Un modèle de connaissances pour mesurer la qualité d'une source d'information

Rémy Choquet, Samiha Qouiyd, Emilie Pasche, Christel Daniel, Omar Boussaïd, Marie-Christine Jaulent

Un modèle de connaissances pour mesurer la qualité d'une source d'information

Rémy Choquet[1], Samiha Qouiyd[1], Emilie Pasche[3], Christel Daniel[1], Omar Boussaid[2] et Marie-Christine Jaulent[1]

[1] INSERM UMRS872 EQ.20, Université Pierre et Marie Curie, 75006 Paris
{remy.choquet, christel.daniel, marie-christine.jaulent}@crc.jussieu.fr
[2] Laboratoire ERIC, Université de Lyon 2 , 69676 Bron
omar.boussaid@univ-lyon2.fr
[3] SIM, Université de Genève et Hôpitaux Universitaires de Genève, Suisse
emilie.pasche@sim.hcuge.ch

Résumé : La multiplication des bases de données pose dans de nombreux domaines la question du caractère informationnel des données et de la connaissance qu'elles peuvent générer. Les standards et autres ressources sémantiques permettent d'apporter un cadre d'interprétation aux données. L'alignement des données à ces ressources reste cependant difficile. Nous proposons un modèle de qualité de l'information permettant de mesurer et d'améliorer le contenu informationnel de données de santé. Nous expérimentons ce modèle sur des données réelles dans le cadre d'un projet européen DebugIT[1] où les données sont partagées afin de générer de nouvelles connaissances dans le domaine de l'antibiorésistance.

Mots-clés : Qualité de l'information, Entrepôts de données, Informatique Médicale. Représentation des connaissances de qualité.

1 Introduction

Les systèmes d'information (SI) stockent des volumes de données de plus en plus importants et hétérogènes. Pour aider à l'exploitation de ces données, plusieurs approches ont été proposées, comme par exemple l'intégration de données dans un même espace (entrepôt de données), et l'utilisation de ressources sémantiques (ontologies, terminologies) pour assister la navigation au sein des données ou leur interrogation. Il reste nécessaire, lors de la mise en œuvre d'entrepôts de données d'évaluer la qualité des données qui y sont stockées et de rendre explicite et non ambigu le sens de ces données. Cette dernière tâche requiert l'alignement de données d'un SI avec des ressources sémantiques de référence, ce qui est une opération difficile.

Dans le domaine de la santé aussi, les données stockées dans les systèmes d'information cliniques sont généralement utilisées dans des entrepôts de données cliniques où l'on agrègera de grands volumes de données à des fins d'analyse de

[1] Detecting and Eliminating Bacteria UsinG Information Technology, http://www.debugit.eu/

données, d'aide à la décision, d'alerte ou de fouille de données pour générer de nouvelles connaissances (Gainer *et al.*, 2007). A ce jour, la mise en œuvre de telles architectures d'analyse de données est une bonne opportunité pour les institutions de santé d'améliorer la qualité de leurs données, et de replacer leurs données dans leur contexte sémantique (Wisniewski *et al.*, 2003).

Nous positionnons ce travail dans le cadre du projet Européen DebugIT (Lovis *et al.*,2008). Le but de ce projet est de construire une plateforme complète d'intégration sémantique de données hétérogènes cliniques pour la surveillance et le contrôle des maladies infectieuses en Europe (Teodoro *et al.*,2009). L'hôpital européen Georges Pompidou est un des fournisseurs de données. Un entrepôt de données nommé TransMED a été développé par notre laboratoire pour ce projet. Cet entrepôt de données contient aujourd'hui 10 ans de données de prise en charge diagnostique et thérapeutique de patients dans le domaine des maladies infectieuses et comportant notamment les résultats d'analyses microbiologiques et les prescriptions d'antibiothérapie dispensées à l'HEGP (Hôpital Européen Georges Pompidou). Dans le cadre de ce projet, un de nos objectifs est de proposer une représentation de la qualité de l'information prenant en compte de manière exhaustive les différents aspects caractérisant l'information de santé, ceci afin d'optimiser la qualité des données intégrées et partagées entre les quatre fournisseurs de données du projet. La motivation médicale du projet est de générer de nouvelles connaissances dans le domaine de l'antibiorésistance.

