Proposition de recherche doctorale

Gestion Contextuelle de la Qualité des Données

Mots clés :

- **Directeur de thèse** : samira SI-SAID CHERFI
- **Co-encadrant(s)** :
- **Unité de recherche** : Centre d'Étude et de Recherche en Informatique et Communications
- **Ecole doctorale** : École Doctorale Informatique, Télécommunications, Électronique de Paris
- **Domaine scientifique principal**: Divers

Résumé du projet de recherche (Langue 1)

Il est aujourd'hui largement reconnu que l'utilisation de données non appropriées, obsolètes ou incomplètes a un impact négatif sur les systèmes d'information et sur la qualité des services qu'ils délivrent. Les problèmes engendrés par une mauvaise qualité des données a un impact évident sur l'image de l'entreprise mais peut à terme avoir un impact sur sa survie. Le travail de recherche adressé dans ce projet de thèse a pour but d'étudier les problèmes de qualité de données dans leur contexte d'utilisation afin d'établir une méthode de gestion de cette qualité en tenant compte de la spécificité du contexte. Les travaux existants sur la qualité des données adressent la définition de la qualité (1, 2), sa modélisation (3, 4, 5) et son évaluation (6, 7). De nombreux travaux reconnaissent la nature multidimensionnelle de la qualité des données (8, 9, 10). La qualité d'une donnée s'évalue selon diverses dimensions telles que la complétude, la fraîcheur, l'actualité, la pertinence etc. Ces dimensions ne sont pas toujours indépendante et la qualité globale nécessite souvent un arbitrage difficile entre ces diverses dimensions. Une des problématiques de recherche qui reste ouverte est la qualification et la quantification de ces interdépendances (11). Un autre facteur qui rentre en compte est le fait que la qualité engendre un coût et que l'arbitrage est souvent décidé par le coût plus que par le besoin de qualité. Cependant, bien qu'il existe des approches adressant l'évaluation du coût de la qualité, il est souvent difficile de juger de ce coût et il serait plus judicieux de le comparer au coût de la non qualité. Enfin, la qualité n'est jamais un objectif absolu et toutes les approches qui s'appuient sur la définition de seuil d'acceptabilité pour les dimensions de la qualité sont contestables puisque la fixation de ces seuils est souvent subjective. Il est plus judicieux de considérer une vision contextuelle de la qualité où les objectifs de qualité devraient être paramétrés par les contextes d'usage. Ceci nécessite d'abord la définition du concept de contexte d'usage et de ses composantes. Il convent ensuite de définir une approche permettant d'élaborer des stratégies de la qualité en fonction du contexte. L'Anses (Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail) est une agence d'évaluation du risque et compte environ 4000 avis et rapports publiés. L'Anses représente un modèle d'approche scientifique et sanitaire intégrée qui associe des activités d'expertise et de recherche et qui prend en compte de manière globale l'ensemble des risques auxquels l'homme est soumis dans son environnement, tant pour la population générale que pour les travailleurs et incluant la santé animale et végétale. La recherche à l'Anses est réalisée dans une douzaine de laboratoires orientés sur les missions de l'Agence (santé animale, santé végétale, sécurité sanitaire des aliments, médicament vétérinaire). Le projet de thèse s'inscrit dans les activités liées à la santé animale et qui comprennent des unités de recherche en épidémiologie (4 unités) ainsi que des laboratoires nationaux ou européens de référence (40 mandats nationaux et 4 mandats européens) qui tous produisent, gèrent ou utilisent des données sanitaires. Ces équipes de recherche sont reliées à tout un ensemble d'acteurs nationaux qui produisent, collectent, gèrent ou utilisent des données sanitaires. Un certain nombre de ces acteurs se sont rassemblés en 2011 pour créer la Plateforme nationale de surveillance épidémiologique en santé animale dans le but de mutualiser, analyser et interpréter des données