Année : 2019 **Thèse N° : 153/ST2I**

École Nationale Supérieure d'Informatique et d'Analyse des systèmes
Centre d'Études Doctorales en Sciences des Technologies de l'Information et de l'Ingénieur

THÈSE DE DOCTORAT

ELABORATION D'UNE DEMARCHE D'EVALUATION DE PORTEFEUILLES DE PROJETS DE QUALITE DES DONNEES, BASEE ARCHITECTURE D'ENTREPRISE

Présentée par

Meryam BELHIAH

Le 26/07/2019

Formation doctorale : Informatique
Structure de recherche : Qualité des Architectures d'Entreprise, Développement et Intégration

JURY

Professeur Abdellatif MEZRIOUI PES, INPT, Rabat	**Président**
Professeur Bouchaïb BOUNABAT PES, ENSIAS, Université Mohammed V de Rabat	**Directeur de thèse**
Professeur Saïd ACHCHAB PES, ENSIAS, Université Mohammed V de Rabat	**Co-Encadrant de thèse**
Professeur Bouchra EL ASRI PES, ENSIAS, Université Mohammed V de Rabat	**Rapporteur**
Professeur Mohammed BERRADA PH, ENSA, Université Sidi Mohamed Ben Abdellah, Fès	**Rapporteur**
Professeur Ibtissam BENMILOUD PES, ENSMR, Rabat	**Examinateur**

ملخص

تمكن المبادرات المتزايدة، الرامية إلى تحسين جودة البيانات، المؤسسات من تحصيل العديد من المزايا المادية والغير مادية، من بينها الرفع من رضا العملاء وخفض تكاليف التسيير مع تحسين الإرادات.

يتميز السياق الحالي أيضا باعتماد المؤسسات على البيانات الرقمية بشكل متزايد ومحوري، مما يحتم ضرورة العمل على تحسين جودتها. على الرغم من وجود العديد من المنهجيات المرتبطة بتحسين جودة البيانات، إلا أنه لا توجد منهجية معتمدة بشكل موحد، تمكن من تقييم الأرباح التي يمكن للمؤسسات جنيها من خلال تطبيقها، إضافة إلى الكلفة والمجهود اللذان يجب بذلهما من أجل تنفيذها.

يقدم هذا البحث مقاربة جديدة تحدد بشكل كيفي وكمي، التكاليف والفوائد المباشرة وغير المباشرة، المترتبة عن مشاريع تحسين جودة البيانات، وذلك في إطار البنية المقاولاتية. حيث تمكن نتائج هذا البحث من تحديد مبادرات تحسين جودة البيانات استنادا إلى فوائدها بالنسبة للمؤسسات المعنية، مقارنة مع تكاليف تفعيلها. لتسهيل فهم هذه المقاربة وتنزيلها على أرض الواقع، تم تطوير تطبيق معلوماتي على الويب.

من جهة أخرى، تتمثل رؤية المبادرات الوطنية للبيانات المفتوحة في تمكين أكبر شريحة ممكنة من المستخدمين من الوصول إلى البيانات الحكومية من أجل تحقيق مجموعة من الأهداف. ويجمع الجزء الثاني من هذا البحث بين قياس وتقييم جودة البيانات من جهة وأنظمة التوصيات من جهة ثانية. الهدف من ذلك هو اقتراح بيانات ذات جودة عالية وقيمة مضافة، إضافة إلى توجيه اهتمام الجهات الحكومية نحو تحسين جودة البيانات ذات أهمية بالنسبة للمستخدمين.

من أجل التحقق من نتائج هذا البحث ومدى فاعليته تم إجراء دراسة تطبيقية بهدف تقييم وتحسين جودة بيانات مؤسسة حكومية.

كلمات مفتاحية: تقييم وتحسين جودة البيانات، دقة البيانات، تحليل التكلفة/ الفائدة، أداء أساليب العمل، بيانات المؤسسات، البيانات الحكومية المفتوحة

Résumé

Le nombre croissant d'initiatives en matière d'amélioration de la qualité des données offre des bénéfices monétaires et non monétaires aux organisations. Ces avantages incluent la satisfaction client, l'optimisation des coûts d'exploitation ainsi que l'accroissement des revenus.

Eu égard aux nouvelles formes des organisations orientées données, l'amélioration des niveaux de qualité des données s'impose inévitablement. Malgré un nombre croissant de méthodologies, il n'existe pas d'approche de référence, pour évaluer les projets d'assainissement des données. L'objectif d'une telle approche serait d'établir une analyse de rentabilisation, en vue d'optimiser le rapport coûts/bénéfices. Ce travail de recherche présente une démarche permettant d'identifier clairement les opportunités d'augmentation des avantages monétaires et non monétaires imputées à l'amélioration de la qualité des données, dans un contexte d'Architecture d'Entreprise. L'objectif de ce travail est d'élaborer et éprouver une démarche d'évaluation de la manière avec laquelle les processus clés permettent d'exécuter la stratégie d'une organisation et de quantifier ensuite, aussi bien les avantages que la complexité de mise en œuvre des actions d'amélioration de la qualité des données, utilisées et produites par ces processus. Ces résultats permettent de sélectionner les projets d'amélioration de la qualité, en fonction des bénéfices pour l'organisation, rapportés aux coûts de mise en œuvre. Une *Web-based* plateforme est également développée dans le but de supporter la démarche.

Le volet « Open Data » de la présente démarche combine les méthodologies de gestion de la qualité des données ainsi que les systèmes de recommandation pour fournir des suggestions de datasets pouvant potentiellement représenter un intérêt pour les usagers de l'Open Data. Cette démarche permet de tirer parti de la valeur des données ouvertes, ainsi que planifier des actions d'amélioration de la qualité des données, qui sont rentables et ayant un impact positif.

Une étude de cas a été également menée, pour illustrer cette démarche : elle porte sur l'application de la présente approche à l'assainissement de *Data Assets* gouvernementaux.

Mots-clés : Evaluation et amélioration de la qualité des données, Précision des données, Analyse coûts-avantages, Performance des processus métier, Données d'entreprise, Données gouvernementales ouvertes

Abstract

Growing data quality initiatives are increasingly offering multitudes of monetary and non-monetary benefits for organizations. These benefits include an increase of revenues, a reduction of operational costs and more satisfied customers.

In fact, the new data-oriented shape of organizations inevitably imposes the need for the improvement or their data quality. Although numerous initiatives have been made, there is still no globally accepted approach for evaluating data quality projects in order to build the optimal business cases for improvements in terms of benefits and costs. This research presents an approach to clearly identify the opportunities for increased monetary and non-monetary benefits from improved Data Quality, within an Enterprise Architecture context. The aim of this paper is to measure, in a quantitative manner, how key business processes help to execute an organization's strategy, and then to qualify the benefits as well as the complexity of improving data, that are consumed and produced by these processes. These findings will allow to select data quality improvement projects, based on the latter's benefits to the organization and their costs of implementation. To facilitate the understanding of this approach, a Java EE Web framework is developed and slightly presented here.

The portion of this research that covers Open Data combines data quality measurement and recommender systems to provide suggestions of items that may represent a potential interest for citizens for leveraging the value of open datasets, as well as planning data quality improvement actions that are cost-effective and have a highly positive impact.

For the purpose of verifying and validating the results of this research, a case study has been conducted to improve government Data Assets.

Keywords: Data Quality Assessment and Improvement, Data Accuracy, Cost/Benefit Analysis, Business Process Performance, Corporate Data, Open Government Data

Remerciements

A l'issue de ma thèse de doctorat, je tiens à remercier les personnes qui ont contribué à l'aboutissement de ce travail.

En tout premier lieu, je tiens à remercier Pr. Bouchaïb BOUNABAT, mon Directeur de thèse, pour sa disponibilité, son encadrement effectif et ses conseils judicieux, qui ont permis de mener ce travail à bon port. Je le remercie également pour son écoute et son empathie.

Pr. Saïd ACHCHAB, co-Directeur de thèse, pour son encadrement et ses conseils.

Pr. Abdellatif MEZRIOUI, pour m'avoir fait l'honneur d'accepter de présider le jury et d'évaluer mon travail de thèse.

Pr. Bouchra EL ASRI, Pr. Mohammed BERRADA et Pr. Ibtissam BENMILOUD, pour avoir accepté d'évaluer mon travail de thèse.

Mes collègues et Professeurs de l'équipe ALQUALSADI pour avoir contribué à ce travail.

Mes collègues à la Direction des Domaines de l'Etat, pour leur participation effective à ce travail.

Mes sincères remerciements et ma profonde gratitude à,

Ma mère et mon père, pour leurs prières, leur soutien et leur amour inconditionnels. Qu'ils trouvent ici le témoignage de ma profonde reconnaissance,

Mes beaux-parents, pour m'avoir accueilli au sein de leur famille les bras ouverts et pour leur soutien. Qu'ils trouvent ici l'expression de mes remerciements les plus sincères,

Moaad, mon mari, pour m'avoir épaulé durant mes années de thèse, pour son amour, ses encouragements et pour avoir toujours cru en moi,

Lina, Safa et Marwa, mes étoiles du Nord, pour m'avoir insufflé l'énergie nécessaire d'aller au bout de ce voyage et pour l'avoir ponctué de beaucoup d'amour et de tendresse,

Mon frère, mes sœurs et leurs familles respectives, pour leurs conseils et leurs efforts inestimables,

Ma famille et mes amies, pour leur présence, leurs conseils et leur soutien infaillible.

Table des matières

Liste des abréviations

Abréviation	Détail
A-B-C	Activity Based Costing
ACB	Analyse Coûts-Bénéfices
AE	Architecture d'Entreprise
AIQM	A Methodology for Information Quality Assessment
ANCFCC	Agence Nationale de la Conservation Foncière du Cadastre et de la Cartographie
ASCQ	American Society for Quality Control
ASQ	American Society for Quality
BCA	Cost Benefit Analysis
BDP	Base de Données du Patrimoine
BO	Business Object
BPML	Business Process Modeling Language
BPMN	Business Process Modeling Notation
CES	Conseil Economique et Social
CDO	Chief Data Officer
CF	Conservation Foncière
CMR	Caisse Marocaine des Retraites
CNI	Carte Nationale d'Identité
CoQ	Cost of Quality
CPM	Conditional Probability Matrix
CRM	Customer Relationship Management
DAO	Data Access Object
DaQuinCIS	Data Quality in Cooperative Information Systems
DDE	Direction des Domaines de l'Etat
DDP	Direction des Dépenses du Personnel
DPE	Domaine Privé de l'Etat
DQ	Data Quality
DQM	Data Quality Management
DSI	Direction des Systèmes d'Information
EAP	Enterprise Architecture Planning
ERP	Enterprise Resource Planning
FEAF	Federal Enterprise Architecture Framework
GBVM	Gartner Business Value Model
GPS	Global Positioning System
ID	Immeubles domaniaux

IP-MAP	Information Product Map
IP-UML	Information Product - Unified Modeling Language
IT	Information Technology
JDBC	Java Database Connectivity
JSP	Java Server Pages
KM	Knowledge Management
KPI	Key Performance Indicator
MVC	Model View Controller
OGD	Open Government Data
OMG	Object Management Group
P-A-F	Prevention - Appraisal - Failure
PIB	Produit Intérieur Brut
PortfolioDQAF	Portfolio Data Quality Assessment Framework
PSI-data	Public Sector Information data
RC	Registre de Commerce
RH	Ressources Humaines
ROI	Return On Investment
RS	Recommender Systems
SCM	Supply Chain Management
SI	Système d'Information
SIDOM	Système d'Information des DOMaines
SLA	Service Level Agreement
TDQM	Total Data Quality Management
TGR	Trésorerie Générale du Royaume
TIQM	Total Information Quality Management
TOGAF	The Open Group Architecture Framework
UML	Unified Modeling Language
WPM	Weighted Product Model

Liste des figures

Liste des tableaux

Introduction

Contexte

Le savoir-faire lié aux données constitue un facteur central et déterminant dans la réussite des organisations, en l'occurrence celles de type GAFA (Google, Apple, Facebook et Amazon). Ces géants de la technologie ont développé un environnement opérationnel qui extrait de la valeur ajoutée des données collectées, pour développer l'efficacité et l'efficience des opérations quotidiennes et des services offerts.

Inversement, les organisations qui n'ont pas exploité les données collectées dont elles disposent et qui les ont laissées à la périphérie de leurs activités quotidiennes ont vu leurs parts du marché s'effriter au profit de nouveaux arrivants dont la culture est orientée données *(data-centric)*. Ces derniers se développent à un rythme effréné et délogent des acteurs longuement établis dans un ensemble de secteurs. A titre d'exemple, Netflix[1] a précipité l'effondrement des parts du marché et la faillite de Blockbuster[2], qui avait pourtant une grande longueur d'avance et le quasi-monopole du marché du contenu vidéo aux Etats-Unis.

Le repositionnement en organisation orientée données ou du moins, la valorisation des données dont dispose l'organisation comme un réel actif (Evans & Price, 2012), impose de manière inévitable le besoin d'améliorer leur qualité. En effet, les organisations s'appuient sur leurs données pour prendre des décisions, dont peut dépendre leur pérennité. Egalement, une meilleure connaissance des clients à travers les données des applications de gestion de la relation client, leur permet de devancer leurs concurrents. Il est à citer aussi les différentes utilisations innovantes de l'information pour l'amélioration de l'efficacité opérationnelle, l'offre de meilleurs produits et services, la réduction des coûts et le contrôle de risques.

Dès lors, l'amélioration de la qualité des données (DQ[3]) est indispensable pour maximiser la valeur ajoutée tirée de la donnée, qui est dès lors considérée comme un actif à gérer, comme tout autre actif de l'organisation.

Au moment où les solutions technologiques sont de plus en plus accessibles en termes de coûts et où les processus métier sont de plus en plus automatisés et optimisés, rien ne semble plus limitant et pénalisant à la performance de ces processus que la valeur informationnelle. Les insuffisances constatées au niveau de la qualité des données impactent négativement les opérations quotidiennes, les objectifs financiers et métier, l'analyse en aval pour une prise de décision avisée et la satisfaction client (Laney, 2017), s'agit-il d'un citoyen, d'un partenaire commercial ou institutionnel ou d'une autorité de régulation.

[1] https://www.netflix.com
[2] http://www.blockbuster.com
[3] *Data Quality*

Les recherches dans le domaine de la qualité des données ont démontré que la non-qualité absorbe une marge considérable des revenus des organisations. Aux Etats-Unis et au terme de l'année fiscale 2012, La Poste *(The Postal Service)* a estimé le coût de traitement du courrier adressé et non livré à 1.5 milliards de dollars (Nixon, 2012). Un rapport publié en 2011 par Gartner, révèle qu'approximativement 40% de la valeur anticipée par les initiatives métier n'est pas atteinte à cause de la mauvaise qualité des données. En effet, cette dernière affecte les opérations quotidiennes, la productivité, la prise de décision et l'analyse en aval (Gartner, 2011).

Les organisations doivent aussi évaluer les différents scénarios liés à l'implémentation des projets de qualité des données. Le scénario optimal doit fournir la meilleure valeur financière et métier et répondre aux spécifications en termes de temps, de ressources et de coût.

Motivations et problématique

Dans une conjoncture économique complexe et dont l'issue n'est pas toujours certaine et avec une comptabilité et gestion financière transparentes, il n'est pas possible de remédier aux problèmes de qualité des données à « coup de dirhams » sans justificatifs et indicateurs mesurables au préalable.

L'objectif global n'est donc pas d'améliorer la qualité des données par n'importe quel moyen, mais de planifier les projets de qualité des données ayant un rapport coûts-bénéfices avantageux pour l'organisation. Ce savoir-faire est particulièrement pertinent pour les structures ayant pas ou peu d'expérience dans le domaine des projets d'assainissement des données.

En effet, il est important que les coûts et les bénéfices associés à la qualité des données soient explicites et surtout quantifiables aussi bien pour les managers métier, que pour les analystes SI.

Le besoin de mesurer et d'évaluer la valeur positive escomptée de l'amélioration de la qualité des données a été extensivement développé, mais d'un point de vue strictement économique (Gartner, 2011 ; Knowledge Integrity, 2011 ; Laney, 2017). Ce qui rend ces modèles difficilement applicables et adaptables à un éventail de contextes, où les objectifs de l'organisation, à travers l'amélioration de la qualité de ses données, ne sont pas d'ordre monétaire (projets en santé, éducation, Open Data, etc.).

Ce travail s'inscrit donc dans le cadre de la mesure de la valeur financière et métier des projets de qualité des données. Il peut donner ensuite lieu à une analyse plus poussée et propre au contexte de chaque organisation. Une méthode d'analyse coûts-avantages peut dès lors assister les organisations bénéficiaires à déterminer l'investissement optimal à allouer aux projets d'amélioration de la qualité des données.

En somme, établir un business case pour toute initiative d'amélioration de la qualité des données au sein d'une organisation, doit pouvoir répondre aux questions suivantes :

- Pourquoi un niveau de qualité élevé est-il important pour l'organisation ?
- Quel est l'impact des niveaux de qualité des données sur les objectifs de l'entreprise ?
- Quels sont les intrants qui affectent le coût total d'amélioration ?
- Comment mesurer de manière quantitative, la valeur positive de l'amélioration de la qualité des données et les coûts de mise en œuvre associés ?

Une attention égale doit donc être portée aux objets de type « données », ainsi qu'aux processus qui les manipulent en création, consultation et modification.

Principes et axes de recherche

Une pléthore de travaux de recherche dans les sphères académique et industriel proposent des approches pour mesurer les coûts imputés à la mauvaise qualité des données ainsi que la valeur financière des initiatives d'amélioration de cette dernière. Il manque cependant des métriques génériques et tangibles, basées sur une analyse coûts-bénéfices, pouvant être adoptées par des organisations évoluant dans des contextes diversifiés.

Partant de ce constat, ce travail de thèse s'est basé sur un framework d'Architecture d'Entreprise (AE), en particulier Archimate (Archimate, 2012) pour analyser comment la qualité des objets métier influent sur la qualité des services métier exposés et délivrés aux interlocuteurs externes et internes.

Par la suite, les bénéfices et les coûts d'implémentation des projets d'assainissement des données ont été assimilés à l'impact positif et à la complexité d'implémentation. Chaque facteur est analysé et décomposé en plusieurs critères pour en faciliter l'évaluation et la sommation par la suite, dans des indices agrégés, alimentant un modèle de coût de la qualité. Ce modèle a pour objectif de rendre aisée l'analyse en aval et l'identification des opportunités pour un bénéfice renforcé.

De manière générale, cette démarche a été développée pour les données d'entreprise, mais s'adapte au cas particulier des données ouvertes, où la valeur positive de l'amélioration de la qualité des données n'est pas d'ordre monétaire. En effet, les volets « données d'entreprise » et « données ouvertes » ont des caractéristiques différentes et les objectifs d'amélioration de la qualité ne partent pas des mêmes exigences.

Afin de rendre cette démarche effective, une application Web java EE a été développée.

Une étude de cas a été également menée, pour illustrer cette démarche : elle porte sur l'application de cette démarche à l'assainissement de *Data Assets* gouvernementaux.

Apport et contributions

Ce travail de thèse permet d'apporter plusieurs contributions aux domaines des modèles de coût de la qualité (CoQ[4]), ainsi qu'à l'évaluation de la valeur financière/métier de la qualité des données, dont :

- L'adaptation d'un méta-modèle de l'AE pour supporter l'analyse de l'attribut « qualité » ;
- La caractérisation de l'impact positif de la qualité des données sur la qualité globale des processus métier clés, et de manière transitive sur l'exécution de la stratégie de l'organisation ;
- La caractérisation de la complexité de mise en œuvre ;
- Le développement d'un framework de calcul et son implémentation à travers une application Web ;
- L'élaboration d'une étude de cas complète de la démarche pour un département gouvernemental.

Plan du travail de thèse

Ce travail de thèse est organisé somme suit :

La chapitre I a pour objectif de définir les concepts objets de ce travail, à savoir : la qualité des données, ses dimensions ainsi que ses méthodologies d'évaluation et d'amélioration. La section 2 développe les différents modèles de coût de la qualité. La section 3 dresse la liste des différentes approches d'évaluation de la valeur de la donnée. Un tableau comparatif des principaux travaux de recherche, qui mettent l'emphase sur l'évaluation de la valeur financière/métier de la qualité des données, ainsi que l'aspect « coût » de cette évaluation conclut ce chapitre.

Le chapitre II introduit l'Architecture d'Entreprise comme cadre de travail pour la démarche *PortfolioDQAF (Portfolio Data Quality Assessment Framework)*. Cette démarche est ensuite décrite en termes de : structure générale et étapes qui la composent.

Le chapitre III développe le volet « données d'entreprise[5] » de *PortfolioDQAF*. Il introduit ensuite le modèle de coût de la qualité de *PortfolioDQAF* pour la précision. La dernière section de ce chapitre présente la plateforme Web développée et qui automatise la démarche *PortfolioDQAF*.

Le chapitre IV développe le volet « Open Data » de *PortfolioDQAF*. Il commence par un benchmark des plateformes existantes les plus en vue pour l'évaluation des données gouvernementales ouvertes, ainsi que le cycle de vie des données ouvertes. La section 3 décrit les

[4] *Cost of Quality*
[5] *Corporate Data*

étapes qui composent le volet « Open Data » de la démarche *PortfolioDQAF*. Ce chapitre se termine par une synthèse des résultats obtenus.

Le chapitre 5 présente le contexte de l'étude de cas au sein de l'Administration Centrale de la Direction des Domaines de l'Etat (DDE), Rabat, Maroc. Après la description de l'étude de cas, le setup expérimental est présenté dans ses aspects : paramétrage, processus et objets métier candidats. Les résultats de l'étude de cas sont ensuite détaillés.

Les perspectives et travaux futurs concluent ce travail de thèse.

Publications

- *Conférences internationales*

Belhiah, M., & Bounabat, B. (2017, October). A User-Centered Model for Assessing and Improving Open Government Data Quality. In *MIT International Conference on Information Quality (ICIQ)*, UA Little Rock, USA

Republication à l'IQ International Journal (2018, December) par l'IAIDQ, l'*International Association for Information and Data Quality*

Belhiah, M., Benqatla, M. S., Bounabat, B., & Achchab, S. (2015). Towards a Context-aware Framework for Assessing and Optimizing Data Quality Projects. *Proceedings of 4th International Conference on Data Management Technologies and Applications*, Colmar, France. doi:10.5220/0005557001890194

Belhiah, M., Bounabat, B., & Achchab, S. (2015). The impact of data accuracy on user-perceived business services quality. *2015 10th Iberian Conference on Information Systems and Technologies (CISTI), Aveiro, Portugal.* doi:10.1109/cisti.2015.7170445

- *Chapitres dans un livre*

Belhiah, M., Benqatla, M. S., & Bounabat, B. (2016). Decision Support System for Implementing Data Quality Projects. *Communications in Computer and Information Science Data Management Technologies and Applications*, 1-16. doi:10.1007/978-3-319-30162-4_1

Benqatla, M. S., Belhiah, M., Chikhaoui, D., & Bouchaïb, B. (2018). IT Collaboration Based on Actor Network Theory: Actors Identification through Data Quality. *Innovations in Smart Cities and Applications Lecture Notes in Networks and Systems*, 95-106. doi:10.1007/978-3-319-74500-8_9

- *Articles de revues nationales/internationales*

Belhiah, M., & Bounabat, B. (n.d.). Open data au Maroc : état des lieux et perspectives à la lumière de l'adoption de la loi 31-13 et de l'adhésion du Maroc à l'OGP.

Accepté pour publication à *The Electronic Journal of Information Technology*

Belhiah, M., & Bounabat, B. (2019). Towards a Context-Dependent Approach for Evaluating Data Quality Cost. *International Journal of Advanced Computer Science and Applications.* 10. 578-584. 10.14569/IJACSA.2019.0100471.

Belhiah, M., Benqatla, M. S., & Bounabat, B. (2018). Evaluating a Collaborative Network of an Inter-organizational Data Exchange Project. *International Journal of Computer Science Issues*, 15(5), 26-35. https://doi.org/10.5281/zenodo.1467650

- *Journées doctorales*

Belhiah, M., Bounabat, B., & Achchab, S. (2014). Modèle basé Architecture d'Entreprise pour l'évaluation orientée donnée, de la qualité des services métiers. *6ème édition des Journées Doctorales en Technologie de l'Information et de la communication (JDTIC'14)*

Prix de la meilleure présentation

"You can't control what you can't measure." ~ Tom DeMarco

Chapitre I : Evaluation coûts-avantages des projets de qualité des données

I.1. Introduction

Le concept large de l'économie de la qualité *(Economics of quality)* remonte au début des années 1950. En effet, le coût de la qualité (CoQ[6]) a été introduit pour la première fois dans « Juran's Quality Control Handbook » (Juran, 1951) et « Total Quality Control » (Feigenbaum, 1956). Depuis, de nombreux experts se sont intéressés aux systèmes de coût de la qualité (Hwang, 1996 ; Kumar, 1998 ; Plunkett, 1988 ; Porter, 1992). L'importance des coûts liés à la qualité des données (DQ[7]) est de plus en plus reconnue. En effet, elle absorbe une proportion considérable des revenus des organisations. Elle empêcherait également ces dernières de tirer pleinement profit des fonds investis au niveau de leurs processus.

Les différentes fonctions de l'entreprise s'accordent donc sur le fait que la qualité des données est cruciale pour les différentes initiatives métier. Cette hypothèse est soutenue par une masse critique et un nombre crédible de sondages et d'études issues des domaines de la recherche et de l'industrie. Il existe cependant un fossé manifeste entre les gestionnaires qui constatent et subissent réellement les effets de cette mauvaise qualité et les décideurs et bailleurs de fonds qui doivent anticiper le ROI des projets d'amélioration de la qualité des données.

En effet, bien que les organisations soient conscientes de la place prépondérante d'une qualité des données élevées pour la concrétisation et la pérennisation de leurs objectifs financiers et métier (Rast, 2015), la difficulté de définir des mesures quantitatives et d'y associer les métriques adéquates, relèguent les projets d'amélioration de la qualité des données à des places reculées, au niveau des priorités des décideurs.

Afin d'améliorer la qualité des données de son Système d'Information (SI), toute organisation doit être en mesure d'évaluer et d'analyser les coûts associés à l'implémentation de niveaux de qualité souhaitables, en comparaison avec les bénéfices à escompter de cette amélioration. Ceci permet de combler l'écart entre les analystes SI et les managers métier, quand il s'agit d'articuler la valeur financière et métier des projets associés à l'amélioration de la qualité des données.

Ce chapitre est organisé comme suit : la première section s'attarde sur la définition des concepts, objets de ce travail de thèse, à savoir : la qualité des données, ses dimensions ainsi que ses méthodologies d'évaluation et d'amélioration. La deuxième section développe les différents modèles de coût de la qualité. La troisième section dresse la liste des différentes approches d'évaluation de la valeur de la donnée. Un tableau comparatif des principaux travaux de recherche, qui mettent l'emphase sur l'évaluation de la valeur financière/métier de la qualité des données, ainsi que l'aspect « coût » de cette évaluation conclut ce chapitre.

[6] *Cost of Quality*
[7] *Data Quality*

I.2. Qualité des données : définitions, évaluation et amélioration

I.2.1. Définitions, dimensions et métriques

I.2.1.1 Définitions

La qualité des données est largement conçue comme un concept multidimensionnel. Elle est communément définie comme « le degré avec lequel l'information répond aux exigences et attentes de toutes les parties prenantes, qui en ont besoin pour exécuter leur processus » (IAIDQ, 2013) ; Ce concept est repris par l'expression « *fitness for use* » (Wang & Strong, 1996). Cette définition figure également dans Strong (1997) et Batini et al. (2006).

Une attention particulière est portée au contexte dans lequel la qualité des données est considérée, puisqu'elle ne peut être évaluée et analysée indépendamment de l'environnement de l'organisation en question. L'environnement fait référence à l'environnement direct de l'organisation, à savoir : ses clients, concurrents, fournisseurs, etc. ainsi que son environnement macro : technologique, géopolitique, économique, social, légal, etc.

Les données sont créées ou collectées, stockées et manipulées par les SI, à travers les différents processus métier déployés. Eu égard à la variété des domaines d'application, l'hétérogénéité des SI et la volumétrie croissante des données disponibles (réseaux sociaux, données ouvertes, données provenant des objets connectés, etc.), diverses problématiques relatives à la qualité des données ont émergées.

I.2.1.2 Caractérisation de la recherche en qualité des données

L'un des premiers cadres de référence qui s'est intéressé à l'analyse et la classification des différents travaux de littérature sur la thématique de la qualité des données date de 1995 (Wang, 1995). Ce framework s'inspire de la norme ISO 9000 (Johnson, 2000) et établit une analogie entre les produits issus d'un processus industriel et la donnée produite dans le SI de l'organisation. Il a servi par la suite de socle à un ensemble de travaux dans ce domaine. Parmi lesquels, le framework présenté par Zhu et al. (2014), qui caractérise la recherche dans le domaine de la qualité des données. Ce framework est exhaustif et intuitif, donc facile à utiliser et à adopter. Il est également flexible, dans la mesure où il prévoit l'extension des sujets de recherche dans le domaine de la qualité des données. De ce fait, il caractérise la recherche selon deux dimensions : d'un côté les thématiques de recherche et de l'autre, les méthodes avec lesquelles il est possible d'appréhender ces thématiques.

Quatre catégories représentant les thématiques de recherche sur la qualité des données peuvent ainsi être recensées : (i) impact de la qualité des données ; (ii) solutions techniques des problèmes

de la qualité des données relatifs aux bases de données ; (iii) qualité des données dans le contexte de l'informatique et des SI ; (iv) qualité des données et assainissement.

Le tableau 1 représente le framework de Zhu et al.

Tableau 1. Framework de Zhu et al. (2014)

Thématiques	**Méthodes**
1. Impact de la qualité des données 1.1. Applications (exemple : CRM, KM, SCM et ERP) 1.2. Performance, coûts-bénéfices et opérations 1.3. Gestion des SI 1.4. Gestion de changement des organisations et processus 1.5. Stratégie et politiques 2. Solutions techniques des problèmes de la qualité des données relatifs aux bases de données 2.1. Intégration de données et entrepôts de données 2.2. Architecture d'entreprise et modèles conceptuels 2.3. Résolution d'entités et couplage d'enregistrements 2.4. Monitoring et assainissement 2.5. Lignage, provenance et marquage de la source 2.6. Incertitude 3. Qualité des données dans le contexte de l'informatique et des SI 3.1. Mesure et évaluation 3.2. Systèmes d'information 3.3. Réseaux 3.4. Vie privée 3.5. Protocole et standards 3.6. Sécurité 4. Qualité des données et assainissement	1. Recherche-action 2. Intelligence artificielle 3. Etude de cas 4. Fouille de données 5. Sciences de la conception 6. Econométrie 7. Méthodes empiriques 8. Méthodes expérimentales 9. Modélisation mathématique 10. Méthodes qualitatives 11. Méthodes quantitatives 12. Analyse statistique 13. Conception de systèmes et implémentation 14. Enquêtes 15. Théories et preuves formelles

Traduction libre
Source : Zhu et al. (2014), p. 5

Comme soulevé par le framework de classification, la qualité des données a des domaines d'application variés, tels que les applications CRM *(Customer Relationship Management)*, la gestion de la connaissance, la gestion de la chaîne logistique et les progiciels de gestion intégrée.

Il est possible de citer également, les données collectées à partir des capteurs, réseaux sociaux, données gouvernementales et données ouvertes.

Pour solutionner les problématiques de recherche émanant de ces domaines d'application, il existe donc une pléthore de méthodes de recherche, qu'il est possible d'utiliser de manière unitaire ou composée, comme le présente le framework de Zhu et al.

Pour toute activité d'évaluation, d'analyse et d'amélioration de la qualité des données, il est primordial de qualifier les niveaux de qualité par un ensemble de dimensions.

De manière générale, les dimensions de la qualité des données sont des catégories utilisées pour caractériser les données et leur aptitudes à l'emploi *« fitness for use »*. Elles permettent de définir de manière tangible les exigences en termes de qualité des données. Elles sont ensuite associées à des métriques quantifiables.

I.2.1.3 Dimensions

Les dimensions de la qualité des données permettent de décrire un aspect de la donnée, qu'il est possible de mesurer et d'évaluer par rapport à un niveau de qualité de référence, pour caractériser le niveau actuel de la qualité.

Initialement, les chercheurs ont recensé 179 attributs de la qualité des données (Wang & Strong, 1996). Comme il s'agit d'un nombre élevé et avec lequel il est difficile d'envisager de travailler, des méthodes statistiques avancées ont été appliquées pour réduire le nombre de dimensions, de manière conséquente, à 15 dimensions, ventilées sur 4 catégories (Wang, 1996).

Les 4 catégories sont les dimensions : (i) intrinsèques ; (ii) contextuelles ; (iii) de représentation ; (iv) d'accessibilité.

Les dimensions intrinsèques informent sur la manière avec laquelle les données possèdent une qualité en soi. En d'autres termes, la manière avec laquelle la donnée doit être précise, crédible, objective et dotée d'une bonne réputation. Les dimensions contextuelles soulignent l'exigence selon laquelle la qualité des données doit être prise en compte dans le contexte de la tâche à accomplir. Les dernières catégories, à savoir les dimensions de représentation et d'accessibilité, mettent l'emphase sur le rôle des systèmes et des outils permettant de faciliter les interactions entre les usagers et les données.

Aussi, les 4 catégories et les 15 dimensions forment-elles en même temps une hiérarchie et un framework pour la gestion de la qualité des données, tel que représenté par la figure 1.

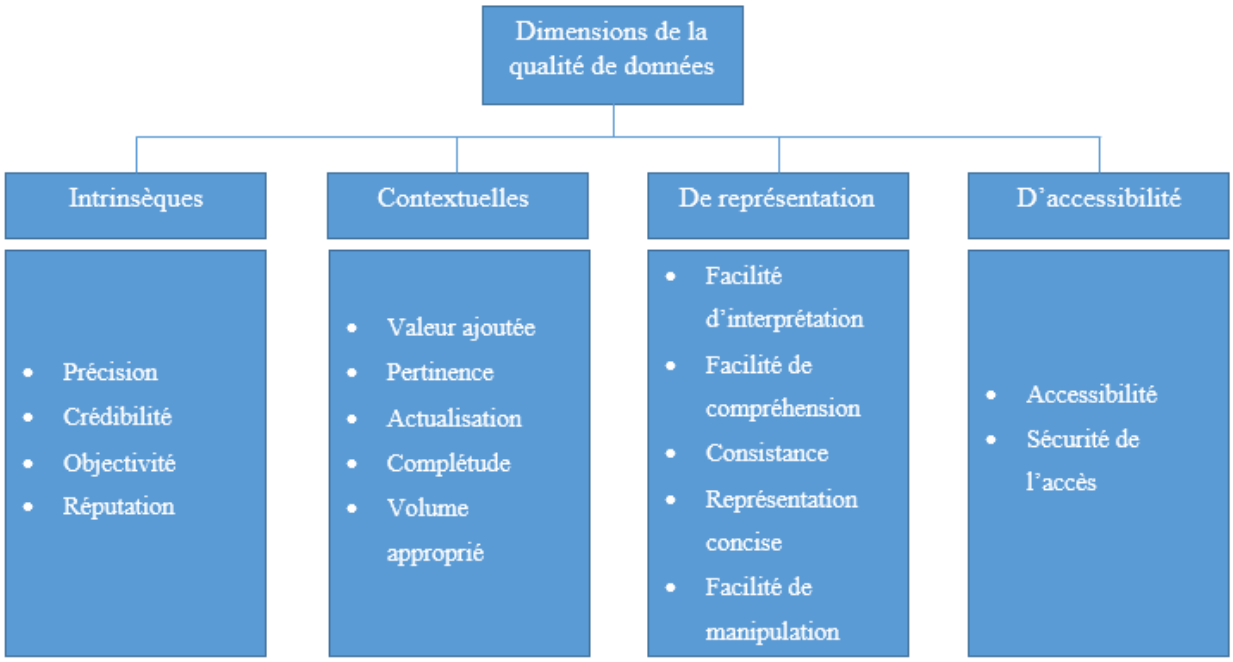


Figure 1. Hiérarchie des catégories et des dimensions de la qualité des données

Le tableau 2 ci-dessous, définit les principales dimensions de la qualité des données, ventilées par catégorie :

Tableau 2. Définitions des principales dimensions de la qualité des données

Catégorie	Dimension	Définition
Dimensions intrinsèques	Précision	La mesure dans laquelle les données sont proches des résultats observés, des valeurs réelles ou acceptées comme étant réelles : - La précision syntaxique décrit à quel point une donnée répond à une norme (format) qui permet de la valider - La précision sémantique décrit à quel point la valeur de la donnée décrit la réalité observée
	Crédibilité	La mesure dans laquelle les données sont considérées comme exactes et crédibles
	Objectivité	La mesure dans laquelle la valeur de la donnée est objective, impartiale et sans biais
	Réputation	La mesure dans laquelle la donnée est hautement estimée eu égard à sa source et son contenu
Dimensions contextuelles	Valeur ajoutée	La mesure dans laquelle la donnée est bénéfique et fournit des avantages à ses utilisateurs
	Pertinence	La mesure dans laquelle la donnée est applicable et utile dans le contexte de la tâche à exécuter

	Actualisation	La mesure dans laquelle la donnée est à jour pour l'usage qui en sera fait
	Complétude	La complétude peut se décliner en complétude pour les schémas de données, colonnes (Codd, 1972) et populations (Pipino et al., 2002) Pour la population, elle est définie par la représentativité d'un phénomène observé dans l'ensemble des données disponibles
	Volume approprié de données	La mesure dans laquelle la couverture de données est suffisante pour la tâche à accomplir
Dimensions de représentation	Facilité d'interprétation	La donnée est exprimée dans un langage approprié. Les symboles, unités et définitions sont claires
	Facilité de compréhension	La donnée est compréhensible
	Consistance	La donnée est uniformément représentée à travers les bases de données ou le SI de l'entreprise. Elle est cohérente et non-redondante
	Représentation concise	La donnée est représentée de manière compacte
Dimensions d'accessibilité	Accessibilité	La donnée est disponible, ou facilement et rapidement retrouvable
	Sécurité de l'accès	L'accès à la donnée est restreint et obéit à des règles bien définies, qui en garantissent la sécurité

Par leur nature même, les dimensions intrinsèques de la qualité des données sont indépendantes de toute représentation de cette dernière, donc indépendante du modèle de données adopté, comme il est le cas de la précision par exemple.

