Automated adaptation of Electronic Heath Record for secondary use in oncology

Vianney Jouhet

université de BORDEAUX

THÈSE

PRÉSENTÉE À

L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE SOCIETE, POLITIQUE, SANTE PUBLIQUE

par **Vianney Jouhet**

POUR OBTENIR LE GRADE DE

DOCTEUR

SPÉCIALITÉ : SANTE PUBLIQUE, OPTION: INFORMATIQUE ET SANTE

Adaptation automatique des données de prises en charge hospitalières pour une utilisation secondaire en cancérologie

Automated adaptation of Electronic Heath Record for secondary use in oncology

---

Date de soutenance : 16 décembre 2016

Devant la commission d'examen composée de :

Roger Salamon PU-PH, Université de Bordeaux,Bordeaux	President
Pierre Ingrand PU-PH, Université de Poitiers, Poitiers	Examinateur
Simone Mathoulin-Pélissier PU-PH, Inserm-U1219,Bordeaux	Examinateur
Anita Burgun PU-PH, Inserm-U1138,Paris	Rapporteur
Marc Cuggia PU-PH, Inserm-U1099,Rennes	Rapporteur
Frantz Thiessard MCU-PH, Inserm-U1219,Bordeaux	Directeur

– 2016 –

Résumé Avec la montée en charge de l'informatisation des systèmes d'information hospitaliers, une quantité croissante de données est produite tout au long de la prise en charge des patients. L'utilisation secondaire de ces données constitue un enjeu essentiel pour la recherche ou l'évaluation en santé.

Dans le cadre de cette thèse, nous discutons les verrous liés à la représentation et à la sémantique des données, qui limitent leur utilisation secondaire en cancérologie. Nous proposons des méthodes basées sur des ontologies pour l'intégration sémantique des données de diagnostics. En effet, ces données sont représentées par des terminologies hétérogènes. Nous étendons les modèles obtenus pour la représentation de la maladie tumorale, et les liens qui existent avec les diagnostics. Enfin, nous proposons une architecture combinant entrepôts de données, registres de métadonnées et web sémantique. L'architecture proposée permet l'intégration syntaxique et sémantique d'un grand nombre d'observations. Par ailleurs, l'intégration de données et de connaissances (sous la forme d'ontologies) a été utilisée pour construire un algorithme d'identification de la maladie tumorale en fonction des diagnostics présents dans les données de prise en charge. Cet algorithme basé sur les classes de l'ontologie est indépendant des données effectivement enregistrées. Ainsi, il fait abstraction du caractère hétérogène des données diagnostiques initialement disponibles.

L'approche basée sur une ontologie pour l'identification de la maladie tumorale, permet une adaptation rapide des règles d'agrégation en fonction des besoins spécifiques d'identification. Ainsi, plusieurs versions du modèle d'identification peuvent être utilisées avec des granularités différentes.

Mots-clés Informatique Médicale, Oncologie, web sémantique, registre de métadonnées, Ontologie, Entrepôt de données clinique

Laboratoire d'accueil Equipe de Recherche en Informatique Appliquée à la Santé, INSERM U1219, Université de Bordeaux, rue Léo Saignat 33000 Bordeaux France

Title Automated adaptation of Electronic Heath Record for secondary use in oncology

Abstract With the increasing adoption of Electronic Health Records (EHR), the amount of data produced at the patient bedside is rapidly increasing. Secondary use is thereby an important field to investigate in order facilitate research and evaluation.

In these work we discussed issues related to data representation and semantics within EHR that need to be address in order to facilitate secondary of structured data in oncology. We propose and evaluate ontology based methods for heterogeneous diagnosis terminologies integration in oncology. We then extend obtained model to enable tumoral disease representation and links with diagnosis as recorded in EHR. We then propose and implement a complete architecture combining a clinical data warehouse, a metadata registry and web semantic technologies and standards. This architecture enables syntactic and semantic integration of a broad range of hospital information System observation. Our approach links data with external knowledge (ontology), in order to provide a knowledge resource for an algorithm for tumoral disease identification based on diagnosis recorded within EHRs. As it based on the ontology classes, the identification algorithm is uses an integrated view of diagnosis (avoiding semantic heterogeneity).

The proposed architecture leading to algorithm on the top of an ontology offers a flexible solution. Adapting the ontology, modifying for instance the granularity provide a way for adapting aggregation depending on specific needs.

Keywords Medical Informatics, Oncology, semantic web, metadata registry, ontology, clinical data warehouse

Remerciements

Au Professeur Roger Salamon
Je suis honoré que vous ayez accepté de présider mon jury de thèse. Je tiens aussi à vous remercier pour m'avoir soutenu lors de mon arrivée à Bordeaux et permis de trouver ma place au sein du Service d'Information Médicale du CHU de Bordeaux, de l'ISPED et de l'équipe ERIAS.

Au Professeur Pierre Ingrand
Je vous remercie d'avoir accepté de participer à mon jury de thèse. Une large partie de ce travail a été initié au sein de votre unité au cours de mon internat de Santé Publique. Je vous suis reconnaissant de m'avoir poussé à m'épanouir pleinement en me faisant découvrir le domaine de l'Informatique Médicale que je continue à explorer avec passion.

Au Professeur Simone Mathoulin-Pélissier
Je vous exprime mes sincères remerciements pour avoir accepté de juger mon travail. Depuis mon arrivée au sein du Registre des Cancers de la Gironde vous m'avez donné de nombreuses opportunités de m'engager sur des projets passionnants. Je vous remercie de l'intérêt que vous portez à ce travail.

Au Professeur Anita Burgun
Je suis honoré que tu aies accepté d'être rapporteur de cette thèse. J'espère que tu jugeras ce travail digne de la formation que j'ai reçue lors de mon Master d'Informatique Médicale à Rennes.

Au Professeur Marc Cuggia
Je te remercie de me faire l'honneur d'être rapporteur de cette thèse. J'espère que nous aurons encore l'occasion de travailler ensemble sur ces sujets qui nous passionnent.

Au Docteur Frantz Thiessard
Je te remercie d'avoir dirigé cette thèse. Merci pour ton aide et ton soutien tout au long de ce travail. Au delà de ce ton encadrement, je te remercie d'avoir

eu confiance en moi lors mon arrivée, et de me permettre de participer au sein de l'équipe ERIAS à l'essor de notre discipline à Bordeaux.

Au Professeur Geneviève Chêne
Je tiens à vous remercier pour m'avoir permis de réaliser ce travail. Soyez assurée de ma reconnaissance pour votre soutien et la confiance que vous m'avez accordée au sein du Service d'Information Médicale du CHU de Bordeaux.

Au Professeur Alain Ravaud
Je vous remercie d'avoir accepté de participer à mon jury de thèse. Je regrette que vous ne puissiez être effectivement présent pour cette soutenance. Je vous remercie également pour l'intérêt que vous avez porté dès le début de ce travail et votre soutien pour sa mise en œuvre effective.

A l'équipe ERIAS
Je remercie vivement les membres de l'équipe ERIAS : Fleur, Valérie, Gayo et Frantz. C'est une chance de pouvoir travailler et échanger au sein d'une équipe ouverte, passionnée et passionnante. Évidemment j'ai également une pensée pour Bruno, Elise, Jean Noël, Clément, Sébastien et Bérénice avec qui j'ai pu partager certains projets.

A l'Unité IAM du CHU de Bordeaux
En particulier à Moufid Hajjar pour ton accueil chaleureux, ton soutien et ta bienveillance dans la direction de notre unité.

Au Service d'Information Médicale du CHU de Bordeaux
Au sein duquel ce travail a été réalisé en grande partie. Merci notamment à Aurélie, Nathalie et Sylvie pour avoir supporté mes longs silences pendant ces derniers mois ...

A l'équipe du Registre Général des Cancers de la Gironde
Pour m'avoir permis d'initier ce travail. Merci notamment à Brice et Gaëlle pour ces échanges informels entre deux réunions ...

A Françoise Colombani
Merci pour tous ces échanges constructifs concernant le système d'information en cancérologie. Merci aussi pour l'intérêt que tu as porté à ce projet dès le début et le soutien permanent que tu m'as apporté.

A Erika, Maïalen et Clémence Pour m'avoir porté (et supporté) tout au long de ce travail. Merci évidemment d'être simplement là à donner du sens à chaque jour. **Mes parents, ma famille, mes amis** et autres musiciens de Los de Seiche, des Roccos pour tout ces bon moments passés et à venir.

Table des matières

Résumé substantiel

Introduction

Utilisation secondaire des données biomédicales

L'utilisation secondaire des données biomédicales est un enjeu reconnu. En effet, avec la montée en charge de l'informatisation des hôpitaux, la quantité de données produites chaque jour est de plus en plus importante. Ainsi, de nombreux projets visant à réutiliser ces données ont vu le jour depuis plusieurs années [1, 2, 3, 4, 5, 6, 7].

En 2007, l'American Medical Informatics Association insistait sur les bénéfices attendus de cette utilisation secondaire :*"Secondary use of health data can enhance healthcare experiences for individuals, expand knowledge about disease and appropriate treatments, strengthen understanding about the effectiveness and efficiency of our healthcare systems, support public health and security goals, and aid businesses in meeting the needs of their customers"* [4]. Par ailleurs, certains centres hospitalo-universitaires aux Etats-Unis et en Europe ont montré les opportunités que génère la mise en place de plateformes pour l'utilisation des données prise en charge hospitalière [8, 9].

Malgré ce constat, la mise en œuvre de solutions dans les centres hospitaliers reste lente, notamment du fait de la complexité des données [10]. Les principaux challenges reconnus sont notamment l'intégration sémantique et syntaxique de données hétérogènes, l'identification de phénotypes (situations cliniques) et l'évaluation de la qualité des données (qui peut concerner les données utilisées, mais aussi les données construites à partir des données brutes) [11].

Le domaine de la cancérologie est, lui aussi, impacté par ce phénomène. Les registres des cancers, pour l'identification de nouveaux cas potentiels, utilisent massivement les données issues des établissements sur les territoires qu'ils couvrent. Cette démarche est notamment mise en œuvre pour faciliter l'enregistrement des cas, en recherchant de façon active les cas qui surviennent sur le territoire couvert. Si certains registres ont une démarche basée sur l'exploration manuelle des dossiers, d'autres ont commencé à utiliser les données disponibles depuis de nombreuses années [12, 13]. Cette démarche a notamment été ren-

due indispensable pour des registres couvrant une population importante [14]. Par ailleurs, dans le contexte de la médecine personnalisée (ou médecine de précision), l'identification de situations clinico-biologiques spécifiques nécessite l'intégration de données toujours plus nombreuses et hétérogènes autour du patient. L'intégration des données de prise en charge avec des données biologiques (génomiques, immunologiques), ou d'imagerie est un enjeu important pour : (i) la visualisation de dossiers complets pour la prise de décision médicale ; (ii) l'exploitation des données dans un objectif de recherche translationnelle. Ces approches nécessitent notamment l'intégration des données de prise en charge avec les bio-banques [15, 16]. Dans ce contexte, l'Institut National du CAncer a notamment financé la mise en œuvre de bases clinico-biologiques. Au même titre que les registres des cancers, ces structures bénéficieraient d'un enrichissement avec les données issues directement des dossiers patient informatisés.

Dans cette partie, nous identifions et discutons les causes de la sous-utilisation des données de prise en charge en cancérologie. Nous analysons en particulier les aspects liés à la représentation des diagnostics de cancer sous forme de données structurées au sein des dossiers patient informatisés, au regard du besoin d'identifier la maladie cancéreuse pour une utilisation secondaire efficiente. Enfin, nous proposons un ensemble de solutions à investiguer pour lever ces verrous.

Représentation du diagnostic et de la maladie

Représentation du diagnostic en cancérologie dans les données de prise en charge.

Tout au long de la prise en charge, les médecins collectent des données concernant le patient. Ces données très hétérogènes peuvent être représentées de façon très structurée (diagnostic codé suivant une terminologie standard) mais aussi sous forme de documents en texte libre. D'une manière générale, chaque élément de données qui concerne le patient peut être considérée comme une observation, dont le niveau de structuration peut varier en fonction des choix des utilisateurs et des outils disponibles pour l'acquisition de ces données.

En cancérologie, trois sources de données majeures produisent des données structurées pour décrire les diagnostics : le Programme de Médicalisation du Système d'Information (PMSI), l' Anatomie et Cytologie Pathologiques (ACP) et les Réunions de Concertation Pluridisciplinaires (RCP). Ces sources utilisent des terminologies différentes pour renseigner leur diagnostic. Chaque fois qu'un patient est vu par un clinicien, un (ou plusieurs) diagnostic(s) sont enregistrés afin de rendre compte du motif d'hospitalisation ou de la conclusion du clinicien (figure 2.1). Ces données de diagnostic ont plusieurs caractéristiques importantes :

— Plusieurs terminologies sont utilisées pour enregistrer le diagnostic.

— Chaque observation est enregistrée de façon indépendante (si ce n'est qu'elle est rattachée à la même venue du patient).
— De nombreux diagnostics peuvent être enregistrés pour un patient. Cependant, d'une manière générale, la maladie n'est elle-même jamais explicitement enregistrée. Ainsi, les informations qui concernent une maladie pour un patient ne sont pas reliées de façon explicite.

Représentation de la maladie dans les données de recherche

Les données produites par les registres des cancers ainsi que les données des bases clinico-biologiques ont de leur côté plusieurs caractéristiques communes :
— Les patients y sont inclus en fonction de l'existence d'une maladie. Pour les registres des cancers, les critères d'inclusion peuvent comprendre les tumeurs solides, ou les hémopathies par exemple. Les bases de données clinico-biologiques sont, quant à elles, centrées sur un type de cancer particulier (soit par son type histologique, soit par son site primitif).
— Ces données sont souvent utilisées pour la mise en œuvre d'études ancillaires. Ces études nécessitent l'identification de situations cliniques précises dont la définition inclut souvent des informations liées à une maladie. Par exemple, l'identification de *patients ayant une tumeur métastatique du colon n'ayant pas fait l'objet d'une chirurgie première* nécessite des liens explicites entre la maladie (tumeur du colon), son évolution (métastase) et les traitements mis en œuvre (chirurgie).

Ainsi, les données de recherche sont organisées autour de la maladie, alors que les données de prises en charge sont organisées autours du patient et de sa venue. Les figures 2.2 et 2.3 montrent les modèles sous-jacents aux données des dossiers patient informatisés et des données de recherche respectivement. Ces caractéristiques sont une différence entre ces données.

L'identification de la maladie et l'explicitation des liens entre la maladies et les éléments qui s'y rattachent constituent donc deux éléments essentiels dans la construction de données adaptées pour un usage secondaire à partir des données de prises en charge.

Problématiques pour l'utilisation secondaire des données en cancérologie

Hétérogénéité des terminologies diagnostiques en cancérologie

En France, trois terminologies sont principalement utilisées pour renseigner les diagnostics en cancérologie : la Classification Internationale des Maladies en Oncologie (CIM-O-3), la Classification Internationale de Maladies (CIM-10) et la classification de l'Association pour le Développement de l'Informatique en Cytologie et en Anatomie Pathologiques (ADICAP). Ces terminologies sont hétérogènes à deux niveaux :

— **Pré-coordination vs post-coordination**. La CIM-10 est une terminologie monoaxiale. Elle représente le diagnostic sous la forme d'un seul code qui porte l'ensemble de l'information (les termes sont représentés de façon pré-coordonnée). La CIM-O-3 (ainsi qu'ADICAP) représente le diagnostic en combinant la morphologie (type histologique) et le site primitif (organe d'origine) de la tumeur (la construction du diagnostic est post-coordonnée en combinant le site primitif et la morphologie). Ainsi, le passage de CIM-10 vers la CIM-O-3 est une tâche de composition (et décomposition). Pour que des données représentées dans ces terminologies puissent être intégrées, il est nécessaire de proposer des méthodes permettant de réaliser une composition automatique.

— **Granularité différente**. La CIM-10 (notamment du fait de son caractère pré-coordonné) propose peu de détails pour l'enregistrement de la morphologie tumorale. Ainsi, en fonction de la terminologie utilisée, il est possible de faire référence à la même maladie avec une précision différente (du fait de la granularité disponible dans les terminologies).

La figure 2.6 présente un exemple de ces deux niveaux d'hétérogénéité. Ainsi, le passage d'une terminologie vers une autre n'est pas une tâche simple. Il est nécessaire de proposer des solutions offrant une vue intégrée de ces terminologies. Il est en particulier important de fournir des accès transparents aux concepts représentés par ces terminologies, quelques soient leurs structures ou leurs granularités d'origine.

Représentation implicite de la maladie tumorale dans les données de prise en charge

Des définitions ont été proposées pour les concepts de maladie et de diagnostic par Scheuermann et al [17] :

— "***Diagnosis* =*def.*** *– A conclusion of an interpretive process that has as input a clinical picture of a given patient and as output an assertion to the effect that the patient has a disease of such and such a type.*

A diagnosis is a continuant entity that, once made, will survive through time, and is often supplanted by further diagnoses. The diagnostic process is thus iterative : the clinician is forming hypotheses during history taking, testing these during physical exam, forming new hypotheses as a result, and so on. "

— "***Disease* =*def.*** *– A disposition (i) to undergo pathological processes that (ii) exists in an organism because of one or more disorders in that organism.* "

Les données de recherche sont produites en analysant les dossiers des patients à inclure. Une des tâches majeures est d'interpréter le dossier pour identifier la maladie, et y rattacher les éléments d'intérêt. Cette tâche est indispensable

pour que les données puissent être exploitées à des fin de recherche.

Les données de prise en charge, quant à elles, n'enregistrent jamais (ou rarement) la maladie elle-même. Les diagnostics sont enregistrés sous la forme d'une succession indépendante d'observations. La maladie n'étant pas elle-même explicitement représentée, les éléments qui devraient s'y rattacher sont également enregistrés sous la forme d'observations indépendantes. L'interprétation et la combinaison de ces observations entre elles est une tâche qu'un médecin peut réaliser rapidement à la lecture d'un dossier. En effet, en s'appuyant sur ses connaissances, il est capable de reconstruire les liens qui existent entre les observations.

La construction de données adaptées à un usage secondaire passe donc par :

— L'identification de la maladie (implicite dans les données de prise en charge).
— L'explicitation des relations entre les observations et la maladie identifiée.

Cette tâche repose notamment sur des connaissances médicales qui ne sont pas présentes dans les données de prise en charge. Il est donc nécessaire de (i) fournir ces connaissances selon un format interprétable par des machines afin de les exploiter et (ii) d'établir les liens entre ces connaissances et les observations effectivement enregistrées dans les données de prise en charge.

Intégration de données et de connaissances pour l'identification de la maladie

Afin de répondre à la problématique d'identification de la maladie à partir des données de prises en charge, il est nécessaire de traiter différents aspects :

— Intégration syntaxique
— Intégration sémantique
— Intégration des données et des connaissances.

Nous présentons dans la partie suivante les solutions et standards qui sont en lien avec ces problématiques.

Intégration syntaxique : les entrepôts de données cliniques

Dans le domaine médical, les entrepôts de données cliniques (EDC) constituent la solution la plus largement adoptée pour l'intégration de données des dossiers patient informatisés. Parmi les solutions actuellement disponibles, i2b2 est un projet open source, qui propose notamment un EDC au sein d'une infrastructure visant à faciliter l'utilisation des données des dossiers patient informatisés.

Intégration sémantique : Le Web Sémantique

Le Web Sémantique vise à partager du contenu directement entre machines [18]. Un des enjeux du web sémantique est la représentation, le stockage et le partage de connaissances. Dans ce contexte, plusieurs standards existent pour représenter ces connaissances de façon plus ou moins formelle :

— **Web Ontology Language (OWL)**. OWL est un langage défini par le W3C pour décrire les ontologies. Les documents OWL sont exploitables par des machines. Notamment, la syntaxe permet de décrire un domaine suivant la logique de description. Cette logique peut être exploitée par des raisonneurs pour vérifier la cohérence de la représentation, et inférer des axiomes supplémentaires.

— **Simple Knowledge Organization System (SKOS)**. SKOS est un langage basé sur OWL qui permet de décrire des terminologies. Il permet de décrire notamment des ressources moins formelles que des ontologies en proposant notamment une sémantique moins contrainte pour les relations hiérarchiques.

Intégration des données et des connaissances : Les registres de métadonnées

Les registres de métadonnées sont des systèmes de stockage et de gestion de données "à propos des données". Ils sont par exemple utilisés pour maintenir une certaine consistance des données au sein d'une organisation. Ils peuvent notamment être utiles pour le partage de données et la mise en oeuvre des entrepôts de données. Une norme (ISO/IEC 11179) permet de définir la représentation des métadonnées au sein de ces registres.

Dans le domaine biomédical, plusieurs projets ont utilisé les technologies et standards du web sémantique et notamment pour faciliter l'utilisation secondaire des données. Par ailleurs, des registres de métadonnées ont été implémentés pour faciliter le partage et la réutilisation de données. Dans le cadre du projet SALUS, Semantic MDR [19] est un registre de métadonnées basé sur ISO/IEC 11179. Il implémente une version OWL du métamodèle ISO/IEC 11179.

Conclusion : solutions à investiguer

Comme nous l'avons discuté ci-dessus, l'utilisation secondaire des données biomédicales est en particulier limitée par le fait que les représentations ne sont pas les mêmes entre les données de prise en charge, et les données de la recherche. L'identification de la maladie à partir de diagnostics et l'explicitation de lien implicites, sont des éléments essentiels pour permettre une adaptation des données de prise en charge à un usage secondaire.

Les registres de métadonnées permettent de décrire les données, leur représentation physique, et de faire des liens vers les concepts qui les représentent. Le Web Sémantique permet de gérer des terminologies qui sont une façon de normaliser les concepts à représenter. Par ailleurs, OWL permet de représenter des modèles formels décrivant un domaine particulier. L'ensemble de ces éléments est spécifié de façon standard et des outils ont été développés pour permettre l'exploitation de ces ressources par des machines. Ainsi, en combinant les registres de métadonnées, les terminologies et les modèles formels, il est possible de faire le lien entre des données et les connaissances nécessaires à leur interprétation. Combinées avec une solution de stockage et d'intégration syntaxique telle que les EDC, ces technologies doivent permettre de répondre au besoin d'intégration de données et de connaissances, pour adapter automatiquement les données de prise en charge à une utilisation secondaire.

Dans la suite de ce document, nous décrivons :

— La construction d'un modèle formel pour l'intégration des terminologies diagnostiques en cancérologie, et la description de la maladie à partir des diagnostics

— Le développement et la mise en œuvre d'une architecture permettant l'intégration syntaxique et sémantique, ainsi que l'identification de la maladie cancéreuse à partir des données de prise en charge.

La construction d'un modèle formel pour l'intégration des terminologies diagnostiques en cancérologie, et la description de la maladie à partir des diagnostics

Dans cette partie, nous traitons de deux problématiques liées à l'intégration de données de prise en charge en cancérologie :

— D'une part, nous proposons une solution basée sur des ontologies pour l'intégration des terminologies diagnostiques en cancérologie.

— D'autre part, nous proposons de modéliser conjointement le diagnostic et la maladie et les liens qui existent entre ces concepts.

Deux approches ont été explorées. La première se base sur une ressource existante (le NCI thésaurus) pour construire un modèle permettant de décrire les concepts à intégrer, et les relations entre eux. La deuxième consiste à construire un modèle *de novo* permettant de décrire des classes correspondant aux règles pour l'enregistrement des tumeurs multiples par les registres des cancers.

Contexte

L'identification des cas incidents de cancer du sein d'une population reste un enjeu important pour faciliter la recherche en cancérologie. De nos jours, avec la montée en charge de l'informatisation des hôpitaux, de plus en plus de données sont produites quotidiennement tout au long de la prise en charge. En cancérologie, des terminologies sont utilisées pour l'enregistrement des diagnostics. Ces données structurées constituent une source d'une grande richesse qui doit être exploitée pour l'identification de ces cas incidents. Cependant, dans le domaine de la cancérologie, de multiples acteurs sont amenés à voir les patients tout au long de leur prise en charge. Ces acteurs, en fonction de leur spécialité, utilisent des terminologies différentes pour le codage du diagnostic. Ces terminologies diffèrent par leur granularité et leur structure :

- La Classification Internationale des Maladies (CIM-10) est une terminologies monoaxiale qui enregistre le diagnostic de cancer sous la forme d'un seul code.
- La Classification Internationale des Maladies en Oncologie (CIM-O-3) est une terminologie multiaxiale qui enregistre le diagnostic de cancer sous la forme d'une combinaison d'un site primitif (origine de la tumeur et d'un type histologique).

Ainsi, il n'est pas possible d'aligner directement la CIM-10 avec les axes de la CIM-O-3 puisqu'un diagnostic ne peut être équivalent à une topographie ou une morphologie. L'objectif de ce travail est de proposer la construction d'un modèle permettant d'intégrer les terminologies diagnostiques en prenant en compte cette hétérogénéité.

Construction d'un modèle pour l'intégration des terminologies diagnostiques en cancérologie. Une approche basée sur le NCI thésaurus

Méthodes

Le NCI thésaurus (NCIt) est une ressource *"ontology like"* qui décrit à la fois le diagnostic, le site anatomique, le type histologique ainsi que les relations entre ces concepts. Notre approche a consisté à construire un modèle formel basé sur les éléments disponibles dans la version OWL du NCIt. Cette construction a été réalisée en plusieurs étapes :

- Définition d'un modèle de haut niveau décrivant les éléments à intégrer (diagnostic, site anatomique, et morphologie).
- Identification de ces éléments au sein du NCIt.
- Construction de classes diagnostiques définies à partir des éléments disponibles au sein du NCIt.
- Instanciation du modèle obtenu avec les codes de la CIM-10 et de

la CIM-O-3 en s'appuyant sur le NCI métathésaurus pour définir les classes anonymes représentant les codes diagnostics au sein de la ressource créée.

Au cours de ces étapes, nous avons notamment proposé d'introduire le type histologique comme de nouvelles classes dans le NCIt en nous appuyant sur des annotations liant le diagnostic avec le type histologique codé en CIM-O-3. Par ailleurs, nous avons utilisé les caractéristiques des terminologies à intégrer pour faciliter l'identification des classes à instancier (par exemple : en fonction du comportement tumoral).

Afin d'évaluer la qualité du modèle obtenu et de l'instanciation des terminologies, nous avons recherché les liens obtenus à partir du modèle entre le code CIM-10 et des combinaisons de codes CIM-O-3 (topographie et morphologie). Ces liens ont été comparés avec le fichier de conversion de la CIM-O3 vers la CIM-10 fourni par le SEER program.

Résultats

Le modèle final comprenait 27 953 classes (6 720 topographies, 1 100 morphologies et 20 133 diagnostics). Au total, 1 440 codes ont pu être utilisés pour instancier le modèle (278 topographies CIM-O-3 soit 68%, 860 morphologies CIM-O-3 soit 98% et 302 codes CIM-10 soit 42%). La comparaison avec les mappings du SEER program montrait que notre modèle permettait de construire des relations cohérentes entre la CIM-10 et la CIM-O-3.

Conclusion

Notre approche a permis de construire un modèle capable d'intégrer les terminologies diagnostiques en prenant en compte leur hétérogénéité. Au cours du processus de construction, nous avons pu mettre en évidence des inconsistances au sein du NCIt. Ces erreurs ont déjà été rapportées par des travaux antérieurs. L'utilisation de caractéristiques propres aux terminologies à intégrer pour les définir au sein du modèle, a permis de mettre en évidence un certain nombre d'inconsistances supplémentaires (lié notamment à du *"is overloading"*). Malgré les erreurs identifiées, le NCIt constitue une bonne base de départ pour construire un modèle permettant d'intégrer les terminologies diagnostiques en cancérologie. Cependant, la couverture reste insuffisante, et les erreurs identifiées ne permettent pas d'envisager une utilisation automatique du modèle. Un travail doit être mené pour combiner la ressource construite avec d'autres modèles semi formels (SNOMED-CT, FMA) et ainsi améliorer la représentation des diagnostics en fonction des sites primitifs et des types histologiques.

