Formalisation des correspondances pour l'optimisation des alignements automatisés de schémas de données : Application au domaine des maladies rares.

Meriem Maaroufi[1,2], Rémy Choquet[1,2], Paul Landais[1,3] et Marie-Christine Jaulent[2]

[1] Banque Nationale de Données Maladies Rares, Hôpital Necker Enfants Malades, Assistance Publique des Hôpitaux de Paris, Paris
meriem.maaroufi@nck-aphp.fr
[2] INSERM, U1142, LIMICS, Paris; Sorbonne Universités, UPMC Université Paris 06, UMR_S 1142, LIMICS, Paris; Université Paris 13, Sorbonne Paris Cité, UMR_S 1142, LIMICS, Villetaneuse
[3] Université Montpellier1, EA2415 & BESPIM, Hôpital Universitaire, Nîmes, France

Résumé : Dans l'ère de partage des données et de l'interopérabilité des systèmes, l'automatisation de l'alignement[1] des schémas de données est devenue une priorité. La découverte des correspondances[2] entre les données est l'objectif de plusieurs approches d'alignement décrites dans la littérature et dont l'efficacité dépend des spécifications des données. Dans ce contexte, nous proposons une méthode pour la formalisation des correspondances qui permet l'optimisation des processus d'alignement automatisé des schémas de données. Cette formalisation, qui intègre le niveau - élément de donnée - ainsi que le niveau - élément de valeur -, permet la déduction automatisée des correspondances exprimées sous forme de règles. Dans cet article, nous commençons par décrire la méthode employée pour parvenir à cette formalisation des correspondances. Nous expliquons par la suite la validation de cette formalisation pour deux cas d'usage. Nous finissons par une discussion des objectifs de la formalisation proposée.

Mots-clés : Intégration de données, alignement de schémas, interopérabilité, formalisation des correspondances, informatique médicale.

1 Introduction

Le projet Banque Nationale de Données Maladies Rares ("Banque Nationale de Données Maladies Rares", 2014) est un projet ambitieux qui vise à identifier les patients souffrant de maladies rares sur tout le territoire français. Cette base de données doit permettre aux institutions d'évaluer l'adéquation entre l'offre et la demande de soins et aux chercheurs d'identifier les patients éligibles de participer à des essais cliniques ou à des cohortes maladies rares. Le projet concerne toutes les maladies rares dont le nombre est estimé entre 6000 et 7000 ("Orphanet", 2012) et doit collecter les données de sources multiples à l'échelle nationale. Les sources de données sont les Centres de Référence et les Centres de Compétence Maladies Rares dont les spécialités couvrent une trentaine de groupes maladies rares.

Connecter ces sources à la base nationale requiert beaucoup d'efforts pour traiter l'hétérogénéité des données. Toutes les sources ne collectent pas les mêmes données, et même si elles partagent certains éléments de données similaires, l'hétérogénéité syntaxique ou

[1] Alignement = processus de détection des correspondances entre les éléments de schémas de données différents. Appelé aussi « Schema matching » (Bellahsene & al., 2011)

[2] Correspondance = relation indiquant une similarité selon une mesure donnée entre deux éléments de deux schémas de données. Appelé aussi « mapping » (Klein, 2001)

sémantique peut persister. Pour surmonter ces problématiques d'interopérabilité, BaMaRa, une application communicante a été conçue pour collecter un jeu de données minimal spécifique aux maladies rares (Choquet & al., 2012). Ce jeu de données recueille principalement des éléments relatifs au diagnostic et à l'activité de soin. Il s'agit jusqu'à présent d'un ensemble de données enrichi avec les types de données et leurs domaines de valeurs respectifs. Des éléments de données standardisés ont été utilisés pour construire cet ensemble de données et des correspondances doivent être identifiées pour connecter les sources qui n'utilisent pas ces standards.

Des approches automatisées d'alignement de schémas ont été proposées dans la littérature afin d'éviter la perte de temps dans les alignements manuels et détecter les similarités entre les éléments de données (ISO & IEC, 2013) construisant les différents schémas (schémas de bases de données, schémas xml, ontologies…).

Comme analysé dans certaines études d'évaluation des techniques d'alignement (Euzenat & Shvaiko, 2013b) (Rahm & Bernstein, 2001) (Kaza & Chen, 2008), l'efficacité des différentes approches dépend des caractéristiques inhérentes aux données, aux schémas et au codage. Dans cet article nous proposons une méthode de caractérisation des correspondances qui permettra la pré analyse des données et l'optimisation de l'application de chaque approche automatisée d'alignement. Nous avons donc étudié la nature des correspondances expérimentales à travers une approche heuristique pour déduire cette caractérisation.