Ce même raisonnement s'applique aux dimensions contextuelles de la donnée, et particulièrement aux aspects temporels, à l'instar de l'actualisation. Le reliquat des dimensions à savoir : les caractéristiques relatives à la représentation et à l'accessibilité ont trait à l'évolution des SI des entreprises et au développement des technologies émergentes.

Pour parvenir à évaluer les dimensions de la qualité des données, des métriques concrètes leur sont associées. Ces métriques sont par la suite mesurées, analysées et interprétées. Elles permettent in fine, de concevoir et d'implémenter les recommandations formulées par rapport aux données. Les dimensions et les métriques qui leur sont associées sont donc au cœur de toute activité d'évaluation et d'amélioration de la qualité des données.

I.2.1.4 Evaluation des dimensions de la qualité des données

Le tableau 3 recense les travaux de recherche les plus relayés sur la thématique des dimensions de la qualité des données et les concepts qui lui sont associés.

Tableau 3. Comparaison des travaux relatifs aux dimensions de la qualité des données

Travaux de recherche	Objectif	Dimensions de la DQ	Proposition de métriques
The Six Primary Dimensions for Data Quality Assessment *(DAMA UK Working Group, 2013)*	Détaille les dimensions clés de la qualité des données, qui sont recommandées par le *DAMA UK Working Group* pour l'évaluation et l'amélioration de la qualité des données.	- Complétude - Consistance - Actualisation - Précision - Unicité - Validité	oui
Introduction to Information Quality - Chapter III *(Fisher et al., 2012)*	Aborde les thématiques fondamentales de la qualité des données dans le contexte des SI. Un chapitre est dédié aux dimensions de la qualité des données. Ce chapitre reprend les travaux de Wang.	Toutes les dimensions	oui
Methodologies for Data Quality Assessment and Improvement - Chapter IV *(Batini et al., 2009)*	Fournit une description systématique et comparative des méthodologies d'évaluation et d'amélioration de la qualité des données. Les méthodologies sont comparées selon plusieurs aspects, y compris les phases et étapes méthodologiques, les stratégies et techniques, les dimensions de la qualité des données, les types de données et finalement, les types des SI abordés par chaque méthodologie.	- Précision - Complétude - Consistance Dimensions liées à l'aspect temporel : - Actualisation - Volatilité	non
Data Quality: concepts, methodologies and techniques - Chapter II *(Batini & Scannapieca, 2006)*	Dépeint l'état de l'art des problématiques liées à la qualité des données. Ce livre présente également des pistes pour analyser et résoudre ces problèmes dans un contexte pratique. Un chapitre est dédié aux dimensions de la qualité des données.	Toutes les dimensions	oui
Beyond Accuracy: What Data Quality Means to Data Consumers *(Wang & Strong, 1996)*	Présente un framework pour conceptualiser la perception de la qualité des données par les utilisateurs.	Toutes les dimensions	non

Friedman et Laney (2012) ont également publié un référentiel de formules et d'exemples pour caractériser chacune des dimensions par des métriques quantifiables.

La section à venir s'intéresse aux cadres de référence les plus répandus pour la gestion de la qualité des données.

I.2.2. Gestion de la qualité des données

De nombreuses méthodologies de gestion de la qualité des données (DQM[8]) peuvent être recensées. Selon leur contexte d'application, elles ont pour objectif de définir, mesurer, améliorer et surveiller les niveaux de qualité des données.

Une méthodologie de la qualité des données peut être définie comme un ensemble de directives et de techniques qui, à partir d'un besoin fonctionnel en entrée, définit un processus logique, pour évaluer et améliorer la qualité des données ; et ce, en passant par plusieurs phases et points de décision. Batini et al. (2009) en proposent un état de l'art complet (13 méthodologies) et citent entre autres : TDQM (Wang, 1998), TIQM (English, 2003), AIMQ (Lee et al., 2002) et DaQuinCIS (Scannapieco et al. 2004). Les méthodologies sont analysées par rapport à plusieurs aspects, parmi lesquels : les étapes qui les composent, les dimensions qui sont couvertes, la nature des données et les types des SI adressés par chaque méthodologie. Ces méthodologies couvrent les aspects liés à l'évaluation de la qualité des données ainsi que son amélioration.

De manière générale, ces méthodologies distinguent les mesures préventives des mesures curatives. Les mesures préventives visent à éviter les défaillances au niveau des données à effets négatifs sur la qualité, alors que les mesures curatives visent à identifier et corriger les inexactitudes au niveau des données en exploitation.

L'approche curative représente ainsi plusieurs inconvénients :

- Les ressources nécessaires pour améliorer la qualité des données ne sont pas planifiées et budgétisées à l'avance, et peuvent donc ne pas être disponibles le moment venu ;
- La gestion réactive de la qualité des données est souvent associée à une absence de métriques permettant de mesurer la valeur financière/métier de la donnée. En d'autres termes, les organisations en question ne disposent pas de valeurs cibles pour la qualité des données et ne sont donc pas en mesure d'estimer si les résultats d'amélioration de la qualité des données sont suffisantes ou s'étendent bien au-delà des objectifs souhaités ;
- L'approche TDQM *(Total Data Quality Management)* a démontré que le coût total de toutes les mesures curatives dépasse le coût de la gestion préventive de la qualité (Reid & Sanders, 2005). Ceci s'applique aussi bien aux ressources matérielles qu'aux produits et ressources immatériels, comme les données.

La première section du chapitre 1 permet de présenter les principaux concepts relatifs à la thématique de la qualité des données, notamment les dimensions de la qualité, la définition des

[8] *Data Quality Management*

métriques associées et les méthodologies de gestion de la qualité des données. Le tableau 3 dresse une comparaison des travaux relatifs aux dimensions de la qualité.

De manière générale, les dimensions de la qualité des données sont des catégories utilisées pour caractériser les données et leur aptitude à l'emploi « *fitness for use* ». Celles-ci permettent de caractériser l'état actuel de la qualité et de communiquer autour de l'état souhaité.

De manière concrète, cette qualification permet de :

- Agir comme cadre de référence et guide des normes de la qualité ;
- Agir comme instrument de segmentation des efforts de l'amélioration de la DQ ;
- Faire correspondre les dimensions de la DQ avec les besoins de l'organisation ;
- Définir les priorités au niveau des scénarios d'amélioration de la DQ.

Dans un contexte économique où les ressources fiscales sont peu abondantes, l'objectif n'est pas de réaliser une qualité des données supérieure, indépendamment du coût. En plus de la caractérisation des niveaux de la qualité, il est donc important d'intégrer la dimension « coût » dans toute stratégie d'amélioration de la qualité des données. Ce qui permettra de donner une valeur à cette initiative en termes monétaires.

Dès lors, la problématique de l'amélioration de la qualité est approchée de manière aussi bien efficiente qu'efficace. C'est dans cette optique que la section suivante de ce chapitre introduit les modèles de coût de la qualité, qu'il est possible d'étendre au domaine de la qualité des données.

I.3. Modèles de coût de la qualité : définitions et applications

Un modèle de coût de la qualité (CoQ) peut être défini comme « une méthodologie qui permet à l'organisation de déterminer dans quelle mesure ses ressources sont consommées par les activités de (i) prévention de la mauvaise qualité ; (ii) évaluation de la qualité des produits et services ; (iii) correction des défaillances internes et externes » (Duffy, 2013).

Les modèles de CoQ offrent une mesure agrégée et globale de la performance de la qualité en évaluant dans un même indicateur, les coûts associés aux activités d'évaluation, de prévention et de gestion des défaillances. Les modèles de CoQ ont l'avantage de porter l'attention des décideurs et des gestionnaires sur les coûts imputés à la mauvaise qualité. Ils permettent aussi une analyse coûts-bénéfices des différentes stratégies ou programmes d'amélioration de la qualité.

Les principaux modèles de CoQ sont :

1. Le modèle P-A-F (*Prevention* - *Appraisal* - *Failure*) de Feigenbaum (Feigenbaum, 1956) ;
2. Le modèle de Juran (Juran et al., 1975) ;
3. Le modèle de Crosby (Crosby, 1980) ;

4. Le modèle coûts-bénéfices (Quah & Haldane, 2007) ;
5. Le modèle A-B-C *(Activity Based Costing Model)* de Cooper et Kaplan (1982) ;
6. Le contrôle des coûts des processus (Aoieong et al., 2002).

La suite de cette section présente une description succincte des 4 premiers modèles. Le modèle A-B-C est en dehors du périmètre de ce travail, alors que le modèle de contrôle des coûts des processus n'est pas adopté à grande échelle.

I.3.1. Le modèle P-A-F de Feigenbaum

D'après la littérature sur les modèles de CoQ, la plupart s'inspirent du modèle P-A-F ; soit la classification de Feigenbaum des coûts de la qualité en coûts de : prévention, évaluation et des défaillances (internes et externes) :

- Les coûts de prévention s'apparentent aux coûts permettant à un processus donné de fournir un produit ou un service de qualité ;
- Les coûts d'évaluation sont ceux associés à la mesure du niveau de qualité d'un processus ;
- Les coûts des défaillances sont les coûts occasionnés par la correction de la qualité des données et services avant la livraison à l'utilisateur final (il s'agit alors de coûts internes) ou après la livraison (il s'agit dans ce cas de figure de coûts externes).

Le modèle P-A-F repose sur les hypothèses suivantes :

1. Les investissements dans les coûts de prévention et d'évaluation permettraient de réduire les coûts de traitement des défaillances ;
2. Des investissements supplémentaires dans les activités de prévention réduiraient les coûts de l'évaluation.

Comme pour les autres modèles de CoQ, l'objectif est de trouver le niveau optimal de la qualité pour réduire la valeur global des coûts.

I.3.2. Le modèle de Juran

Juran s'est intéressé aux aspects économiques de la qualité ainsi qu'aux représentations graphiques des modèles de CoQ. Juran a mis également en exergue le compromis entre les coûts de prévention et d'évaluation d'un côté et les coûts des défaillances de l'autre. Ce qui conduit à un optimum économique de la qualité. Au-delà de cet optimum, les coûts d'amélioration de la qualité dépasseraient les bénéfices escomptés.

Le schéma élaboré par Feigenbaum et Juran, représenté par la figure 2, a été adopté par l'*American Society for Quality*, l'ASQ anciennement ASQC (ASQ, 2018), ainsi que par le *British Standard Institute* (BIS) (BS 6143, 1992).

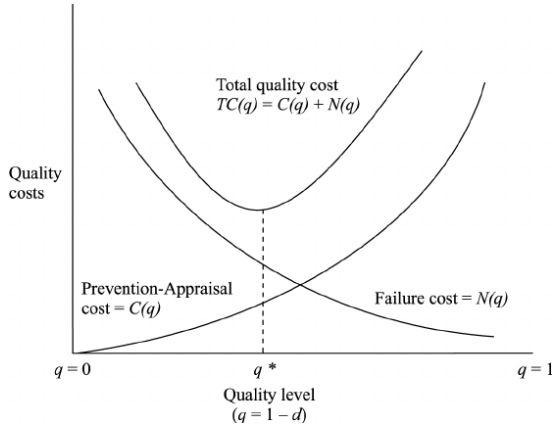


Figure 2. La représentation graphique du modèle P-A-F (Juran et al., 1975)

I.3.3. Le modèle de Crosby

Le modèle de Crosby s'inspire du modèle P-A-F. Cependant, il utilise une terminologie différente. Pour Crosby, la qualité correspond à « la conformité aux spécifications » (Crosby, 1980). De ce fait, il définit le coût de la qualité comme étant la somme des coûts de conformité et de non-conformité. Le coût de conformité est l'investissement nécessaire pour réaliser un produit ou un service qui répond aux spécifications dès la première livraison. Tandis que le coût de non-conformité correspond aux coûts de la correction, de la reprise de travail et de l'abandon le cas échéant quand le produit/service échoue à remplir le cahier des charges.

Il ressort de la comparaison des deux modèles, à savoir le modèle P-A-F et le modèle de Crosby, que les coûts de conformité correspondraient aux coûts de la prévention et de l'évaluation, tandis que les coûts de non-conformité correspondraient aux coûts des défaillances.

La sous-section suivante introduit l'évaluation des problèmes de décision par l'analyse multicritères ainsi que le modèle coûts-bénéfices.

I.3.4. L'analyse multicritères d'aide à la décision et le modèle coûts-bénéfices

I.3.4.1 Analyse multicritères d'aide à la décision

Bien que les problèmes qui impliquent une prise de décision soient intrinsèquement différents, ils partagent cependant quelques caractéristiques communes, à savoir :

- Une décision doit porter sur au moins deux alternatives pour résoudre le problème de décision ;
- Les alternatives sont évaluées selon la valeur qu'elles procurent par rapport aux critères de décision ;

- Les critères correspondent aux facteurs importants à la décision du point de vue du décideur. Ces critères sont influencés par les alternatives.

En effet, l'analyse devient multicritères lorsqu'elle fait intervenir plusieurs critères parfois mutuellement exclusifs.

La définition suivante a été retenue pour l'évaluation multicritères ; c'est « un outil d'aide à la décision qui permet de classer plusieurs alternatives en ordre de préférence sur la base de plusieurs critères dont les unités peuvent être différentes » (Zopounidis & Doumpos, 2012).

Pour évaluer deux offres d'emploi des compagnies A et B, les critères de décision seraient : le salaire de départ, les perspectives d'évolution, la proximité géographique, la taille de l'organisation, etc. La nature des critères ainsi que l'importance accordée à chacun d'eux sont différentes selon le candidat concerné par la prise de décision.

Dans la pratique, un décideur utilise plus d'un critère pour évaluer les différents scénarios plausibles pour un problème de décision. Souvent, ces critères sont mutuellement exclusifs, par exemple la rentabilité d'un investissement et son niveau de risque : un décideur souhaiterait une rentabilité élevée avec un niveau de risque réduit. Il est cependant admis qu'un retour sur investissement (ROI[9]) élevé correspondrait à un niveau de risque élevé et inversement. Le décideur doit alors trouver des compromis entre le risque et le retour sur investissement, pour identifier le (les) scénario(s) qui représentent l'équilibre le plus satisfaisant entre le rendement et le risque.

Il existe plusieurs variantes de l'analyse multicritères d'aide à la décision, parmi lesquelles, l'analyse multicritères par la méthode de multiplication de ratios (WPM[10]).

L'analyse multicritères par multiplication de ratios **-** Dans cette méthode, chaque alternative donne lieu à un score. Le score est calculé à partir des notations attribuées aux critères, et des coefficients pondérateurs qui caractérisent l'importance relative de chaque critère, aux yeux du décideur.

On dénote par :

- S_{ij}, le score attribué au critère i dans l'alternative j ;
- w_i le poids assigné au critère i. Le poids est invariable pour l'ensemble des alternatives.

Le score pondéré pour l'alternative j, sera calculé selon la formule suivante :

$$\sum_i \text{Sij} * wi \qquad (1)$$

[9] *Return On Investment*
[10] *Weighted Product Model*

Par la suite, les alternatives sont ordonnancées selon la somme pondérée résultante. Le décideur sélectionnerait les alternatives avec la somme pondérée la plus élevée.

La démarche globale pour formaliser un problème multicritères d'aide à la décision est représentée par la figure 3 :

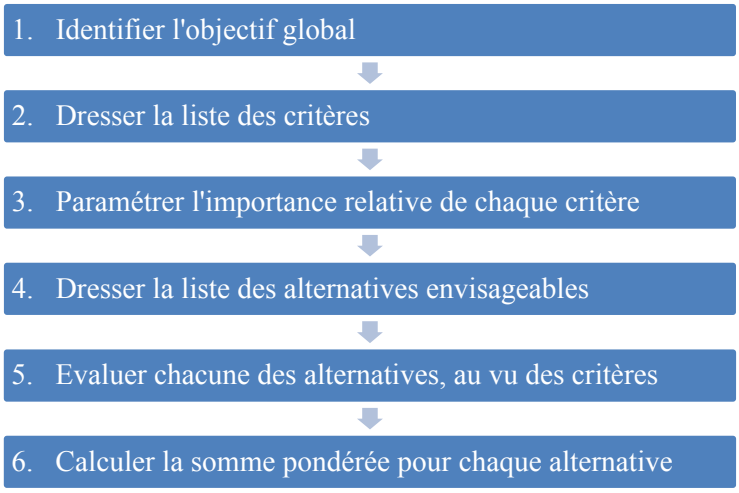


Figure 3. Démarche globale d'une analyse multicritères d'aide à la décision par la méthode WPM

Tenant compte du large spectre des problèmes de décision, il est à noter que les critères ne sont pas toujours quantifiables en termes monétaires, ce qui rend la tâche de l'évaluation plus complexe. C'est le cas par exemple dans le domaine de la santé, où le problème de décision peut s'agir du choix d'un traitement ou d'une intervention médicale et où les critères décisionnels sont : le coût du traitement et son efficacité ainsi que le nombre d'année de vie gagnée en bonne santé (Dionne, 2015). Dans ce cas, les coûts sont monétaires, mais les bénéfices correspondent à des mesures ou valorisations non-monétaires.

I.3.4.2 Modèle coûts-bénéfices

Dans le même registre, l'analyse coûts-bénéfices (ACB) est un outil d'aide à la décision qui peut faciliter la discussion entre parties prenantes. Elle fournit un cadre structuré permettant de représenter l'ensemble des éléments de la décision et discuter de leur pondération respective, favorisant ainsi la transparence du processus décisionnel (Meunier, 2009).

Dans le domaine de la gestion de programmes ou de projets, l'objectif de l'analyse coûts-bénéfices est de sélectionner un portefeuille de projets qui produit des bénéfices ou une utilité élevée (si le bénéfice est exprimés en termes non-monétaires), eu égard à des budgets restreints.

Le tableau 4 est un récapitulatif des principaux modèles de coût de la qualité développés dans cette section.

Tableau 4. Tableau récapitulatif des principaux modèles de coût de la qualité

Modèle de coûts	Classification des coûts/activités
Le modèle P-A-F Le modèle de Juran	- Coûts de prévention - Coûts d'évaluation - Coûts des défaillances (internes et externes)
Le modèle de Crosby	- Coûts de conformité - Coûts de non-conformité
Le modèle coûts-bénéfices	- Coûts directs de la qualité - Coûts indirects de la qualité - Bénéfices directs de la qualité - Bénéfices indirects de la qualité

Bien que les modèles de coût de la qualité développés plus haut ne soient pas destinés à priori à l'évaluation des coûts liés à la gestion de la qualité des données, ils sont cependant applicables à ce domaine. En effet, l'évaluation particulièrement économique, des coûts et avantages de l'amélioration de la qualité des données couvre les aspects suivants :

- Les coûts associés à la mauvaise qualité ;
- Les coûts associés à l'amélioration de la qualité des données ;
- La valeur financière/métier associée à l'amélioration de la qualité des données.

Les modèles de coût susmentionnés ainsi que la classification qu'ils proposent sont donc transposables aux aspects de l'évaluation des stratégies/programmes l'amélioration de la qualité des données.

La dernière section du chapitre I est consacrée aux travaux de recherche les plus proéminents dans le domaine de l'évaluation, principalement économiques, des coûts et avantages financiers/métier associés aux projets de qualité des données et plus particulièrement, la valeur tangible et intangible escomptée de ces initiatives.

I.4. Evaluation de la valeur financière/métier de la qualité des données

L'une des thématiques les plus importantes de la gestion de la qualité des données est comment définir et mesurer la valeur de la donnée. Cette section présente une vue d'ensemble des travaux des principaux experts aussi bien dans l'industrie que dans le domaine académique, sur la manière avec laquelle la données referme ou produit de la valeur.

I.4.1. Recherche dans le domaine de l'industrie

Dans *« Data Quality for the Information Age »*, Thomas Redman identifie plusieurs manières avec lesquelles la mauvaise qualité des données affectent les résultats financiers d'une organisation.

Parmi lesquelles : l'attrition des clients, l'induction de coûts accessoires, la diminution de la satisfaction des employés, l'impact négatif sur la réputation de l'organisation, l'impact négatif sur la prise de décision, l'induction de coûts liés à la réingénierie des processus et l'impact négatif sur la stratégie à long terme de l'organisation (Redman, 1996, p. 6-11), à ne citer que ceux-là. Il met également l'emphase sur le fait que la production et la maintenance de données de qualité peut être une source unique d'avantage compétitif (Redman, 1996, p. 12-13).

Dans « *Improving Data Warehouse and Business Information Quality : Methods for Reducing Costs and Increasing Profits* », Larry English s'attarde sur le coût élevé lié à des données de qualité médiocre. Il y cite des exemples de coûts directs et indirects induits par des données inexactes et incomplètes et par des informations fausses et trompeuses (English, 1999, pp. 7–12). Il conclue ce chapitre par : *« Le coût élevé d'une mauvaise qualité des données menace la pérennité de l'entreprise… La gestion et l'amélioration de la qualité de l'information sont une nécessité pour cette dernière »* (English, 1999, p. 13). English fournit également des recommandations sur la manière avec laquelle il est possible de calculer les coûts associés à une mauvaise qualité des données (English, 1999, p. 221–235).

Dans *« Enterprise Knowledge Management: The Data Quality Approach »*, David Loshin décrit les coûts essentiels et accessoires liés à la mauvaise qualité des données sur les niveaux opérationnel, tactique et stratégique (Loshin, 2001, p. 83-93). Les catégories relevées forment un framework qui peut être utilisé pour identifier et évaluer les coûts imputés à une mauvaise qualité et dans la même mesure, les bénéfices relatifs à un niveau de qualité élevé au sein d'une organisation. Il définit les coûts accessoires comme étant ceux qui sont clairement identifiés, mais qui demeurent difficiles à mesurer, comme la difficulté à prendre des décisions ainsi que les conflits organisationnels. Par opposition, les coûts essentiels comme l'attrition de la clientèle, les rebuts et reprise de tâches déjà effectuées ainsi que les délais opérationnels, sont des coûts qui peuvent être estimés et mesurés (Loshin, 2001, p. 84). Les impacts opérationnels relatifs à ces coûts englobent les coûts de détection et de correction des erreurs, tandis que les impacts opérationnels accessoires concernent les investissements en termes de relations publiques pour soigner l'image de l'organisation. Loshin présente également un processus pour utiliser ce framework afin de créer un tableau de bord agrégé qui *« synthétise le coût associé à une qualité des données médiocre »* (Loshin, 2001, p. 93).

Dans *« Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information (TM) »*, Dannette McGilvray présente plusieurs techniques pour mesurer l'impact des problèmes de qualité des données, dans leurs déclinaisons quantitatives et qualitatives. Pour ceci, elle présente un cadre de référence pour faciliter cette analyse (McGilvray, 2008, p. 163-198). Ces techniques incluent :

- La collecte et l'analyse des antécédents et des exemples d'impact qu'a la mauvaise qualité des données sur l'organisation ;
- La création d'un référentiel des utilisations courantes et futures des données ;

- L'analyse des problèmes liés à la qualité des données ;
- La création d'une matrice de bénéfices versus de coûts pour comprendre les effets relatifs à la mauvaise qualité des données ;
- La classification et l'ordonnancement par importance, des problèmes ainsi que des solutions plausibles.

L'objectif de cette évaluation est de construire un business case optimal et garantir l'adhésion nécessaire du management, pour mener à bien le projet d'amélioration de la qualité des données. Aussi, le processus de décision par rapport aux investissements lié à cette activité est-il amélioré.

Dans *« Information quality applied: Best practices for improving business information, processes and systems »*, Larry English fournit un récapitulatif des coûts liés à une mauvaise qualité de l'information (English, 2009). Il y présente des cas largement médiatisés :

- en 1999, la NASA a perdu la sonde spatiale *Mars Climate Orbiter* d'un coût total de 125 millions de dollars, ainsi que tout le savoir que devait recueillir cette sonde ;
- en 2000, la Cour Suprême des Etats-Unis a discrédité le vote de 4.6 millions d'électeurs à cause de problèmes liés à la qualité des données.

La liste consolidée par English des coûts de la mauvaise qualité des logiciels et des données englobe des références de plus de 120 organisations, pour un total de 1.25 billion de dollars. English conclut que, dans de nombreux secteurs d'activités, ces coûts représentent 20 à 25% des revenus d'exploitation de l'entreprise. Ces coûts sont ventilés sur les coûts de la reprise après une défaillance des processus et sur les actions correctives.

« DAMA Book of Knowledge » utilise une approche similaire pour décrire la valeur de la donnée en termes de bénéfices tirés de l'utilisation de données de qualité et des coûts associés à la détérioration des données, du fait d'utiliser des données d'une qualité médiocre. DAMA recommande d'évaluer les effets liés aux potentiels changements en termes de revenus, coûts et d'exposition aux risques divers (DAMA, 2009).

D'après les statistiques collectées par English, ainsi que les catégories et techniques présentées par Redman, Loshin et McGilvray, il est nécessaire que la première étape pour l'amélioration de la qualité des données au sein d'une organisation soit d'en comprendre la valeur. La valeur de la donnée peut être soit négative, à travers les coûts engendrés par la mauvaise qualité, ou positive à travers les bénéfices d'un niveau de qualité élevé. La qualité des données a donc un impact direct sur la valeur de la donnée.

D'autres approches dans le domaine industriel se sont également intéressées à cette problématique, comme le synthétise le tableau 5.

Tableau 5. Analyse coûts-bénéfices de l'amélioration de la qualité des données - Domaine industriel

Publication	Objectif	Modèle/Méthodologie /Approche	Métriques
Understanding the Financial Value of Data Quality Improvement *(Knowledge Integrity, 2011)*	Mapper chaque impact de la mauvaise qualité des données à un indicateur financier	Taxonomie des impacts financiers de la mauvaise qualité des données	oui
Measuring the Business Value of Data Quality *(Gartner, 2011)*	- Communiquer la valeur de la contribution du SI aux services métier - Mesurer l'impact de la qualité des données sur : les processus métier, la productivité et la prise de décision - Ordonnancer les initiatives d'amélioration de la qualité des données	*Gartner Business Value Model (GBVM)*	oui
Infonomics: How to Monetize, Manage, and Measure Information as an Asset for Competitive Advantage – Chapter 11 *(Laney, 2017)*	Mesurer la valeur d'un actif de type information, pour justifier les initiatives parmi lesquelles l'amélioration de la qualité des données	*Gartner Information Asset Valuation Models*	oui

Les chercheurs en milieux industriels ont particulièrement mis l'emphase sur l'évaluation de l'impact positif associé à l'amélioration de la qualité des données. Les différentes approches susmentionnées proposent des formules pour calculer les bénéfices financiers qui découleraient potentiellement de cette amélioration. Seulement, ces approches sont valables uniquement pour les organisations où les effets de l'amélioration de la qualité des données influencent seulement les résultats financiers.

En effet, ces modèles excluraient une frange importante d'entreprises qui ne cherchent pas à maximiser leurs gains et pour lesquelles, les bénéfices escomptées de l'amélioration de la qualité des données ne seraient pas convertibles en termes monétaires. A ne citer que :

- Les entreprises du secteur public, qui ne cherchent pas à maximiser leur gain ;
- Les projets à vocation sociale comme les projets de scolarisation, de santé publique, etc. ;
- Les projets d'ouverture des données publiques (Open Data).

I.4.2. Recherche dans le domaine académique

En plus des références citées plus haut, d'autres travaux de recherche dans le domaine académique, synthétisés dans le tableau 6, se sont intéressés à l'évaluation de la valeur de la donnée à travers

l'analyse des coûts associés à la mauvaise qualité et ceux associés à l'amélioration de la qualité des données.

Tableau 6. Analyse coûts-bénéfices de l'amélioration de la qualité des données – Domaine académique

Publication	Objectif	Modèle/Méthodologie/Approche	Métriques
The Costs of Poor Data Quality *(Haug, 2011)*	Comment identifier le niveau optimal de la qualité des données	Classification des coûts de la mauvaise qualité des données : - coûts directs - coûts cachés Définition de l'effort optimal de maintenance de la qualité des données	aucune
A Classification and Analysis of Data Quality Costs *(Eppler & Helfert, 2004)*	Identifier et catégoriser les coûts potentiels associés aux données de mauvaise qualité	Taxonomie des coûts de la qualité des données. Analyse de la progression des coûts, basée sur : - coûts de la mauvaise qualité - coûts de la détection des erreurs - coûts de la correction des erreurs - coûts de la maintenance préventive.	aucune
Towards Quantifying Data Quality Costs *(Kim & Choi, 2003)*	Quantifier les coûts de la mauvaise qualité et ceux de la réalisation d'une bonne qualité des données	Classification des coûts liés à la mauvaise qualité des données. Classification des coûts liés à l'assurance d'une bonne qualité des données.	aucune

A partir des références développées au début de cette section et dans le tableau 6, il serait juste d'affirmer que les recherches dans la sphère académique ont particulièrement porté leur attention sur la définition et l'évaluation des coûts attribuables à la mauvaise qualité des données.

Toutefois, ces approches ne proposent guère de mesures quantifiables ou de valorisations pour exprimer ces coûts en termes monétaires, permettant ainsi de qualifier l'importance et l'urgence des initiatives d'amélioration de la qualité des données.

I.4.3. Tableau comparatif

Le tableau 7 compare les aspects couverts par les travaux de recherche abordés dans ce chapitre, dans les sphères académique et industrielle, par les principaux instigateurs de la DQ.

Tableau 7. Tableau récapitulatif des travaux de recherche

Publication	Données	Processus	Aspect				Métriques
			Coûts de la mauvaise qualité	Valeur de l'amélioration de la qualité		Coûts de l'amélioration de la qualité	
				financière	non-financière		
Data Quality for the Information Age *(Redman, 1996)*	x		x				
Enterprise Knowledge Management: The Data Quality Approach *(Loshin, 2001)*	x		x				
Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information (TM) *(McGilvray, 2008)*	x		x	x			
Information Quality Applied *(English, 2009)*	x		x				
Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits *(English, 2011)*	x	x	x				
DAMA Book of Knowledge *(DAMA, 2009)*	x		x	x			
The Costs of Poor Data Quality *(Haug, 2011)*	x		x				
Towards Quantifying Data Quality Costs *(Kim & Choi, 2003)*	x		x			x	
A Classification and Analysis of Data Quality Costs *(Eppler & Helfert, 2004)*	x		x			x	
Understanding the Financial Value of Data Quality Improvement *(Knowledge Integrity, 2011)*	x	x		x			x
Measuring the Business Value of Data Quality *(Gartner, 2011)*	x			x			x
Infonomics: How to Monetize, Manage, and Measure Information as an Asset for Competitive *(Laney, 2017)*	x			x			x

Les critères de comparaison étant les aspects traités, à savoir : les coûts de la mauvaise qualité, l'impact d'une qualité des données améliorée et les coûts de l'amélioration de la qualité. Ceci et pour les données et pour les processus métier, producteurs et consommateurs de ces données.

A partir du tableau 7, il est possible de conclure que la littérature dans le domaine de l'évaluation de l'impact et l'analyse de l'efficience d'un avant-programme/projet d'amélioration de la qualité des données est ventilée sur les aspects suivants :

- La valeur négative de la donnée, à savoir les coûts associés à la mauvaise qualité ;
- La valeur financière/métier de l'amélioration de la qualité des données ;
- Les coûts associés à l'amélioration de la qualité des données.

Comme le montre la section 4 de ce chapitre, les coûts associés à la mauvaise qualité des données ont été développés de manière extensive, à travers l'établissement de classifications et de taxonomies pour catégoriser ces coûts (Eppler & Helfert, 2004 ; Haug et al., 2011). Un sous-ensemble de ces travaux a également étudié les coûts associés à l'amélioration de la qualité des données. Ces travaux ne proposent cependant pas de métriques ni d'indicateurs agrégés qui permettent d'analyser ces aspects, pour déboucher sur des choix d'implémentation d'initiatives d'amélioration de la qualité des données.

D'autre part, le besoin de mesurer et d'évaluer la valeur positive escomptée de l'amélioration de la qualité a été extensivement développé, mais d'un point de vue strictement économique. Ce qui rend ces modèles difficilement applicables et adaptables à un éventail de contextes, où les objectifs de l'organisation, à travers l'amélioration de la qualité de ses données, ne sont pas d'ordre pécuniaire. Ces travaux sont d'ailleurs les seuls à proposer des métriques quantifiables, qui sont directement liées aux objectifs financiers de l'organisation.

En plus de la représentation en interne de la donnée, au niveau du SI de l'organisation par exemple, la qualité des données concerne en grande proportion les processus et les services métier (English, 2008) qui manipulent ces données, en entrée et en sortie. D'ailleurs, l'évaluation de la valeur financière/métier de l'amélioration de la qualité des données est une responsabilité partagée entre la DSI (Division des Systèmes d'Information) et les services métier.

Le composant « processus », manipulant ces données en entrée et en sortie, est sous-évalué. Les processus revêtent pourtant une grande importance du fait que pour s'assurer d'un alignement business-IT, il est primordial de :

- Identifier les processus métier clés qui dépendent de données de très bonne qualité ;
- Comprendre comment les dits processus utilisent ces données ;
- Traduire les attentes en termes de qualité des données, en des règles de validité implémentées par ces processus.

Comme les services métier sont au cœur de l'Architecture d'Entreprise et que cette dernière permet l'alignement business-IT, l'analyse de l'AE va être utilisée comme cadre de référence pour la démarche proposée dans le chapitre II.

I.5. Synthèse

Même si les travaux de recherche susmentionnés établissent la méthodologie globale pour mesurer les coûts de la mauvaise qualité des données ainsi que la valeur financière des initiatives de l'amélioration de cette dernière, il manque des métriques génériques et concrètes, basées sur une analyse coûts-bénéfices, pouvant être utilisées par différentes organisations évoluant dans des contextes diversifiés. Ces métriques concerneraient aussi bien les processus que les données et faciliteraient l'identification des opportunités pour un bénéfice renforcé, avant d'entamer des analyses plus fines, utilisant différents indicateurs de performance (KPI[11]) spécifiques à chaque organisation.

En somme, établir un business case pour toute initiative d'amélioration de la qualité des données au sein d'une organisation, doit pouvoir répondre aux questions suivantes :

- Pourquoi un niveau de qualité élevé est-il important pour l'organisation ?
- Quel est l'impact des niveaux de qualité des données sur les objectifs de l'entreprise ?
- Quels sont les intrants qui affectent le coût total d'amélioration ?
- Comment mesurer de manière quantitative, la valeur positive de l'amélioration de la qualité des données et les coûts de mise en œuvre associés ?

L'agrégation des différents facteurs d'impact et de coût est requise dans l'objectif d'avoir une vue synthétisée et de simplifier ainsi le processus de prise de décision par rapport aux initiatives d'amélioration de la qualité à implémenter.

Une attention égale doit être portée aux objets de type « données », ainsi qu'aux processus qui les manipulent en création, consultation et modification.

La définition des facteurs d'impact positif et de coût qui construisent le business case relatif à l'amélioration de la qualité des données, le paramétrage de l'importance relative de chaque facteur et la définition d'indicateurs agrégés d'impact et de coût constituent les principaux objectifs de la démarche *PortfolioDQAF (Portfolio Data Quality Assessment Framework)*, présentée dans le chapitre II.

[11] *Key Performance Indicator*

I.7. Conclusion

Ce chapitre introduit une revue de littérature de la qualité des données, en s'intéressant aux composants les plus importants de cette thématique. Le focus est porté principalement sur les différentes approches de l'évaluation des valeurs positive et négative de la donnée, en l'occurrence : les coûts imputés à la mauvaise qualité ainsi que la valeur monétaire de l'amélioration de cette dernière.

En effet, l'objectif global n'est pas d'améliorer la qualité des données par n'importe quel moyen, mais de planifier les projets de qualité des données ayant un rapport coûts-bénéfices avantageux pour l'organisation. Ce savoir est particulièrement pertinent pour les structures ayant pas ou peu d'expérience dans le domaine des projets d'assainissement des données.

Le chapitre II décrit la démarche *PortfolioDQAF*, objet de ce travail de thèse. Cette démarche permet l'alignement business-IT, par l'évaluation de projets d'amélioration de la DQ, dans l'objectif de prioriser les initiatives de DQ et mesurer la contribution de cette amélioration à la stratégie de l'organisation.

"Essentially, all models are wrong, but some are useful." ~

Dr. George Edward Pelham Box (1919-2013)

Chapitre II : Proposition d'une démarche d'évaluation de portefeuilles de projets de qualité des données, basée Architecture d'Entreprise

II.1. Introduction

Comme présenté au niveau du premier chapitre, différentes approches existent pour évaluer les projets de qualité des données ; cependant aucune démarche n'est communément admise pour construire un business case optimal, en termes de coût et de ROI, pour améliorer la qualité des données.

Ce chapitre présente la démarche *PortfolioDQAF*. Cette démarche est basée Architecture d'Entreprise et permet d'identifier les projets d'amélioration de la qualité des données les plus efficients, à travers l'établissement de deux indicateurs agrégés d'impact positif et de complexité d'implémentation.

L'objectif de la démarche *PortfolioDQAF* est de :

- Evaluer l'impact positif de la qualité des données sur la qualité globale des processus métier clés, et de manière transitive sur l'exécution de la stratégie de l'organisation ;
- Evaluer la complexité d'opérationnalisation des actions d'amélioration de la qualité des données ;
- Recommander à travers l'analyse d'un modèle de coût de la qualité proposé, le business case optimal pour l'amélioration.

