

Approche préventive de la qualité des données d'importation dans le contexte de la protéomique clinique

Pierre Naubourg, Marinette Savonnet, Eric Leclercq, Kokou Yétongnon

▶ **To cite this version:**

Pierre Naubourg, Marinette Savonnet, Eric Leclercq, Kokou Yétongnon. Approche préventive de la qualité des données d'importation dans le contexte de la protéomique clinique. Revue des Nouvelles Technologies de l'Information, 2011, E22, pp.189-205. hal-00710932

HAL Id: hal-00710932

https://u-bourgogne.hal.science/hal-00710932

Submitted on 22 Jun 2012

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

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

Approche préventive de la qualité des données d'importation dans le contexte de la protéomique clinique

Pierre Naubourg, Marinette Savonnet, Éric Leclercq et Kokou Yétongnon

Université de Bourgogne, Laboratoire LE2I - UMR5158
9 Avenue Alain Savary
21000 DIJON, FRANCE
{pierre.naubourg, marinette.savonnet, eric.leclercq, kokou}@u-bourgogne.fr
http://le2i.cnrs.fr

Résumé. Dans le domaine biomédical, la protéomique est confrontée à des sources de données de plus en plus nombreuses et à des volumes de données très importants du fait de la multiplication des technologies dites à haut débit. L'hétérogénéité de la provenance des données implique de fait une hétérogénéité dans la représentation et le contenu de ces données. Les données peuvent aussi se révéler incorrectes ce qui engendre des erreurs sur les conclusions des expériences protéomiques. Notre approche a pour objectif de garantir la qualité initiale des données lors de leur importation dans un système d'information dédié à la protéomique. Elle est basée sur le couplage entre des modèles représentant les sources et le système protéomique, et des ontologies utilisées comme médiatrices entre les modèles. Les différents contrôles que nous proposons de mettre en place garantissent la validité des domaines de valeurs, la sémantique et la cohérence des données lors de l'importation.

1 Introduction

Notre contexte de travail est le domaine biomédical et plus précisément la protéomique clinique. La particularité de la protéomique clinique est la recherche de caractéristiques protéiniques d'échantillons issus de groupes de patients participant à une étude. Parmi ces caractéristiques, nous pouvons donner comme exemple la découverte de biomarqueurs permettant d'identifier une pathologie et ainsi de la classifier, d'effectuer un diagnostic précoce, d'étudier la réponse du patient au traitement, etc. Le travail des plateformes protéomiques est centré sur la réalisation d'études mettant en jeu un grand nombre d'échantillons dont on essaie d'extraire des caractéristiques via des expérimentations. Outre les données nécessaires à l'analyse des échantillons sur les spectromètres de masse par exemple, la réalisation d'études statistiques en aval de ces expériences nécessite l'utilisation de données cliniques. Les données cliniques englobent des données aussi larges que les caractéristiques du patient, la description des pathologies diagnostiquées, les caractéristiques des échantillons prélevés, les conditions de transport et de stockage.

Approche préventive de la qualité des données d'importation protéomiques

Afin de gérer les données biomédicales des études protéomiques, les plateformes utilisent communément un Système Informatique de Gestion de Laboratoire (SIGL [1]). Les SIGL protéomiques gèrent tous les aspects des études protéomiques, depuis le stockage des informations cliniques jusqu'à la réalisation d'études statistiques suite à l'analyse sur différents équipements. La qualité des données gérées au sein des SIGL est assurée par l'automatisation des tâches et la réduction des manipulations directes entre l'utilisateur et les données. Néanmoins, comme il existe un lien entre les données cliniques importées dans le SIGL et les conclusions des études protéomiques, augmenter la qualité des données durant le processus d'importation garantit la pertinence des conclusions des études.

La qualité des données intéresse de plus en plus de chercheurs dans les laboratoires publics et privés comme en témoignent les travaux de Redman (2001) et de Dasu et Johnson (2003). Sans la mise en place d'une gestion rigoureuse de la qualité de données, le SIGL pourra être rapidement pollué par des données superflues, manquantes voire incorrectes. Depuis de nombreuses années, des méthodes de prévention, d'audit et de nettoyage des données améliorent leur qualité au sein des systèmes d'information. Berti-Équille (2007) a référencé quatre approches complémentaires de gestion de la qualité des données. Les approches diagnostiques s'appuient principalement sur des méthodes statistiques détectant les erreurs présentes dans de grandes quantités de données. Les approches adaptatives proposent des traitements dynamiques de vérification en temps réel de contraintes garantissant la qualité des données. Les approches correctives essaient de détecter les erreurs en les comparant à des valeurs issues de la réalité (dites « données de terrain ») et proposent des corrections. Les approches préventives effectuent en amont du stockage des données des évaluations au niveau des modèles et des processus mis en œuvre.

Notre contribution à la qualité des données dans ce contexte est de proposer une méthode d'importation semi-automatisée garantissant la qualité initiale des données importées au sein du SIGL. Notre méthode d'importation se situe dans les approches préventives car elle utilise les modèles et les ontologies lors de sa mise en œuvre. Notre approche se propose de gérer cette problématique à l'aide de différents contrôles prenant en compte les problèmes de provenance des données et de contexte d'utilisation. Ces problèmes de provenance sont le plus souvent liés aux particularités propres à chacun des systèmes d'information des partenaires, ils laissent apparaître des conflits entre les concepts manipulés par les différents systèmes. Afin de régler ces conflits, nous proposons de mettre en place une architecture basée sur des modèles, des ontologies et des mappings. Les problèmes liés au contexte d'utilisation des données apparaissent dans le cas où les données sont en contradiction avec le contexte. Le contexte d'utilisation correspond d'une part à la gestion des données au sein des modèles du SIGL et d'autre part au contexte métier représenté par une ontologie et des règles.

Nous commencerons par exposer, en section 2, la problématique liée à l'importation des données cliniques que nous illustrerons à partir d'exemples. Ensuite, nous décrirons les outils et les méthodes utilisés dans notre approche en section 3. La section 4 présentera notre approche ainsi que sa mise en œuvre au sein de la protéomique clinique. Avant de conclure sur les perspectives envisagées pour ces travaux, nous présenterons plus en détail le prototype d'implémentation de notre approche.

1. Ces systèmes sont plus connus sous leur nom anglais : *Laboratory Information Management System* (LIMS)

2 Problèmes liés à l'importation des données

Cette section présente le contexte dans lequel nous menons nos travaux ainsi que les problèmes soulevés par les difficultés rencontrées dans le cadre d'une importation de données cliniques. Les données sont collectées par des intervenants extérieurs à la plateforme, que nous appellerons dans la suite de cet article des *partenaires*. Par exemple, les cliniciens possèdent des fichiers pathologiques, les Centres Hospitaliers Universitaires (CHU) possèdent les caractéristiques biologiques des patients, les Centres de Ressources Biologiques (CRB) organisent la conservation des échantillons. Ainsi, pour chaque étude protéomique, les échantillons à analyser sont accompagnés de nombreuses données cliniques devant être importées au sein du SIGL. Les expertises des résultats d'expériences nécessitent des données de qualité pour effectuer des conclusions pertinentes.

Nous présentons quelques exemples de jeux de données reçus par les plateformes protéomiques. Les tableaux 1 et 2 sont des extraits de données cliniques fournis à la plateforme protéomique par deux cliniciens (respectivement C1 et C2). Ces deux jeux de données définissent sur chaque ligne les données cliniques associées à un des échantillons devant être analysés. Ils illustrent l'hétérogénéité et les problèmes présents au sein des jeux de données. Les problèmes rencontrés peuvent être divisés en deux catégories : les difficultés liées à la multiplicité des sources de données et celles liées à l'utilisation des données.

NumEch	NumPat	Sexe	DNaissance	Maladie	Organe
S124	HG65	F	26-mai-2007	LAL	moelle osseuse
S125				LAL	moelle osseuse
S126	HG65	G	26-mai-2007	LAL	moelle osseuse
S127	YK37	G	01-juil-2007	LAL	moelle osseuse

TAB. 1: Jeu de données provenant du clinicien C1 (extrait).

N°Ech	DateN	N°Pat	Genre	Pathologie	Prélèvement
654	16/08/1948	hj25	F	néoplasme sein	sein
HG12	01/02/1962	hu65	F	néoplasme sein	sein
S7	12/04/1956	JH34	H	néoplasme sein	foie
YK37	29/02/1945	dv12	F	néoplasme sein	sein

TAB. 2: Jeu de données provenant du clinicien C2 (extrait).

2.1 Problèmes liés à la multiplicité des sources de données

Les données à importer proviennent de partenaires ayant des méthodes de travail propres qui conduisent à des disparités au niveau des jeux de données qu'ils transmettent au SIGL cible. On distingue généralement deux catégories de conflits : les conflits syntaxiques et les conflits sémantiques.

