

Évaluation de la qualité des systèmes multisources

Une approche par les patterns

J. Akoka*, L. Berti-Équille**, O. Boucelma***, M. Bouzeghoub****, I. Comyn-Wattiau*, M. Cosquer[#], V. Goasdoué[#], Z. Kedad****, S. Nugier[#], V. Peralta****, M. Quafafou***, S. Sisaïd-Cherfi*

* CNAM-CEDRIC, Paris, France, {akoka, wattiau, sisaid}@cnam.fr
<http://deptinfo.cnam.fr/quadris/>

** IRISA, Université de Rennes 1, France, berti@irisa.fr

*** LSIS, Aix-Marseille Université, France, {prénom.nom}@lsis.org

**** PRISM, Université de Versailles Saint-Quentin, France
{prénom.nom}@prism.uvsq.fr

[#]EDF-R&D, Clamart, France, {prénom.nom}@edf.fr

[#] Institut Curie, Paris, France, mireille.cosquer@curie.net

Résumé. L'article décrit la problématique et les solutions proposées par le projet QUADRI (ARA-05MMSA-0015) dont l'objectif est d'offrir un cadre d'évaluation de la qualité dans les systèmes d'information multisources (SIM). Ce cadre a permis de définir un méta-modèle pour étudier en particulier les interdépendances entre les dimensions de la qualité d'un modèle conceptuel de données et celles de la qualité des données instanciant ce modèle. Nous étudions la possibilité de définir des patterns d'évaluation de la qualité dans le but de : 1) formaliser les corrélations entre les facteurs de qualité, 2) représenter les processus, et 3) analyser la qualité des données, du système et son évolution. Le projet QUADRI s'est engagé à valider ses propositions dans les trois domaines d'application suivants : le domaine biomédical, le domaine commercial et le domaine géographique.

1 Introduction

Les problèmes de qualité des données (tels que les erreurs typographiques, les doublons, les incohérences, les valeurs manquantes, incomplètes, incertaines, obsolètes, ou peu fiables) se posent de façon récurrente dans tous les systèmes d'information, bases et entrepôts de données et pour tous les domaines d'application. Ces problèmes nuisent considérablement au résultat d'une recherche d'information (même efficace) ou d'une analyse de données préalable à toute prise de décision. En réponse à ces problèmes, l'article décrit la problématique et les solutions proposées dans le cadre du projet QUADRI (ARA-05MMSA-0015), Action de Recherche Amont financée par l'ANR (2005-2008) dont l'objectif principal est d'offrir un cadre d'évaluation de la qualité dans les systèmes d'information multisources (SIM). Ce cadre a notamment permis de définir un méta-modèle pour étudier les interdépendances entre les dimensions de la qualité d'un modèle conceptuel de données et celles de la qualité des données qui l'instancient.

Nous étudions la possibilité de définir des patterns d'évaluation de la qualité à différents niveaux (niveau modèle, niveau données et niveau processus de traitement) et cela constitue une perspective prometteuse du projet visant à formaliser les corrélations entre les facteurs de qualité pour analyser la qualité des données comme celle du SIM et son évolution. Dans un axe applicatif, le projet QUADRIS s'est engagé à valider ses propositions dans trois domaines d'application représentatifs pour leurs volumes conséquents de données, pour la complexité des modèles conceptuels de données sous-jacents et leurs problèmes de qualité nombreux et souvent spécifiques. Ces domaines sont : le domaine biomédical (dossiers médicaux informatisés par les professionnels de la santé de l'Institut Curie), le domaine commercial (avec des données d'EDF en gestion de la relation clientèle - GRC) et dans le domaine géographique (LSIS).

L'article s'organise de la façon suivante : la section 2 présente la problématique et les objectifs généraux du projet ainsi que les points saillants qui en font son originalité. La section 3 présente les récentes contributions du projet avec la proposition d'un méta-modèle et de patterns pour évaluer la qualité. La section 4 présente la mise en œuvre de ces propositions dans une démarche préparatoire à la validation des contributions dans les trois domaines cités. La section 5 resitue le projet par rapport à d'autres projets liés à la qualité des données et menés dans le domaine des systèmes d'information multisources. Enfin, la section 6 conclut l'article et présente nos perspectives de recherche.