Qualifier une information en fonction de sa qualité et de sa représentation sémantique partagée, constitue, pour nous, une connaissance. Pour permettre de générer cette connaissance à partir de données, nous proposons dans ce papier une méthode de représentation de la qualité de l'information basée sur notre interprétation du triangle sémiotique grâce à laquelle, il nous sera possible d'enrichir la sémantique des données et aussi, de faciliter l'extraction de connaissances que l'on fera à partir de ces données. Notre travail est présenté de la manière suivante. Tout d'abord, nous allons introduire le modèle tridimensionnel de qualité de l'information ainsi que les méthodes de mesures associées à celui-ci. Ensuite, nous présentons notre expérimentation sur des données de santé, qui est composée de 4 phases : *l'audit*, la *qualification*, *l'alignement* et la *surveillance*. Enfin, nous discutons les résultats obtenus et l'apport de ce modèle de connaissances.

2 La qualité de l'information

La notion de qualité des données a d'abord été étudiée par les statisticiens vers la fin des années 60 (Fellegi & Sunter, 1969). Au début des années 90, les sciences de l'information ont commencé à formaliser la problématique de la mesure et de l'amélioration de la qualité des données. L'ISO définit la qualité comme « la totalité des spécificités et des caractéristiques d'une entité qui permettent de satisfaire aux usages implicites et explicites définis » (ISO 8402-1986 Vocabulaire Qualité). Wang (1998) propose de définir la qualité d'une donnée en fonction de l'usage attendu que l'on en a. Bien que ce concept d'utilité attendue soit assez générique pour définir un

principe, il faudra attendre Redman (1996) pour une caractérisation du concept de qualité des données suivant 4 dimensions : *exactitude*, *perfection*, *fraîcheur* et *uniformité*. D'autres facteurs de qualité ont été proposés afin de mesurer la qualité des données en fonction de processus (Naumann & Rolker, 2000) et de leur but (Peralta, 2006). Wang propose dans (Wang, 1998) une méthodologie basée sur la roue de Deming (Deming, 1982) (*définir*, *réaliser*, *contrôler*, *agir*) nommée TDQM[2] qui définit un processus itératif d'amélioration de la qualité des données. De la même manière, la communauté des entrepôts de données a proposé des approches afin de mesurer, d'améliorer et de surveiller la qualité des données (Weikum, 1999) (Berti-Equille & Moussouni, 2005).

En marge de ces travaux, l'ingénierie des modèles apporte des méthodes afin de mesurer la qualité des modèles d'information (Krogstie *et al.*, 1995) (Moody & Shanks, 1998). Il est cependant difficile de mesurer la qualité d'expression d'un modèle d'information avec seulement des métriques (Moody, 2003). Moody définit alors des facteurs de qualité qui servent de support à leur méthode d'évaluation subjective (ou empirique) d'un modèle d'information dans laquelle un expert note de 0 à 5 chacun des 7 critères suivants : *l'exactitude*, *l'applicabilité*, *la complétude, l'intelligibilité, l'intégration, la flexibilité* et *la simplicité.*

Dans le domaine de la santé, des travaux spécifiques ont aussi été proposés du fait du besoin croissant d'utiliser les données du dossier patient à des fins d'analyses épidémiologiques ou de santé publique. Néanmoins, l'utilisation de ces données est souvent freinée, là aussi, par la mauvaise qualité des données (Goldberg *et al.*, 2008). Les causes de défaillance de qualité dans les données de santé ont été classifiées en deux types d'erreurs : systématiques et hasardeuses (Arts *et al.*, 2002) aux différents niveaux du processus de saisie. Kerr *et al* (2007) présente un cadre de mesure de la qualité des données basées sur les recommandations du CIHI[3]. Cette étude débouche sur la classification de 69 critères de qualité regroupés dans 24 caractéristiques qui se subdivisent en 6 dimensions : *précision*, *ponctualité*, *comparabilité*, *utilisabilité*, *pertinence* et *sécurité*. Nous pensons que ces méthodes se basent essentiellement sur des méthodes computationnelles qui prennent peu en compte la connaissance associée aux données.