Construction d'une ontologie basée sur les règles pour l'enregistrement des sites primitifs multiples au sein des registres des cancers

Méthodes

L'Internationnal Agency of Cancer Registries (IACR) a publié une série de règles pour définir l'enregistrement des sites primitifs multiples chez un patient. Ces règles visent à harmoniser le recueil de données en vue de produire des données d'incidence comparables d'un pays à l'autre. Ces règles se basent sur le regroupement des topographies et des morphologies tumorales codées en CIM-O-3 pour définir l'agrégation de diagnostics autour d'une maladie ou l'enregistrement d'une nouvelle maladie. D'un point de vue formel, nous avons considéré que les agrégats de codes diagnostics décrits dans les règles de l'IARC décrivaient un processus tumoral sous-jacent. Ce processus pouvant lui-même être décrit par plusieurs diagnostics au cours de la prise en charge du patient. Afin de construire le modèle correspondant, nous avons procédé en plusieurs étapes :

- Définition d'un modèle de haut niveau décrivant les classes à intégrer et les relations permettant la composition et la décomposition des classes.
- Déclinaison au sein de ce modèle des classes correspondant aux groupes topographiques et morphologiques tels que définis au sein des règles IACR.
- Construction de classes *diagnostics* décrites à partir des groupes topographiques et morphologiques.
- Construction de classes *maladie* décrites en fonction des diagnostics.
- Instanciation du modèle avec les codes issus de trois terminologies (ADICAP, CIM-10 et CIM-O-3). Cette instanciation c'est faite en s'appuyant sur des mappings avec la CIM-O-3 (qui permet de définir les classes du modèle IACR).

Résultats

Le modèle de haut niveau a été construit avec 16 classes et 7 propriétés. Il est constitué de 101 axiomes . Le modèle complet, déclinant les groupes IACR a été construit avec 2 389 classes et est constitué de 9 582 axiomes.

Le modèle exploite les définitions formelles des classes pour composer des classes définies en fonction d'éléments atomiques (i.e. *histologie*, *site anatomique* et *comportement tumoral*). Les concepts composés comme le *diagnostic* et la *morphologie* ont été classifiés par raisonnement automatique en se basant sur leur définition formelle. Cette approche permet d'éviter la problématique du "is_a overloading" [20] et permet d'assurer une classification consistante de ces éléments au sein du modèle.

Un total de 190 classes de *morphologies* ont été construites basées sur les 19 classes d'*histologies* et les 10 classes de *comportement tumoral*. Combinées avec les 53 *topographies*, ces *morphologies* ont permi la construction de 1 569 classes de *diagnostics* liés à 538 classes de *maladies*.

Notre approche a permi d'intégrer dans le modèle 88,0% des codes des terminologies ciblées. La CIM-10 n'était pas complètement intégrée avec 67,3% des codes. Parmi les codes non intégrés une part importante était représentée par des codes de tumeurs hématopoïétiques et de tumeurs bénignes pour lesquelles aucun alignement avec la CIM-O-3 n'avait été défini dans les ressources utilisées.

Architecture pour l'intégration du dossier patient informatisé et l'identification automatique de la maladie tumorale

Dans cette partie, nous décrivons le développement, l'implémentation et l'évaluation d'une architecture pour l'intégration des données du dossier patient informatisé et l'identification automatique de la maladie tumorale.

Méthodes

Cette architecture repose sur trois briques informatiques indépendantes :
- La couche d'intégration syntaxique et de stockage qui repose sur i2b2.
- La couche de gestion des métadonnées et d'intégration sémantique.
- Un algorithme d'identification de maladie tumorale à partir des diagnostics.

Intégration syntaxique et stockage

Dans le domaine biomédical, les entrepôts de données sont majoritairement utilisés comme solution pour l'intégration de données hétérogènes [10]. Dans ce contexte, i2b2 est une solution reconnue et largement adoptée pour l'utilisation secondaire des données hospitalières, [21, 22] mais aussi pour l'intégration de données issues de différents projets de recherche [7, 10].

Nous avons mis en œuvre i2b2 au sein du Centre Hospitalier Universitaire de Bordeaux, ainsi qu'un processus d'extraction transformation et chargement (ETL) pour trois dimensions du système d'information :
- Programme de Médicalisation du Système d'Information (PMSI)
- Dossier Patient Informatisé (DPI)
- Anatomie et Cytologie Pathologiques (ACP)

Gestion des métadonnées et intégration sémantique

Cette couche constitue le cœur du système. Elle vise à stocker les informations permettant de décrire la sémantique des données enregistrées dans l'entrepôt de données. L'objectif principal de cette couche est de faire le lien entre les valeurs enregistrées dans i2b2 d'une part, leur origine dans le SIH, et d'autre part la sémantique qui peut leur être rattachée.

La gestion des métadonnées se fait au sein d'un triple store et les métadonnées sont enregistrées au format RDF. Nous avons mis en œuvre une infrastructure basée sur 3 standards :

— Ontology Web language (OWL) : langage permettant de décrire les ontologies formelles.
— Registre de métadonnées (ISO/IEC 11179) qui permettent de décrire les données, leur format et leur lien vers les concepts qui les représentent. Nous avons utilisé sa représentation en OWL mis à disposition par le projet SALUS, et sur lequel repose le semanticMDR [19].
— Simple Knowledge Organization System : langage basé sur OWL qui permet de décrire les terminologies.

A partir de ces 3 standards, nous avons construit un modèle qui permet de décrire les éléments de données et leur lien avec des terminologies, lorsque celui-ci existe. Enfin, nous avons mis en œuvre le modèle décrit plus haut permettant de décrire les règles IACR au-dessus des terminologies diagnostiques.

Identification de la maladie tumorale à partir des diagnostics

En s'appuyant sur la description de la couche d'intégration sémantique, nous avons construit un algorithme qui permet d'identifier la maladie tumorale à partir des diagnostics d'un patient. Cet algorithme utilise la hiérarchie des maladies tumorales construites par raisonnement, pour rattacher des diagnostics imprécis avec des maladies plus précises en l'absence d'ambiguïté.

Evaluation

Architecture

Nous avons évalué la capacité de l'architecture à intégrer des données issues du dossier patient informatisé du CHU de Bordeaux. Les patients inclus étaient l'ensemble des patients présentant un code diagnostic de tumeur (bénigne ou maligne) entre 2000 et 2016. Pour ces patients, nous avons extrait l'ensemble des données de trois dimensions du système d'information du CHU de Bordeaux :

— PMSI
— Anatomie et Cytologie Pathologiques
— Dossier patient (formulaire clinique)

Ces données ont été représentées au sein de la brique sémantique (gestion des métadonnées et des connaissances).

Identification des tumeurs

A partir des données intégrées, nous avons exécuté puis évalué l'algorithme d'identification de la maladie tumorale. Cette évaluation a été réalisée en s'appuyant sur les données du Registre Général des Tumeurs de la Gironde. Nous avons extrait du registre les tumeurs incidentes de 2013, chez les patients ayant eu une visite au CHU de Bordeaux. Ces tumeurs ont été comparées aux tumeurs de 2013 identifiées automatiquement par l'algorithme à partir des données de prise en charge du CHU de Bordeaux.

Résultat

Architecture

Au total, 95 969 patients ont été intégrés. Ces patients totalisaient 418 163 venues au CHU de Bordeaux. Ces patients correspondaient à 12 536 256 enregistrements dans l'EDC. Ces enregistrements se répartissaient en 8 471 130 observations (44,9% PMSI, 1,2% ACP et 53,8% formulaires cliniques) et 4 065 126 modificateurs. Le nombre moyen d'observations par patient s'élevait à 88,27.

Au total, 1 552 237 triplets ont été construits pour l'implémentation de la brique sémantique et la représentation des données intégrées. Ces triplets décrivaient :

— 578 *dataElements* (99,5% pour la description des formulaires cliniques).
— 578 *valueDomains* dont la moitié était énuméré (51,6%).
— 29 617 *permissibleValues*.

Nous avons intégré 8 terminologies correspondant à 47 109 *skos:concept*. Parmi les termes de ces terminologies, 16 066 étaient utilisés comme des *ValueMeaning* (signifiant qu'ils étaient alignés sur des jeux de valeurs enregistrés dans l'EDC).

Les données de la brique sémantique ont été utilisées pour construire des *ontologies i2b2* pour les 3 dimensions du système d'information intégrées. La majorité des nœuds construits correspondaient à des données structurées. Cependant, les formulaires cliniques incluaient à la fois des données structurées et des données en texte libre. Les hiérarchies construites pour cette dimension incluaient l'ensemble des types de données gérés au sein de i2b2 (Large String, String, valeurs numériques et données structurées). Ainsi, les *DataElements* utilisés pour construire les hiérarchies à partir de la brique sémantique couvraient un large champs de type de données et un grand nombre de nœuds. Le processus de construction automatique des hiérarchies permettait donc d'interroger l'ensemble des types de données disponibles au sein de l'EDC.

Identification des tumeurs

Au total 51 572 patients ont été identifiés à partir de l'EDC comme ayant au moins une tumeur codée entre 2012 et 2014. Parmi ces patients, 18 120 ont été identifiés par l'algorithme comme ayant une tumeur maligne survenue au cours de l'année 2013.

A partir du Registre Général des Cancers de la Gironde, 2 589 ont été extraits correspondant à 2 618 tumeurs. Au final,2 436 patients ont été retrouvés au sein des deux jeux de données correspondant à 3 610issues de l'EDC et 2 465 issues du registre des cancers.

En combinant ces jeux de données, un jeu d'évaluation a été construit, prenant en compte l'ensemble des patients issus du registre des cancers et uniquement les patients issus de l'EDC identifiés au sein du registre. Ce jeu d'évaluation incluait les 2 589 patients du registre des cancers (soit 2 618 tumeurs issues du registre des cancers et 3 610 issues de l'EDC).

L'algorithme basé sur la modélisation des règles IACR était capable d'identifier les topographies et morphologies tumorales avec une F-mesure respective de 0,76 et 0,68 . La F-mesure pour l'identification de la maladie complète (combinant topographie et morphologie) s'élevait à 0,53 avec une variabilité en fonction de la tumeur à découvrir (e.g. 0,91 et 0,59 pour l'adénocarcinome de la prostate et du sein respectivement).

Conclusion

Nous avons proposé une architecture complète, combinant des technologies et standards existants. Cette architecture permet la gestion des aspects liés à :

— L'intégration syntaxique,
— L'intégration sémantique,
— L'intégration de connaissances et de données,
— L'identification de la maladie à partir des diagnostics.

L'approche basée sur une ontologie pour l'identification de la maladie tumorale, permet une adaptation rapide des règles d'agrégation en fonction des besoins spécifiques d'identification. Ainsi, plusieurs versions du modèle d'identification peuvent être utilisées avec des granularités différentes.

Dans le cadre de ce travail nous nous sommes intéressé aux aspects liés à la sémantique des données pour identifier un événement sous-jacent (la maladie tumorale) et rattacher des observations.

Des travaux antérieur se sont intéressés au phénotypage [23, 24] et des approches plus spécifiques à la cancérologie cancérologie [25, 26] ont été proposées. Notre travail est complémentaire de ces travaux antérieurs mis en œuvre pour l'exploitation secondaire des données biomédicales. En effet, l'intégration sémantique ainsi que l'ajout de connaissances sous la forme d'ontologies permettent d'identifier des liens entre les observations et les événements qu'ils

décrivent. Cette structuration des données autours de la maladie est une étape essentielle dans l'acquisition de données pour la recherche. Des algorithmes de phénotypage manuels ou par apprentissage automatique pourraient exploiter ces liens pour optimiser l'identification de situations cliniques recherchées.

Introduction

Chapitre 1

Background

1.1 Secondary use of electronic health record

With the increasing adoption of electronic health records (EHRs), the amount of data produced at the patient bedside is rapidly increasing. These data provide new perspectives to : create and disseminate new knowledge ; consider the implementation of personalized medicine ; offer to patients the opportunity to be involved in the management of their own medical data [2]. Since 2007, the American Medical Informatics Association emphasized the value of secondary use of medical data : *"Secondary use of health data can enhance healthcare experiences for individuals, expand knowledge about disease and appropriate treatments, strengthen understanding about the effectiveness and efficiency of our healthcare systems, support public health and security goals, and aid businesses in meeting the needs of their customers"* [4]. Indeed, secondary use of biomedical data produced throughout patient care is an essential issue [1] and is the subject of numerous studies since several years [1, 2, 3, 4, 5, 6, 7]. Although centers such as Harvard and Vanderbilt have raised opportunities for researcher to settle platforms for secondary use of EHR [8, 9], its adoption and implementation within health care remain slow due to its complexity [10]. Secondary use of EHRs is not a straightforward task. Indeed, when reusing data in EHRs, multiple challenges are to be addressed including semantic and syntactic data normalization, phenotype identification and data quality evaluation [11].

Data integration (semantic and syntactic) aims at providing comparable and consistent data from a broad range of heterogeneous (both syntactically and semantically) data. Phenotyping for Hripcsak et al. *"transforms the raw EHR data into clinically relevant features"* [23]. Even if clinical care are now providing a large amount of data, using these raw EHR for phenotyping purpose remains challenging. Indeed, when reusing EHRs, one will have to deal with multiple issues : completeness, complexity and bias that limit feasibility [23]. These issues have to be taken into account, and data produced based on

EHRs need to be evaluated [27].

1.2 Secondary use in the oncology field

The oncology field is not spared by these issues. Cancer registries have the task of exhaustively recording incident cases of cancer in a given territory. In order to be able to register cases, cancer registries staff, must be informed that possible new cases should be registered. This process, called notification, was historically based on voluntary practitioners that declared all new cases they encountered. As early as 1998, an IARC technical report was drawn up describing the methods used by different registries for the establishment of automated notification procedures [12]. To ensure adequate cover of a population of several million people, at a very early stage the Ontario registry was obliged to develop methods for data acquisition [13]. Both notification and record tasks in this registry were automated with no clinical interventions in routine processing.

The job of cancer registries extends well beyond the mere recording of incident cases. To enable the registries to make full use of their expertise and research function in the area of cancer epidemiology, the optimization of registration procedures for incident cases of cancer is crucial, and has been recalled in the two French national cancer plans. The implementation of automated and semi-automated procedures to assist in detecting and documenting incident cases of cancer based on EHR is therefore an attractive approach in this setting.

In the context of personalized medicine, integrating EHRs with biobanks data and biological data (genomics, proteomics etc...) is a major issue for clinical decision support and translational research [15, 16]. In france in 2011 and 2012 l'*Institut national du cancer* has founded bio-clinic databases and networks. The main goal of these networks are to enable transverse analysis of patient data [28] including clinical and biological data. These databases are built by reviews of patient charts and manual data entry. These database may benefit from being enriched with EHRs based clinical data.

Olive et al. have presented a critical analysis of French hospital discharge data for the epidemiology of cancers. In particular, they noted difficulties relating to the use of data in isolation to detect incident cases [29]. This finding underlines the importance of using diverse sources for the notification of new incident cases.

The integration of multiple data sources data from information systems that are structured around the patient makes it possible to optimize automatic processing for the identification of incident cases, and the recording of complementary data [30]. The use of this accumulated information for case notification is a logical strategy in a perspective of exhaustiveness. This view has

led registries to increasingly diversify notification sources. The mean number of notification sources has thus become a criterion for exhaustiveness, and a percentage of histological confirmation of cases that is too high should lead to suspicion of non-exhaustiveness [31]. The price to pay for this approach is an excess notification of false incident cases following coding inaccuracies by the different data sources. These false cases require manual processing to be removed. Methods have been proposed in order to take into account existing noise within the data [25] and to optimize incident cases detection using machine learning [32].

Thereby, secondary use of EHR is attractive in he oncology field, and has been subject to active research since several years. The common goal of these previous work is to identify clinically relevant features (namely the occurring tumoral disease). This task falls in the scope of phenotyping as defined by Hrispcsak et al [23]. Multiple methods have been proposed for phenotyping purpose ranging from manually defined algorithm [33] to unsupervised phenotype identification learned from the data [24]. All these approach are focusing building algorithm (manually defined or learned from data) on the top raw EHR. However EHRs are structured for care purpose and not for secondary use. EHRs structure and its semantic need to studied in the light of needs for secondary use in oncology in order to identify issues still preventing efficient secondary use of EHRs in the oncology field.

1.3 Objectives

In this chapter, we analyze structure and semantic of EHRs in the light of the needs for secondary use in oncology. We focus our work on diagnosis recorded within structured EHRs. The document is structured as follow :

- Discussion of two main issues preventing secondary use of structured EHRs in oncology (namely terminologies heterogeneity and implicit disease representation within EHR).
- Identification of three aspects that must be treated in order to address these issues (namely syntactic integration, semantic integration, and data and knowledge integration)
- Presentation of a set of state of art technologies and standards that exists and their implementation in the context of the three aspects to be treated.
- Proposition of an architecture that could facilitate secondary use of EHRs in oncology by addressing the two main issues identified.

Chapitre 2

Issues for secondary use of structured EHRs in oncology

In the remainder of this document, we will use definitions proposed by Scheuermann et al. in [17] for diagnosis and disease :

— "***Diagnosis =def.*** *– A conclusion of an interpretive process that has as input a clinical picture of a given patient and as output an assertion to the effect that the patient has a disease of such and such a type.*

A diagnosis is a continuant entity that, once made, will survive through time, and is often supplanted by further diagnoses. The diagnostic process is thus iterative : the clinician is forming hypotheses during history taking, testing these during physical exam, forming new hypotheses as a result, and so on. "

— "***Disease =def.*** *– A disposition (i) to undergo pathological processes that (ii) exists in an organism because of one or more disorders in that organism.* "

We will first introduce diagnosis representation within structured EHR. Base on this description we introduce two issues in the light of the needs for secondary use :

— Implicit disease description in EHRs.

— Diagnosis terminologies heterogeneity in oncology.

2.1 Diagnosis representation within structured EHRs

All along patient care, data are collected by clinicians (i.e clinical notes, billing codes etc.) within EHRs.These data are recorded with multiple goals : sharing information about patients (i.e. retained diagnosis, procedures and treatments) ; enabling reimbursement. . . EHR can range from free-text to struc-

tured and coded data. Regardless of this level of structure, EHR are composed of observations. For instance clinical notes are free text documents reporting reason for hospitalization, final diagnosis, patient history, comorbidities, family history and so on.

We will now focus on diagnosis recording within structured EHR.

2.1.1 Structured and coded diagnosis available in EHRs

In france, standards exists at the national level for recording clinical information about tumoral disease. This standards that urge physician to record diagnosis in a structured way lead to a large amount of structured data produced at the patient bedside. Three major hospital activities are producing structured data using various terminologies to describe diagnosis :

- Anatomical pathology (AP) data that includes in addition to free-text reports one or several diagnostic codes usually recorded ADICAP (Association pour le Développement de l'Informatique en Cytologie et en Anatomie Pathologique – French classification of lesions with topographical and histological axis) or ICD-O-3 (International Classification of Diseases for Oncology) [34].
- Hospital discharge (HD) data recorded in the French medical information program that includes ICD-10 diagnostic codes and CCAM medical procedure coded fields (Classification Commune des Actes Médicaux – the health insurance classification).
- Multidisciplinary staff (MS). Each patient presumed or stated to have a cancer must be discussed in a multidisciplinary board in order to make a multidisciplinary decision. A report is produced for each discussed cases. The report is standardized at the national level and include diagnostic codes recorded with ICD-O-3.

However, these structured EHRs remain underused for the aforementioned purpose in the oncology field. In what follow, we discuss two issues that may explain why using these structured data remains challenging namely : (i)heterogeneity of diagnosis terminologies leading to semantic integration issues ; (ii)implicit disease description within EHRs explicitly needed for secondary use purpose.

2.1.2 EHR data organization

Figure 2.1 presents an example of data recorded within, HD, AP and MS for a treated breast cancer. Each time a patient comes to hospital a record will be produced for HD data. This record will contain, among other things, one or more diagnosis code and, if necessary, procedures codes. In this example we can see that the patient has had a biopsy which was analyzed by a pathologist and a diagnosis code was recorded in PA. A decision was then made to treat the

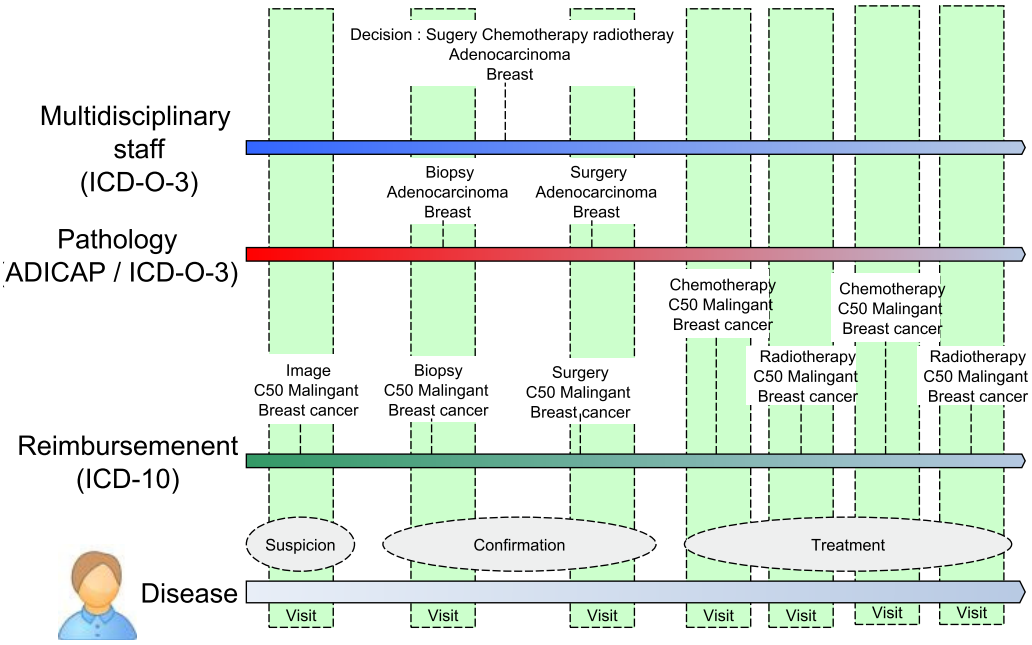

FIGURE 2.1 – Temporal representation of recorded diagnosis in structured EHR. Example for a treated breast cancer.

patient with surgery, chemotherapy and radiotherapy. These information were recorded in the MS report. The surgery was performed and reported in HD data and after being analyzed by a pathologist, a diagnosis was reported in PA. The patient was treated in hospital so that chemotherapy and radiotherapy were reported in HD data.

Based on this example, we can notice two important facts about data structure in EHR's :

— Even in a unique structure, multiple terminologies are used in order to report the same information. Diagnosis are recorded with ICD-10, ADICAP and ICD-O-3 depending the data source.

— There is no link between the records (except that they are related to the same visit). For instance, the link between the biopsy and the result of pathologist analysis is not explicit within the data. The underlying conceptual model is presented figure 2.2. Moreover, the disease itself is not explicitly recorded (only diagnosis referring to the disease are recorded independently).

2.2 Implicit disease description within EHRs

In oncology, research databases (such as cancer registries and bio-clinical databases) are recording data about specific diseases depending on their own

FIGURE 2.2 – Electronic health record model. Every information is related to a visit. Observation are not directly related

FIGURE 2.3 – Disease centered model. This model is availble within research databades such as cancer registries or bio-clinic databases. It enables to select patients depending on their disease.

goals. For instance cancer registry may record only solid tumors or concentrate on liquid tumors. Bio-clinical databases are specific to an organ (such as liver or kidney) or a type of cancer (such as sarcomas [35]). As a result disease assessment is central in the recording process in order to meet inclusion criteria.

Main goal of cancer registries is to provide epidemiological data (such as cancer incidence) about cancer. These data are often presented depending on tumor site [36]. Bio-clinical databases (and usually cancer registries) aims at enabling phenotype identification in order to build ancillary studies (often needing supplementary data collection for identified patient). In this process patient identification mainly rely on the assessment of a disease and its course. For instance a typical phenotype identification problem in oncology would be : *«Retrieve every patient in my health information system with a metastatic colon adenocarcinoma that has not been excised»*. When analyzing this phenotype query, we can identify that we need to :

1. Describe patient with a specific disease (colon-adenocarcinoma).
2. Find events related to the disease's clinical course (metastasis)
3. Know if the disease was treated by a surgical procedure.

As a result, disease is a key feature for this kind of query and observation need to be related to it.

In order to enable (i) disease assessment for inclusion criteria validation and epidemiological data production ; (ii) phenotype query based on disease, its course and treatments ; it is necessary to record the disease (described by the retained diagnosis) and to relate observations to the disease. The underlying conceptual model needed presented in figure 2.3. Thereby research data are produced by analyzing patient chart (usually manually). This analyze of patient chart aims at identifying the disease based on *phenotype* (as defined

in [17]) features. Because it is crucial for research databases, identifying and recording the disease is an unavoidable task.

On the other hand, as discussed in 2.1.2, EHRs record data as succession of independent observations (except that they belong to the same visit for a given patient). Moreover, the disease is rarely explicitly recorded. It is in fact represented by recorded *diagnosis* during the *diagnosis process*. Indeed a single occurring *disease* may be referenced by multiple distinct *diagnosis* depending on the *diagnosis process* and physician's conclusion at a given time of the process.

As a result, there is a mismatch between models underlying EHRs and models underlying research data in oncology. Hence building research data based on EHRs necessarily implies :

- Building a comprehensive view of EHR data.
- Interpreting unrelated observations in order to :
  - Identify tumors as diseases based on reported diagnosis
  - Explicitly relate observations to identified tumors

Even in unstructured hospital reports (i.e. letters, surgery report. . .), physician often let observation unrelated . Indeed, information exchange between physicians relies on shared medical knowledge not available within data. For instance when viewing a patient chart a physician, based on his knowledge, can infer that a treatment is for a specific disease. Fully understanding such patient chart without medical knowledge remains difficult (if possible) partly due to these missing relationships. As he features his own medical knowledge, a clinician can easily interpret EHRs as a clinical history for an occurring disease (except that observations may be split and difficult to retrieve [37]).

Providing methods for such tasks automation falls in the scope of *phenotyping*. As discussed above, multiple approaches ranging from expert consensus [38] to automated phenotype identification learn from data themselves [24] have been proposed. Even if different, these approach are implicitly trying to add medical knowledge to data so that they can be interpreted. This knowledge is contained in the proposed algorithm of the eMERGE and extracted from the data themselves when using machine learning approaches.

However, when extracted from data, even simple knowledge can remain hard to acquire. For instance in [24] authors report that *"[...]there are several diagnosis codes which are highly clinically relevant with each other, and yet do not get coded together in patient records : different stages of pancreatic cancer for example, would make sense in a single phenotype for the disease, but will not be seen jointly over many patients at a time"* as a possible explanation for poor correlation between normalized pointwise mutual information (an automated coherence metric) and human judgment about relevance of obtained phenotype.

As a conclusion, our hypothesis is that medical knowledge is not always available within data and remains one of the key features that enables EHR

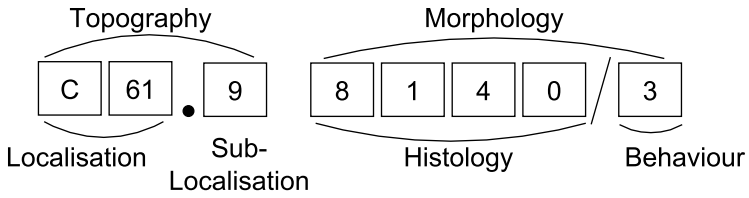

FIGURE 2.4 – Structure of the ICD-O-3 code (with example of a prostate adenocarcinoma)

interpretation. It is necessary to (i) provide machine readable knowledge resources ; (ii) link these knowledge resources with EHR data elements in order to facilitate both manual and automated phenotyping approaches.