2 Problématique et objectifs du projet QUADRIS

La qualité des systèmes d'information dans les organisations est devenue un enjeu essentiel. Elle est aussi maintenant une source de compétitivité. Le problème de la qualité se pose avec d'autant plus d'acuité que les applications tendent à se diversifier et que les volumes de données tendent à augmenter. Outre cela, les pressions réglementaires et les exigences de contrôle interne obligent les entreprises à s'intéresser de plus en plus à la qualité de leurs systèmes et de leurs données. Il ne suffit plus de constater la qualité ou la non qualité des systèmes mis en place, mais il devient urgent de mettre en place des méthodes, des outils et des standards pour anticiper les problèmes de qualité et pour les corriger le plus tôt possible. Il existe une variété de recherches sur la qualité dans les différentes phases du processus de développement (qualité des données, qualité des modèles conceptuels, qualité des processus de développement, qualité des processus de traitement de données, qualité des processus métier, etc.).

Le premier objectif du projet QUADRIS est de proposer des méthodes et des outils (métriques et prototypes) permettant la mesure de la qualité des modèles, des données et des processus liés aux SIM. Un second objectif est de valider ces contributions au moyen d'expérimentations dans les trois domaines d'application mentionnés. Notre travail de recherche est de mettre en évidence les corrélations pouvant exister entre les différentes dimensions de la qualité des modèles conceptuels de données et celle des données. L'une des faiblesses des travaux actuels est le manque de validation par des expérimentations, ce qui constitue un point fort de notre projet. Pour résumer, le projet propose :

1. l'élaboration d'un cadre théorique et pratique pour l'évaluation de la qualité des SIM,

2. l'étude des interrelations entre la qualité des modèles et celles des données au moyen de patterns,
3. la validation des patterns sur des données et modèles réels pour étudier leur capacité de généralisation.

3 Méta-modélisation et patterns pour évaluer la qualité

Une première contribution du projet est la proposition d'un méta-modèle de la qualité (Akoka et al., 2007). Ce méta-modèle est centré sur la description des différentes dimensions de qualité. Il s'inspire de l'approche *Goal-Question-Metric (GQM)* de (Basili et al., 1994) et du méta modèle élaboré dans le projet DWQ (Vassiliadis et al., 2000). Chaque dimension peut être déclinée en plusieurs facteurs. Ainsi, une dimension de qualité possible pour les données est la précision, qui peut se décliner en différents facteurs, comme par exemple la précision syntaxique, qui évalue la conformité syntaxique d'une donnée par rapport à un modèle de référence (masque particulier pour une adresse par exemple). À chaque facteur peut être associé un ensemble de métriques différentes, et à une métrique donnée peuvent correspondre différentes méthodes de mesure. Par exemple, une métrique possible pour le facteur de qualité précision syntaxique pourrait être le pourcentage de valeurs non conformes syntaxiquement au modèle de référence. Plusieurs méthodes de mesure peuvent être utilisées pour évaluer cette métrique : la mesure de la proportion de caractères identiques ou le comptage du nombre d'opérations nécessaires pour rendre les deux chaînes identiques (par distance d'édition).

Dans un premier temps, le méta-modèle est instancié avec un certain nombre de dimensions, facteurs, métriques et méthodes. Dans un deuxième temps, ce méta-modèle est utilisé pour évaluer un certain nombre de scénarii construits à partir des problématiques de qualité spécifiques aux trois domaines d'application considérés dans le projet : GRC, données provenant de dossiers médicaux et données géographiques. L'évaluation de ces scénarii est effectuée au moyen de la boîte à outils QBOX, actuellement en cours de développement. Cette boîte à outils reprend les principes du modèle *Goal-Question-Metric*. Elle permet de spécifier des buts de qualité, qui sont déclinés en un ensemble de questions. Chaque question concerne un facteur de qualité particulier et porte sur un objet particulier du système considéré (attribut, table, source, ...). La boîte à outils propose une bibliothèque de méthodes permettant l'évaluation d'un scénario.

L'objectif de la formalisation de scénarii en termes de buts et de questions liés à des facteurs et métriques déterminés est : 1) de faire émerger des situations types pour traiter un problème de qualité, et 2) d'identifier pour ces situations les facteurs et métriques pertinents. Ceci constitue un patron (*pattern*) de qualité, qui peut être à nouveau instancié dans une situation similaire. Cette notion de pattern de qualité s'apparente à celle de pattern de conception (Kuchana, 2004) ou de pattern d'architecture (Dikel et al., 2001). De façon très générale, un tel pattern peut être représenté par le modèle de la figure 1.

