Méthodes de mapping situées aux niveaux instance et schéma pour l'intégration de sources de données hétérogènes

Fleur Mougin, Julie Chabalier, Olivier Bodenreider, Anita Burgun

Méthodes de mapping situées aux niveaux *instance* et *schéma* pour l'intégration de sources de données hétérogènes

Fleur Mougin[1], Julie Chabalier[1], Olivier Bodenreider[2] et Anita Burgun[1]

[1] EA 3888, IFR 140, Faculté de Médecine, Université de Rennes I, France
{fleur.mougin, anita.burgun, julie.chabalier}@univ-rennes1.fr
[2] National Library of Medicine, Bethesda, Maryland, USA
olivier@nlm.nih.gov

Résumé : Un des obstacles principaux à l'intégration de sources de données est l'hétérogénéité des schémas des sources et en particulier la disparité des éléments de données (EDs) associés. Dans ce cadre, nous proposons des approches visant à intégrer les EDs de onze sources biomédicales. Tout d'abord, au niveau *schéma*, nous cherchons à mettre en correspondance les EDs avec des concepts d'une ressource terminologique biomédicale de référence : l'UMLS. En pratique, nous réalisons un mapping direct et un mapping via une ressource externe, WordNet. Nous proposons ensuite des méthodes au niveau *instance* en exploitant les valeurs associées aux EDs au sein des sources. Nous montrons en quoi le niveau *instance* permet de compléter les mappings obtenus au niveau *schéma* et nous soulignons son aspect essentiel pour considérer la sémantique des EDs.

Mots-clés : bioinformatique, représentation des connaissances, système d'intégration, hétérogénéité sémantique, mapping de schémas.

1 Introduction

La problématique d'intégration de sources de données hétérogènes s'inscrit dans de nombreux domaines. Différentes approches ont été proposées pour construire des systèmes permettant de faciliter au mieux le travail de recherche et de collecte d'informations par les utilisateurs. Cependant, les systèmes d'intégration existants sont souvent créés de manière manuelle, ce qui rend la gestion de leur maintenance particulièrement délicate. Il est pourtant indispensable de pouvoir faire évoluer ce type de système en fonction des besoins des utilisateurs d'une part et de l'intégration de nouvelles données d'autre part. Dans ce cadre, des méthodes automatiques pour réaliser les tâches nécessaires à la création et/ou la maintenance des systèmes d'intégration doivent être mises en œuvre. En particulier, dans les systèmes de médiation, la mise en correspondance (ou *mapping*) des éléments présents dans les schémas des sources avec ceux du schéma global peut être en partie automatisée.

En général, les sources sont développées indépendamment les unes des autres et ne partagent donc ni dictionnaire, ni schéma commun. Leur hétérogénéité se situe

notamment au niveau des attributs, aussi connus sous le nom d'éléments de données (EDs) selon la norme ISO/IEC 11179. Ils peuvent être définis ainsi : "une unité d'information de base construite sur des structures standard ayant un sens unique et des valeurs distinctes[1]". Les ensembles de valeurs associées aux EDs correspondent aux *instances*. En pratique, l'obstacle principal à l'intégration des sources de données est l'hétérogénéité de leurs schémas et en particulier la disparité des EDs associés.

Dans le cadre du développement d'un système de médiation de sources de données biomédicales, nous avons considéré l'automatisation du mapping des EDs à un schéma global utilisant des ressources terminologiques existantes. Nous présentons ici les méthodes que nous avons mises en œuvre pour cela. Plus précisément, nous considérons tout d'abord le niveau *schéma* en cherchant des correspondances entre les EDs de sources biomédicales et des concepts présents dans une terminologie biomédicale de référence : l'UMLS. Ensuite, nous exploitons le niveau *instance* et montrons en quoi cela permet de compléter les correspondances obtenues au niveau *schéma*.

2 État de l'art

Le cadre général de ce travail est l'intégration de sources de données au travers de l'identification de correspondances entre leurs schémas. Ce processus considère deux ensembles d'éléments en entrée (attributs ou EDs et valeurs) constituant deux schémas et détermine les relations (e.g. équivalence, subsomption) existant entre des paires d'éléments des deux schémas. De nombreuses approches ont été développées et classées suivant différents critères (Rahm & Bernstein, 2001) et (Shvaiko & Euzenat, 2005). La différence principale entre ces méthodes concerne le niveau auquel elles sont appliquées. Plus précisément, elles se situent au niveau *schéma* lorsqu'elles n'exploitent que les informations existant dans le schéma des sources, ou au niveau *instance*, si elles utilisent les valeurs associées aux EDs.