Les conflits syntaxiques relèvent le plus souvent de différences de format ou de structure dans la conception des systèmes d'information. Les conflits de format sont conséquents avec le choix de terminologies et nomenclatures différentes entre les systèmes.

Format des données

Nous pouvons remarquer des divergences sur les formats de données. Si nous prenons l'exemple de la date de naissance, le clinicien C1 choisit le format `JJ-mmm-AAAA` alors que le clinicien C2 choisit le format `JJ/MM/AAAA`. Afin que les traitements sur les données soient possibles dans ces deux cas, il est nécessaire de réaliser des opérations de conversion.

Les conflits sémantiques, étudiés depuis longtemps (Kim et Seo (1991); Siegel et Madnick (1991); Naiman et Ouksel (1995)) ont été résumés par Goh (1997) qui identifie trois types de conflits induits par l'hétérogénéité sémantique : 1) les conflits de nommage qui apparaissent en présence d'homonymes et de synonymes, 2) les conflits d'échelle qui apparaissent lorsque les granularités de description choisies ne sont pas les mêmes, et 3) les conflits de confusion qui apparaissent lorsque un mot est utilisé pour deux significations différentes.

Sémantique des données

Lors de l'étude des jeux de données, nous pouvons remarquer des différences entre les nomenclatures utilisées. Par exemple, les deux cliniciens indiquent la date de naissance du patient dans leur tableau en utilisant des dénominations différentes. Le clinicien C1 utilise la dénomination *DNaissance* alors que le clinicien C2 utilise la dénomination *DateN*. La sémantique de ces deux champs est la même : *signifier la date de naissance du patient*.

Domaine des valeurs et échelle

Prenons l'exemple du sexe des patients (*Sexe* pour le clinicien C1 et *Genre* pour le clinicien C2), nous pouvons remarquer que les domaines des valeurs ne sont pas compatibles. Le clinicien C1 travaillant sur des échantillons provenant d'enfants a décidé d'utiliser la notation G pour garçon et F pour fille alors que le clinicien C2 a utilisé la notation H pour Homme et F pour Femme.

Les problèmes d'échelle entre les données sont divisibles en deux catégories. Le problème de mesure se rencontre, par exemple, lorsqu'un volume est exprimé en μl et un autre en ml. Le problème de granularité se rencontre lorsqu'une information est soit vue globalement soit détaillée. Par exemple, le même stade d'évolution d'un cancer peut être décrit par plusieurs champs détaillant les caractéristiques d'évolution ou bien par un seul champ qui englobe toutes les caractéristiques par concaténation.

Degoulet et al. (1997), travaillant sur les échanges de messages entre acteurs du domaine biomédical, ont mis en exergue la possibilité de résoudre les problèmes de sémantique par l'utilisation de vocabulaires normalisés ou de référentiels. Lors du processus d'importation, ces problèmes peuvent être résolus si, d'une part, le format et le domaine de valeurs des données des différents systèmes sont connus et, d'autre part, si des méthodes de transformation automatiques sont disponibles. Cependant, les problèmes liés à la sémantique sont beaucoup plus complexes et nécessitent des techniques de représentation de la connaissance du domaine.

2.2 Problèmes liés à l'utilisation des données

La représentation et la gestion des données biologiques et biomédicales posent des problèmes aux concepteurs de systèmes d'information. Chen et Carlis (2003) ont identifié quatre verrous technologiques dans le domaine génomique : 1) les données sont complexes car elles reflètent les points de vue particuliers des différents spécialistes, 2) la connaissance nécessaire à leur compréhension est importante, 3) la connaissance évolue constamment et 4) les personnes travaillant en bioinformatique ont divers profils, le plus souvent ce sont des personnes ayant des compétences différentes essayant de faire consensus afin de répondre à un objectif commun. Parmi ces verrous, la complexité des données est celui qui pose le plus de problèmes dans notre contexte. La complexité des données biologiques provient de leurs caractères hétérogènes (Davidson et al. (1995)), incomplets, incertains et inconsistants (Willson (1998)) mais aussi de leur différence de granularité. Par exemple, la granularité des données peut varier d'un niveau macroscopique avec les données générales sur les patients jusqu'au niveau des acides aminés constituant les protéines.

La complétude et la consistance de données sont au cœur des préoccupations de nombreux chercheurs. Le chapitre 2 du livre de Han et Kamber (2006) fait une synthèse des différentes solutions de résolutions de valeurs manquantes ou incorrectes. Une solution proposée consiste à ignorer le tuple ou l'objet. La solution inverse consiste à remplir manuellement les données manquantes. D'autres solutions intermédiaires comme l'utilisation d'une constante de remplacement, d'une valeur moyenne ou d'un arbre de décision sont également proposées. Cependant, la complétude et la consistance des données ne peuvent être évaluées que par rapport à un contexte métier clairement spécifié.

Complétude des données

Le tableau 1 présente les données cliniques de patients associées à chaque échantillon. Dans ce fichier, nous pouvons remarquer que la deuxième ligne (correspondant à l'échantillon S125) ne fournit pas les données nécessaires pour *identifier* le patient. Il est alors nécessaire soit de rejeter cette donnée par manque d'information, soit de l'annoter afin de la distinguer des données validées.

Consistance des données

Certaines données décrivant le même concept peuvent parfois définir des caractéristiques différentes pour ce concept. Par exemple, l'étude des données du tableau 1 révèle que les deux échantillons S124 et S126 proviennent du même patient HG65. Néanmoins, nous pouvons remarquer que pour un échantillon le sexe du patient est Garçon et pour l'autre Fille. Ces deux données sur le *sexe* définissent des caractéristiques différentes pour un même patient.

Contexte métier

La prise en compte de la connaissance du domaine permet de souligner un autre problème dans la description des données cliniques du tableau 2. Ces données concernent l'étude de la pathologie « cancer du sein », la plupart des échantillons proviennent ainsi de prélèvements effectués sur le sein du patient. Cependant, l'échantillon correspondant à la troisième ligne provient du foie du patient. S'agit-il d'une erreur du clinicien ou d'une particularité que l'on veut étudier ? Seul un expert du domaine peut répondre à cette question. La mise en place

d'une représentation de la connaissance du domaine, sur laquelle seront exprimées des règles métier, permet de détecter d'éventuelles incohérences au sein des données.

Pour mettre en place le contrôle de la qualité des données, nous proposons un processus reposant sur trois niveaux (figure 1). Le premier niveau met en relation les modèles des données sources des partenaires avec le modèle du SIGL. Il traite les problèmes de formats, de domaines et d'échelles en utilisant une représentation de la connaissance reposant sur des modèles et des ontologies. Le second niveau vérifie la complétude et la consistance des données par rapport à leur contexte d'utilisation dans le SIGL en utilisant le modèle de données du SIGL. Enfin le troisième niveau assure la cohérence des données en s'appuyant sur les règles métier, les ontologies et les modèles. En conséquence, pour mettre en place les trois niveaux de contrôle nous avons besoin : 1) d'une modélisation de la connaissance du domaine, 2) d'une modélisation des règles métier des acteurs de la plateforme protéomique et 3) d'une modélisation de la structure des données au sein du SIGL.

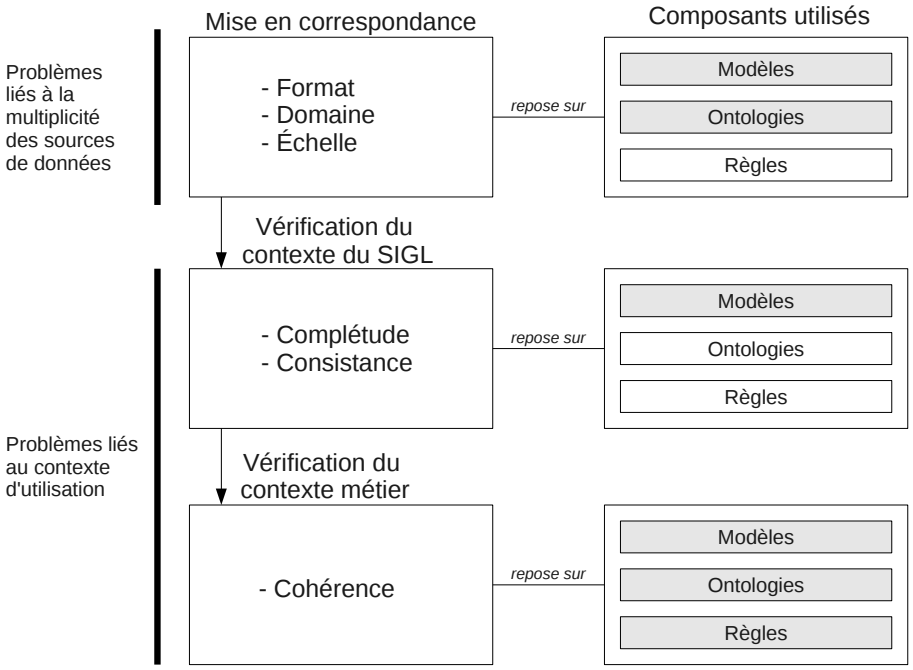


FIG. 1: Niveaux de contrôle garant de la qualité des données.