2.3 Diagnosis terminologies heterogeneity in oncology

2.3.1 Terminologies characteristics

ICD-O-3

ICD-O-3 is a multi-axial classification used in cancer registries in order to record the anatomic site (topography) and the morphology of a neoplasm. Figure 2.4 presents the structure of ICD-O-3 code. The morphology is coded with five digits. The first four digits represent the histological description and the fifth digit indicates the behavior (i.e. whether benign or malignant) of a neoplasm. As a result, it is not possible for a morphology to have multiple behaviors. *"The topography code indicates the site of origin of a neoplasm ; in other words, where the tumor arose"* [34]. From the ICD-O-3 "point of view", any morphology code can be associated with any topography code. Some tumor morphologies have a *"usual primary site"* but it is expressly stated that these associations are provided only to help coders and should not be considered as systematic (and unique) topography-morphology combinations. An example is given in (14) : *"An unusual, but possible, example would be the diagnosis 'osteo-sarcoma of kidney', for which the kidney topography code (C64.9) would be used instead of 'bone, NOS' (C41.9) [...]"*. Thus, ICD-O-3 describes a disease by combining the morphology of the tumor and the topography from where the tumor arose. As a result each neoplastic disease is not described as a whole concept entailed by a unique code within ICD-O-3.

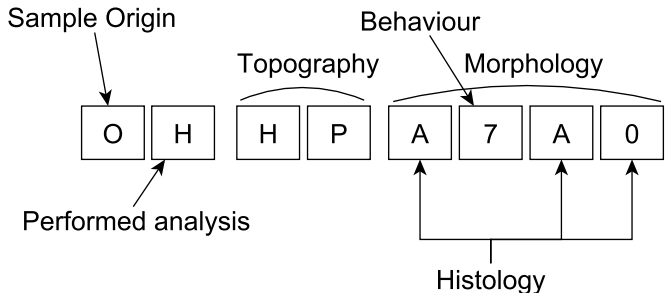

FIGURE 2.5 – Structure of the ADCIAP code (with example of a prostate adenocarcinoma identified in an histological analysis on a surgical sample)

ICD-10

Within ICD-10, the chapter 2 regroups neoplasms. It is divided into four axes depending on the behavior of the tumor (namely *Malignant neoplasms*, *In situ neoplasms*, *Benign neoplasms* and *Neoplasms of uncertain or unknown behavior*). Within the *Malignant neoplasms* block, ICD-10 categories differentiate primary tumors from metastatic secondary tumors. In the same way as for ICD-O-3, a neoplasm cannot have multiple behaviors. ICD-10 describes each neoplastic disease as a whole concept entailed by a unique code. For instance, C50.2 : *Malignant neoplasm upper-inner quadrant of breast* describes two characteristics of the cancer disease :

— The behavior (*Malignant*) which is part of the morphology description.
— The site of origin (*upper-inner quadrant of breast*) which corresponds to the topography.

ADICAP

ADICAP is a multi-axial french terminology. It is widely used by french pathologist in order to standardize pathology report's annotations. Figures 2.5 presents the multi-axial structure of ADICAP codes. Although the complete code can reach 15 digits, in most case, only the required first 8 digits are recorded by pathologist. Thus we describe only structure of these 8 first digit :

— First digit corresponds to the sample origin (i.e. : surgical excision, biopsy . . .).
— Second digit corresponds to the performed analysis (i.e. : histology, immunohistochemistry . . .)
— Third and fourth corresponds to the topography of origin of the sample.
— Fifth to eighth digits corresponds to the morphology. When the diagnosis is a neoplasm, the morphology code can be divided as follow :
  — Sixth digit (second of the morphology part) corresponds to behaviour.

— Fifth seventh and eighth digits corresponds to the histological description.

The complete codes is used as an annotation of a pathology report the code is to be interpreted as a whole describing the conclusion of the pathologist made over a sample using a specific analysis method. The relationship between the topography and the morphology part is slightly different from ICD-O-3's. Indeed, ADICAP describes the finding site of the described morphology whereas ICD-O-3 describes the site of origin of the tumor.

2.3.2 Semantic integration issues

Figure 2.6 presents how a prostate adenocarinoma can be recorded using ICD-10 and ICD-O-3. Two issues can be identified when integrating these terminologies :

— **Pre-coordination versus post-coordination**. On the one hand ICD-10 is a mono-axial terminology. The diagnosis is represented as whole "pre-coordinated" concept. It entails the whole information within one code (the primary site and the behavior of the tumor). On the other hand, ICD-O-3 is a multi-axial terminology. The diagnosis is represented as the combination two distinct codes recording independently the primary site and the morphology of the tumor. The diagnosis is deduced by association of the two codes. The assocation task is called composition. Within ICD-O-3, the complete diagnosis (combining topography and morphology is not available. As a result it is not possible two map a ICD-O-3 code (a topography or a morphology) with an ICD-10 code. There is a need for methods enabling automatic composition in order to build diagnosis based on ICD-O-3 code combinations that can be mapped to ICD-10 codes.

— **Granularity heterogeneity**. Using ICD-10, it is not possible to record the complete information about the morphology of the tumor. As a specialized terminology, ICD-O-3 provide precise codes for morphology thus enabling a finer grained recorded information. However, the ICD-O-3 composed diagnosis and the ICD-10 diagnosis may refer to the same disease for a patient within EHR. Hence it is necessary to link them.

This example shows that relationships between these terminologies are complex. Translating data from one terminology two another is not straightforward and may involve complex relationships and mechanisms. There is a need for solution enabling an integrated view of diagnosis data no matter the terminology.

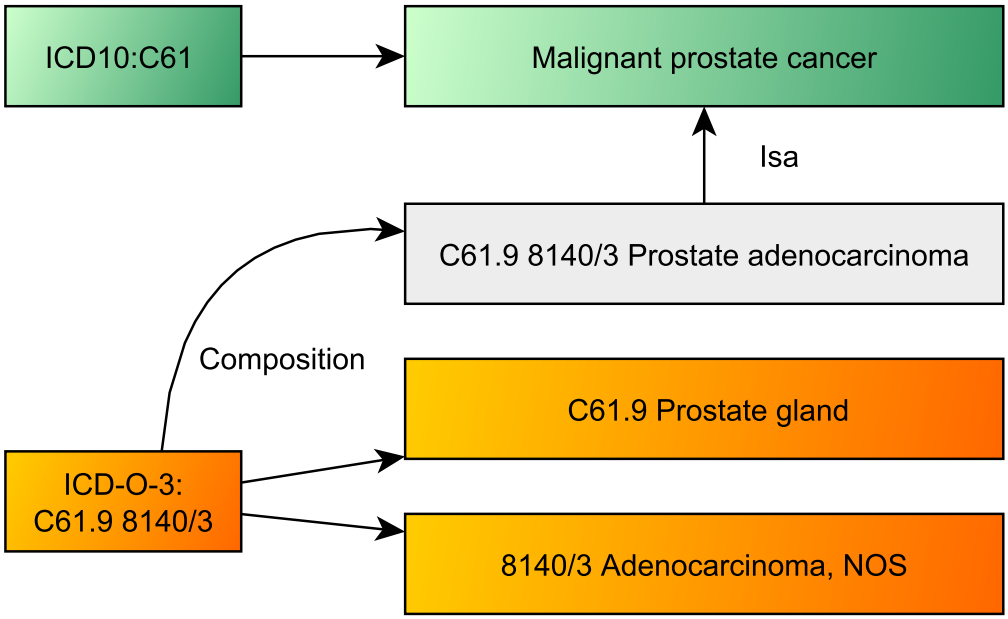

FIGURE 2.6 – Diagnosos terminologies heterogeneity. Example of ICD-O-3 and ICD-10 for recording a prostate adenocarcinoma.

Chapitre 3

Integrating EHR with external knowledge resources for disease identification in oncology

We have identified two major issues to be addressed in order to adapt data for secondary use in oncology. In order to provide methods for addressing this issues, three aspects have to be treated :

- **Syntactic integration**. In order to enable an integrated access to heterogeneous EHRs data.
- **Semantic integration**. In order to take into account heterogeneous terminologies used to represent data.
- **Data and external knowledge integration**. In order to bind external medical knowledge to raw EHRs.

3.1 Existing standard and tools

In this part we present, existing state of art standards and technologies that may be leveraged in order to treat these three aspects.

3.1.1 Syntactic integration : Clinical data warehouse

Syntactic integration is a precondition for secondary use of EHR. Indeed, it is noteworthy that a phenotype algorithm built on the top of heterogeneous system need at least to be aware of how data can be accessed. In the bio-medical domain, clinical data warehouse (CDW) are now largely adopted as a solution for EHR integration [10]. Integrating Biology and the Bedside (i2b2) an NIH-funded National Center for Biomedical Computing at Partners Healthcare System in Boston [39]. It has developed an open source CDW infrastructure [40] for EHR integration and secondary use. I2b2 has been widely adopted by academic hospital [21, 22] and research project [7, 10].

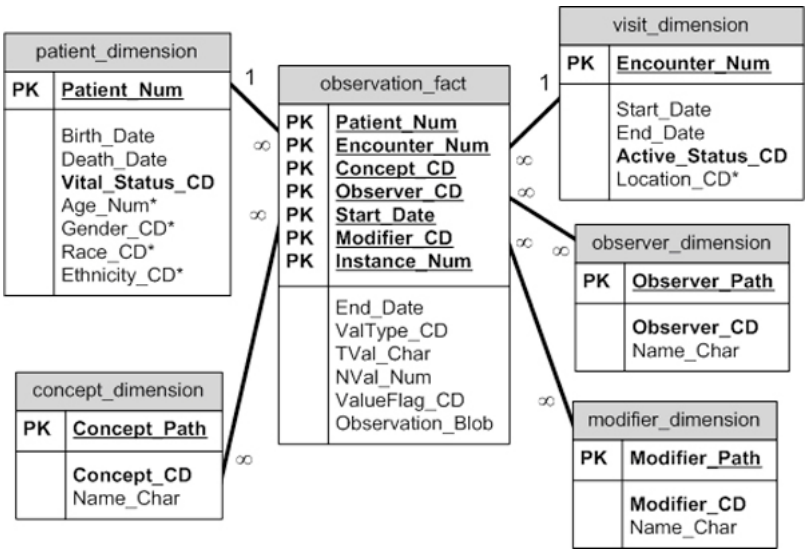

FIGURE 3.1 – i2b2 Start schema

I2b2 infrastructure (so called the i2b2 Hive) consists of Cells providing services. Within the i2b2 Hive, i2b2 core Cells correspond to a minimal set providing basic services for the Clinical Research Chart. The Data Repository Cell (called Clinical Research Chart or CRC) is one of the core Cells. CRC is in charge of storing *"medical and medically oriented genomic data"* [40]. This Cell is built on the top of an Entity-Attribute-Value data model (Figure 3.1), enabling integration of various type of clinical data. Tools are also available in order to build and execute query in a "user friendly environment".

Querying this data, is a task that falls within the scope of phenotyping. It aims at identifying clinical situation (such as disease and so on) based on observation available in integrated EHR. However as the main goal of i2b2 is to enable EHR integration, its data model exposes limitations for relating observations. Observations are related to a patient, a concept (representing the meaning of an observation), a visit, a provider (representing the origin of an observation) and a modifier (which can be used to add complementary information to a given observation). Even when integrated within an i2b2 warehouse, clinical situation identification can lead to complex query because implicit information remain unavailable.

The "Ontology" Cell is in charge of managing meaning of observations. As meaning is separated from data, it is possible to access a single observation multiple way within i2b2 (for instance with multiple aggregation rules). It can provide convenient services to help semantic integration preserving the initial meaning of an observation. However, this Cell provide only hierarchical and mapping mechanism which is not sufficient for addressing aforementioned semantic heterogeneity issues.

3.1.2 Semantic integration : Semantic Web technologies and standards

Semantic Web aims at building and share content meaningful to computer as an extension of the World Wide Web [18]. Semantic Web technologies, are tackling challenges underlying exploitation of heterogeneous distributed data by enabling machines to understand and share content meaning [41]. A key issue in Semantic Web is about knowledge representation, storage and sharing. Indeed, sharing meaning is about representing it consistently. Semantic Web tackles this issue by providing common standardized formats to describe heterogeneous data and corresponding domain knowledge [42].

The World Wide Web consortium (W3C) provide and promotes standards and technologies for Semantic Web. In the Semantic Web, meaning is expressed by the Resource Description Framework (RDF). RDF, captures information in *statements*. *Statements* are triples comprehending three parts : *Subject*, *Predicate*, *Object* [43, 18, 42]. *"A set of such triples is called an RDF graph. An RDF graph can be visualized as a node and directed-arc diagram, in which each triple is represented as a node-arc-node link."* [43]. Nodes can be either IRIs (Internationalized Resource Identifier), or literals, blank nodes. A triple states a relationship between the two nodes. "*The predicate itself is an IRI and denotes a property, that is, a resource that can be thought of as a binary relation*") [43]. Based on RDF abstract syntax, W3C publish multiple standards for linked data serialization and publication and knowledge representation.

Web Ontology Language

The Web Ontology language (OWL) is a standardized language defined by the W3C for ontology specification [44, 45]. Gruber defines an ontology in the context of computer and information sciences as follow : "*an ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members)*".[46].

OWL documents (ontologies) are machine readable. OWL Reasoner are computer programs exploiting OWL logic based language in order to : check owl documents consistency and infer new relation between primitives ("*to make implicit knowledge explicit*" [45]).

Simple Knowledge Organization System

The Simple Knowledge Organization System (SKOS) is a "*common data model for sharing and linking knowledge organization systems via the Web*". It is a representation of common element *"shared by many knowledge organization systems such as thesauri, taxonomies, classification schemes and subject*

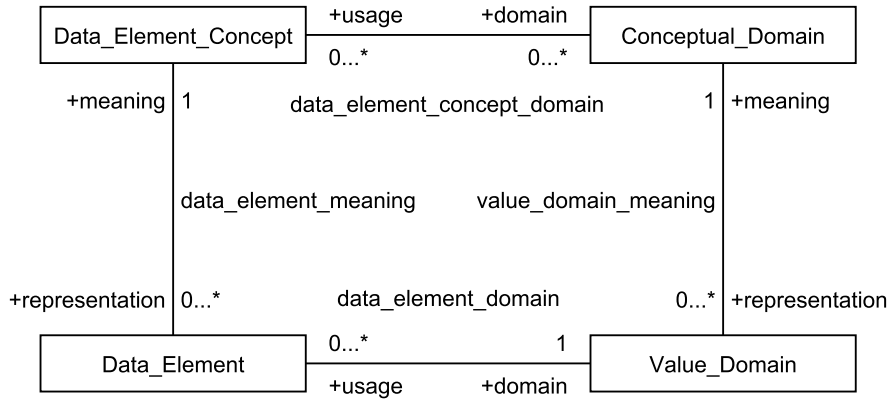

FIGURE 3.2 – ISO/IEC 11179 High-level Data Description metamodel [48]

heading systems". SKOS aims at enabling *"low-cost migration path for porting existing knowledge organization systems to the Semantic Web"* [47]. SKOS model is defined using OWL.

3.1.3 Integrating knowledge with data : Metadata registries

Metadata registries are centralized metadata storage system. They are used in order to maintain data consistency within an organization. They can be used for enabling data sharing and data warehousing by breaking down silos of information. A metadata registry standard is defined in ISO/IEC 11179 specification [48]. In this definition, a metadata registry stores information about *data elements*. *"Examples of data element include : a column in a table of a relational database, a field in a record or form, an XML element, the attribute of a Java class, or a variable in a program"* [48].

Figure 3.2 shows a high level view of the metadmodel specified in ISO/IEC 11179. This metamodel can be split in two parts. The upper part corresponds to *" a conceptual (semantic) level"*. The lower part corresponds to *"a representational level"*. A *data element* is a representation of a *data element concept* and a *data element* has a *value domain* which is a representation of the corresponding *conceptual domain*. Associations such as *data_ element_ meaning* and *value_ domain_ meaning* aims at linking the conceptual level and the representational level. Through this model, multiple *data element* represented within multiple heterogeneous system can be bind to a single *data element concept*. Hence, metadata registry can bring semantics on the top of a represented *data element*.

3.1.4 Semantic Web and Metadata registry usage in the bio-medical field

In the bio-medical domain, many project have been using semantic web technologies and standards for data secondary use. In [49] authors propose to leverage openEHR archetypes [50] as an EHR standard combined semantic web technologies (namely OWL representation and reasoning for inclusion criteria) for phenotyping purpose. Liaw et al. implemented an ontological approach to improve the accuracy of diabetes disease registers [51].

Metadata registries ISO/IEC 11179 standard has been used to build solution for metadata management [52, 53, 19, 54, 7]. A significant part of these project use Semantic Web technologies in conjunction with ISOIEC standard. Within the SALUS project, Semantic MDR is an ISO/IEC 11179 based metadata registry. It uses an OWL representation of ISO/IEC 11179 standard and triplestore technology as the backbone for metadata storage, and semantic services providing [19]. CDISC uses a a comparable approach for metadata management [54].

3.2 Solution to be investigated

Phenotyping implementation mainly depends on syntactic and semantic data integration. An implementation of phenotype algorithm is at least dependent of the way data are represented and stored. Data warehouse solution for syntactic integration have been widely used and have demonstrated benefits for secondary use. Semantic Web technologies have been increasingly used for semantic integration purpose in the biomedical domain. Semantic Web has also been leveraged in order to model inclusion criteria and classify patient base on their EHR. Moreover W3C semantic web standards offers tools for representing both formal models and structured terminologies. Modeling diagnosis and disease in oncology depending on topography and morphology is to be investigated in order to address diagnosis terminology integration issue. Metadata registries enable to record data representation and their link with concepts that represent them. ISO/IEC 11179 metamodel has a structure that allow joint management of data representation and knowledge ressources.

Combining these three components in a single infrastructure is to be investigated as a solution for automated adaptation of EHRs for secondary use in oncology. Indeed, when combined, CDW, Semantic Web technologies and metadata registries may provide solutions in order to (i) store and retrieve heterogeneous EHRs (ii) integrate heterogeneous diagnosis terminologies in oncology ; (iii) bring necessary knowledge on the top of EHR for implicit disease identification based on diagnosis available within EHR.

Our objective is to investigate how metadata registry, Semantic Web and data warehousing approach can be combined to settle tools for automatic can-

cer identification based on EHR.

We will first explore ontology based solutions for disease terminology integration in oncology. Then we will propose an architecture combining :

- Data warehousing based on i2b2 infrastructure for syntactic integration.
- Semantic Web technologies for external knowledge integration :
  - Terminology representation (using SKOS)
  - Heterogeneous terminology integration based on ontologies (using OWL).
  - Diagnosis and disease in oncology modeling (using OWL).
- Metadata registries for linking data representation with terminologies.

Model for diagnosis integration and disease identification

Chapitre 4

Building a model for disease classification integration in oncology. An approach based on the National Cancer Institute thesaurus

Abstract

Background

Identifying incident cancer cases within a population remains essential to enable research in oncology. Thus data produced within electronic health records should be used for this purpose. Due to the multiplicity of providers, heterogeneous terminologies are used for oncology diagnosis recording purpose. To enable disease identification base on these diagnosis, there is a need for integrating disease classification in oncology. Our aim was to build a model integrating concepts involved in two disease classification, namely ICD-10 (diagnosis) and ICD-O-3 (topography and morphology combinations), despite their structural heterogeneity. Based on the NCIt, a "derivative" model for linking diagnosis and topography-morphology combinations was defined and built. ICD-O-3 and ICD-10 codes were then instantiated in the "derivative" model. Links between terminologies obtained through the model were then compared to mappings provided by the SEER Program.

Results

The model integrated 98% of morphology codes (excluding metastasis), 68% of topography codes and 42% of neoplasm ICD-10 codes (excluding metas-

tasis). When codes were integrated, all SEER mappings were related through the model.

Conclusions

We have proposed a method to automatically build a model for integrating ICD-10 and ICD-O-3 based on the NCIt. The resulting "derivative" model is a machine understandable resource that enables an integrated view of these heterogeneous terminologies. The NCIt structure and the available relationships can help to bridge disease classification taking into account their structural and granular heterogeneity. However, (i) inconsistencies exist within the NCIt leading to misclassifications in the "derivative" model, (ii) the "derivative" model only integrates a part of ICD-10 and ICD-O-3. The NCIt is not sufficient for integration purpose and further work based on other termino-ontological resources is needed in order to enrich the model and avoid identified inconsistencies.

abstract

4.1 Background

With the increasing adoption of electronic health records (EHRs), the amount of data produced at the patient bedside is rapidly increasing. These data provide new perspectives to : create and disseminate new knowledge ; consider the implementation of personalized medicine ; offer to patients the opportunity to be involved in the management of their own medical data [2]. Indeed, secondary use of biomedical data produced throughout patient care is an essential issue [1] and is the subject of numerous studies since several years [1, 2, 3, 4, 5, 6]. Since 2007, the American Medical Informatics Association emphasized the value of secondary use of medical data : *"Secondary use of health data can enhance healthcare experiences for individuals, expand knowledge about disease and appropriate treatments, strengthen understanding about the effectiveness and efficiency of our healthcare systems, support public health and security goals, and aid businesses in meeting the needs of their customers"* [4].

In the oncology field, it is necessary to identify and describe incident cancer cases within a population in order to facilitate research and public health monitoring. For instance, cancer registries have to exhaustively record incident cases of cancer in a given territory and this task remains time consuming if it is performed manually. As early as 1998, a technical report was drawn up by the International Agency for Research on Cancer describing the methods used by different registries for establishing automated procedures to identify new cases using available data [12]. Methods have been proposed for automatically

identifying and registering cancers using structured data indexed with standard terminologies [14, 32, 55, 30, 26].

However, multiple actors with many different medical specialties are providing information in EHRs. As a result, within EHRs, data describing diseases are recorded according to multiple heterogeneous terminologies even for a single disease happening to a single patient. For instance, in France, reimbursement data use the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10) [56] to describe diseases, whereas pathology data use either ADICAP (a French pathology terminology) or the 3rd edition of the International Classification of Diseases for Oncology (ICD-O-3) [34] and data from multidisciplinary meetings in oncology use ICD-O-3. Providing an integrated access to these disease classification may improve automated cancer identification.

Although ICD-10 and ICD-O-3 both describe cancer diseases, they exhibit differences in terms of structure and granularity. Thus, it is necessary to identify or to build a resource that allows to integrate cancer disease classification taking into account these heterogeneities. To achieve this goal, composite relations must be defined between the involved concepts, such as "a neoplasm is a disease and has a specified morphology, as well as a specified topography". The National Cancer Institute thesaurus (NCIt) *"provides reference terminology covering vocabulary for clinical care, translational and basic research, and public information activities"*(cited from http://ncit.nci.nih.gov/, visited 2015-01-22) . It is described as *"a controlled terminology which exhibits ontology-like properties in its construction and use"* [57]. These characteristics *"open up the possibility [...] in linking together heterogeneous resources created by institutions external to the NCI"* [58]. Thus, the NCIt could be used as a resource to bridge the gap between disease classification, which are structurally heterogeneous.

However, since 2005, it has been shown in many occasions that the NCIt remains flawed [58, 20, 59] and especially that logic-based reasoning over the NCIt should be used cautiously. However, re-building a model "from scratch" would be time consuming and comes with no guaranty of avoiding inconsistencies. Despite of the limitations described above, the NCIt contains knowledge that could be useful for our integration purpose. In this manuscript we propose an approach to build a resource based on a subset of the NCIt, linking the three axes that refer to diseases as described in ICD-10 and ICD-O-3, i.e., the disease itself as well as its morphology and topography.

4.1.1 ICD-O-3

ICD-O-3 is a multi-axial classification used in cancer registries in order to record the anatomic site (topography) and the morphology of a neoplasm. The morphology is coded with five digits. The first four digits represent the

histological description and the fifth digit indicates the behavior (i.e. whether benign or malignant) of a neoplasm. As a result, it is not possible for a morphology to have multiple behaviors. *"The topography code indicates the site of origin of a neoplasm ; in other words, where the tumor arose"* [34]. From the ICD-O-3 "point of view", any morphology code can be associated with any topography code. Some tumor morphologies have a *"usual primary site"* but it is expressly stated that these associations are provided only to help coders and should not be considered as systematic (and unique) topography-morphology combinations. An example is given in (14) : *"An unusual, but possible, example would be the diagnosis 'osteo-sarcoma of kidney', for which the kidney topography code (C64.9) would be used instead of 'bone, NOS' (C41.9) [...]"*. Thus, ICD-O-3 describes a disease by combining the morphology of the tumor and the topography from where the tumor arose. As a result each neoplastic disease is not described as a whole concept entailed by a unique code within ICD-O-3.

4.1.2 ICD-10

Within ICD-10, the chapter 2 regroups neoplasms. It is divided into four axes depending on the behavior of the tumor (namely *Malignant neoplasms*, *In situ neoplasms*, *Benign neoplasms* and *Neoplasms of uncertain or unknown behavior*). Within the *Malignant neoplasms* block, ICD-10 categories differentiate primary tumors from metastatic secondary tumors. In the same way as for ICD-O-3, a neoplasm cannot have multiple behaviors. ICD-10 describes each neoplastic disease as a whole concept entailed by a unique code. For instance, C50.2 : *Malignant neoplasm upper-inner quadrant of breast* describes two characteristics of the cancer disease :

— The behavior (*Malignant*) which is part of the morphology description.
— The site of origin (*upper-inner quadrant of breast*) which corresponds to the topography.

4.1.3 Concepts involved in ICD-10 and/or ICD-O-3

Even if called "disease classification", ICD-10 and ICD-O-3 are in fact used within EHR, for recording diagnosis. The diagnosis, is a way for the physician to describe the disease but it is not the disease itself which corresponds to an evolving process. A single disease may have multiples diagnosis all along its clinical course (for instance an in situ neoplasm may evolve and become a malignant invasive neoplasm) but the disease (process) remains the same. Indeed, when used in this context, disease classification, are in fact kinds of diagnosis which can be viewed as representations of the disease at a given time. In the remaining part of the manuscript :

— *Diagnosis*, stands for the opinion at a given time and by a given person about the neoplastic disease.

— *Morphology*, stands for the opinion at a given time and by a given person about the morphology of the neoplastic disease.
— *Topography*, stands for the opinion at a given time and by a given person about the arising site of the neoplastic disease.

None of these concepts are representing the disease, or the actual morphology of the disease. They are representing reported information about the disease. Within ICD-10 and ICD-O-3, three different kinds of concepts are thus involved :

— The morphology of the tumor, which is a representation of the pathological description of the tumor reported at a given time. These concepts are available within the ICD-O-3 morphology axis.
— The topography of the tumor, which is a representation of the site of origin of the tumor reported at a given time. These concepts are available within the ICD-O-3 topography axis.
— The diagnosis, which is a is a representation if the reported description of tumor and entails information about both the topography and the morphology of the tumor. These concepts are available as such within ICD-10 and can be built by combining ICD-O-3 topographies and morphologies.

Because it is not possible to state that a *diagnosis* is equivalent to a *topography* or a *morphology*, it is obviously not possible to find equivalences between concepts represented within these two terminologies.

4.1.4 The National Cancer Institute thesaurus (NCIt)

"NCI Thesaurus (NCIt) is NCI's reference terminology. NCIt provides the concepts used in caCORE and caBIG to establish data semantics. It covers terminology for clinical care, translational and basic research, and public information and administrative activities. NCIt is also a widely recognized standard for biomedical coding and reference, used by a broad variety of public and private partners both nationally and internationally".

In the NCIt, topographies are described in the *Anatomic structure, system, or substance* axis. Morphologies and diagnosis are represented within the same hierarchy, subsumed by *Neoplasm*. Thus, no specific axis for tumor morphologies is defined and diseases are modeled as anatomic specializations of morphologies. For example, *Breast adenocarcinoma is_ a Adenocarcinoma* is stated in :

$$\text{Breast adenocarcinoma} \equiv \text{Adenocarcinoma} \cap \text{Breast carcinoma}$$

Some of the NCIt concepts are annotated as being mapped to some ICD-O-3 morphologies. For example, *Invasive ductal carcinoma, not otherwise specified* is annotated as being mapped to two ICD-O-3 morphology codes (8500/3 *Infiltrating duct carcinoma, NOS* and 8521/3 *Infiltrating ductular carcinoma*).