sanitaires pour une meilleure gestion de la surveillance épidémiologique à l'échelon national. L'un des objectifs de la Plateforme étant de mutualiser des données propriété de divers organismes ou sur lesquelles interviennent à des degrés divers ces différents organismes, les méthodes et les outils pour permettre l'assemblage puis l'utilisation de ces données sont un enjeu majeur de la Plateforme. Le travail qui sera mené dans cette thèse vise à proposer une approche de gestion contextuelle de la qualité des données. Cette approche devra être : - générique permettra de disposer d'une méthodologie pouvant être utilisée aussi bien lors de la construction d'un nouvel entrepôt de données (approche préventive) que pour gérer la qualité d'un système existant (approche corrective), - contextuelle en tenant compte du contexte d'usage des données dans l'évaluation et l'amélioration de leur qualité. Une telle approche vise à élaborer des stratégies personnalisées pour la gestion de la qualité des données. Cette personnalisation portera sur le choix des dimensions de qualité à considérer, la manière de les évaluer et le poids à leur affecter dans l'évaluation globale tout en tenant compte de leur contexte d'usage, - outillée en implantant les outils méthodologiques proposés par des outils adéquats, - validée sur des données réelles. Le projet de cette thèse s'inscrit dans le cadre d'un projet réel d'entreprise avec des données existantes et des besoins précis de gestion de la qualité. De plus, les réflexions qui conduiront à l'élaboration d'une approche adaptée alimenteront les travaux actuellement conduits par le ministère de l'agriculture pour la construction de son nouveau système d'information (RESYTAL) qui est amené à remplacer le système d'information existant (SIGAL). Références 1. Wang., Richard Y., Strong., Diane M.(Spring 1996): Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems;; 12, 4; ABI/INFORM Global 2. Won Kim, Byoung-Ju Choi, Eui-Kyeong Hong, Soo-Kyung Kim, and Doheon Lee. A Taxonomy of Dirty Data. Data Min. Knowl. Discov. 7, 1 (January 2003), 81-99. 3. Theodoratos, D.; Bouzeghoub, M.: "Data Currency Quality Factors in Data Warehouse Design". In Proc. of the Int. Workshop on Design and Management of Data Warehouses (DMDW'99), Germany, 1999. 4. Vassiliadis P., Bouzeghoub M., Quix C.:Towards Quality-Oriented Data Warehouse Usage and Evolution 5. Akoka J., Berti-Équille L., Boucelma O., Bouzeghoub M., Comyn-Wattiau I., Cosquer M., Goasdoué V., Kedad Z., Nugier S., Peralta V., Quafafou M., Sisaïd-Cher? S.:Évaluation de la qualité des systèmes multisources Une approche par les patterns 6. Kurian R.(2009):The benefits to management of using GQM, continous GQM, and V-GQM in a measurement program.A thesis submitted to kent state University in partial fulfillment of the requirement for the degree of Master's of science in computer science 7. Peralta, V.; Bouzeghoub, M.: "Evaluating Data Freshness in Data Integration Systems". Technical report, Université de Versailles, France, 2004. 8. Redman, T.C., ed. Data Quality for the Information Age. Artech House: Boston, MA., 1996. 9. Wand, Y. and Wang, R.Y. Anchoring data quality dimensions in ontological foundations. Commun. ACM 39,11 (1996), 86–95. 10. Berti-Equille L.,Comy-Wattiau I .,Cosquer M., Kedad Z., Nugier S., Paralta V.,Si-Said Cherfi S ., Thion-Goasdoué V. (2011): Assessment and analysis of information quality: a mutlidimension model and case studies. Int. J. Information Quality., Vol. 2., No 4. 11. Akoka J., Berti-Équille L., Boucelma O., Bouzeghoub M., Comyn-Wattiau I.,Cosquer M., Goasdoué-Thion V., Kedad Z., Nugier S., Peralta V., Sisaid-Cherfi S.: A framework for quality evaluation in data integration systems. ICEIS (3) 2007: 170-175. 12. Pilar Angeles M ., Mhor MacKinnon L .: Assessing Data Quality of Integrated Data by Quality Aggregation of its Ancestors. Computacion y Sistemos ., Vol. 13., No 3. 13. Pilar Angeles M .,Javier Garcia-Ugalde F. : Relevance of Quality Criteria According to the Type of Information Systems. IARIA, 2012.