2 Etat de l'art

Les classifications des approches d'alignement automatisés diffèrent dans la littérature (Rahm & Bernstein, 2001) (Euzenat & Shvaiko, 2013a). Toutefois, nous pouvons identifier 4 grandes classes : approches linguistiques, approches niveau structure, approches basées sur les contraintes et les approches niveau instances.

La première approche consiste à comparer à un niveau linguistique les éléments des différents schémas. Que cela soit par détection d'une égalité entre les chaines ou les sous-chaines de caractères ou par mesure de similarité plus complexe entre les libellés des éléments et de leurs descriptions. Cette approche est souvent enrichie par l'utilisation de ressources externes tels que des dictionnaires pour reconnaître les synonymes ou des bases de données enregistrant les anciennes correspondances.

La deuxième approche opère à un niveau structurel. Elle compare les combinaisons des éléments qui donnent des structures complexes. Par exemple, si deux éléments ont les mêmes sous-classes, ils pourraient être mis en correspondance.

La troisième approche est basée sur les contraintes qui définissent par exemple les types de données ou les domaines de valeurs. Combinée à d'autres approches, cette méthode permet de détecter les correspondances erronées ou de confirmer les correctes, mais elle reste incapable de détecter des correspondances par elle même. En effet, deux éléments de données peuvent avoir le même type et la même plage de valeurs sans pour autant faire référence à un même concept, tel est le cas de la date de naissance d'un patient et de la date de l'acte médical dont il a bénéficié.

La quatrième approche est basée sur les instances. Elle est particulièrement utile lorsque des données semi structurées sont traitées et que l'information sur la structure des schémas n'est pas suffisante. Par exemple, la récurrence de l'instance « Cystinose » dans chacun des éléments « Maladie » et « Diag » de deux schémas de données différents peut inférer une correspondance entre les deux.

Chacune de ces approches est plus ou moins adaptée à un type de données. Nous avons choisi la dernière approche pour illustrer certaines limites liées aux spécificités des données (Rahm & Bernstein, 2001).

En utilisant une approche basée sur les instances on admet implicitement que les domaines de valeurs des éléments de données que nous alignons sont similaires (des domaines de valeurs basés sur la même référence ou partageant certains éléments de valeurs

comme la nomenclature Orphanet et OMIM). Par ailleurs, plus le domaine de valeurs est grand plus il devient difficile de détecter les correspondances. Par exemple, il est difficile pour ces outils de détecter une similarité entre des éléments de données contenant des identifiants patients puisqu'il n'y a pas assez de redondance pour un identifiant donné. Les approches basées sur les instances peuvent aussi inférer des correspondances erronées lorsque les éléments traités contiennent des valeurs booléennes.

Ainsi, nous nous sommes intéressés à l'étude des natures des correspondances issues de différents alignements dans le but de trouver un moyen d'optimiser les processus automatisés d'alignement de schémas de données.

3 Méthodes

3.1 Première classification des correspondances

Notre première expérience a été d'intégrer les données issues de la Banque Nationale Alzheimer (BNA) (Le Duff et al., 2010). Suite à un alignement manuel, nous avons obtenu moins de 50% de recouvrement. L'étude des résultats a permis l'identification de cinq différents types de correspondances liant les éléments de données de la source (BNA) à ceux de la cible (BaMaRa) :

- Correspondance exacte : l'élément de la source est lié à un élément de la cible et leurs domaines de valeurs correspondent parfaitement.
  Ex : L'élément de la source « nom de naissance » est lié à l'élément de la cible « nom patronymique » sans transformation de codage.
- Correspondance partielle : l'élément de la source est lié à un élément de la cible mais leurs domaines de valeurs correspondent partiellement.
  Ex : L'élément de la source « patient envoyé par » est lié à l'élément de la cible « patient adressé par » mais leurs listes de codage se chevauchent.
- Correspondance conditionnée : l'élément de la source est lié à un élément de la cible si une certaine condition est vérifiée.
  Ex : L'élément de la source « nom d'usage » est lié à l'élément de la cible « nom marital » seulement lorsqu'il est différent du « nom de naissance ».
- Agrégation : Deux ou plusieurs éléments de la source sont liés à un élément cible.
  Ex : Les éléments de la source « code département » et « code commune » sont agrégés pour donner l'élément de la cible « code pays de naissance ».
- Eclatement : un élément de la source est lié à deux ou plusieurs éléments de la cible.
  Ex : L'élément de la source « type de l'acte » est lié aux trois éléments de la cible « contexte de l'activité », « objectif de l'activité » et « profession du personnel réalisant l'activité ».