The semantics of this mapping annotation is not defined (i.e., exact match or another type of relationship). In the NCIt, even if the term *disease* is employed, it is not clear whether *Neoplasm* represents the disease or the diagnosis. For instance, in the NCI term Browser, *Neoplasm* is defined as *"A benign or malignant tissue growth..."* and *"An abnormal mass of tissue..."*. As discussed above, disease classification are mainly used in EHR for diagnosis recording and in the remaining part of this manuscript we use NCIt *Neoplasm* as a kind of diagnosis describing the disease.

An OWL-DL representation of the NCIt is freely available in the Web ontology Language (OWL) format on the NCI website (`http://cbiit.nci.nih.gov/evs-download`). Although logic-based reasoning can be made with this OWL-DL representation, some inconsistencies have been discussed and it has been shown that the NCIt should be used cautiously for this purpose [58, 20, 59].

4.1.5 NCI Metathesaurus

NCI Metathesaurus (NCIm) is a biomedical terminology database *"that covers most terminologies used by NCI for clinical care, translational and basic research, and public information and administrative activities"* [60]. It has been built and maintained by the NCI. Its structure and a significant part of concepts are based on the UMLS Metathesaurus [61]. Inside NCIm, identical elements from different terminologies are related to the same Concept Unique Identifier (CUI).

4.2 Methods

We focused our study on primary tumor descriptions, ignoring metastases and uncertain behaviors. The ICD-10 and ICD-O-3 terminologies do not have a formal representation. In [62], authors recommend to use SKOS to describe the knowledge of such resources. In order to bridge these two terminologies, it is necessary to identify how concepts that are represented within them (diagnosis, morphology, topography) are related. These relationships should therefore be represented at the conceptual level so that they could be machine readable. Moreover, the concept represented by terminologies should be conceptually defined on the top of the corresponding codes in order to be independent of terminology to integrate, thus enabling the integration of other disease classification. Our approach was to follow the W3C recommendations to define formal and semi-formal hybrid models [63] in order to build a model combining SKOS for terminologies description and OWL for representing concepts involved and defining of relationships between these concept as proposed in [62].

Figure 4.1 presents the organization of the proposed model using Graffoo

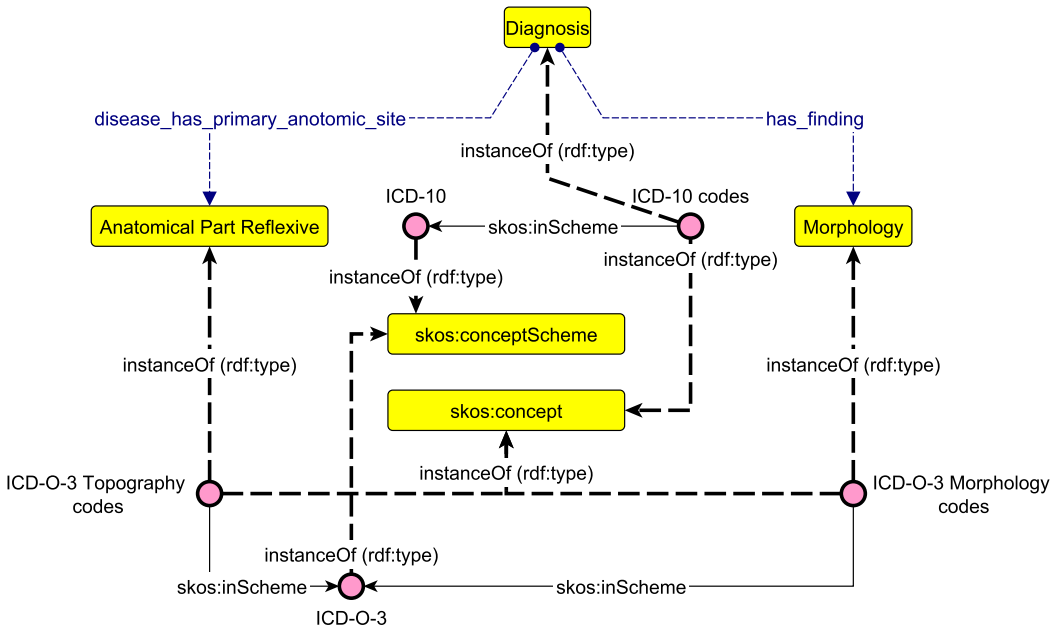

FIGURE 4.1 – Graffoo [64] representation of the proposed model. The model is formal and semi-formal hybrid. Terminologies (ICD-10 and ICD-O-3) are represented in SKOS. Above them, a formal model is represented in OWL. Every OWL class of the formal model are subclasses of skos :Concept so that they can be instantiated by terminological artifacts.

[64].

The methods can be divided in three steps :

— Defining a formal pattern for linking diagnosis, topography and morphology
— Building a model based on the NCIt corresponding to the formal pattern
— Instantiating the model with terminologies

4.2.1 Defining a formal pattern for linking diagnosis, topography and morphology

In order to link ICD-O-3 and ICD-10 concepts, it is necessary to determine which relationships are involved and how these relationships associate concepts with each other. A topography-morphology combination in ICD-O-3 leads to a diagnosis description. ICD-O-3 axes can be viewed as descriptors that, when combined, provide necessary and sufficient information to represent a diagnosis. For instance, the diagnosis *Malignant neoplasm of lower-outer quadrant of breast* in ICD-10 can be defined as a malignant neoplasm arising from the lower-outer quadrant of breast (because it is defined as a presumed or stated primary malignant tumor within ICD-10). The mention of “arising from” is

ambiguous because this relationship implies the fact that the tumor arises from the topography as a whole (lower-outer quadrant of breast) or from a part of this topography (a part of lower-outer quadrant of breast). Indeed, if a digestive system's tumor is reported, it may refer to a tumor that originates from a part of the digestive system and not from the whole digestive system. In order to capture the fact that a primary tumor refers to a primary site as a whole and all its parts, we need specific topography classes. In [65], the W3C describes a way to represent those reflexive parts (e.g., *"Class(CarPart_reflexive complete unionOf(Car CarPart))"*). This pattern has been proposed (as S-node) for biomedical domain in [66, 67]. Formally, we can define *Malignant neoplasm of lower-outer quadrant of breast* as a diagnosis whose morphology is a malignant neoplasm and whose primary site is the reflexive part of lower-outer quadrant of breast. In description logics, this can be stated as :

Malignant neoplasm of lower outer quadrant of breast ≡

Diagnosis

∩∃ has_morphology.Malignant neoplasm

∩∃ has_primary_site.Lower outer quadrant of breast Reflexive part

In addition, because of its expressivity, ICD-O-3 provides finer-grained information about the morphology of diagnosis than ICD-10 does. For instance, an adenocarcinoma arising from the lower-outer quadrant of breast can be reported using ICD-O-3. In ICD-10, there is no code corresponding to this diagnosis. However, an adenocarcinoma being a type of malignant neoplasm, an adenocarcinoma arising from the lower-outer quadrant of breast can be defined as a type of malignant neoplasm arising from the lower outer quadrant of breast (which is a coarser grained concept that exists in ICD-10). Formally, this definition is equivalent to :

Malignant neoplasm of lower outer quadrant of breast ⊆

Diagnosis

∩∃ has_morphology.Adenocarcinoma

∩∃ has_primary_site.Lower outer quadrant of breast Reflexive part

4.2.2 Building a model based on the NCIt corresponding to the formal pattern

Building a part-whole lattice

In order to address the integration of diagnosis (ICD-10) with topographies and morphologies (ICD-O-3), the NCIt relationship *disease_has_primary_anatomic_site* is of particular interest. The NCIt's definition of this relationship is : *"A role used to relate a disease to the anatomical site where the originating pathological process is located. The domain and the range for this role*

are 'Disease, Disorder or Finding' and 'Anatomic Structure, System, or Substance'". This relationship is equivalent to the *has_primary_site* relationship defined in the previous subsection. As discussed in page 42, we should consider that the primary anatomic site of a tumor encompasses the site itself and all its parts (this definition is in accordance with "is located", which is mentioned in the NCIt definition of the *disease_has_primary_anatomic_site* relationship). In order to make this description possible, we have built a subsumption lattice composed of classes defined as the reflexive part of each *Anatomic Structure, System, or Substance*. For instance Lower outer quadrant of breast Reflexive part was defined as follows :

Lower outer quadrant of breast Reflexive part ≡

Lower outer quadrant of breast

∪∃ part_of.Lower outer quadrant of breast

The lattice was built with DL-reasoning over classes defined using two part-whole relationships available in the NCIt (namely anatomic_structure_is_physical_part_of and anatomic_structure_has_location).

Isolating morphologies

In contrast, no morphology axis is distinguished as such within the NCIt, and it is not possible to find a relationship equivalent to the aforementioned has_morphology. However, the NCIt provides a mapping between diagnosis and ICD-O-3 morphology codes. We have added classes corresponding to ICD-O-3 morphologies as types of NCIt *Findings* and, based on the NCIt mappings, we have built new "refined" concepts defined according to the following model :

[NCIt concept (refined)] ≡

[NCIt concept]

∩∃ disease_has_finding [Morphology mapped to the NCIt concept]

For instance, in the NCIt, *adenocarcinoma* is mapped to 8140/3 ICD-O-3 morphology (*Adenocarcinoma, NOS*) the corresponding "refined" NCIt concept was defined as follow :

adenocarcinoma (refined) ≡

adenocarcinoma

∩∃ disease_has_finding Adenocarinoma,NOS (ICD-O-3 morhology)

Each morphologies were classified depending on their tumoral behavior as described in ICD-O-3 (i.e., benign, malignant primary, in situ, malignant metastatic, unknown whether benign or malignant, unknown whether primary or metastatic).

Building the model

Using reflexive part anatomic concepts and morphologies and adapting the description logics' expressions proposed in page 42, we have defined a pattern to formally describe relationships between diagnosis, morphologies and topographies within the "derivative" NCIt. In description logics, the pattern is the following :

Diagnosis ≡
∃ disease_has_finding.Morphology
∩∃ disease_has_primary_anatomic_site.Topography_Reflexive_Part

Based on the defined pattern, we implemented and executed the following algorithm :

+ For each (Morphology identified → [Morphology])
  + For each (Topography_Reflexive_Part identified → [Topography])
    + Build [expression] of the form :
      ∃ disease_has_finding.[Morphology]
      ∩∃ disease_has_primary_anatomic_site.[Topography]
    + If at least one subclass of [expression] exists in the NCIt then
      + Build the *Diagnosis* class defined as equivalent to [expression]

We have then implemented the model presented in Figure 4.1 containing :

1. Morphologies
2. Reflexive part topographies
3. Diagnosis identified by the aforementioned algorithm.

Instantiating the model with disease classification

ICD-O3 ICD-O3 morphologies were represented as instances of the built-in morphology classes. ICD-O-3 topographies were represented as instances of the built-in reflexive part topographies. Reflexive part topographies to be instantiated by ICD-O-3 topographies were identified as follow :

1. Identify mappings between ICD-O-3 topographies and NCIt concepts having the same CUI within NCIm
2. Define these codes as instances of the corresponding NCIt concept
3. Retrieve the corresponding reflexive part topography after DL-reasoning

ICD-10 ICD-10 codes were represented as instances of built-in diagnosis classes. Diagnosis to be instantiated by ICD-10 codes were identified as follow :

1. Identify mappings between ICD-10 codes and NCIt concepts having the same CUI within NCIm
2. Define concepts corresponding to ICD-10 codes based on the NCIt definition (by adding a restriction for the primary site based on NCIt concept formal definition) and ICD-10 (by adding a restriction for the behavior) semantics. For instance :
   — *Breast, Unspecified* (C50.9) is a malignant primary neoplasm within the ICD-10 classification.
   — *Breast, Unspecified* (C50.9) has the same CUI as *Malignant Breast Neoplasm* (C9335) within NCIm.
   — *Malignant Breast Neoplasm* (C9335) has the associated primary site Breast (C12971) within NCIt.
   — The built expression describing *Breast, Unspecified* (C50.9) was then :

     Malignant Breast Neoplasm
     ∩∃ disease_has_finding.Maligant primary neoplasm
     ∩∃ disease_has_primary_anatomic_site.Breast
3. Retrieve the corresponding Diagnosis after DL-reasoning

4.2.3 Evaluation of the model

The National Cancer Institute provides, within the Surveillance, Epidemiology, and End Results (SEER) Program, a set of tools for ICD conversions [68]. We used the 2014-05-08 conversion file of ICD-O-3 to ICD-9-CM, to ICD-10 (Causes of Death) and to ICD-10-CM (available at `http://seer.cancer.gov/tools/conversion/`) as a gold standard for evaluating how ICD-O-3 and ICD-10 could be related. Based on this file, we have rebuilt ICD-O-3 topography-morphology combinations mapped to ICD-10 codes. A 2-steps evaluation was then performed :

— For each combination, we queried the proposed model in order to evaluate how many branches of the diagnosis lattice were instantiated by both the ICD-10 code and the topography-morphology combination.
— We tried to build mappings based on the proposed model with a simple algorithm (ICD-10 codes and topography-morphology combinations with the minimum hierarchical edge-based distance were considered as mapped) and compared it with the gold standard.

	Total number	Instantiating the final model
ICD-O-3 Topographies	409	278 (68.0%)
ICD-O-3 Morphologies	873	860 (98.5%)
ICD-10	727	302 (41.5%)
ICD-10 Benign	180	73 (40.5%)
ICD-10 In situ	66	22 (33.3%)
ICD-10 Malignant	481	207 (43.0%)

TABLE 4.1 – Part of ICD-O-3 and ICD-10 terminologies integrated within the final model.

4.3 Results

All the analyses were processed over the OWL-DL version of the NCIt (14.11d) available at http://evs.nci.nih.gov/ftp1/NCI_Thesaurus/.

4.3.1 Built model based on the NCIt

A total of 6 720 topographies involved in at least one diagnosis definition were identified and the corresponding topographies reflexive parts were introduced in the NCIt in order to build the topography reflexive parts lattice (section : Building a part-whole lattice). A total of 1 120 NCIt codes were identified as being related to 1 094 ICD-O-3 morphology codes. The 1 094 corresponding morphology classes were added to the model and automatically classified under six general morphology classes depending on their behavior leading to a set of 1 100 possible morphologies. Combining the 1 100 morphology classes with the 6 720 reflexive part topographies, 7 392 000 expressions were built. A total of 20 133 (0,27%) expressions subsuming at least one NCIt code were identified and the corresponding classes were introduced in the model as diagnosis.

4.3.2 Instantiating the model with disease classification

Table 4.1 presents the part of each terminology that was covered by the final model. The numbers of codes to be integrated were :

— 409 ICD-O-3 topographies
— 873 ICD-O-3 morphologies (excluding /6 *Malignant neoplasms, stated or presumed to be secondary* and /1 *Neoplasms of uncertain and unknown behavior*)
— 727 ICD-10 neoplasms (excluding C81-C96 *Malignant neoplasms, stated or presumed to be secondary* and D37-D48 *Neoplasms of uncertain or unknown behavior*)

Using NCIm, 298 ICD-O-3 topography codes were linked to 540 NCIt codes. Within these NCIt codes, 29 were not subclasses of *Anatomic Structure, System, or Substance*. Among the 298 ICD-O-3 topography codes, 20 were related

	N	Instances of multiple classes
ICD-O-3 Topographies	278	79 (28.4%)
ICD-O-3 Morphologies	860	0 (- %)
ICD-10	302	153 (50.7%)
ICD-10 Benign	73	26 (35.6%)
ICD-10 In situ	22	22 (100%)
ICD-10 Malignant	207	105 (50.7%)

TABLE 4.2 – Number of codes instantiating multiple classes in the model.

only to these 29 codes and were then excluded (e.g., C05.1 *Soft palate, NOS* was erroneously mapped to *Malignant Soft Palate Neoplasm*). Thus, 278 topography codes were finally included within the model as instances of the corresponding NCIt codes and classified as instances of topography reflexive parts after DL-reasoning. Using NCIm, 302 ICD-10 codes were linked to NCIt codes. Building the corresponding expressions and after DL-reasoning, we were able to add 302 ICD-10 codes as instances of 380 diagnosis.

4.3.3 Characteristics of the final model

The resulting model is constituted of 113 643 axioms, including 27 953 classes (6 720 topographies, 1 100 morphologies and 20 133 diagnosis). A total of 1 440 codes were instantiated (278 ICD-O-3 topographies, 860 ICD-O-3 morphologies and 302 ICD-10 codes).

Within the model, a significant part of ICD-10 (51%) and ICD-O-3 topography codes (28%) are instances of multiple classes (Table 4.2). This situation arises when the hierarchy of diagnosis within the NCIt is not in accordance with the topography or the morphology that we used to describe them. For example *Colon Cavernous Hemangioma* is a direct subclass of the following expressions :

- ∃ disease_has_finding.Cavernous hemangioma
  ∩ ∃ disease_has_primary_anatomic_site.Colorectal Region Reflexive part
- ∃ disease_has_finding.Cavernous hemangioma
  ∩ ∃ disease_has_primary_anatomic_site.Colon Reflexive part
- ∃ disease_has_finding.Hemangioma,NOS
  ∩ ∃ disease_has_primary_anatomic_site.Colorectal Region Reflexive part
- ∃ disease_has_finding.Hemangioma,NOS
  ∩ ∃ disease_has_primary_anatomic_site.Colon Reflexive part

The explanation for this situation is twofold : (1) there is neither an *anatomic_structure_has_location* relationship, nor an *anatomic_structure_is_physical_part_of* relationship between *Colon* and *Colorectal Region* within

Tumors	All N=42 260 (%)	Hematopoietic N=14 213 (%)	Solid N=28 047 (%)
Related in the model*	42 260 (100.0)	14 213 (100.0)	28 047 (100.0)
More than 1 branch[a]	15 234 (36.1)	9 910 (69.7)	5 324 (18.9)
Rebuilt from the model**	17 766 (42.0)	739 (5.2)	17 027 (60.7)
Non unique mappings[b]	4 886 (27.5)	333 (45.1)	4 553 (26.7)

*Related in the model means that there is at least a common diagnosis inside the model that is instantiated by both the ICD-10 code and the ICD-O-3 combination.
**Mappings rebuilt from the model corresponds to the mappings that we were able to rebuild automatically from the model.
[a]More than 1 branch means that there is more than one branch of the diagnosis lattice that was instantiated by both the ICD-10 code and the ICD-O-3 combination.
[b]Non unique mappings means that the topography-morphology combination was also mapped to another ICD-10 code (inconsistent with the SEER file).

TABLE 4.3 – Comparison with the SEER conversion program according to the tumor type (hematopoietic and solid tumors) and the number of branches of the diagnosis lattice that are identified for an ICD-10 code / ICD-O-3 combination.

the NCIt ; (2) *Colon Cavernous Hemangioma* is described as having these two anatomic structures as a primary site. On the other hand, through the NCIt diagnosis lattice, *Colon Cavernous Hemangioma* is described as being a subclass of the concepts *Cavernous hemangioma* and *Hemangioma, NOS*.

4.3.4 Comparison with the SEER conversion file

Based on the SEER conversion file, excluding metastatic and uncertain behaviors from ICD-10 and ICD-O-3 morphologies, we were able to build 103 950 mappings between an ICD-O-3 topography-morphology combination and an ICD-10 code. Because some codes were missing in our model, 59% of these mappings could not be evaluated. Table 4.3 presents the results of the evaluation over the 42 260 mappings combining codes which were instantiated within the resulting model. The model relates 100% of the mappings through at least a diagnosis. A significant part of these mappings (36%) are related by more than one branch of the diagnosis lattice, especially for hematopoietic tumors (70%). Using a simple algorithm, the model was able to identify 42% of the mappings of the SEER file (61% for solid tumors and 5% for hematopoietic tumors). A quarter of these topography-morphology combinations were also mapped to another ICD-10 code, which is not consistent with the SEER file.

4.4 Discussion

4.4.1 Implemented methods to build the model

We achieved to automatically build a model based on the NCIt, describing topographies, morphologies and diagnosis that can be instantiated by both ICD-O-3 and ICD-10 codes. As no morphological axis is available within the NCIt, we have adapted the NCIt by adding concepts corresponding to ICD-O-3 morphologies, which we related to the corresponding diagnosis (based on the ICD-O-3 annotation of the NCIt).

For the description of topographies, we have built an organ reflexive part lattice that enables the description of a primary site as encompassing the site itself and all its parts. These reflexive parts have been proposed for the biomedical domain representation in [66].

The diagnosis lattice was then automatically generated by DL-reasonners based on the topography and morphology lattices preventing is_a overloading [20].

In order to instantiate the model, we have used NCIm to identify links between NCIt and the terminologies (namely ICD-O-3 topography and ICD-10). For ICD-10, we had to add a restriction based on the semantics available within the ICD-10 classification in order to ensure that primary tumors were described according to a primary site. The resulting model could not be instantiated completely by ICD-O-3 and ICD-10 codes for different reasons :

— NCIt completeness for describing diagnosis. As our method rely on NCIt diagnosis, the resulting classes which were built depend on their existence within the NCIt (e.g., C00.1 *External Lower Lip malignant neoplasm* is not available within the NCIt).

— NCIm provides a way for identifying common concepts using the CUI but it can be incomplete or wrong (e.g., 20 ICD-O-3 topographies were mapped erroneously to non-anatomic concepts).

However, when the codes were found, we were able to identify a common diagnosis for all cases described in the SEER conversion program and, using a simple algorithm, 42% of the SEER mappings (corresponding to codes instantiating the model) could be rebuilt from the model. Our aim is not to enable conversion between codes but to provide a machine usable and semantically integrated view over them. From this perspective, the model seems to be consistent for addressing this integration purpose. Moreover, the model describes much more possible relationships between diagnosis then the SEER conversion does. For instance, in the SEER conversion program, there is no relationship between *Adenocarcinoma, NOS – Colon, NOS* (C18.9 – M8140/3) and *Malignant neoplasm of rectosigmoid junction* (C19.9) whereas our model identifies successfully that they are both instances of *Malignant, primary site - Large Intestine Reflexive part.*

4.4.2 Choice of the NCIt

In the biomedical field, other description logics-based terminologies exist. Specifically, SNOMED-CT® provides not only topography, morphology and diagnosis dimensions but also implements relationships between these concepts. However, the NCIt is specific to the oncology field and provides useful knowledge related to neoplasm diagnosis. In addition, it is freely and easily accessible. On the other hand, SNOMED-CT® has a much more restrictive affiliate license agreement and it is not easily accessible for countries which are not members of the IHTSDO . In addition, it has been shown that SNOMED-CT®'s formal representation suffers from the same flaws [69, 70] as the NCIt and has to be used cautiously while needing logic-based reasoning. Thus, a similar evaluation could be carried out on SNOMED-CT® in order to estimate whether it could be useful for integrating disease classification in oncology and to compare the results with what we found when using the NCIt.

4.4.3 Limitations of the NCIt for integration purposes

In [59], Schultz et al. discussed that the OWL Description Logics (OWL-DL) version of the NCIt may lead to unexpected results which were not visible due to the lack of use case needing logic-based reasoning over the OWL-DL version of the NCIt. Integration of heterogeneous disease classification corresponds to one of such a use cases. Indeed, we've have identified, some limitations due to inconsistencies.

On the one hand, the NCIt provides, concepts describing cancer diagnosis and, on the other hand, concepts describing the tumor topography. It also provides relationships which are involved in topography-morphology combinations which are expected to be equivalences of diagnosis. Its formal representation and the availability of an OWL version enable reasoning and the implementation of DL-queries. However, some intrinsic characteristics prevent its direct use for the integration of cancer disease classification : (i) the absence of distinction between morphologies and diagnosis ; (ii) diagnosis concepts described as having a specific primary site but not its parts. We have proposed a method to address these issues and to automatically build a consistent model based on the NCIt and the intrinsic semantics available within ICD-O-3 and ICD-10.

During the building process, a significant part of NCIt concepts were retrieved as subclasses of multiple diagnosis classes. As a result, the corresponding ICD-10 codes were defined as instances of multiple diagnosis classes and 36% of SEER mappings evaluated were retrieved as being related to more than one branch in the diagnosis lattice. For instance, the ICD-10 code C18.0 *Malignant neoplasm : Caecum* was mapped to the NCIt concept C9329 *Malignant Cecum Neoplasm*, which is related to multiple anatomic sites : *Gastrointestinal System*, *Cecum*, *Colon*, *Intestine* and *Colorectal Region*. As there is no relationship

between *Cecum*, *Colorectal Region* and *Colon* within the NCIt (except that they are part of the *large intestine*), C18.0 instantiates the following classes :

— *Malignant, primary site – Cecum Reflexive part*
— *Malignant, primary site – Colon Reflexive part*
— *Malignant, primary site – Colorectal Region Reflexive part*

Two issues can be identified : (i) *Malignant Cecum Neoplasm* should not have *Colon* as an associated anatomic site within the NCIt because *Cecum* is neither a part, nor a subclass of *Colon*, (ii) *Cecum* and *Colon* should be related to *Colorectal Region*. The former is due to is_a overloading and has been discussed in [20]. The latter issue is due to the lack of part_of relationships within the NCIt. Another important issue can be identified for in situ neoplasms. 100% of the in situ ICD-10 codes are instances of more than one diagnosis within the model. The NCIt asserts that a *Carcinoma In situ* is a *Carcinoma*, which seems to be true. However, in the NCIt, *Carcinoma* is related to the *Carcinoma, NOS* ICD-O-3 morphology (having an invasive behavior) and *Carcinoma In situ* is related to the *Intraepithelial carcinoma, NOS* ICD-O-3 morphology (having an in situ behavior). Consequently, the subsumption relationship between *Carcinoma In situ* and *Carcinoma* is not consistent because a tumor cannot be both invasive and in situ at the same time. For instance, D05 *Carcinoma in situ of breast* is mapped to the NCIt concept C3641 *Stage 0 Breast Cancer*, which is related to the *Intraepithelial carcinoma, NOS* and *Epithelioma, NOS* ICD-O-3 concepts through the NCIt lattice. As a result, D05 instantiates the following classes :

— *Intraepithelial carcinoma, NOS - Breast Reflexive part*
— *Epithelioma, NOS - Breast Reflexive part*

Because *Intraepithelial carcinoma, NOS* has an in situ behavior and *Epithelioma, NOS* has a malignant, invasive behavior, it is not consistent to be an instance of these two diagnosis. This issue emphasizes erroneous mappings that may exist between ICD-O-3 and the NCIt due to ambiguous labels. A simple solution to this problem would be to add a concept representing the "Carcinoma" category of which both *Carcinoma* a *Carcinoma In situ* should be subclasses.

It is noteworthy that these patterns, which are mainly due to is_a overloading, can easily be retrieved by searching for codes which are instances of multiple diagnosis. By linking ICD-O-3 and ICD-10 terminologies to the NCIt and adding some restrictions based on their own semantics, our method may provide a useful auditing solution. Indeed, identifying those codes within the resulting model may enable to find within the NCIt : (i) structural inconsistencies (e.g., *Malignant cecum neoplasm* related to emphColon), (ii) missing concepts (e.g., Carcinoma invasive that can be related to the ICD-O-3 concept Carcinoma, NOS) and (iii) missing relationships between concepts (e.g., emphCecum which should be defined as a part of emphColorectal Region).