Ces résultats vont permettre de sélectionner les projets de qualité des données, sur la base des bénéfices apportés à l'organisation, ramenées à la complexité d'implémentation.

Comme les bénéfices de la qualité ne sont pas toujours mesurables en termes monétaires (projets en santé, éducation, Open Data, etc.), un indicateur de niveau d'impact est plutôt proposé. Ceci est d'autant plus pertinent, étant donné que l'impact positif et la complexité d'implémentation peuvent être transformés en valeurs pécuniaires de coûts et de bénéfices.

Afin de rendre cette démarche effective, une plateforme Web java EE a été implémentée, avec un ensemble de métriques intermédiaires et agrégées, qui permettent l'analyse en aval et une prise de décision rapide.

Une étude de cas a été également menée : elle porte sur l'application de la démarche *PortfolioDQAF* à l'assainissement de *Data Assets* gouvernementaux. La plateforme développée ainsi que les conclusions de l'étude de cas sont présentées au niveau des chapitres III et V.

Ce chapitre est segmenté comme suit : la section 2 introduit l'AE comme cadre de travail pour la démarche *PortfolioDQAF*. La section 3 présente le cycle de vie des projets de qualité des données. La section 4 décrit ensuite la démarche *PortfolioDQAF* en termes de : structure générale et étapes qui composent cette démarche.

II.2. Analyse de l'AE pour l'évaluation de l'impact de la qualité des données

Dans un contexte d'AE, la stratégie business d'une organisation est portée par ses processus métier clés. La qualité de ces processus repose sur la qualité des données critiques qu'ils manipulent. En effet, les applications opérationnelles, analytiques et décisionnelles reposent sur des données de qualité élevée.

En effet, les processus métier sont de plus en plus automatisés, la qualité des données devient donc un facteur bloquant à la productivité, à la performance des processus et à la satisfaction client.

Au cours de leur cycle de vie, les données sont créées ou acquises, puis réutilisées à plusieurs reprises. La détérioration de la qualité des données peut impacter négativement l'exécution des processus métier qui les manipulent et dès lors, les objectifs financiers/métier de l'organisation.

Il est donc primordial de :

- Identifier les processus métier clés qui dépendent d'une qualité de données élevée ;
- Comprendre comment ces processus utilisent lesdites données ;
- Comprendre les attentes des usagers des données.

L'AE permet de faire communiquer les différentes parties prenantes de l'organisation et de faciliter la compréhension des systèmes complexes qu'elle permet de modéliser. Elle permet surtout l'alignement business-IT, dans la mesure où elle représente les relations entre la structure organisationnelle, les processus métier et la stratégie SI. La notion de service serait donc comprise de manière uniforme par les divisions métier et la DSI. L'AE peut donc être utilisée comme un outil de management stratégique et de gestion des processus en vue de répondre aussi bien aux besoins des clients qu'aux attentes des organisations partenaires et autorités de régulation.

Cette partie analyse l'attribut « qualité » à travers un modèle d'AE. L'objectif étant de mettre en évidence l'incidence de la qualité des données et celle des processus métier sur la qualité globale des services métier et dès lors, les objectifs de l'organisation.

En premier lieu, les principaux concepts de l'AE, ses cadres de référence ainsi que la discipline de l'analyse de l'AE sont présentés de manière succincte. Dans un deuxième temps, les modèles d'AE sont utilisés pour ressortir les dépendances entre les attributs « qualité » des processus et des objets métier. La section 4 présente la démarche *PortfolioDQAF*, à travers la présentation d'un modèle et des étapes qui composent cette démarche. Les métriques intermédiaires et agrégées qui composent le volet quantitatif de *PortfolioDQAF* sont également développées.

II.2.2. Architecture d'Entreprise : concepts, cadres de référence et analyse

II.2.1.1 Concepts et définitions

L'AE peut être définie comme « un ensemble cohérent de principes, méthodes et modèles qui sont utilisés dans la conception et l'implémentation de la structure organisationnelle d'une entreprise, de ses processus métier, de son système d'information et de son infrastructure » (Lankhorst, 2009).

Selon Ross et al. (2006), l'AE est « la logique structurante pour les processus métier et l'infrastructure informatique, reflétant les exigences d'intégration et de standardisation du modèle opératoire de l'entreprise. L'architecture d'entreprise fournit une vision à long terme des processus, des systèmes et des technologies de l'entreprise afin que les projets individuels puissent construire des capacités et non pas simplement répondre à des besoins immédiats ».

Dépendamment des références, l'AE est composée de plusieurs domaines d'architecture, variant de 4 à 5 couches. Certains des premiers modèles, à savoir EAP[12] (Spewak, 1993) et *The Zachman Framework* (Inmon et al., 1997), divisent l'AE en 4 couches :

- **La couche métier -** les fonctions métier qui exposent des services aux entités externes et internes ;
- **La couche données -** les données métier ainsi que toutes les autres données stockées au niveau du SI de l'entreprise et qui ont de la valeur ajoutée ;
- **La couche application -** les applications métier qui offrent des services aux autres applications ainsi qu'aux fonctions métier ;
- **La couche infrastructure SI -** le matériel, le réseau ainsi que les plateformes applicatives.

En plus des 4 couches susmentionnées, Le FEAF[13] (1999), l'AE de référence du gouvernement fédéral américain, cite également la couche « Environnement », en dessus de la couche métier. Cette couche est assimilée aux entités et activités externes, qui sont dirigées et supervisées par le métier.

La figure 4 représente les couches de l'AE.

[12] *Enterprise Architecture Planning*
[13] *The Federal Enterprise Architecture Framework*

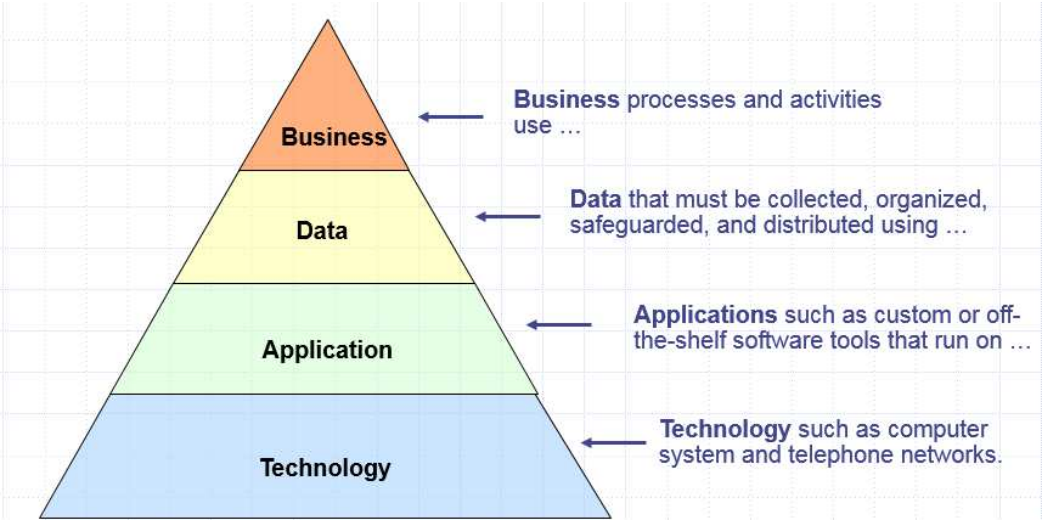


Figure 4. Couches de l'Architecture d'Entreprise (USDA, 2006)

Le concept des services métier est au cœur de l'AE, du fait qu'il s'agit d'un concept assimilable par les managers métier et les analystes SI. Un service métier peut être interne ou externe. Il est assimilé à une unité de fonctionnalités qu'un fournisseur de services expose à son environnement, via une interface, tandis que les processus internes sont cachés, ce qui produit une valeur ajoutée directe au client. Cette valeur peut être de nature matérielle ou immatérielle.

Un service métier est concrètement implémenté par un ou plusieurs processus métier automatisés ou semi-automatisés. Un processus métier est soit unitaire, soit décomposable en plusieurs sous-processus.

Un objet métier est un ensemble d'attributs qui représente une pièce d'information et qui est accédé en mode lecture ou écriture par un processus. Les objets métier produits par un processus peuvent servir de données d'entrées à un autre processus en aval.

Du fait que les processus métier accèdent en mode lecture ou écriture aux objets métier, il est naturel que la qualité des données ait un impact sur le résultat d'exécution de ces processus métier et inversement.

II.2.1.2 Cadres de référence

Un modèle d'AE permet de représenter l'architecture actuelle *(as-is)* et l'état futur *(to-be)* d'une entreprise ou d'une administration. Cependant, un cadre de référence d'AE offre :

- Un ou plusieurs méta-modèles pour la description de l'AE (détail des différentes couches) ;
- Une ou plusieurs méthodes pour la conception de l'AE ;
- Un vocabulaire commun pour l'AE ;

- Des modèles de référence qui peuvent être utilisée comme des *templates* pour la conception de l'AE.

De nombreux cadres de référence existent pour l'AE, dont les plus répondus sont regroupés dans le tableau 8 :

Tableau 8. Outils de modélisation de l'Architecture d'Entreprise

Outil	Nature
Archimate *(The Open Group)*	Langage de modélisation
TOGAF *(The Open Group)*	Framework
Zachman Framework *(Zachman, 1997)*	Framework
UML *(OMG[14])*	Langage de modélisation pour le génie logiciel
BPMN *(OMG)*	Langage de modélisation

II.2.1.3 Analyse de l'AE

La modélisation de l'AE permet donc de dresser la structure de l'entreprise, ses produits, opérations, technologies, les liens entre ces éléments et les éléments environnants. L'analyse de l'AE permet quant à elle, d'effectuer une analyse basée sur les modèles[15], de comparer des designs alternatifs et de prendre des décisions éclairées, en tenant compte d'aspects tels que : la qualité, le coût et la performance (Lankhorst, 2009).

Comme susmentionné, il existe plusieurs outils graphiques pour modéliser l'AE, parmi lesquels le langage UML et les frameworks TOGAF et Zachman. Cependant l'outil Archimate *(the Open Group)* est plus approprié pour l'analyse de l'AE, puisque :

- Il a une portée plus large qu'UML ;
- Il est particulièrement adapté à la modélisation abstraite ;
- Il permet de quantifier aisément les relations d'accès, d'utilisation, de réalisation, etc., de son méta-modèle ;
- Il intègre des supports pour les frameworks TOGAF et Zachman.

D'autres outils et langages peuvent également être utilisés comme IP-UML (Batini & Scannapieca, 2006) ou le BPML *(Business Process Modeling Language)* pour modéliser les processus métier.

[14] *The Object Management Group*
[15] *Model-based analysis*

La sous-section II.2.3 propose une méthode d'analyse de l'AE, basée sur les réseaux de Bayes, pour évaluer, de manière quantitative, l'impact de la qualité des données et plus particulièrement l'impact de la précision des objets données, sur la qualité des processus métier.

Cette méthode est orientée modèle et utilise le langage Archimate pour modéliser les objets les plus pertinents pour cette méthode, à savoir les objets de la couche « métier » et « données ».

II.2.3. Analyse de l'AE : application à la précision des données

II.2.3.1 Méta-modèle Archimate

Le méta-modèle Archimate permet de reprendre les entités les plus pertinentes pour l'évaluation de l'impact de la qualité des données. Ces entités correspondent aux : clients (internes ou externes à l'organisation), services métier, processus et objets métier ainsi que les objets données. Ces entités ainsi que les liens entre elles sont représentés dans la figure 5.

Dans le cadre de cette méthode, le méta-modèle Archimate est utilisé également comme base pour le développement du modèle logique, qui est une vue de haut niveau de l'espace d'analyse.

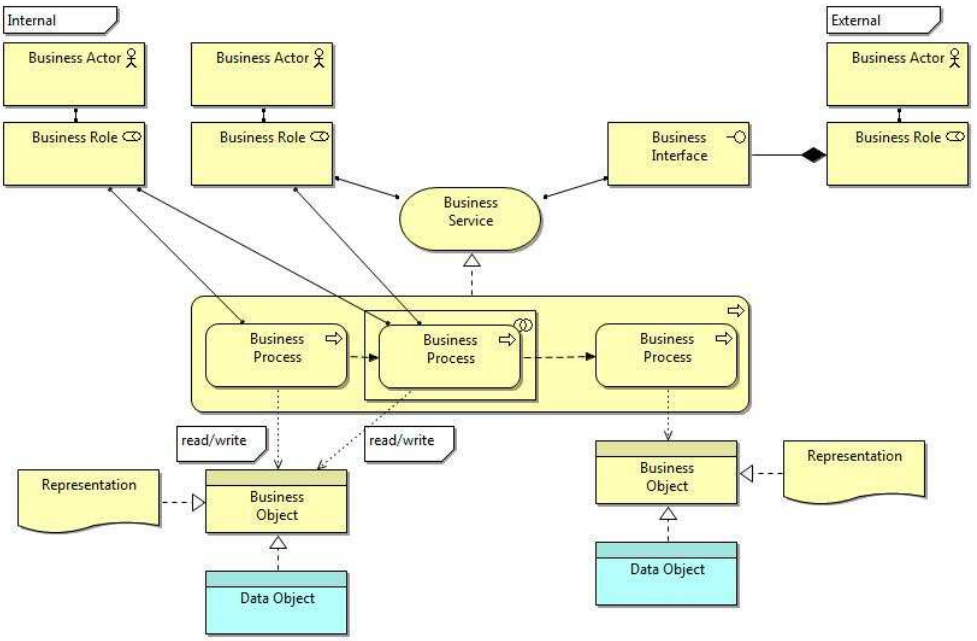


Figure 5. Méta-modèle Archimate

Selon ce méta-modèle, un service métier peut être sollicité par un client interne ou externe via une interface métier. En interne, un service métier est implémenté par un ou plusieurs processus, qui manipulent des données en entrée pour fournir un résultat. Deux processus peuvent accéder à la même donnée. Une donnée en sortie d'un processus peut également alimenter un autre processus en aval.

Au niveau du méta-modèle Archimate, la relation de réalisation *« realized by »* exprime un lien entre une entité logique, le service métier dans ce cas à une entité concrète qui la réalise. Dans le cas de la présente modélisation, il s'agit du processus métier. La relation « *accessed by »* modélise l'utilisation des objets métier par les services métier.

Il est supposé que la précondition de la disponibilité du service métier est réalisée. La précision du service métier dépend de la précision d'exécution du processus métier qui réalise le service métier, ainsi que de l'objet métier qui est accédé en mode lecture/écriture.

Le diagramme d'activités permet de modéliser les étapes de l'interaction entre le client et un service métier, de relever les entités les plus pertinentes qui participent à améliorer ou détériorer la qualité de livraison du service métier et enfin de développer un framework de calcul pour qualifier l'état actuel de la qualité de livraison. Le diagramme de la figure 6 décrit les étapes de l'interaction entre le client et le fournisseur du service. Une réponse adéquate à la requête du client implique :

- La disponibilité du service, autrement le client ne peut pas initier la requête ou son expérience client est dégradé au point d'abandonner la requête ;
- La précision, complétude et actualisation des données nécessaires au traitement de la requête ;
- La précision d'exécution du processus qui implémente le service ;
- La livraison à temps du service qui est la dernière étape en aval et qui dépend de toutes les étapes en amont.

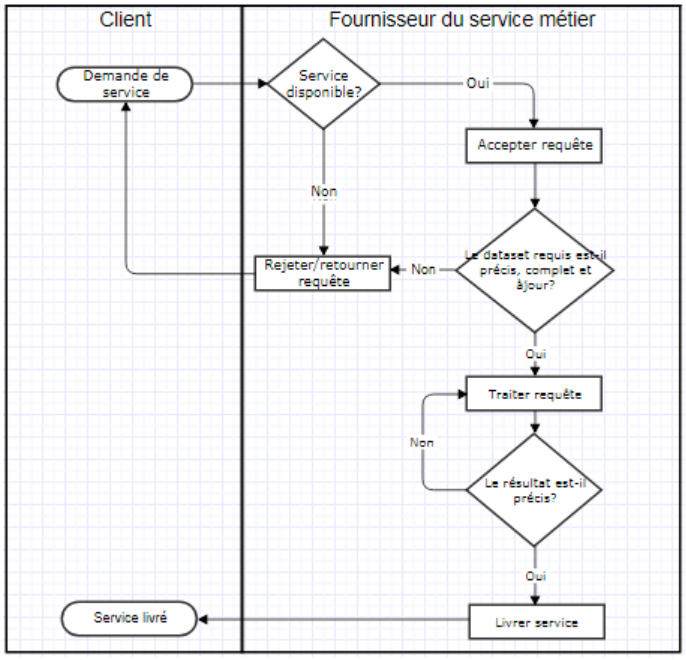


Figure 6. Diagramme d'activités (Belhiah et al., 2015a)

Le diagramme d'activités montre comment la qualité de l'expérience client peut être rehaussée ou détériorée selon la précision, la complétude et l'actualisation des données manipulées ainsi que la précision d'exécution des processus.

Il est possible donc de déduire à partir du méta-modèle Archimate et du diagramme d'activités, comment la qualité des données et la qualité d'exécution des processus affecte la qualité du service métier.

En effet, l'objectif étant d'évaluer l'état actuel de la qualité du service métier d'un point de vue client (*as-is*) pour évoluer vers un état optimisé (*to-be*), qui répond aux attentes dudit client, il est important d'associer des métriques à la disponibilité, la précision et la livraison à temps du service. Les métriques proposées pour le calcul de ses aspects sont la disponibilité du service, la précision et la livraison à temps.

- Disponibilité du service métier :

$$\frac{\text{nombre de requêtes prises en charge}}{\text{nombre total de requêtes reçues}} \quad (2)$$

- Précision du service métier :

$$\frac{\text{nombre de requêtes traitées avec précision}}{\text{nombre total de requêtes}} \quad (3)$$

- Livraison à temps du service métier :

$$\frac{\text{nombre de requêtes livrés à temps selon le SLA}}{\text{nombre total de requêtes}} \quad (4)$$

La sous-section suivante présente comment les réseaux de Bayes peuvent être combinés avec le modèle logique, issu du méta-modèle Archimate pour analyser la qualité des données et particulièrement la dimension de la précision. Cette approche met l'accent sur la précision du résultat d'un service métier et comment cette précision est paramétrée par la précision du processus métier qui l'implémente et de la donnée qu'il consomme.

II.2.3.2 Framework de calcul de la précision

Modèle logique d'évaluation de la précision - Un modèle logique est une vue statique des entités et relations qui forment l'espace d'analyse. C'est une vue de haut niveau. Elle appartient au niveau M1 de l'architecture 4-tiers de l'OMG (OMG, 2017).

La figure 7 représente les classes et les entités du modèle logique, objet de cette démarche.

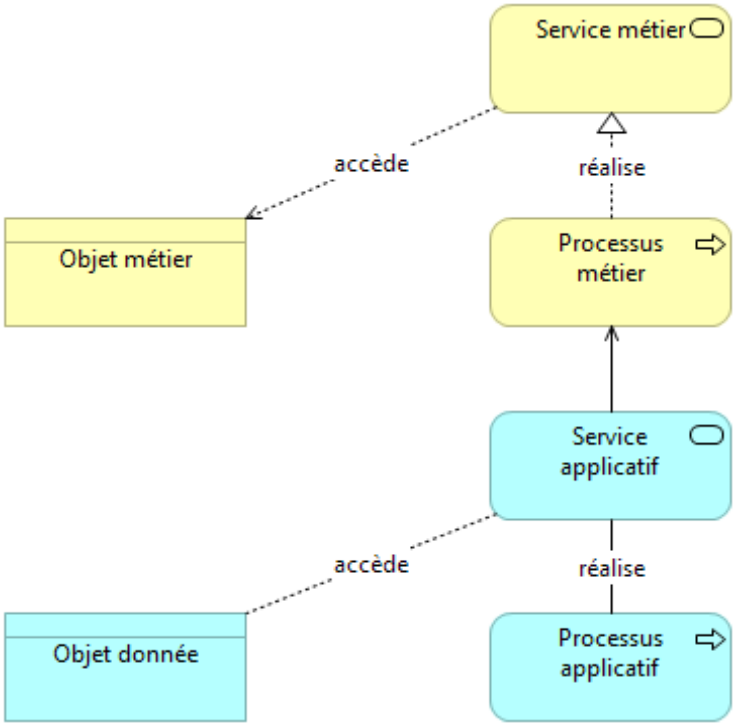


Figure 7. Modèle logique Archimate

A partir de la figure 7, il est possible de conclure que les classes et les relations dans le modèle abstrait sont conformes au méta-modèle Archimate, comme représenté par la figure 6.

Modèle concret - Le modèle concret correspond à l'instanciation du modèle logique. Cette vue appartient au niveau M0 de l'architecture 4-tiers de l'OMG (OMG, 2017). Si les entités de la couche métier, à savoir : le service métier, le processus métier et l'objet métier, sont uniquement considérés, un exemple d'instanciation serait comme illustré par la figure 8 :

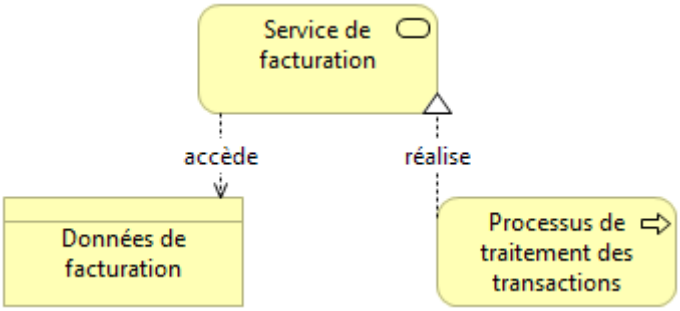


Figure 8. Modèle concret Archimate : exemple de l'instanciation du modèle logique

Adaptation du méta-modèle Archimate pour évaluer l'impact de la précision des données - La figure 9 représente l'adaptation du modèle logique illustré au niveau de la figure 7. Les entités objets de cette analyse ont été annotées avec les attributs exprimant la qualité.

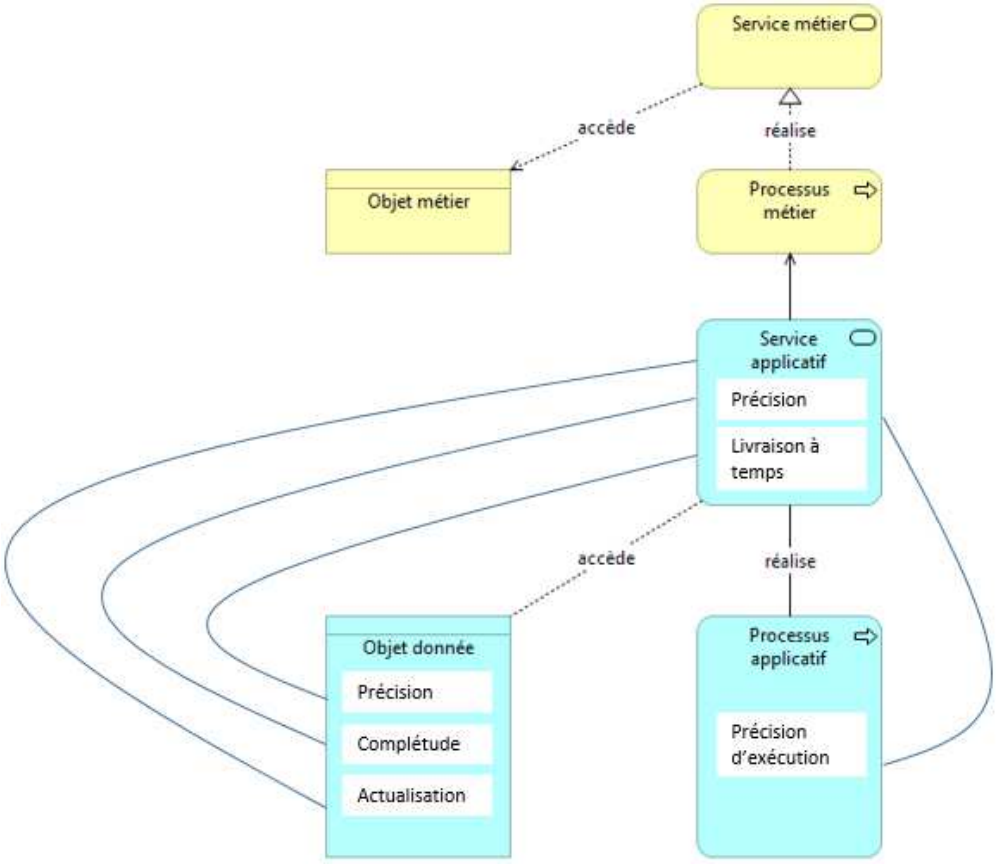


Figure 9. Adaptation du modèle Archimate

Réseau de Bayes - Un réseau bayésien appartient à la famille des modèles graphiques probabilistes. Il constitue un ensemble de variables aléatoires respectant une structure de dépendances/indépendances conditionnelles traduite par un graphe acyclique dirigé.

Les attributs et les relations entre les attributs du méta-modèle d'Archimate s'apparentent aux nœuds et aux relations causales des réseaux de Bayes.

Friedman et al. (2000) et Nielsen (2001) décrivent les réseaux de Bayes, B= (G, P), comme une représentation d'une distribution de probabilité conjointe, où G= (V, E) est un graphe acyclique dirigé, constitué de sommets V et d'arêtes E. Les sommets désignent un domaine de variables aléatoires $X_1, \ldots, X_n$, également appelées nœuds aléatoires. Dans le contexte des modèles abstraits, chaque nœud aléatoire correspond à un attribut. Chaque nœud, X_i, peut prendre une valeur x_i du domaine fini $Val(X_i)$. Les arêtes désignent des dépendances de causalité entre les nœuds, en d'autres termes, elles désignent les relations causales entre les nœuds. Le second composant du réseau, P, décrit une distribution de probabilité conditionnelle pour chaque nœud aléatoire, P (X_i), étant donné ses parents P_a (X_i) dans le graphe G. Il est possible de décrire la distribution de la

probabilité conjointe des domaines $X_1,\ldots, X_n$ en utilisant la règle de probabilité en chaîne, sous la forme du produit :

$$P(X1, \ldots, Xn) = \prod_{i=1}^{n} P(Xi \setminus \text{Pa(Xi)}) \quad (5)$$

Afin de spécifier la distribution conjointe, les probabilités conditionnelles respectives qui participent au produit doivent être définies. Ces probabilités sont représentées dans la matrice de probabilités conditionnelles (CPM[16]). En ayant recours aux réseaux de Bayes, il est possible de répondre aux questions de type : quelle est la probabilité que la variable X soit dans l'état x_1, sachant que $Y=y_2$ et $Z=z_3$. Le tableau 9 contient une description générale d'une matrice de probabilité conditionnelle, pour les trois nœuds X, Y et Z, le premier étant affecté de manière causale par les deux autres.

Tableau 9. Description générale de la matrice de probabilité conditionnelle

Z		z_1		z_2	
Y		y_1	y_2	y_1	y_2
X	x_1	$P(x_1 \mid y_1,z_1)$	$P(x_1 \mid y_2,z_1)$	$P(x_1 \mid y_1,z_2)$	$P(x_1 \mid y_2,z_2)$
	x_2	$P(x_2 \mid y_1,z_1)$	$P(x_2 \mid y_2,z_1)$	$P(x_2 \mid y_1,z_2)$	$P(x_2 \mid y_2,z_2)$

Une littérature plus complète et exhaustive sur les réseaux de Bayes peut être trouvée dans (Jensen, 1996 ; Naïm, 2011 ; Pearl, 2003).

Framework de calcul de la précision - La précision d'un processus métier peut être exprimée par le degré par lequel le système, étant donné des données précises en entrée, produit des données correctes en sortie.

Habituellement, la précision est associée avec les objets de type « données », mais dans les faits la précision concerne également la fonctionnalité. En effet, si l'implémentation de l'algorithme de traitement est erronée, le résultat ne serait pas précis, malgré une entrée précise. Le service applicatif est l'entité de haut niveau pour laquelle il est possible d'évaluer la précision.

A partir de la figure 10, il est possible de dire que la précision du service applicatif est dépendante de la précision d'exécution du processus applicatif et de la précision de l'objet donnée.

Le réseau de Bayes représentant ceci est illustré par la figure 10.

[16] *Conditional Probability Matrix*

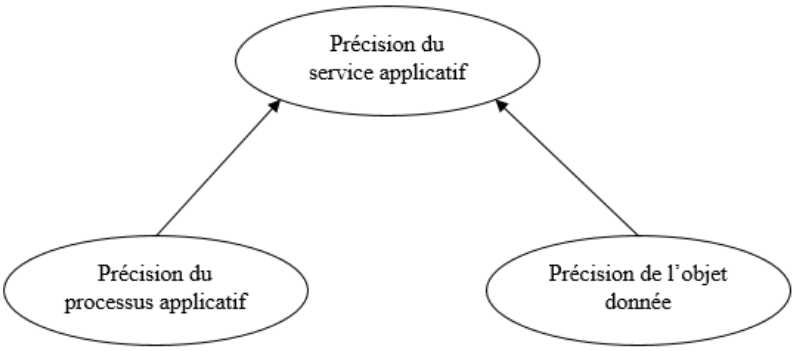


Figure 10. Réseau de Bayes

Soit :

- X : la probabilité que le résultat du service applicatif soit précis
- Y : la précision d'exécution du processus applicatif
- Z : la précision de l'objet donnée

La probabilité qu'Y soit précis, sachant que Z est précis, est exprimée en termes de probabilités élémentaires et conditionnelles :

$$P(X) = P(Y \setminus Z) = \frac{P(Y \cap Z)}{P(Z)} \qquad (6)$$

Dans l'équation susmentionnée, le numérateur est la probabilité qu'Y et Z se réalisent en même temps et le dénominateur est la probabilité que Z se réalise. La translation de ce modèle en une matrice de probabilité conditionnelle, utilisant la relation « AND » est démontrée dans le tableau 10. La matrice de probabilité conditionnelle contient les statuts « précis » et « non précis ».

Tableau 10. Matrice de probabilité conditionnelle

Précision d'exécution du processus applicatif		**Précis**		**Non précis**	
Précision de l'objet donné		Précis	Non précis	Précis	Non précis
Précision du service applicatif	Précis	x			
	Non précis		x	x	x

Après avoir démontré le lien causal entre la qualité des données d'un SI et la qualité des services applicatifs délivrés, la section suivante consiste à définir de manière quantitative comment la qualité de ces mêmes données influent les objectifs financiers et métier d'une organisation, ainsi que le coût associé à l'amélioration de cette qualité.

Avant de présenter la démarche *PortfolioDQAF* dans la section II.4, il est convenable de présenter les phases communes qui composent le cycle de vie d'un projet de qualité des données.

II.3. Cycle de vie des projets de qualité des données

Les composantes principales du processus d'amélioration de la qualité sont : la définition et la caractérisation de la qualité des données, l'analyse de l'impact des niveaux de qualité sur l'organisation et l'amélioration de la qualité des données via des solutions techniques et métier.

L'approche TDQM (Madnick & Wang, 1992) englobe les phases : (i) définir ; (ii) mesurer ; (iii) analyser ; (iv) améliorer. Ce cycle exprime l'amélioration continue de la qualité.

La figure 11 schématise ce processus.

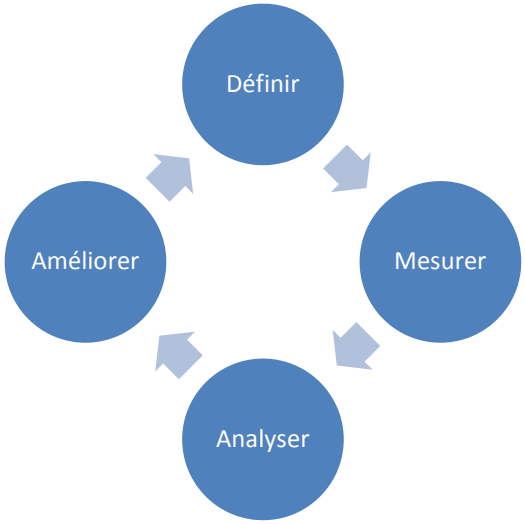


Figure 11. Cycle de vie d'un projet de qualité des données

- **Définir -** Définir la liste des dimensions de la qualité des données à partir du point de vue des rôles ou des personnes qui vont utiliser la donnée, en utilisant les outils appropriés : sondages, questionnaires, interviews, etc. ;
- **Mesurer -** Associer des métriques de la qualité des données aux dimensions définies dans la première étape ;
- **Analyser -** Interpréter les résultats des mesures ;
- **Améliorer -** Concevoir et implémenter les scénarios d'amélioration des données et des processus pour répondre aux spécifications en termes de qualité des données.

II.4. Démarche *PortfolioDQAF*

PortfolioDQAF est une démarche orientée évaluation, basée sur un modèle qualtimétrique, qui s'inspire du modèle « facteurs-critères-métriques » (McCall, 1979). Pour les projets d'évaluation de la qualité des données, les facteurs évalués sont :

- L'impact positif sur les objectifs financiers et métier de l'organisation, qui pourrait être assimilé à la valeur positive de l'amélioration de la qualité ;
- La complexité d'implémentation des initiatives d'amélioration de la qualité, associée aux coûts d'amélioration de la qualité.

Chaque facteur est décomposé en critères, définis de concert avec les analystes SI et les managers métier. Selon son importance, le critère participe à construire des indicateurs quantifiables qui permettent d'avoir des valeurs agrégées de l'impact positif et de la complexité d'implémentation. *PortfolioDQAF* s'adresse aussi bien aux managers métier qu'aux analystes SI et permet à chacun depuis sa position, d'évaluer l'utilité de l'engagement de ressources fiscales et autres sur la stratégie ou le programme d'amélioration de la qualité des données.

II.4.1. Modèle *PortfolioDQAF*

II.4.1.1 Structure générale

Le modèle *PortfolioDQAF* est conçu pour caractériser l'analyse coûts-avantages d'un portefeuille de projets d'amélioration de la qualité des données. Il s'agit d'un modèle orienté usager, identifiant un ensemble de critères qui décrivent les facteurs de l'impact positif et de la complexité d'implémentation du point de vue de l'utilisateur des données. Le modèle *PortfolioDQAF* permet également de nuancer ces facteurs, selon le contexte propre à chaque organisation. Ceci est possible en ayant recours à la pondération des critères selon leur importance, paramétrée par les usagers des processus et des données.

Aussi, la structure générale de *PortfolioDQAF* s'intéresse-t-elle à :

1. L'identification des objectifs financiers/métier dépendant d'un niveau de qualité élevé. Ces objectifs sont identifiés à partir du point de vue de l'organisation ;
2. L'identification des exigences par rapport à la qualité des données. Il s'agit des dimensions de la qualité des données objets de l'amélioration. Ces dimensions sont orientées usager des données ;
3. La caractérisation de l'impact positif par des critères ;
4. La définition d'un indicateur agrégé de l'impact positif ;
5. Pour chaque aspect de la qualité des données, la définition des intrants qui affectent la complexité d'implémentation des projets d'amélioration ;
6. La définition d'un indicateur agrégé de la complexité d'implémentation.

Pour l'ensemble des objets métier à améliorer, *PortfolioDQAF* permet in fine, d'évaluer les ratios coûts-bénéfices pour établir un business case optimal pour l'amélioration de la qualité des données. Ce business case correspondrait au niveau optimal d'investissement dans la qualité des données.

Le modèle *PortfolioDQAF* schématisé dans la figure 12, est doté d'une approche structurée d'évaluation, prenant en considération les particularités de chaque organisation en termes de ses objectifs financiers/métier, à travers l'amélioration de la qualité de ses objets métier critiques. *PortfolioDQAF* est doté également d'un mécanisme d'agrégation de métriques, permettant ainsi d'associer des mesures quantifiables aux facteurs coûts et avantages. Ceci rend aisées l'analyse rapide et la prise de décision.

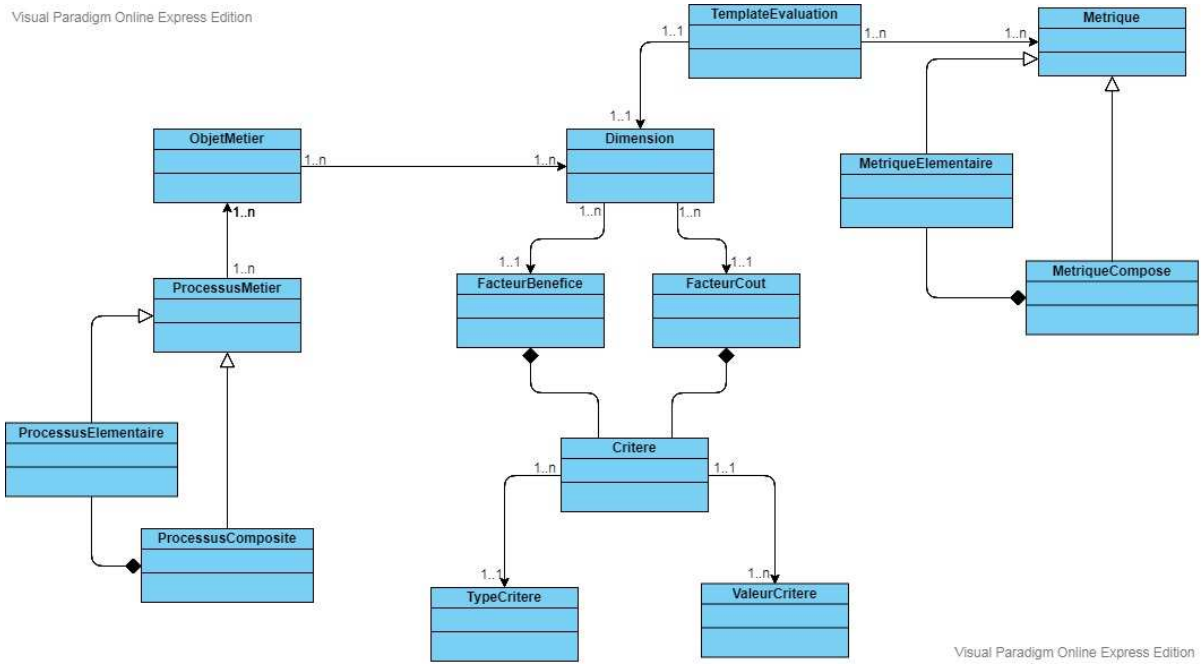


Figure 12. Modèle *PortfolioDQAF*

Ce modèle permet de fournir une vue détaillée des critères d'évaluation de l'impact positif et de la complexité d'implémentation des projets d'amélioration de la qualité des données. L'objet « aspect » du modèle *PortfolioDQAF* correspond à la dimension de la qualité des données que l'organisation souhaite évaluer et améliorer.