Ce modèle est similaire à un modèle multidimensionnel en constellation utilisé dans les entrepôts de données. Il est centré sur deux mesures principales : *i*) une mesure élémentaire (*quality value*) qui donne la valeur de qualité associée à une question (ou un facteur) et évaluée à l'aide d'une métrique particulière et d'une méthode de calcul particulière ; *ii*) une mesure agrégée (*measurement scenario*) qui correspond à une synthèse des mesures faites, par période

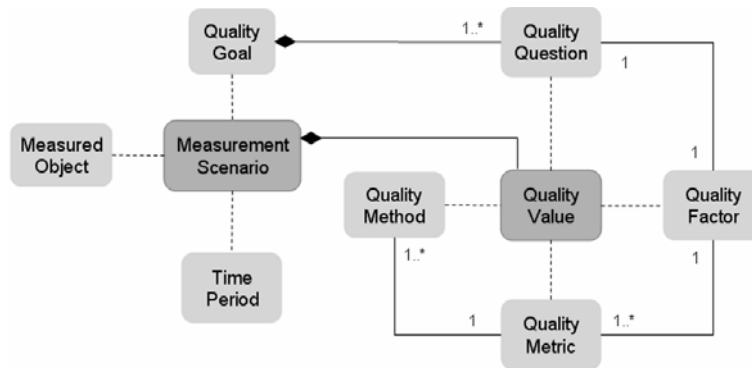


FIG. 1 – Forme générique d'un pattern de qualité

et par objet mesurable, sur l'ensemble des questions définissant le but de qualité. Les liens en pointillés dans le modèle de la figure correspondent à des dimensions de regroupement des valeurs élémentaires ou agrégées. Ce pattern pourrait être décomposé en deux patterns élémentaires correspondant aux deux structures en étoile déterminées par les mesures simples et les mesures agrégées. D'autres patterns élémentaires pourraient être définis pour représenter les corrélations entre plusieurs facteurs de qualité ou pour exprimer des comportements remarquables comme la cohésion d'un modèle conceptuel ou l'amélioration/dégradation de la qualité d'un fichier en fonction de sa fréquence de mise à jour.

4 Mise en œuvre et validation par scénario applicatif

4.1 Impact de la qualité d'un modèle sur la qualité des données dans un contexte GRC

Il existe de nombreux travaux visant à définir la qualité d'un modèle conceptuel (Batini et al., 1992), (Sisaïd-Cherfi et al., 2002), (Sisaïd-Cherfi et al., 2006). Ces travaux mettent en évidence différentes dimensions de la qualité des modèles conceptuels. Nous pouvons citer la clarté (mesurant la facilité à lire le modèle, selon une considération visuelle), la simplicité (selon la nature des concepts), l'expressivité (richesse du modèle), la justesse (correction du modèle), la complétude (niveau de couverture des besoins), la compréhension, etc. L'évaluation de la qualité d'un modèle conceptuel intervient lors de la définition d'un nouveau modèle et permet par exemple de choisir entre plusieurs modèles conceptuels concurrents ou bien lors de la mise à jour d'un modèle et permet alors de quantifier son évolution selon ses différentes dimensions de qualité. La mise à jour du schéma conceptuel peut être entraînée par un besoin d'amélioration de la qualité des données du système d'information (par exemple pour améliorer l'accessibilité des données ou leur complétude). Le choix du modèle conceptuel d'une base peut jouer un grand rôle dans la qualité des données qui seront stockées dans la base. Par exemple, intuitivement, on peut penser qu'augmenter la minimalité du modèle conceptuel peut amener à dégrader la complétude des données et que l'augmentation de l'expressivité du

modèle (en ajoutant par exemple des contraintes d'intégrité) peut faire décroître le nombre de doublons et augmenter la cohérence des données. Modéliser les relations entre qualité d'un modèle conceptuel et qualité des données est un problème complexe car selon le contexte (système d'information, données, mesures choisies pour définir la qualité, etc.), les influences observées peuvent beaucoup varier.

L'approche choisie dans le cadre du projet QUADRIIS consiste en la définition de patterns décrivant des situations types. Des scénarii applicatifs (guidés par des buts opérationnels) sur des cas "réels" de bases GRC ont été définis. Chaque scénario permet la définition de patterns de qualité (facteurs et métriques de la qualité du modèle et des données). L'expérimentation consiste à rejouer le calcul des métriques sur différentes instances opérationnelles de la base de données. Les résultats obtenus permettent d'inférer un nouveau pattern de qualité décrivant le type d'influence de certaines métriques de dimensions de la qualité des modèles sur certaines métriques de dimensions de la qualité des données. Les patterns d'influence seront ensuite soumis à des experts métier qui en jugeront la pertinence et le niveau de réutilisation.