In order to build the organ reflexive part lattice, parts of anatomical concepts

were identified using transitive part-whole properties available within the NCIt (namely anatomic_structure_is_physical_part_of and anatomic_structure_has_location). This results in including *Cell Part* as (indirect) sub classes of topographies (i.e. *Birbeck Granule* part of *Langerhans Cell* part of *Epidermis* part of *Skin*). This would suggest that we allow a neoplasm to have *Birkbeck Granule* as primary site. Since the range of the disease_has_primary_anatomic_site property include *cells parts*, such an assertion is allowed in the NCIt. Thereby, the built hierarchy is in accordance with NCIt representation of primary sites. As discussed in [67], the transitivity of the part of property remains controversial. For instance in [71] Rescher stated that *"A part (i.e., a biological sub-unit) of a cell is not said to be a part of the organ of which that cell is a part"* which is in contradiction with what is stated within the NCIt. However, diagnosis built in the 'derivative' model depend on their existence of within the NCIt and as NCIt describe only existing (even if sometimes rare) tumors, diagnosis definitions remain realistic. Nevertheless further work should be done in order address this issues. In this context patterns proposed by Schulz and Hahn in [67] are to be investigated.

4.4.4 Perspectives

This work confirms that inconsistencies present in the NCIt lead to sub optimal (if not erroneous) classification when using logic-based reasoning over the NCIt in the context of a specific use case. Thus, OWL-DL version of the NCIt should be used cautiously. However, this resource may be helpful in order to build a formal model for integrating cancer disease classification. Using NCIm CUI to map the NCIt to ICD-O-3 and ICD-10 can be useful but is not enough because mappings are missing and some are inconsistent. We are currently working on a method based on the NCIm to identify additional mappings.

SNOMED-CT® exposes comparable structural characteristics with diagnosis, anatomic and even morphological concepts as well as relationships between them. Future work will explore SNOMED-CT as a resource for integration purpose. As SNOMED-CT is known to have the same inconsistencies as NCIt, we will study the feasibility of using both SNOMED-CT and the NCIt to build a consistent model addressing semantic and structural heterogeneities between disease classification in oncology.

As discussed above, topographies representation need to be refined in order to avoid inconsistencies and define consistent level of granularity for propagation of the disease_has_primary_anatomic_site property. The Foundational Model of Anatomy (FMA) Ontology [72] *"is a domain ontology that represents a coherent body of explicit declarative knowledge about human anatomy"* [73]. Further work will explore the ability to define these topographies based on the FMA.

The "derivative" model can contribute to EHR data integration for secondary use in the context of oncology data. For instance it can be used to build an algorithm for neoplastic disease identification based on diagnosis available within EHRs. Indeed, this task can be challenging as soon as diagnosis are recorded with heterogeneous terminologies. This resource can manage this heterogeneity by providing an integrated view diagnosis recorded in EHRs.

4.5 Conclusion

We have proposed a method to automatically build a model for integrating ICD-10 and ICD-O-3 based on the NCIt. The resulting "derivative" model is a machine understandable resource that enables an integrated view of these heterogeneous terminologies. The NCIt structure and the available relationships can help to bridge disease classification taking into account their structural and granularity heterogeneity. However, (i) inconsistencies exists within the NCIt leading to miss classifications in the "derivative" model, (ii) the "derivative" model only integrates a part of ICD-10 and ICD-O-3. The NCIt is not enough for the integration purpose and further work based on other termino-ontological resources is needed in order to enrich the model and avoid identified inconsistencies.

Chapitre 5

Building an ontology based on IACR rules for multiple primary tumor registration

We have proposed an approach for building a model for diagnosis terminology integration in oncology. Even if the obtained model could be used as a baseline, terminology coverage remains insufficient. Completing such a detailed model manually or using other resources such as FMA or SNOMED-CT is out of the scope of our work. In the following part, we will report the implementation of a high level model describing diagnosis, disease and terminology axis. We will focus on modeling classes corresponding to concepts described in the registration rules for multiple primary cancer published by the International Agency for Cancer Registry.

5.1 background

In order to harmonize data collection, the International Association for Cancer Registries (IACR), the International Agency for Research on Cancer (IARC) and the World Health Organization (WHO) specify that registered cases should be coded according to the International Classification of Diseases in Oncology, 3rd edition (ICD-O-3) [75, 34]. Further to this, recommendations have been issued in collaboration with IACR, IARC, WHO and the European Network of Cancer Registries (ENCR) concerning registration rules for multiple primary cancers [74]. These recommendations define when a record should be considered to contribute to a new case and when it should be considered to contribute to an already registered case, and the level at which data are to be aggregated for followup of incidence and survival data.

In [74], primary cancer is defined : *"A primary cancer is one that originates in a primary site or tissue and is not an extension, nor a recurrence, nor*

ICD-O-2/3 site code	Label
C01	Base of tongue
C02	Other and unspecified parts of tongue
C00	Lip
C03	Gum
C04	Floor of mouth
C05	Palate
C06	Other and unspecified parts of mouth
C09	Tonsil
C10	Oropharynx
C12	Pyriform sinus
C13	Hypopharynx
C14	Other and ill-defined sites in lip, oral cavity and pharynx
C19	Rectosigmoid junction
C20	Rectum
C23	Gallbladder
C24	Other and unspecified parts of biliary tract C24.9
C33	Trachea
C34	Bronchus and lung
C40	Bones, joints and articular cartilage of limbs
C41	Bones, joints and articular cartilage of other and unspecified sites
C65	Renal pelvis
C66	Ureter
C67	Bladder
C68	Other and unspecified urinary organs

TABLE 5.1 – Groups of topography codes considered a single site in the definition of multiple cancers. Adapted from [74]

a metastasis.”. Following this definition, we can state that primary tumor is a kind of *disease* as defined in [17]. This document is therefore defining rules for identifying *neoplastic disease* based on *diagnosis* describing them. The *neoplastic disease* is considered to be defined by a topography-morphology combination. Namely, a *neoplastic disease* has a *primary site* from which it originates and a *morphology* depending on the cell it is composed of.

For a single patient, only one *neoplastic disease* is to be identified for each topography-morphology combination (if the same combination arise multiple times it is considered to be a local recurrence of the same disease). Topography and morphology are described using ICD-O-3. Moreover, based on ICD-O-3 codes, “*some groups of codes are considered to be a single organ for the purposes of defining multiple tumors*” (table 5.1) and the same rule is applied for morphology (table 5.2). Kaposi sarcoma and tumors of haematopoietic system being systemic tumors and are considered as a unique disease no matter the *primary site*. Considering morphology, imprecise groups are defined (numbers between brackets in table 5.2). Given a tumor arising from a single primary site (organ group), if “*one morphology is not specific [...]and a specific morphology is available, the case should be reported with the specific histology and the non-specific diagnosis should be ignored*”.

5.2 Methods

Our approach was to :

— Build a core model by :
  — Defining concept involved in disease classifications, namely diagnosis, anatomic site and morphology.
  — Defining disease following rules described in [74]
  — Defining relationship between the define concepts.
— Describe IARC groups inside the core model
— Model diagnosis by combining topography and morphology groups
— Instantiate the model with terminologies (ICD-O-3, ADICAP and ICD-10).

5.2.1 Core model for disease classification integration

Classes definition

This core model were built on three axis corresponding to concepts involved in disease classifications (section 4.1.3) enriched with two axis for morphology description :

— *Diagnosis*, defined in [17] as “*a conclusion of an interpretive process that has as input a clinical picture of a given patient and as output an*

Group	Codes
Carcinomas	
1. Squamous and transitional cell carcinoma	8051-8084, 8120-8131
2. Basal cell carcinomas	8090-8110
3. Adenocarcinomas	8140-8149, 8160-8162, 8190-8221, 8260-8337, 8350-8551, 8570-8576, 8940-8941
4. Other specific carcinomas	8030-8046, 8150-8157, 8170-8180, 8230-8255, 8340-8347, 8560-8562, 8580-8671
(5) Unspecified carcinomas (NOS)	8010-8015, 8020-8022, 8050
6. Sarcomas and soft tissue tumours	8680-8713, 8800-8921, 8990-8991, 9040-9044, 9120-9125, 9130-9136, 9141-9252,9370-9373, 9540-9582
7. Mesothelioma	9050-9055
Tumours of haematopoietic and lymphoid tissues	
8. Myeloid	9840, 9861-9931, 9945-9946, 9950, 9961-9964, 9980-9987
9. B-cell neoplasms	9670-9699, 9728, 9731-9734, 9761-9767,9769, 9823-9826, 9833, 9836, 9940
10. T-cell and NK-cell neoplasms	9700-9719, 9729, 9768, 9827-9831, 9834, 9837, 9948
11. Hodgkin lymphoma	9650-9667
12. Mast-cell Tumours	9740-9742
13. Histiocytes and Accessory Lymphoid cells	9750-9758
(14) Unspecified types	9590-9591, 9596, 9727, 9760, 9800-9801, 9805, 9820, 9832, 9835, 9860, 9960, 9970, 9975, 9989
15. **Kaposi sarcoma**	9140
16. **Other specified types of cancer**	8720-8790, 8930-8936, 8950-8983, 9000-9030, 9060-9110, 9260-9365, 9380- 9539
(17) **Unspecified types of cancer**	8000-8005

Table 5.2 – Groups of malignant neoplasms considered to be histologically 'different' for the purpose of defining multiple tumours. Adapted from [74]

assertion to the effect that the patient has a disease of such and such a type".

— *Anatomic site*. It can be defined as equivalent to the *Anatomical structure* (http://purl.org/sig/ont/fma/fma305751) of the FMA [72] :"*Material anatomical entity which is generated by coordinated expression of the organism's own genes that guide its morphogenesis ; has inherent 3D shape ; its parts are connected and spatially related to one another in patterns determined by coordinated gene expression. Examples : embryo, mesoderm, heart, right ventricle, mitral valve, myocardium, endothelium, lymphocyte, fibroblast, thorax, cardiovascular system, hemoglobin, T cell receptor*".

— *Morphology* defined in [76] as *"the kind of tumor that has developed and how it behave"*.

— *Histology* corresponding to the tumor/cell type (independent of the behavior) as defined in [76].

— *Behavior* corresponding to the way the tumor *"acts within the body"* (see SEER Training:Morphology).

— *Disease*, defined in [17] as a "*disposition (i) to undergo pathological processes that (ii) exists in an organism because of one or more disorders in that organism*".

Property definitions

In order to address composition issues, we have defined properties relating *diagnosis*, *topography* and *morphology* :

— *Diagnosis has associated anatomic site* relates a *diagnosis* with an *anatomic site* associated with the underlying pathological process described by the *diagnosis*. For example, the association can be a finding site, a primary anatomic site or a metastatic site. The domain of this property is *diagnosis* and the range is *Anatomic site*.

— *Diagnosis has finding site*, relates a *diagnosis* with the *anatomic site* where the underlying pathological process (described by the *diagnosis*) was found. It is defined as a sub-property of *Diagnosis has associated anatomic site*.

— *Diagnosis has metastatic anatomic site*,relates a *diagnosis* with the *anatomic site* where the underlying pathological process (described by the *diagnosis*) has secondarily spread. It is defined as a sub-property of *Diagnosis has associated anatomic site*.

— *Diagnosis has primary anatomic site*, relates a *diagnosis* with the *anatomic site* from where the underlying pathological process (described by the *diagnosis*) arose. It is defined as a sub-property of *Diagnosis has associated anatomic site*.

— *Diagnosis has morphology*, relates a *diagnosis* with the *morphology* de-

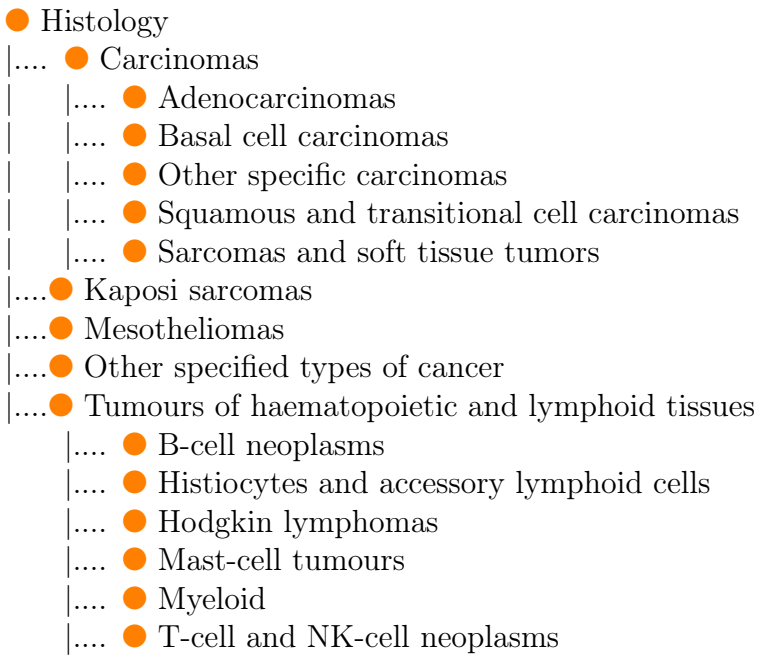

FIGURE 5.1 – Histology subsomption lattice

veloped in the underlying pathological process. The domain of this property is *diagnosis* and the range is *Morphology*.

— *Morphology has histologic type*, relates a *morphology* with the cell type (*histology*) of the corresponding pathological process. The domain of this property is *morphology* and the range is *histology*.

— *Morphology has behavior*, describe how the pathological process underlined by a *morphology* acts within the body (*behavior*). The domain of this property is *morphology* and the range is *behavior*.

— *Disease is described by*, relates a *disease* to *diagnosis* that refer to it.

5.2.2 Modeling IARC groups with the core model

IARC topography groups were modeled as sub-classes of *Anatomical structure*.As described in page 42, topography groups should be considered as reflexive parts of these *Anatomical structures*. These groups are based on ICD-O-3 topography classification, hence they correspond to non overlapping parts of the body. As a result, the groups were all defined as disjunctive from each other.

IARC morphology groups were modeled as sub-classes of *Histology*. Figure 5.1 presents the classification for morphology groups in the histology lattice. This classification is based on table 5.2, however, imprecise groups were not represented as such in the model.

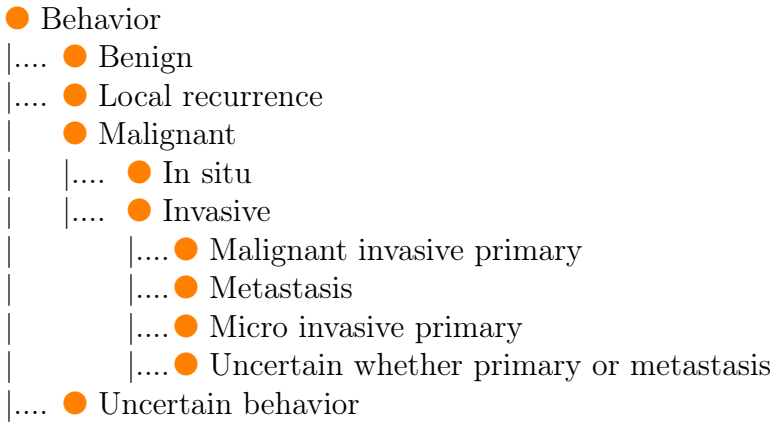

FIGURE 5.2 – Behavior subsomption lattice.

5.2.3 Modeling morphology, diagnosis and disease based on IARC groups

Morphology is composed of an histology and a behavior. In order to describe morphology, it is necessary to describe behavior in addition to histology. Based on ICD-O-3 and ICD-10 classifications, 10 *behavior* classes were described and classified in a subsomption lattice (Figure 5.2).

In order to build the *Morphology* lattice, a pattern was designed as presented in equation 5.1

$$\begin{aligned} \text{Morphology} \equiv \\ \exists\ \text{Morphology\_has\_behavior.Behavior} \\ \cap \exists\ \text{Morphology\_has\_histologic\_type.Histology} \end{aligned} \tag{5.1}$$

This pattern was applied for each *histology* and *behavior* combination and the corresponding morphology classes were introduced in the model.

Two *diagnosis* lattices were built separately :

— Neoplasm diagnosis classified depending on *histology* and primary site of the tumor. This lattice was built based on the pattern presented in equation 5.2 :

$$\begin{aligned} \text{Diagnosis} \equiv \\ \exists\ \text{Diagnosis\_has\_primary\_anatomic\_site.Anatomic\_site} \\ (\cap \exists\ \text{Diagnosis\_has\_morphology} \\ \exists\ \text{Morphology\_has\_histologic\_type.Histology}) \end{aligned} \tag{5.2}$$

This pattern was applied for each *anatomic site* and *histology* combination and the corresponding diagnosis classes were introduced in the model as sub-classes of *Diagnosis*.

— Neoplasm diagnosis classified depending on *behavior* and primary site of the tumor. This lattice was built based on the pattern presented in equation 5.3 :

$$\begin{aligned} &\text{Diagnosis} \equiv \\ &\exists\ \text{Diagnosis\_has\_primary\_anatomic\_site.Anatomic\_site} \\ &(\cap \exists\ \text{Diagnosis\_has\_morphology} \\ &\exists\ \text{Morphology\_has\_Behavior.Behavior}) \end{aligned} \tag{5.3}$$

This pattern was applied for each *anatomic site* and *behavior* combination and the corresponding diagnosis classes were introduced in the model as sub-classes of *Diagnosis*.

Diseases was then built as being described by *diagnosis*. In order to take into account systemic tumors as described in IARC rules, we proposed two different patterns to describe *disease* :

— For Systemic tumors [74]. *Disease* was defined as described by *diagnosis* having a specified *histology* (no matter the *topography*). The pattern is presented in equation 5.4

$$\begin{aligned} &\text{Disease} \equiv \\ &\exists\ \text{Disease\_is\_described\_by.Diagnosis} \\ &(\exists\ \text{Morphology\_has\_histologic\_type.Histology}) \end{aligned} \tag{5.4}$$

— For other tumors disease, we introduced the disease corresponding to each diagnosis applying the pattern presented in equation 5.5.

$$\text{Disease} \equiv \exists\ \text{Disease\_is\_described\_by.Diagnosis} \tag{5.5}$$

5.2.4 Instantiating the model with disease terminologies

ICD-O-3

Anatomic site were instantiated with ICD-O-3 topographies. Classifying ICD-O-3 topographies were straightforward since *anatomic site* were defined following IARC groups which depend on ICD-O-3 topographies (table 5.1).

Morphology were instantiated with ICD-O-3 morphologies. In order to discover witch *morphology* class should be instantiated by an ICD-O-3 morphology code we have built an expression describing the ICD-O-3 morphology within the model. The expression was built based on IACR group related to the code (table 5.2) and the behavior (equation 5.6). The expression was then used to query the model and retrieve the *morphology* class. ICD-O-3 morphology were then set as an instance of the retrieved *morphology* class.

$$\begin{aligned} &\exists\ \text{Morphology\_has\_behavior.Behavior} \\ &\cap \exists\ \text{Morphology\_has\_histologic\_type.Histology} \end{aligned} \tag{5.6}$$

ADICAP

In order to instantiate the model with ADICAP organ and morphologies our approach was to map these codes with ICD-O-3 topography and morphology. We have re-used existing mappings available from the Poitou-Charentes cancer registry [14] and the biobanque transcoder (http://transcoder.ebiobanques.fr/index.php?r=site/docs, resources and documentations are available at http://transcoder.ebiobanques.fr/index.php?r=site/docs). These mappings were reviewed and represented in RDF triples using SKOS mapping relations (namely skos:exactMatch, skos:broadMatch, skos:narrowMatch). All ADICAP codes mapped to an ICD-O-3 code were considered to instantiate the class to which the ICD-O-3 code belong.

ICD-10

In order to instantiate the model wih ICD-10 diagnosis, our approach was to identify which ICD-O-3 topography-morphology combination can be used to describe the ICD-10 code. We have used mapping table available in the Poitou-Charentes cancer registry terminology server [14] and enrich them with the 2014-05-08 conversion file of ICD-O-3 to ICD-9-CM, to ICD-10 (Causes of Death) and to ICD-10-CM (available at http://seer.cancer.gov/tools/conversion/). Link between ICD-O-3 and ICD-10 were manually reviewed. For each ICD-10 code, we have built an expression (equation 5.7) based on classes instantiating ICD-O-3 codes. These expression where then used to query the model and retrieve the corresponding *diagnosis*. Retrieved *diagnosis* were instantiated with the targeted ICD-10 code.

$$\begin{aligned} &\exists \text{ Diagnosis\_has\_primary\_} \\ &\text{anatomic\_site.[class of ICD-O-3 topography]} \\ \cap \exists &\text{ Diagnosis\_has\_morphology.[class of ICD-O-3 Morphology]} \end{aligned} \tag{5.7}$$

5.2.5 Material

The core ontology, histology, behavior and topography subsomption lattices were built using Tawny-owl [77]. Pre-coordinated classes (such as diagnosis, disease, and morphology) were built with OWL API 3.4.8 [78]. Reasoning over the ontology and DL-Query performing for instantiating classes with terminologies were implemented using OWL-API and HermiT reasonner [79].

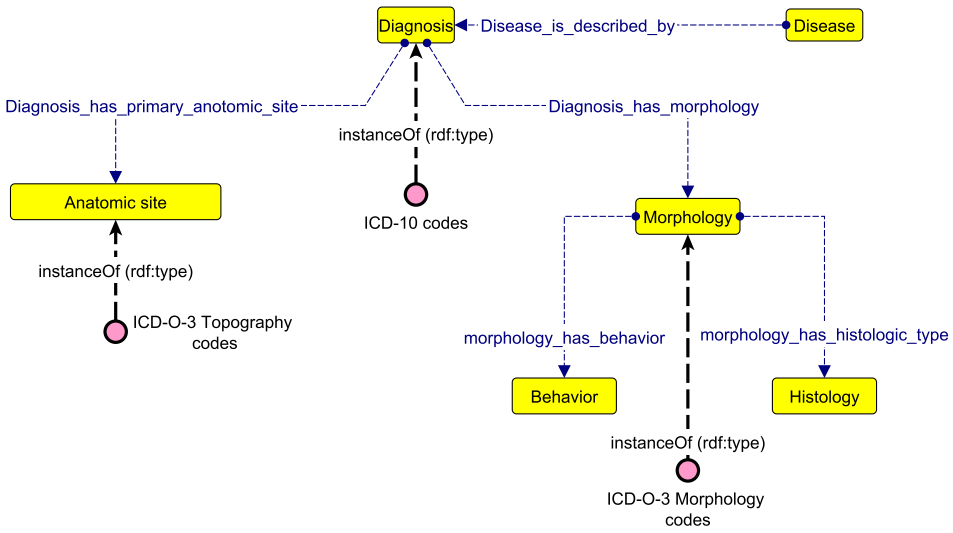

FIGURE 5.3 – Graphoo [64] representation of the core model for disease classification integration based on IACR rules for multiple primary tumor registration. Terminology codes are modeled as instances of classes representing, *diagnosis*,*morphology* and *topography*. The *disease* is considered to be implicitly described by *diagnosis*

	Core model	Complete model
Classes	16	2 389
Properties	7	7
Axioms	101	9 582

TABLE 5.3 – Metrics of the core model and its specification for IACR groups

5.3 Results

5.3.1 Obtained model

Figure 5.3 presents a high level representation of the core model using graphoo [64] formalism. The core model were built with 16 classes and 7 properties and was made of 101 axioms . The complete model specifically describing IACR groups were built with 2 389 classes and was made of 9 582 axioms (table 5.3).

The model leverages class definition in order to compose concept based on atomic classes (i.e. *histology*, *anatomic site* and *behavior*). Composed concepts such as *diagnosis* and *morphology* are classified depending on their formal definition. This approach prevents the "is_a overloading" [20] issue and ensure consistent classification within the model. A total of 190 *morphology* classes where built based on the 19 *histology* classes and the 10 *behavior* classes. Combined with the 53 *topography* classes these *morphology* classes lead to 1 569 *diagnosis* classes where built and related to 538 *disease* classes.

	N	%
ICD-O-3 Topography	398	99,5
ICD-O-3 Morphology	1 032	100,0
ICD-10	573	67,3
Adicap Organ	148	94,9
Adicap Lesion	1 343	87,7
Total	**3 494**	**88,0**

TABLE 5.4 – Code instantiating the model depending on the terminology of origin

5.3.2 Instantiating the model with disease classifications

Table 5.4 presents the number and percentage of codes integrating depending on the terminology of origin. Our approach enabled to integrate 88,0% of the targeted disease classifications codes. ICD-10 codes were not fully integrated with 67,3% of the targeted codes. Within the non integrated codes, a significant part were due to non mapped haematopoietic tumors and benign tumor. The 12% of ADICAP lesion codes were corresponding to codes that were not mapped to ICD-O-3 morphologies within the mapping resources.

Figure 5.4 presents an example of the model for prostate adenocarcinoma integration. Adicap organ and ICD-O-3 topography are instantiating the *prostate* (part-reflexive) class. Adicap lesion and ICD-O-3 morphology are instantiating the *adenocarinoma class*. The ICD-10 code instantiates the *prostate cancer diagnosis* class. After classification all these codes can be identified as describing at least a prostate cancer disease. Moreover, prostate cancer disease subsumes prostate adenocarcinoma disease in the model enabling to bridge these codes.

5.4 Discussion

The major issue to address is that ICD-10 and ICD-O-3 are not representing the diagnosis in the same way. Indeed, while ICD-10 represents the diagnosis as a whole, ICD-O-3 is a post-coordinated terminology that records diagnosis as a combination of topography and morphology. As a result, it is not possible to map ICD-O-3 topographies (or morphologies) to ICD-10 codes. Moreover, representing all possible values of topography-morphology combination would lead to a large amount of pre-coordinated terms not existing inside ICD-10 due to its granularity.

In order to address this issue, we have proposed to build a model on the top of these two terminologies. Our approach was to build a high level resource (avoiding precise representation of specific tumors) that corresponds to the specific use case of tumor identification in cancer registry. Based on

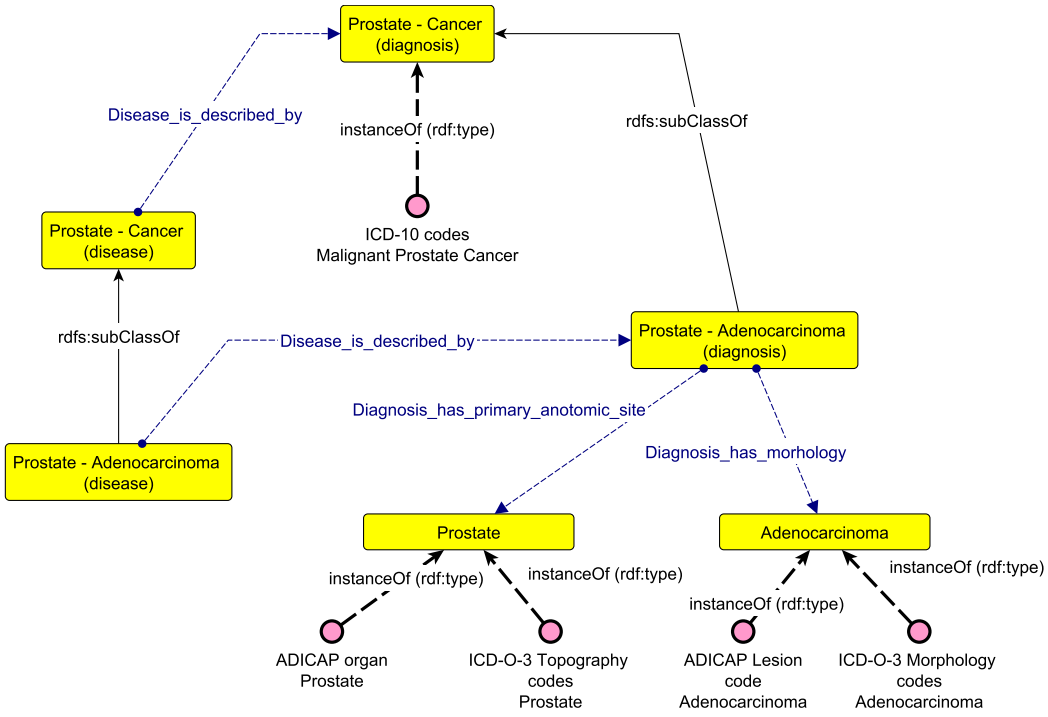

FIGURE 5.4 – Example of the model for representing a prostate adenocarcinoma. *Prostate* and *adenocarcinoma* are combined through to build the corresponding *diagnosis* using *Prostate adenocarcinma (diagnosis)* description logic definition. ICD-10 code does not provide information about morphology. However, the model enables two link it with diagnosis deduced from ICD-O-3 (or ADICAP) combination through the subsumption lattice

.

rules for multiple primaries registration [74], we have modeled an ontology representing *histotype* and *anatomic site* and necessary relationship to combine them (e.g *Diagnosis has primary anatomic site*). Based on these atomic concepts, we have defined needed concepts for ICD-10, ICD-O-3 integration purpose and disease representation (namely *morphology*, *behavior*, *diagnosis*, *disease*). Resulting defined concepts were automatically classified with a reasoner. This approach avoids "*is_ a overloading*" and ensure the consistency of formal definitions.