Afin de nous aider à définir notre cadre d'évaluation, il convient de se tourner vers la communauté de l'informatique médicale qui a, depuis quelques années, mis en œuvre des bases de connaissances diverses comme les modèles d'information standardisés, des terminologies ou thesaurus et enfin, des ontologies (Brown & Sonksen, 2000). Les terminologies de référence ont été largement adoptées par les communautés de recherche du domaine (Cornet *et al.*, 2004), elles sont reconnues comme une ressource clé aidant à l'interopérabilité des systèmes de santé. Plus récemment, le web sémantique a proposé les standards nécessaires à l'adoption massive des ontologies comme un des supports de la connaissance. De même, HL7[4]

[2] Total Data Quality Management

[3] The Canadian Institute for Health Information

[4] Health Level 7 : http://www.hl7.org

et le CEN TC 251[5] sont deux organismes de standardisation en informatique de santé qui proposent des modèles d'information de référence (validés par des experts) du domaine. Ces ressources définissent un cadre sémantique pour les données de santé et nous les utiliserons dans le cadre de notre modèle qualité comme référence de mesure de distance.

3 Matériel et méthodes

3.1 Le modèle qualité

L'information stockée dans le système d'information clinique peut être définie suivant trois dimensions : 1) les données, ou les instances d'objets du monde réel, sont stockées physiquement dans les bases de données de santé, 2) les modèles d'information représentent des concepts et des relations (parmi d'autres propriétés) qui permettent d'organiser et de structurer l'information et 3) les systèmes terminologiques en santé fournissent les termes selon lesquels sont représentés ces concepts. L'ISO distingue les terminologies (listes de termes), les thésaurus (index et synonymes), les classifications (avec des relations génériques) ou les vocabulaires (avec des définitions) et les ontologies (ISO TS17117). A la différence des autres systèmes terminologiques, une ontologie peut représenter les 3 sommets du triangle (*concepts*, *termes* et *instances*). D'une manière générale, on utilise une ontologie dans le domaine de la santé pour représenter une formalisation du domaine (*concept*) et des *termes* d'un système d'information clinique. Nous proposons d'effectuer une classification des mesures de qualité de la littérature en fonction de ces trois dimensions définies de la manière suivante : *concepts, termes* et *objets*. Dans cet article, nous ne discutons pas des rapports mouvants qu'il peut y avoir entre les *objets*, les *termes* et les *concepts* (Bourigault & Slodzian, 1999). Nous sommes dans un système d'information réel pour lequel les rapports sont fixés par la pratique.

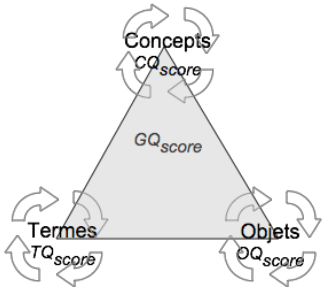

Fig. 1 –Le triangle de qualité de l'information (TQI)

Les trois dimensions sont représentées par les sommets du triangle. Un score de qualité est associé à chaque sommet, ainsi qu'une méthode de mesure inspirée de TDQM. Les scores de chacun des sommets peuvent être obtenus suivant différentes

[5] Open Electronic Healthcare Record : http://www.openehr.org/home.html

méthodes de la littérature ; nous en choisirons certaines que nous avons jugées adéquates à notre domaine d'application. La moyenne des 3 scores définira un score global de qualité (GQ) qui permettra de déterminer le niveau de qualité de la source de données. La mise en œuvre de ce modèle vise à mesurer la distance entre des données et leur domaine sémantique de référence.

3.2 Matériel

La source de données principale de notre expérimentation est le Dossier Patient Informatisé (DPI) déployé à l'HEGP depuis 2000. Il contient 10 ans de données dans divers domaines médicaux. Nous avons particulièrement travaillé avec les données de prise en charge diagnostique et thérapeutique de patients présentant des maladies infectieuses et notamment sur les résultats de microbiologie (cultures bactériennes et tests de résistance aux antibiotiques (antibiogrammes)) et les prescriptions d'antibiotiques. Le jeu de données décrit l'historique de 1 200 000 patients, au cours de 1 600 000 séjours hospitaliers, comprenant 3 200 000 antibiogrammes et 24 000 prescriptions médicamenteuses.