3 Ontologies et modèles

Linster (1992) présente deux paradigmes de modélisation complémentaires : l'un pour signifier et l'autre pour implémenter. Modéliser pour signifier consiste à structurer formellement les connaissances des experts du domaine. Modéliser pour implémenter est la modélisation la plus commune, elle consiste à organiser les composants d'un système de façon à les exécuter sur une machine. Dans notre approche, nous utilisons de manière complémentaire ces deux paradigmes : les ontologies pour représenter la connaissance et les modèles pour implémenter le module de contrôle de la qualité des données.

3.1 Ontologies

En informatique, les ontologies sont des spécifications explicites et formelles d'une conceptualisation (Gruber (1995)). Dans la pratique, elles peuvent être utilisées pour représenter la connaissance du domaine ou pour aider à la compréhension d'un système en séparant les données du système des concepts du domaine.

De nombreux efforts sont réalisés au sein du domaine biomédical pour structurer la connaissance sous forme d'ontologies afin de créer des standards. Le consortium *Gene Ontology* (GO [2]) s'attache à produire un vocabulaire contrôlé sous forme ontologique des rôles des gènes dans l'expression des protéines (Ashburner et al. (2000)). *Medical Subject Headings* (MeSH [3]) est une autre initiative d'acteurs du domaine biomédical s'attachant à regrouper et structurer tout le vocabulaire du domaine biomédical. Les termes proposés concernent aussi bien l'anatomie que les disciplines ou encore les administrations de santé.

Il existe différents types d'ontologies utilisés afin de servir un but précis. Van Heijst et al. (1997) définissent quatre types d'ontologies : les ontologies génériques, les ontologies de représentation, les ontologies de domaine et les ontologies d'application. Nous détaillerons dans cet article uniquement les ontologies de domaine et d'application.

Ontologie de domaine

Une ontologie de domaine est utilisée afin de représenter la connaissance consensuelle d'un domaine (Salem et AbdelRahman (2010)). Elle représente les concepts clés du domaine organisés entre eux par différentes relations. Les principales relations utilisées sont les spécialisations (*is-a*), les synonymies (*synonym*), les relation méronymiques (*part-of*) et les relations génériques (*related-to*). Ce type d'ontologie est utilisé pour garantir une homogénéité sémantique (aussi appelée *semantic net* par Wiederhold (1994)) entre les différents systèmes s'y référant. Une ontologie de domaine peut servir de référence scientifique lors de discussions ou d'échanges avec des partenaires de ce domaine. Les concepts et les relations servent alors de consensus syntaxique et sémantique.

Ontologie d'application

Une ontologie d'application est utilisée afin de représenter des connaissances spécifiques à un sous-domaine, un système ou une application (Guarino (1998)). Elle diffère de l'ontologie de domaine par sa représentation de la réalité du système d'information auquel elle est liée.

2. `http://www.geneontology.org`
3. `http://www.nlm.nih.gov/mesh`

Une ontologie de ce type peut être utilisée lors d'une coopération entre les systèmes des différents partenaires d'un domaine. Elle sert souvent de référence lors de discussions techniques entre les différents utilisateurs des systèmes d'un domaine afin de savoir si tel concept d'un système correspond bien à tel concept d'un autre système. Par exemple, deux systèmes avec des identifiants de patients différents, NumDossier et CodePatient, feront référence au concept IdPatient afin de signifier ce que représentent ces identifiants.

Règles et ontologies

Afin de traiter le caractère dynamique de la connaissance de notre contexte nous avons choisi de mettre en place un système de gestion évolutif de la logique métier. Ce système repose sur la définition de règles s'appuyant sur les concepts et les relations entre les concepts de notre ontologie de domaine. Dans le domaine des systèmes d'information, les règles métiers sont des expressions formelles qui définissent ou contraignent certains aspects d'un métier. Elles structurent, contrôlent et influencent un système (Hall et al. (2000); Ross (2003)). De récents travaux ont montré l'intérêt des règles au sein du Web Sémantique (Horrocks et Patel-Schneider (2004); Motik et Rosati (2008); Krötzsch et al. (2008)). Ces travaux ont abouti à la définition d'un langage formel d'écriture des règles nommé *Semantic Web Rule Language* [4] (SWRL) combinant les langages OWL-DL [5] et RuleML [6]. Les règles que nous avons retenues permettent de définir des connaissances qui ne sont pas directement modélisables dans l'ontologie. Seuls des experts du domaine peuvent définir quelles sont les règles devant être prises en compte afin de s'approcher au plus près de la réalité du travail des plateformes protéomiques. Les règles écrites en SWRL peuvent se révéler indécidables et rendre l'ontologie inconsistante (section 6 de l'article de Horrocks et al. (2005)). Afin de ne traiter que des règles décidables, nous avons choisi de travailler en respectant les recommandations des règles *DL-Safe* (Motik et al. (2005)). Ces règles sont décidables si elles travaillent sur des classes nommées de l'ontologie et sur un ensemble d'individus connus.

Au sein de notre approche, les ontologies sont utilisées de deux manières : 1) comme médiatrices entre les systèmes partenaires et le SIGL et 2) comme support de la connaissance. La médiation entre les systèmes est réalisée aussi bien sur le plan sémantique que sur le plan syntaxique. Sur le plan sémantique, les ontologies servent à mettre en correspondance des descripteurs de données appartenant à différents schémas. Sur le plan syntaxique, les ontologies servent à déterminer les opérations à effectuer afin de convertir une valeur en une autre. Enfin nous utilisons les ontologies comme support de la connaissance lors de la vérification de la cohérence par rapport au contexte métier. L'évolutivité du système est apportée par le découplage entre la connaissance (ontologies et règles) et l'implémentation du système qui l'exploite.

4. *A Semantic Web Rule Language Combining OWL and RuleML* : `www.w3.org/Submission/SWRL`

5. *Ontology Web Language - Description Logics* est une version intermédiaire de OWL permettant de restreindre certains constructeurs. *OWL 2 Web Ontology Language Document Overview* : `www.w3.org/TR/owl2-overview`

6. *Rule Markup Language* est un langage de balisage basé sur XML qui permet le stockage, l'échange, la récupération et la vérification de règles : `www.ruleml.org`

3.2 Modèles

Les modèles sont des représentations des systèmes selon un certain point de vue. Parmi les langages de modélisation, l'un des plus utilisés est sans doute Unified Modeling Language (UML). UML définit plusieurs diagrammes permettant de décrire chaque aspect (structurel, comportemental, temporel, etc.) d'un système ou d'une application. Fowler (2003) définit dans son livre *UML Distilled* les trois modes d'utilisation des modèles UML : comme un croquis (*sketch*), comme un plan (*blueprints*) ou comme un langage de programmation. Selon Fowler, les modèles UML sont principalement utilisés « *as sketches* ». Les croquis aident principalement à la communication des idées entre les différents acteurs d'un projet au sein des réunions de travail, ils ne sont en aucun cas axés sur le développement ultérieur. Par analogie avec Linster (1992), ces modèles sont utilisés *pour signifier*. Les modèles dits « *blueprints* » sont réalisés afin d'être assez précis en vue d'une implémentation par un développeur. Enfin certaines approches permettent d'utiliser les diagrammes UML comme un langage de programmation à part entière, les diagrammes sont transformés afin d'obtenir le code source du programme (par exemple Executable UML (Starr (2001)). Ces deux derniers points de vue se rapprochent de la *modélisation pour implémenter* de Linster (1992).

Dans le contexte de notre approche, les modèles UML sont définis *as blueprints*, ils sont assez précis pour être utilisés comme une spécification dans la phase d'implémentation. Ils permettent de plus de déterminer la structure et le comportement prévu par le SIGL lors de l'utilisation des données. Lors de l'importation des données, le modèle de structure permet de vérifier la complétude et la consistance des données fournies par les partenaires.

3.3 Complémentarité des ontologies et des modèles

Spear (2006) définit deux dimensions pour la construction de la description de la connaissance d'un domaine :

- la dimension horizontale (ou pertinence) a pour objectif de déterminer l'étendue de l'information qui devra être incluse dans la représentation de la connaissance ;
- la dimension verticale (ou granularité) a pour objectif de déterminer le degré de précision de la représentation des connaissances.