À ces deux niveaux, nous détaillons deux groupes principaux de méthodes existantes pour réaliser le mapping : méthodes lexicales et structurelles. Les approches lexicales au niveau *schéma* ont été développées pour mettre en correspondance les libellés des EDs en exploitant leur morphologie. Par exemple, une petite distance d'édition (somme minimale du coût des opérations élémentaires, e.g. l'ajout ou la suppression d'un caractère, pour transformer une chaîne de caractères en une autre) ou une grande proportion de N-grammes communs (sous-séquence de *N* caractères - ou mots - construite à partir d'une séquence donnée) entre deux chaînes de caractères révèlent une ressemblance lexicale (Do & Rahm, 2002). Les méthodes structurelles considèrent notamment les schémas comme des graphes et appliquent des approches classiques de comparaison de graphes, e.g. en déterminant la similitude entre des nœuds ayant des ancêtres ou descendants communs (Ehrig & Sure, 2004).

Les informations disponibles sur les schémas étant parfois insuffisantes ou ambiguës, il est indispensable d'exploiter les informations situées au niveau *instance*.

[1] http://www.atis.org/tg2k/_data_element.html

Ici aussi, des méthodes lexicales peuvent être utilisées pour mettre en correspondance des instances avec des ressources externes (Xu & Embley, 2003). Des approches structurelles peuvent également être appliquées si des contraintes existent sur les EDs (Miller *et al.*, 2001). Par exemple, il est possible d'identifier la portée des valeurs associées si ce sont des données numériques, ou encore des termes récurrents dans le cas de données textuelles. Notons que des techniques d'apprentissage ont également été utilisées pour mettre en correspondance des EDs via leur ensemble de valeurs (Doan *et al.*, 2004). L'inconvénient de ces dernières est qu'elles nécessitent généralement de créer un jeu d'entraînement avant de pouvoir être appliquées.

Dans le domaine biomédical, de nombreux systèmes d'intégration ont été développés (Hernandez & Kambhampati, 2004) mais le problème est qu'ils ont généralement été conçus de manière manuelle, e.g. TAMBIS (Stevens *et al.*, 2000). La gestion de leur maintenance est ainsi problématique puisqu'aucune tâche ne peut être automatisée. Récemment, certains systèmes se sont malgré tout penchés sur ces questions, e.g. BACIIS (Ben-Miled *et al.*, 2004), mais uniquement de manière partielle. Le problème est que ces systèmes ignorent la sémantique des EDs. Par exemple, des EDs de même nom, e.g. Symbol, sont automatiquement mis en correspondance alors que l'un d'eux pourrait concerner des symboles de gènes et l'autre des symboles de protéines.

C'est pour pallier les limites des systèmes existants en terme d'automatisation des tâches de conception et de maintenance que nous avons réalisé ce travail. Nous montrons en quoi l'exploitation du niveau *instance* est indispensable pour compléter, voire corriger les mappings obtenus au niveau *schéma*.

3 Matériels et méthodes

3.1 Ressources terminologiques

L'**UMLS**® (Unified Medical Language System®) (Lindberg *et al.*, 1993) est un système incluant trois composants : le Metathesaurus®, le Réseau Sémantique et des outils lexicaux (McCray *et al.*, 1994). Le Metathesaurus, intégrant plus de 100 vocabulaires sources, est constitué de plus d'un million de nœuds, appelés concepts, et de 16 millions de relations entre ces concepts. Chaque concept regroupe les termes synonymes provenant des vocabulaires sources. Le Réseau Sémantique est un réseau beaucoup plus restreint de 135 types sémantiques organisés de manière arborescente. Chaque concept du Metathesaurus est catégorisé par au moins un type sémantique du Réseau Sémantique. Les outils lexicaux de l'UMLS offrent différentes méthodes pour identifier des concepts dans des textes par des recherches exactes et normalisées (e.g. en supprimant des différences telles que l'inflexion, la casse, la ponctuation ou encore des variations dans l'ordre des mots). Le programme MetaMap (Aronson, 2001) permet d'extraire des concepts d'un texte en exploitant les variantes des termes.

WordNet (WN ; Miller, 1998) est une base de données lexicale de langue anglaise. Noms, verbes, adjectifs et adverbes sont organisés en ensembles de

synonymes, appelés synsets, chacun représentant un concept. Elle contient plus de 155 000 items lexicaux regroupés dans près de 117 000 synsets.

3.2 Éléments de données et valeurs associées

Notre ensemble de test est constitué d'éléments de données (EDs) extraits de onze sources de données biomédicales vérifiant les critères suivants : les sources doivent être accessibles sur Internet, complémentaires et fournir des informations concernant des entités biomédicales de référence. Onze sources génomiques[2], protéomiques[3] et médicale[4] ont ainsi été choisies.