Afin de mesurer la qualité *conceptuelle* de notre entrepôt de données, nous avons défini un modèle d'information de référence. Cette ressource de référence a été obtenue à partir de 6 modèles d'information issus d'HL7 que nous avons intégrés en un modèle d'information unique: *Encounter universal*, *Result Event*, *Composite Order*, *Common Observation*, *Adverse Reaction*, *BillableClinicalService Encounter*. Le modèle d'information conceptuel résultant couvre le périmètre conceptuel du projet DebugIT, il comporte 61 classes et 262 propriétés.

Afin de mesurer la qualité du référentiel de termes du système DPI, nous avons utilisé deux ressources terminologiques de référence: 1) l'ATC : la classification internationale des médicaments et substances et 2) NEWT : une taxonomie des bactéries (et autres organismes).

Des routines PERL, développées aux Hôpitaux Universitaires de Genève (HUG), ont été utilisées afin d'aligner les termes stockés en texte libre du système DPI de l'HEGP avec les référentiels. A titre d'exemple, la correspondance des noms de médicament du DPI avec l'ATC est effectuée en plusieurs étapes. Une méthode de similarité par substitution de lettres dans la chaîne de caractères est appliquée. Par exemple, le terme *ac.fus* n'est similaire à aucun terme de l'ATC. Lorsque l'on cherche avec seulement *fus,* l'algorithme de similarité trouve plusieurs candidats, mais un seul est un antibiotique : *fusidic acid.* Une approche similaire est utilisée pour établir une correspondance entre les noms de bactéries du DPI et NEWT.

Talend© OpenProfiler[6] et des procédures stockées en SQL[7] ont été utilisés pour la plupart des critères *objets* à évaluer.

[6] http://www.talend.com/products-data-quality/talend-open-profiler.php

[7] Structured Query Language

3.3 Méthodologie

Nous avons mis en œuvre la méthodologie TDQM en quatre étapes pour évaluer le score qualité de chaque sommet (Wang, 1998). Ces quatre étapes peuvent être regroupées en deux processus complémentaires : l'évaluation (*audit* et *qualification)* de la source d'information en amont du processus de chargement des données, et l'amélioration ou l'alignement (*standardisation* et *surveillance)* lors du chargement des données dans l'entrepôt de données. Nous présentons d'abord, pour chaque sommet du triangle, les méthodes d'évaluation que nous avons utilisées, puis, les méthodes d'alignement utilisées.

3.3.1 Evaluation

Audit
La phase d'*audit* de chacun des sommets s'effectue grâce à diverses méthodes de mesure résumées dans la Table 1. Chaque critère de qualité est généralement mesuré grâce à des méthodes algorithmiques, sauf pour la mesure de qualité du modèle d'information qui s'appuiera sur la méthode proposée par Moody (2003). L'usage d'expressions régulières a par exemple été nécessaire afin de vérifier que le format des dates était uniforme. La distance terminologique est une distance statistique entre la terminologie locale et celle de référence.

Sommet	Critère	Méthode
Concepts	Domaine	Méthode subjective de mesure de qualité d'un modèle d'information
Objets	Complétude	Nombre d'enregistrements corrects / Nombre total d'enregistrements
Objets	Précision	Adéquation du Format et/ou du Type des données
Objets	Unicité	Un algorithme qui vérifie l'unicité des données
Objets	Cohérence	Un algorithme qui vérifie la cohérence, par exemple si la date de départ d'une prescription est inférieure à la date de fin
Termes	Cohérence	Mesure de distance aux référentiels standardisés

Table 1. Critères et méthodes d'évaluation de la qualité de l'information

Chaque critère est mesuré en fonction d'une référence généralement consensuelle. Pour la dimension *objet*, les références sont des jeux de règles validées par des experts comme par exemple la *date de décès* est plus récente que la *date de naissance*. Concernant la dimension *concept,* nous avons utilisé les modèles d'information HL7 version 3 spécialisés depuis le modèle d'information de référence qui proposent une représentation conceptuelle des données du dossier patient informatisé. Pour les *termes*, nous avons utilisé comme référence NEWT et WHO-ATC.