4.2 Interdépendances entre dimensions de la qualité des données médicales et celle des processus de traitement des données

À l'Institut Curie, le dossier patient est considéré comme un élément clé dans la prise en charge d'un patient de l'hôpital, outil permettant d'assurer la continuité, la sécurité et l'efficacité des soins. Son amélioration a toujours été une thématique constante des démarches d'amélioration continue de la qualité des soins dans les établissements de santé. Aujourd'hui, dans un contexte d'informatisation, mais aussi de montée en charge de la tarification à l'activité, il s'avère important de piloter la qualité des données médicales dans un système d'information hospitalier. Notre démarche engagée sur la problématique de la qualité des données du Dossier Patient s'oriente selon les deux axes suivants : *i*) l'identito-vigilance ciblant la qualité de l'identité et donc l'unicité des dossiers administratifs (élimination des doublons). Des indicateurs de qualité ont notamment été élaborés par le Groupement pour la Modernisation du Système d'Information Hospitalier¹ : par exemple, le taux de doublons, de collisions, de modification d'identité, de fusion, et *ii*) l'infovigilance portant essentiellement sur la traçabilité, vise la complétude et la fraîcheur des informations du dossier médical et du dossier de soins. Des scores de conformité de la tenue et du contenu du dossier des patients par rapport à des critères d'évaluation réglementaires et non réglementaires ont été établis par le projet COMPAQH² et sont repris par le projet QUADRIIS afin de calculer ces indicateurs.

La boîte à outils QBox, en s'intégrant au système d'information existant, devrait permettre de suivre ces indicateurs au fil des traitements des dossiers médicaux. A partir d'un tel outil opérationnel de pilotage de la qualité des données médicales, la maîtrise des flux de données inter-applications, et donc la qualité des processus, pourrait être à terme, mise en corrélation avec celle des données. Tel est l'objectif du projet dans ce contexte applicatif.

¹Groupement pour la modernisation du système d'information hospitalier : <http://www.gmsih.fr>

²Projet COMPAQH : <http://ifr69.vjf.inserm.fr/compaqh/>

4.3 Interdépendances entre plusieurs dimensions de la qualité des données géo-spatiales

Renseigner la qualité des données géographiques a été au cours de ces dernières années une préoccupation constante de la part des utilisateurs (publics notamment). Par exemple, le FGDC³ (*Federal Geographic Data Committee*) est un organisme interministériel américain qui a développé dès 1994 le standard CSGDM (*Content Standards for Digital Geospatial Metadata*). Plus récemment, en 2003, le comité technique TC 211 du groupe ISO a proposé la norme ISO 19115 (2003) qui définit le schéma requis pour la description de l'information et des services dans le domaine géospatial. La norme contient de l'information sur l'identification, la portée, la qualité, les schémas spatial et temporel, la référence spatiale et la distribution des données géographiques. La figure 2 illustre la problématique d'intégration de deux sources de données avec prise en compte de trois dimensions de qualité que sont : *i*) la complétude : mesure principalement des taux de déficit (données manquantes par rapport au terrain nominal de référence) ou d'excès (données présentes dans le jeu de données mais manquantes ou indéterminées sur le terrain nominal), *ii*) la précision sémantique : mesure de la conformité des valeurs des éléments de la source de données par rapport à celles du terrain nominal, *iii*) la cohérence logique : mesure qui quantifie le degré de cohérence interne des données selon des règles de modélisation. L'agrégation de ces trois dimensions est non-monotone ce qui amène l'utilisateur à faire des compromis lors de la prise de décision. Dans le projet QUADRIS, sur la base du calcul de ces dimensions sur des données géospatiales du Cemagref, nous travaillons à la construction de services Web pour mesurer cette agrégation et aider à la réalisation de ces compromis entre dimensions de qualité.