La méthode utilisée pour extraire les EDs est automatique et exploite la redondance structurelle des sources d'information (i.e., la répétition des EDs alors que les valeurs qui leurs sont associées changent), également utilisée par RoadRunner (Crescenzi *et al.*, 2001). En pratique, pour extraire les EDs, nous avons interrogé chacune des 11 sources dynamiquement à partir d'une liste de 100 noms et symboles de gènes extraits de manière aléatoire sur le site Web du Genetics Home Reference[5] (Mougin *et al.*, 2004). Pour chaque source, un ensemble de 100 pages HTML partageant la même structure a été obtenu. Les éléments communs à au moins 75% des 100 pages HTML ont été extraits. Cela permet d'éliminer l'information spécifique (e.g. "Inhibin, beta B") tout en gardant l'information d'ordre général (e.g. le terme "Nom de gène"). Des heuristiques sont ensuite appliquées, imposant par exemple une longueur maximale de 50 caractères pour les EDs. L'ensemble des EDs restants après ce filtrage sont gardés et considérés comme pertinents. Au total, 548 EDs (474 distincts) ont été extraits des onze sources de données. Ces EDs sont souvent ambigus et sans savoir dans quel contexte ils s'expriment, on ne peut pas les mettre en correspondance directement avec un concept commun. Une illustration des EDs extraits de la source HGNC est donné Fig. 1.

Cette méthode permet également d'extraire les valeurs associées aux EDs qui correspondent aux éléments variables d'une page à l'autre pour un ED donné. Des exemples de valeurs associées à l'ED `Approved Name` (HGNC) sont "Breast Cancer 1, early onset", "Inhibin, beta B" et "Tenascin XB" (Fig. 1). Notons que certaines valeurs peuvent être vides, e.g. l'ED `Previous Symbols` qui n'a pas de valeur associée pour les gènes "BRCA1" et "INHBB".

3.3 Méthodes de mapping

3.3.1 Mapping au niveau *schéma*

[2] GeneCards (http://bioinformatics.weizmann.ac.il/cards/), Entrez Gene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene), Geneloc (http://genecards.weizmann.ac.il/geneloc/), HGNC (http://www.gene.ucl.ac.uk/nomenclature/), HGMD (http://www.hgmd.org/) et MGI (http://www.informatics.jax.org/)

[3] Swiss-Prot (http://www.expasy.org/sprot/), PDB (http://www.rcsb.org/pdb/), HPRD (http://www.hprd.org/) et Interpro (http://www.ebi.ac.uk/interpro/)

[4] OMIM (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM)

[5] http://ghr.nlm.nih.gov/

Mapping direct dans l'UMLS

Pour mettre en correspondance les EDs avec des concepts de l'UMLS, une recherche exacte est d'abord réalisée puis si aucun concept n'est trouvé, l'ED est normalisé pour identifier une correspondance proche. Enfin, une correspondance approximative est recherchée avec MetaMap pour tous les EDs auxquels aucun concept n'a été associé. Différents types de correspondances sont ainsi obtenus :

- *unique* ou de cardinalité 1-1, i.e. un ED est mis en correspondance avec un unique concept UMLS. Par exemple, mRNA sequence est associé au concept RNA, Messenger ;
- *multiple* ou de cardinalité 1-n, i.e. à un ED sont associés plusieurs concepts UMLS. Par exemple, l'ED Interactions est mis en correspondance avec deux concepts : Social Interactions et Drug Interactions ;
- *aucune correspondance*. Certains EDs, tels que Topology, Product, Keywords et Domains, sont simplement absents des terminologies médicales intégrées dans l'UMLS.

Les résultats sont exploitables tels quels pour les correspondances uniques mais incomplets dans les autres cas. Les correspondances multiples nécessitent d'être désambiguïsées et pour les EDs non trouvés dans l'UMLS, il est nécessaire d'utiliser une ressource externe permettant de les mettre en correspondance de manière indirecte dans l'UMLS.

