Définition de l'architecture des données pour la gestion contextuelle de leur qualité dans le secteur bancaire

- **Financement :** bourse CIFRE (salaire annuel brut à partir de 23 484 €) : https://www.anrt.asso.fr/sites/default/files/cifre-plaquette-2019.pdf
- **Lieu de thèse :** Conservatoire National des arts et Métiers, CEDRIC Lab, 2, rue Conté 75003 Paris France and BNP Paribas Real Estate, 167 Quai de la bataille de Stalingrad, 92130 Issy les Moulineaux
- **Mots clés :** Données, Qualité, Dimensions de la qualité, Evaluation de la qualité, Architecture d'entreprise, Contextualisation

1 Objectifs et contexte de la thèse :

Il est aujourd'hui largement reconnu que l'utilisation de données non appropriées, obsolètes ou incomplètes à 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 ont un impact évident sur l'image de l'entreprise mais peut à terme avoir un impact sur sa survie. Dans le secteur bancaire, assurance et immobilier, la qualité des données est un sujet critique et les exigences de la qualité sont encore plus élevées vu la nature des données manipulées (données privées, financières) et les réglementations en cours.

La donnée est au cœur de nombreux processus opérationnels et de décision. Le rôle de l'évaluation de la qualité est alors d'ajouter des informations permettant d'éviter le biais ou les fausses conclusions que pourrait induire des données erronées ou imprécises. Cependant, cette évaluation est complexe et multidimensionnelle nécessitant l'intégration du contexte (de production, d'acquisition et d'usage de la donnée). Une démarche contextuelle de gestion de la qualité des données nécessite une vision globale du système information afin de mieux déterminer et comprendre, les risques liés à la non qualité, les divers acteurs impliqués dans le processus de transformation de la donnée et leur impact sur la qualité et les dimensions de qualité et leur qualification dans le contexte. Une telle vision peut s'appuyer sur l'architecture d'entreprise (AE) qui adresse les problématiques de gestion de l'entreprise et de son système d'information dans son ensemble.

L'architecture d'entreprise est une approche globale de l'entreprise qui permet, grâce à la mise en place d'une discipline d'architecture, d'aligner le système d'information aux objectifs stratégiques et aux besoins métiers. Elle fournit aux différents acteurs une description structurée de l'ensemble de ses ressources sous forme d'un cadre (cartographie, cible et référentiel). Les référentiel 'architecture contiennent, entres autres, les règles et les principes pour qu'une entreprise puisse compléter sa mission de façon pérenne (1, 13). Différentes approches adressent ces problèmes notamment les approches d'urbanisation des SI (2).

L'architecture des données fait partie de l'architecture du système d'information de toute entreprise. La démarche d'AE comprend la création d'un référentiel des données et l'établissement des principes spécifiques à cette dimension du SI (par exemple, « les données sont saisies une fois », « les données sont fournies par leur source ») afin d'assurer la cohérence, la pérennité et l'adaptabilité d'utilisation de la gestion des données.

Concernant la qualité dans le domaine d'AE, peu de travaux s'y intéressent de façon holistique. (3) définit un ensemble d'attributs de qualité adaptés au domaine de l'architecture d'entreprise en faisant une extension du standard ISO 9126 et souligne que des critères de qualité doivent être établis pour l'ensemble de principes d'architecture ainsi que pour chaque dimension de l'architecture (y compris celle des données). Appliqué à l'AE, (4) définit deux types de qualité, interne et externe, en parlant de la qualité des modèles de l'AE en général.

Le travail de recherche adressé dans ce projet de thèse a pour but de définir l'architecture de données correspondant au niveau de qualité attendu et en fonction du contexte donné. Plusieurs objectifs devront être atteints afin de résoudre le problème posé :

Les travaux existants sur les données adressent la définition de la qualité (5, 6), sa modélisation (7, 8, 9) et son évaluation (10, 11). De nombreux travaux reconnaissent la nature multidimensionnelle de la qualité des données (12). 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 indépendantes et la **qualité globale** nécessite souvent un juste équilibre entre ces diverses dimensions. Une des problématiques de recherche qui reste ouverte est la qualification et la quantification de ces interdépendances.