Les ontologies, de part leur mécanisme de raffinement et de spécialisation des concepts sont les plus adaptées à la description verticale d'un domaine. L'axe horizontal est quant à lui mieux supporté par les modèles qui permettent l'agrégation des connaissances sur de grandes étendues.

Ashenhurst (1996) pense que l'utilisation d'ontologies afin de guider la sémantique et donc la connaissance du domaine est pertinente, notre proposition reprend ses conclusions et propose l'utilisation des ontologies comme support de la modélisation de la connaissance et l'utilisation des modèles UML (principalement le diagramme de classes) comme définition de la structure des composants du système. La gestion de la qualité des données étant sujette à dépendre de ces deux aspects (connaissance et structure), le couplage de ces deux types de paradigme de modélisation apporte des solutions en matière de qualité des données.

4 Composants support de la qualité des données

Notre approche consiste à utiliser les atouts propres aux ontologies et aux modèles afin de garantir la qualité des données lors du processus d'importation. Pour cela, nous avons défini le modèle de données utilisé au sein du SIGL afin de garantir la structure attendue des données importées. Nous proposons d'utiliser deux ontologies : une ontologie de domaine et une ontologie d'application. L'ontologie de domaine, construite à partir de normes et nomenclatures, représente la connaissance générale utilisée lors du processus d'importation. L'ontologie d'application sert plus spécifiquement à modéliser les connaissances nécessaires aux systèmes cible et sources. Une médiation entre les modèles des partenaires et le SIGL, au moyen des ontologies, est réalisée via des mappings. La figure 2 présente une synthèse de l'organisation des modèles et des ontologies utilisés au sein de notre approche.

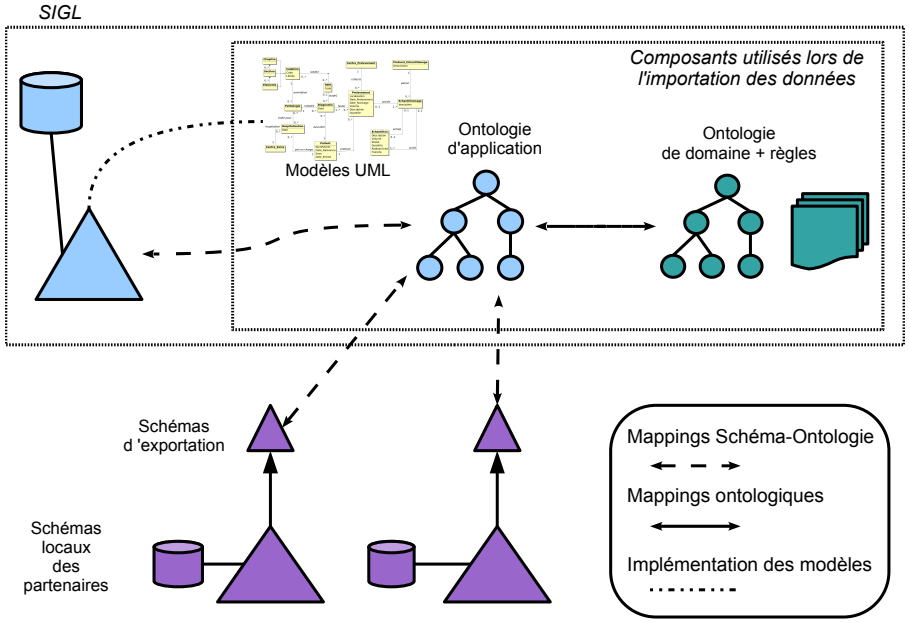


FIG. 2: Organisation des composants utilisés lors de l'importation des données.

4.1 Modèle des données cliniques du SIGL

Le SIGL utilisé par la plateforme protéomique stocke les données reçues au sein d'une base de données relationnelles. La base de données cliniques doit pouvoir assurer la persistance des informations *identifiées* et le cas échéant *transformées* afin de garantir la pertinence des outils de recherche et la qualité des données.

Le modèle de données cliniques a été réalisé grâce au langage de modélisation UML et principalement à l'aide du diagramme de classes. La figure 3 présente un extrait du modèle

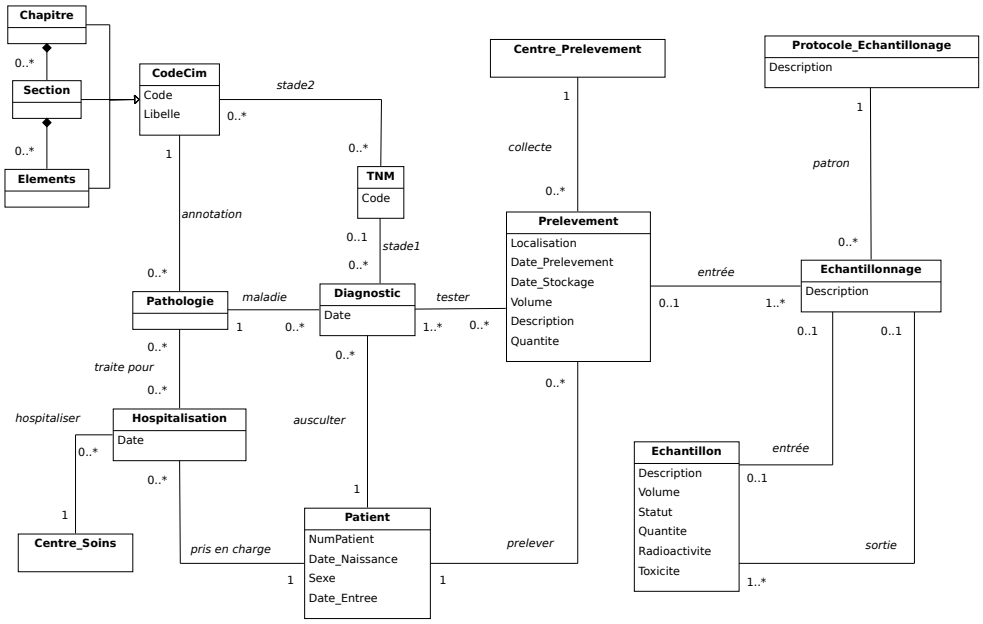


FIG. 3: Modèle des données cliniques (extrait).

concernant les données propres aux patients (telles que la date de naissance et le sexe) et leurs associations d'une part à des pathologies via une date de diagnostic (un patient peut présenter plusieurs pathologies au cours d'une étude), et d'autre part aux données biologiques des prélèvements. Les prélèvements sont effectués en respectant un protocole puis subissent un processus d'échantillonnage avant de produire des échantillons analysables. Les pathologies sont associées à un code respectant la Classification Internationale des Maladies [7] (CIM) proposée par l'OMS. Le diagramme de classes reprend la structure *Chapitre - Section - Element* de la CIM permettant une description plus ou moins fine des pathologies. Par exemple un clinicien peut définir une pathologie par le code CIM `C78.7` (Tumeur maligne secondaire du foie) ou par le code `C00-D48` (Tumeurs malignes) selon le degré de précision des informations fournies. Les prélèvements cancéreux peuvent être associés à un code TNM (Tumor, Nodes, Metastasis) permettant de définir les stades de développement des tumeurs (Belleannée (2006)).

4.2 Ontologies

La création de deux ontologies est nécessaire au fonctionnement de notre approche : une ontologie de domaine support de la connaissance du domaine et une ontologie d'application support de la connaissance du métier des partenaires. L'ontologie de domaine représente l'envi-

7. International Classification of Diseases (ICD), `http://www.who.int/classifications/icd`

ronnement (ici la protéomique clinique) du SIGL et l'ontologie d'application représente l'utilisation effectuée par le SIGL de son environnement.