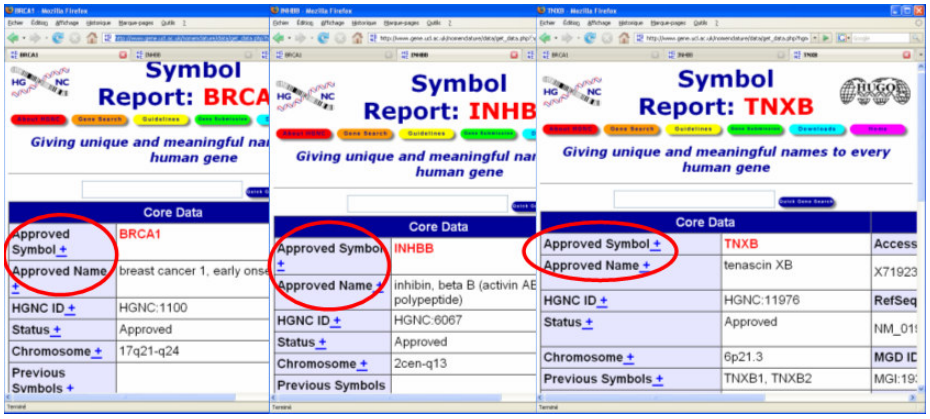

Fig. 1 - Pages Web obtenues par interrogation de la source HGNC pour les gènes "BRCA1", "INHBB" et "TNXB". Des exemples d'EDs sont entourés dans chaque page.

Mapping indirect via WordNet

Pour mettre en correspondance les EDs avec des synsets de WN, nous avons utilisé le programme *wn* permettant d'y faire une recherche lexicale, i.e., de vérifier l'existence d'un syntagme associé à chaque ED dans WN. Quand un ED est constitué de plus d'un mot, la correspondance couvrant le plus long syntagme est sélectionnée. Si des correspondances multiples dans WN sont identifiées, les synsets du domaine

biomédical sont sélectionnés (i.e. de domaines prédéfinis comme *Biology*, *Medicine* ou dont la définition (*gloss*) contient des mots spécifiques du domaine).

Pour mettre en correspondance les synsets (préalablement associés aux EDs) avec des concepts de l'UMLS, nous avons réalisé une recherche exacte puis normalisée. Là encore, les correspondances obtenues sont de différentes cardinalités. Afin de désambiguïser les correspondances multiples, nous proposons des méthodes terminologiques et structurelles pour associer des concepts et des synsets en comparant leurs propriétés par rapport aux critères suivants : 1) similitude des définitions, 2) présence de synonymes communs, et 3) présence d'ancêtres communs. Il suffit qu'un des critères soit vérifié pour associer un concept à un synset. Pour le critère 1, les définitions sont d'abord découpées en mots. Les mots les plus fréquents dans les définitions sont éliminés car ils n'apportent pas d'information pertinente (e.g. *a, of, or, that, something* et *be)*. Ensuite, chaque mot constituant les définitions est normalisé grâce au logiciel TreeTagger[6]. Enfin, les définitions sont comparées mot à mot et leur similitude est déterminée par le coefficient de Dice (Rasmussen, 1992) :

$$Sim_{Dice} = \frac{NbMotsCommuns*2}{NbMotsTotal} \quad (1)$$

Les mots non pertinents ayant été éliminés, le critère 1 est vérifié si ce coefficient est différent de 0 (i.e. si les définitions ont au moins un mot significatif en commun). Pour les critères 2 et 3, des concepts UMLS sont associés aux synonymes et hypernymes de WN grâce aux méthodes de recherches exacte et normalisée. La présence de synonymes ou d'ancêtres communs entre le concept et le synset permet de vérifier respectivement les critères 2 et 3.

Comparaison des deux approches

L'approche directe est plus pertinente quand elle permet de trouver l'ED tel quel dans l'UMLS alors que l'approche via WN identifie souvent plusieurs synsets (et donc plusieurs concepts). Par exemple, Northern Blot[7], qui existe dans l'UMLS, est mis en correspondance partiellement dans WN au travers des deux mots *northern* et *blot* (six synsets au total).

L'approche indirecte présente différents avantages par rapport à l'approche directe :

- pour les EDs qui n'ont pas pu être trouvés directement dans l'UMLS, l'approche via WN propose deux solutions : si un des synonymes ou hypernymes directs des synsets associés à ce type d'ED peut être mis en correspondance avec un concept UMLS alors celui-ci est candidat ;
- pour les correspondances multiples obtenues par l'approche directe, l'approche via WN permet parfois de les désambiguïser. Cela est possible si l'on parvient à déterminer si une paire (concept, synset) est meilleure que les autres (suivant les trois critères présentés précédemment). Le concept faisant partie de cette paire est alors sélectionné et la correspondance devient unique ;

[6] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

[7] Technique employée en biologie moléculaire pour étudier l'expression de gènes

- pour les correspondances uniques obtenues par l'approche directe, l'approche via WN permet parfois de les valider. Pour cela, on vérifie que le concept identifié est le même que celui obtenu avec l'approche directe. Cela peut notamment être utile pour les EDs comportant des acronymes, que les outils lexicaux de l'UMLS interprètent parfois de manière erronée.