Cette classification a permis d'affecter toutes les correspondances obtenues suite à l'alignement des schémas de la BNA et de BaMaRa. Elle a permis de donner une vue d'ensemble sur les relations liant les éléments de données de la source à ceux de la cible et sur les cardinalités impliquées.

Cependant, les classes proposées n'étaient pas souvent disjointes, une correspondance entre des éléments de la source et un élément de la cible peut effectivement être un éclatement conditionné.

Notre démarche a donc été de proposer par la suite une formalisation générique, et non une classification, qui puisse être traitée par la machine et qui prenne en considération aussi bien le niveau éléments de données que le niveau éléments de valeurs. Etant donné un schéma de données de la source, l'objectif est de trouver toutes les correspondances le liant au schéma de données cible. Il s'agit d'un processus unidirectionnel visant à intégrer les données issues de la source au schéma de données cible. Cette approche peut toutefois être appliquée dans l'autre sens afin d'obtenir un alignement bidirectionnel.

3.2 Formalisation des correspondances

Dans ce qui suit nous proposons une méthode de formalisation des correspondances qui prend en compte le niveau élément de donnée, le niveau élément de valeur, et la relation exacte liant les éléments de la source et de la cible.

Soit $S=\{E^S_i;i=1..n;n=card(S)\}$ le set de données source, et $T=\{E^T_j;j=1..m;m=card(T)\}$ le set de données cible, avec E^S_i et E^T_j leurs éléments de données constitutifs. Les domaines de valeurs de E^S_i et E^T_j peuvent être finis (une liste prédéfinie de valeurs) ou infinis (un entier ou du texte). Nous notons e^S_{ik} et e^T_{jl} les éléments de valeurs respectifs de E^S_i et E^T_j. Un élément de valeur peut représenter un item d'un domaine de valeurs fini (e^S_{ik}, k=1..p avec p=card(E^S_i)) ou n'importe quelle valeur d'un domaine infini (pour e^S_{ik}, nous posons k=0 et e^S_{i0} est traité indépendamment de la valeur prise).

E^S_i peut ainsi être lié à E^T_j par une ou plusieurs relations binaires e^S_{ik}-e^T_{jl}. Chaque relation binaire e^S_{ik}-e^T_{jl} est définie par une ou plusieurs règles r. Une correspondance est donc définie pour chaque paire d'éléments de valeurs et non pour chaque paire d'éléments de données.

Pour résumer, une correspondance de S à T peut être caractérisée par le triplet $\{E^S_i$-E^T_j ; e^S_{ik}-e^T_{jl} ; $r\}$:

- Une relation binaire E^S_i-E^T_j liant l'élément de donnée source à l'élément de donnée cible.
- Une relation binaire e^S_{ik}-e^T_{jl} liant un élément de valeur source de E^S_i à un élément de valeur cible de E^T_j.
- Un règle exprimée dans le format « si… alors… ».

Une illustration de ce formalisme est proposée dans la table 1 de la section Résultats.

4 Résultats

La formalisation des correspondances décrite plus haut est un résultat qui a été validé par la caractérisation d'un ensemble de correspondances issues de deux cas d'usage.

Le premier alignement effectué est celui liant la base de données CEMARA (Messiaen et al., 2008) à BaMaRa. CEMARA est une base de données universitaire collectant les données de 240,000 patients souffrant de maladies rares. Elle contient des données de diagnostic et d'actes médicaux réalisés sur 240,000 patients souffrants de maladies rares.

Le deuxième schéma de données à aligner avec le jeu de données de BaMaRa est celui de la BNA. Il s'agit d'une banque nationale créée pour collecter l'ensemble des actes réalisés sur les différents patients souffrant d'Alzheimer. Depuis sa création, elle a cumulé des données de 479,000 patients.

L'objectif des bases CEMARA et BaMaRa étant le même, collecte de données patients maladies rares, leurs schémas de données ne sont pas très différents et partagent un grand nombre d'éléments de données. Chaque correspondance a pu être caractérisée par le triplet : « paire d'éléments de données, paire d'éléments de valeurs et règle » du formalisme proposé.