Ontologie de domaine

La construction de notre ontologie de domaine a suivi une méthode reposant sur deux étapes : 1) la découverte des concepts basée sur la réponse aux « questions pertinentes » et 2) la recherche de concepts communs. Selon Brusa et al. (2006) les questions pertinentes sont des questions que se posent les spécialistes du domaine lors de leurs « investigations » et auxquelles l'ontologie peut apporter une réponse, comme par exemple « Puis-je connaître le degré d'évolution de cette tumeur ? ». L'autre étape de la construction de cette ontologie est basée sur la recherche de concepts communs (Sugumaran et Storey (2002)). En effet, l'analyse de différentes sources de données d'un même domaine fait apparaître l'utilisation d'un grand nombre de concepts communs souvent dissimulés par des termes synonymes. La figure 4 présente un

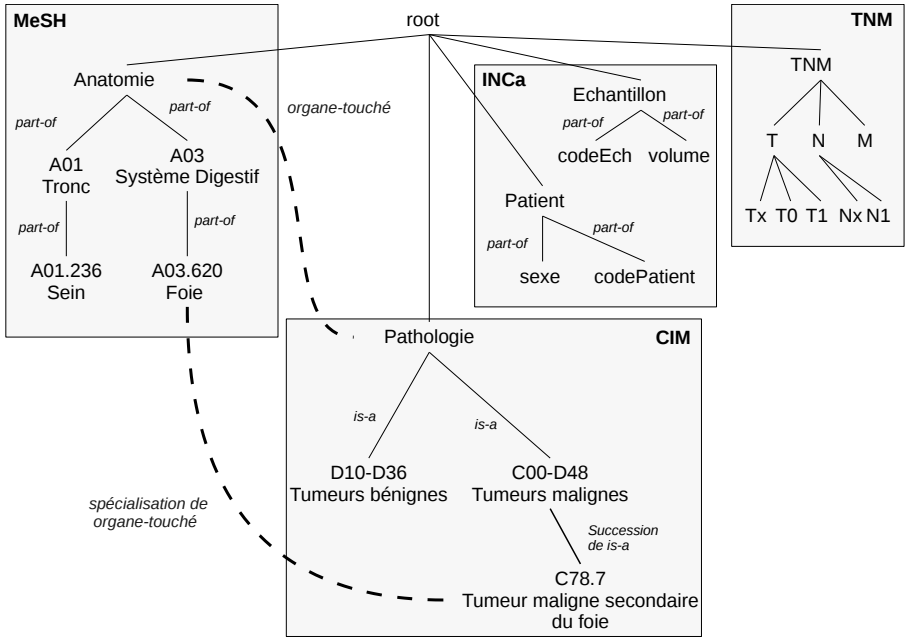


FIG. 4: Ontologie de domaine (extrait).

extrait de notre ontologie de domaine. Les ressources consensuelles que nous avons retenues, afin de répondre aux questions pertinentes, sont la CIM, la nomenclature TNM, la branche anatomie de la classification MeSH et les recommandations de l'Institut National du Cancer (INCa) aux tumorothèques [8]. Cette recommandation englobe les concepts communs des données cliniques utilisés dans notre domaine (les concepts *patient*, *échantillon*, *prélèvement*, etc.).

8. Les tumorothèques sont des banques de tissus tumoraux cryopréservés.

Nous avons mis en place des relations entre les différentes branches de l'ontologie. Ces relations, dont un exemple est représenté en pointillé sur la figure 4, rendent compte de la connaissance métier, par exemple, en spécifiant quels sont les organes qui sont touchés par une pathologie. Pour cela nous définissons une relation générique *organeTouché* qui relie le concept *Anatomie* de la branche Mesh et le concept *Pathologie* de la branche CIM, permettant ainsi de spécifier les organes touchés par une pathologie. Les experts doivent ensuite « spécialiser » la connaissance en précisant les organes touchés par une pathologie donnée : le *Foie* est un organe susceptible d'être touché par la pathologie *C78.7* qui correspond à une tumeur maligne secondaire du foie.

Les règles que nous définissons s'appuient sur les concepts et relations entre les concepts de l'ontologie de domaine. Elles représentent des connaissances que l'on ne peut pas modéliser avec les concepts, les relations et les propriétés. Par exemple, une règle métier énonçant « *un prélèvement est valide si la pathologie pour lequel il est étudié et l'organe dont il provient sont mutuellement pertinents* », sera définie ainsi :

```
Prelevement(?p), organeTouché(?p,?m), pathologie(?m)
    => PrelevementValide(?p)
```

Ontologie d'application

L'ontologie d'application est utilisée comme médiatrice entre les schémas des partenaires et le schéma du SIGL. Elle a été réalisée en accord avec les partenaires principaux des plateformes protéomiques et représente une synthèse des différents schémas de ces derniers. Il s'agit d'une représentation de la sémantique des données sur laquelle repose le SIGL et qui est commune à toutes les études protéomiques gérées par le SIGL. Ces concepts sont partagés par tous les partenaires, ils sont stables car utilisés par toutes les études, et fréquemment référencés. La figure 5 présente un extrait de notre ontologie d'application et la figure 6 présente un extrait de sa branche décrivant les types et formats de données. Cette branche de l'ontologie possède des relations entre les concepts définissant les opérations de conversion entre les types et les opérations de transformation entre les formats.

4.3 Mappings entre les composants

Nous avons choisi d'emprunter aux travaux sur l'alignement d'ontologies (Shvaiko et Euzenat (2008)) le terme *mapping* afin de désigner la mise en place de correspondances, d'une part, entre les concepts des deux ontologies et, d'autre part, entre les concepts de l'ontologie d'application et les descripteurs des schémas. Nous utilisons donc deux types de mappings : les mappings ontologiques rapprochant deux concepts issus des deux ontologies et les mappings schéma-ontologie liant des concepts ontologiques à des descripteurs de schéma.

Mappings ontologiques

Les mappings ontologiques M_O liant des concepts d'ontologies sont des mappings d'équivalence entre les concepts. Dans notre approche ce type de mapping est utilisé pour mettre en correspondance un concept de l'ontologie d'application avec un concept de l'ontologie de domaine. Ces mappings sont réalisés lors de la construction des deux ontologies et doivent être mis à jour lors de l'évolution d'une (ou des deux) ontologies.

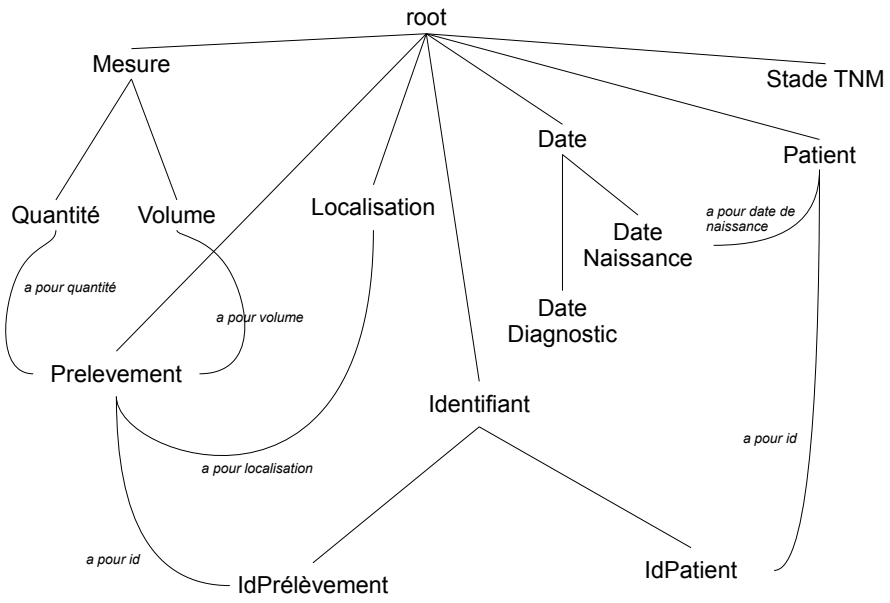


FIG. 5: Ontologie d'application (extrait).

Définition 1 *Un mapping ontologique M_O est une paire $\langle C_{o1}, C'_{o2}\rangle$ où C est un concept de l'ontologie o1 et C' est un concept de l'ontologie o2.*

Par exemple, nous avons mis en place le mapping ontologique suivant :
$M_O1\ (Anatomie_{OD}, Localisation_{OA})$ qui permet de faire correspondre le concept *Anatomie* de l'ontologie de domaine OD et le concept *Localisation* de l'ontologie d'application OA.

Mappings schéma-ontologie

Les mappings schéma-ontologie M_{SO} lient les concepts d'une ontologie d'application aux schémas de données. Ces mappings peuvent être : 1) de type 1..1 liant un concept de l'ontologie à un descripteur du schéma, 2) de type 1..n liant un concept de l'ontologie à plusieurs descripteurs du schéma ou encore 3) de type n..1 liant plusieurs concepts de l'ontologie à un seul descripteur du schéma. Les maping n..m sont décomposables en mappings n..1 et 1..m. Les mappings mis en place définissent quelle est la signification exacte de chaque descripteur de schéma.