3.3.2 Mapping au niveau *instance*

Typage des valeurs associées à chaque élément de données

On dispose pour chaque source d'un ensemble d'EDs auxquels sont associées entre 0 et 100 valeurs, que l'on va chercher à mettre en correspondance dans l'UMLS.

Pour cela, nous effectuons des recherches exactes et normalisée pour chaque valeur. On obtient ainsi 1) un ensemble de concepts associés à chaque ED et 2) les types sémantiques catégorisant ces EDs. Ensuite, le type sémantique permettant de catégoriser au moins 50% des concepts correspondant aux valeurs est retenu car il permet de préciser le sens des EDs. Par exemple, notre méthode détermine ainsi que l'ED `Approved Name` réfère à des noms de **gènes** (et non pas de protéines par exemple) puisque la majorité de ses valeurs sont catégorisées par le type sémantique *Gene or Genome* (Fig. 2 (a)). Il est ainsi possible d'identifier une nouvelle correspondance entre l'ED `Approved Name` et le concept UMLS `Genes`.

Cette méthode n'est efficace que dans les cas où les valeurs des EDs appartiennent à l'UMLS, ce qui est limitatif. Nous proposons donc une solution alternative pour ces cas. Nous cherchons ainsi à assigner des types plus "larges" que nous avons choisis par rapport aux valeurs qui existent dans les EDs de notre ensemble :

- les EDs dont les valeurs contiennent des termes tels que "ID(s)", "identifier" ou "accession", sont tout d'abord isolés et typés en tant que *Identifier* ;
- les EDs dont les valeurs sont des chaînes de caractères constituées uniquement des lettres "A", "T", "G", "C" sont typés *Sequence* (Fig. 2 (b)) ;
- les EDs restants sont typés *Integer* ou *String* en fonction de leurs valeurs.

Comparaison des valeurs de paires d'éléments de données

Des EDs issus de différentes sources, et contenant des valeurs similaires, peuvent être associés et permettre parfois de trouver de nouvelles correspondances. Pour identifier ce type de cas, nous proposons de comparer l'ensemble de valeurs pour chaque paire d'EDs issus de sources différentes. Nous avons choisi pour cela d'utiliser l'indice de Jaccard qui détermine la similarité entre deux ensembles de valeurs de cardinalité respective c_1 et c_2 (Van Rijsbergen, 1979), définie par :

$$Sim_{Jaccard} = \frac{c_1 c_2}{c_1 + c_2 - c_1 c_2} \quad (2)$$

$c_1 c_2$ correspondant à la cardinalité de l'ensemble de valeurs communes.

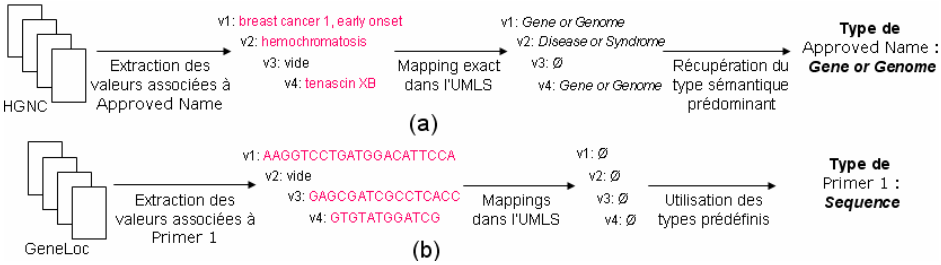

Fig. 2 - Exemples de typage d'EDs au travers de leurs valeurs. (a) Approved Name est typé par le type sémantique *Gene or Genome*. (b) Primer 1 est typé avec *Sequence*

4 Résultats

4.1 Mapping au niveau *schéma*

Mapping direct dans l'UMLS

Parmi les 474 EDs distincts extraits des onze sources biomédicales considérées dans cette étude, 387 ont été mis en correspondance directement dans l'UMLS. Plus précisément, 187 correspondances uniques et 200 multiples ont été identifiées. Ainsi, 87 EDs (18,4%) n'ont pas pu être mis en correspondance avec un concept de l'UMLS.

Mapping indirect via WordNet

Parmi les 474 EDs, 394 ont été mis en correspondance dans WN. Dans un deuxième temps, parmi ces 394 EDs trouvés dans WN, 339 ont été mis en correspondance indirectement dans l'UMLS. Parmi les 87 EDs non trouvés dans l'UMLS avec la méthode directe, 36 ont pu l'être via WN. Ainsi, le nombre d'EDs trouvés dans l'UMLS passe de 387 à 423.