Le modèle permet d'abord de caractériser l'impact et la complexité à travers plusieurs critères et de les évaluer ensuite. Il est possible donc de dire que le modèle couvre les 2 volets :

- Qualitatif correspondant pour chaque dimension évaluée de la qualité, à l'éclatement des facteurs suscités en critères d'évaluation ;
- Quantitatif incluant les templates d'évaluation ainsi que les métriques génériques.

Cette section décrit comment :

  - le facteur d'impact agrège l'impact des niveaux de qualité des données sur les objectifs de l'organisation ;

- le facteur de complexité agrège ce facteur à partir de l'ensemble des critères qui le caractérisent.

II.4.1.2 Arbre qualimétrique

Les aspects de la qualité des données pris en charge par *PortfolioDQAF* correspondent aux 15 dimensions de la qualité (voir Tableau 2). La figure 13 correspond à l'arbre qualimétrique pris en charge par *PortfolioDQAF*.

Figure 13. Arbre qualimétrique de *PortfolioDQAF*

Comme susmentionné, l'évaluation dans le cadre de *PortfolioDQAF* porte sur les facteurs de coûts et bénéfices. Les critères qui composent le facteur de coûts sont de nature tangible et intangible, tandis que ceux qui composent le facteur de bénéfices couvrent les avantages d'ordre financier et métier.

La sous-section suivante traite le volet quantitatif de *PortfolioDQAF*, soit la manière avec laquelle *PortfolioDQAF* associe des métriques quantifiables aux facteurs d'impact et de complexité d'implémentation, ainsi qu'aux critères qui leur sont associés.

II.4.1.3 Métriques PortfolioDQAF d'évaluation

Dans le but de rendre aisée la construction d'un business case optimal pour l'amélioration de la qualité des données, *PortfolioDQAF* s'appuie sur la collecte des notations attribuées à chaque critère. Ces notations correspondent aux valeurs intermédiaires de l'évaluation, persistées au niveau de *PortfolioDQAF*, et agrégées ensuite pour construire les deux facteurs d'évaluation, à savoir l'impact positif et la complexité d'implémentation.

En effet, pour parvenir à évaluer les deux facteurs suscités, ils sont décomposés en plusieurs critères. Selon son importance relative, chaque critère participe à la construction de la métrique associée. Quantitativement, cette importance est exprimée à travers des coefficients de pondération configurables et sensibles au contexte. Ceci permet également d'adopter l'approche *PortfolioDQAF* dans des environnements variés, avec pas ou peu d'ajustements.

En d'autres termes, l'objectif derrière l'utilisation d'un coefficient de pondération est de permettre à chaque organisation d'exprimer l'importance de chacune des dimensions de la qualité des données, selon son environnement et sa stratégie. Ci-après, un ensemble de situations qui montrent la pertinence de l'utilisation des coefficients pondérateurs :

- Les producteurs de données dont la nature est « *master data* » (données géographiques par exemple), qui ne sont pas soumises à des mises-à-jour fréquentes, peuvent choisir d'accorder une plus grande importance à la précision et à la complétude, plutôt qu'à l'actualisation ;
- Les départements publics qui produisent des données financières comme les départements d'économie et des finances, peuvent accorder une importance égale à la précision et à l'actualisation, à cause de la nature volatile de ces données ;
- D'autres organismes, issus d'autres domaines, peuvent accorder le même poids à toutes les dimensions.

Pour agréger les notations intermédiaires en un seul indicateur quantitatif, l'analyse multicritères par la méthode de multiplication de ratios (voir sous-section I.3.4) est utilisée.

In fine, cette étape permet d'avoir une valeur pour l'impact positif et la complexité d'implémentation. Ce qui facilite par la suite, l'analyse rapide et la prise de décision par rapport aux initiatives de qualité des données à adopter.

II.4.2. Approche *PortfolioDQAF*

L'approche *PortfolioDQAF* (Belhiah et al., 2015b) permet de caractériser les critères qui permettent d'évaluer les indicateurs de bénéfices et de coûts liés à un programme d'amélioration de la qualité des données.

Le point de départ de l'approche est la définition des objectifs financiers/métier de l'organisation : ceux-ci dépendent du domaine d'activités de l'organisation et de sa stratégie. Sur le plan qualitatif, il faut ensuite arrêter la liste des dimensions de la qualité des données qu'il faut améliorer. Ensuite et pour pouvoir déterminer la valeur des améliorations de la qualité des données ainsi que le coût de réalisation, les facteurs de coût et d'impact positif sont décomposés en critères.

Le volet quantitatif consiste à définir les métriques d'évaluation comme développé dans la section II.5. Il en résulte l'agrégation des mesures intermédiaires (quoique persistées pour les besoins d'analyse), en deux indicateurs d'impact positif et de complexité d'implémentation.

La dernière phase consiste à analyser les mesures intermédiaires et consolidées pour construire le business case optimal pour l'amélioration de la qualité des données, en utilisant les avantages déterminés à l'étape (6) et les coûts associés, déterminés à l'étape (7), comme le montre la figure 14. Cette figure illustre le séquencement des phases de la démarche *PortfolioDQAF*.

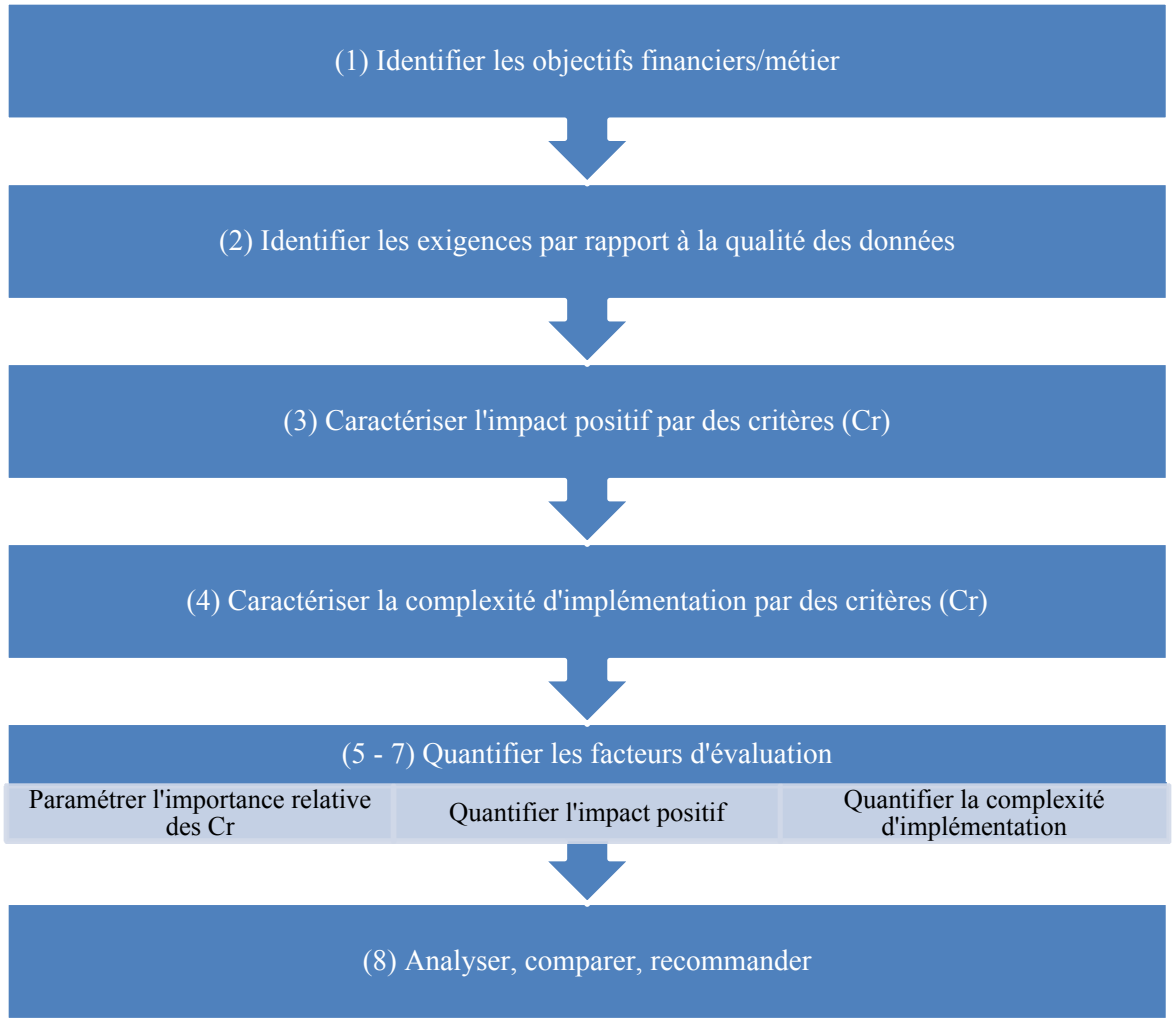


Figure 14. Approche *PortfolioDQAF*

(1) Etape d'identification des objectifs financiers/métier - La première phase de la démarche *PortfolioDQAF* consiste à définir les objectifs financiers/métier de l'organisation. L'objectif étant de comprendre par la suite comment ces mêmes objectifs sont portés par les processus métier, qui dépendent d'une qualité de données élevées pour leur exécution. Parmi les objectifs financiers/métier qu'il est possible de considérer (Gartner, 2011) :

- Efficacité opérationnelle ;
- Augmentation des revenus ;
- Amélioration de la productivité ;
- Réduction des coûts ;
- Amélioration de la satisfaction client ;
- Efficacité des fournisseurs ;

- Réactivité du marché ;
- Conformité aux autorités de régulation ;
- Amélioration de la prise de décision ;
- Amélioration de l'analyse en aval ;
- Réactivité des fonctions support (RH et SI) ;
- Autre.

(2) Etape d'identification des dimensions de la DQ - Même s'il existe une quinzaine de dimensions de la qualité des données, quasi toutes les études de cas qui avaient pour objectif l'évaluation et l'amélioration de la qualité des données, choisissent un sous-ensemble des dimensions de la qualité, dépendamment des objectifs de l'étude (Aladwani et al., 2002 ; Batini et al., 2012 ; Belhiah & Bounabat, 2017 ; Catarci & Scannapieco, 2002 ; Haug et al., 2011 ; Narman et al., 2009).

Cette démarche est également préconisée lors du développement de *PortfolioDQAF* pour les volets : données d'entreprise et Open Data présentés au niveau des chapitres III et IV.

Au niveau de cette phase, l'état actuel de la qualité des données est évalué à travers le prisme des dimensions de la qualité arrêtées, telles que : la précision, la complétude, l'actualisation, la consistance, etc.

(3-4) Etape d'identification des critères d'évaluation - Cette étape consiste à construire les facteurs d'impact positif et de complexité d'implémentation. La construction d'un indicateur n'est pas neutre. S'il permet de synthétiser des problèmes multidimensionnels pour la prise de décision, de faire une comparaison rapide et de faciliter l'annotation des objets auxquels il est rattaché pour le grand public, il est important qu'il soit bien construit pour ne pas conduire à des résultats et décisions erronés. En effet, il ne doit pas cacher les faiblesses de certaines dimensions si les pondérations sont mal choisies ; à cet égard, les pondérations doivent faire l'objet d'ateliers de travail avec les experts du domaine pour les définir convenablement.

(5-7) Etape de quantification des facteurs d'évaluation - Afin de rendre aisée l'analyse et l'interprétation des facteurs d'impact et de complexité d'implémentation, des niveaux d'impact et de complexité leur ont été associés :

- **Paramétrer l'importance relative des critères** : en raison des spécificités de chaque organisation ainsi que ses propres facteurs de succès, et dans l'optique de fournir une approche générique qui peut être implémentée avec un moindre ajustement, cette étape introduit les coefficients de pondération configurables et sensibles au contexte ;
- **Quantifier l'impact positif** : pour identifier le niveau d'impact, l'approche *PortfolioDQAF* propose un score d'impact allant de 0 à 5, où 0 correspond à un impact nul

et 5 à un impact très élevé. Le tableau 11 illustre la correspondance entre le score et les niveaux d'impact ;

Tableau 11. Niveaux de l'impact positif

Score d'impact	Niveau d'impact
0 – 1.5	0 – Impact imperceptible à bas
1.5 – 3	1 – Impact moyen
3 – 4.25	2 – Impact élevé
> 4.25	3 – Impact très élevé

- **Quantifier la complexité d'implémentation** : selon les mêmes règles de définition du niveau d'impact, un score de complexité allant de 0 à 5 est proposé, où 0 correspond à une complexité imperceptible et 5 correspond à une complexité très élevée. Le tableau 12 illustre la correspondance entre le score et les niveaux de complexité. Cette évaluation est généralement effectué par les analystes SI ;

Tableau 12. Niveaux de la complexité d'implémentation

Score de complexité	Niveau de complexité
0 – 1.5	0 – complexité très faible à faible
1.5 – 3	1 – complexité moyenne
3 – 4.25	2 – complexité élevée
> 4.25	3 – complexité très élevée

(8) Etape d'analyse, de comparaison et de recommandation - la dernière étape permet d'analyser le rapport coûts-bénéfices, en utilisant les avantages déterminés à l'étape 6 et les coûts déterminés à l'étape 7. Les managers SI prennent généralement le relais à cette étape.

II.5. Conclusion

La démarche *PortfolioDQAF* proposé, met en exergue les projets d'amélioration de la qualité des données, les plus rentables en termes de ratio coûts-bénéfices. Il propose de définir et d'évaluer les facteurs de l'impact positif et de la complexité d'implémentation, caractérisant un projet d'amélioration de la qualité des données. *PortfolioDQAF* prend en considération les objectifs de l'organisation, dans le but d'apprécier correctement l'impact de la qualité des données sur la réalisation de ces objectifs.

Le modèle implémenté par *PortfolioDQAF* est un modèle orienté évaluation, incluant deux facteurs, décomposables en un ensemble de critères. C'est un modèle d'évaluation qualitatif et quantitatif, avec des métriques intermédiaires et agrégées. Il offre différentes grilles de lecture aux managers métier et aux analystes SI.

PortfolioDQAF est composé de deux niveaux : (i) le premier permet d'énumérer les facteurs et les critères d'évaluation ; (ii) le deuxième décrit la démarche et les mécanismes pour quantifier ces facteurs. Il s'appuie sur les modèles de l'Architecture d'Entreprise et décline un processus d'évaluation en 8 étapes (figure 14).

Le résultat du travail de cette thèse, développe donc comment mesurer, de manière quantitative, la valeur métier des projets d'amélioration de la précision des données, en établissant deux indicateurs globaux de l'impact positif et de la complexité d'implémentation.

Comme chaque organisation évolue dans un environnement qui lui est propre, il est intéressant de voir comment *PortfolioDQAF* s'adapte à différents contextes. Les chapitres III et IV développent respectivement les volets « données d'entreprise » et « Open Data ».

Chapitre III : Démarche *PortfolioDQAF* - volet « données d'entreprise »

III.1. Introduction

Ce chapitre décrit l'adaptation de la démarche *PortfolioDQAF* pour les données d'entreprise. Ce volet est développé et appliqué à l'étude de cas portant sur l'assainissement de *Data Assets* gouvernementaux.

Les données d'entreprise se distinguent des données de l'Open Data par plusieurs aspects : (i) les interlocuteurs sont les ressources internes de l'organisation ; (ii) les données d'entreprise critiques pour la stratégie de l'organisation sont connues à priori ; (iii) les processus qui utilisent les données ainsi que les flux de données sont cartographiées. Inversement, pour l'Open Data : (i) les données sont destinées à des utilisateurs externes à l'organisation ; (ii) l'utilisation, la réutilisation et l'intégration qui seront faites des données ne sont pas connues à priori ; (iii) les données qui sont importantes pour les usagers ne sont pas également connues au moment de la publication de l'Open Data.

La démarche *PortfolioDQAF* est adoptée sans ajustements pour l'évaluation des projets d'assainissement des données d'entreprise. Pour les données ouvertes, des étapes préparatoires permettent de caractériser les données critiques et les niveaux de qualité de l'Open Data, sont prévues. Pour la suite, la démarche *PortfolioDQAF* est inchangée.

Les sections de ce chapitre reprennent les phases de l'approche *PortfolioDQAF* en mettant le focus sur les spécifications du volet « données d'entreprise ». Les étapes qui concernent l'identification des objectifs financiers/métier et la caractérisation de l'impact positif de l'amélioration de la qualité des données sont intrinsèquement indépendantes de l'aspect (dimension) de la qualité des données, sujet d'amélioration. Cependant, l'étape d'évaluation de la complexité d'implémentation est fortement couplée avec la dimension de la qualité des données à améliorer.

De ce fait, l'exercice dans le présent chapitre couvre en premier lieu la dimension de la précision. D'après la revue de littérature, cette dimension semble représenter, avec la complétude et l'actualisation, les aspects les plus critiques de la qualité des données.

Ce chapitre se présente comme suit : la section 2 développe le volet « données d'entreprise » de *PortfolioDQAF*. La section 3 introduit le modèle de coût de la qualité de *PortfolioDQAF* proposé pour la précision. La section 4 présente la plateforme Web développée et qui automatise la démarche *PortfolioDQAF*.

III.2. PortfolioDQAF - volet données d'entreprise

III.2.1. Etape d'identification des objectifs financiers/métier

Dans le but de cerner comment la performance d'exécution et la qualité des processus métier influencent le succès d'une organisation, ses objectifs financiers/métier et ses résultats, des facteurs clés de la section 6 du chapitre 2 sont repris.

Parmi ce facteurs : l'efficacité opérationnelle, l'augmentation des revenus, l'amélioration de la productivité, la réduction des coûts, l'amélioration de la satisfaction client, la conformité aux autorités de régulation et l'amélioration de la prise de décision

III.2.2. Etape d'identification des exigences par rapport à la qualité

Comme mentionné au niveau de l'introduction de ce chapitre, parmi les 15 dimensions (figure 1) de la qualité des données, ce volet concerne la dimension de la précision.

Il est à retenir cependant que la méthodologie est réutilisable et transposable aux autres dimensions de la qualité, dans le cadre des données d'entreprise. Ceci est illustré par l'arbre qualtimétrique présenté au niveau du chapitre II.

III.2.3. Etape d'identification des critères d'évaluation

Comme développé à la sous-section II.4.2, l'objectif de cette étape est d'éclater les facteurs d'impact positif et de complexité d'implémentation en des critères qui en permettent l'analyse.

En plus des critères mentionnés dans la section II.4.2 et qui influencent principalement les objectifs de l'organisation, d'autres critères ont été additionnées pour l'évaluation de l'impact positif.

III.2.3.1. Critères de l'impact positif

Le facteur de l'impact positif est décomposable en plusieurs critères. Il s'agit des facteurs de succès de l'étape (1) de la figure 14, augmentés d'autres critères, à savoir :

- **Le caractère transverse du processus -** en effet, l'amélioration de la qualité d'une donnée critique, utilisée par un processus transverse a plus d'impact que pour un processus vertical ;
- **La nature de la donnée -** les données sont classifiées en : (i) master data ; (ii) données transactionnelles ; (iii) données d'historique. Il est possible de présumer que l'amélioration de la qualité des données de nature master data a plus d'impact que l'amélioration des données transactionnelles ou d'historique ;

- **La fréquence d'accès à la donnée -** si la donnée critique est utilisée plusieurs fois par le processus, l'amélioration de sa qualité aura plus d'impact positif ;
- **Le délai de réalisation -** à l'image des autres projets technologiques, un délai de réalisation court permet d'avoir des résultats rapidement. En effet, les délais allongés risquent d'induire la démotivation de l'équipe projet, le changement du périmètre du projet, l'évolution de la réglementation, entre autres.

III.2.3.2. Critères de la complexité d'implémentation

De même, le facteur de la complexité d'implémentation est éclaté en critères. Les critères suivants concernent l'évaluation de la complexité d'amélioration de la précision. Ces critères sont issus de la revue de littérature, mais également d'entretiens avec des managers SI.

Les critères considérés pour la complexité d'amélioration sont les :

- **Niveau de risques -** un niveau de risque élevé est proportionnel à un niveau de complexité d'implémentation considérable. Les risques peuvent correspondre à la perte de données, l'arrêt des opérations, les risques systémiques, les réactions en chaîne ainsi que le dépassement des coûts, délais, ressources allouées, etc. ;
- **Existence de normes pour valider les données -** l'existence de normes pour vérifier et valider les données permet de réduire la complexité de détection des valeurs erronées ;
- **Existence d'un référentiel de données -** l'existence d'une source de données de référence, même à l'extérieur du SI de l'organisation, permet de réduire la complexité de correction des inexactitudes au niveau de donnée, à travers un travail de confrontation ;
- **Potentiel d'identification par clé -** l'existence d'une clé primaire/identifiant global et consistant à travers les différentes sources de données permet de réduire la complexité d'assainissement des données, en rendant aisés la confrontation et le recoupement de données ;
- **Nature du traitement des données -** sur le plan technique, le projet d'amélioration de la précision peut consister en des traitements automatiques, semi-automatiques ou manuels. Un traitement manuel, dépendamment de la volumétrie des données à traiter, correspond à une charge et une complexité élevée ;
- **Volumétrie des données à traiter -** un volume de données élevé induit une charge et une complexité élevée.

III.2.4. Etape de quantification des facteurs d'évaluation

L'objectif de cette étape est de paramétrer l'importance relative de chaque critère, à travers les coefficients de pondération. Ceci permet à *PortfolioDQAF* de collecter des métriques

intermédiaires, avant de les agréger dans des indicateurs d'impact positif et de complexité d'implémentation, facilitant l'analyse et la prise de décision par rapport aux initiatives d'amélioration de la qualité des données.

III.2.4.1. Paramétrage de l'importance relative des critères

Pour opérationnaliser l'analyse multicritères par multiplication de ratios, un template illustré par le tableau 13 est utilisé pour évaluer l'impact positif. En effet, chaque critère d'impact répond à un ensemble de valeurs. A chaque valeur correspond une notation.

Pour chaque critère, un coefficient pondérateur est défini par les managers métier, permettant ainsi d'exprimer l'importance de la contribution de chaque critère à construire le facteur agrégé d'impact. Les coefficients de pondération sont propres à chaque organisation et décrivent son contexte et sa stratégie.

Tableau 13. Template d'évaluation de l'impact positif

Facteur	Valeurs	Notation (R)	Pondération (C)
Impact sur les opérations quotidiennes	- Vrai - Faux	1 0	
Impact sur les objectifs financiers/métier à court terme	- Augmenter les revenus - Améliorer la productivité - Réduire les coûts - Améliorer la satisfaction client - Se conformer aux autorités de régulation - Autres	0.15 0.15 0.15 0.15 0.15 0.15	
Impact sur la prise de décision	- Vrai - Faux	1 0	
Impact sur l'analyse en aval	- Vrai - Faux	1 0	
Caractère transverse du processus	- Vrai - Faux	1 0	
Délai de réalisation opportun	- Vrai - Faux	1 0	
Nature de la donnée	- Master data - Donnée transactionnelle - Donnée d'historique	4 2 0	

Un template similaire est adopté pour l'évaluation de la complexité d'implémentation. A la différence du paramétrage de l'analyse multicritères du facteur d'impact, la définition des coefficients de pondération des critères du facteur de complexité d'implémentation incombe aux responsables SI.

Il est à noter également que les activités de profilage des données permettent de renseigner convenablement le template illustré par le tableau 14. Ce tableau représente le template de paramétrage du facteur de la complexité d'implémentation :

Tableau 14. Template d'évaluation de la complexité d'implémentation

Facteur	Valeurs	Notation (R)	Pondération (C)
Niveau de risque	- Sévère/inacceptable - Majeur - Moyen - Mineur - Imperceptible	1 0,75 0,5 0,25 0	
Existence de normes pour valider les données	- Faux - Vrai	1 0	
Existence d'un référentiel de données	- Faux - Vrai	1 0	
Potentiel d'identification par clé primaire	- Faux - Vrai	0 1	
Nature du traitement des données	- Manuel - Semi-automatique - Automatique	1 0.5 0.25	
Volumétrie des données à traiter	- Volume très large - Large - Moyen - Faible	1 0.75 0.5 0.25	

Comme il est le cas pour le facteur de l'impact positif, les coefficients pondérateurs du tableau 14 permettent de prendre en considération les particularités de chaque organisation.

III.2.4.2. Quantification de l'impact positif

Les managers métier et les responsables SI, qui sont en charge des projets de qualité des données doivent :

- Lister l'ensemble des processus métier clés ;
- Configurer l'importance de chaque facteur en agissant sur le poids associé au coefficient pondérateur. La somme de tous les coefficients pondérateurs doit être égale à 100 ;
- Pour chaque facteur de la colonne 1, choisir la valeur correspondante de la colonne 2 (à chaque valeur est associée une notation de la colonne 3).

Dans le cas d'une organisation avec plusieurs processus métier clés, l'impact positif de chaque processus est calculé avec la formule de la somme pondérée ci-après :

$$\sum_{i=1}^{m}(Ri * Ii) / \sum_{i=1}^{m}(Ii) \quad (7)$$

Où R_i est la notation attribuée au facteur "i" et I_i le coefficient pondérateur qui est associé au facteur "i", comme il a été défini au préalable par les managers métier. Le score obtenu se situe entre 0 et 5, où "0" fait référence à un "impact imperceptible" et "5" fait référence à un "impact positif très élevé" (voir Tableau 11).

III.2.4.3. Quantification de la complexité d'implémentation

Pour une donnée particulière, utilisée par un processus métier clé, la complexité d'implémentation sera calculée comme suit :

$$\sum_{i=1}^{m}(Ri * Ci) / \sum_{i=1}^{m}(Ci) \quad (8)$$

Où R_i est la notation attribuée au facteur "i" et C_i le coefficient pondérateur qui est associé au facteur "i", comme il a été défini au préalable par les responsables métier et SI. Le score obtenu se situe entre 0 et 5, où "0" fait référence à une "complexité minimale" et "5" fait référence à une "complexité très élevée". La table 4 décrit la correspondance entre le score et le niveau de la complexité (se référer au Tableau 12).

III.2.5. Etape d'analyse, de comparaison et de recommandation

Après avoir parcouru l'ensemble des processus métier et calculé le score d'impact positif associé, les processus métier sont automatiquement ordonnés par priorité, dans l'objectif de déterminer le point de départ pour identifier les opportunités de bénéfices issues de l'amélioration de la qualité des données.

Comme les processus métier consomment et produisent des données, classifier les processus métier clés par leur impact positif sur les objectifs à court terme d'une organisation ainsi que sur ses résultats, doit être suivi par l'identification des options d'amélioration de la qualité des données avec la meilleure valeur financière/métier et au moindre coût.

Les processus métier accèdent aux données en modes lecture et écriture, il en résulte que la qualité de ces données a un impact sur le résultat d'exécution des processus métier et inversement.

Définir les processus clés revient ainsi à définir ceux qui participent à la réalisation des objectifs de l'organisation. Ces processus dépendent de données nécessaires à l'atteinte de ces objectifs.

La figure 15 déclinée ci-dessous, schématise les étapes principales de la démarche *PortfolioDQAF* dans son volet « données d'entreprise » :

1. Déterminer les processus clés, qui contribuent majoritairement aux objectifs et résultats de l'organisation ;
2. Déterminer la complexité de l'amélioration des objets données ;
3. Recommander le scénario optimal pour l'amélioration de la qualité des données.

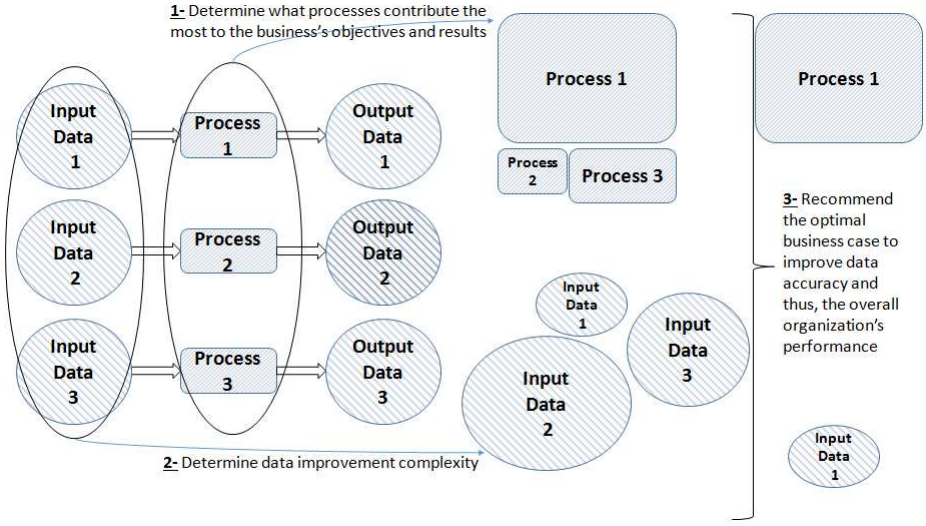


Figure 15. Etapes principales de la démarche (Belhiah et al., 2016)

La section II du chapitre III s'attarde sur le volet « données d'entreprise » de la démarche globale *PortfolioDQAF*.

Les métriques quantitatives de la démarche *PortfolioDQAF* correspondent aux facteurs d'impact positif et de complexité d'implémentation. Chaque facteur est exprimé avec un score, allant de 0 à 5, correspondant à un niveau d'impact ou de complexité.

L'objectif de la section suivante du volet « données d'entreprise » de *PortfolioDQAF* est de traduire de manière empirique, la complexité d'implémentation aux coûts de mise en place, exprimés en termes monétaires.

III.3. Modèle de coût de la qualité des données

Comme présenté au niveau du chapitre II, l'objectif de l'analyse multicritères des projets d'amélioration de la qualité des données est de sélectionner un portfolio de projets, qui produit le meilleur ratio coûts-bénéfices, sujet à plusieurs contraintes d'ordre financier et humain. Dans ce type de problèmes, il y a une incertitude quant au ROI de ces projets. La programmation non-linéaire permet d'aider à résoudre le caractère incertain de ce problème (Ragsdale, 2014).

Même si cette section ne prétend pas présenter une théorie de coût exhaustive, elle essaye d'introduire quelques éléments qui pourraient constituer le point de départ d'un modèle de coût pour la qualité des données.

Les modèles classiques de coût de la qualité sont le modèle P-A-F *(Prevention-Appraisal-Failure),* qui a été publié la première fois par 1956 (Feigenbaum, 1956) et le modèle hyponyme Juran (Juran, 1991) dont la première édition date de 1951. Dans le contexte de la qualité des données, les livres de référence en termes de l'évaluation des initiatives de la qualité des données sont (English, 1999), (Loshin, 2001) et (Redman, 1996).

Actuellement, la majorité des modèles destinés à mesurer la valeur métier de la qualité des données sont développés dans le contexte de l'industrie : (Gartner, 2011 ; Knowledge Integrity, 2011 ; Laney, 2017).

III.3.1. Définition du problème de décision

La démarche *PortfolioDQAF* qualifie l'impact positif de la qualité des données ainsi que sa complexité d'implémentation par des indicateurs quantitatifs, il est maintenant important de caractériser le coût financier traduisant la complexité d'implémentation.

Actuellement, il n'y pas d'informations disponibles au grand public qui portent sur le lien entre l'investissement en termes de coût de la qualité des données et les niveaux de qualité escomptés. Ceci rend donc cette caractérisation difficile.

Il est cependant possible, à partir de l'expérience dans le domaine de l'industrie d'émettre les hypothèses suivantes :

- La courbe de la qualité des données en fonction du coût serait convexe ;
- Le coût d'amélioration est nul si le même niveau de qualité des données est maintenu ;
- Le coût de qualité est exponentiellement élevé à l'approche d'une précision de 100%. Cependant le gradient devient plus important vers le maximum de la qualité ;
- Le gradient serait fonction de la complexité.

Ce problème mathématique a les caractéristiques suivantes : il s'agit d'un modèle prédictif et la forme de la fonction mathématique est faiblement définie.

Pour optimiser ce problème mathématique, il faut d'abord identifier les variables de décision ainsi que les contraintes. La fonction objective est construite à partir de ces mêmes variables de décision et contraintes.

III.3.3. Définition des variables de décision

Pour la sélection de projets, les managers doivent faire deux choix indépendants mais reliés :

- Parmi un portefeuille de projets, quels projets sélectionner en vue de les implémenter ? Pour modéliser cette décision, les variables de décision binaires suivantes seront utilisées : Y_i, où i= 1, 2, 3…
- En second lieu, les managers doivent déterminer le niveau de qualité des données optimal qu'il est possible de réaliser. Cette variable va être caractérisée par les variables a_i, où i= 1, 2, 3 …

III.3.4. Définition des contraintes

Dans un cas réel, cette fonction objective serait l'objet de plusieurs contraintes, parmi lesquelles :

- La somme des coûts des projets retenus ne doit pas dépasser le budget alloué au programme d'amélioration de la qualité des données :

$$\sum_i Yi * Ci \leq C \quad (9)$$

  Où :

  - C_i : est le coût associé à l'amélioration de la précision de l'objet métier i
  - C : est le coût global de l'amélioration de la précision des données.
- Les contraintes sur les ressources humaines allouées aux projets.

Ce modèle est implémenté dans l'étude de cas, objet du chapitre V.

III.3.5. Définition empirique de la fonction objective

L'objectif des managers est de minimiser les coûts associés aux projets d'amélioration de la qualité des données, tout en maximisant l'impact escompté de cette amélioration. Ceci implique que l'indice de complexité compose le numérateur et l'indice de l'impact positif fasse partie du dénominateur. La fonction objective est donc composée de :

- Le facteur de l'impact positif ;
- Le facteur de la complexité d'implémentation ;
- La précision initiale de l'objet métier ;
- La précision cible de l'objet métier.

$$\min_i \sum Yi * \frac{Ci * (ai - a_0 1)}{Ii * (1 - ai)} \quad (10)$$

Où :

- Y_i : est la variable binaire qui définit si l'objet métier i va être sélectionné ;
- Ci : fait référence à la complexité d'implémentation de l'objet métier i ;
- Ii : fait référence à l'impact positif de l'amélioration de l'objet métier i ;
- a_0i : fait référence à la précision initiale de l'objet métier i ;
- ai : fait référence à la précision cible de l'objet métier i.

III.3.6. Identification et résolution des problèmes de qualité des données

La dernière étape de l'analyse de rentabilisation consiste à comprendre les causes à l'origine des problèmes de la qualité des données et à déterminer comment elles peuvent être adressés. Typiquement, cela signifie revoir le flux d'information depuis le point de création de la donnée au niveau du processus pour déceler à quel moment l'erreur a été introduite. Une fois la source de la défaillance de la donnée est identifiée, l'analyste de données peut considérer les alternatives pour éliminer les sources des erreurs, en instituant des mesures préventives ou actions correctives. Chacune de ces alternatives aura un impact et va introduire un coût financier ou autre (organisationnel, risques, etc.) qu'il est possible de mesurer, même de manière conservatrice, avec l'analyse coûts-avantages de *PortfolioDQAF*. Les techniques de la section III.3 permettront de prioriser les actions qui ont un ratio coûts-bénéfices intéressant, sujettes aux différentes contraintes financières et humaines.

En outre, et dans l'objectif de recommander le scénario optimal pour améliorer la précision des données et ainsi la performance globale de l'organisation, le modèle prend en considération :

1. Le niveau initial de la qualité des données *(as-is)* ;
2. L'impact positif des processus clés qui utilisent les données à améliorer ;
3. La complexité d'implémentation de l'amélioration de la qualité des données.

Selon les valeurs de ces indicateurs et selon la valeur de la précision cible *(to-be)*, un ou plusieurs scénarii d'amélioration sont à envisager.

Une plateforme Web a été développée pour implémenter la démarche *PortfolioDQAF* et calculer les différentes métriques, de manière automatique. Cette plateforme est présentée au niveau de la section III.4.

III.4. Automatisation de la démarche PortfolioDQAF

« *PortfolioDQAF-tool* » a été développé pour automatiser la démarche d'évaluation de portefeuille de projets de qualité des données. Ses fonctionnalités aident à exécuter les étapes de la démarche et à automatiser les calculs des différentes métriques. Il permet également de garder l'historique des différentes instances d'évaluation pour les analyser à postériori.

III.4.1. Phases en amont du développement de « PortfolioDQAF-tool »

« *PortfolioDQAF-tool* » permet de prendre en charge l'évaluation de l'aspect précision. Cependant son architecture technique et son modèle de données ont été développés de manière à supporter sans une grande maintenance, d'autres aspects de la qualité des données, notamment : la complétude, l'actualisation et la consistance, à ne citer que ceux-là.

« PortfolioDQAF-tool » a été adapté par Maqboul & Bounabat (2017) pour évaluer la dimension de la complétude.

Les fonctionnalités actuelles de *« PortfolioDQAF-tool* » sont :

1. Paramétrer le modèle d'évaluation (facteurs, critères, pondération) ;
2. Créer la définition des processus métier ;
3. Lister l'ensemble des processus métier configurés ;
4. Ajouter de nouveaux objets métier (implémentés physiquement par des objets de données), qui sont utilisés par les processus métier préalablement enregistrés ;
5. Lister l'ensemble des objets métier enregistrés ;
6. Evaluer les projets d'amélioration de la qualité des données ;
7. Lister l'historique des évaluations.

