thérapeutiques d'un médicament en se basant sur des preuves recueillies	http://dbmi-icode-01.dbmi.pitt.edu/dikb-evidence/front-page.html
MeSH	MeSH est le thésaurus de vocabulaire contrôlé de médecine. Il se compose d'ensembles de termes de nommage descripteurs dans une structure hiérarchique qui permet la recherche à différents niveaux de spécificité	https://www.nlm.nih.gov/mesh/
PubMed	PubMed permet aux auteurs de partager des opinions et des informations sur les publications scientifiques dans PubMed .	http://www.ncbi.nlm.nih.gov/pubmedcommons

Medline	MedlinePlus est le site Web américain de Santé pour les patients et leurs familles. Il est conçu et mis en œuvre par la Bibliothèque nationale de médecine, la plus grande bibliothèque médicale au monde. Il fournit des informations sur les maladies, les conditions et les questions de bien-être dans la langue souhaitée.	http://www.nlm.nih.gov/medlineplus/

7 TABLE DES FIGURES

8 Tables des tableaux

9 BIBLIOGRAPHY

9.2 WIKIPEDIA

Apache Cassandra. (2015, June). Consulté le 16 Juillet 2015, sur Wikipedia: https://en.wikipedia.org/wiki/Apache_Cassandra

Apache HBase. (2015). Consulté le 16 Juillet 2015, sur Wikipedia: https://en.wikipedia.org/wiki/Apache_HBase

Linked Data. (2014, 8 30). Consulté le 14 Juillet 2015, sur linkeddata.org: http://linkeddata.org/

RIAK_Wikipedia. (March 2015). Consulté le 15. Juillet 2015 sur RIAK_Wikipedia: https://en.wikipedia.org/wiki/Riak

Wikipedia. (May 2015). MongoDB. Consulté le 16. Juillet 2015 sur Wikipedia: https://en.wikipedia.org/wiki/MongoDB

Wikipedia. (2015). Triplestore. Consulté le 14 juillet 2015, sur Wikipedia: https://en.wikipedia.org/wiki/Triplestore

9.3 SITES BIOMEDICAUX

CHEBI. (s.d.). Chemical Entity. Consluté en Juillet 2015, sur Chebi: http://www.ebi.ac.uk/ebisearch/search.ebi?query=24431&submit=&db=allebi&requestFrom=global-masthead

ClinicalTrials.gov. (n.d.). consuté le 14 juillet 2015 sur https://clinicaltrials.gov/: https://clinicaltrials.gov/

DIKB. (n.d.). Evidence Type. Consulté le 20 juillet 2015 sur Drug Interaction Knowledge Base Ontology: Drug Interaction Knowledge Base Ontology

DOID. (June 2015). DOID: 305. Consulté en mai 2015 sur Disease Ontology: http://disease-ontology.org/

NCIT. (n.d.). NCIT. Consulté en mai 2015, sur National Cancer Institute Thesarus: https://ncit.nci.nih.gov/ncitbrowser/

OBI. (2011). Main Page. Consulté en mai 2015 sur OBI-Ontology: http://obi-ontology.org/page/Main_Page

VARIO. (kein Datum). Vario. Consulté le 12. avril 2015 sur Variation Ontology: http://variationontology.org/

9.4 PUBLICATION

Berners-Lee, T. (June 1994). Request for Comment 1630. Universal Resource Identifiers in WWW . Geneva, Switzerland: W3C.

TIM BERNERS-LEE, JAMES HENDLER and ORA LASSILA. (2001). The Semantic Web. Scientific American .