Apports de l'approche directe

Pour 89 cas de correspondances uniques dans l'UMLS, l'approche directe est plus pertinente. Par exemple, l'ED All beta proteins est mis en correspondance avec le concept UMLS Beta Protein de manière unique alors qu'il est associé à quatre synsets.

Apports de l'approche indirecte

L'approche indirecte a permis :

- d'identifier 36 nouvelles correspondances d'EDs dans l'UMLS dont 16 grâce aux synonymes présents dans WN et 20 via les hypernymes directs (Fig. 3 (a) et (b), respectivement) ;
- de désambiguïser 95 des 200 correspondances multiples d'EDs trouvées dans l'UMLS avec l'approche directe (Fig. 3 (c));
- de valider 98 des 187 correspondances uniques obtenues dans l'UMLS avec l'approche directe (Fig. 3 (d)).

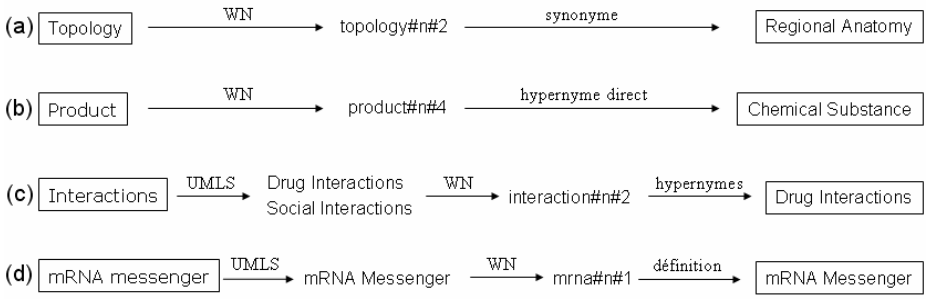

Fig. 3 - Exemples pour les cas où l'approche indirecte est plus efficace que l'approche directe. (a) et (b) identification de nouvelles correspondances, (c) désambiguïsation de correspondance multiple, (d) validation de correspondance unique

Validation

L'approche indirecte permet de valider de manière automatique 193 correspondances[8] sur les 423 trouvées au total dans l'UMLS. Ainsi, les 230 correspondances restantes nécessitent une intervention humaine. Nous avons vérifié manuellement leur validité et compte tenu de la redondance existant entre les EDs issus de sources distinctes, nous avons eu à valider uniquement 52 correspondances uniques et 92 multiples. Au total, nous sommes parvenus à mettre en correspondance 394 EDs dans l'UMLS (seules 23 correspondances étaient incorrectes).

4.2 Mapping au niveau *instance*

Typage des valeurs associées à chaque élément de données

D'une manière globale, seuls 62 EDs (11,3% de l'ensemble) ont pu être caractérisés par des types autres que *String* (Table 1).

Table 1. Résultats obtenus pour le typage des éléments de données extraits des sources.

Type	**Nombre d'EDs de ce type**	**Exemples d'EDs**	**Exemples de valeurs associées à cet ED**
Type sémantique	36	From *(Organism)*	Rattus norvegicus, Homo sapiens
Integer	18	Molecular Weight	207732, 464482
Identifier	6	Accession Numbers	U14680, X71923
Sequence	2	Primer 2	GAGATCGCCTCACC
String	486	Bibliography	(Earliest) J:31493 Hall JM et al., Linkage of early-onset familial breast cancer to chromosome 17q21. Science 1990. 21;250:1684-9

[8] Ce nombre correspond aux cas suivants : 1) 95 correspondances multiples désambiguïsées par l'approche indirecte; 2) 98 correspondances uniques validées par l'approche indirecte avec au moins un des critères de ressemblance vérifié entre concepts et synsets

Comparaison des valeurs de paires d'éléments de données

11 paires d'EDs ont un indice de Jaccard de plus de 0,5. Celles-ci ont permis d'identifier de nouvelles correspondances dans l'UMLS. Par exemple, les valeurs de l'ED Official Symbol (Entrez Gene) sont similaires à celles de l'ED Gene Symbol (HPRD) (indice de 0,55). L'ED Official Symbol contient donc des symboles de **gènes** officiels (et non de protéines par exemple). Une nouvelle correspondance est ainsi identifiée entre cet ED et le concept UMLS Genes. Cet exemple démontre également que cette approche permet de valider des correspondances existantes comme celles de Official Symbol et Gene Symbol avec le concept Symbols (il s'agit bien de symboles).