En plus des besoins fonctionnels susmentionnés, *« PortfolioDQAF-tool* » est doté des caractéristiques non-fonctionnelles suivantes : (i) multiplateformes, (ii) globalisation, (iii) maintenabilité (Architecture MVC[17], Open Source) ; (iv) réutilisabilité par paramétrage.

III.4.1.1. Processus d'évaluation implémenté par PortfolioDQAF

La figure 16 représente le processus d'évaluation de l'impact positif et de la complexité de mise en œuvre implémenté par *« PortfolioDQAF-tool* ».

[17] *Model View Controller*

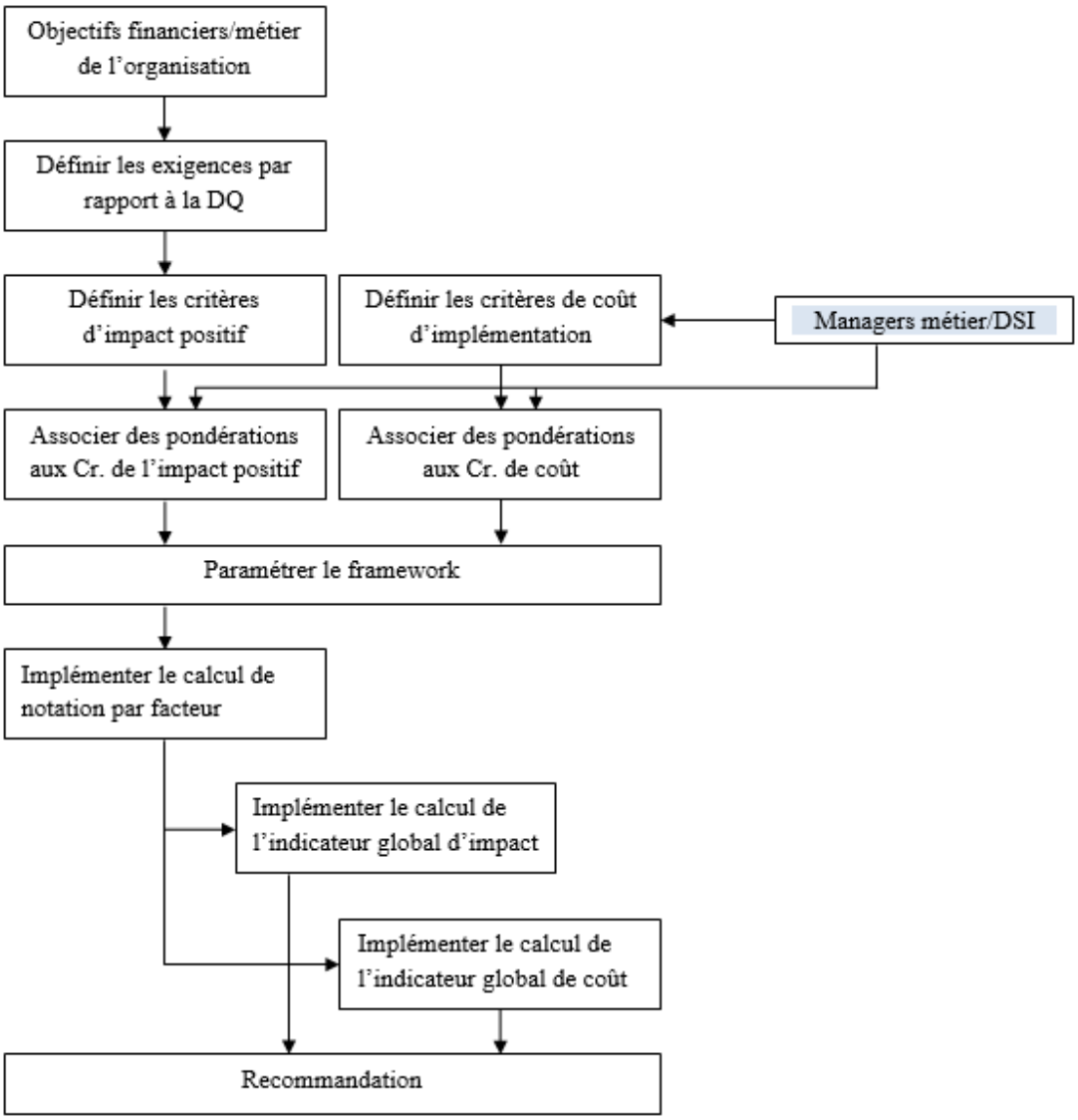


Figure 16. Processus d'évaluation implémenté par *PortfolioDQAF*

Le processus implémenté par l'outil informatique reprend les étapes de l'approche *PortfolioDQAF*.

III.4.1.2. Architecture technique

Au niveau de cette sous-section, l'architecture technique de la plateforme *« PortfolioDQAF-tool »* est présentée.

Cette architecture se décline en 3 couches : la couche Web qui assure l'interfaçage avec l'utilisateur, la couche métier qui referme les traitements métier et la couche DAO[18] qui, à travers

[18] *Data Access Object*

un connecteur, se charge des opérations de gestion des enregistrements au niveau de la base de données. Comme susmentionné, cette architecture en couche permet une maintenabilité et une extensibilité aisée. Toutes les technologies utilisées sont Open Source.

La figure 17 illustre l'architecture technique.

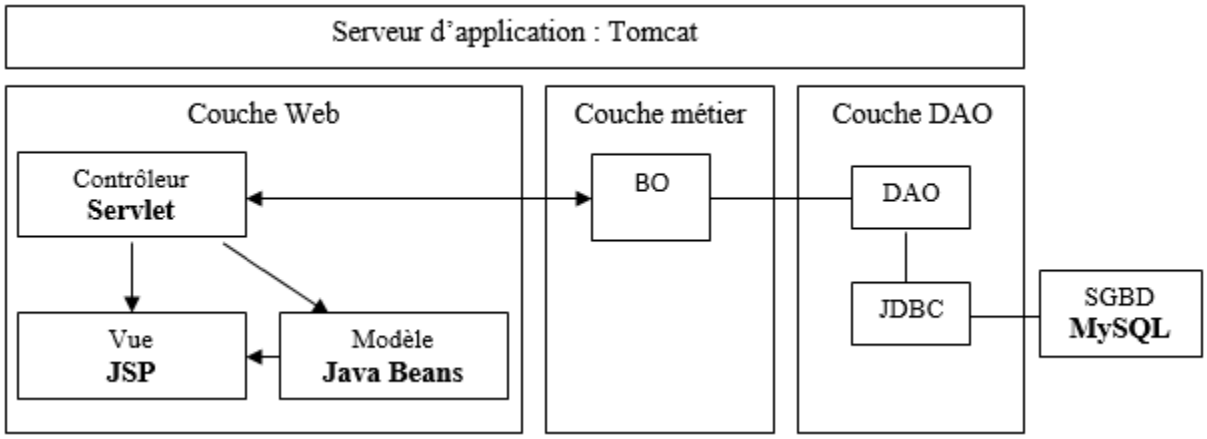


Figure 17. Architecture technique de *PortfolioDQAF*

III.4.2. Phase d'implémentation

Cette sous-section présente des prises d'écran de la plateforme *« PortfolioDQAF-tool »*.

La figure 18 correspond au menu principal de :

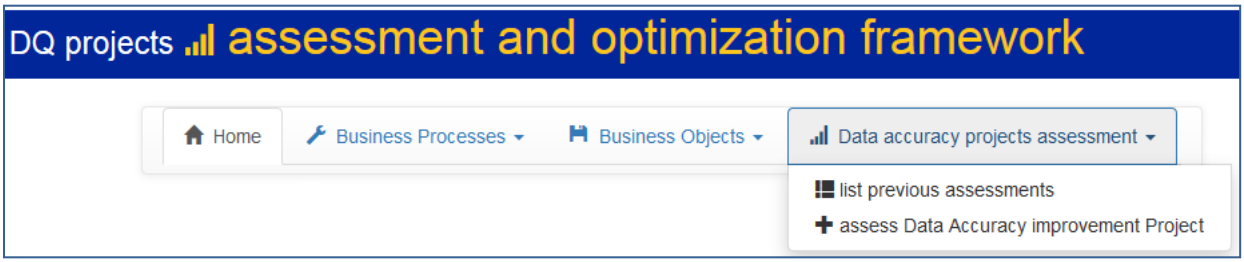


Figure 18. Menu principal

La figure 19 ci-après correspond à l'écran d'évaluation du facteur de l'impact positif par *« PortfolioDQAF-tool »*.

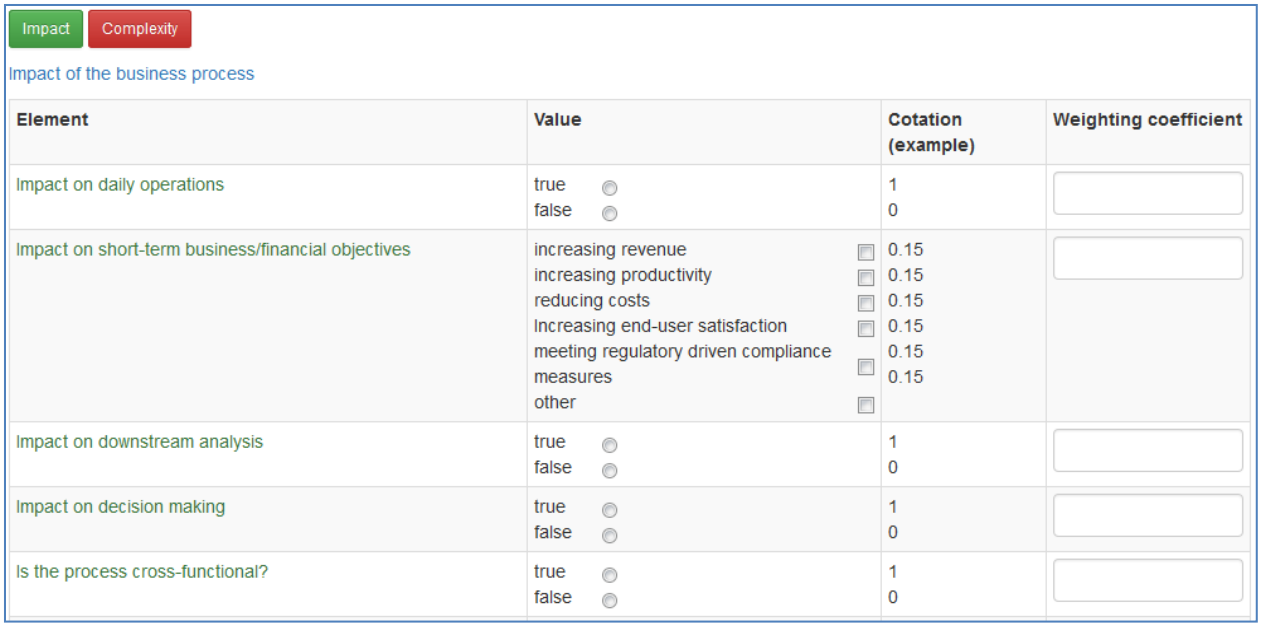


Figure 19. Evaluation du facteur d'impact positif

Enfin la figure 20 correspond à l'écran d'évaluation de la complexité d'implémentation par *« PortfolioDQAF-tool »*.

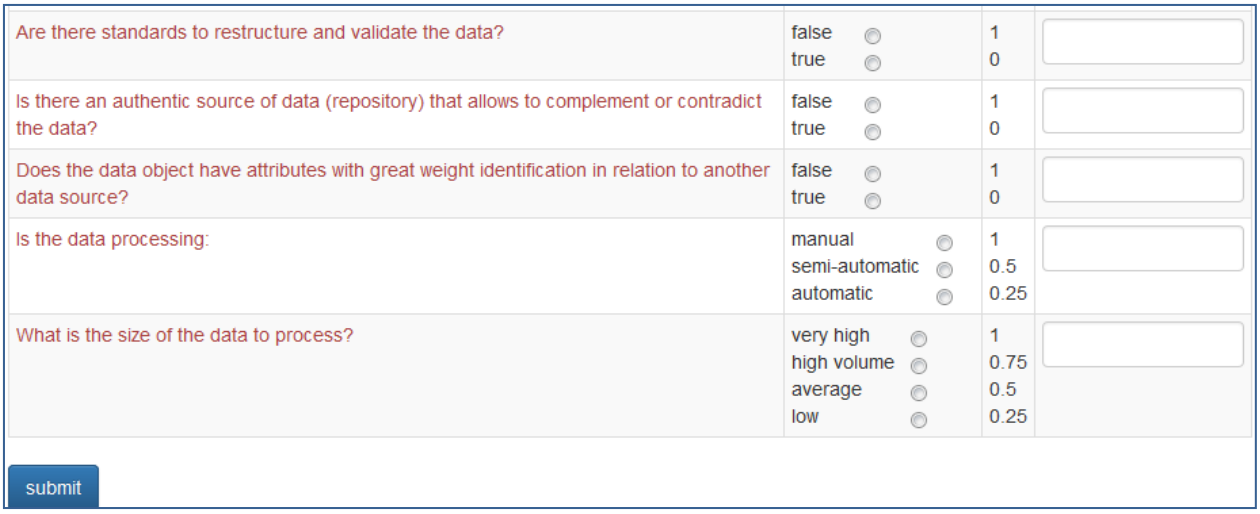


Figure 20. Evaluation de la complexité d'implémentation

Pour automatiser l'application de la présente démarche, la plateforme *« PortfolioDQAF-tool »* est utilisée dans le cadre de l'étude de cas, portant sur l'assainissement de *Data Assets* gouvernementaux, objet du chapitre V.

III.6. Conclusion

Ce chapitre présente le volet « données d'entreprise » du modèle *PortfolioDQAF*. Les facteurs d'impact positif et de complexité d'implémentation ont été décomposés en critères. Le deuxième facteur a été particulièrement développé pour la dimension « précision » de la qualité des données.

Un modèle de coût, qui prend en compte les facteurs d'impact et de complexité d'implémentation, ainsi que la précision initiale de l'objet donnée, sujet à l'amélioration est également présenté. L'outil *« PortfolioDQAF-tool »* automatisant la démarche est présentée dans la section III.4.

La chapitre IV développe le volet « Open Data » de *PortfolioDQAF*. En effet, à la différence des données d'entreprise, les données les plus critiques ainsi que l'utilisation qui en sera faite ne sont pas connues au moment de leur publication par les productions de données. Les données gouvernementales ouvertes sont naturellement destinées à être accédées et réutilisées par des acteurs externes à l'instance gouvernementale qui les produit et les publie. L'enjeu est donc important pour cerner les jeux de données ouvertes pour lesquelles l'amélioration de la qualité est opportune.

Chapitre IV : Démarche PortfolioDQAF - volet « Open Data »

IV.1. Introduction

Les projets d'ouverture de données offrent des bénéfices d'ordre monétaire et non monétaire pour différents acteurs, parmi lesquels : les gouvernements, les entreprises, les startups technologiques, la société civile et les citoyens. Ces bénéfices incluent : la conformité aux autorités de régulation, l'accroissement de la transparence et l'accroissement des opportunités pour le développement social et économique. Au Maroc, l'ouverture des données est actée par diverses instances gouvernementales (CES, 2013 ; Constitution, 2011 ; MMSP, 2018).

Un réseau pluraliste et mondial d'acteurs travaille de manière assidue, pour étendre la disponibilité des données ouvertes, à travers l'établissement des fondements légaux et l'augmentation des capacités technologiques des départements et agences gouvernementaux, et ce pour différents pays. Indépendamment du nombre de ces initiatives, elles se sont pour la plupart, concentrées sur l'évaluation de la préparation *(readiness assessment)* et de l'implémentation, par rapport aux aspects légaux et technologiques, et fournissent peu ou pas d'assistance quant aux aspects relatifs à la qualité des données.

En plus des difficultés inhérentes à leur nature, à savoir : leur volumétrie, origines diverses et ouverture à tous, les données ouvertes souffrent également de problèmes liés à leur qualité (Zaveri et al., 2016). Ces niveaux de qualité insatisfaisants peuvent miner les efforts des gouvernements d'ouvrir leurs données, et surtout ne pas être en mesure d'en tirer les profits économiques et sociaux escomptés (Chignard & Benyayer, 2015 ; Laney, 2017).

Le but de ce chapitre est de développer le volet « Open Data » de *PortfolioDQAF*. Ce volet permet d'orienter les producteurs de données ouvertes vers les datasets qui ont de la valeur ajoutée pour les usagers, et qui semblent ne pas se conformer à un niveau acceptable de qualité ; ce qui est plus optimal en termes de coût et d'impact positif.

Le plan de ce chapitre est comme suit : la section 2 présente un benchmark des plateformes existantes les plus proéminentes pour l'évaluation des données gouvernementales ouvertes. La section 3 présente le cycle de vie des données ouvertes. La section 4 décrit les étapes qui composent le volet « Open Data » de la démarche *PortfolioDQAF*. Ce chapitre se termine par une synthèse des résultats obtenus.

IV.2. Revue de littérature

Le processus d'évaluation de la disposition des départements et des agences gouvernementaux à ouvrir leurs données, ainsi que l'évaluation de l'implémentation (Bogdanović-Dinić et al., 2014 ; Ceolin et al., 2013 ; Harper, 2012 ; Reiche et al., 2013), fait l'objet d'une attention croissante dans la littérature scientifique mais également au niveau des pratiques industrielles et du conseil (The World Bank, 2018 ; Tim Burners-Lee, 2015). L'évaluation de la disposition à l'ouverture des

données s'attarde sur l'existence des préconditions nécessaires pour un projet réussi d'ouverture de données, tandis que l'évaluation de l'implémentation essaie de répondre à la question : à quel point l'implémentation actuelle est réussie par rapport à un ensemble de critères. Ces critères incluent : l'accessibilité et le format des données entre autres.

Le travail cité plus haut, établit la méthodologie globale pour évaluer la préparation à l'ouverture et l'évaluation de l'implémentation des initiatives d'ouverture de données. Néanmoins, il évalue uniquement certains aspects de la qualité des données et ceci du point de vue du producteur des données ouvertes. En effet, ce travail manque de point de vue du public général par rapport à l'aspect de la qualité des jeux de données publiées. Ceci est d'autant plus pertinent car le succès d'un projet d'ouverture de données ne se mesure pas uniquement pas la quantité des données qui sont publiées mais par leur qualité et par les applications et réutilisation qui en découlent.

IV.2.1. Concepts : Données gouvernementales ouvertes

IV.2.1.1 Données ouvertes « Open Data »

Le savoir ouvert correspond à tout contenu, information ou donnée que les usagers peuvent utiliser, réutiliser et redistribuer sans aucune restriction légale ou technologique, sans discrimination vis-à-vis de leur appartenance et indépendamment de l'utilisation qui sera faite de ce contenu.

Selon la définition de l'*Open Knowledge Foundation Network* (OKFN, 2005), l'un des organismes les plus importants pour le mouvement Open data, une donnée ouverte doit satisfaire les spécifications suivantes :

- ***Disponibilité et accès -*** doit être accessible via internet sans aucun coût, autre que le coût de reproduction et dans un format numérique, lisible par une machine ;
- ***Réutilisation et distribution -*** doit être disponible sous une licence qui en permet la réutilisation et la distribution ;
- ***Participation universelle -*** doit être libre de toute restriction liée à l'identité de l'usager (appartenance à un groupe de personnes ou à une géographie particulière) et à l'usage qui sera fait des données (usage commercial par exemple).

La *Sunlight Foundation (*2014) a de son côté défini dix critères d'ouverture, à savoir : complétude, format brute ou primaire, actualisation, accessibilité physique ou électronique, accès non-discriminatoire, utilisation de standards ouverts, format lisible par une machine, licence ouverte, permanence et gratuité.

A titre d'exemple, un document au format « pdf », même publié sur le Web, n'est pas considéré compatible avec cette définition. En effet, l'application « Acrobate » de l'éditeur « Adobe » qui permet d'y accéder est propriétaire ; une machine sur laquelle « Acrobate » n'est pas installé ne sera pas en mesure d'accéder au document.

IV.2.1.2 Données gouvernementales ouvertes

Les données gouvernementales ouvertes (OGD[19]) font référence aux données qui sont produites ou commanditées par les départements et agences gouvernementaux, et qui sont également ouvertes, selon les principes de l'Open Data.

La figure 21 montre les données publiques ouvertes au croisement des données ouvertes et de celles du secteur public.

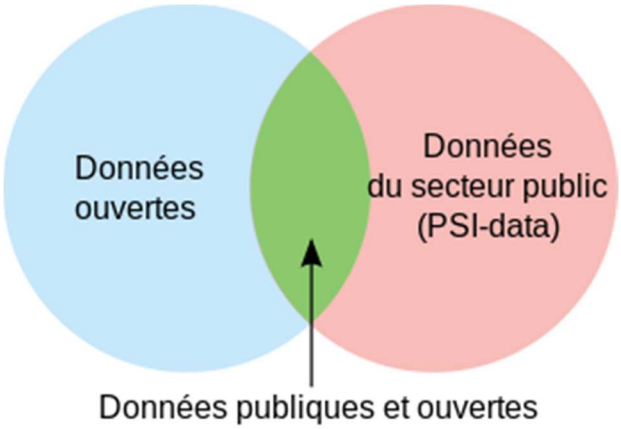


Figure 21. Données publiques ouvertes[20] (Crédit photo : Peter Krantz)

En rendant leurs données ouvertes, les institutions publiques deviennent plus transparentes et rendent des comptes aux contribuables, par rapport aux actions gouvernementales. En encourageant l'utilisation, la réutilisation et la distribution libre de ces données, ces mêmes institutions instaurent un climat d'affaires favorables à la création et à l'innovation. En effet, la publication des données ouvertes par les institutions gouvernementales devraient paver la voie aux entreprises, aux startups technologiques, aux acteurs de la société civile et aux citoyens pour bénéficier de ces données ainsi que de leurs supports technologiques, pour les réutiliser, en extraire de la valeur ajoutée et contribuer à l'enrichissement des jeux de données déjà publiés.

La sous-section suivante explore, de manière plus détaillée, les enjeux politiques, économiques et citoyens liés à l'ouverture des données gouvernementales.

IV.2.1.3 Enjeux politiques, économiques et citoyens de l'Open Data

La gouvernance ouverte et l'accès à l'information offrent des bénéfices d'ordre monétaire et non monétaire pour différents acteurs, parmi lesquels : les gouvernements, les entreprises, les startups technologiques, la société civile et les citoyens. Ces bénéfices incluent : la conformité aux

[19] *Open Government Data*
[20] https://upload.wikimedia.org/wikipedia/commons/thumb/a/ae/Oppen-data-definition_fr.svg/800px-Oppen-data-definition_fr.svg.png

autorités de régulation, l'accroissement de la transparence de l'action gouvernementale et l'accroissement des opportunités pour le développement socioéconomique.

Enjeux politiques publiques – Comme suscité, l'engouement pour les données ouvertes est porté par la volonté des institutions gouvernementales d'instaurer la transparence dans la gestion de la chose publique et la reddition des comptes par les élus et les responsables politiques.

En effet :

- L'Open Data permet d'évaluer les politiques publiques et de mesurer leur impact, en surveillant le processus démocratique depuis la prise de décision jusqu'à l'exécution et la publication des rapports associés (rapport financier, etc.) ;
- L'accès égal à l'information réduit les opportunités d'accès à des données privilégiées, uniquement par un groupe donné de personnes, ce qui permet de combattre la corruption et renforce un climat de concurrence sain ;
- En se basant sur des données précises, complètes et à jour, les décisions économiques et financières prises par l'Etat et le secteur privé sont améliorées.

Il est à noter que le *data journalism*, branche très spécialisée du journalisme d'investigation, utilise les dimensions traditionnelles du journalisme, sur le terrain particulier des données gouvernementales ouvertes pour analyser et interpréter de manière indépendante les informations officielles. Ceci a donné lieu à plusieurs investigations, parmi lesquelles : la publication des salaires des fonctionnaires gouvernementaux aux Etats-Unis par *The Texas Tribute* (2018) ainsi que les dépenses des députés de la Chambre des Commune du Royaume-Uni par *The Guardian* (2018).

Enjeux économiques – Depuis 2006, le mouvement Open data est en plein essor et sa valeur économique n'est plus à démontrer. Ci-après, quelques statistiques compilées par la Banque Mondiale (Stott, 2014) :

- Selon des enquêtes commandités par l'Union-Européenne (UE), les données ouvertes ont boosté l'activité économique de l'ordre de 40 milliards d'euros/an, ce qui représente des bénéfices directs et indirects de l'ordre de 200 milliards d'euros/an (1.7% du PIB de l'UE) ;
- Aux Etats-Unis, l'ouverture des données de la météo a permis la création de 400 compagnies, employant 4000 personnes ;
- En Espagne, le rapport annuel de 2012, qui étudie le secteur infomédiaire, a recensé au moins 150 compagnies, qui emploient 4000 personnes et qui génèrent entre 330 et 350 millions d'euros annuellement ;
- Particulièrement, les données ouvertes du transport et de la cartographie ont participé à la création de nouvelles industries comme les produits GPS et de nouveaux « *mashups* » qui combinent les données de la cartographie, transport public et trafic routier pour suggérer les itinéraires, éviter les embouteillages et émettre des alertes relatives au trafic.

« *Infonomics: How to Monetize, Manage, and Measure Information as an Asset for Competitive Advantage* » (Laney, 2017) explore toutes les facettes de la monétisation de l'Open Data.

Enjeux citoyens **–** Au-delà des aspects politiques publiques et avantages économiques, la gouvernance ouverte et l'accès aux données représentent des bénéfices immatériels pour l'ensemble de la collectivité, et ceci à travers :

- La contribution à la création d'une société du savoir ;
- La contribution à la recherche scientifique en mettant les données ouvertes ainsi que les publications scientifiques à la disposition des chercheurs ;
- La création d'applications numériques d'intérêt général par les acteurs de l'économie sociale et solidaire.

Actuellement, il n'existe pas de mesure quantitative formalisée pour mesurer l'impact, particulièrement économique, de l'Open Data. Cependant, toutes les études référencées plus haut, permettent de conclure que les bénéfices politiques, économiques et sociales de l'Open Data sont tangibles et tendraient à s'accroitre dans les années à venir.

IV.2.2. Frameworks existants pour l'évaluation des OGD

Comme susmentionné, plusieurs frameworks sont conçus dans l'objectif d'évaluer la disposition à l'ouverture des données ainsi que le succès de l'implémentation. Ces évaluations sont destinées à couvrir les projets d'ouverture des données depuis les phases initiales jusqu'à leur mise en production, en passant par la conception, le développement, la phase de recettes et le déploiement. Pour chaque framework, les projets d'ouverture de données sont évalués par rapport à un ensemble de critères.

Cette section synthétise les aspects les plus importants qui sont couverts par les frameworks les plus proéminents, dans les domaines de la recherche et de l'industrie.

Certains de ces frameworks évaluent uniquement des aspects spécifiques à l'ouverture des données, pour la plupart la disponibilité, comme « *The Five-star model of Open Data* » par l'inventeur du Web Sir. Tim Burners-Lee (2010) ou encore Socrata (2016) et Osimo (2008). D'autres s'intéressent à quelques aspects de la qualité des métadonnées ou des données comme Reiche et al. (2013) qui proposent une démarche, qui en plus de la disponibilité, permet de tacler la complétude, la complétude pondérée, la précision et l'exhaustivité *(Richness of Information)*.

Pour une revue systématique de la littérature relative aux plateformes d'évaluation des données ouvertes, il est possible de se référer aux travaux suivants (Attard et al., 2015) et (Barrau et al., 2016).

Les tableaux suivants décrivent les principes qui définissent l'ouverture par rapport aux métadonnées et au contenu (données proprement dites), pour les standards internationaux les plus utilisés. Le tableau 15 décrit ceci pour les standards internationaux.

Tableau 15. Aperçu des différents standards pour l'évaluation des données gouvernementales ouvertes

Standards	**Aspects couverts de la qualité des données**	**Qualité des données**	**Qualités de métadonnées**
opendefinition.org *(OKFN, 2005)*	- Licence ouverte - Format lisible par une machine - Gratuité - Disponibilité - Format ouvert	x	x
5-Star Deployment Scheme of Open Data *(Tim Burners-Lee, 2015)*	- Disponibilité		x
Sunlight Principles for Opening Up Government Information *(Sunlight Foundation, 2007)*	- Complétude - Format brute ou primaire - Actualisation - Accessibilité physique ou électronique - Format lisible par une machine - Licence non-discriminatoire - Utilisation de standards ouverts - Licence ouverte - Permanence - Gratuité	x	x
Data Catalog Vocabulary *(DCAT, 2014)*	- Découvrabilité *(Discoverability)* - Unicité		x
The ODI Open Data Certificate *(The ODI, 2018)*	- Disponibilité - Licence ouverte - Actualisation		x

Le tableau 16 recense les travaux de recherche pour l'évaluation des OGD. Ceci concerne l'évaluation de la préparation à l'ouverture des données et l'évaluation de l'implémentation des projets d'ouverture de données.

Tableau 16. Aperçu des différentes approches dans la recherche pour l'évaluation des OGD

Evaluation	Aspects couverts de la qualité des données	Qualité des données	Qualités de métadonnées
How Open Are Public Government Data? An Assessment of Seven Open Data Portals *(Bogdanović-Dinić et al., 2014)*	- Complétude - Actualisation - Accessibilité - Format lisible par une machine - Licence non discriminatoire - Format non propriétaire - Licence gratuite	x	x
Reliability Analyses of Open Government Data *(Ceolin et al. 2013)*	- Fiabilité		x
Grading the Government's Data Publication Practices *(Harper, 2012)*	- Réputation - Disponibilité - Découvrabilité - Format lisible par une machine		x
Implementation of Metadata Quality Metrics and Application on Public Government Data *(Reiche et al., 2013)*	- Complétude - Complétude pondérée - Précision - Richesse de l'information - Accessibilité		x

Dans le cadre de ce travail de recherche, une analyse des données qui sont publiées par le Gouvernement du Royaume du Maroc[21], via le portail national des données publiques, a été entreprise[22]. Les informations minimales qui sont requises et qui doivent figurer dans ces datasets sont : (i) les codes postaux ; (ii) les adresses ; (iii) les coordonnées géographiques. Et ceci à l'échelle nationale.

Cependant les données publiées sont :

- **Incomplètes**, car les datasets contiennent uniquement les codes postaux et les adresses : les coordonnées géographiques ont été omises. D'un autre côté, les datasets publiés couvrent uniquement les 24 plus grandes villes au Maroc : des villes de plus petite taille et des localités ont été omises ;
- **Ne sont pas à jour** : selon les spécifications du « *Global Open Data Index* », les jeux de données doivent être mis à jour au moins une fois par an. Seulement, les datasets publiés correspondent aux données datant de 2011. Hors ces données ne sont pas à jour, vu que la

[21] http://data.gov.ma/fr

[22] Selon, le classement « Global Open Data Index » de l'année 2015, le Maroc est classé parmi les meilleurs pays au monde dans la catégorie « Jeux de données géographiques ».

carte administrative au Maroc a été changée en 2015, suite à un nouveau découpage territorial ;

- **précises** : les données publiées sont précises car elles correspondent à d'autres données issues d'une autre autorité réputée, notamment Poste Maroc (Barid Al-Maghrib, 2015).

La figure 21 synthétise cette analyse :

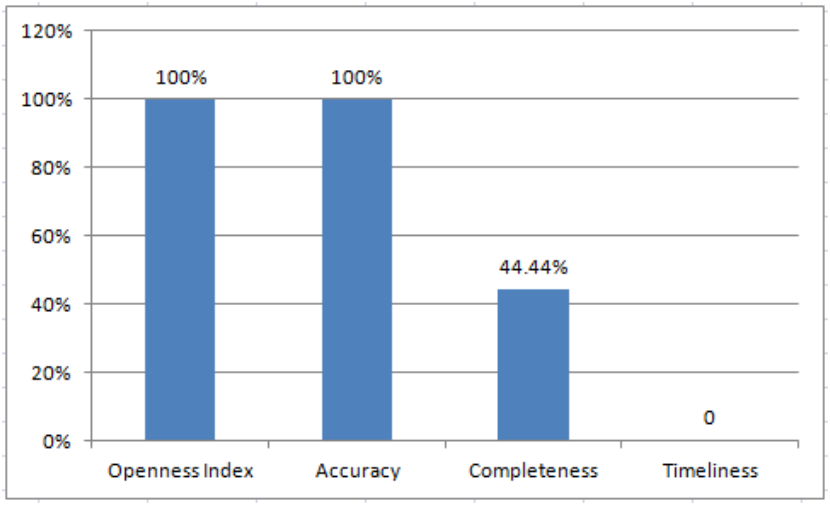


Figure 22. Comparaison entre les niveaux d'ouverture et de qualité des données

Cet exemple montre comment les indicateurs d'ouverture de données informent peu sur la qualité des données publiées : même si un datasets est ouvert à 100% selon les critères d'un certain framework ou standard, il peut manquer de précision, complétude et mise-à-jour, donc d'utilité pour les utilisateurs finaux.

IV.3. Cycle de vie des OGD

La recherche dans le domaine de la gestion de projets a démontré que 40% de la valeur ajoutée anticipée par l'implémentation des projets mis en place par les entreprises, n'est pas atteintes. Ceci est attribué en premier lieu à la mauvaise qualité des données dans les phases de planification et d'exécution (Gartner, 2011). La mauvaise qualité des données affecte également l'analyse en aval, la prise de décision et la satisfaction des usagers.

De ce fait, les producteurs de données ouvertes, doivent évaluer les différents scénarios pour l'implémentation de projets potentiels d'amélioration de leur qualité des données. Seulement, il n'y pas un commun accord sur l'ensemble de critères qui définissent le meilleur scénario avec le meilleur coût et le plus grand retour sur investissement.

Avant de développer le deuxième volet de la démarche *PortfolioDQAF*, il convient de présenter les phases communes du cycle de vie des données gouvernementales ouvertes.

Quoiqu'il existe un nombre croissant de modèles de gestion du cycle de vie des données, ces derniers ne sont pas adaptés entièrement aux données gouvernementales ouvertes ; en effet, ces modèles décrivent les étapes les plus communes de l'ouverture des données et omettent d'autres étapes importantes. Dans la mesure où la gestion de la qualité des données est au cœur de la démarche *PortfolioDQAF*, un modèle qui adapte quelques modèles existants (Auer et al., 2012) est proposé :

Ce modèle présente un processus standard que tout producteur de données ouvertes peut adopter.

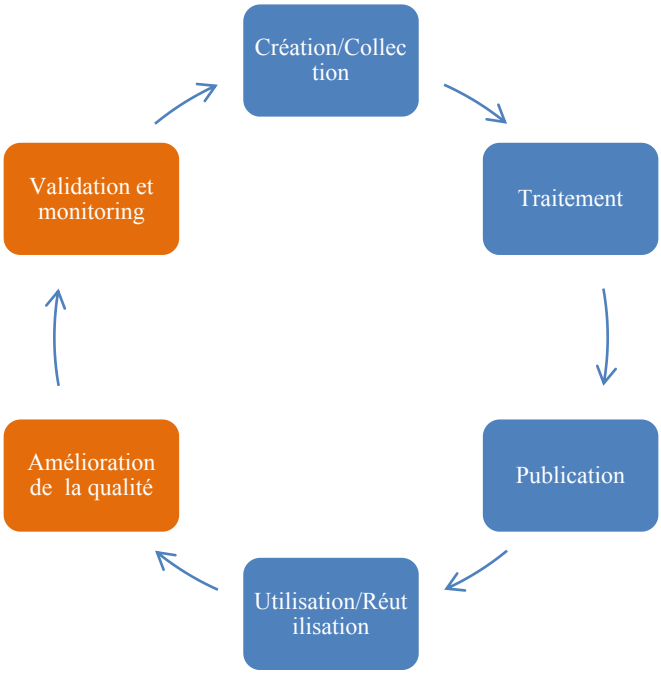


Figure 23. Cycle de vie des données gouvernementales ouvertes

- **Création/Collection des données -** Le cycle de vie des données gouvernementales ouvertes commence communément par cette phase. Les données qui sont destinées à être publiées peuvent exister au préalable dans le SI du producteur de données, dans le cadre de ses activités quotidiennes. Elles peuvent cependant être récoltées à partir de sources externes pour le but de les publier : données massives *(Big data)*, autres jeux de données, etc. ;
- **Traitement des données -** Cette phase consiste en la sélection des données à publier ; en effet, pour être publiées, les données sélectionnées doivent réussir l'évaluation de la préparation. Pendant cette phase, les données sont également harmonisées pour répondre à un format adéquat pour la publication ;
- **Publication des données -** Cette phase consiste en l'ouverture de l'accès des données depuis l'extérieur, via un portail sous forme de jeux de données, rapport technique, etc. ;
- **Utilisation/Réutilisation des données -** Cette phase correspond à la période pendant laquelle les données sont en ligne et mises à la disposition du public pour leur utilisation, réutilisation

et intégration dans des applications tierces. C'est pendant cette phase que la valeur économique et sociale est extraite ou créée depuis les données ;

- **Amélioration de la qualité des données** – Pendant leur publication, certaines anomalies relatives à la qualité des données peuvent être soulevées. De ce fait, il sera nécessaire de concevoir et d'implémenter des solutions au niveau des données et des processus, pour répondre aux spécifications de la qualité des données ;
- **Validation et monitoring du niveau de la qualité des données** – Cette dernière phase consiste en la définition du seuil d'acceptabilité de la qualité des données. L'acteur approprié : responsable DSI, CDO (*Chief Data Officer*) ou une personne désignée sera notifiée dans le cas de l'échec de se conformer aux spécifications relatives au niveau de la qualité des données. Les actions appropriées doivent être alors mises en place pour remédier à la mauvaise qualité des données.

Les autres phases qui ne sont pas couvertes par ce modèle comme l'interconnexion des données *(data interlinking)*, sont à l'extérieur du périmètre de ce travail de recherche.