Ce formalisme a aussi permis la caractérisation des correspondances issues de l'alignement des schémas de données de la BNA et de BaMaRa.

TABLE 1 – *Exemples de triplets (correspondances) issus des deux cas d'usage*

E^S_i - E^T_j	e^S_{ik} - e^T_{jl}	r
"décès" – "status vital"	"oui" - "oui"	si e^S_{ik} ="oui" alors e^T_{jl} ="oui"
"décès" – "status vital"	"non" - "non"	si e^S_{ik} ="non" alors e^T_{jl} ="non"
"venu CPC" – "patient adressé par"	"oui" - "CPC"	si e^S_{ik} ="oui" alors e^T_{jl} ="CPC"
"nom d'usage" – "nom marital"	chaîne de caractères – chaîne de caractères	si $e^S_{i0} \neq e^S_{c0}$ alors $e^T_{j0} = e^S_{i0}$ (E^S_c= "nom de naissance")

De plus, cette formalisation des correspondances a permis la définition d'un processus d'alignement automatisé où les deux paires d'éléments du triplet (éléments de données et les éléments de valeurs) sont les données en entrée et où le troisième élément du triplet, la règle définissant la relation entre ces éléments est le résultat en sortie. La méthodologie a été la suivante :

- Pré-analyse des données : recueil des descriptions des schémas de données source et cible et identification des éléments de données (E^S_i, E^T_j) et des éléments de valeurs correspondants (e^S_{ik}, e^T_{jl}). Création de groupes de données homogènes se basant sur les types de données et la nature des domaines de valeurs.
- Définition des processus : définition des stratégies (processus et algorithmes impliquant une ou plusieurs approches d'alignement de schémas) à mettre en œuvre pour chaque groupe de données et router les données au traitement adéquat.

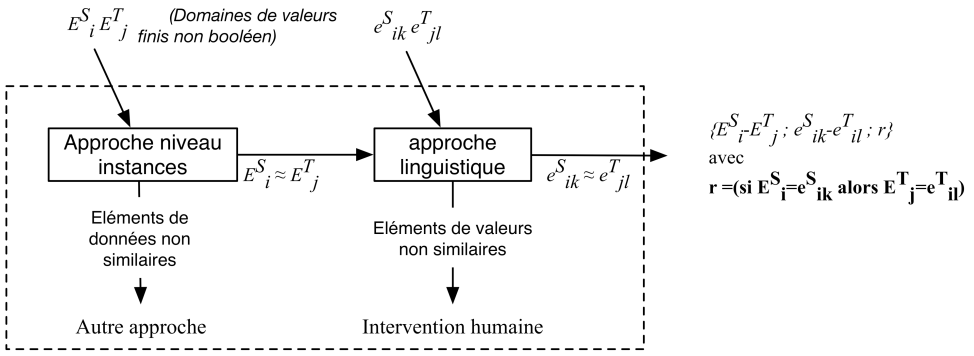

FIGURE 1 – Exemple d'un processus d'alignement optimisé

Un exemple (voir figure 1) d'un processus optimisé peut être le suivant :

- En entrée du processus ne traiter que les sous-ensembles des éléments de données de $\{E^S_i\}$ et $\{E^T_j\}$ qui ont des domaines de valeurs finis et non booléens.
- Appliquer une approche niveau instances ou sous-ensembles $\{E^S_i\}$ et $\{E^T_j\}$ afin de détecter les similarités entre les éléments de données en se basant sur la redondance des instances.
- Pour les paires similaires E^S_i-E^T_j, appliquer une approche linguistique aux ensembles de valeurs correspondants $\{e^S_{ik}\}$ and $\{e^T_{jl}\}$ afin de détecter les paires e^S_{ik}-e^T_{jl}.

Les correspondances issues du processus peuvent être caractérisées par le triplet $\{E^S_i\text{-}E^T_j ; e^S_{ik}\text{-}e^T_{jl} ; r\}$ avec r = (si E^S_i=e^S_{ik} alors E^T_j=e^T_{jl}).

5 Conclusion

La formalisation des correspondances présentée dans ce papier est basée sur une description à deux niveaux des relations entre les éléments de deux schémas de données à aligner. Cette formalisation en triplets, paire éléments de données, paire éléments de valeurs et règle, a été validée par la caractérisation de l'alignement de deux schémas de données avec celui de BaMaRa. Par ailleurs elle nous permet de penser les processus d'alignement de schémas d'une manière plus optimisée pour les raisons suivantes:

- La pré-analyse des données et l'importance des dualités données sources/données cibles et niveau éléments de données/niveau éléments de valeurs, ce qui permettra la construction de groupes de données homogènes.