Obtained model were representing 2 389 classes and was able to integrate 88,0% of the codes represented in targeted terminologies. Missing codes were due to the resources used in order to instantiate the model. As the model is a high level representation of tumoral diagnosis, it enables to integrate a broad range of codes from various terminologies.

By using a model on the top of these terminologies we manage granularity heterogeneity. Thus, the codes meaning is kept unchanged but the model aggregates information at the class level. This approach is flexible. One can easily add narrower classes in the model and instantiate them with appropriate codes. The subsumption lattice can therefore manage semantic links between codes.

When mapping ICD-O-3 with ICD-10, the classical approach is to link ICD-10 code to a morphology code on the one hand and a topography code on the other hand. This kind of approaches is used for instance in the SEER Program 2014-05-08 conversion file of ICD-O-3 to ICD-9-CM, to ICD-10 (Causes of Death) and to ICD-10-CM file (available at `http://seer.cancer.gov/tools/conversion/`). It offers mappings that enable to from ICD-O-3 to ICD-10. In this file for instance :

— *malignant adenocarcinoma of prostatic gland* (C61.9 - 8140/3) is mapped to *Malignant neoplasm of prostate* (C61).
— *malignant neoplams of prostatic gland* (C61.9 - 8000/3) is mapped to *Malignant neoplasm of prostate* (C61).

While these mappings are true, they are oriented and defined for the specific purpose of converting ICD-O-3 to ICD-10. Reversing the conversion process will lead to inconsistencies with multiple ICD-O-3 combinations mapped to a single ICD-10 code (here C61 mapped to both C61.9 - 8140/3 and C61.9 - 8000/3). A solution would be to specify that C61 is mapped to C61.9 - 8000/3. But this approach will result in loosing semantic of origin if going from ICD-O-3 to ICD-10 and then back to ICD-O3 (for instance converting C61.9 - 8140/3 to C61 and then C61 to C61.9 - 8000/3). On the other hand, our modeling approach enables to have an integrated view of terminology codes keeping their semantic fully available inside the model (because they are representing as is inside the model). The model can then evolve depending on the specific need without changing or suppressing the initial semantic of the terminology code. In Figure 5.4, codes are bind to classes that are linked through the model. One

can easily build set of tools that manipulate classes so that it is possible to implement methods on the top of these terminologies and if needed retrieve the initially coded information.

Architecture for integrating EHR and disease identification

We have identified two issues for secondary use of structured EHR in oncology. In order to address semantic heterogeneity, previous chapter was presenting an ontology based approach for semantic integration of heterogeneous diagnosis terminology in oncology.

Another issue need to be addressed, indeed as discussed before, disease is implicitly represented within EHRs. In the biomedical domain, phenotype retrieval often lead to searching for patient with a specified disease. These kind of queries or algorithm need an to have explicitly recorded disease and related information.

In this part, we present a complete architecture that address this issue. We propose to build a layered framework enabling to bind data with external knowledge resource. We then build an algorithm for disease identification based on conceptual representation of data and apply to the data. Our approach leverage multiple existing technologies and standards and combine them in a layered architecture.

Chapitre 6

Methods

6.1 Integrating EHR

We have proposed and implemented an architecture based on existing open-source tools. The main goal is to provide methods for semantic and syntactic integration of EHR in oncology. Although our work focuses on automated structured data processing, we have settled tools for a wider range of data types. The proposed architecture mainly relies on three layers (Figure 6.1) :

1. Data warehouse (storage) layer
2. Semantic integration layer
3. Rule base Neoplasm identifier layer

Each layer leverages existing tools and standards in order to manage its data and implements its methods. Methods presented focuses on our approach for combining these tools and provide a conceptual access to observations syntactically integrated.

6.1.1 Data warehouse (storage) layer and syntactic integration

The storage layer is a data warehouse solution. In the bio-medical domain clinical data warehouse (CDW) is often implemented as a solution for EHR integration [10]. Within this scope i2b2 team [39] has developed and implemented a CDW infrastructure for EHR integration. I2b2 is widely adopted by academic hospitals [21, 22] and research projects [7, 10].

I2b2 architecture

I2b2 infrastructure (so called the i2b2 Hive) consists of Cells providing services. Within the i2b2 Hive, two cells are of particular interest for data storage and retrieval :

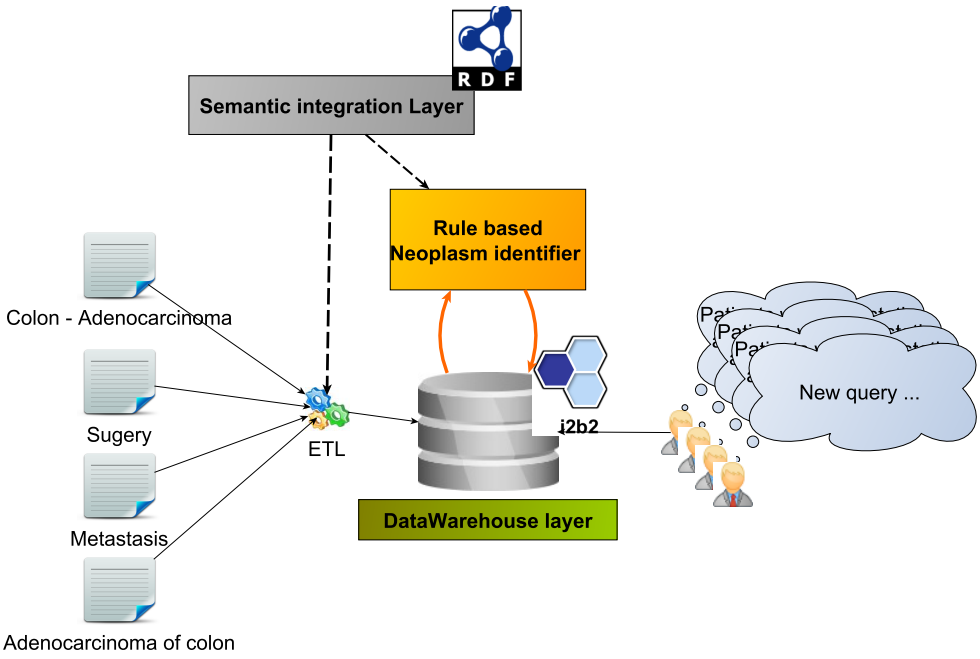

FIGURE 6.1 – Architecture for physical and semantic integrating of EHR in oncology.

— **CRC cell**. This cell aims at storing EHR data (CRC : Clinical Research Chart). It relies on *Entity-Attribute-Value* data model (i2b2 star schema - see figure 3.1). The data model enables to integrate almost all kind of data related to a patient and its interaction with a care provider. The CRC data are stored as observations. Observations can range from coded value (with *ad'hoc* or standard terminology) to numerical value and large free text reports. Within the CRC cell, only syntactic integration is managed. The meaning of observations stored in the Entity-Attribute-Value model is not available within this cell. However, semantic of origin is stored in the observation table using the *concept_cd* column.
Concept_cd is describe in i2b2 documentation as a "*code for the observation of interest (i.e. diagnoses, procedures, medications, lab tests)*". As a result it can represent an heterogeneous set of values ranging from an coded value (e.g. diagnosis code) to an open question (free text field in a form). The former corresponds to a recorded value whereas as the latter corresponds to a field without the effective value recorded. Indeed the value is stored in an other column in the observation table. An observation can be extended with a modifier. Modifier semantic of origin is stored in the *modifier_cd* column. *Modifier_cd* has the same characteristics as *concept_cd*, it can store both a value or a field.
I2b2 CRC cell web services provide methods for patient, visit and obser-

vation retrieval by querying observations. The research engine can manage complex query combining multiple observations (managing transparently data format) and taking into account temporal selection criteria if needed.

— **Ontology Cell**. This cell aims at storing meaning and syntax of concepts recorded in the CRC as observations. It stores hierarchical lattices, pointing to concepts representing observations as recorded in the CRC Cell. These lattice provide a way for aggregating sibling concepts in a broader concept so that they can be queried together transparently. The ontology cell can manage multi-hierarchical lattices so that a single concept can be accessed within multiple path. Within i2b2 lattices format, and query pattern are recorded for each concepts. These annotations provide information to the CRC query engine in order to build queries.

Despite of its name, the Ontology cell does not provide tools for ontology management. Hence, its aim is to pilot the i2b2 query engine and not to fully record observation's semantics. However, by separating, data storage (for syntactic integration) and data meaning (for semantic integration) I2b2 offers opportunities for building layered infrastructure for EHR integration

Extract Transform and Load Process

The Extract, transform and load (ETL) process aims at integrating EHR data into the i2b2 CDW. EHR are scattered in multiple databases within Hospital Information System (HIS). These databases correspond specific applications (called HIS dimension) adapted to care specific activities.

We have implemented extraction processes for three HIS dimensions recording oncology data :

— Reimbursement data
— Pathology data
— Clinical forms containing multidisciplinary staff for decision making in oncology.

These extraction processes are followed by a transformation process in order to provide a specific data format. These data are consumed by the load process in order to load data in the CDW.

During the ETL process we keep data semantic unchanged. Only syntactic adaptation operations are performed. Each observations are based on data element combined with a recorded value. This combination is represented and recorded within i2b2 CDW without meaning transformation. For instance :

— An ICD-10 diagnostic code from reimbursement data is recorded by combining the code itself and the hierarchical position of diagnosis within reimbursement data (primary, secondary or associated diagnosis)
— A free text within a form is recorded by combining the form identifier, the question identifier and the free text.

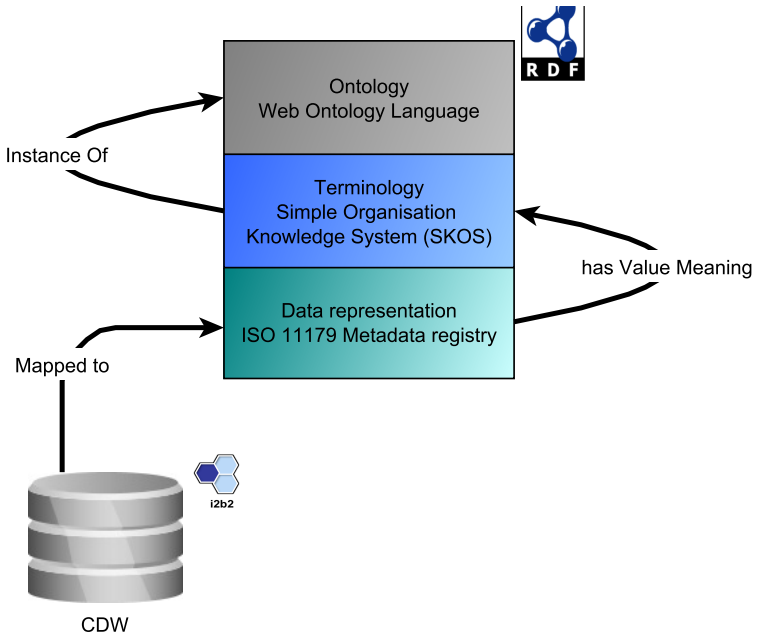

FIGURE 6.2 – Standard abstract models leveraged in the semantic layer and link between elements

— A coded value in a form is recorded by combining the form identifier, the question identifier and the code associated with the value.

6.1.2 Semantic integration layer

The semantic integration layer is the key component of the architecture. The main goal of this layer is to store and manage semantics of data elements and provide semantic services to other layer (namely Data warehouse layer and Rule base Neoplasm identifier layer). The semantic integration layer uses Resource Description Framework standard (RDF) to represent three kind of semantic resources :

— Data elements (representation of observations recorded within the CRC)
— Terminologies
— Ontologies

These resources are persisted using a triplestore. Our approach was to leverage existing standard for representing and linking : data representations, terminologies and ontologies. The semantic layer is based on three standard abstract models (Figure 6.2) :

— **ISO 11179 metadata registry (ISO/MDR)** [48]. This standard is used, within the semantic layer, in order to represent data as recorded in the CDW. Moreover, we use relationships provided by the model in order to bind data representations to conceptual level representation.

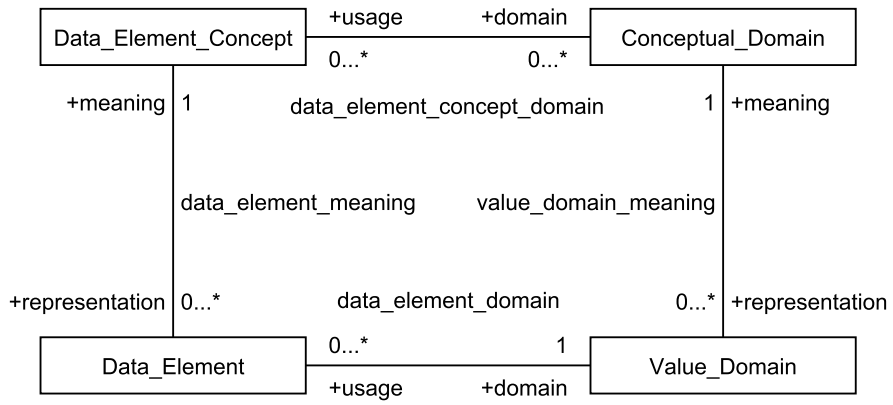

FIGURE 6.3 – ISO/IEC 11179 High-level Data Description metamodel [48]

— **Simple Knowledge Organization System (SKOS)**[47]. This language, extending OWL, is used in order to represent terminologies within the semantic layer.
— **Web Ontology Language (OWL)** [44, 45]. This language is used in order to represent ontologies within the semantic layer.

Data representation (ISO/IEC 11179)

Data representation is mainly based on ISO/IEC 11179. We have used the OWL representation of the ISO/IEC 11179 provided by the Semantic MDR developed in the SALUS project [19] (available on github). ISO/IEC 11179 specifies more than only data and concept description. However, our aim was to represent data and bind its semantics to conceptual representations. Thereby we focus here on data representations.

Figure 6.3, presents ISO/IEC 11179 High-level Data Description metamodel [48]. According to ISO/IEC 11179 documentation a "*data element is a basic unit of data of interest to an organization, for which the definition, identification, representation, and permissible values are specified by means of a set of attributes. Examples of data element include : a column in a table of a relational database, a field in a record or form, an XML element, the attribute of a Java class, or a variable in a program. The description of data elements is a major purpose of ISO/IEC 11179 Metadata Registries*". Following this definition we have instantiated the *dataElement* class with three kind of information depending on the integrated dimension :

— Column table for reimbursement data (e.g. diagnosis, procedures etc...) and pathology data (e.g. ADICAP code).
— Fields in clinical forms (e.g. patient history in the multidisciplinary staff report) .

As stated above, in i2b2, an observation can correspond to both a coded

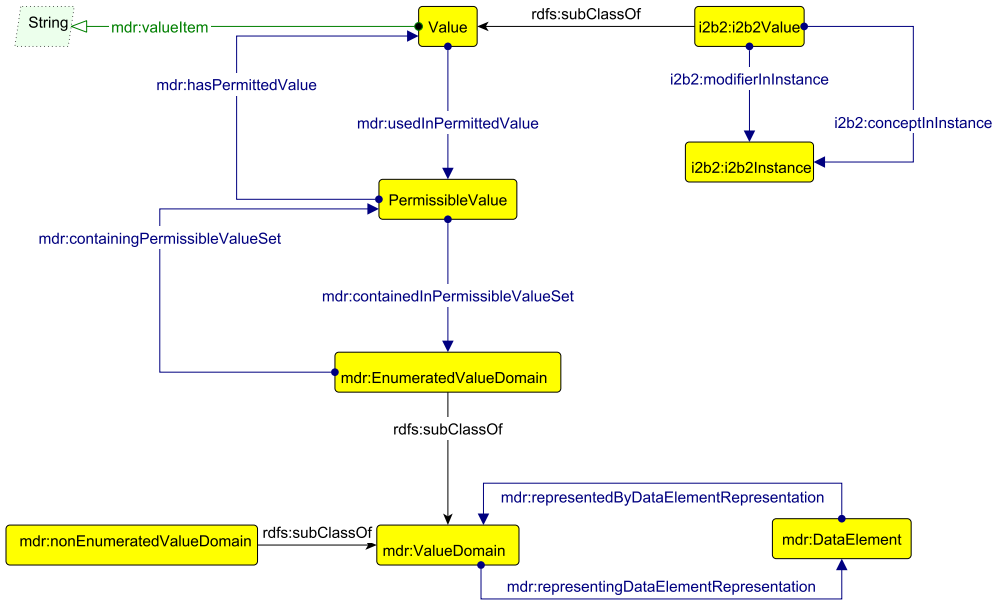

FIGURE 6.4 – ISO/IEC 11179 Value representation extended with i2b2 specific classes

value and a field. These types of observations are to be distinguished within ISO/IEC 11179 metamodel. Indeed the former can be considered as a *value* whereas the latter is a *dataElement*. However, they are all recorded as a *concept_cd* or *modifier_cd* in I2b2. As a result these observations are to be considered as types of *value* within i2b2's scope.

In order to manage observation enabling both i2b2 observation representation and HIS element description, we have describe each concept_cd and modifier_cd depending on the context. Our approach was to extend ISO/IEC 11179 metamodel with two classes (figure 6.4) :

- *I2b2Instance*. Individual of this class correspond to CRC concrete implementations.
- *I2b2Value*. Individual of this class correspond to the union of all concept_cd and modifier_cd in a CRC concrete implementation. *I2b2Value* is a *subclass of* the ISO/IEC 11179 *value* class.

Each *i2b2Value* is related to at least an i2b2Instance by two kind of relationships :

- *conceptInInstance*. Specifies that an *i2b2Value* is used as a *concept_cd* in an *i2b2Instance*.
- *modifierInInstance*. Specifies that an *i2b2Value* is used as a *modifier_cd* in an *i2b2Instance*.

These *i2b2Value* are then represented in two different ways depending on

their origin within the HIS :

- Represented as a *DataElement* when it corresponds to a *non enumerated* item (e.g. free text, lab result).
- Represented as a *Value* when it corresponds to a structured value (e.g. coded element, standard or non standard value list). In this situation, the observation is related to a *PermissibleValueSet* which is used by an *enumeratedValueDoamin* representing a *dataElement*.

Thereby, observations to be stored are all represented as being related to a *dataElement* or as being themselves a *dataElement*

Termino-ontological resources representation

Terminology representation

Terminologies are represented using SKOS [47]. The terminology is considered to be a *skos :conceptScheme* identified by an URI. Each term is considered to be a *skos :concept*. Each *skos :concept* is related to a *skos :conceptScheme* using *skos :inScheme*. Therefore, the complete terminology is represented by a set of *skos :concept* related to a *skos :conceptScheme*. Within each terminology, *skos :concept* are hierarchically organized using *skos :broader* and *skos :narrower* relationships. In addition, terminologies where represented inside the ISO/IEC 11179 model as *ConceptualDomain* and terms as *ValueMeaning*.

Needed SKOS representations were built automatically using available resources (e.g. UMLS for ICD-10, NCI metathesaurus for ICD-O-3) or manually when no resource were available (e.g. ADICAP). During the building process two versions of each terminology were built and stored in the triple store :

- **Simple version** : This version contains only stated relationships
- **Inferred version** : This version contains inferred SKOS relationships using Hermit reasoner based on OWL representation of the SKOS specification (available at `https://www.w3.org/2009/08/skos-reference/skos.rdf`).

Ontology representation

Needed ontologies were built (e.g. diagnosis and disease representation in chapter 5) and directly integrated in the triple store as OWL files. No specific modification was applied during the import process.

Linking data, terminology and ontology representations

As presented in figure 6.2 we use a bottom up approach going from observation recorded in i2b2 to formal domain model representation. Here we present only links for structured data elements that corresponds to coded information

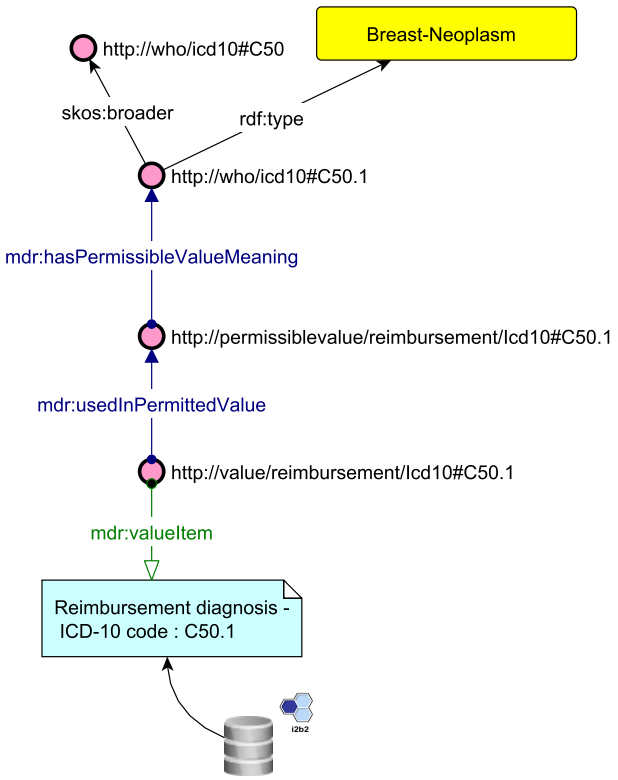

FIGURE 6.5 – Example of integration an ICD-10 diagnosis coded in reimbursement data. The observation is linked to a value within the ISO/IEC 11179 metamodel. The meaning of the observation is managed by terminologies represented in SKOS and instantiating the conceptual part of ISO/IEC 11179 metamodel. The terminology term can then instantiate a class of a specific ontology.

within the HIS. Figure 6.5 presents an example for integration of an ICD-10 diagnosis coded in reimbursement data

Each observation (namely a *concept_cd* or a *modifier_cd*) has its physical representation encoded using the *valueItem* ISO/IEC 11179 data property (figure 6.4). The corresponding *i2b2 :i2b2Value* is treated as a *Value* within the HIS scope of the metadata repository (because we focus on coded information). The *Value* is linked a *permissibleValue* with the *usedInPermittedValue* relationship. Following on ISO/IEC 11179 model (figure 6.3) *permissibleValue* is linked to the corresponding *valueMeaning* using the *permissible_value_meaning* relationship. As discussed above, *skos :concept* (representing terminology terms) are also instances of *valueMeaning* in ISO/IEC 11179 so that the *permissible_value_meaning* relationship is used to bind data representation with meaning representation.

The last stage enabling a formal representation of the domain correspon-

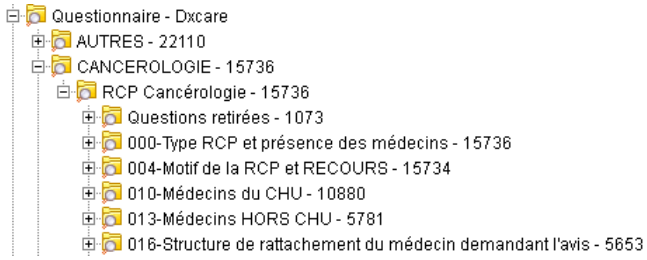

FIGURE 6.6 – Example of hierarchical lattice inside i2b2 representing clinical forms

ding to the recorded observation is to link terminology term with its class representation in an ontology. Our approach was to follow the W3C recommendations to define formal and semi-formal hybrid models [63] in order to build a model combining SKOS for terminologies description and OWL for representing concepts involved and defining of relationships between these concepts as proposed in [62].

6.1.3 Implemented methods based on the semantic layer

I2b2 metadata builder

Has stated above, the i2b2 query engine relies on metadata recorded in the *Ontology cell*. These metadata are structured as hierarchical lattices. Data within ontology cell describes :

- Hierarchical relationships between nodes
- Type of data recorded for each node (e.g. Numerical value, short free text, long text, structured value).
- Visual attributes of nodes.

Figures 6.6 presents an example of a lattice describing i2b2 nodes for clinical forms representation. The hierarchical relation in the lattice aims at grouping nodes when building query. Thus, when every instance of a node is considered to be an instance of the father node. For instance in figure 6.6 a query using the *RCP Cancérologie* node will retrieve all patient having at least one observation corresponding to one of its child node (e.g. *Motif de la RCP et RECOURS*).

Based on the semantic layer, we have implemented methods in order to automatically generate hierarchical lattices compatible with the i2b2 *Ontology Cell*. These lattice enable two different kinds of data access :

1. **Access through the HIS data structure**. In this situation, lattices are build for each dimension seperatly. For each HIS dimension, we retrieve the corresponding *dataElement*. Based on the *dataElement's valueDomain* we retrieve its type (Enumerated or not) and its format. For *EnumeratedValueDomain*, we search for a *ConceptualDomain* representing the *valueDomain*. The *ConceptualDomain* is then used to build a

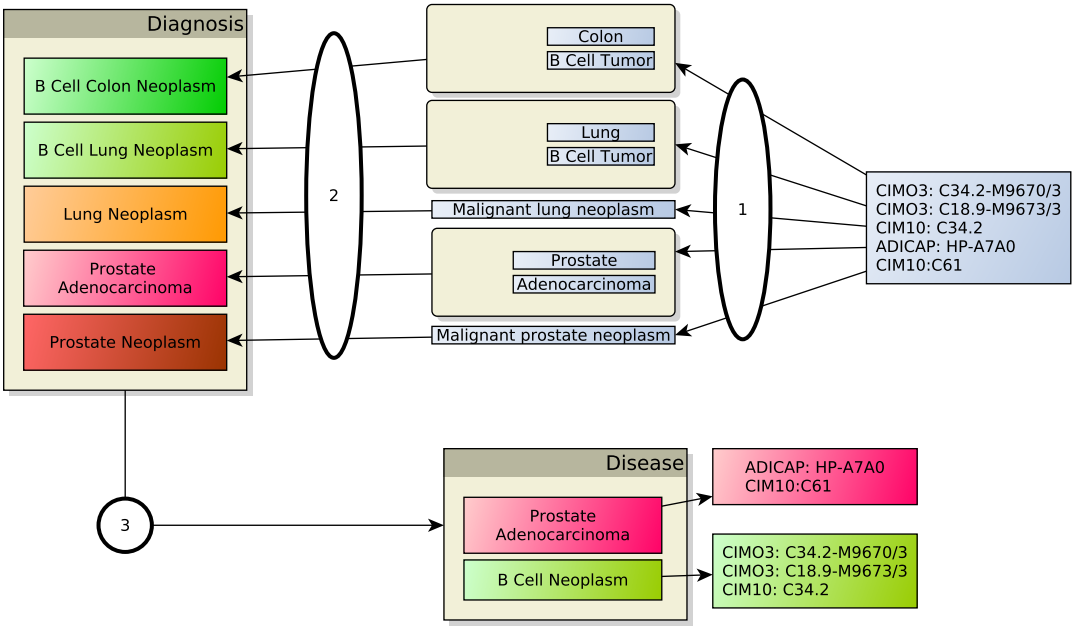

FIGURE 6.7 – Example of execution of the rule base neoplasm identifier. **Step 1** Identifies atomic element describing diagnosis. **Step 2** identifies diagnosis classes based on coded observation.**Step 3** identifies disease based diagnosis for a given patient. Disease are linked to observation that produced them.

lattice based on SKOS hierarchical relationships (namely *skos :broader* and *skos :narrower*). These skos lattices are built only for *permissibleValues* of the *valueDomain*. As a result we obtain a full lattice representing each types of observation and leveraging terminology hierarchies for coded *dataElements*.