Par ailleurs, cette approche permet d'éliminer des correspondances identifiées au niveau *schéma*. En effet, l'ED Gene Name (Entrez Gene), et l'ED Approved Symbol (HGNC) ont une similarité de 0,92[9]. Le libellé du premier ED indique qu'il décrit des noms de gènes alors que le second contient des symboles. Il y a donc une incohérence sur le nom d'un des EDs. Or, l'ED Approved Symbol est également mis en correspondance avec les EDs Official Symbol (Entrez Gene) et Gene Symbol (HGMD et HPRD), ce qui signifie que cet ED est correct. Gene Name porte donc un nom inadapté à son contenu. Dans ce cas, cette approche permet deux choses : éliminer la correspondance entre l'ED Gene Name et le concept Names et ajouter une correspondance entre ce même ED et le concept Symbols.

En résumé, cette approche permet de compléter les résultats obtenus au niveau *schéma*. Lorsqu'une incohérence est repérée entre les deux niveaux, c'est la correspondance identifiée au niveau *instance* qui est sélectionnée automatiquement (car l'indice de Jaccard est élevé). Ce choix est fait ainsi car nous considérons que les valeurs associées aux EDs portent une information qui a plus de sens que leurs libellés. Dans les autres cas, les correspondances supplémentaires identifiées au niveau *instance* sont ajoutées systématiquement à celles obtenues au niveau *schéma*.

5 Discussion et conclusion

Nous proposons ici une approche permettant d'automatiser la mise en correspondance des EDs issues de sources biomédicales dans une terminologie commune. Pour cela, nous avons développé des méthodes situées au niveau *schéma* qui ont pu être complétées à l'aide de techniques exploitant le niveau *instance*.

Nos méthodes situées au niveau *schéma* permettent de traiter des correspondances de tout type de cardinalités. La plupart des travaux existants se focalisent principalement sur les correspondances de cardinalité 1-1 (Rahm & Bernstein, 2001), ce qui constitue une limite. Notons cependant que des travaux plus éloignés, tels que Xtab2SML qui vise à découvrir des relations dans des tableaux HTML en comparant les colonnes avec des concepts d'une ontologie de domaine et instancie ces relations avec les valeurs associées (Gagliardi *et al.*, 2006), ont traité des cas de mappings 1-0 et 1-n. Notre approche permet également de résoudre des correspondances de

[9] Sur les 100 pages obtenues lors de l'interrogation de ces sources, ces EDs contenaient chacun 96 valeurs non vides et 92 sont communes aux deux EDs. Leur indice de Jaccard est donc de 0,92.

cardinalité 1-1, 1-n et 1-0 entre les EDs et les concepts de l'UMLS. De plus, nos méthodes permettent de valider et de désambiguïser des correspondances si au moins un des trois critères que nous avons définis est vérifié. Cependant, si la similarité entre un concept et un synset est très basse, il peut être erroné de considérer cette condition comme suffisante pour garantir qu'une correspondance est correcte. Il serait nécessaire de compléter notre travail avec des mesures de similarité robustes en fixant un seuil en dessous duquel les correspondances ne pourraient être acceptées (Kefi *et al.*, 2006). Par ailleurs, une perspective possible pour compléter nos méthodes au niveau *schéma* serait d'intégrer d'autres techniques structurelles basées sur les graphes. En l'occurrence, l'utilisation des relations est potentiellement prometteuse, comme dans (Maedche & Staab, 2002) pour comparer des ontologies. L'UMLS contient différents types de relations dans le Metathesaurus, certaines d'entre elles étant définies de manière formelle (e.g. issues de SNOMED CT[10]). WN contient également des relations pouvant être mises en correspondance avec certaines relations de l'UMLS.

Les méthodes que nous proposons au niveau *instance* permettent non seulement de préciser certains EDs qui sont ambigus ou mal nommés mais aussi d'identifier de nouvelles correspondances. En effet, l'ED From (Swiss-Prot) n'est présent ni dans l'UMLS ni dans WN mais ses valeurs permettent de déterminer qu'il indique l'organisme pour lequel est définie une protéine donnée. 100% de ses valeurs correspondent à des concepts catégorisés par le type sémantique *Organism*, indiquant qu'une correspondance existe avec ce dernier. De plus, ces méthodes permettent d'identifier des correspondances erronées obtenues au niveau *schéma*. Quand deux EDs ont les mêmes valeurs mais que les concepts avec lesquels ils ont été mis en correspondance sont incompatibles, il existe une incohérence au niveau *schéma*. En effet, les instances portent l'information sémantique concernant les EDs, contrairement à leurs libellés dont on exploite l'aspect lexical. Par contre, les méthodes situées au niveau *instance* fournissent pour l'instant peu de résultats. Cela peut s'expliquer par le fait que les sources biomédicales fournissent souvent des données peu structurées. Il est nécessaire de compléter le typage des EDs quand les valeurs ne sont pas présentes dans l'UMLS. Pour cela, il faudrait définir des patrons pouvant identifier un type complexe mais connu d'informations, telles que des dates ou encore des références bibliographiques dont le format est généralement le même. Cela permettrait, par exemple, de typer plus précisément l'ED Bibliography, extrait de Entrez Gene, qui contient des informations bibliographiques comme les EDs Primary Citation ou References, dont les valeurs sont du même style. Une autre perspective serait d'utiliser des techniques d'apprentissage, comme dans (Doan *et al.*, 2003).