Le volet « Open Data » complète l'ossature générique de la démarche *PortfolioDQAF*. Ceci est dû à la nature même des données ouvertes qui se caractérisent par :

- La volumétrie conséquente des datasets, d'où l'impossibilité d'améliorer la qualité de tous les datasets ;
- L'usage future des données qui n'est pas connue a priori ;
- Les jeux de données qui représentent de l'intérêt pour les usagers ne sont pas connus a priori ;
- Les objectifs des producteurs derrière l'ouverture des données n'est pas d'ordre financier ;
- L'impact positif s'apparente à la popularité des datasets.

Si le premier volet de la démarche est orienté organisation, dans le mesure où ce sont les managers métier qui définissent les données critiques dont dépendent leur processus métier clés, et ainsi la réalisation de leurs objectifs financiers/métier, le deuxième volet est orienté usager. En effet, l'impact positif s'apparente à l'évaluation de la popularité des datasets. Egalement, les données critiques ayant une mauvaise qualité sont déterminées par les usagers des données ouvertes.

IV.4. Démarche, orientée usager pour l'évaluation et l'amélioration de l'OGD

IV.4.1. Etape d'identification des exigences par rapport à la DQ

Après l'examen des standards existants et des approches académiques pour l'évaluation des données ouvertes, le volet « Open Data » de *PortfolioDQAF* pour l'évaluation et l'amélioration de la qualité des données ouvertes est développé.

Les recherches en systèmes d'information et gestion des bases de données fournissement un cadre général de recherche en qualité des données (Batini & Cappiello, 2009 ; Scannapieco & Catarci, 2002), qui est applicable aux données ouvertes. Comme mentionné dans le chapitre II, il est possible d'évaluer la qualité des données par rapport à différentes dimensions.

Même s'il est difficile de s'accorder sur les dimensions qui vont définir les niveaux de qualité des données, il est cependant possible, si le point de vue des utilisateurs est pris en considération (Belhiah et al., 2015a ; Frank & Walker, 2016 ; Zaveri et al., 2013), de se confiner à un nombre restreint de dimensions. Le modèle *PortfolioDQAF* évalue la qualité des données à travers les dimensions suivantes : (i) la précision ; (ii) la complétude ; (iii) l'actualisation. Ces dimensions ont été arrêtées après avoir effectué une évaluation subjective auprès d'usagers (citoyens, développeurs et intégrateurs), qui ont utilisé des jeux de données ouverts dans le cadre d'un *hackaton*[23] qui a été organisé à Rabat, Maroc, et dont l'objectif a été de promouvoir l'utilisation des données ouvertes et le développement d'une communauté autour de ces données. Un questionnaire a été également administré à une communauté sur le Web qui évolue autour de la thématique de l'Open data. Ces dimensions sont problématiques pour les utilisateurs finaux. Un feedback objectif d'un producteur de données ouvertes a permis de corroborer ces résultats.

Comme à travers le sondage, il a été difficile de demander aux utilisateurs de pointer les problèmes de qualité des données directement en termes de dimensions et de métriques, un ensemble de questions a été posé ; les réponses ont été associées aux dimensions correspondantes de la qualité des données.

Le tableau 17 ci-dessous liste les difficultés qui ont été soulevées et les relie aux dimensions correspondantes de la qualité des données.

Tableau 17. Problèmes relatifs à la qualité des données

Problème	Dimension de la qualité des données
- Y a-t-il des données dupliquées ? - Les données sont-elles normalisées, de manière à permettre une validation syntaxique ?	***Précision -*** Les données ont-elles une marge acceptable d'erreur ?
- Y a-t-il des données manquantes ? - Les cellules sont-elles complètes ? - Les lignes sont-elles complètes ?	***Complétude -*** Les données sont-elles complètes ?
- Est-ce que la date de publication correspond à la date figurant dans la documentation ? - Y a-t-il des délais dans la publication ? - Est-ce que les données sont obsolètes ? - Est-ce que les données sont rapidement mises-à-jour (mise en ligne aussitôt qu'elles sont disponibles) ?	***Actualisation -*** Les données sont-elles à jour ?

[23] http://www.diplohack.org/diplohack-maroc.html

L'approche permet ensuite aux usagers de noter (évaluer) la qualité des jeux de données auxquels ils ont accédés en termes de : précision, complétude et actualisation. En effet, le feedback des usagers permet de :

- Recommander les datasets qui ont le meilleur score en termes de qualité des données ;
- Fournir un feedback sur les jeux de données les plus utilisés et un aperçu de leurs problèmes de qualité. Ce feedback va permettre aux producteurs de données ouvertes de planifier les projets d'assainissement des données avec le meilleur rapport coûts-bénéfices.

Il est supposé néanmoins que ces jeux de données sont accessibles et disponibles selon la définition de base, des données ouvertes (*OKFN, 2005*) Le schéma de la figure 24 décrit les itérations du volet « Open Data » de la démarche *PortfolioDQAF*.

Il est également important de garder à l'esprit que cette approche est basée sur un nombre réduit d'aspects de la qualité des données, qui ont été relevés suite aux entretiens avec un échantillon d'usagers ; ces aspects étaient bien évidemment les plus représentatifs pour l'échantillon considéré.

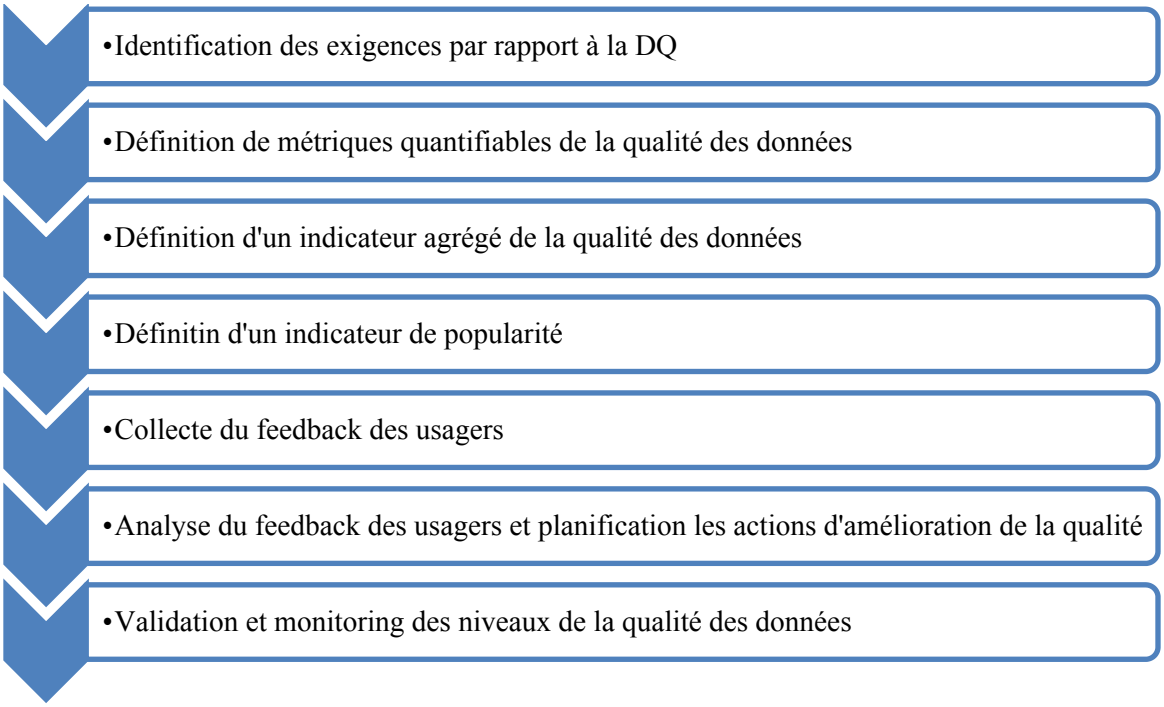


Figure 24. Approche *PortfolioDQAF* - volet « Open Data »

IV.4.2. Etape d'identification des critères d'évaluation

Quand les usagers accèdent aux portails de données gouvernementales ouvertes, ils sont confrontés à une myriade de possibilités et à un large choix de datasets. A titre d'exemple, le portail des

données ouvertes du gouvernement du Royaume-Unis[24] regroupe, à la date d'écriture de ces lignes, 1.410 producteurs pour 43.111 jeux de données. Comme la présente démarche est orientée usager, elle permet à ce dernier, de filtrer l'information pour trouver celle qui correspondrait le mieux à ses besoins, en termes d'utilité et de qualité.

Dans cette optique, les systèmes de recommandation *(recommender systems)* sont l'une des méthodes d'intelligence artificielle qui a été largement utilisée, pour le filtrage des données massives *(Big data)*. Cette technologie a été appliquée avec succès aux différents contextes du commerce électronique, pour suggérer les meilleures options quant à l'achat de livres, visualisation de contenu et visite de lieux, transformant ainsi comment les gens magasinent (exemples : Ebay.com, Amazon.com, Netflix.com, Tripadvisor.com, etc.). Ceci rend cette technique possiblement applicable aux domaines des données gouvernementales ouvertes. Les systèmes de recommandation suggèrent donc les items qui semblent représenter de l'intérêt pour l'usager (Burke, 2010).

Les systèmes de recommandation couvrent un large spectre de techniques, parmi lesquels : le filtrage collaboratif, démographique, la recommandation basée sur le contenu, basée sur la connaissance ainsi que les techniques hybrides, issues de la composition de plusieurs technologies de recommandation.

Dans ce qui suit un bref aperçu de ces techniques :

- **Filtrage collaboratif -** Il s'agit de la technique la plus familière, la plus implémentée et la plus mûre des techniques précitées. Le système de recommandation collaboratif agrège les notes attribuées aux items, décèle les points communs entre les utilisateurs sur la base de la notation qu'ils ont attribuée et génère en dernier lieu, de nouvelles recommandations basées sur la comparaison inter utilisateurs. Le point fort des techniques collaboratives est qu'elles sont totalement indépendantes de la représentation des objets à recommander. En effet, elles affichent de bonnes performances indépendamment de la complexité des items, comme il est le cas pour les sites de streaming de contenu (iTunes, spotify) ;
- **Système de recommandation démographique -** Ces systèmes cherchent à classifier les utilisateurs à partir d'attributs personnels, et génèrent des recommandations sur la base de classes démographiques. Ils ressemblent au filtrage collaboratif dans la mesure où ils procèdent à la comparaison inter utilisateurs ; seulement, en utilisant des entrées différentes. L'avantage de l'approche démographique est qu'elle ne requiert pas l'historique de la notation des utilisateurs comme il est le cas pour l'approche collaborative ;
- **Recommandation basée sur le contenu -** Dans un système basé sur le contenu, des caractéristiques sont associées aux objets à recommander. En effet, le système construit un

[24] https://data.gov.uk/

profil de l'utilisateur, dont la composante principale est les caractéristiques qui l'intéressent, sur la base de sa notation d'objets utilisés ou visionnés par le passé ;

- **Recommandation basée sur l'utilité -** Contrairement aux techniques précitées, la recommandation basée sur l'utilité ne cherche pas à générer un profilage des utilisateurs, mais construit ses suggestions sur la base des besoins du client et des options disponibles. Sa particularité est qu'elle permet de tenir compte lors du calcul de la fonction d'utilité, d'attributs non fonctionnels du produit, comme la fiabilité du fournisseur et la disponibilité du produit, attribut révélateur pour un client pressé ;
- **Recommandation basée sur la connaissance -** Cette technique essaye d'inférer les besoins et les préférences des utilisateurs. Sa particularité est qu'elle construit une connaissance fonctionnelle sur comment un item donné répond aux besoins d'un utilisateur en particulier, et résonne dorénavant sur la relation entre le besoin du client et une éventuelle recommandation. Le profil de l'utilisateur peut être de n'importe structure de connaissance qui supporte ces inférences.

Pour la recommandation de jeux de données ouverts, les algorithmes les plus adaptés seraient les moins invasifs, par rapport aux données personnelles des usagers. En effet, à l'exception des données de navigation, les producteurs de données ont peu ou pas de connaissance sur l'utilisateur final. Les techniques à envisager et à privilégier sont celles liées au contenu, soit la représentation des attributs fonctionnels et non fonctionnels des jeux de données.

Au niveau de ce volet, deux mesures sont associées à chaque jeu de données : la popularité et l'indicateur de qualité.

IV.4.2.1 Indicateur de popularité

Pour chaque jeu de données, l'indicateur de popularité est défini, en utilisant une ou plusieurs combinaisons des mesures suivantes :

- Nombre de vues ;
- Nombres de téléchargements ;
- Nombres de réutilisation ;
- Indicateur de qualité des données (voir Etape 3).

Plusieurs portails de données ouvertes disposent d'une ou plusieurs mesures parmi les 3 premières, à savoir : le nombre de vues, le nombre de téléchargements et le nombre de réutilisation. Seulement, ces portails n'utilisent pas ces statistiques de manière sophistiquée, pour recommander les jeux de données.

Le tableau 18 dresse un portrait des métriques de popularité qui sont déjà implémentées par plusieurs fournisseurs de données à l'échelle internationale et qui sont en même temps les leaders en matière d'ouverture de données.

Tableau 18. Indicateurs de popularité comme intégrés par 5 portails de données gouvernementales

Pays	URL	*Global Open Data Index*[25] (2016)	Métriques de popularité
Taiwan	data.gov.tw	1	- Nombre de vues - Nombre de téléchargements
Royaume-Unis	data.gov.uk	2	- Nombre de vues
Danemark	opendata.dk	3	- Néant
Finlande	data.gouv.fi	5	- Nombre de vues
France	data.gouv.fr	10	- Nombre de réutilisation

Comme ces portails de données ont par défaut accès à ces métriques, introduire l'indicateur de popularité nécessiterait peu d'ajustements. Accessoirement, il est possible d'incorporer des données contextuelles au niveau des recommandations, en utilisant la localisation et l'historique de recherche par exemple.

Il est important de souligner que ces techniques de filtrage préservent les données de l'usager, qui sont à caractère personnel, comme leur exécution ne requiert aucune information privée de l'utilisateur, autre que l'historique de navigation.

IV.4.2.2 Indicateur de qualité

Durant la phase de lancement *(cold start)*, l'indicateur de qualité ne peut encore être calculé et n'est donc pas encore connu. Il l'est, une fois le feedback est collecté à travers les notations des usagers.

Les deux métriques susmentionnées, seront rattachées aux jeux de données et affichées sous forme d'annotations. L'indicateur de popularité qui est fonction de l'indicateur de qualité, va contrôler le listing et l'affichage des jeux de données.

IV.4.3. Etape de quantification des facteurs d'évaluation

La qualité des données a été définie par Wang (1996), de façon mémorable, par l'expression : *« fitness for use* », qu'il est possible de traduire par l'aptitude à l'emploi. Par ailleurs, la précision a été définie par (Pipino & al., 2002) comme étant « la proximité des résultats des observations des vraies valeurs ou des valeurs acceptées comme étant vraies ». Wang et al. (1996) définit la précision comme « la mesure dans laquelle la donnée est correcte, fiable et certifiée ».

[25] https://index.okfn.org/place/

La complétude spécifie comment « les données ne sont pas manquantes et sont suffisantes pour les besoins de la tâche présente » (Batini & Scannapieco, 2006). Comme la complétude s'intéresse toujours au sens des valeurs nulles, elle peut être exprimée comme « le rapport entre le nombre de valeurs non-nulles dans une source de données et la taille de la relation universelle » (Naumann, 2002).

L'actualisation (aspect temporel de la qualité), exprime quant à elle « à quel point la donnée est à jour pour les besoins de la tâche à accomplir » (Batini & Scannapieca, 2006). Dans le contexte des données ouvertes, elle peut être définie par la comparaison entre la date réelle de publication et la date documentée. Il est évident que même si une donnée est précise et complète, elle peut être non utilisable s'elle est obsolète.

Pour rendre opérationnelle la qualité des données, il faut être en mesure de l'évaluer de manière tangible ; en effet, quand un producteur de données ouvertes effectue une évaluation objective, il doit développer des métriques qui sont spécifiques et adaptées à son besoin, car l'évaluation de la qualité des données n'est pas indépendante du contexte dans lequel évolue l'organisation et surtout de l'utilisation qui en sera faite.

Il est à noter que pour l'ensemble de ces aspects de la qualité des données, l'objectif n'est pas d'avoir une qualité impeccable indépendamment du coût, mais d'équilibrer la qualité par rapport au coût pour le producteur de données ; les données ouvertes doivent être assez bonnes pour l'usage auquel elles sont destinées. Pour l'actualisation à titre d'exemple, au niveau des données en outputs, la fréquence des mises-à-jour doit être adaptée à la nature de la donnée et aux applications à lesquels elles sont destinées.

Pour les deux premières dimensions susmentionnées, à savoir : la précision et la complétude, une méthode de calcul par ratio est utilisée pour leur associer des mesures quantitatives. Cependant, pour la métrique de l'actualisation, une variable binaire (0 ou 1) est utilisée, 1 étant : le jeu de données est mis-à-jour selon la documentation associée.

Les indicateurs de mesure de la qualité des données sont présentés dans le tableau 19.

Tableau 19. Indicateurs de mesure de la qualité des données

Dimension	Nature de la dimension	Forme fonctionnelle	Nature du score	Intervalle du score
Précision	- Intrinsèque	$\frac{\text{Nombre de valeurs précises}}{\text{Nombre total de valeurs}}$	Pourcentage	[0,1]
Complétude	- Dépendante du jeu de données - Dépendante du domaine	$\frac{\text{Nombre de valeurs non} - \text{nulles}}{\text{Nombre total de valeurs}}$	Pourcentage	[0,1]
Actualisation	- Dépendante du jeu de données - Dépendante du domaine	oui/non	Binaire	(0,1)

Ce tableau fournit également une brève description de la structure des dimensions de la qualité, notations et forme fonctionnelle. Il est important de noter que ces mesures sont définies pour les valeurs des jeux de données et non pas pour les métadonnées ou schémas de données.

Comme susmentionné, il ne s'agit pas de valeurs absolues pour la qualité des données, mais de l'évaluation des niveaux de qualité requis par les utilisateurs finaux selon leurs intérêts au moment d'accéder aux données et les applications qu'ils vont construire à partir de ces dernières.

A la lumière de ces résultats, une évaluation plus objective peut être faite par le producteur de données ouvertes pour corroborer les niveaux de qualité comme exprimées par les utilisateurs finaux.

IV.4.3.1 Définition d'un indicateur agrégé « DQ index »

Vu les particularités de chaque domaine, et pour fournir une approche générique qui peut être implémentée sans ajustements majeurs, l'étape suivante de l'approche consiste à définir un indicateur agrégé de qualité des données, désigné par « DQ index » qui est composé des mesures définies dans l'étape précédente. A chaque indicateur de mesure, correspondrait un coefficient pondérateur qui ne cause pas de biais dans l'interprétation.

L'idée derrière l'indicateur unique « DQ index » est d'agréger les « n » critères afin de les réduire en un critère unique en utilisant la méthode de la somme pondérée.

« DQ index » doit représenter aussi précisément que possible et que nécessaire le phénomène qu'il mesure ; Dans un contexte multi varié, le producteur de données doit avoir une très bonne compréhension de l'importance de chaque dimension, respectivement : la précision, la complétude et la ponctualité et ce, dépendamment de son domaine d'activité. Le coefficient pondérateur correspond à la contribution de chaque mesure au calcul de l'indicateur agrégé.

$$\text{DQ index} = \sum_{i=1}^{3}(Ri * wi)/100) \quad \text{sujet à : } R_i > \text{seuil} \tag{11}$$

Où R_i est la notation pour la dimension « i » de la qualité des données, et w_i est le coefficient pondérateur, comme il a été défini par le producteur de l'Open Data. La notation finale (score) obtenue se situe entre 0 et 5, où « 0 » fait référence à « Aucune qualité » et « 5 » fait référence à une « Qualité élevée ».

Pour l'actualisation, il est cependant suggéré, d'introduire en plus du coefficient pondérateur, un paramètre de sensibilité et qui prend en considération la vitesse avec laquelle la donnée devient obsolète (Pipino et al., 2002).

Cependant, un indicateur demeure une représentation simplifiée de la réalité et un outil simple et accessible à tous. S'il permet d'avoir un aperçu rapide du niveau de la qualité globale, ceci ne

dispense par le producteur de données d'analyser chacun de ses composants régulièrement, à savoir : la précision, la complétude et l'actualisation, en corrélation avec la nature de la donnée et l'usage qui en est fait, pour identifier précisément quelle dimension est la plus pertinente pour l'utilisateur et laquelle est problématique. A la lumière de cette analyse, il serait probablement amené à revoir les pondérations attribuées à ses mesures de la qualité.

IV.4.3.2 Collecte du feedback des usagers

Le volet « Open Data » du modèle *PortfolioDQAF* aide à fournir une compréhension de la demande, côté usager, des données gouvernementales ouvertes, de manière à ce que les efforts d'amélioration de la qualité des données soient alignés avec ses besoins.

Les utilisateurs expriment leurs avis de leurs expériences pendant qu'ils accèdent aux portails de données. Les notations (scores) sont exprimées sur une échelle de 1 à 5, « 5 » étant le meilleur score. Ceci peut être implémenté en utilisant un système de classement par étoiles :

Tableau 20. Niveau de qualité des données

Score	Niveau de qualité
★★★★★	Très grande qualité
★★★★	Grande qualité
★★★	Qualité moyenne
★★	Qualité faible
★	Qualité inexistante

Chaque aspect de la qualité des données, à savoir : la précision, la complétude et l'actualisation doit être noté de manière séparée. Le cas échéant, dans une étape ultérieure, l'indice agrégé de qualité des données, sera calculé et associé aux jeux de données correspondants.

Au niveau de l'algorithme de classement, les anciens avis (critique, notation) peuvent se voir attribués un poids faible, surtout pour les jeux de données qui sont sujets à des contraintes de temps (données transactionnelles v/s master data). Le nombre de revues est également un facteur à considérer : en effet, un jeu de données ayant un nombre élevé de revues devraient avoir un score différent d'un autre dataset ayant la même notation mais avec un moindre nombre de revues.

Il en résulte que l'indice global de qualité est une sommation pondérée de notations, basé sur la qualité, la quantité et la récence des scores.

IV.4.4. Etape d'analyse, de comparaison et de recommandation

Les écarts entre l'état « *as-is* » et l'état « *to-be* » peuvent être identifiés à cette étape ; en effet, après avoir collecté la fréquence d'accès aux jeux de données mis en ligne (utilité pour les usagers) et les indicateurs de qualité comme perçue par les usagers, les jeux de données peuvent être classés par priorité et placés dans des catégories, pour planifier les actions d'amélioration de la qualité des données.

En plus des paramètres de sélection définis par l'utilité et le niveau de qualité des données, une évaluation plus élaborée des projets d'amélioration de la qualité à implémenter peut inclure l'effort estimé pour adresser ces écarts.

En effet, les investissements en termes de temps, ressources humaines et financières seront dirigés vers l'assainissement des données qui sont le plus utilisées par les utilisateurs finaux et qui ont des problèmes de qualité. De ce fait, les options d'amélioration de la qualité avec la meilleure valeur métier et aux meilleurs coûts sont repérées.

Il en résulte une carte de route d'implémentation, sur plusieurs étapes en commençant par une itération du projet avec un retour sur investissement élevé. Les prochaines itérations correspondraient aux données dont l'amélioration de la qualité est plus coûteuse et qui ont un intérêt moindre pour l'utilisateur final.

Du fait que les processus métier accèdent en mode lecture/écriture aux données, il est normal que la qualité des données ait un impact sur les résultats d'exécution des processus métier et vice-versa. Il en résulte deux scenarios à considérer :

- Le premier scénario est basé sur l'amélioration de la qualité des données, en déterminant et en analysant les sources de la mauvaise qualité, comme l'acquisition incontrôlée de données, les problèmes de mise-à-jour, etc. Il est nécessaire ensuite d'éliminer les sources identifiées des problèmes de qualité ;
- Le deuxième scénario est basé sur l'amélioration des processus (réingénierie des processus, introduction de points de contrôle, etc.), en améliorant la précision d'exécution.

Un input nécessaire aux deux scénarios mentionnés plus haut est la compréhension des sources d'erreur de données. Les recommandations qui peuvent être émises sont présentées dans le tableau 21.

Tableau 21. Correspondance des problèmes de qualité des données et des recommandations

Problèmes	Recommandations
Précision - Les données ont-elles une marge acceptable d'erreurs ?	- Valider les données de manière syntaxique - Confronter les données à une autre source authentique de données, dans le SI du producteur de données ou à partir d'une source externe
Complétude - Les données sont-elles complètes ?	- Atteindre la couverture souhaitée par l'utilisateur final - Atteindre la granularité souhaitée par l'utilisateur finale
Actualisation - Les données sont-elles à jour ?	- Publier les données aussitôt qu'elles sont disponibles - Documenter les délais de publication - Concevoir et implémenter les processus de mise-à-jour

La composante la plus importante des mesures de la qualité des données est la capacité à collecter les statistiques relatives à cette dernière, de les rapporter dans un format facile à interpréter et qui permet de prendre des actions et fournir un historique de l'évolution à travers le temps.

Le producteur de données doit également être en mesure de lier l'impact de l'amélioration de la qualité à ses objectifs financiers et métier (voir Chapitre II). Dans le cas des producteurs des OGD, ces objectifs seraient : (i) améliorer la gouvernance ; (ii) améliorer la transparence ; (iii) extraire de la valeur ajoutée à partir de données (nombre de réutilisation qui en ont été faites par exemple).

Le schéma 24 illustre la planification des itérations d'assainissement des données

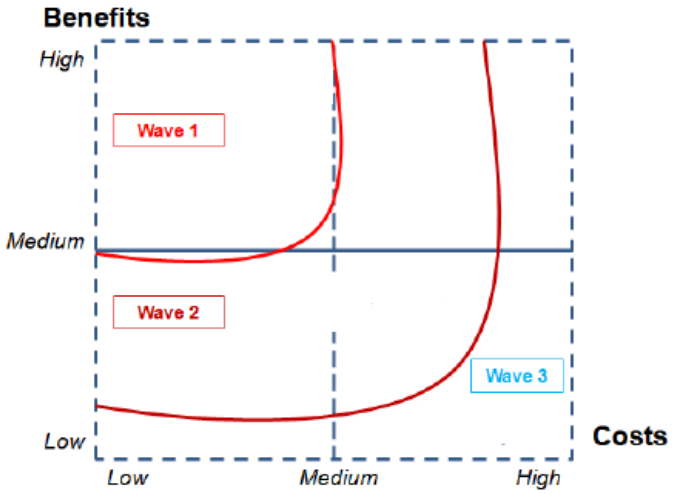


Figure 25. Planification des itérations d'assainissement des données

IV.4.5. Etape de validation et monitoring des niveaux de la qualité des données

Cette étape consiste en la définition des seuils d'acceptabilité du niveau de la qualité des données. Les niveaux de qualité pour la précision, la complétude et l'actualisation des jeux de données publiés, doivent correspondre aux attentes des utilisateurs finaux. Des tableaux de bord

automatisés permettront de notifier les personnes adéquates quand les seuils d'acceptabilité ne sont pas respectés. La figure 26 synthétise la validation et le monitoring des niveaux de la qualité des données en automatisant le calcul et le reporting.

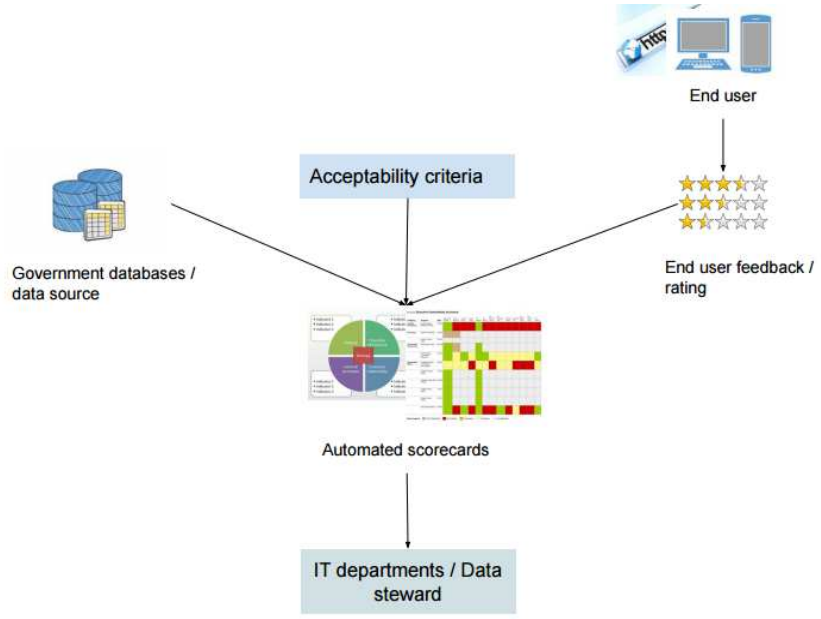


Figure 26. Informatisation des reportings des niveaux de la qualité des données

La figure suivante illustre l'architecture du modèle :

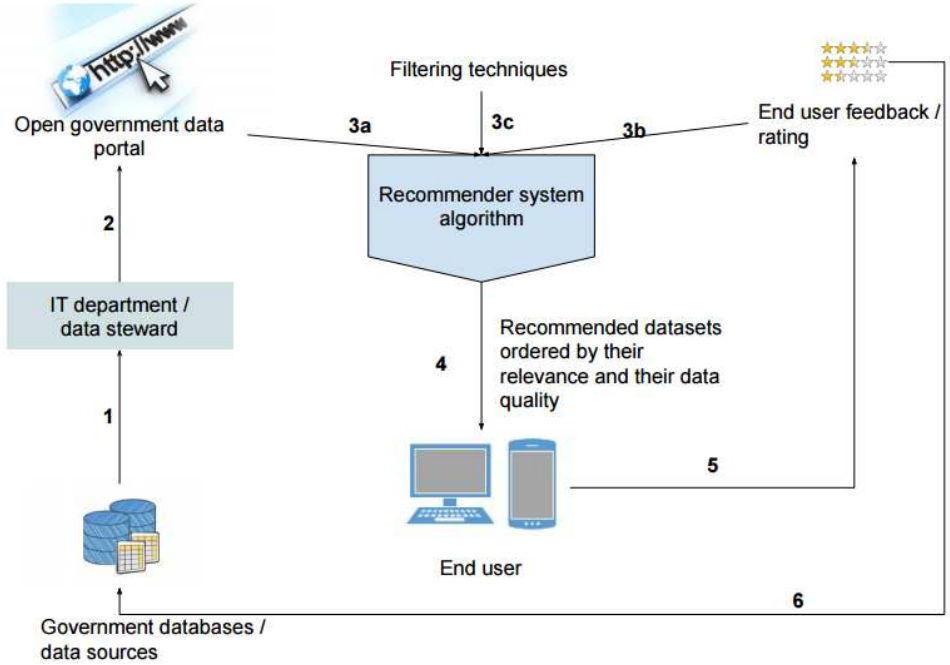


Figure 27. Architecture du volet « Open Data » de *PortfolioDQAF* (Belhiah & Bounabat, 2017)

IV.5. Conclusion

Comme les départements et les agences gouvernementaux commencent à reconnaître l'information comme étant parmi les actifs les plus précieux et que la mauvaise qualité des données a un impact significatif sur leurs plans de transformation digitale, la demande pour les frameworks d'évaluation de la préparation à publier les données ouvertes et d'évaluation de l'implémentation devient mature. La gestion de la qualité des données ajoute un autre niveau de complexité à une activité qui est déjà très exigeante, dans un contexte d'entreprise (contexte fermée et données propriétaires). En effet, au moment de publier de vastes volumes de données, les organismes publics n'ont aucune connaissance sur la manière avec laquelle ces informations vont être utilisées et pour quelle intention, quelle est la donnée la plus utile pour les usagers et quel est le niveau de qualité requis pour profiter complètement de ses bénéfices ?

Les gouvernements collectent, assainissent, transforment et publient périodiquement des giga-octets de données brutes, les projets d'assainissement de données doivent être planifiés avec parcimonie. Il en résulte que ces projets doivent avoir un coût raisonnable et la meilleure contribution et utilité aux usagers. Cette instruction est particulièrement cruciale dans le contexte des données gouvernementales ouvertes où il y a peu ou pas d'informations sur quels sets de donnée compte le plus pour les usagers. La démarche met en avant les actions d'assainissement de données les plus rentables et les plus utiles.

Le volet « Open Data » de *PortfolioDQAF* établit un indicateur global de la qualité des datasets ouverts. Cet indicateur agrège les niveaux de qualité de la précision, complétude et l'actualisation, comme exprimés par les utilisateurs finaux et collectés au niveau de l'entité responsable de la production des données. Lorsque cet indicateur est adopté comme annotation et est intégré aux portails de données gouvernementales, il permet aux utilisateurs d'identifier les données fiables. Egalement, les problèmes liés à la qualité des données qui peuvent émerger sont transférés en amont des utilisateurs finaux vers les producteurs de données gouvernementales. Lorsque ces données sont couplées avec les données relatives à la fréquence d'usage des données (nombre de vues, nombre de téléchargements, nombre de réutilisations, etc.), elles permettent aux producteurs de données d'implémenter efficacement les projets de qualité des données.

La dernière étape de la démarche recommande de mettre les bonnes pratiques en place pour définir les seuils d'acceptabilité pour les niveaux de qualité des données et de développer des tableaux de bord automatisés pour surveiller le niveau de qualité des données. Ainsi, les actions d'évaluation et d'amélioration de la qualité des données sont dictées par les besoins exprimés par les utilisateurs et deviennent une pratique continue, dans l'optique de délivrer des données de qualité et donner confiance aux instances publiques par rapport à la valeur économique et sociale de ces données.

Chapitre V : Etude de cas - Application à l'assainissement de *Data Assets* gouvernementaux

V.1. Introduction

Les managers métier et les analystes SI peinent à justifier de manière viable, les coûts associés à la qualité des données au niveau de leurs organisations. Ceci, Malgré une pléthore d'études sur la criticité de la qualité des données dans la réussite de projets tels que les entrepôts de donnée, les projets de migration, etc. En effet, un projet avec un revenu annuel approximatif de 5 milliards de dollars US perd en moyenne 2.1 millions de dollars US à cause de mauvaises décisions prises sur la base de données de mauvaise qualité. 40% des initiatives métier non abouties sont le résultat d'une mauvaise qualité des données (Gartner, 2011 ; Laney, 2000).

Pour bon nombre d'organisations ayant un volume critique de données, le coût de la fiabilisation de l'ensemble de ses données ne serait pas absorbable ou au mieux nécessiterait une réorganisation à grande échelle. La clé est donc la priorisation des projets de qualité des données.

En d'autres termes, comprendre quelles sont les informations qui sont utilisées de manière récurrente par quels processus métier, lesquelles ont un niveau inacceptable de qualité des données, quels processus pèsent le plus dans la balance des objectifs financiers/métier et lesquels présentent le plus de risques par rapport aux cahiers des charges des autorités de régulation.

Dans un premier temps, la démarche *PortfolioDQAF* relie les caractéristiques des données, principalement en termes de précision aux objectifs de l'organisation. Elle mesure ensuite de manière quantitative, l'impact positif et la complexité d'implémentation, traduisible en coûts monétaires. Les modèles sont censés être une approximation de la réalité à travers les inputs disponibles et les fonctions qu'ils implémentent. L'étude de cas présentée et détaillée dans ce chapitre permet de vérifier jusqu'à quel point *PortfolioDQAF* est pertinent pour conduire des programme de qualités de données et prioriser ainsi que les initiatives d'amélioration de la qualité et en justifier les coûts associés.

L'organisation de ce chapitre est comme suit : la section 2 présente le contexte de cette étude de cas au sein de la Direction des Domaines de l'Etat (DDE), Ministère de l'Economie (MEF) et des Finances, Rabat, Maroc. Après la description de l'étude de cas, le setup expérimental est présenté dans ses aspects : paramétrage, processus et objets métier candidats. Les résultats de l'étude de cas sont ensuite détaillés.

V.2. Evaluation du portefeuille des projets de qualité des données

Comme étayé au niveau du chapitre II, la démarche *PortfolioDQAF* permet d'identifier clairement les meilleures initiatives pour optimiser les projets d'amélioration de la qualité des données. L'objectif étant de mesurer, de manière quantitative, la façon avec laquelle les processus clés participent à mettre en œuvre la stratégie d'une organisation, puis de qualifier l'impact ainsi que la complexité d'implémentation de l'amélioration de la qualité des données, consommées ou produites par ces processus. Les outputs permettent d'identifier clairement les projets

d'amélioration de la qualité des données, en fonction des avantages pour l'organisation et des coûts de mise en œuvre.

Dans le but de vérifier et de valider cette démarche d'évaluation de portefeuilles de projets de qualité des données, dans un contexte d'AE, une étude de cas a été conduite au sein de la DDE.

V.2.1. Description de l'étude de cas

La DDE gère le domaine privé de l'Etat (DPE), qui est constitué de l'ensemble des biens immobiliers et mobiliers dont l'Etat est propriétaire et ne faisant pas partie de son domaine public. Sa principale mission est de mettre à contribution le patrimoine immobilier de l'Etat pour apporter des bénéfices économiques, sociaux et environnementaux, tout en optimisant la valeur des actifs de la DDE.

Pour atteindre ses objectifs, la DDE avait implémenté un ensemble de processus totalement automatisés, à travers son système d'information SIDOM (Système d'Information des DOMaines), parmi lesquels : la cession dans le cadre de l'investissement, la gestion du programme annuelle des acquisitions en faveur des départements ministériels (éducation, santé, habitat, etc.), la comptabilité domaniale, la vente et location aux fonctionnaires de l'Etat, etc.

La figure 28 illustre la cartographie applicative de SIDOM.