Qualification
Le processus de *qualification* a pour but d'établir le score de chaque dimension grâce à la phase d'*audit.* La plupart des scores issus des mesures utilisées sont numériques, cependant, nous préférons qualifier le score global de la source de données de manière plus qualitative. Nous utiliserons des degrés de qualité variant de A à D pour chaque sommet. Lorsqu'un score est un pourcentage, nous rapportons ce pourcentage à son degré correspondant (par exemple : si 73% de termes s'alignent au référentiel de termes NEWT alors la note sera B). Nous proposons l'interprétation suivante :

- A : La qualité de l'information est excellente. La source d'information est cohérente en termes de sémantique et d'organisation des données et peut être interrogée sans être adaptée.
- B : La qualité est bonne cependant il faudra améliorer la qualité d'une des dimensions du TQI.
- C : La qualité de l'information est faible. La source d'information peut être utilisée mais un effort conséquent doit être mis en œuvre pour améliorer celle-ci.
- D : La source d'information ne présente pas la qualité nécessaire pour espérer extraire de celle-ci de la connaissance et donc pour être une source de données potentielle pour un projet d'aide à la décision à partir de l'entrepôt de données.

3.3.2 Alignement et surveillance

La phase de *standardisation* a été mise en œuvre au niveau du chargement des données depuis la source vers l'entrepôt de données de santé, TransMED. La figure 2 représente la vue logique de l'architecture de mise en œuvre dans TransMED. Tout d'abord, le DPI est évalué grâce aux processus d'audit des *concepts,* des *termes* et des *objets.* Ensuite, lors du chargement des données et leur adaptation au modèle d'information cible (HL7), les termes sont exportés dans un référentiel de termes qui sera aligné avec les référentiels standards. Un expert validera les termes.

Lors du chargement des données, des routines permettent de contrôler continuellement le vocabulaire chargé dans l'entrepôt. Si un nouveau terme est introduit, il sera présenté à un expert si l'alignement n'est pas automatiquement fait. De cette manière, l'entrepôt de données ainsi créé présentera les caractéristiques nécessaires à l'extraction de connaissances depuis des données.

4 Résultats

Pour chaque étape de la méthodologie TDQM, nous présentons ici les résultats de notre expérimentation pour les 3 sommets du triangle de la qualité de l'information.

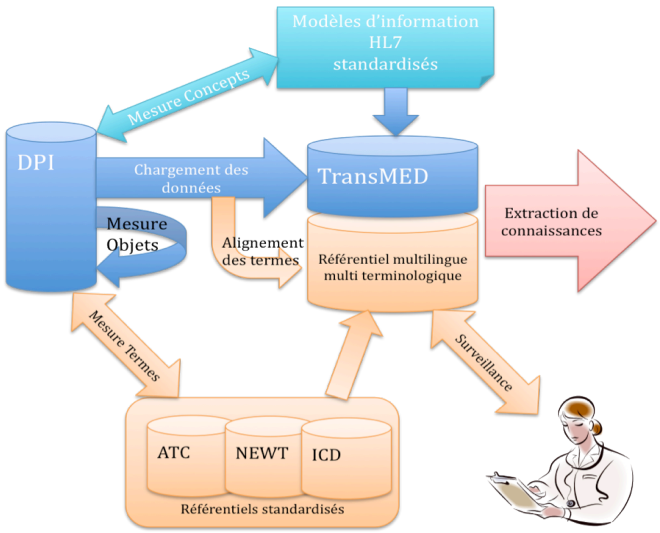

Fig. 1 –Vue logique de l'architecture de qualité de TransMED

4.1 Audit

Les *objets* sont mesurés à l'aide de procédures stockées sur la base de données source. Nous avons défini des procédures pour chaque champ que nous voulions tester. Les scores de qualité sont stockés dans une table *indicateurs*. La table 2 présente un extrait des résultats ainsi obtenus.