Introduction to Triplestores. (2013, January 31). Consulté le 14 juillet 2015, sur Dataversity: http://www.dataversity.net/introduction-to-triplestores/

Warshaw, R. (May 2015). Precision Medicine: Paving the Way for a New Era. Consulté le 14. Juillet 2015 sur AAMC.ORG: https://www.aamc.org/newsroom/reporter/may2015/431964/precision-medicine.html

Virginie Fortineau. Contribution à une modélisation ontologique des informations tout au long du cycle de vie du produit. Chemical and Process Engineering. Ecole nationale supérieure d'arts et métiers - ENSAM, 2013. French. <NNT : 2014ENAM0049>. <pastel-01064598> (page 35)

9.5 AUTRES DOCUMENTATIONS

ICE_Project_Team. (2014). *ICE Dossier D'architecture.*

9.6 APERCU FICHIER OWL DE L'ONTOLOGIE ICE

```
<!-- http://www.semanticweb.org/lgineico/ontologies/2015/4/ICE_OntologyV2#Case-Control_Study -->

<owl:Class rdf:about="&ICE_OntologyV2;Case-Control_Study">
    <rdfs:subClassOf rdf:resource="&ICE_OntologyV2;Retrospective_Study"/>
</owl:Class>


<!-- http://www.semanticweb.org/lgineico/ontologies/2015/4/ICE_OntologyV2#Clinical_Trials -->

<owl:Class rdf:about="&ICE_OntologyV2;Clinical_Trials">
    <rdfs:subClassOf rdf:resource="&ICE_OntologyV2;Treatment"/>
    <Definition>A research study using human subjects to evaluate the effect of interventions or exposures on biomedical or
</owl:Class>


<!-- http://www.semanticweb.org/lgineico/ontologies/2015/4/ICE_OntologyV2#Copy_Type_variation -->

<owl:Class rdf:about="&ICE_OntologyV2;Copy_Type_variation">
    <rdfs:subClassOf rdf:resource="&ICE_OntologyV2;Number_Variation"/>
</owl:Class>


<!-- http://www.semanticweb.org/lgineico/ontologies/2015/4/ICE_OntologyV2#DNA -->

<owl:Class rdf:about="&ICE_OntologyV2;DNA">
    <rdfs:subClassOf rdf:resource="&ICE_OntologyV2;Gene"/>
    <Definition>DeoxyriboNucleic Acid is a molecule that encodes the genetic instructions used in the development and funct
```

Figure 87 : Exemple de Classe

change.log | LGI_Sujet_1.py | List_exo.py | List_exo1.py | List_exo1_2.py | ice_v2

```
59
60
61      <!-- http://www.semanticweb.org/lgineico/ontologies/2015/4/ICE_OntologyV2#causes -->
62
63      <owl:ObjectProperty rdf:about="&ICE_OntologyV2;causes">
64          <rdfs:range rdf:resource="&ICE_OntologyV2;Cancer"/>
65          <rdfs:domain rdf:resource="&ICE_OntologyV2;Variation"/>
66      </owl:ObjectProperty>
67
68
69
70      <!-- http://www.semanticweb.org/lgineico/ontologies/2015/4/ICE_OntologyV2#evaluates -->
71
72      <owl:ObjectProperty rdf:about="&ICE_OntologyV2;evaluates">
73          <rdfs:domain rdf:resource="&ICE_OntologyV2;Clinical_Trials"/>
74          <rdfs:range rdf:resource="&ICE_OntologyV2;Drug"/>
75      </owl:ObjectProperty>
76
77
78
79      <!-- http://www.semanticweb.org/lgineico/ontologies/2015/4/ICE_OntologyV2#has_condition -->
80
81      <owl:ObjectProperty rdf:about="&ICE_OntologyV2;has_condition"/>
82
83
84
85      <!-- http://www.semanticweb.org/lgineico/ontologies/2015/4/ICE_OntologyV2#has_gene_variation -->
86
87      <owl:ObjectProperty rdf:about="&ICE_OntologyV2;has_gene_variation">
88          <rdfs:domain rdf:resource="&ICE_OntologyV2;Patient"/>
89          <rdfs:range rdf:resource="&ICE_OntologyV2;Variation"/>
```