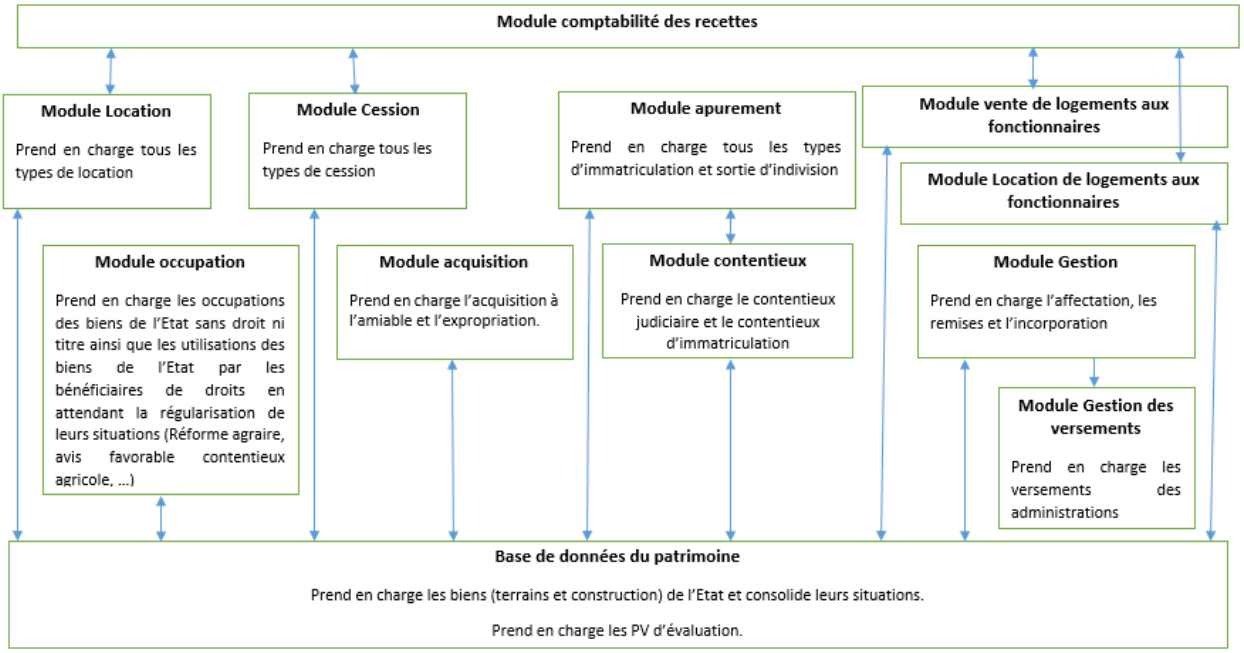


Figure 28. Cartographie applicative de SIDOM

Dans un environnement dynamique et changeant, et pour permettre à la DDE de répondre de manière efficace et efficiente à ses attributions, il est vital qu'elle dispose de données précises.

Parallèlement, la DDE mène un chantier d'envergure pour la refonte de son SI, qui est à coupler avec le chantier de migration des données vers le nouveau système cible. Ces deux systèmes sont différents en termes de technologies et de structures. Seulement pour bénéficier pleinement des avantages qu'offre le nouveau système, sur les plans opérationnel, stratégiques et décisionnel, il est important de l'alimenter avec des données qui répondent aux seuils de qualité définis par la DDE.

Etant donné les challenges économiques actuels et les pressions budgétaires auxquels doivent faire face les organisations, parmi lesquelles la DDE, il existe un désir substantiel pour éradiquer ou du moins, réduire les problèmes liés à la qualité des données, avec un budget raisonnable et des changements non critiques.

L'objectif général du projet d'assainissement des données de la DDE est d'améliorer la qualité des données au niveau de son SI. Cependant, les objectifs spécifiques permettent d'avoir une vue plus détaillée, afin de mieux paramétrer la démarche. Ces objectifs spécifiques sont :

- Connaître la réserve foncière disponible ;
- Disposer de tableaux de bord fiables ;
- Connaître les immeubles domaniaux (ID mobilisés par certaines procédures, à classer par importance par rapport aux missions de la DDE) ;
- Répondre aux besoins à court et moyen termes des départements gouvernementaux, investisseurs et citoyens.

V.2.1.1 Aspects et coûts de la non-qualité des données

Le déploiement progressif du système SIDOM, a permis d'alléger les tâches opérationnelles et de normaliser les circuits des procédures de gestion au niveau de la DDE. Cependant, SIDOM enregistre plusieurs anomalies liées à la qualité des données manipulées.

Ces problèmes se manifestent par :

1. La fréquence des maintenances correctives des données. En effet, malgré les efforts de la DSI pour cerner la qualité des données saisies au niveau des systèmes à travers différents contrôles et règles de gestion, il y a un nombre croissant d'appels dont l'objet est la correction de données. Ceci a évidemment un impact sur le déroulement normal des opérations en attente de la correction de l'anomalie ainsi que sur le planning de l'équipe qui prend en charge leur correction, les tests et le déploiement des correctifs.
   D'après une situation extraite du système de *bug tracking* de la DDE (sur une période d'une année entre 05/08/2013 et 05/08/2014), le volume des anomalies relatives à la correction de données représentent 35% du volume global des appels. Cette situation est synthétisée dans le tableau 22.

Tableau 22. Répartition de la charge de développement et de maintenance

Nature des développements	Charge en J*H
Maintenance corrective de SIDOM	364
Maintenance évolutive de SIDOM	587
Nouveaux développements	263
Total	1114

D'autant plus qu'à partir de 2015, il y'a eu peu de maintenance évolutive, par contre la maintenance corrective a maintenu une cadence croissante. Cette activité concerne 260 appels en 2016 et 274 appels en 2017 ;

2. L'exploitation partielle des tableaux de bord alimentés par le système : pour certaines données, la non-précision empêche d'exploiter la puissance de ces outils d'aide à la décision, ce qui aurait permis un gain en temps pour les managers ;
3. Les rejets des prises en charge par les organismes de précompte et par la TGR à cause de la non-précision des données d'identification des bénéficiaires, à savoir : nom, prénom, n° CNI et matricule, ainsi que le montant des redevances ;
4. La limitation du nombre des services électroniques ouverts aux citoyens, ce qui entrave la réalisation de la stratégie de l'e-gov.

Après avoir parcouru les différents aspects et coûts engendrés par la mauvaise qualité des données, particulièrement la non-précision, il est maintenant intéressant d'investiguer ses causes. Comme soulevé à la fin du chapitre II, il est important d'identifier les causes à l'origine des problèmes de qualité des données et de déterminer comment elles peuvent être adressées.

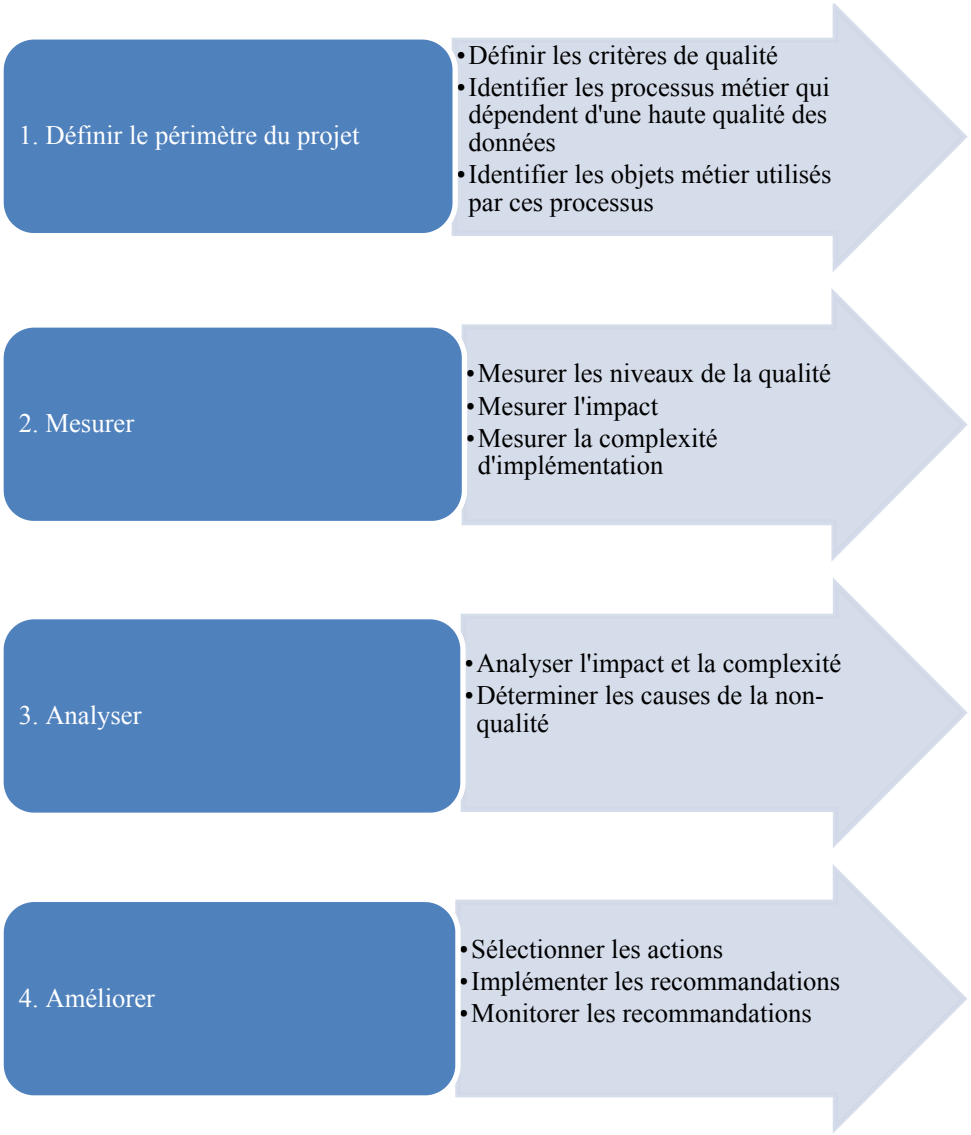
V.2.1.2 Causes de la non-qualité des données

Il est possible d'imputer les niveaux insatisfaisants de la qualité des données à un certain nombre de causes, qui sont comme suit : (i) la non-saisie au niveau de SIDOM, (ii) les erreurs de saisie, (iii) la non-maîtrise du système, (iv) la non-conformité de quelques règles de gestion du système avec les règles métier, (v) les problèmes techniques, (vi) les problèmes de gestion.

Le périmètre de cette étude de cas concerne la précision des données, objet de la démarche *PortfolioDQAF* développée au niveau des chapitres II et III. Le périmètre de cette étude concerne donc les données non-précises, soit les anomalies liées aux erreurs de saisie.

V.2.3. Cycle de vie du projet d'assainissement des données de la DDE

Les phases qui composent la démarche continue de la DDE pour améliorer la qualité de ses données sont illustrées par la figure 29. Ce cycle de vie s'inspire du cycle général des projets de qualité des données, présentée à la section II.3 (figure 11).



.

Figure 29. Phases de l'approche d'amélioration de la qualité, adoptées par la DDE

La figure 29 établit un rapprochement entre le cycle de vie des projets de qualité des données et les étapes qui composent la démarche *PortfolioDQAF*. A chaque étape du cycle de vie, correspond une ou plusieurs phases de la démarche *PortfolioDQAF*.

La présente étude de cas couvre l'ensemble des phases. En effet, la démarche *PortfolioDQAF* est utilisée pour identifier les projets d'amélioration de la qualité des données avec une meilleure

valeur ajoutée, en conformité avec les objectifs de la DDE et qui sont moins complexes à implémenter, ce qui peut être directement corrélé avec les coûts financiers.

Le processus de calcul de l'impact positif et la complexité d'implémentation est déroulé automatiquement, à travers l'application Web « *PortfolioDQAF-tool* » présentée au niveau du chapitre III.

V.2.4. Setup expérimental : paramétrage, processus et objets métier

V.2.3.1 Paramétrage de la solution par les gestionnaires métier

Le modèle *PortfolioDQAF* intègre un aspect paramétrage, qui exprime l'importance relative de chacun des facteurs de l'impact positif des processus métier et de la complexité d'amélioration de la qualité des données. En effet, ce que met *PortfolioDQAF* en avant est sa capacité à s'adapter à différents environnements des organisations en proposant un template paramétrable.

Vu que l'entité qui est chargée du pilotage stratégique du projet d'assainissement des données de la DDE est la division de l'Organisation et du Contrôle de Gestion et particulièrement le service de l'Organisation et de l'Analyse de Données (SOAD), en étroite collaboration avec la DSI, ces deux entités ont été approchées pour paramétrer, les templates de « *PortfolioDQAF-tool* ». Le service SOAD a pour principales missions :

1. La normalisation de l'activité de la DDE ;
2. La collecte, traitement et analyse des données statistiques ;
3. La conception et développement des outils d'aide à la décision.

La cadre logique du projet d'assainissement de données, représenté par le tableau 23, permet de comparer les objectifs généraux et spécifiques du projet d'assainissement avec les facteurs d'impact et de complexité du modèle *PortfolioDQAF*.

Tableau 23. Cadre logique simplifié du projet d'assainissement des données de la DDE

Objectifs généraux	- Moderniser la gestion du patrimoine de l'Etat - Implémenter les bonnes pratiques de la gouvernance - Offrir des outils fiables d'aide à la décision
Objectifs spécifiques	- Anticiper la reprise et le transfert de données vers le nouvel ERP de la DDE - Préparer les données de la DDE en vue d'échanges automatisés avec ses partenaires (TGR, CMR, DD, etc.) - Réduire le temps des traitements correctifs des anomalies - Avoir des statistiques précises pour la prise de décision - Améliorer la qualité des services électroniques à ouvrir aux citoyens

Le paramétrage du framework doit mettre en exergue les missions de la DDE ainsi que ses objectifs stratégiques et tactiques.

La table 24 effectue une correspondance entre les objectifs spécifiques du projet d'assainissement de données de la DDE et quelques critères du modèle *PortfolioDQAF* :

Tableau 24. Mapping des objectifs spécifiques du cadre logique avec les critères du framework

Objectif spécifique du cadre logique	Critère du framework
Anticiper la reprise et le transfert de données vers le nouvel ERP de la DDE	Impact sur les objectifs financiers/métier suivants : - Autre : réussir la refonte du SI
Préparer les données de la DDE en vue d'échanges automatisés avec ses partenaires (TGR, CMR, DDP entre autres)	Impact sur l'efficacité opérationnelle Impact sur les objectifs financiers/métier suivants : - Générer plus de revenus - Améliorer la productivité - Minimiser les coûts
Réduire le temps des traitements correctifs des anomalies	Impact sur l'efficacité opérationnelle Impact sur les objectifs financiers/métier suivants : - Améliorer la productivité - Minimiser les coûts
Avoir des statistiques précises pour la prise de décision	- Impact direct sur l'analyse en aval - Impact direct sur la prise de décision
Améliorer la qualité des services électroniques à ouvrir aux citoyens	Impact sur les objectifs financiers/métier suivants : - Améliorer la satisfaction client

Comme développé dans des chapitres II et III, les facteurs d'impact positif et de complexité sont éclatés en plusieurs critères, qui chacun selon son importance absolue et surtout relative par rapport aux autres critères, possède une pondération.

Pour simplifier les calculs, la somme des coefficients pondérateurs doit être égale à 100. La notation est par la suite ramenée à une échelle de 0 à 5 pour exprimer l'interprétation de l'impact positif et de la complexité d'implémentation.

Les scores d'impact positif et de complexité d'implémentation, varient ainsi entre 0 et 5. Pour l'impact positif, la valeur « 0 » correspond à un impact imperceptible et « 5 » fait référence à un impact positif très élevé. Idem pour la complexité d'implémentation.

Le tableau 25 illustre cette démarche.

Tableau 25. Tableau de paramétrage de l'impact positif par les gestionnaires métier

Facteur	Valeurs	Notation (R)	Pondération (C)
Impact sur l'efficacité opérationnelle	- Vrai - Faux	1 0	**20**
Impact sur les objectifs financiers/métier à court terme	- Augmenter les revenus - Améliorer la productivité - Réduire les coûts - Améliorer la satisfaction client - Se conformer aux autorités de régulation - Autres	0.15 0.15 0.15 0.15 0.15 0.15	**20**
Impact sur la prise de décision	- Vrai - Faux	1 0	**10**
Impact sur l'analyse en aval	- Vrai - Faux	1 0	**15**
Caractère transverse du processus	- Vrai - Faux	1 0	**5**
Délai de réalisation opportun	- Vrai - Faux	1 0	**20**
Nature de la donnée	- Master data - Donnée transactionnelle - Donnée d'historique	4 2 0	**10**

A partir du paramétrage susmentionné, il en ressort que les critères d'impact positif les plus pesants pour la DDE sont : l'impact sur l'efficacité opérationnelle, les objectifs financiers/métier à court terme et la durée de réalisation. Cette importance est exprimée à travers l'attribution d'une pondération plus élevée. Les autres critères participent moins à la construction de ce facteur.

La même démarche a été réitérée pour le paramétrage de la complexité d'implémentation. Cependant, ce paramétrage a été accompli par les responsables SI plutôt que les gestionnaires métier. Le tableau 26 illustre le paramétrage de la complexité d'implémentation.

Tableau 26. Tableau de paramétrage de la complexité d'implémentation par les responsables SI

Facteur	Valeurs	Notation (R)	Pondération (C)
Existence de normes pour valider les données	- Faux - Vrai	1 0	**15**
Niveau de risque	- Sévère/inacceptable - Majeur - Moyen - Mineur - Imperceptible	1 0,75 0,5 0,25 0	**15**

Niveau de changement	- Sévères - inacceptable - Majeurs - Moyens - Mineurs - Pas d'impact	1 0,75 0,5 0,25 0	**15**
Existence d'un référentiel de données	- Faux - Vrai	1 0	**10**
Potentiel d'identification par clé primaire	- Faux - Vrai	0 1	**25**
Nature du traitement des données	- Manuel - Semi-automatique - Automatique	1 0.5 0.25	**5**
Volumétrie des données à traiter	- Volume très large - Large - Moyen - Faible	1 0.75 0.5 0.25	**15**

A partir du tableau, il est possible de dire que les analystes SI à la DDE accordent une très grande importance au potentiel d'identification par clé primaire. Effectivement, ceci permet de rendre possible d'identifier les enregistrements correspondants à partir d'un autre référentiel de données et ainsi de réduire la complexité de l'assainissement des données.

V.2.3.2 Processus et objets métier candidats

En premier lieu, *« PortfolioDQAF-tool »* calcule l'impact positif des processus métier en input. L'étape suivante consiste à calculer la complexité d'implémentation de la précision des données pour un sous-ensemble de données qui sont manipulées par les processus métier en question. La troisième étape est de recommander le scénario optimal, soit le scénario avec le meilleur rapport impact positif et complexité d'implémentation. La liste des processus métier et des objets métier qui ont été sélectionnés pour le setup expérimental a été suggéré par les responsables des domaines métier. Le tableau 27 décrit les processus métier candidat à l'évaluation, ainsi que les différentes expressions de la non-qualité et des actions correctives à considérer.

Tableau 27. Description des processus métier candidats

Processus métier	Description	Coût de la non-qualité	Actions candidates
Apurement (enregistrement des titres fonciers)	Procédure visant à immatriculer les biens constituant le Domaine Privé de l'Etat (DPE), à lever les charges le grevant et à corriger les superficies et la localisation (coordonnées géographiques)	- Mauvaise maîtrise du patrimoine - La non valorisation du patrimoine - Erreurs lors de la prise de décision relatives au patrimoine disponible	(A1) : Rectification des numéros et indices des titres fonciers de la BDP[26] conformément aux données de la Conservation Foncière
Comptabilité des recettes et produits domaniaux	La comptabilité domaniale gère deux flux financiers majeurs : les recettes et les dépenses. Une recette périodique est un paiement, régulier (mensuel ou annuel), effectué par le redevable, elle est liée à un dossier de procédure. Le redevable est principalement identifié par son numéro de CNI	- Rejet par les comptables assignataires - Temps pour recouvrer les recettes - Risque de prescription après dépassement du délai réglementaire de recouvrement - Charges opérationnelles pour traiter les rejets	(A2) : Rectification des numéros des CNI
Locations	Gestion de la location des terrains domaniaux avec ou sans réalisation de projets	- Perte de revenus	(A3) : Redressement des dossiers de location avec un taux de révision présumé erroné
Gestion des fonctionnaires logés	Gestion des dossiers de location et de vente de logements aux fonctionnaires logés, ainsi que la prise en charge des différentes pièces comptables	- Perte de revenus - Statistiques erronées sur les logements présumés disponible - Non-maîtrise du patrimoine - Surestimation du nombre de logements gérés par la DDE - Statistiques erronées sur les logements présumés disponible	(A4) : Redressement de la situation des parcs de logements des fonctionnaires pour les logements vacants sans prise en charge active (A5) : Redressement de la situation des parcs de logements des fonctionnaires pour les logements vacants sans prise en charge initiale

Le tableau 28 représente pour chaque objet métier, manipulé par les processus métier clés, ses principaux attributs, ainsi que leur définition.

[26] Base de Données du Patrimoine

Tableau 28. Description des objets métier candidats

Objet métier	Principaux attributs	Définition
Bien immobilier[27] – Données légales	- référence foncière - dénomination - superficie - liste propriétaires - quoteparts - document propriété	Données qui décrivent la situation légale d'un bien immobilier appartenant à l'Etat, en termes de sa référence foncière (titre foncier, réquisition ou non immatriculé), ses propriétaires ainsi que leurs quoteparts dans le cas d'un bien en copropriété
Bien immobilier – Données urbanistiques	- référence foncière - nature périmètre - zoning - coordonnées géo. - adresse	Données qui décrivent la situation des biens immobiliers, en termes de situation par rapport au périmètre urbain et données géographiques
PV d'expertise	- référence foncière - valeur bien - date évaluation - nature évaluation	Données qui décrivent le résultat de l'évaluation des terrains appartenant à l'Etat, en vue de les céder ou les louer
Dossier de procédure	- identification dossier procédure - procédure - localisation - date ouverture - date fermeture - phase en cours	Données qui décrivent le dossier de procédure, à savoir son identifiant unique, son emplacement, ses dates d'ouverture et de clôture, la phase en cours ainsi que la procédure à laquelle il est lié
Pièce comptable	- num. pièce comptable - type pièce comptable - << Dossier procédure >> - << Redevable >> - montant	Données qui décrivent le type de la pièce comptable (dépense ou revenu), son identifiant, son montant et la procédure à laquelle elle est liée
Redevable	- qualité interlocuteur - nom, prénom - CNI - matricule registre de commerce - num. de passeport - adresse - << Mandataire >>	Données d'identification de la personne physique ou morale (publique ou privée)
Contrat	- num. contrat - nature opération - périodicité - date début - date fin - << Bénéficiaire >>	Données d'un contrat de location

[27] Bien immobilier fait référence aux biens de type terrain ou construction

Les tableaux 27 et 28 décrivent donc les processus et les objets métier candidats qui ont été considérés pour l'amélioration de la précision des données. Le tableau 29, quant à lui, décrit le mode d'accès des processus métier aux données. Il montre comment ces objets métier sont manipulés par les processus métier critiques, en termes d'opérations de bases de données.

Tableau 29. Matrice d'accès

Objet métier	**Processus métier**			
	Apurement	**Comptabilité des recettes et produits domaniaux**	**Locations**	**Gestion des fonctionnaires logés**
Bien foncier – Données légales	CRUD[28]	R	R	R
Bien foncier – Données urbanistiques	CRUD	R	R	R
Valeur vénale/locative	-	R	R	R
Dossier de procédure	-	R	CRUD	CRUD
Pièce comptable	-	CRUD	R	R
Redevable	-	RU	CRUD	CRUD
Contrat de location	-	R	CRUD	-

Le déroulement de l'étude de cas au niveau du framework permettra de voir comment chaque processus et chaque donnée scorent au niveau de l'impact et de la complexité d'implémentation de l'amélioration de la qualité des données. En complément du tableau 29, le tableau 30 offre un aperçu sur la volumétrie des données au démarrage du projet en 2015 (15/12/2015).

Tableau 30. Volumétrie des données au démarrage du projet

Action	**Taille des enregistrements**	**Volumétrie**
(A1)	1005 enregistrements	Faible
(A2)	8542 enregistrements	Très élevée
(A3)	234 enregistrements	Moyenne
(A4)	10 795 enregistrements	Très élevée
(A5)	62 enregistrements	Très faible

[28] *Create, Read, Update, Delete*

V.2.5. *« PortfolioDQAF-tool »* en action

Chacun des responsables des domaines fonctionnels était invité à évaluer pour le couple processus métier et objet métier, l'impact positif et la complexité d'implémentation. Cette évaluation a été menée à travers *« PortfolioDQAF-tool »*. Les résultats intermédiaires ainsi que les valeurs finales sont stockées pour l'analyse en aval et la prise de décision.

La figure 30 illustre pour chaque objet métier, l'impact positif du processus clé qui l'utilise.

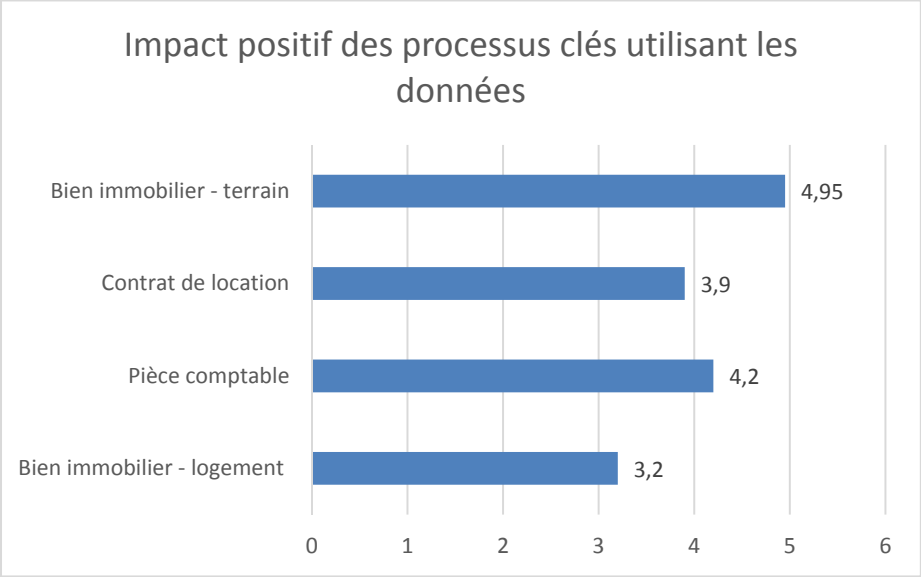


Figure 30. Score de l'impact positif pour les processus métier analysés

La figure 31 illustre le résultat de l'évaluation de la complexité de la mise en place des actions d'amélioration pour chacun des objets métier susmentionné.

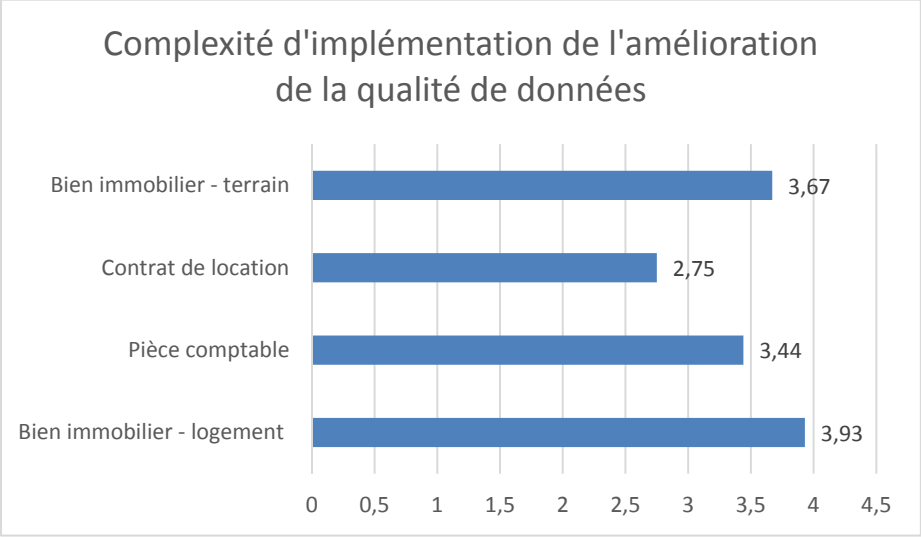


Figure 31. Score du facteur de complexité pour les objets métier analysés

A partir de la figure 30, il est possible de constater que le score de l'impact positif le plus élevé est lié au processus « Apurement », suivi de la « Comptabilité des recettes ». La « Vente et location de logements » a eu le moins d'impact positif.

La figure 31 illustre les résultats du calcul du score de complexité de la mise en œuvre. A partir de cette figure, il est possible de déduire que le score le plus élevé de complexité de mise en œuvre est associé à l'objet « Bien immobilier – logement », manipulé principalement par le processus « Vente et location de logements ».

Un examen plus approfondi des résultats des évaluations administrées via la plateforme Web fournit des informations détaillées sur les éléments ayant contribués aux niveaux élevés ou faibles de l'impact positif et de la complexité d'implémentation. A titre d'exemple, l'amélioration de l'objet métier « Bien immobilier – terrain » a un impact important car le processus « Apurement » influence quasiment tous les objectifs financiers et métier de la DDE. D'autre part, l'amélioration de la précision de l'objet métier « Bien immobilier – logement » est très complexe, car il n'existe aucune norme pour restructurer et valider les données, ni une source authentique de données permettant de les compléter ou de les contredire. De plus, le traitement des fichiers est manuel et la taille des données à traiter est élevée.

Le travail accompli plus haut permet d'associer des mesures quantitatives à l'impact positif et à la complexité d'implémentation des projets d'amélioration de la précision des données. Pendant l'évaluation des coûts-bénéfices associés à ces projets, il est important de prendre en compte la précision initiale des objets métier. Cela est pertinent car il est moins coûteux d'améliorer un objet de données avec une précision initiale plus élevée.

Les résultats de l'application du modèle de coût de la qualité, objet de la section III.3 (chapitre III) sont représentés au niveau de la figure 32 : l'axe des abscisses représente la précision, tandis que l'axe des ordonnées représente le coût associé à l'amélioration de la précision des données.

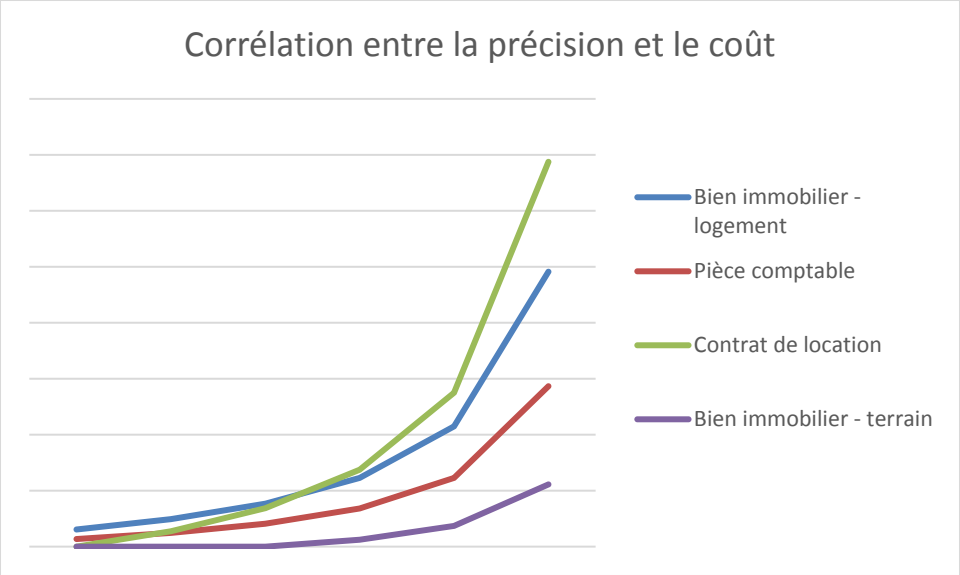


Figure 32. Corrélation entre la précision et la complexité de mise en œuvre

Afin de construire le scénario optimal pour l'amélioration de la précision des données et ainsi améliorer l'exécution des objectifs stratégiques de la DDE, les gestionnaires métier utiliseront les résultats extraits des figures 30, 31 et 32 qui représentent les avantages et coûts inhérents à l'amélioration de la précision des données.

V.2.6. Recommandation

A partir de ce qui précède, les actions du tableau 27 sont exécutées dans l'ordre suivant :

- (A1) : Rectification des numéros et indices des titres fonciers de la BDP conformément aux données de l'ANCFCC [29] ;
- (A2) : Rectification des numéros des CNI ;
- (A4) : Redressement de la situation des parcs de logements des fonctionnaires pour les logements vacants sans prise en charge active ;
- (A5) : Redressement de la situation des parcs de logements des fonctionnaires pour les logements vacants sans prise en charge initiale ;
- (A3) : Redressement des dossiers de location avec un taux de révision présumé erroné.

Pour l'action (A1), une interface de recoupement des données de la DDE avec les données de l'ANCFCC est développée. Pour le reliquat des actions, des tableaux de bord sont prévu pour remonter les données non-précises et faire le suivi de leur correction.

[29] Agence Nationale de la Conservation Foncière du Cadastre et de la Cartographie

V.4. Volet « Open Data »

Le mouvement Open data au Maroc a été propulsé par la nouvelle Constitution de juillet 2011, et notamment l'article 27 consacrant le droit d'accès à l'information.

Cependant, le Maroc n'a pas pu en tirer les avantages escomptés à cause du retard qu'accusait le texte de loi relatif au droit d'accès à l'information dans le circuit législatif. En outre, l'Open Data ne faisait partie d'aucune stratégie gouvernementale claire et ne bénéficiait par de la communication et la sensibilisation adéquate, au sein des départements ministériels producteurs de données et auprès des acteurs pouvant utiliser ces mêmes données.

Selon la dernière édition de l'*Open Data Barometer*[30], mise en ligne par l'*Open Data Institute* (ODI), le Maroc a perdu 15 places au classement mondial. Cette situation est illustrée par la figure 33. Selon l'ODI, ceci est dû à la qualité, fréquence d'actualisation et pertinence des jeux de données publiées, qui demeurent limitées. L'Open Data ne faisait pas partie d'une stratégie gouvernementale claire et ne bénéficiait pas de la communication et la sensibilisation adéquate.

Country	Barometer Rank	ODB Scaled	Readiness (Scaled)	Implementation (Scaled)	Impact (Scaled)	2013 ODB	ODB Change	2013 Rank	Rank Change
Costa Rica	41	31.26	56	33	6	31.21	0.05	36	-5
Turkey	41	31.24	47	35	6	27.58	3.66	37	-4
Malaysia	41	30.76	44	37	3				
South Africa	41	30.7	48	31	15	19.2	11.5	52	11
Tunisia	45	28.57	58	19	30	21.02	7.55	50	5
China	46	28.12	52	24	19	11.82	16.3	61	15
Rwanda	46	28.05	36	35	3	24.27	3.78	45	-1
Ghana	46	27.99	35	36	0	21.6	6.39	47	1
Jamaica	49	26.26	42	27	11	22.69	3.57	46	-3
Kazakhstan	49	25.87	40	30	3	27.61	-1.74	37	-12
Kenya	49	25.8	42	23	20	43.06	-17.26	22	-27
UAE	52	24.86	53	22	8				
Philippines	53	23.19	58	18	8	21.91	1.28	47	-6
Mauritius	54	21.86	35	25	3	26.08	-4.22	42	-12
Ukraine	55	21.23	37	23	6				
Morocco	55	21.11	47	15	18	27.24	-6.13	40	-15

Figure 33. Classement du Maroc selon l'*Open Data Barometer* – Edition 2016

Avec l'adoption définitive de la loi 31-13, du droit d'accès à l'information en février 2018 (Bulletin Officiel, 2018) et l'adhésion du Maroc à l'*Open Government Partnership* (OGP), le « club des gouvernements transparents » en avril 2018, l'Open Data du gouvernement du Maroc est devenu un chantier prioritaire au sein du programme e-gov.

Actuellement, aucun jeu de données de la plateforme « data.gov.ma » n'est attribué à la Direction des Domaines de l'Etat. En dehors des jeux de données, les seuls documents, qu'il est possible de trouver sur le Web et qu'il est possible de qualifier de données ouvertes est le rapport d'activités de la DDE et qui date de l'année 2011 (MEF-DDE, 2011).

[30] https://opendatabarometer.org/2ndEdition/analysis/rankings.html

Les autres rapports d'activités annuels sont uniquement disponibles au niveau du site intranet du MEF.

Cette situation serait attribuable à l'absence d'une politique d'ouverture des données, à la non-identification d'un point focal pour gérer le cycle de vie des données à ouvrir et à l'absence de la communication et la sensibilisation autour de l'intérêt et des bénéfices de l'Open Data et des mesures de la publication proactive comme le stipule la loi 31-13.

Au vu du contexte actuel du Maroc et particulièrement du cadre légal adapté, et dans l'optique d'une publication proactive des données de la DDE, le volet « Open Data » de la démarche *PortfolioDQAF* est parfaitement opportun, dans le cadre de cette publication proactive.

V.5. Conclusion

Dans le cadre de l'amélioration de la précision de ses données critiques, prérequis de la réussite de la refonte de son SI et sa migration vers un système de gestion intégré, la démarche *PortfolioDQAF* a été appliquée au sein de la DDE pour analyser les coûts-avantages des différents scénarios d'amélioration de la qualité. En effet, l'impact financiers/métier ainsi que la complexité d'implémentation sont mesurés de manière quantitative, en établissant deux indicateurs globaux.

Les recommandations issues de l'application de la démarche *PortfolioDQAF* permettent de mettre en exergue les actions d'amélioration de la qualité des données les plus avantageuses en termes de rapport coûts-bénéfices et qui sont en phase avec les orientations et les objectifs actuels de la DDE.