Définition 2 *Un mapping schéma-ontologie M_{SO} est une paire $\langle \{D_S\}, \{C_o\}\rangle$ constituée d'un ensemble de descripteurs D du schéma S et d'un ensemble de concepts C de l'ontologie o.*

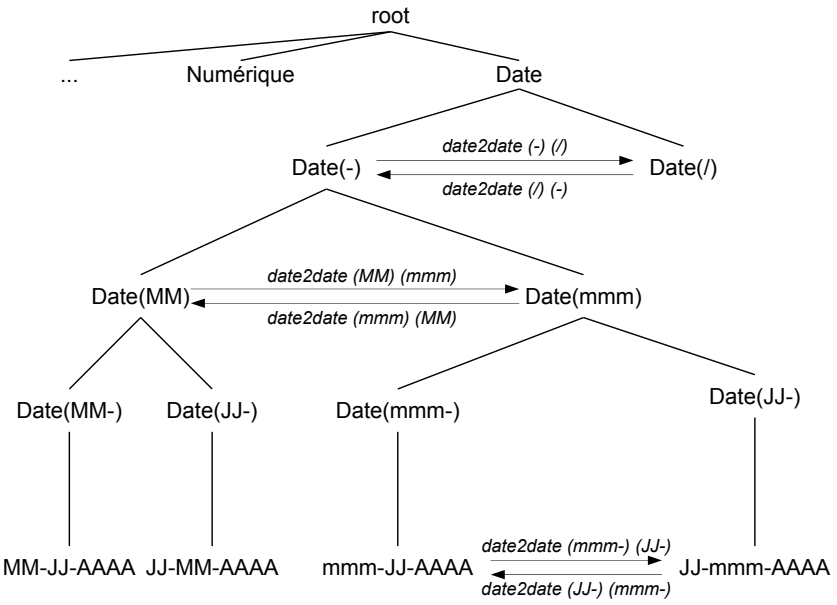


FIG. 6: Branche des types et formats de données de l'ontologie d'application (extrait).

Par exemple, nous avons mis en place les mappings schéma-ontologie suivants afin de mettre en correspondance les descripteurs représentant les identifiants des patients :

- $M_{SO}1$ $(IdPatient_{OA}, NumPatient_{SIGL})$ permet de faire la correspondance entre le concept *IdPatient* de l'ontologie d'application OA et le descripteur *NumPatient* du schéma du SIGL ;
- $M_{SO}2$ $(IdPatient_{OA}, NumDossier_{P1})$ permet de faire la correspondance entre le concept *IdPatient* de l'ontologie d'application OA et le descripteur *NumDossier* du schéma du Partenaire 1. Le partenaire 1 représentant un CHU, les patients sont identifiés selon leur numéro de dossier.

La figure 7 est une capture d'écran de notre prototype montrant des mappings schéma-ontologie (partie droite de la figure) entre les données à importer (partie gauche de la figure) et l'ontologie d'application. La première ligne définit le mapping entre le descripteur *NumPrel* et le concept de l'ontologie d'application *IdPrélèvement*. La deuxième ligne correspond au mapping $M_{SO}2$ liant *NumDossier* à *IdPatient*. Les deux dernières lignes correspondent aux mappings de la date de naissance et du sexe du patient.

Le mappping suivant est un exemple de mapping de type 1..n :

- $M_{SO}3$ $(StadeTNM_{OA}, \{TailleT_{P1}, Noeud_{P1}, Metas_{P1}\})$ permet de faire la correspondance entre le concept *StadeTNM* de l'ontologie d'application OA et les descripteurs *TailleT*, *Noeud* et *Metas* du schéma du partenaire P1.

Les descripteurs des schémas sont aussi liés par des mappings 1..1 schéma-ontologie avec la branche des formats de données. Par exemple, au sein de notre SIGL, le descripteur de schéma *Date_Naissance* est mis en correspondance avec le concept *DateNaissance* et avec

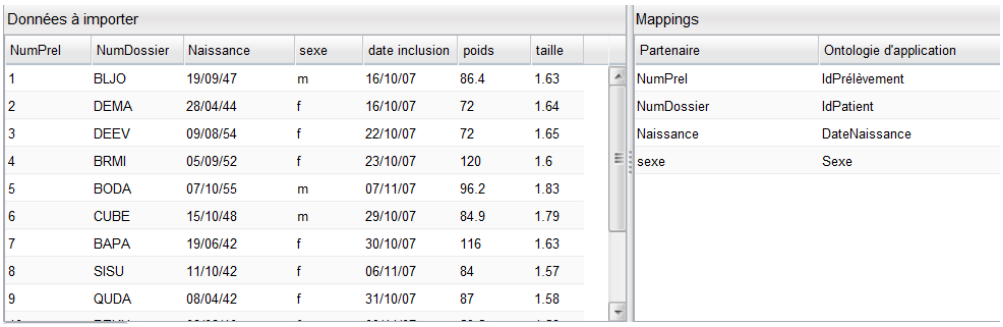


FIG. 7: Création des mappings schéma-ontologie au sein de notre prototype.

le concept désignant le format `JJ-MM-AAAA` alors que la date de naissance du schéma du clinicien C1 (*DNaissance*) est mis en correspondance avec le concept *DateNaissance* et avec le concept désignant le format `JJ-mmm-AAAA`. Nous avons donc deux types de mappings schéma - ontologie : les mappings définissant la signification des descripteurs et ceux définissant leur format. L'utilisation conjointe de ces deux types de mappings nous permet de retrouver quelle est la séquence de fonctions de conversion nécessaires pour passer de la valeur de *DNaissance* provenant du clinicien C1 (`JJ-mmm-AAAA`) à la valeur de *Date_Naissance* adéquate pour le SIGL (`JJ-MM-AAAA`).

Chaque schéma des partenaires comporte ses spécificités. L'entrée d'un nouveau partenaire dans ce système nécessite seulement de réaliser les mappings schéma-ontologie entre les descripteurs du nouveau schéma et l'ontologie d'application. Les mappings schéma-ontologie des autres partenaires ne seront ainsi aucunement impactés par ces changements.

5 Mise en œuvre de l'approche

Comme nous l'avons vu, la mise en œuvre de notre approche comporte trois niveaux de contrôle. Le premier niveau consiste en la création des tuples sémantiques correspondant aux données (Nardini et al. (2011)). Un tuple sémantique est une représentation des données sous la forme d'individus au sein de l'ontologie d'application. Ils permettent de raisonner sur les données au niveau structurel et logique. La création de ces tuples s'appuie sur les mappings spécifiant la sémantique des descripteurs de schéma, le domaine et le format des données. Le deuxième niveau consiste à vérifier la complétude et la consistance des tuples en accord avec le schéma du SIGL. Le troisième et dernier niveau consiste à vérifier la cohérence des tuples selon le contexte métier et la connaissance du domaine. La figure 8 synthétise les différents niveaux de notre approche, pour des raisons de lisibilité, nous n'avons pas fait apparaître les mappings présents sur la figure 2. Lorsque la plateforme protéomique reçoit les données, elle ne reçoit pas des copies conformes des schémas des systèmes des partenaires, mais des extraits de ces schémas comportant uniquement les descripteurs pertinents des données.

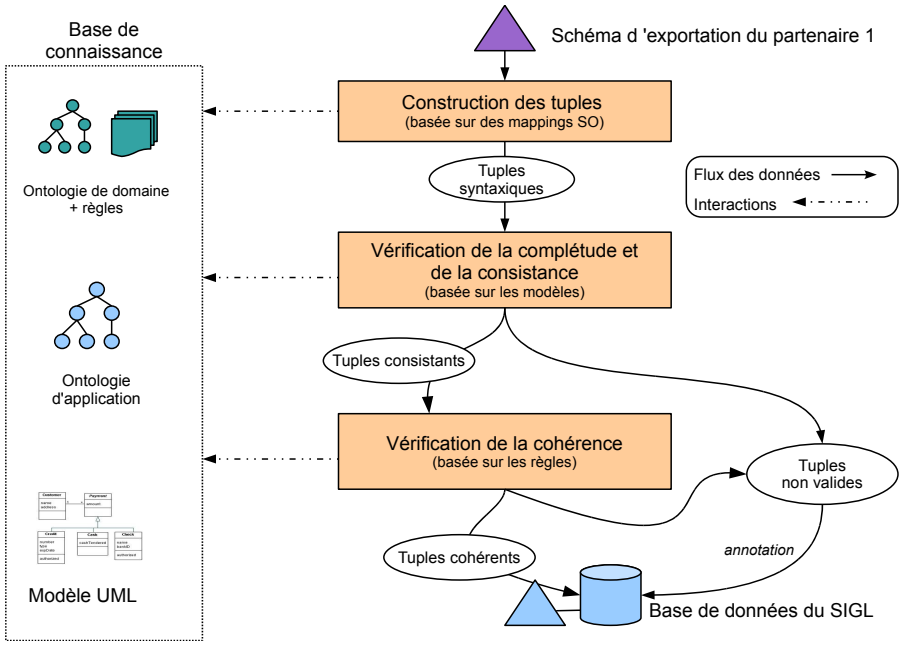


FIG. 8: Flux des données au sein de notre approche.