Objet	Critère Qualité	Score	Commentaires
Date_Fin_Sejour	Complétude	69,9%	Aide à calculer la durée du séjour
CodeUCD	Cohérence	75,6%	L'UCD est une classification française des noms de médicaments
Patient_ID	Unicité	100%	L'identifiant patient doit être unique

Table 2. Extrait des scores relatifs aux *objets*

Concernant les *termes,* la table 3 montre un exemple de mesure statistique de distance entre le référentiel des données, et le référentiel NEWT, ainsi que 2 référentiels locaux construits par un expert concernant les localisations et les types de prélèvement biologiques. La distance ainsi mesurée représente le pourcentage de termes alignés avec la ou les terminologies de référence grâce aux scripts PERL développés aux HUG.

Terminologies (Référentiels Standards)	Distance au standard
Noms des bactéries (NEWT)	85,53%
Localisation de prélèvement (local)	93,11%
Type de prélèvement (local)	36,77%

Table 3. Distance des référentiels

Enfin, concernant les *concepts* ou la modélisation conceptuelle du domaine d'intérêt de notre étude, nous avons appliqué la méthode d'évaluation subjective des modèles d'information. La figure 3 présente le résultat obtenu sur le modèle d'information du DPI de l'HEGP, avec le modèle d'information HL7 comme référence. Chaque axe est mesuré de façon empirique (avec l'aide d'un expert du domaine), et noté de 1 (mauvais) à 5 (excellent). Beaucoup d'exemples peuvent illustrer la distance entre le modèle du DPI et celui d'HL7. Par exemple, dans le domaine des résultats de laboratoire, le modèle « Result Event » d'HL7 propose une classe spécifique pour gérer les groupes de tests à faire sur un spécimen dérivé d'une culture donnée, ce qui permet une gestion plus fine des résultats microbiologiques comparé à celle en vigueur à l'HEGP à ce jour. Du point de vue de la compréhension du modèle d'information, *C_SPECIALITE*, qui en fait est une table contenant les prescriptions médicamenteuses par spécialité, n'est pas un terme parlant. D'un autre côté, une force du modèle d'information du DPI est qu'il intègre un référentiel d'éléments de données partagés, qui permet de lier le référentiel du DPI à d'autres référentiels de termes, ceci aidant à l'intégration.

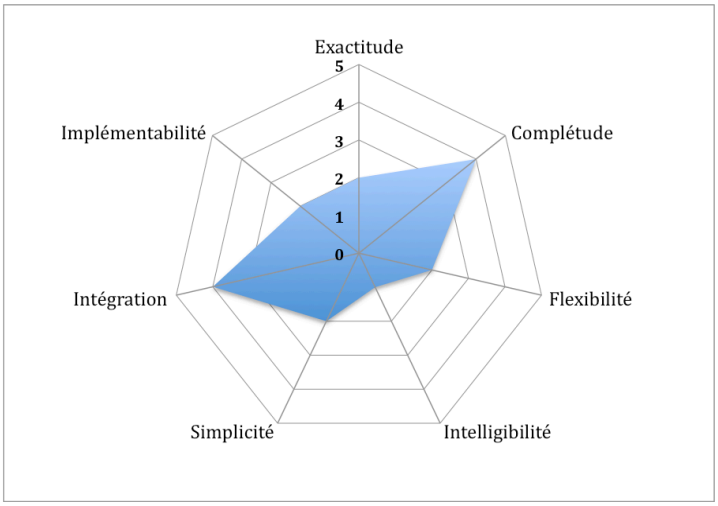

Fig. 2 –Résultat de la méthode d'évaluation subjective d'un modèle d'information

4.2 Qualification

La qualification de la source d'information pour le domaine restreint de notre étude est présentée dans la table 4.