2. **Integrated access for ICD-10 diagnosis**. In this situation we leverage terminological representation of ICD-10 and its relationship with ISO/IEC 11179 *valueItems* to provide a query tool for ICD-10 diagnosis through every integrated dimensions.

Rule base Neoplasm identifier layer

These methods aims at identifying possible disease arising for an individual based on diagnosis recorded within EHR. The neoplasm identifier, takes as input the set of all cancer diagnosis recorded for a patient. These diagnosis can be recorded using ICD-10, ICD-O-3 or ADICAP.

I2b2Values corresponding to tumor diagnosis (identified by the ontology) are retrieved for each patient. A diagnosis is to be a set of atomic observation (which can have a length of 1). An atomic observation may be a diagnosis code or a morphology code (recorded alone) or a topography morphology combination (recorded together as a diagnosis description within data sources). As a

result tumor diagnosis may be sent to the algorithm as set containing one or more atomic observation of all types). A 3 steps algorithm has been implemented as follow :

1. **Identify classes corresponding to recorded observations** (step 1 in figure 6.7). Each atomic code is bind to the corresponding class though the semantic layer as shown in figure 6.5. Through the ontology these atomic codes are classified as *diagnosis*, *morphology* or *Topography* (high level types).
2. **Build diagnosis based on atomic observations** (step 2 in figure 6.7). Each classified atomic observation, is used to build DL Queries depending on its high level type (composition) :
   — If the tumor diagnosis contains *morphologies*, then the diagnosis to retrieve is described as having the specified *morphologies*.
   — If the tumor diagnosis contains *Anatomic sites*, then the diagnosis to retrieve is described as having the specified *Anatomic sites* as primary sites.
   — If the tumor diagnosis contains *diagnosis*, then the diagnosis to retrieve is described as being a kind of the specified *diagnosis*.

   The obtained DL-Query is executed on the model in order to retrieve the corresponding diagnosis.
3. **Identify disease based on diagnosis** (step 3 in figure 6.7). For each diagnosis retrieved, a DL-Query is built executed in order to retrieve disease described by the corresponding diagnosis. If diseases subsumption exists between retrieved disease, only the deepest disease is retained and the diagnosis is related to it.

6.2 Implementation

The full architecture was implemented inside Bordeaux University hospital's HIS. The implementation environment was settled as follow :

— I2b2 was deployed with a Oracle 11g® database backend.
— ETL were implemented using Talend Open Studio®
— The semantic layer was build on the top of Apache Marmotta© with PostgreSQL based Kiwi-triple store backend.
— SKOS serialization for ontologies were performed using OWL API 3.4.8 [78].
— Neoplasm identifier and metadata builder were implemented using OWL API 3.4.8 [78] and HermiT reasonner for inferences and DL-query performing [79].

6.3 Evaluation

6.3.1 Data used

We have loaded a set of the three integrated HIS dimension (namely reimbursement data, pathology data and multidisciplinary staff data). Data were including every information within these dimension concerning patient having at least a cancer diagnosis code. This set included structured and unstructured data.

6.3.2 I2b2 metadata builder

We have then built *ontology cell* metadata for the data warehouse based on the semantic layer. These lattice were then tested to query multiple type of data.

6.3.3 Rule base neoplasm identifier

Gold standard

We have used the Gironde solid tumor registry as a gold standard in order to evaluate performances of the rule base neoplams identifier. The Gironde solid tumor cancer registry, records malignant solid tumors of adult patients (more than 18 years old at diagnosis). This registry records malignant solid tumor occurring for patient living in Gironde at the date of diagnosis. The registry excludes tumors of haematopoietic and lymphoid tissues and tumors of central nervous system because of the existence of two specialized registry covering the same territory.

We have extracted data of the Gironde solid tumor cancer registry for tumors occurring in 2013 and corresponding to patient having visited in Bordeaux university hospital. These data were used as a gold standard.

Evaluation methods

Based on data available on the CDW we have used the neoplams identifier algorithm in order to built tumors for patient having at least cancer diagnosis referred during 2012, or 2013, 2014. Among identified neoplasm we excluded :

— Non malignant tumors
— Tumors corresponding only to metastasis diagnostic code
— Tumors referred by at least a diagnosis recorded before June 2012
— Tumors referred by only diagnosis recorded after June 2014
— Tumours of haematopoietic and lymphoid tissues
— Tumors of Central nervous system

Based on the cancer registry diagnostic codes (coded using ICD-O-3), we have build the corresponding IARC disease. The obtained data sets were merge by patient identifiers in order to produce performance metrics. For each individual a sparse binary matrix was constructed representing presence or absence of each type of tumour. From this matrix, the contingency table was derived for each type of tumour so as to calculate locally recall (sensitivity) and precision (positive predictive value). The F-measure, the harmonic mean of recall and precision, was deduced from these values [80]. These measures were produced for :

- Topography identification
- Morphology identification
- Disease identification declined with the 6 most frequent tumors retrieved in cancer registry data (namely Lung - Adenocarcinoma, Prostate - Adenocarcinoma, Thyroid - Other carinoma, Skin - Squamous carcinoma, Colon - Adnocarcinoma, Breast - Adnocarcinoma)

Chapitre 7

Results

7.1 Data integrated

Table 7.1 presents the data integrated in the i2b2 data warehouse for evaluation purpose. A total of 95 969 patients were integrated, corresponding to 418 163 visits. Among these patient 49 196 (51,3%) were men and 6 947 (7,2%) were recorded as deceased in the HIS.

A total of 12 536 256 records were integrated in the CDW. These data were corresponding to 8 471 130 observations (44,9% of reimbursement data, 1,2% of pathology data and 53,8% of clinical form data) modified by 4 065 126 modifiers. Figure 7.1 presents the distribution of the number of observations per patient integrated in the CDW. Mean number of observations per patient was 88,27.

7.2 Semantic layer

Table 7.2 presents data integrated in the semantic layer. A total of 1 552 237 triples were used to implement the semantic layer. These triples were describing :

— 578 *dataElements* (99,5% for clinical form description).
— 578 *valueDomains* enumerated in half (51,6%).
— 29 617 *permissibleValues*.

A total of 8 terminologies were integrated as *skos:ConceptScheme* corresponding to 47 109 *skos:concept* :

— WHO version of ICD-10 (12 301 *concepts*)
— Bordeaux university hospital version of ICD-10 (including morphology codes and specific ICD-10 subdivisions - 20 683 *concepts*)
— ADICAP Lesion (2 451 *concepts*)
— ADICAP Organe (175 *concepts*).

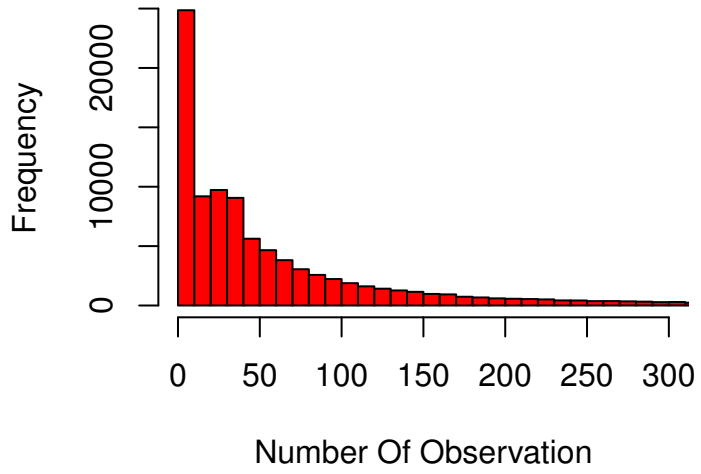

FIGURE 7.1 – Distribution of the number of observation per patient

	N	Percent
Total		
Patient	95 969	
Visit	418 163	
Observation	8 471 130	
Clinical forms		
Patient	60 677	63,2
Visit	208 076	49,8
Observation	4 559 763	53,8
Reimbursement data		
Patient	67 492	70,3
Visit	285 534	68,3
Observation	3 802 747	44,9
Pathology data		
Patient	59 213	61,7
Visit	81 619	19,5
Observation	97 533	1,2

TABLE 7.1 – Data integrated in i2b2 data warehouse

	N	Percent
DataElement	**578**	
Clinical forms	575	99,5
Reimbursement data	2	0,3
Pathology data	1	0,2
valueDomain	**578**	
Enumerated	298	51,6
Described (non enumerated)	280	48,4
PermissibleValue	**29 617**	
Clinical forms	13 109	44,3
Reimbursement data	14 652	49,5
Pathology data	1 856	6,3
***ValueMeaning* (skos concept)**	**16 066**	

TABLE 7.2 – Data integrated in the semantic layer

ConceptScheme	***ValueMeaning***	***Permissible***	***ValueDomain***
ADICAP Lesion	1 267	1 267	1
ADICAP Organe	145	145	1
CCAM	4 238	4 238	1
ICD-10 (Bordeaux HIS)	10 410	16 722	23
Data Sources	6	6	2

TABLE 7.3 – Number of *skos:Concept* used as *ValueMeaning* and *permissibleValues* bind to these *ValueMeanings* depending on the *skos:ConceptScheme*

— CCAM (Classification Commune des Actes Médicaux - http://www.ameli.fr/accueil-de-la-ccam/index.php - 9 990 *concepts*).
— ICD-O-3 Topography (410 *concepts*)
— ICD-O-3 Morphology (1 092 *concepts*)
— Data sources for diagnosis codes (an *ad'hoc* terminology developed for representing data sources of diagnostic codes - 7 *concepts*)

A total of 16 066 of these *skos:concepts* were used as *ValueMeaning*. Table 7.3 presents the number of *skos:Concept* used as *ValueMeaning* and *permissibleValues* bind to these *ValueMeanings* depending on the *skos:ConceptScheme*. Among these *ValueMeaning* Bordeaux's HIS version of ICD-10 was covering 23 *ValueDomains* and its *skos:concept* where bind to 16 722 *premissibleValues*. Among the overall 29 617 *PermissibleValue*, 22 378 (75,6%) were bind to a *ValueMeaning*. Among the 298 *EnumeratedValueDomain*, 28 (9,4%) were covered by these *conceptualDomain*.

	Clinical*	Adm**	Pathology***
Node	**16 994**	**18 254**	**1 575**
Class node	4 220	4 039	138
Leaf node	12 774	14 215	1 437
Structured node	16 714	18 254	1 575
Unstructured node	280	0	0
Large string node (Blob)	116	0	0
String node	56	0	0
Numerical node	77	0	0

TABLE 7.4 – Number of node built by the metadata feeder based on the semantic layer depending on represented HIS dimension. ***Clinical** : Clinical forms data ****Adm** : Administrative reimbursement data *****Pathology** : Pathology data

7.3 I2b2 metadata builder

Metadata integrated in the semantic layer were used to build *i2b2 ontologies* for the three integrated HIS dimensions (namely reimbursement data, pathology data and clinical forms).Table 7.4 presents the number of nodes for each HIS dimension *i2b2 ontology* representation. The majority of nodes were corresponding to structured data. Clinical Form data representation encompasses both structured and unstructured data. As a result the corresponding *ontology* incorporate a broad range of data types representation (Large String, String, and numerical values). On the other hand reimbursement data and pathology data corresponds only to structured data relying on terminologies. *DataElements* used in order to evaluate the i2b2 metadata feeder cover a large number of nodes and a broad range of data types. The process for building i2b2 ontology was fully automatic based on data available in the semantic layer.

Figure 7.2 presents an example of a clinical form *i2b2 ontology* representation. In this example, the hierarchy is built for a *dataElements* recording ICD-10 diagnosis code. As a result, the metadata feeder uses *valueMeanings* bind to *permissibleValues* in order to build the lattice corresponding to skos hierarchical structure. Thereby each *permissibleValue* can be accessed through ICD-10 lattice within the *i2b2 Ontology*. The proposed method enables to combine multiple models (such as terminology and metadata registry) in order to build a consistent hierarchical lattice.

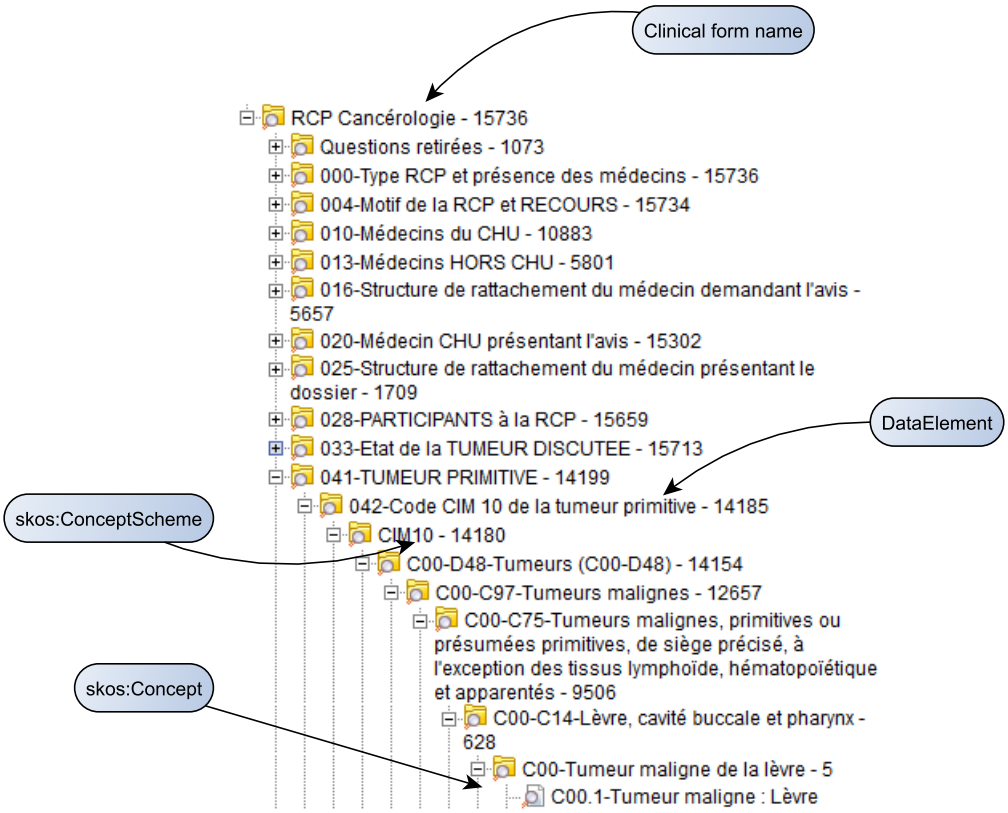

FIGURE 7.2 – Example of clinical form i2b2 *ontology* built based on the semantic layer. The root node is the clinical form itself. For each *DataElement* of the clinical form, a node is built. When the *DataElement* has an *EnumeratedValueDomain* related to a *ConceptualDomain*, skos hiearchy is used to build the lattice for accessing *valueMeaning* nodes (*skos:Concept*)

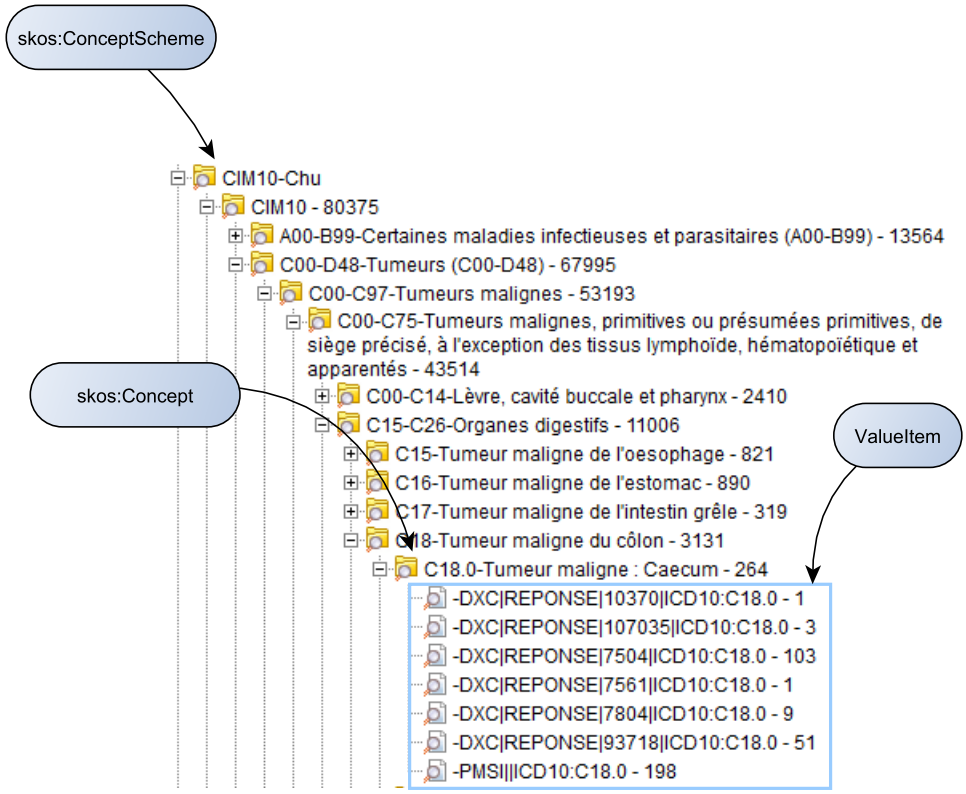

FIGURE 7.3 – Example of an i2b2 ontology built *"on the fly"*. the strategy is to start from ICD-10 (Bordeaux University hospital version). I2b2 ontology is built based on SKOS hierarchical lattice of ICD-10 *skos:ConceptScheme* and bind to *ValueItem* through ISO/IEC 11179 metamodel. As a result, this ontology can query transparently any *dataElement* recorded using ICD-10 within the HIS.

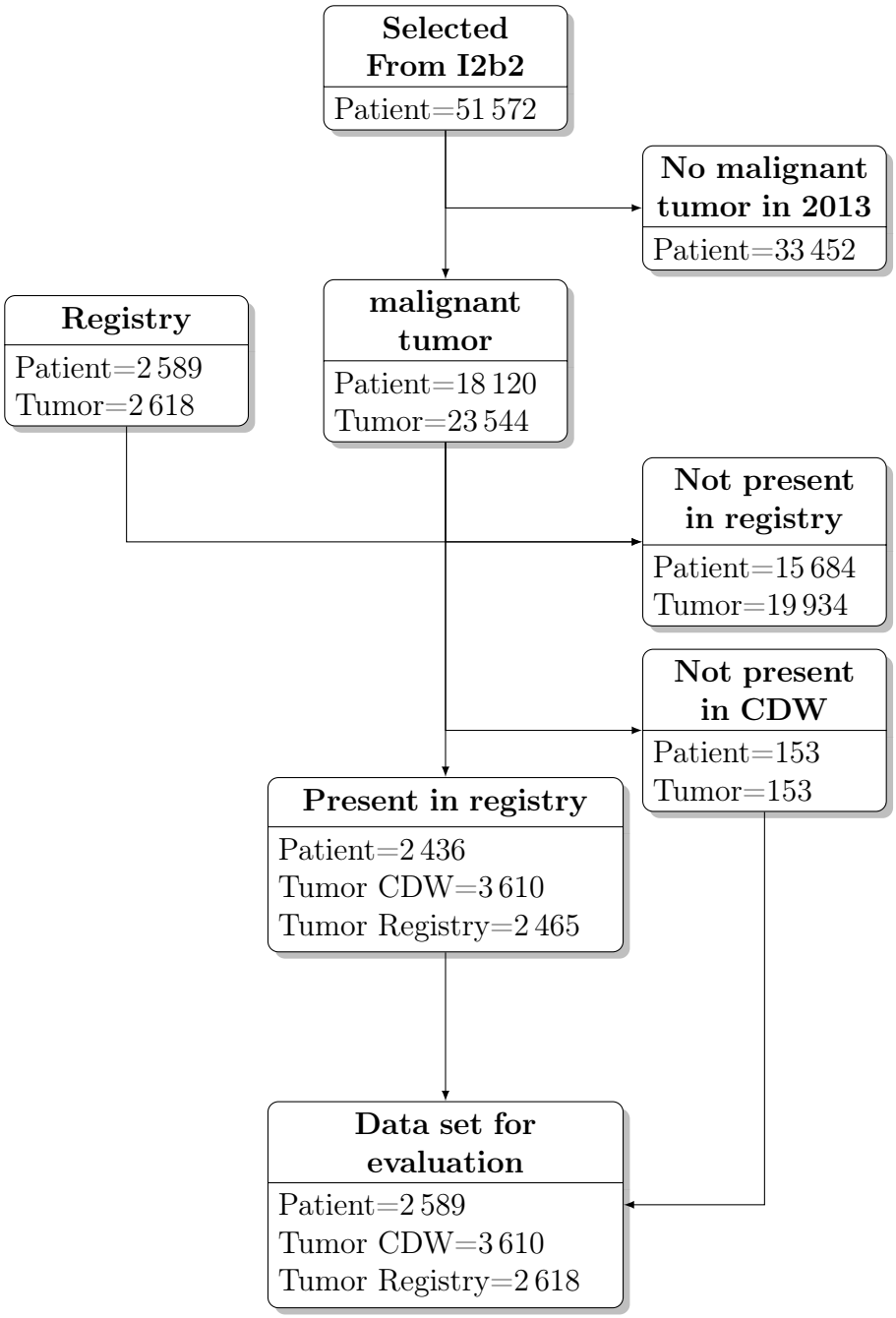

FIGURE 7.4 – Flow chart for patient of extracted from the clinical data ware house (CDW).

	Rappel	Precision	F-measure
Topography	0,89	0,67	0,76
Morphology	0,75	0,62	0,68
Disease	0,63	0,46	0,53
Lung - Adenocarcinoma (247)	0,66	0,80	0,72
Prostate - Adenocarcinoma (203)	0,86	0,97	0,91
Thyroid - Other carinoma (184)	0,45	0,96	0,61
Skin - Squamous (159)	0,87	0,90	0,88
Colon - Adnocarcinoma (136)	0,70	0,80	0,75
Breast - adenocarcinoma (110)	0,43	0,98	0,59

TABLE 7.5 – Evaluation of the rule base neoplasm identifier vs cancer registry data

7.4 Rule based neoplasm identifier

7.4.1 Data used

Figure 7.4 presents the flow chart of patient selected for neoplasm identifier evaluation. A total of 51 572 patients were identified as having at least a tumor coded within the CDW between 2012 and 2014. Among these patients, 18 120 were retrieved by the neoplasm identifier as having a malignant tumor in 2013. A total of 2 589 patients were extracted from the cancer registry corresponding to 2 618 tumors. A total of 2 436 patients were retrieved in the two data sets corresponding to 3 610in the CDW data set and 2 465 in the registry data set.

Combining these data sets an evaluation was built. This data set were including the 2 589 patients of the cancer registry (corresponding to 2 618 tumor recorded by the cancer registry and 3 610 tumor identified by the rule based neoplasm identifier.

7.4.2 Evaluation of the rule based neoplasm identifier

Table 7.5 presents evaluation metrics for the rule based neoplasm identifier. Using the IACR model, the neoplasm identifier was able to identify topographies and morphologies with a F-measure of 0,76 and 0,68 respectively. Identifying disease (as a combination of topography and morphology) was leading to a 0,53 F-measure. Looking at the 6 more frequent tumors, the Performances were highly variable depending on the tumor (ranging from a 0,72 F-measure for lung adenocarcinomas to a 0,91 F-measure for prostate adenocarcinomas.

Chapitre 8

Discussion

8.1 Architecture for Integration

We have proposed an architecture based on three layer :

— **Storage layer**. Based on i2b2 CDW, with a built in ETL strategy.

— **Semantic layer**. Using web semantic standards, we have adapted and bind models for metadata registration (ISO/IEC 11179), terminology representation (SKOS) and ontology representation (OWL).

— **Neoplasm identifier**. A rule based algorithm which uses a model in order to identify possible diseases depending on coded diagnosis.

Our architecture enables integration of a broad range of HIS data (from structured data coded with a terminology to unstructured free text record). Both syntactic and semantic integration are performed during the process. Syntactic integration capability is mainly due to i2b2 CDW ability to integrate these kind of heterogeneous data in the bio-medical domain. Semantic integration capability is performed using semantic web technologies and standards.

By using ISO/IEC 11179, we bind data values and data elements to terminology concepts representing them. Moreover, terminology concepts can be bind to formal model (ontology) enabling reasoning and DL-query performing. Based on the semantic layer we have developed a process enabling automatic *i2b2 ontology* building based on semantic layer's metadata representation. With such a process, *i2b2 ontology* can be built *"on the fly"*. As a result we were able to build *i2b2 ontologies* representing HIS elements (figure 7.2). Moreover this approach can be used to build multiple hierarchical lattices representing the same set of data. Following this principle, multiple aggregation strategies can be proposed to users without changing data registration. This approach can be used for instance in order to query transparently a broad range of data elements described with the same terminology as shown in the example based on ICD-10 figure 7.3.

Previous work have successfully leveraged semantic web technologies for

phenotyping purpose [49, 51]. These works focused on modeling purpose and proposed to build specific architectures for binding ontologies to data. Here we focus on using and combining existing standards in order to bridge data to concepts represented in ontologies. Thereby the semantic layer is a generic solution and can be used with other CDW systems. As we used SALUS's ISO/IEC 11179 OWL representation, both data and conceptual representation can be shared with other ISO/IEC 11179 implementation.

8.2 Rule based neoplasm identifier

We have developed an algorithm based on IARC classes represented using a formal ontology. The main goal of this work was to propose a tool for semantic integration and possible disease identification. Moreover, the algorithm relates diagnostic codes with identified disease. Performance of the neoplasm identifier are similar with the baseline in [25]. In this previous work we were using data from multiple care centers so that data were more complete. Indeed here we used data from only one care center leading to incomplete care data when patient is treated in multiple care centers. For instance pathology data may not be available if the exam was performed outside of the care center.

This approach enables to build disease centered data based on diagnosis recorded within EHR. This structure is in conformance with secondary use needs. For instance one can easily use data related to disease in order to build a phenotype query.

In this work we focused on semantic, avoiding noise management or a full phenotyping framework. Our aim was to provide solutions for semantic integration and disease (as an underlying implicit process occurring for a patient) identification. We have implemented an approach based on IARC rules formal representation. The developed algorithm only uses classes hierarchical structure in order to identify links between coded diagnosis. As a result, modifying the formal model (e.g. adding narrower morphology classes) can be easily settled. However, it is noteworthy that the more precise the class are, the more difficult the task would be.

Our work is complementary with previous work on secondary use. Selection methods [25] for noise reduction,supervised [32] and unsupervised [24] machine learning approaches for phenotype identification could be implemented using external knowledge for disease representation, terminology integration and diagnosis aggregation. Moreover, while identifying the disease we have linked it with diagnosis describing them. As a result, all data related to visit corresponding to the diagnosis can be used as descriptor of the disease.

Chapitre 9

Conclusion and perspectives

We have identified two issues that have to be addressed in order to facilitate secondary use of EHR data in the oncology field. We have proposed a complete architecture that aims at addressing these issues.

Previous work have focus non phenotyping purpose. Approaches ranging from manual algorithm settings [38] to fully unsupervised phenotype identification [24] have been proposed. In the oncology field, methods have been developed for tumor identification [25, 26, 55, 32].

While these methods focuses on data, our work, focuses on semantic integration and data representation topics. Diagnosis terminology heterogeneity and disease implicit representation have to be addressed. Indeed significant part of research queries rely on the existence a particular disease and observation related to it. We have proposed a method to address these issues by binding knowledge on the top of data. Semantic web technologies and methods are of particular interest in this context. Previous work have leveraged semantic web for phenotyping purpose [81, 51, 82, 49]. These methods are modeling specific clinical situations in order to classify patient depending on their EHRs. In contrast, we have proposed to build a more generic domain model, representing diagnosis and disease in oncology. The aim of this model is not to classify patient but to identify the implicit disease. Obtained result can therefore be used as input features for phenotyping purpose.