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.

Ensuite, **la qualité n'est jamais un objectif absolu** et toutes les approches qui s'appuient sur la définition de seuils 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 convient ensuite de définir une approche permettant d'élaborer des stratégies de la qualité en fonction du contexte.

Enfin, une architecture de données doit être définie afin d'établir les services d'architecture d'entreprise pour la gestion de cette qualité (14). Cette architecture devrait prendre en compte la nature des données, les processus impliqués et/ou influencés par ces données. Ceci passe par la formalisation du concept de qualité contextuelle et la définition d'une démarche permettant de développer une telle architecture.

2 Entreprise partenaire

BNPParibas Real Estate

3 Contributions scientifiques

Le travail qui sera mené dans cette thèse vise à proposer une approche de gestion contextuelle de la qualité des données en se basant sur les principes de l'architecture des données. Cette approche sera développée pour être utilisée dans le secteur bancaire.

Cette approche devra :

- Comporter un ensemble de solutions comprenant un cadre méthodologique et des modèles. Ces solutions devront être suffisamment génériques pour permettre leur utilisation dans divers contextes lors de la construction de nouveaux systèmes,

- Fournir une démarche permettant d'assister l'application de ces solutions pour gérer la qualité d'un système existant,

- Etre 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,

- S'inscrire dans la démarche d'architecture d'entreprise, afin d'assurer la cohérence de la gestion de la qualité des données avec les exigences d'évolution du système d'information de l'organisation,
- Etre outillée en implémentant la démarche méthodologique proposée par des outils adéquats,
- Etre validée sur des cas d'application et des données réelles.

Références :

1. TOGAF 9.1. The OpenGroup Architecture Framework. http://www.opengroup.org/togaf/, 2011.
2. Longépé C. (2004) Le projet d'urbanisation du SI, 2ème édition, Dunod, Paris.
3. Greefhorst, D., Proper, E. (2011). Architecture Principles: The Cornerstones of Enterprise Architecture, Springer, Berlin.
4. Lankhorst, M. et al. (2013). Enterprise Architecture at Work: Modelling, Communication and Analysis, Springer, Berlin.
5. Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 5-33.
6. Kim, W., Choi, B. J., Hong, E. K., Kim, S. K., & Lee, D. (2003). A taxonomy of dirty data. Data mining and knowledge discovery, 7(1), 81-99. Heinrich, B., & Klier, M. (2011).
7. Fan, W., & Geerts, F. (2012). Foundations of data quality management. Synthesis Lectures on Data Management, 4(5), 1-217.
8. Lemaitre, J., & Hainaut, J. L. (2011, January). Quality evaluation and improvement framework for database schemas-using defect taxonomies. In Advanced Information Systems Engineering (pp. 536-550). Springer Berlin Heidelberg.
9. Basili, V., Heidrich, J., Lindvall, M., Münch, J., Regardie, M., Rombach, D., ... & Trendowicz, A. (2014). GQM+ Strategies: A comprehensive methodology for aligning business strategies with software measurement. arXiv preprint arXiv:1402.0292.
10. Peralta, V. (2008). Data quality evaluation in data integration systems (Doctoral dissertation, Universidad de la República, Uruguay). Redman, T.C., ed. Data Quality for the Information Age. Artech House: Boston, MA., 1996.
11. Wand, Y. and Wang, R.Y. Anchoring data quality dimensions in ontological foundations. Commun. ACM 39,11 (1996), 86–95.
12. 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
13. Qurratuaini, H. (2018, September). Designing enterprise architecture based on TOGAF 9.1 framework. In IOP Conference Series: Materials Science and Engineering (Vol. 403, No. 1, p. 012065). IOP Publishing.
14. Van den Berg, M., Slot, R., van Steenbergen, M., Faasse, P., & van Vliet, H. (2019). How enterprise architecture improves the quality of IT investment decisions. Journal of Systems and Software, 152, 134-150.