Sommet	Score
Objet	C
Terme	B
Concept	C
Global	**C**

Table 4. Evaluation de la source de données DPI de l'étude

Cette phase de qualification peut être considérée comme une validation de notre TQI. Le résultat de l'évaluation de la dimension conceptuelle du DPI évalué était 2.42, son score qualité sera donc C. Le score global reflète la qualité du DPI étudié dans le cadre de notre domaine des maladies infectieuses. Il représente aussi la distance entre l'information source et son domaine sémantique.

4.3 Normalisation et surveillance

Le processus de normalisation a été effectué lors du chargement des données vers notre entrepôt de données de santé pour la recherche clinique TransMED. Il a été effectué à deux niveaux. Avec l'aide de Talend© OpenStudio[8], nous avons mis en œuvre des procédures de standardisation des termes grâce au référentiel multi terminologies et multilingue basé sur le modèle d'information proposé dans CTS[9]. Dans un deuxième temps, nous avons mis en œuvre 2 scripts PERL pour aligner les termes correspondant aux noms de médicaments et aux noms des bactéries. 76% des noms de médicaments ont pu être alignés avec l'ATC. Concernant les noms des bactéries, 99% des noms ont été alignés avec NEWT. Le modèle d'information de TransMED a été créé grâce à OMDF[10], un outil d'aide à la modélisation développé dans notre laboratoire. Cet outil permet la sélection des modèles HL7, leur spécialisation en modèles conceptuels, puis la génération du modèle physique pour la base de données.

Enfin, nous avons mis en œuvre les procédures nécessaires pour continuer à surveiller le vocabulaire lors de chargements futurs de données. Comme la Figure 2 l'indique, un expert est alerté si un nouveau terme (nom de médicament, nom de bactérie, type de prélèvement, diagnostique) est détecté. Ceci afin de garder un espace terminologique contrôlé.

---

[8] http://fr.talend.com/products-data-integration/talend-open-studio.php

[9] Common Terminology services : Specification HL7 pour le management des terminologies dans les systèmems d'information clinique

[10] Open Medical Development Framework : https://gforge.spim.jussieu.fr/projects/omdf-hl7/

5 Discussion et Conclusion

Un nombre important de projets de recherche actuels vise à intégrer des données et leur sémantique à travers des approches diverses visant à améliorer l'interopérabilité de l'information, à des fins d'analyses et d'extraction de nouvelles connaissances. Se pose dans chacun de ces projets la problématique de la qualité de l'information. Afin de réduire la distance entre les données issues des systèmes d'information opérationnels et la représentation sémantique partagée de ces données, nous proposons un modèle de mesure de la qualité de l'information prenant en compte l'aspect terminologique et conceptuel de l'information stockée dans une base de données. Nous pensons que le triangle de la qualité de l'information offre une grille de lecture générique pour aborder la problématique de la mesure de la qualité dans le contexte des bases de données, en amont du processus d'extraction de connaissances à partir de données. La mesure de cette qualité permet de mesurer la distance entre l'information d'un système d'information et de son domaine sémantique. Dans notre expérimentation, la dimension conceptuelle est caractérisée par un modèle d'information de référence, dans une autre expérimentation, une ontologie aurait pu être prise comme référence.

L'apport de ce travail est triple. Premièrement, ce travail a permis à l'HEGP de confronter son système d'information clinique et de mesurer la distance entre les modèles ou référentiels standardisés et son DPI après 10 ans de production. Deuxièmement, nous avons amélioré la qualité de TransMED et nous disposons aujourd'hui d'un entrepôt de données de santé où l'information est contrôlée et ou la sémantique de cette information est partageable. Enfin, cet entrepôt de données a permis à l'HEGP d'interopérer avec d'autres instituts de santé à travers, notamment, le projet européen DebugIT.

Conceptuellement, nous avons aidé au positionnent de méthodes d'évaluation de la qualité des données en fonction des 3 dimensions qui caractérisent les données stockées dans des bases de données.

Notre expérimentation a été mise en œuvre avec succès dans le cadre du projet DebugIT. Toute l'expérimentation n'est pas automatisée, l'aide d'experts du domaine est toujours requise, cependant cela n'invalide pas notre modèle. L'entrepôt de données de santé que nous avons mis en œuvre est un pas vers une meilleure interopérabilité des systèmes d'information d'une part, et vers une meilleure qualité des connaissances extraites depuis les DPI d'autre part. Il serait intéressant d'appliquer notre méthode sur d'autres hôpitaux. Enfin, nous aimerions travailler sur la qualification des données jusqu'à leur utilisation dans un système décisionnel ou d'extraction de connaissances.