Our architecture aims at binding data, terminologies that represent them and external knowledge. This kind of architecture was proposed by Fernandez-Breis et al. in [49]. Our approach is similar in that we aims at performing *"each activity at the abstraction level with the most appropriate technology"* available . However, our implementation of this principle is slightly different. Indeed (i) we propose to use ISO/IEC 11179 as a model for linking data and terminologies while they implement direct mapping for this purpose ; (ii) Moreover we do not bind directly data to OWL model but we use skos serialization of terminologies to support this link. Thereby the resulting semantic layer is a rich knowledge resource providing knowledge about domain, data and data representation,

This knowledge can therefore (as we demonstrate) be used for different task such as :

- i2b2 ontology building (providing multiple views of a single set of data)
- Ontology based semantic integration of date represented with heterogeneous terminologies.
- Algorithm implementation based on formal model (Leveraging DL-reasoning and applying to data rule defined at the conceptual level).

We have proposed a method adapting EHRs for secondary use in oncology, however further work is needed in order to fully enable secondary use in oncology. We have modeled diagnosis representation but other elements related to identified disease also need to be modeled (*i.e.* treatments, clinical courses ...). Data produced need to be validated so that link with existing research databases is necessary for cross validation. Moreover, phenotype algorithm or contextual selection of relevant information given a disease should be evaluated using disease centered data as input feature.

Annexe A

Publications

Publications en lien avec le travail

Jouhet V, Defossez G, Ingrand P. Automated Selection of Relevant Information for Notification of Incident Cancer Cases within a Multisource Cancer Registry. Methods Inf Med. 24 avr 2013 ;52(4).

Bigeard E, Jouhet V, Mougin F, Thiessard F, Grabar N. Automatic extraction of numerical values from unstructured data in EHRs. Stud Health Technol Inform. 2015 ;210 :50-4.

Toulmonde M, Ducimetière F, Jouhet V, Gaudin T, Malfilatre A, Laizet Y'han, et al. Les grandes bases de données nationales et les études dans la vraie vie : l'exemple des sarcomes en France. Bulletin du Cancer. juin 2016 ;103(6, Supplement 1) :S71-5

Jouhet V, Mougin F, Brechat B, Thiessard F. Building a model for disease classification integration in oncology. An approach based on the National Cancer Institute thesaurus. Journal of Biomedical semantics (accepté).

Communications lors de congrès avec commité de lecture

Communications orales

Jouhet V, Bréchat B, Mougin F, Thiessard F. Intégration de terminologies diagnostiques en cancérologie. Revue d'Épidémiologie et de Santé Publique. sept 2014 ;62 :S185.

Jouhet V, Bréchat B, Mougin F, Thiessard F. Intégration de terminologies diagnostiques en cancérologie : le NCI thésaurus comme pivot ? Revue d'Épidémiologie et de Santé Publique. 2014 ;62 :S125–S126.

Brechat, Bérénice, Fleur Mougin, Frantz Thiessard, et Vianney Jouhet. «Mapping de terminologies diagnostiques en cancérologie par l'intermédiaire du NCI Metathesaurus». In Actes des Proceedings of 15es Journées francophones d'informatique médicale co-organisées avec co-located with 2e Congrès National d'Informatique Médicale (CNIM 2014), Fes, Morocco, June 12th-13th, 2014, édité par Pierre Zweigenbaum et Cheick Oumar Bagayoko, 1379 :34–43. CEUR Workshop Proceedings. CEUR-WS.org, 2014. http://ceur-ws.org/Vol-1379/paper-04.pdf.

Posters

Jouhet V, Amadeo B, Maurisset S, Tranchet E, Mathoulin-Pélissier S, Coureau G. Système d'information pour l'enregistrement des cas incidents de cancer–Déploiement d'une solution permettant d'optimiser le traitement automatique de l'information. Revue d'Épidémiologie et de Santé Publique. 2014 ;62 :S138.

Jouhet V, Renaud-Salis J-L, Gaudin T, Malfilatre A, Mathoulin-Pélissier S, Coindre J-M. Plate-forme pour l'intégration et l'exploitation des bases de données clinico-biologiques du groupe Sarcome Français. Revue d'Épidémiologie et de Santé Publique. 2014 ;62 :S135–S136.

Cossin S, Amadeo B, Jouhet V, Maurisset S, Mathoulin-Pélissier S, Coureau G. Géocodage et localisation spatio-temporelle des cas de cancer : analyse exploratoire, Gironde. Revue d'Épidémiologie et de Santé Publique. 2013 ;61 :S320–S321.

Autres publications

Petit-Monéger A, Saillour-Glénisson F, Nouette-Gaulain K, Jouhet V, Salmi L-R. Comparing Graphical Formats for Feedback of Clinical Practice Data. A Multicenter Study among Anesthesiologists in France. Methods Inf Med. 7 oct 2016 ;55(6).

Annexe B

Financement obtenus en lien avec le sujet

iBCB - Integrating Biological and Clinical data for Biobanks (Appel à projet CRB IBiSA).

InB2 - Integrating Brio's BCB (SIRIC BRIO)

Bibliographie

[1] H U Prokosch and T Ganslandt. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. *Methods of information in medicine*, 48(1) :38–44, 2009.

[2] Travis B Murdoch and Allan S Detsky. The inevitable application of big data to health care. *JAMA : the journal of the American Medical Association*, 309(13) :1351–1352, April 2013.

[3] Charles Safran, Meryl Bloomrosen, W Edward Hammond, Steven Labkoff, Suzanne Markel-Fox, Paul C Tang, Don E Detmer, and Expert Panel. Toward a national framework for the secondary use of health data : an American Medical Informatics Association White Paper. *Journal of the American Medical Informatics Association : JAMIA*, 14(1) :1–9, February 2007.

[4] Taxiarchis Botsis, Gunnar Hartvigsen, Fei Chen, and Chunhua Weng. Secondary Use of EHR : Data Quality Issues and Informatics Opportunities. *AMIA Summits on Translational Science Proceedings*, 2010 :1–5, March 2010.

[5] Cynthia Barton, Crystal Kallem, Patricia Van Dyke, Donald Mon, and Rachel Richesson. Demonstrating "Collect once, Use Many" – Assimilating Public Health Secondary Data Use Requirements into an Existing Domain Analysis Model. *AMIA Annual Symposium Proceedings*, 2011 :98–107, 2011.

[6] AbdenNaji El Fadly, Bastien Rance, Noël Lucas, Charles Mead, Gilles Chatellier, Pierre-Yves Lastic, Marie-Christine Jaulent, and Christel Daniel. Integrating clinical research with the Healthcare Enterprise : from the RE-USE project to the EHR4cr platform. *Journal of biomedical informatics*, 44 Suppl 1 :S94–102, December 2011.

[7] Georges De Moor, Mats Sundgren, Dipak Kalra, Andreas Schmidt, Martin Dugas, Brecht Claerhout, Töresin Karakoyun, Christian Ohmann, Pierre-Yves Lastic, Nadir Ammour, Rebecca Kush, Danielle Dupont, Marc Cuggia, Christel Daniel, Geert Thienpont, and Pascal Coorevits. Using electronic health records for clinical research : The case of the EHR4cr project. *Journal of Biomedical Informatics*, 53 :162–173, 2015.

[8] Ioana Danciu, James D. Cowan, Melissa Basford, Xiaoming Wang, Alexander Saip, Susan Osgood, Jana Shirey-Rice, Jacqueline Kirby, and Paul A. Harris. Secondary use of clinical data : the Vanderbilt approach. *Journal of Biomedical Informatics*, 52 :28–35, December 2014.

[9] Ruth Nalichowski, Diane Keogh, Henry C. Chueh, and Shawn N. Murphy. Calculating the benefits of a Research Patient Data Repository. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, page 1044, 2006.

[10] Bastien Rance, Vincent Canuel, Hector Countouris, Pierre Laurent-Puig, and Anita Burgun. Integrating Heterogeneous Biomedical Data for Cancer Research : the CARPEM infrastructure. *Applied Clinical Informatics*, 7(2) :260–274, May 2016.

[11] Christopher G Chute, Jyotishman Pathak, Guergana K Savova, Kent R Bailey, Marshall I Schor, Lacey A Hart, Calvin E Beebe, and Stanley M Huff. The SHARPn Project on Secondary Use of Electronic Medical Record Data : Progress, Plans, and Possibilities. *AMIA Annual Symposium Proceedings*, 2011 :248–256, 2011.

[12] Roger J Black, L. Simonato, H. H. Storm, E. Démaret, and International Agency for Research on Cancer. *Automated data collection in cancer registration*. IARC, Lyon, 1998.

[13] E N MacKay and A H Sellers. The Ontario cancer incidence survey, 1964-1966 : a new approach to cancer data acquisition. *Canadian Medical Association journal*, 109(6) :489 passim, September 1973.

[14] V Jouhet, G Defossez, and P Ingrand. Automated Selection of Relevant Information for Notification of Incident Cancer Cases within a Multisource Cancer Registry. *Methods of information in medicine*, 52(4), April 2013.

[15] Casey Lynnette Overby, Peter Tarczy-Hornoch, James I. Hoath, Ira J. Kalet, and David L. Veenstra. Feasibility of incorporating genomic knowledge into electronic medical records for pharmacogenomic clinical decision support. *BMC Bioinformatics*, 11(9) :1–9, 2010.

[16] K. Stephen Suh, Sreeja Sarojini, Maher Youssif, Kip Nalley, Natasha Milinovikj, Fathi Elloumi, Steven Russell, Andrew Pecora, Elyssa Schecter, and Andre Goy. Tissue banking, bioinformatics, and electronic medical records : the front-end requirements for personalized medicine. *Journal of oncology*, 2013, 2013.

[17] Richard H. Scheuermann, Werner Ceusters, and Barry Smith. Toward an Ontological Treatment of Disease and Diagnosis. *Summit on Translational Bioinformatics*, 2009 :116–120, March 2009.

[18] Tim Berners-Lee, James Hendler, Ora Lassila, and others. The semantic web. *Scientific american*, 284(5) :28–37, 2001.

[19] A. Anil Sinaci and Gokce B. Laleci Erturkmen. A federated semantic metadata registry framework for enabling interoperability across clinical research and care domains. *Journal of Biomedical Informatics*, 46(5) :784–794, October 2013.

[20] Anand Kumar and Barry Smith. Oncology ontology in the NCI thesaurus. In *Artificial Intelligence in Medicine*, pages 213–220. Springer, 2005.

[21] Eric Zapletal, Nicolas Rodon, Natalia Grabar, and Patrice Degoulet. Methodology of integration of a clinical data warehouse with a clinical information system : the HEGP case. *Studies in health technology and informatics*, 160(Pt 1) :193–197, 2010.

[22] T. Ganslandt, S. Mate, K Helbing, U. Sax, and H.U. Prokosch. Unlocking Data for Clinical Research – The German i2b2 Experience. *Applied Clinical Informatics*, 2(1) :116–127, March 2011.

[23] George Hripcsak and David J. Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association : JAMIA*, 20(1) :117–121, January 2013.

[24] Rimma Pivovarov, Adler J. Perotte, Edouard Grave, John Angiolillo, Chris H. Wiggins, and Noémie Elhadad. Learning probabilistic phenotypes from heterogeneous EHR data. *Journal of Biomedical Informatics*, 58 :156–165, 2015.

[25] V Jouhet, G Defossez, A Burgun, P le Beux, P Levillain, P Ingrand, and V Claveau. Automated classification of free-text pathology reports for registration of incident cases of cancer. *Methods of information in medicine*, 51(3) :242–251, 2012.

[26] Giovanna Tagliabue, Anna Maghini, Sabrina Fabiano, Andrea Tittarelli, Emanuela Frassoldi, Enrica Costa, Silvia Nobile, Tiziana Codazzi, Paolo Crosignani, Roberto Tessandori, and Paolo Contiero. Consistency and accuracy of diagnostic cancer codes generated by automated registration : comparison with manual registration. *Population Health Metrics*, 4 :10, September 2006.

[27] Nicole Gray Weiskopf and Chunhua Weng. Methods and dimensions of electronic health record data quality assessment : enabling reuse for clinical research. *Journal of the American Medical Informatics Association : JAMIA*, 20(1) :144–151, January 2013.

[28] Les bases de données clinico-biologiques - Recherche translationnelle | Institut National Du Cancer, September 2016.

[29] F. Olive, F. Gomez, A. M. Schott, L. Remontet, N. Bossard, N. Mitton, S. Polazzi, M. Colonna, and B. Trombert Paviot. Analyse critique des données du PMSI pour l'épidémiologie des cancers : une approche longitudinale devient possible. *Revue d'Epidémiologie et de Santé Publique*, 59 :53–58, 2011.

[30] Paolo Contiero, Andrea Tittarelli, Anna Maghini, Sabrina Fabiano, Emanuela Frassoldi, Enrica Costa, Daniela Gada, Tiziana Codazzi, Paolo Crosignani, Roberto Tessandori, and Giovanna Tagliabue. Comparison with manual registration reveals satisfactory completeness and efficiency of a computerized cancer registration system. *Journal of Biomedical Informatics*, 41(1) :24–32, 2008.

[31] Cancer incidence in five continents. Volume VIII. *IARC scientific publications*, (155) :1–781, 2002.

[32] Sandro Tognazzo, Bovo Emanuela, Fiore Anna Rita, Guzzinati Stefano, Monetti Daniele, Stocco Cramen Fiorella, and Zambon Paola. Probabilistic classifiers and automated cancer registration : An exploratory application. *Journal of Biomedical Informatics*, 42(1) :1–10, 2009.

[33] Luke V. Rasmussen, Will K. Thompson, Jennifer A. Pacheco, Abel N. Kho, David S. Carrell, Jyotishman Pathak, Peggy L. Peissig, Gerard Tromp, Joshua C. Denny, and Justin B. Starren. Design patterns for the development of electronic health record-driven phenotype extraction algorithms. *Journal of Biomedical Informatics*, 51 :280–286, October 2014.

[34] April G Fritz and World Health Organization. *International classification of diseases for oncology : ICD-O*. 2013.

[35] Maud Toulmonde, Françoise Ducimetière, Vianney Jouhet, Thomas Gaudin, Arnaud Malfilatre, Yec'han Laizet, Simone Mathoulin-Pélissier, Jean-Yves Blay, and Jean-Michel Coindre. Les grandes bases de données nationales et les études dans la vraie vie : l'exemple des sarcomes en France. *Bulletin du Cancer*, 103(6, Supplement 1) :S71–S75, 2016.

[36] D. Max Parkin, Jacques Ferlay, Maria-Paula Curado, Freddie Bray, Brenda Edwards, Hai-Rim Shin, and David Forman. Fifty years of cancer incidence : CI5 I–IX. *International Journal of Cancer*, 127(12) :2918–2927, 2010.

[37] Frantz Thiessard, Fleur Mougin, Gayo Diallo, Vianney Jouhet, Sébastien Cossin, Nicolas Garcelon, Boris Campillo-Gimenez, Wassim Jouini, Julien Grosjean, Philippe Massari, and others. RAVEL : retrieval and visualization in ELectronic health records. In *MIE*, pages 194–198, 2012.

[38] Omri Gottesman, Helena Kuivaniemi, Gerard Tromp, W. Andrew Faucett, Rongling Li, Teri A. Manolio, Saskia C. Sanderson, Joseph Kannry, Randi Zinberg, Melissa A. Basford, Murray Brilliant, David J. Carey, Rex L. Chisholm, Christopher G. Chute, John J. Connolly, David Crosslin, Joshua C. Denny, Carlos J. Gallego, Jonathan L. Haines, Hakon Hakonarson, John Harley, Gail P. Jarvik, Isaac Kohane, Iftikhar J. Kullo, Eric B. Larson, Catherine McCarty, Marylyn D. Ritchie, Dan M. Roden, Maureen E. Smith, Erwin P. Böttinger, Marc S. Williams, and and The eMERGE Network. The Electronic Medical Records and Genomics

(eMERGE) Network : past, present, and future. *Genetics in Medicine*, 15(10) :761–771, October 2013.

[39] Shawn N. Murphy, Griffin Weber, Michael Mendis, Vivian Gainer, Henry C. Chueh, Susanne Churchill, and Isaac Kohane. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association*, 17(2) :124–130, March 2010.

[40] Shawn N Murphy, Michael Mendis, Kristel Hackett, Rajesh Kuttan, Wensong Pan, Lori C Phillips, Vivian Gainer, David Berkowicz, John P Glaser, Isaac Kohane, and Henry C Chueh. Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 548–552, 2007.

[41] María del Carmen Legaz-García, Catalina Martínez-Costa, Marcos Menárguez-Tortosa, and Jesualdo Tomás Fernández-Breis. A semantic web based framework for the interoperability and exploitation of clinical models and EHR data. *Knowledge-Based Systems*, 105 :175–189, 2016.

[42] Alan Ruttenberg, Tim Clark, William Bug, Matthias Samwald, Olivier Bodenreider, Helen Chen, Donald Doherty, Kerstin Forsberg, Yong Gao, Vipul Kashyap, June Kinoshita, Joanne Luciano, M. Scott Marshall, Chimezie Ogbuji, Jonathan Rees, Susie Stephens, Gwendolyn T. Wong, Elizabeth Wu, Davide Zaccagnini, Tonya Hongsermeier, Eric Neumann, Ivan Herman, and Kei-Hoi Cheung. Advancing translational research with the Semantic Web. *BMC Bioinformatics*, 8(Suppl 3) :S2, May 2007.

[43] David Wood, Markus Lanthaler, and Richard Cyganiak. RDF 1.1 Concepts and Abstract Syntax. W3c Recommendation, W3C, 2014. http ://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/.

[44] Sean Bechhofer. OWL : Web ontology language. In *Encyclopedia of Database Systems*, pages 2008–2009. Springer, 2009.

[45] Jie Bao, Elisa Kendall, Deborah McGuinness, and Peter Patel-Schneider. OWL 2 Web Ontology Language Quick Reference Guide (Second Edition). Technical report, W3C, 2012. http ://www.w3.org/TR/2012/REC-owl2-quick-reference-20121211/ bibtex[bibsource=http ://w2.syronex.com/jmr/w3c-biblio] bibtex : bao_owl_2012.

[46] Tom Gruber. Ontology. *Encyclopedia of database systems*, pages 1963–1965, 2009.

[47] Alistair Miles and Sean Bechhofer. SKOS Simple Knowledge Organization System Reference. W3c Recommendation, W3C, 2009. http ://www.w3.org/TR/2009/REC-skos-reference-20090818/ bibtex[bibsource=http ://w2.syronex.com/jmr/w3c-biblio] bibtex : miles_skos_2009.

[48] International Electrotechnical Commission and others. ISO/IEC 11179 : information technology-Metadata registries (MDR). *Geneva : International Electrotechnical Commission*, 2005. bibtex : commission_iso/iec_2005.

[49] Jesualdo Tomás Fernández-Breis, José Alberto Maldonado, Mar Marcos, María del Carmen Legaz-García, David Moner, Joaquín Torres-Sospedra, Angel Esteban-Gil, Begoña Martínez-Salvador, and Montserrat Robles. Leveraging electronic healthcare record standards and semantic web technologies for the identification of patient cohorts. *Journal of the American Medical Informatics Association*, 20(e2) :e288–e296, December 2013.

[50] Thomas Beale, Thomas Beale, and Thomas Beale. Archetypes Constraint-based Domain Models for Futureproof Information Systems. 2000.

[51] Siaw-Teng Liaw, Jane Taggart, Hairong Yu, Simon de Lusignan, Craig Kuziemsky, and Andrew Hayen. Integrating electronic health record information to support integrated care : Practical application of ontologies to improve the accuracy of diabetes disease registers. *Journal of Biomedical Informatics*, 52 :364–372, 2014.

[52] George A. Komatsoulis, Denise B. Warzel, Francis W. Hartel, Krishnakant Shanbhag, Ram Chilukuri, Gilberto Fragoso, Sherri de Coronado, Dianne M. Reeves, Jillaine B. Hadfield, Christophe Ludet, and Peter A. Covitz. caCORE version 3 : Implementation of a model driven, service-oriented architecture for semantic interoperability. *Journal of Biomedical Informatics*, 41(1) :106–123, 2008.

[53] Christel Daniel, Anil Sinaci, David Ouagne, Eric Sadou, Gunnar Declerck, Dipak Kalra, Jean Charlet, Kerstin Forsberg, Landen Bain, Charlie Mead, Sajjad Hussain, and Gokce B. Laleci Erturkmen. Standard-based EHR-enabled applications for clinical research and patient safety : CDISC – IHE QRPH – EHR4cr & SALUS collaboration. *AMIA Summits on Translational Science Proceedings*, 2014 :19–25, April 2014.

[54] Frederik Malfait and Scott Bahlavooni. Semantic Technology and CDISC Standards. In *PhUSE 2013 - ninth PhUSE Annual Conference*, Brussels, October 2013.

[55] S. Tognazzo, A. Andolfo, E. Bovo, A. R. Fiore, A. Greco, S. Guzzinati, D. Monetti, C. F. Stocco, and P. Zambon. Quality control of automatically defined cancer cases by the automated registration system of the Venetian Tumour Registry Quality control of cancer cases automatically registered. *The European Journal of Public Health*, 15(6) :657–664, December 2005.

[56] World Health Organization. *International statistical classification of diseases and related health problems*, volume 1. World Health Organization, 2004.

[57] F. Hartel, D. B. Warzel, and P. Covitz. OWL/RDF/LSID Utilization in NCI Cancer Research Infrastructure. In *W3C Workshop on Semantic Web for Life Sciences*, 2004.

[58] Werner Ceusters, Barry Smith, and Louis Goldberg. A terminological and ontological analysis of the NCI Thesaurus. *Methods of information in medicine*, 44(4) :498, 2005.

[59] Stefan Schulz, Daniel Schober, Ilinca Tudose, and Holger Stenzhorn. The Pitfalls of Thesaurus Ontologization – the Case of the NCI Thesaurus. *AMIA Annual Symposium Proceedings*, 2010 :727–731, 2010.

[60] NCI Metathesaurus, March 2015.

[61] Olivier Bodenreider. The Unified Medical Language System (UMLS) : integrating biomedical terminology. *Nucleic acids research*, 32(Database issue) :D267–270, January 2004.

[62] Cui Tao, Jyotishman Pathak, Harold R. Solbrig, Wei-Qi Wei, and Christopher G. Chute. Terminology representation guidelines for biomedical ontologies in the semantic web notations. *Journal of Biomedical Informatics*, 46(1) :128–138, 2013.

[63] Using OWL and SKOS, April 2016.

[64] Riccardo Falco, Aldo Gangemi, Silvio Peroni, David Shotton, and Fabio Vitali. Modelling OWL Ontologies with Graffoo. In Valentina Presutti, Eva Blomqvist, Raphael Troncy, Harald Sack, Ioannis Papadakis, and Anna Tordai, editors, *The Semantic Web : ESWC 2014 Satellite Events*, volume 8798, pages 320–325. Springer International Publishing, Cham, 2014.

[65] Simple part-whole relations in OWL Ontologies, April 2016.

[66] S. Schulz, M. Romacker, and U. Hahn. Part-whole reasoning in medical ontologies revisited–introducing SEP triplets into classification-based description logics. *Proceedings of the AMIA Symposium*, pages 830–834, 1998.

[67] Stefan Schulz and Udo Hahn. Part-whole representation and reasoning in formal biomedical ontologies. *Artificial Intelligence in Medicine*, 34(3) :179–200, 2005.

[68] ICD Conversion Programs - SEER, February 2016.

[69] Gergely Héja, György Surján, and Péter Varga. Ontological analysis of SNOMED CT. *BMC Medical Informatics and Decision Making*, 8(Suppl 1) :S8, October 2008.

[70] Olivier Bodenreider, Barry Smith, Anand Kumar, and Anita Burgun. Investigating subsumption in SNOMED CT : An exploration into large description logic-based biomedical terminologies. *Artificial intelligence in medicine*, 39(3) :183, March 2007.

[71] Nicholas Rescher. Axioms for the Part Relation. *Philosophical Studies*, 6(1) :8–11, 1955.

[72] Cornelius Rosse and José L. V. Mejino Jr. The Foundational Model of Anatomy Ontology. In Albert Burger BSc MSc, Duncan Davidson BSc, and Richard Baldock BSc, editors, *Anatomy Ontologies for Bioinformatics*, number 6 in Computational Biology, pages 59–117. Springer London, 2008. DOI : 10.1007/978-1-84628-885-2_4.

[73] Foundational Model of Anatomy | Structural Informatics Group, August 2016.

[74] Curado, M., Okamoto, N., Ries, L., Sriplung, H., Young, J., Carli, M., Izarzugaza, I., Koscianska, B., Demaret, E., Ferlay, J., Parkin, M., Tyczynski, J., and Whelan, S. International rules for multiple primary cancers (ICD-O Third Edition), 2004.

[75] Antoine Buemi. Pathology of Tumours for Cancer Registry Personnel. *IARC, Lyon*, 2008.

[76] Constance Percy, Valerie van Holten, Calum S. Muir, World Health Organization, and others. International classification of diseases for oncology/editors, Constance Percy, Valerie Van Holten, Calum Muir. 1990.

[77] Phillip Lord. The semantic web takes wing : Programming ontologies with Tawny-OWL. *arXiv preprint arXiv :1303.0213*, 2013.

[78] Matthew Horridge and Sean Bechhofer. The OWL API : A Java API for OWL ontologies. *Semantic Web*, 2(1) :11–21, January 2011.

[79] Birte Glimm, Ian Horrocks, Boris Motik, Giorgos Stoilos, and Zhe Wang. HermiT : an OWL 2 reasoner. *Journal of Automated Reasoning*, 53(3) :245–269, 2014.

[80] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1) :1–47, March 2002.

[81] Olivier Dameron, Paolo Besana, Oussama Zekri, Annabel Bourdé, Anita Burgun, and Marc Cuggia. OWL model of clinical trial eligibility criteria compatible with partially-known information. *Journal of Biomedical Semantics*, 4 :17, September 2013.

[82] Cui Tao, Guoqian Jiang, Thomas A. Oniki, Robert R. Freimuth, Qian Zhu, Deepak Sharma, Jyotishman Pathak, Stanley M. Huff, and Christopher G. Chute. A semantic-web oriented representation of the clinical element model for secondary use of electronic health records data. *Journal of the American Medical Informatics Association*, December 2012.

Glossaire

ACP	Anatomie et Cytologie pathologique
ADICAP	Association pour le Développement de l'Informatique en Cytologie et en Anatomie Pathologique
CDISC	Clinical Data Interchange Standards Consortium
CDW	Clinical Data Warehouse
CHU	Centre Hospitalier Universitaire
CRC	Clinical Research Chart
CUI	Concept Unique Identifier
DPI	Dossier Patient Informatisé
EDC	Entrepôt de Données Clinique
EHR	Electronic Health Record
ENCR	European Network of Cancer Registries
ETL	Extract Transorm and Load
FMA	Foundational Model of Anatomy
HIS	Hospital Information System
i2b2	Informatics for Integrating Biology & the Bedside
IACR	International Association of Cancer Registries
ICD-10	International Classification of Diseases 10^{th} Edition
ICD-O-3	International Classification of Diseases in Oncology 3^{rd} Edition
IHTSDO	International Health Terminology Standards Development Organisation
IRI	Internationalized Resource Identifier
NCI	National Cancer Institute
NCIm	National Cancer Institute's Metathesaurus
NCIt	National Cancer Institute's Thesaurus
OWL	Web Ontology Language
OWL-DL	Ontology Web Language Description Logics
PMSI	Programme de Médicalisation des Systèmes d'Information
RCP	Réunion de Concertation Pluridisciplinaire
RDF	Resource Description Framework
SEER Program	Surveillance, Epidemiology, and End Results Program

SKOS	SKOS Simple Knowledge Organization System
SNOMED-CT	Systematized Nomenclature of Medicine - Clinical Terms
UMLS	Unified Medical Language System
W3C	World Wide Web Consortium
WHO	World Health Organization