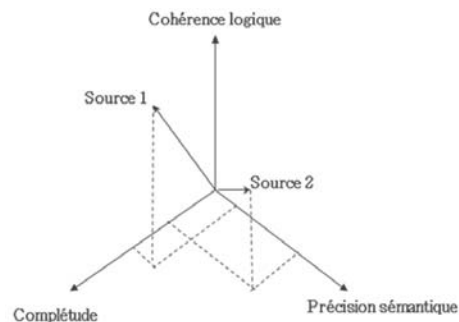


FIG. 2 – Interactions des facteurs de qualité

5 Positionnement du projet

Plusieurs projets tels que le projet TDQM (Wang et al., 1995), TQdM (English, 1999) ou les recommandations et retours d'expérience pour la gestion de projet sur la qualité des informations de (Redman, 2001) ont pu aborder tous les aspects méthodologiques liés à la

³FGDC : http://eden.ign.fr/wg/fgdc/index_html/fr/view

mise en œuvre de projets d'amélioration de la qualité ou d'assurance qualité dans les systèmes d'information d'entreprise. Sur ce dernier point, nous renvoyons le lecteur au chapitre 7 du livre de (Batini et Scannapieco, 2006) qui propose une description détaillée des stratégies adoptées et des principales méthodologies de gestion et d'évaluation de qualité des informations dans un contexte plutôt managérial. Dans le domaine des bases et entrepôts de données, des projets se sont consacrés aux techniques d'évaluation de la qualité des données : *i*) dans les entrepôts tel que le projet européen DWQ (*Data Warehouse Quality*) (Jarke et al., 1999), ou bien *ii*) dans les systèmes d'information coopératifs tel que le projet italien de DaQuinCis (Santis et al., 2003), ou encore au requêtage des données avec des contraintes notamment sur le langage ou la confiance à accorder aux données avec le projet Trio (Widom, 2005).

Dans le projet DWQ, (Jarke et al., 1999) ont proposé un méta-modèle de la qualité des données ainsi qu'une architecture d'entrepôt de données et ses composants incluant une base de méta-données sur la qualité associée à chaque méta-objet de l'entrepôt. Le projet DWQ a également fourni, en complément opérationnel, une méthodologie sur la façon d'employer des facteurs caractérisant la qualité des données et atteindre des buts de qualité du point de vue de l'utilisateur. Cette méthodologie étend l'approche GQM de (Basili et al., 1994) déjà citée, qui capture les corrélations entre différents facteurs de qualité et les organise pour atteindre des objectifs de qualité particuliers.

Le projet italien de DaQuinCIS (2001-2003) a étudié la coopération des systèmes d'information et ce en quoi celle-ci peut jouer un rôle en améliorant la qualité de données des différents systèmes d'information coopérant. (Santis et al., 2003) ont proposé une méthodologie intégrée pour concevoir une architecture distribuée avec : *i*) la définition d'un modèle de représentation de la qualité des données échangées entre les différents systèmes d'information coopératifs (CIS) et *ii*) la conception d'un logiciel dédié qui offre des services d'évaluation de la qualité des données (*quality factory* pour la génération de méta-données) pour chaque CIS et de requêtage selon la qualité au niveau du schéma global (*quality broker*).

Le projet QUADRIS rejoint ces deux projets : l'un sur le méta-modèle et la méthodologie GQM appliquée à la spécification des dimensions et des facteurs de la qualité, l'autre sur le composant logiciel qui assure la génération de méta-données d'évaluation de la qualité des données du système. Toutefois, l'originalité du projet QUADRIS réside en deux points essentiels que sont : *i*) la recherche de généricité en proposant des patterns de qualité pour la conception de systèmes d'information multisources possédant de façon native des primitives de gestion et de contrôle de la qualité de leurs données, *ii*) l'analyse fine et expérimentale des interdépendances entre les dimensions de la qualité à différents niveaux du modèle, des données, des processus et du système complet pour différents scénarii d'application, ce qui, à notre connaissance, n'a jamais été étudié auparavant dans ce domaine.

6 Conclusion

Assurer et maximiser la qualité des données dans un système d'information nécessite une compréhension fine des interdépendances entre les diverses dimensions qui caractérisent la qualité des données (QoD), la qualité du modèle conceptuel des données sous-jacent (QoM), et la qualité des processus de traitement des données (QoP). L'amélioration d'une dimension de qualité (par exemple, l'exactitude des données ou l'expressivité du modèle de données) peut