Références

ARTS D., DE KEIZER N AND SCHEFFER G.J. (2002). Defining and Improving Data Quality in Medical Registries: A Literature Review, Case Study, and Generic Framework. *Journal of the American Medical Informatics Association.* vol. 9 (6) pp. 600-611.

BERTI-EQUILLE L AND MOUSSOUNI, F. (2005). Quality-aware integration and warehousing of genomic data. In *Proc. of the 10th Intl. Conference on Information Quality* (IQ'05). MIT, Cambridge, U.S.A.

BOURIGAULT D ET SLODZIAN M. (1999). Pour une terminologie textuelle. *Terminologies Nouvelles.*

BROWN PJB AND SONKSEN P. (2007). Evaluation of the quality of information retrieval of clinical findings from a computerized patient database using a semantic terminological model. *Journal of the American Medical Informatics Association*. vol. 7 (4) pp. 392-403.

CORNET ET AL. (2004). Overcoming Barriers to Evaluation of Terminological Systems. *Studies in health technology and informatics.*

DEMING, WE. (1982). Out of the Crisis, *MIT Press*, Cambridge.

FELLEGI I.P. AND SUNTER A.B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, vol. 64.

GAINER ET AL. (2007). Using the i2b2 Hive for Clinical Discovery: an Example. *AMIA Annual Symp Proceedings*

GOLDBERG SI, NIEMIERKO A AND TURCHIN A. (2008). Analysis of data errors in clinical research databases. *AMIA Annual Symp Proc*. Nov 6:242-6.

KERR K, NORRIS A AND STOCKDALE, R. (2007). Data Quality Information and Decision Making: A Healthcare Case Study. *Proc. 18th Australasian Conference on Information Systems.*

KROGSTIE J. LINDLAND O.I. AND SINDRE, G. (1995). Towards a Deeper Understanding of Quality in Requirements Engineering. *Proceedings of the 7th International Conference on Advanced Information Systems Engineering (CAISE)*, Jyvaskyla, Finland.

LOVIS C ET AL. (2008). DebugIT for patient safety - improving the treatment with antibiotics through multimedia data mining of heterogeneous clinical data. *Stud Health Technol Inform*. 136, 641-6

MOODY D.L. (2003). Measuring the quality of data models: an empirical evaluation of the use of quality metrics in practice. *Proceedings of the Eleventh European Conference on Information Systems, ECIS.*

MOODY D.L. AND SHANKS G.G. (1998). What Makes A Good Data Model? A Framework For Evaluating And Improving The Quality Of Entity Relationship Models. *Australian Computer Journal.*

NAUMANN F. AND ROLKER C. (2000). Assessment Methods for Information Quality Criteria. In *Proc. Of the MIT Conf. on Information Quality(IQ'00)*. Cambidge, USA.

PERALTA V. (2006). Data quality evaluation in data integration systems. *PhD Thesis*, Université de Versailles (France) and Universidad de la República. Uruguay.

REDMAN T.C. (1996). Data Quality for the Information Age. *Artech House*.

TEODORO D ET AL. (2009). Biomedical Data Management: A Proposal Framework. *Proceedings of the Medical Information for Europe congress*.

WANG R.Y. (1998). A product Perspective on Total Data Quality Management. *Communication of the ACM*, vol. 41, no.2.

WEIKUM G. (1999). Towards guaranteed quality and dependability of information systems. In *Proc. of the Conf. Datenbanksysteme inB*uro, Technik und Wissenschaft*, Freiburg, Germany.

WISNIEWSKI MF, KIESZKOWSKI P, ZAGORSKI BM, TRICK WE, SOMMERS AND M WEINSTEIN RA. (2003). Development of a clinical data warehouse for hospital infection control. *Journal of the American Medical Informatics Association*. vol. 10 (5) pp. 454-462.