- La possibilité de proposer des processus fiables d'intégration de données pour chaque groupe de données afin de déduire le troisième élément du triplet ; la règle spécifiant la correspondance.

Le test d'un exemple d'un processus complet constituera l'étape suivante de nos travaux de recherche.

Il est à préciser qu'adopter cette nouvelle méthodologie n'améliorera pas l'efficacité des approches automatiques d'alignement appliquées aux données. Nous n'introduisons pas une nouvelle approche ou un nouvel algorithme qui détectera des correspondances auparavant indécelables sans une intervention humaine à cause de la sémantique. En effet, cette proposition vise à optimiser l'utilisation des approches d'alignement et à limiter les interventions humaines.

Dans la pratique, aligner les schémas en utilisant des approches automatisées reste une tâche supervisée, non seulement pour valider les similarités proposées mais aussi pour choisir le « bon » résultat. En effet, les utilisateurs ont l'habitude d'appliquer les différentes approches à toutes les données, de comparer les résultats et de pondérer selon les spécificités de leurs données. La méthodologie que nous proposons permettra la réutilisation des processus ayant prouvé leur efficacité pour chaque groupe spécifique de données, et introduira une certaine confiance dans les travaux qui ont été réalisés ainsi qu'une confiance dans les résultats proposés.

Références

BANQUE NATIONALE DE DONNEES MALADIES RARES. (2014). RETRIEVED FROM HTTP://WWW.BNDMR.FR/

BELLAHSENE, Z., BONIFATI, A., & RAHM, E. (2011). SCHEMA MATCHING AND MAPPING. SPRINGER BERLIN HEIDELBERG.

CHOQUET, R., MESSIAEN, C., PRIOUZEAU, A., DE CARRARA, A., & LANDAIS, P. (2012). UN JEU DE DONNÉES MINIMUM POUR FACILITER L'INTEROPÉRABILITÉ DES BASES DE DONNÉES POUR LES MALADIES RARES (PP. 1–6). PRESENTED AT THE CONFERENCE IC 2012, PARIS.

EUZENAT, J., & SHVAIKO, P. (2013A). CLASSIFICATIONS OF ONTOLOGY MATCHING TECHNIQUES. IN ONTOLOGY MATCHING (PP. 73–84). SPRINGER BERLIN HEIDELBERG.

EUZENAT, J., & SHVAIKO, P. (2013B). EVALUATION OF MATCHING SYSTEMS. IN ONTOLOGY MATCHING (PP. 285–317). SPRINGER BERLIN HEIDELBERG.

ISO, & IEC. (2013, FEBRUARY 15). ISO/IEC 11179-3.

KAZA, S., & CHEN, H. (2008). EVALUATING ONTOLOGY MAPPING TECHNIQUES: AN EXPERIMENT IN PUBLIC SAFETY INFORMATION SHARING. DECIS. SUPPORT SYST., 45(4), 714–728.

KLEIN, M. (2001). COMBINING AND RELATING ONTOLOGIES: AN ANALYSIS OF PROBLEMS AND SOLUTIONS. IN IJCAI-2001 WORKSHOP ON ONTOLOGIES AND INFORMATION SHARING (PP. 53–62).

LE DUFF, F., DUPORT, N., GONFRIER, S., LAFAY, P., TEXIER, N., SCHÜCK, S., … ROBERT, P. (2010). PLAN NATIONAL ALZHEIMER 2008-2012 - MESURE 34 MISE EN PLACE DU RECUEIL EPIDEMIOLOGIQUE NATIONAL ET PREMIERES TENDANCES. LA REVUE DE GERIATRIE, 35(8), 575–582.

MESSIAEN, C., LE MIGNOT, L., RATH, A., RICHARD, J.-B., DUFOUR, E., BEN SAID, M., … LANDAIS, P. (2008). CEMARA: A WEB DYNAMIC APPLICATION WITHIN A N-TIER ARCHITECTURE FOR RARE DISEASES. STUDIES IN HEALTH TECHNOLOGY AND INFORMATICS, 136, 51–56.

ORPHANET. (2012). RETRIEVED FROM HTTP://WWW.ORPHA.NET/

RAHM, E., & BERNSTEIN, P. A. (2001). A SURVEY OF APPROACHES TO AUTOMATIC SCHEMA MATCHING. THE VLDB JOURNAL, 10(4), 334–350.