Figure 88 : Exemple d'Object Property

change.log | LGI_Sujet_1.py | List_exo.py | List_exo1.py | List_exo1_2.py | ice_v2

```
299         <rdfs:domain rdf:resource="&ICE_OntologyV2;Cancer"/>
300         <rdfs:subPropertyOf rdf:resource="&ICE_OntologyV2;Cancer"/>
301         <rdfs:range rdf:resource="&xsd;string"/>
302     </owl:DatatypeProperty>
303
304
305
306     <!-- http://www.semanticweb.org/lgineico/ontologies/2015/4/ICE_OntologyV2#chromosome -->
307
308     <owl:DatatypeProperty rdf:about="&ICE_OntologyV2;chromosome">
309         <rdfs:subPropertyOf rdf:resource="&ICE_OntologyV2;gene_location"/>
310     </owl:DatatypeProperty>
311
312
313
314     <!-- http://www.semanticweb.org/lgineico/ontologies/2015/4/ICE_OntologyV2#clinical_trial_ID -->
315
316     <owl:DatatypeProperty rdf:about="&ICE_OntologyV2;clinical_trial_ID">
317         <rdfs:subPropertyOf rdf:resource="&ICE_OntologyV2;Treatment"/>
318     </owl:DatatypeProperty>
319
320
321
322     <!-- http://www.semanticweb.org/lgineico/ontologies/2015/4/ICE_OntologyV2#clinical_trial_conditio
323
324     <owl:DatatypeProperty rdf:about="&ICE_OntologyV2;clinical_trial_condition">
325         <rdfs:domain rdf:resource="&ICE_OntologyV2;Clinical_Trials"/>
326         <rdfs:subPropertyOf rdf:resource="&ICE_OntologyV2;Treatment"/>
327     </owl:DatatypeProperty>
328
```

Figure 89 : Exemple de Data Property

change.log | LGI_Sujet_1.py | List_exo.py | List_exo1.py | List_exo1_2.py | ice_v2

```
1505    <owl:NamedIndividual rdf:about="&ICE_OntologyV2;CD95L">
1506        <owl:sameAs rdf:resource="&ICE_OntologyV2;LG_gene1"/>
1507    </owl:NamedIndividual>
1508
1509
1510
1511    <!-- http://www.semanticweb.org/lgineico/ontologies/2015/4/ICE_OntologyV2#CT1 -->
1512
1513    <owl:NamedIndividual rdf:about="&ICE_OntologyV2;CT1">
1514        <rdf:type rdf:resource="&ICE_OntologyV2;Interventional_Study"/>
1515        <Definition>Evaluate Risk/Benefit of Nab Paclitaxel in Combination With Gemcitabine and Carboplatin Compared to Gemcit
1516        <clinical_trial_start_date>June 2013</clinical_trial_start_date>
1517        <clinical_trial_end_date>April 2016</clinical_trial_end_date>
1518        <clinical_trial_condition>Breast Tumor
1519 Breast Cancer
1520 Cancer of the Breast
1521 Estrogen Receptor- Negative Breast Cancer
1522 HER2- Negative Breast Cancer
1523 Progesterone Receptor- Negative Breast Cancer
1524 Recurrent Breast Cancer
1525 Stage IV Breast Cancer
1526 Triple-negative Breast Cancer
1527 Triple-negative Metastatic Breast Cancer
1528 Metastatic Breast Cancer</clinical_trial_condition>
1529        <clinical_trial_title>A Phase 2/3, Multi-Center, Open-Label, Randomized Study of Weekly Nab®-Paclitaxel in Combination
1530        <clinical_trial_sponsor>Celgene Corporation</clinical_trial_sponsor>
1531        <clinical_trial_identifier>NCT01881230</clinical_trial_identifier>
1532        <evaluates rdf:resource="&ICE_OntologyV2;dr4"/>
1533        <evaluates rdf:resource="&ICE_OntologyV2;dr5"/>
1534    </owl:NamedIndividual>
1535
```

Figure 90 : Instance d'Essai clinique (individu)