Détails sur les trois niveaux

La première vérification concerne la sémantique des descripteurs, le domaine et le format des données. Elle utilise les deux types de mappings schéma-ontologie pour déterminer la sémantique de chaque descripteur ainsi que son format. La comparaison des mappings réalisés sur le schéma du SIGL à ceux réalisés sur les schémas des partenaires nous permet de découvrir les correspondances entre les descripteurs des schémas. Cette comparaison donne les opérations de conversion nécessaires à la transformation des valeurs des données des partenaires vers les valeurs acceptées au sein du SIGL. Chaque valeur provenant d'un partenaire est transformée dans le bon format et attribuée à un tuple correspondant au schéma de notre SIGL. À la fin de cette étape, nous obtenons ce que nous appelons des tuples syntaxiques.

Une fois les tuples correspondant aux données créés, nous pouvons en vérifier la complétude et la consistance. L'utilisation du diagramme de classes UML comme modèle structurel de notre système permet de spécifier les associations facultatives ou obligatoires entre les tuples. Ainsi nous pouvons repérer les erreurs d'association entre les tuples. Nous pouvons aussi vérifier la consistance de certaines données au sein des tuples. Si nous reprenons l'exemple du paragraphe « Consistance des données », nous détectons que le patient HG65 a déjà été créé avec un sexe différent. Le matériel biologique étant rare, nous ne pouvons pas nous permettre de rejeter les données, les tuples non valides sont insérés dans la base de données avec une annotation. La résolution de ce problème sera réalisée manuellement avec l'aide des différents partenaires. Pour notre exemple, le clinicien à la source du jeu de données sera interrogé pour déterminer le sexe du patient. L'annotation de donnée non valide empêche l'utilisation de

l'échantillon biologique au sein d'une expérimentation.

Une fois les tuples vérifiés, le moteur de règles prend en compte les faits, c'est-à-dire les tuples nouvellement créés et la connaissance, ainsi que les règles à vérifier. À la fin de ce processus nous obtenons des tuples cohérents qui ont passé les trois vérifications avec succès ou des tuples non valides également annotés.

Architecture applicative

Les travaux présentés dans cet article sont intégrés dans le prototype du module clinique eClims du SIGL open source ePims™ développé conjointement par le CEA de Grenoble, la société *Advanced Solutions Accelerator* [9], la plateforme protéomique CLIPP [10] et le laboratoire LE2I de l'Université de Bourgogne. Des informations supplémentaires sur eClims et des captures d'écrans sont disponibles à l'adresse suivante `http://eclims.u-bourgogne.fr`. L'architecture de notre prototype est composée de trois couches : 1) l'interface utilisateur, 2) les services métiers et 3) la couche de persistance. La persistance est réalisée grâce à une base de données PostgreSQL, un dépôt d'ontologies sous la forme d'un triple store et de règles dans un schéma relationnel. La couche métier est entièrement réalisée en Java avec l'aide du framework Hibernate [11] pour le mapping objet-relationnel et de l'API Java de Pellet [12] pour la gestion du moteur de règles. L'interface de l'application a été réalisée grâce au framework GWT [13]. La figure 9 présente une vue synthétique des différentes couches applicatives composant le prototype eClims.

La figure 10 présente une capture d'écran de données importées au sein du SIGL via notre module d'importation. La partie gauche montre l'organisation hiérachique des données et la partie droite montre le détail des informations du Patient 2.

Évaluation du prototype

Du fait du caractère confidentiel des données présentes sur la plateforme protéomique CLIPP, nous n'avons pu testé notre module d'importation que sur un seul jeu de données. Ce jeu de données présenté sous la forme d'un fichier CSV comporte 345 échantillons et 64 descripteurs de données pertinentes. Nous avons repéré 114 échantillons ne répondant pas aux critères de qualité de la plateforme. Sur ces 114 échantillons seulement 9 d'entre eux ne sont pas cohérents. L'objectif de l'étude protéomique est la recherche de biomarqueurs de rechute sur des patients de sexe féminin souffrant du cancer du sein. Les 9 échantillons incohérents proviennent de patients de sexe masculin ou inconnu. Les 105 échantillons restant présentent des problèmes de complétude ou d'inconsistance. La plupart des échantillons incomplets ne référence aucun patient. La plupart des échantillons inconsistants référence une pathologie ne correspondant à aucune dénomination liée au cancer du sein.

9. ASA `www.advancedsolutionsaccelerator.com`
10. CLinical Innovation Proteomic Platform `www.clipproteomic.fr`
11. Hibernate est un framework open source de persistance des objets en base de données relationnelle. `www.hibernate.org`
12. Pellet est un raisonneur OWL2. `http://clarkparsia.com/pellet`
13. *Google Web Toolkit* est un framework de création d'applications web dynamiques. `http://code.google.com/webtoolkit`

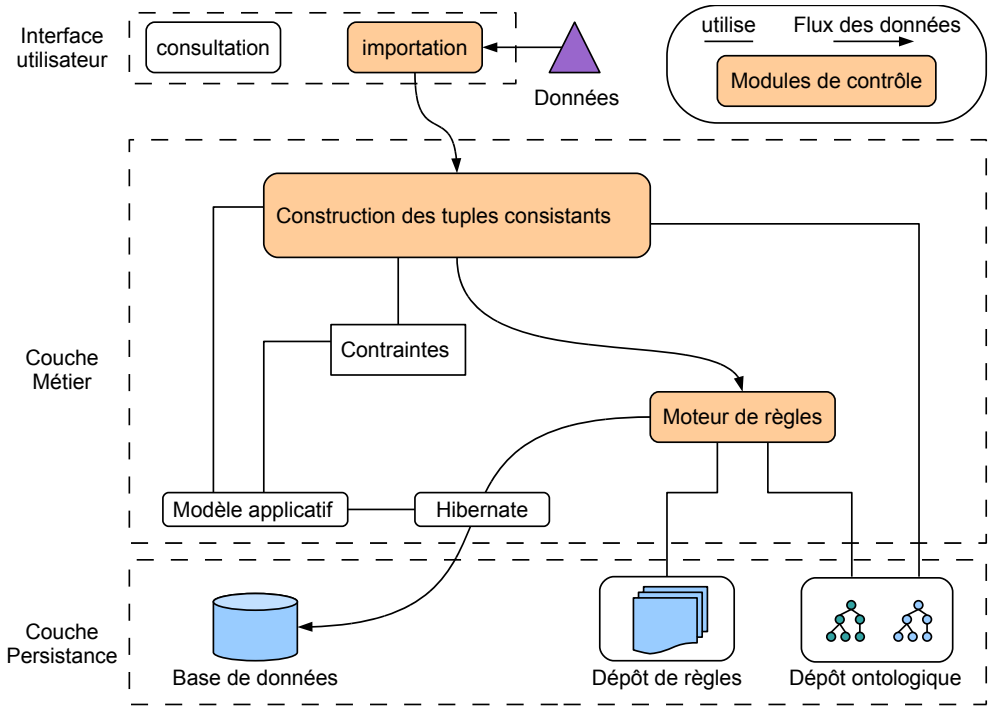


FIG. 9: Couches applicatives du prototype eClims.

6 Conclusion

Le système de gestion de la qualité des données que nous avons décrit dans cette article permet de garantir la qualité initiale des données de protéomique clinique lors de leur importation. Sa mise en œuvre peut nécessiter un gros investissement humain lors de la création des ontologies d'application et de domaine. Cependant cet investissement initial permet de garantir à chaque importation de données provenant d'une même source la même qualité globale. Dans notre approche centrée sur le système cible, le passage à l'échelle de cette méthode est acceptable du fait de la centralisation de l'importation. L'ajout d'une nouvelle source ne nécessite que la réalisation de mapping schéma-ontologie entre le schéma source et l'ontologie d'application. Les perspectives majeures que nous envisageons sur ces travaux concernent :

- la mise en place de règles *SWRL-DL-Safe* pour améliorer la cohérence des données. Les règles *SWRL-DL-Safe*, bien que non contraignantes pour l'utilisateur, nécessitent un gros travail de création des individus au sein de l'ontologie. Une des pistes envisagées est d'améliorer cet aspect par l'implémentation du langage ELP (Krötzsch et al. (2008)) ;
- l'aspect découverte automatique des mappings schéma-ontologie lors de l'ajout d'un nouveau partenaire. Pour cela, nous nous intéressons aux travaux menés sur l'alignement automatique des ontologies (Rahm et Bernstein (2001)).