avoir des conséquences négatives sur d'autres dimensions de qualité (par exemple, l'amélioration de l'expressivité du modèle de données peut améliorer la cohérence des données jusqu'à un certain degré mais, dans le même temps, elle peut dégrader la lisibilité et la clarté du modèle). Améliorer les processus de nettoyage de données par des opérations plus sophistiquées et plus complexes peut également dégrader des facteurs liés à la fraîcheur des données en retardant la présentation des données depuis leur extraction. Dans ce contexte, le projet QUADRIS a pour but une évaluation fine des facteurs qui caractérisent la qualité des données, des modèles, des processus de traitement des données et, plus globalement, des systèmes d'information multisources ainsi que les interdépendances de ces dimensions pour être en mesure d'adopter des stratégies réfléchies d'amélioration de la qualité sur une ou plusieurs dimensions corrélées de la qualité selon les différents niveaux (données, modèle, processus, système) avec une connaissance des effets collatéraux prévisibles sur d'autres facteurs de qualité et des coûts engendrés.

Références

- Akoka, J., L. Berti-Équille, O. Boucelma, M. Bouzeghoub, I. Comyn-Wattiau, M. Cosquer, V. Goasdoué, Z. Kedad, S. Nugier, V. Peralta, et S. Sisaïd-Cherfi (2007). A Framework for Quality Evaluation in Data Integration Systems. In *Proc. of the 9th Intl. Conf. on Enterprise Information Systems (ICEIS 2007), Madeira, Portugal*.
- Basili, V., G. Caldiera, et H. Rombach (1994). *The Goal Question Metric Approach*. John Wiley & Sons, Inc.
- Batini, C., S. Ceri, et S. Navathe (1992). *Conceptual Database Design: An Entity Relationship Approach*. Benjamin Cummings.
- Batini, C. et M. Scannapieco (2006). *Data Quality: Concepts, Methodologies and Techniques*. Data-Centric Systems and Applications. Springer-Verlag.
- Dikel, D., D. Kane, et J. Wilson (2001). *Software Architecture Organizational Principles and Patterns*. Prentice Hall Computer software.
- English, L. P. (1999). *Improving Data Warehouse and Business Information Quality*. Wiley.
- Jarke, M., M. A. Jeusfeld, C. Quix, et P. Vassiliadis (1999). Architecture and Quality in Data Warehouses: An Extended Repository Approach. *Inf. Syst.* 24(3), 229–253.
- Kuchana, P. (2004). *Software Architecture Design Patterns in Java*. Computers & CRC Press, ISBN 0849321425.
- Redman, T. (2001). *Data Quality: The Field Guide*. Digital Press, Elsevier.
- Santis, L. D., M. Scannapieco, et T. Catarci (2003). Trusting Data Quality in Cooperative Information Systems. In *Proceedings of CoopIS, DOA, and ODBASE - OTM Confederated International Conferences*, Catania, Sicily, Italy, pp. 354–369.
- Sisaïd-Cherfi, S., J. Akoka, et I. Comyn-Wattiau (2002). Conceptual Modeling Quality - From EER to UML Schemas Evaluation. In *Proc. of the Intl. ER 2002 Conf.*
- Sisaïd-Cherfi, S., J. Akoka, et I. Comyn-Wattiau (2006). Use Case Modeling and Refinement: A Quality-Based Approach. In *Proc. of the Intl. ER 2006 Conf.*

- Vassiliadis, P., M. Bouzeghoub, et C. Quix (2000). Towards Quality-oriented Data Warehouse Usage and Evolution. *Inf. Syst.* 25(2), 89–115.
- Wang, R. Y., V. C. Storey, et C. P. Firth (1995). A Framework for Analysis of Data Quality Research. *IEEE Trans. Knowl. Data Eng.* 7(4), 623–640.
- Widom, J. (2005). Trio: A System for Integrated Management of Data, Accuracy, and Lineage. In *Proce. of 2nd Biennial Conference on Innovative Data Systems Research*, Asilomar, CA, USA, pp. 262–276.

Summary

The article describes the operational problems and the solutions proposed by the QUADRI project (ARA-05MMSA-0015) whose objective is to offer a framework for the evaluation of quality in multisource information systems (MISs). This framework is based on the definition of a meta-model reflecting the interdependencies between the dimensions of quality of the conceptual data model and the dimensions of quality of the data instantiating this model. The proposition of patterns for quality evaluation is a promising perspective for the project in order to formalize the correlations between the factors of quality, to represent the processes, and to analyze the quality of data, the quality of the system, and its evolution. The QUADRI project gets under way for validating its proposals in three application domains: the biomedical domain, the commercial domain, and the geographical domain.