Enfin, les méthodes présentées ici sont transposables à d'autres domaines sous condition qu'une ontologie ou au moins une ressource terminologique suffisamment complète (i.e. contenant des concepts organisés en hiérarchie et pour lesquels on dispose de synonymes et définitions) existent pour ces domaines spécifiques. Même si certaines phases restent manuelles, notamment la validation de certaines correspondances, ces approches de mapping permettent cependant de limiter des tâches fastidieuses et délicates de manière significative.

[10] http://www.snomed.org/snomedct/index.html

Remerciements
Ce travail a été financé en partie par le PRIR (Région Bretagne), l'ACI 045507 et le programme intramural de recherche des NIH, National Library of Medicine (NLM).

Références

ARONSON A. (2001) Effective Mapping of Biomedical Text to the UMLS Metathesaurus : The MetaMap Program. Proc AMIA Symp, 17-21

BEN-MILED Z., LI N., LIU Y., HE Y., LYNCH E., BUKHRES O. (2004). On the Integration of a Large Number of Life Science Web Databases. In DILS, 172-186

CRESCENZI V., MECCA G., MERIALDO P. (2001) RoadRunner: Towards Automatic Data Extraction from Large Web Sites. In VLDB, 109-118

DO HH. AND RAHM E. (2002). COMA - a system for flexible combination of schema matching approaches. In VLDB, 610-621

DOAN A., MADHAVAN J., DOMINGOS P., HALEVY A. (2004) Ontology matching : A machine learning approach. Handbook on Ontologies in Information Systems : 397-416

EHRIG M., SURE Y. (2004). Ontology mapping - an integrated approach. In ESWS:76-91

GAGLIARDI H., HAEMMERLE O., PERNELLE N., SAIS F. (2006) Découverte de relations candidates à l'enrichissement d'un entrepôt thématique. Atelier Fouille du web de EGC06

HERNANDEZ T., KAMBHAMPATI S. (2004) Integration of biological sources: current systems and challenges ahead. Proc. ACM Sigmod conf; 33(3):51-60

KEFI H., SAFAR B., REYNAUD C. (2006) Alignement de taxonomies pour l'interrogation de sources d'information hétérogènes. Actes du congrès francophone Reconnaissance des Formes et Intelligence Artificielle (RFIA)

LINDBERG DA., HUMPHREYS BL., MCCRAY AT. (1993) The Unified Medical Language System. Methods Inf Med;32(4):281-291

MAEDCHE A. , STAAB S. (2002) Measuring Similarity between Ontologies. In EKAW, 251-263

MCCRAY AT., SRINIVASAN S., BROWNE AC. (1994). Lexical methods for managing variation in biomedical terminologies. Proc AMIA Symp, 235-239

MILLER G. (1998) Wordnet: An electronic lexical database (language, speech, and communication). The MIT Press

MILLER RJ., HERNANDEZ MA., HAAS LM., YAN L., HOWARD HO CT., FAGIN R., POPA L. (2001) THE Clio project : managing heterogeneity. SIGMOD Rec.,(30)1:78-83

MOUGIN F., BURGUN A., LORÉAL O., LE BEUX P. (2004) Towards the automatic generation of biomedical sources schema. Medinfo;783-787

RAHM E., BERNSTEIN PA. (2001) A survey of approaches to automatic schema matching. In VLDB; 10(4):334-350

RASMUSSEN E. (1992) Clustering algorithms. Information retrieval : data structures and algorithms, 419-442

SHVAIKO P., EUZENAT J. (2005) A survey of schema-based matching approaches. Journal on data semantics; 4:146-171

STEVENS R., BAKER PG., BECHHOFER S., NG G., JACOBY A., PATON NW., GOBLE CA., BRASS A. (2000) TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. Bioinformatics, (16)2:184-186

VAN RIJSBERGEN CJ. (1979). Information retrieval. Butterworth-Heinemann, Newton, USA

XU L., EMBLEY DW. (2003) Discovering direct and indirect matches for schema elements. In DASFAA;39-46