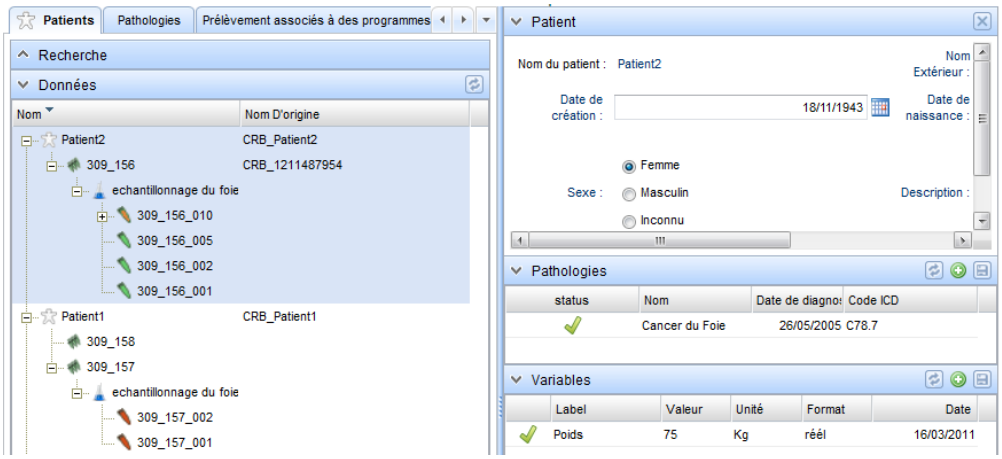


FIG. 10: Arborescence des données du module clinique eClims.

Remerciements

Les auteurs tiennent à remercier la plateforme protéomique CLIPP (CLinical Innovation Proteomic Platform), la société ASA (Advanced Solutions Accelerator) ainsi que le Conseil Régional de Bourgogne pour leur soutien à ces travaux.

Références

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, et G. Sherlock (2000). Gene ontology : tool for the unification of biology. the gene ontology consortium. *Nature genetics vol. 25*(1), 25–29.

Ashenhurst, R. L. (1996). Ontological aspects of information modeling. *Minds and Machines vol. 6*, 287–394.

Belleannée, G. (2006). Le système TNM : trois lettres de base pour un langage riche mais parfois ambigu. *Annales de pathologie vol.26*, 435–479.

Berti-Équille, L. (2007). *Quality Awareness for Data Managing and Mining*. Habilitation à diriger les recherches, Université de Rennes 1, France.

Brusa, G., M. L. Caliusco, et O. Chiotti (2006). A process for building a domain ontology : an experience in developing a government budgetary ontology. In *Proceedings of the second Australasian workshop on Advances in Ontologies*, Darlinghurst, Australia, pp. 7–15.

Chen, J. Y. et J. V. Carlis (2003). Genomic data modeling. *Inf. Syst. vol. 28*, 287–310.

Dasu, T. et T. Johnson (2003). *Exploratory Data Mining and Data Cleaning*. John Wiley.

Davidson, S., C. Overton, et P. Buneman (1995). Challenges in Integrating Biological Data Sources. *Journal of Computational Biology vol. 2*(4), 557–572.

Degoulet, P., M. Fieschi, et C. Attali (1997). Les enjeux de l'interopérabilité sémantique dans les systèmes d'information de santé. *Informatique et gestion médicalisée vol. 9*, 203–212.

Fowler, M. (2003). *UML Distilled : A Brief Guide to the Standard Object Modeling Language* (Third ed.). Boston, MA, USA : Addison-Wesley Longman Publishing Co., Inc.

Goh, C. H. (1997). *Representing and reasoning about semantic conflicts in heterogeneous information systems*. Thèse de doctorat, Massachusetts Institute of Technology, USA.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies vol. 43*(5-6), 907–928.

Guarino, N. (1998). Formal ontology and information systems. In *Proceedings of Formal ontology in information systems*, Volume 46 of *FOIS'98*, Trento, Italy, pp. 3–15.

Hall, J., K. Healy, et R. Ross (2000). *Defining Business Rules : What Are They Really ?* Rapport technique, Business Rules Group.

Han, J. et M. Kamber (2006). *Data mining : concepts and techniques* (Second ed.). Morgan Kaufmann.

Horrocks, I. et P. F. Patel-Schneider (2004). A proposal for an OWL rules language. In *Proceedings of the 13th international World Wide Web Conference (WWW 2004)*, New York, NY, USA, pp. 723–731.

Horrocks, I., P. F. Patel-Schneider, S. Bechhofer, et D. Tsarkov (2005). OWL rules : A proposal and prototype implementation. *Web Semantics : Science, Services and Agents on the World Wide Web vol. 3*, 23–40.

Kim, W. et J. Seo (1991). Classifying schematic and data heterogeneity in multidatabase systems. *Computer vol. 24*, 12–18.

Krötzsch, M., S. Rudolph, et P. Hitzler (2008). ELP : Tractable Rules for OWL 2. In A. P. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. Finin, et K. Thirunarayan (Eds.), *The Semantic Web - ISWC 2008*, Volume 5318 of *Lecture Notes in Computer Science*, pp. 649–664. Springer Heidelberg.

Linster, M. (1992). Viewing knowledge engineering as a symbiosis of modeling to make sense and modeling to implement systems. In H. J. Ohlbach (Ed.), *GWAI*, Volume 671 of *Lecture Notes in Computer Science*, pp. 87–99. Springer.

Motik, B. et R. Rosati (2008). Reconciling description logics and rules. *J. ACM vol. 57*, 30 :1–30 :62.

Motik, B., U. Sattler, et R. Studer (2005). Query Answering for OWL DL with rules. *Web Semantics vol. 3*(1), 41–60.

Naiman, C. F. et A. M. Ouksel (1995). A classification of semantic conflicts in heterogeneous database systems. *J. Organ. Comput. vol. 5*, 167–193.

Nardini, E., A. Omicini, M. Viroli, et M. Schumacher (2011). Coordinating e-health systems with tucson semantic tuple centres. *ACM Applied Computing Review vol. 11*(2), 43–52.

Rahm, E. et P. A. Bernstein (2001). A survey of approaches to automatic schema matching. *The VLDB Journal vol. 10*, 334–350.

Redman, T. C. (2001). *Data quality : the field guide*. Newton, MA, USA : Digital Press.

Ross, R. G. (2003). *Principles of the Business Rule Approach*. Boston, MA, USA : Addison-Wesley Longman Publishing Co., Inc.

Salem, S. et S. AbdelRahman (2010). A multiple-domain ontology builder. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, Stroudsburg, PA, USA, pp. 967–975. Association for Computational Linguistics.

Shvaiko, P. et J. Euzenat (2008). Ten challenges for ontology matching. In R. Meersman et Z. Tari (Eds.), *On the Move to Meaningful Internet Systems : OTM 2008*, Volume vol. 5332 of *Lecture Notes in Computer Science*, pp. 1164–1182. Springer Berlin / Heidelberg.

Siegel, M. et S. E. Madnick (1991). A metadata approach to resolving semantic conflicts. In *Proceedings of the 17th International Conference on Very Large Data Bases*, VLDB '91, San Francisco, CA, USA, pp. 133–145. Morgan Kaufmann Publishers Inc.

Spear, A. D. (2006). *Ontology for the twenty first century : An introduction with recommendations*. Institute for Formal Ontology and Medical Information Science, Saarbrücken, Germany.

Starr, L. (2001). *Executable Uml : How to Build Class Models*. Upper Saddle River, NJ, USA : Prentice Hall PTR.

Sugumaran, V. et V. C. Storey (2002). Ontologies for conceptual modeling : their creation, use, and management. *Data Knowl. Eng. vol. 42*, 251–271.

Van Heijst, G., A. T. Schreiber, et B. J. Wielinga (1997). Using explicit ontologies in KBS development. *Int. J. Hum.-Comput. Stud. vol. 46*, 183–292.

Wiederhold, G. (1994). Interoperation, mediation, and ontologies. In *Proceedings International Symposium on Fifth Generation Computer Systems (FGCS94), Workshop on Heterogeneous Cooperative Knowledge-Bases*, Volume 3, pp. 33–48.

Willson, S. J. (1998). Measuring inconsistency in phylogenetic trees. *J Theor Biol vol. 190*, 15–36.

Summary

Biomedical domain and proteomics in particular are faced with an increasing volume of data. The heterogeneity of data sources implies heterogeneity in the representation and in the content of data. Data may also be incorrect, implicate errors and can compromise the analysis of experiments results. Our approach aims to ensure the initial quality of data during import into an information system dedicated to proteomics. It is based on the joint use of models, which represent the system sources, and ontologies, which are use as mediators between them. The controls, we propose, ensure the validity of values, semantics and data consistency during import process.