Les données et les calculs intermédiaires recueillis et persistés dans « *PortfolioDQAF-tool »* permettent de mieux analyser comment les processus clés contribuent à la réalisation de la mission de la DDE ainsi que la compréhension des défis techniques et métier relatifs à l'amélioration de la qualité des données.

Conclusion

Récapitulatifs du travail

Les problèmes de qualité des données illustrent parfaitement le principe clé de tout effort de performance : *« On ne peut pas contrôler ce qu'on ne peut pas mesurer ».* En effet, sans une analyse de rentabilisation des efforts d'amélioration de la qualité des données, les organisations risquent d'engager leurs ressources financières et humaines dans les problèmes perçus, plutôt que les problèmes réels. Elles continueraient alors à subir les effets de la mauvaise qualité des données sur l'efficacité opérationnelle, les revenus, la conformité aux autorités de régulation et la satisfaction des clients. Il en découle qu'un processus optimal et efficient de l'amélioration de la qualité des données est une condition préalable d'une valeur ajoutée exceptionnelle des processus métier implémentés par les organisations.

Ce travail de thèse présente la démarche *PortfolioDQAF*, qui est une approche basée sur des métriques, pour évaluer et analyser l'impact positif et la complexité de mise en œuvre des projets d'amélioration de la qualité des données. Cette démarche offre une base factuelle qui permet d'identifier et de justifier les investissements en termes de qualité des données. Elle constitue également un médium de communication entre les managers métier et les analystes SI.

La revue de littérature réalisée porte sur les modèles de coût de la qualité ainsi que l'évaluation de la valeur financière et métier de la qualité des données. L'objectif étant de comparer les approches selon différents critères et en ressortir les aspects à combler.

Ce travail de thèse repose sur un méta-modèle de l'Architecture d'Entreprise pour analyser l'incidence de la qualité des processus et particulièrement celle des objets métier sur la qualité globale des services métier exposés aux clients externes et internes par l'organisation.

PortfolioDQAF développe un modèle coûts-bénéfices, basé sur l'analyse multicritères d'aide à la décision, pour évaluer les portfolios de projets d'amélioration de la qualité des données. *PortfolioDQAF* se décline en deux volets, à savoir les données d'entreprise et l'Open Data. L'outil *« PortefolioDQAF-tool »* implémente cette démarche.

La démarche est ensuite illustrée par un cas pratique pour les besoins de planification d'action d'amélioration des *Data Assets* d'un département gouvernement dans le cadre de la refonte de son Système d'Information.

Contributions

Le travail accompli dans le cadre de la présente thèse consiste en la proposition d'une démarche d'évaluation qualitative et quantitative, de la valeur financière et métier des projets d'amélioration de la précision des données, en établissant deux indicateurs globaux : l'impact positif et la complexité d'implémentation.

Peuvent aussi être mentionnées comme contributions de ce travail :

- L'adaptation d'un méta-modèle de l'Architecture d'Entreprise pour supporter l'analyse de l'aspect qualité des services, processus et objet métier ;
- La caractérisation de l'impact positif de la qualité des données sur la qualité globale des processus métier clés, et de manière transitive sur l'exécution de la stratégie de l'organisation ;
- La caractérisation de la complexité de mise en œuvre et la proposition d'un modèle de coût de la qualité ;
- Le développement d'un framework de calcul et son implémentation à travers une *Web-based* plateforme ;
- L'élaboration d'une étude de cas complète et d'envergure de la démarche pour un département gouvernemental.

Perspectives et travaux futurs

Le travail de thèse réalisé ainsi que les résultats obtenus permettent d'identifier de nombreuses voies de recherche et de développement, parmi lesquelles :

- La généralisation de la démarche *PortfolioDQAF* pour un plus grand nombre des dimensions de la qualité des données, en plus de la précision et de la complétude. Par exemple, l'actualisation et la cohérence ;
- L'enrichissement de la démarche par un mécanisme automatisée de réévaluation de la progression des coûts dans le cadre d'une amélioration continue de la qualité ;
- A l'instar de l'étude de cas du chapitre V, la conduite d'une deuxième étude de cas dans le contexte des données gouvernementales ouvertes au Maroc. En effet, dans le cadre de ce travail, des niveaux disparates de la qualité des données ont été identifiés (figure 23). Il est maintenant intéressant d'analyser comment la mise en place de la démarche proposée, dans sa déclinaison Open Data, permettrait de faciliter la publication de données d'une qualité répondant aux exigences et attentes des usagers.

Bibliographie

Aladwani, A. M., & Palvia, P. C. (2002). Developing and validating an instrument for measuring user-perceived web quality. *Information & management*, *39*(6), 467-476.

A.V, Feigenbaum. (1956). *Total Quality Control*: Harvard Business Review

Aoieong, R. T., Tang, S. L., & Ahmed, S. M. (2002). A process approach in measuring quality costs of construction projects: model development. *Construction Management & Economics*, *20*(2), 179-192.

Archi – Open Source ArchiMate Modelling. (2012). Retrieved from https://www.archimatetool.com/

ASQ. (n.d.). Cost of Quality (COQ). Retrieved December 25, 2018, from https://www.informatica.com/downloads/1530_KnowledgeIntegFinValueDQ.pdf

Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015). A systematic review of open government data initiatives. *Government Information Quarterly*, *32*(4), 399-418.

Auer, S., Bühmann, L., Dirschl, C., Erling, O., Hausenblas, M., Isele, R., ... & Stadler, C. (2012, November). Managing the life-cycle of linked data with the LOD2 stack. In *International semantic Web conference* (pp. 1-16). Springer, Berlin, Heidelberg.

Barid Al-Maghrib. (n.d.). Annuaire du code postal. Retrieved December 25, 2018, from http://www.codepostal.ma/annuaire.pdf

Barrau, D., Barthélémy, N., Kedad, Z., Laboisse, B., Nugier, S., & Thion, V. (2016). Gestion de la qualité des données ouvertes liées-État des lieux et perspectives. *Revue des Nouvelles Technologies de l'Information*.

Batini, C., & Scannapieca, M. (2006). Data quality: Concepts, methodologies and techniques. Berlin: Springer.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)*, *41*(3), 16.

Batini, C., Comerio, M., & Viscusi, G. (2012, October). Managing quality of large set of conceptual schemas in public administration: methods and experiences. In *International Conference on Model and Data Engineering* (pp. 31-42). Springer, Berlin, Heidelberg.

Belhiah, M., & Bounabat, B. (2017, October). A User-Centered Model for Assessing and Improving Open Government Data Quality. In *MIT International Conference on Information Quality (ICIQ)*.

Belhiah, M., Benqatla, M. S., & Bounabat, B. (2015a, July). Decision support system for implementing data quality projects. In *International Conference on Data Management Technologies and Applications* (pp. 1-16). Springer, Cham.

Belhiah, M., Benqatla, M. S., Bounabat, B., & Achchab, S. (2015b, July). Towards a Context-aware Framework for Assessing and Optimizing Data Quality Projects. In *DATA* (pp. 189-194).

Belhiah, M., Bounabat, B., & Achchab, S. (2015a, June). The impact of data accuracy on user-perceived business service's quality. In *Information Systems and Technologies (CISTI), 2015 10th Iberian Conference on* (pp. 1-4). IEEE.

Bogdanović-Dinić, S., Veljković, N., & Stoimenov, L. (2014). How open are public government data? An assessment of seven Open Data portals. In *Measuring E-government efficiency* (pp. 25-44). Springer, New York, NY.

British Standards Institution. (1992). *BS 6143. Guide to the economics of quality*. London: BSI.

Bulletin Officiel 6340. (2015, March). Retrieved from http://www.sgg.gov.ma/BO/FR/2015/BO_6340_Fr.pdf

Bulletin Officiel 6670. (2018, May). Retrieved from http://www.sgg.gov.ma/BO/FR/2018/BO_6670_Fr.pdf?ver=2018-05-14-102617-547

Burke, R. (2010, September). Evaluating the dynamic properties of recommendation algorithms. In *Proceedings of the fourth ACM conference on Recommender systems* (pp. 225-228). ACM.

Ceolin, D., Moreau, L., O'Hara, K., Schreiber, G., Sackley, A., Fokkink, W., ... & Shadbolt, N. (2013). Reliability analyses of open government data.

Chignard, S., & Benyayer, L. D. (2015). *Datanomics. Les nouveaux business models des données*. FYP editions.

Codd, E. F. (1972). *Relational completeness of data base sublanguages* (pp. 65-98). IBM Corporation.

Conseil Economique et Social (CES). (2013). Retrieved from http://www.ces.ma/Documents/PDF/Rapport-AS_14_2013_VF.pdf

Constitution. (2013, July). Retrieved from http://www.maroc.ma/fr/content/constitution-0

Cooper, R., & Kaplan, R. S. (1988). Measure costs right: make the right decisions. *Harvard business review*, *66*(5), 96-103.

Crosby, P. B. (1980). *Quality is free: The art of making quality certain*. Signet.

DAMA UK Working Group. (2009). *The Dama Guide to the Data Management Body of Knowledge*. Technics Pubns Llc.

DAMA UK Working Group. (2013, October). The Six Primary Dimensions for Data Quality Assessment. Retrieved from https://www.dqglobal.com/wp-content/uploads/2013/11/DAMA-UK-DQ-Dimensions-White-Paper-R37.pdf

Dionne P-A., (2015). Analyses des coûts et méthodes d'évaluation économique. Séminaire COPEP. Institut Universitaire en santé mentale de Montréal.

Duffy, G. L. (Ed.). (2013). *The ASQ Quality Improvement Pocket Guide: Basic History, Concepts, Tools, and Relationships*. ASQ Quality Press.

English, L. P. (1999). *Improving data warehouse and business information quality: methods for reducing costs and increasing profits* (Vol. 1). New York: Wiley.

English, L. P. (2003). Total information quality management: A complete methodology for IQ management. *Dm Review*, *9*(03), 7320-1.

English, L. P. (2009). *Information quality applied: Best practices for improving business information, processes and systems*. Wiley Publishing.

English, L. P. (2011, October). Knowledge Articles. Retrieved from https://www.iqint.org/publication2011/doc2/english-2011-10.shtml

Eppler, M., & Helfert, M. (2004, November). A classification and analysis of data quality costs. In *International Conference on Information Quality* (pp. 311-325).

Evans, N., & Price, J. (2012). Barriers to the Effective Deployment of Information Assets: An Executive Management Perspective. *Interdisciplinary Journal Of Information, Knowledge & Management*, *7*.

Federal enterprise architecture framework. (1999). Washington, D.C.: Chief Information Officers Council.

Fisher, C., Lauría, E., & Chengalur-Smith, S. (2012). *Introduction to information quality*. Authorhouse.

Frank, M., & Walker, J. (2016). User Centred Methods for Measuring the Value of Open Data. *The Journal of Community Informatics*, *12*(2).

Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of computational biology*, *7*(3-4), 601-620.

Gartner Group. (2011). Measuring the Business Value of Data Quality. Retrieved from https://www.data.com/export/sites/data/common/assets/pdf/DS_Gartner.pdf

Harper, J. (2012). *Grading the governments data publication practices*. Washington, D.C.: Cato Institute.

Haug, A., Zachariassen, F., & Van Liempd, D. (2011). The costs of poor data quality. *Journal of Industrial Engineering and Management*, *4*(2), 168-193.

Hwang, G. H., & Aspinwall, E. M. (1996). Quality cost models and their application: a review. *Total Quality Management*, *7*(3), 267-282.

Inmon, W. H., Zachman, J. A., & Geiger, J. G. (1997). *Data stores, data warehousing and the Zachman framework: managing enterprise knowledge*. McGraw-Hill, Inc..

International Association for Information and Data Quality. (2013). Retrieved from http://iaidq.org/main/glossary.shtml#I

J. M, Juran. (1951). *Juran's Quality Control Handbook*

Jensen, F. V. (1996). *An introduction to Bayesian networks* (Vol. 210, pp. 1-178). London: UCL press.

Johnson, P. L., & Johnson, P. (2000). *ISO 9000: The year 2000 and beyond*. McGraw Hill Professional.

Juran, J.M., Gryna, F.M. and Bingham, R. (1975). *Quality Control Textbook.* McGraw-Hill, New York

Kim, W., & Choi, B. (2003). Towards Quantifying Data Quality Costs. *Journal of Object Technology*, *2*(4), 69-76.

Kumar, K., Shah, R., & Fitzroy, P. T. (1998). A review of quality cost surveys. *Total Quality Management*, *9*(6), 479-486.

Laney, D. (2000). Customer Data Quality: Part I –A Taxonomy of Troubles.META Group Research, Stamford, Connecticut.

Laney, D. B. (2017). *Infonomics: How to monetize, manage, and measure information as an asset for competitive*. S.l.: ROUTLEDGE.

Laney, D., & Friedman, T. (2012, September 21). Toolkit: Assessing Key Data Quality Dimensions. Retrieved from https://www.gartner.com/doc/2171520/toolkit-assessing-key-data-quality

Lankhorst, M. (2009). *Enterprise architecture at work: Modelling, communication and analysis*. Springer Science & Business Media.

Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: a methodology for information quality assessment. *Information & management*, *40*(2), 133-146.

Loshin, D. (2001). *Enterprise Knowledge Management: The Data Quality Approach*. Morgan Kaufmann.

Knowledge Integrity. (2011, January). Understanding the Financial Value of Data Quality Improvement. Retrieved from https://www.informatica.com/downloads/1530_KnowledgeIntegFinValueDQ.pdf

Madnick, S. E., & Wang, Y. Y. R. (1992). *Introduction to the TDQM research program*. Massachusetts Institute of Technology.

Maqboul, J., & Bounabat, B. (2017, March). Towards a Completeness Prediction Based on the Complexity and Impact. In *International Conference on Information Technology and Communication Systems* (pp. 108-116). Springer, Cham.

McCall, J. A. (1979). An introduction to software quality metrics. In *Software quality management* (pp. 127-142). New York: Petrocelli Books.

McGilvray, D. (2008). *Executing data quality projects: Ten steps to quality data and trusted information*. Amsterdam: Morgan Kaufmann/Elsevier.

Meunier, V. (2009). *Analyse coût-bénéfices : guide méthodologique*. Number 2009-06 of the Cahiers de la Sécurité Industrielle, Institute for an Industrial Safety Culture, Toulouse, France (ISSN 2100-3874). Available at http: //www.icsi-eu.org/francais/dev_cs/cahiers/

Ministère de l'Economie et des Finances. Direction des Domaines de l'Etat (MEF-DDE). (2011). RAPPORT D'ACTIVITE : Direction des Domaines de l'Etat. Retrieved from http://www.abhatoo.net.ma

Ministère de la Fonction Publique et de la Modernisation de l'Administration (MMSP). (2018). Retrieved from https://www.mmsp.gov.ma/fr/decline.aspx?m=5&r=382

Naïm, P., Wuillemin, P. H., Leray, P., Pourret, O., & Becker, A. (2011). *Réseaux bayésiens*. Editions Eyrolles.

Narman, P., Johnson, P., Ekstedt, M., Chenine, M., & Konig, J. (2009, September). Enterprise architecture analysis for data accuracy assessments. In *Enterprise Distributed Object Computing Conference, 2009. EDOC'09. IEEE International* (pp. 24-33). IEEE.

Naumann, F. (2003). *Quality-driven query answering for integrated information systems* (Vol. 2261). Springer.

Nielsen, T. D., & Jensen, F. V. (2009). *Bayesian networks and decision graphs*. Springer Science & Business Media.

Nixon, R. (2012, November 15). U.S. Postal Service Reports $15.9 Billion Loss. Retrieved from https://www.nytimes.com/2012/11/16/us/politics/postal-service-reports-a-nearly-16-billion-loss.html

Object Management Group. (2011, January). Retrieved from https://www.omg.org/spec/BPMN/2.0/About-BPMN/

Object Management Group. (2017, December). Retrieved from https://www.omg.org/spec/UML/About-UML/

OKFN. (2005). Retrieved from https://okfn.org/opendata/

Osimo, D. (2008). Benchmarking eGovernment in the Web 2.0 era: what to measure, and how. *European Journal of ePractice*, *4*, 33-43.

Pearl, J. (2003). Causality: models, reasoning, and inference. *Econometric Theory*, *19*(675-685), 46.

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, *45*(4), 211-218.

Plunkett, J. J., & Dale, B. G. (1988). Quality costs: a critique of some 'economic cost of quality'models. *The International Journal of Production Research*, *26*(11), 1713-1726.

Porter, L. J., & Rayner, P. (1992). Quality costing for total quality management. *International Journal of Production Economics*, *27*(1), 69-81.

Quah, E., & Haldane, J. B. S. (2007). *Cost-benefit analysis*. Routledge.

Ragsdale, C. (2014). *Spreadsheet modeling and decision analysis: A practical introduction to business analytics*. Nelson Education.

Rast, C. (2015, October). D&A: A new driver of performance and valuation. Retrieved from https://home.kpmg.com/xx/en/home/insights/2015/06/data-new-driver-of-performance.html

Redman, T. C., & Blanton, A. (1997). *Data quality for the information age*. Artech House, Inc..

Reiche, K. J., & Höfig, E. (2013, July). Implementation of metadata quality metrics and application on public government data. In *Computer Software and Applications Conference Workshops (COMPSACW), 2013 IEEE 37th Annual* (pp. 236-241). IEEE.

Reid, R. D., & Sanders, N. R. (2005). *Operations management: an integrated approach*. Hoboken, NJ: John Wiley.

Ross, J. W., Weill, P., & Robertson, D. (2006). *Enterprise architecture as strategy: Creating a foundation for business execution*. Harvard Business Press.

Scannapieco, M., & Catarci, T. (2002). Data quality under a computer science perspective. *Archivi & Computer*, *2*, 1-15.

Scannapieco, M., Virgillito, A., Marchetti, C., Mecella, M., & Baldoni, R. (2004). The DaQuinCIS architecture: a platform for exchanging and improving data quality in cooperative information systems. *Information systems*, *29*(7), 551-582.

Socrata. (2016, October). 2016 Open Data Benchmark Report. Retrieved from https://socrata.com/webinar/2016-open-data-benchmark-report/

Spewak, S. H., & Hill, S. C. (1993). *Enterprise architecture planning: developing a blueprint for data, applications and technology*. QED Information Sciences, Inc..

Stott, A. (2014, July 23). Open data for economic growth. Retrieved from http://documents.worldbank.org/curated/en/131621468154792082/Open-data-for-economic-growth

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. *Communications of the ACM, 40*(5), 103-110. doi:10.1145/253769.253804

Sunlight Foundation. (2007). Ten Principles For Opening Up Government Information. Retrieved from https://sunlightfoundation.com/policy/documents/ten-open-data-principles/

Sunlight Foundation. (2014). Open Data Policy Guidelines. Retrieved from https://sunlightfoundation.com/opendataguidelines/

The Guardian. (n.d.). MPs' expenses | Politics. Retrieved December 25, 2018, from https://www.theguardian.com/politics/mps-expenses

The ODI. (n.d.). The mark of quality and trust for Open Data. Retrieved December 25, 2018, from https://certificates.theodi.org/en/

The Texas Tribute. (n.d.). Government Salaries Explorer. Retrieved December 25, 2018, from https://salaries.texastribune.org/

The TOGAF® Standard - Version 9.2. (n.d.). Retrieved December 26, 2018, from https://www.opengroup.org/togaf

The World Bank. (2018, December 06). Menu Open Government Data Toolkit. Retrieved from http://opendatatoolkit.worldbank.org/en/

Tim Burners-Lee. (2015). 5 ★ Open Data. Retrieved from https://5stardata.info/en/

USDA. (2006). The USDA Enterprise Architecture Program. Retrieved from https://www.fsa.usda.gov/online-services/sdlc/enterprise-architecture-program/index

W3C. (2014). Data Catalog Vocabulary (DCAT). Retrieved from https://www.w3.org/TR/vocab-dcat/

Wang, R. Y. (1998). A product perspective on total data quality management. *Communications of the ACM*, *41*(2), 58-65.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, *12*(4), 5-33.

Wang, R., Storey, V., & Firth, C. (1995). A framework for analysis of data quality research. *IEEE Transactions on Knowledge and Data Engineering, 7*(4), 623-640. doi:10.1109/69.404034

Zachman, J. A. (1997). The Zachman Framework: a primer for enterprise engineering and manufacturing. *Zachman International*.

Zaveri, A., Kontokostas, D., Sherif, M. A., Bühmann, L., Morsey, M., Auer, S., & Lehmann, J. (2013, September). User-driven quality evaluation of dbpedia. In *Proceedings of the 9th International Conference on Semantic Systems* (pp. 97-104). ACM.

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality assessment for linked data: A survey. *Semantic Web*, *7*(1), 63-93.

Zhu, H., Madnick, S. E., Lee, Y. W., & Wang, R. Y. (2014). Data and Information Quality Research: Its Evolution and Future.

Zopounidis, C., & Doumpos, M. (2002). Multicriteria classification and sorting methods: a literature review. *European Journal of Operational Research*, *138*(2), 229-246.

Annexes

Annexe A – The Business Value of Data Quality Projects

G00218962

Measuring the Business Value of Data Quality

Published: 10 October 2011

Analyst(s): Ted Friedman, Michael Smith

Research shows that 40% of the anticipated value of all business initiatives is never achieved. Poor data quality in both the planning and execution phases of these initiatives is a primary cause. Poor data quality also effects operational efficiency, risk mitigation and agility by compromising the decisions made in each of these areas.

Key Findings

- Poor data quality is a primary reason for 40% of all business initiatives failing to achieve their targeted benefits.
- Data quality effects overall labor productivity by as much as a 20%.
- As more business processes become automated, data quality becomes the rate limiting factor for overall process quality.

Recommendations

Business leaders and IT leaders focused on data quality improvement and information governance should:

- Measure the business value of improved data quality by focusing on business processes, investment decisions and overall productivity.
- Qualify the business value of improved data quality using business metrics that are correlated with financial outcomes.
- Share the findings of this research with your finance department to receive feedback and guidance on where to begin identifying opportunities for increased business value from improved data quality.

Table of Contents

List of Figures

Analysis

Organizations are discovering that data quality deficiencies have a significant impact on their most strategic business initiatives, often holding them back from achieving the growth, agility and competitiveness that they desire. In addition to challenges with growth and agility, compliance and transparency pressures increasingly bring data quality issues to the fore — it is no longer acceptable to ignore flaws in data, and organizations must prove the accuracy of information that they report to auditors, regulators and the public. Given the current economic challenges and budgetary pressures facing most organizations, there is a substantial desire to eradicate quality issues in data as a way to reduce costs and improve efficiency (see "Findings From Primary Research Study: Organizations Perceive Significant Cost Impact From Data Quality Issues").

Historically, a minority of organizations have taken a proactive approach to managing data quality — the majority have endured the pains of poor-quality data and dealt with it in a reactive manner. However, with recognition of the impact of poor-quality data increasing as a result of all of these contemporary business drivers, many are undertaking improvement efforts within the context of individual projects or occasionally on a broader, enterprisewide basis.

They can address these challenges through the discipline of data quality, which includes the approaches, organizational models, techniques and technologies used to assess and ensure the "fitness for purpose" of all types of data for use by the various applications and business processes across the enterprise (see "Key Issues for Data Management and Integration Initiatives, 2010"). While technology plays a key role in data quality improvement, changes in work processes and behavior of people are critical.

The demand for data quality best practices is maturing — more organizations are focused on fine-tuning their efforts by increasing the precision with which they measure and monitor data quality, the breadth of their scope (looking across multiple domains, processes and groups) and creating data quality-specific roles and organizational structures. Tools for measuring, monitoring, and

improving data quality can help enterprises improve effectiveness and deploy controls in approaching data quality challenges.

How can CIOs and CFOs quantify the impact on their organizations? Each organization is different, but many different examples can be seen across a variety of industries (see Note 1). CIOs can use similar examples that are relevant to their organization to quantify the financial impact of improving data quality and gain buy-in from business leaders in sales, marketing, customer service and other parts of the organization. One of the new realities of the global economic environment is the desire of business executives to manage risk more effectively. This has created the need to shift from "gut-feel" to fact-based decision making (see "Survey Analysis: How Executives Use Business Metrics"). Gartner conducted a survey of business executives before and after the financial crisis of 2008. The results showed a 30% increase in the use of business metrics as the basis for setting strategic direction. So what is the measurable impact of data quality?

The Impact of Data Quality on Business Processes

The effects of data quality on business processes can be estimated based on Six Sigma. Harry and Schroeder estimate that the average successful business process maintains 3.5 sigma value (see Note 2). This equates to 22,800 defects per million. At this rate, the average cost of quality is 20% of the overall business process costs. For example, if the business process were sales and a company spent $200 million to maintain its sales process throughout the year, then at 3.5 sigma the cost of quality would be $40 million annually. In the past 10 years data quality has begun to set the basis for process quality.

Business processes are becoming digital. Over the past 10 years, information technology has been used to automate transaction that go well beyond accounting. Automation of sales, product development, manufacturing, and recruitment processes, to name a few, is expanding each year to lower costs. When processes are automated to the maximum extent possible, data quality issues become the limiting factor in maximizing process quality.

Consider the following example. A $10 billion consumer goods distribution company recently implemented a sales force automation system. All leads, qualified leads, sales resources, and campaign strategies are recorded in the system. The quality of prospect and product data are critical to the overall quality of the sales process. Data quality defects become process quality defects as mistakes are made when data is wrong. It does not take much. To maintain average process quality of 3.5 sigma, data quality must remain below 22,800 defects per million, or 97.72%, in highly automated processes such as automated material handling systems. The 20% cost of quality results from flaws in data. In effect, data quality becomes the limiting factor in process quality.

The Impact of Data Quality on Productivity

The top half of Figure 1 shows the impact of information technology on overall labor productivity in the U.S. for the past 10 years. The Bureau of Labor Statistics estimates that 70% of the annual productivity growth of 2.7%, or 1.9%, comes from IT. This is the good news; the bad news is

shown on the bottom half of Figure 1. The productivity benefits are not shared equally; in fact, there exists substantial variation. Why is it that one enterprise can benefit while others do not?

Figure 1. The Productivity Benefits From IT

"IT is correlated with productivity, but there are substantial variations among companies."

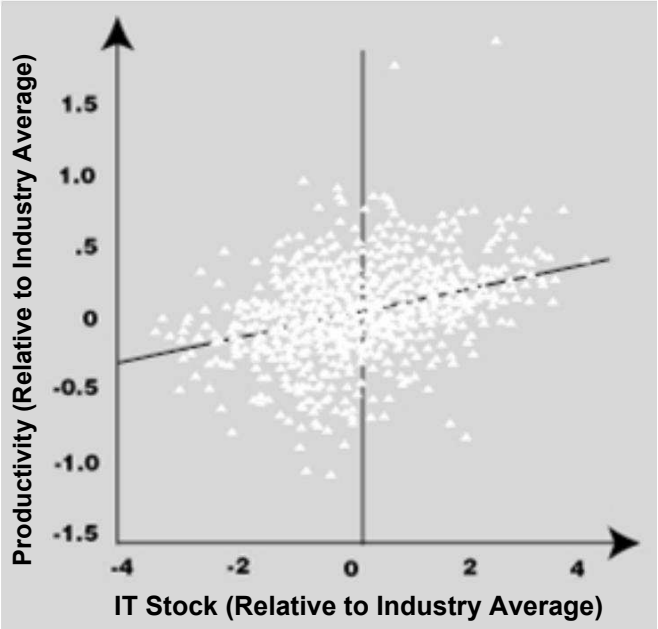
http://ebusiness.mit.edu/erik/Optimize/pr_roi.html

IT stock is measured as the current replacement cost of IT hardware stock per worker.

Productivity is measured as real output divided by a weighted average of all inputs, including labor and non-IT capital.

Base: 1,167 companies

Data Erik Brynjolfsson and Lorin Hitt



Source: Gartner (October 2011)

One reason is poor data quality. Through client interactions and studies, Gartner continues to observe a high correlation between significant data quality issues and degraded productivity of key resources. A survey of business executives in your company can be used to estimate the impact of poor data quality on labor productivity.

The average 1.9% increase in labor productivity from IT saves approximately $21.7 million for the average enterprise. If poor data quality affects 10% of this savings, the result is an annual $2.2 million loss on that saving.

Average Enterprise:

Annual Revenue: $2,857 billion

Labor Costs: $1,143 billion

Average Annual IT Productivity: 1.9%

Benefit From IT Productivity: $21.7 million

10% of IT Productivity From Data Quality: $2.12 million

The Impact of Data Quality on Decision Making

Our research, and that of the Standish Group, shows that 40% of projects succeed (achieve targeted benefits), 40% are challenged (achieve half the benefit) and 20% fail completely (no benefit). This results in an average 60% benefit realization.

Benefit realization rate = (40% x 100%) + (40% x 50%) + (20% x 0%) = 60%.

This means that 40% of the benefits targeted in the annual capital budget are never realized. There are many reasons for projects not achieving targeted benefits, but consider the impact of data quality on the assumptions and projections made. One way to assess this in your company is to take the list of approved capital projects for the current year. Ask the members of the approving authority (capital expenditure committee) how accurate they believe the data was that was used to make the assumptions and projections in the business case for each project. Also ask them for their estimate of the impact that the level of data quality will have on benefit realization. A 10% impact of data quality on benefit realization would mean $2.1 million per year for the average enterprise.

Average Enterprise:

Annual Revenue: $5 billion

IT Spend: $175 million (3.5% of revenue)

IT Capex: $52.5 million (30% of total IT spend)

Benefit Realization: 60% (weighted average for benefit realization, see above)

Lost Benefits: 40% x $46.2 million = $21 million

10% Impact From Data Quality: $2.1 million

Process for Measuring the Specific Business Value of Data Quality

Measuring the impact of data quality on business processes and capital investments requires that we use both accounting and non-accounting metrics. Accounting metrics are required by various regulators around the world for reporting the financial position of an enterprise. This is true of both public and private sector organizations. They are quite useful for comparison purposes because the definitions are clear and unambiguous.

Much of the value of information technology, like the value of data quality, is consumed in building and growing branding, intellectual property, unique business processes and knowledge. Yet this value cannot be measured with accounting metrics. Recognizing this measurement gap, Gartner spent several years developing a set of extensions to accounting metrics called the Gartner Business Value Model (see "The Gartner Business Value Model: A Framework for Measuring Business Performance"). The metrics in the model measure activities and events that precede accounting results and are correlated with actual accounting results. Each metric in the model is mathematically linked with an accounting metric in either the balance sheet or the income statement. This is how "buy-in" with senior business executives, particularly financial executives, is

achieved. By showing these executives how the non-accounting metrics are correlated to and have a causal relationship with accounting results, the extension becomes part of the lexicon for business value.

Using extensions like those defined in the Gartner Business Value Model (GBVM), we can more precisely measure the value of data quality. Figure 2 is a graphical representation of the model.

Figure 2. The Gartner Business Value Model

Business Aspect	Desired Business Outcome	Key Risk Indicators			
Demand Management	Market Responsiveness	Target Market Index	Market Coverage Index	Market Share Index	Opportunity/Threat Index
		Product Portfolio Index	Channel Profitability Index	Configureability Index	
	Sales Effectiveness	Sales Opportunity Index	Sales Cycle Index	Sales Close Index	Sales Price Index
		Cost of Sales Index	Forecast Accuracy	Customer Retention Index	
	Product Development Effectiveness	New Products Index	Feature Function Index	Time-to-Market Index	R&D Success Index
Supply Management	Customer Responsiveness	On-Time Delivery	Order Fill Rate	Material Quality	Service Accuracy
		Service Performance	Customer Care Performance	Agreement Effectiveness	Transformation Ratio
	Supplier Effectiveness	Supplier On-Time Delivery	Supplier Order Fill Rate	Supplier Material Quality	Supplier Service Accuracy
		Supplier Service Performance	Supplier Care Performance	Supplier Agreement Effectiveness	Supplier Trans-formation Ratio
	Operational Efficiency	Cash-to-Cash Cycle Time	Conversion Cost	Asset Utilization	Sigma Value
Support Services	Human Resources Responsiveness	Recruitment Effectiveness Index	Benefits Administration Index	Skills Inventory Index	Employee Training Index
		HR Advisory Index	HR Total Cost Index		
	Information Technology Responsiveness	Systems Performance	IT Support Performance	Partnership Ratio	Service Level Effectiveness
		New Projects Index	IT Total Cost Index		
	Finance & Regulatory Responsiveness	Compliance Index	Accuracy Index	Advisory Index	Cost of Service Index

Source: Gartner (October 2011)

The following 10 steps summarize the process for measuring the business value of data quality.

1. Choose an area within your enterprise where data quality is an issue. Using the GBVM, sales might be an example.
2. Engage the management team in that area to explore the potential benefits that improvements in data quality may have. Include a representative from finance in the discussion, someone familiar with the business area being explored. Explain this process to these stakeholders and ask for their involvement as outlined below.
3. Select between two and six metrics from that sales area that are leading indicators of revenue. The GBVM is ideal for this purpose.
4. Baseline the current performance levels, using the definitions, for each metric selected.
5. Discuss the current state of data quality in the area being explored, and the expectations for data quality held by business stakeholders. Current state should be assessed across a range of data quality dimensions, such as validity, completeness, consistency and accuracy. Also discuss how this level of data quality affects decision making in the area being explored. Estimate the impact of various levels of improvements to data quality on the metrics selected. These are estimates and assumptions will need to be documented.
6. Determine the value of improvements in data quality by converting improvements in performance of the metrics selected using the financial sensitivity calculations. The financial sensitivity calculations were developed as part of the research for the GBVM and must be validated with the financial representative on your team. This has been done hundreds of times with financial representatives at Gartner clients. A common result is for the representative to tweak the calculation based on specific characteristics found in the enterprise, but acceptance of the calculation is achieved.
7. Determine the cost for achieving the targeted improvements in data quality with the stakeholder. Information technology professionals usually take the lead on this step.
8. Build the business case for improvements in data quality using the benefits determined in step 6 and the costs determined in step 7. Be sure to include all assumptions.
9. Present the business case with the value quantified to the approving authority. Be sure to have all stakeholders with you.
10. Implement the solution, including a change management focus to ensure data quality controls, metrics, and the roles participating in data quality improvement evolve over time as business requirements evolve.

Example of Measuring Business Value of Data Quality

The following is a simple example of applying the above process to data quality concerns in a sales process. While not relevant for all organizations in all industries, this example is meant to demonstrate the basic steps and principles of applying the GBVM concepts to identifying and addressing data quality issues in a typical enterprise. The figures cited in the example are fictitious, but are representative for efforts of this type.

1. For this example, we will focus on the sales area. Sales growth is an important objective, and many organizations will want to link their data quality focus to this objective.
2. The key stakeholders that must participate in this effort will be IT leaders focused on data quality, executives responsible for the sales organization and process, and a representative from finance.
3. After general discussion, the stakeholders agree that sales opportunity index (SOI), forecast accuracy and customer retention are the key leading indicators of sales growth. These will be the metrics that will be used to measure the business value of data quality.
4. The baseline for current performance, using the definitions of these sales-related metrics in the GBVM was:
   - SOI = 0.1
   - Forecast accuracy = 0.40
   - Customer retention = 0.70
5. The team then identified specific ways in which data quality issues negatively impacted these key metrics. Their main findings showed:
   - SOI was weak due to poor quality of prospect data (demographics, credit history, past purchase history, and so on) which caused fewer leads to be qualified than expected.
   - Forecast accuracy was low due to poor quality of inventory data (for example, duplicate stock-keeping unit numbers) and incomplete/missing sales history data.
   - Customer retention was low because a lack of quality in customer returns data inhibited the ability to identify dissatisfied customers.
6. In order to determine the business impact of these data quality issues, the team then used specific data quality dimensions to quantify the magnitude of data quality flaws affecting each metric.
   - Related to SOI, the quality of prospect data was measured to be 50% complete (based on the definition of completeness requiring valid values for 10 specific attributes about a prospect) and 30% accurate (in many cases, the revenue and employee size figures for a prospect were measured to be inaccurate).
   - Related to forecast accuracy, the product master was found to have 10% duplicate records. Combined with the impact of some missing sales history data, this contributed to an overall quality level of 70% for key data used in forecasting.
   - Related to customer retention, only 60% of returns had accurate customer details, prohibiting the other 40% from being linked to customer accounts.
   - Using the financial sensitivity calculations in the GBVM, the finance representative determined that these data quality flaws created a loss of $1 million of sales growth opportunity.

7. The team estimated that the cost for resolving these data quality issues would be $300,000. This was based on the expected time and effort to implement required changes to existing systems, as well as for training of pertinent individuals involved in the related business processes.
8. Using the costs and benefits identified, the finance representative determined that there was clearly a solid business case for implementing these changes.
9. The team then presented the business case to the approving authorities and received funding and support to execute the effort.
10. Led by IT, modifications to various sales and inventory systems were made to introduce improved data quality controls. This eliminated the possibilities of the identified data quality issues from occurring in the future. In addition, sales and inventory personnel were trained in data quality best practices and given specific work tasks to continually monitor and validate data quality levels. Ongoing, the team will measure both the specific data quality metrics and the key sales growth metrics to ensure the changes are having a positive impact.

Bottom Line

Data quality issues are a perfect example of the business cliché, "You can't manage what you don't measure." Because most organizations do not make the effort to measure the quality of their data in any objective or quantitative way, they fail at building a business case for formal data quality improvement efforts, expend energy on problems that are perceived rather than real, and never have a clear understanding of whether their efforts are making a difference. Even worse, organizations make (often erroneous) assumptions about the state of their data and continue to experience inefficiencies, excessive costs, compliance risks and customer satisfaction issues as a result. In effect, data quality in their business goes unmanaged.

A metrics-based approach to assessing data quality helps remove the assumptions, politics and emotion often associated with this issue, thereby giving organizations a factual basis on which to justify, focus and monitor their efforts. In addition, the identification, communication and analysis of metrics on an ongoing basis provides a tangible indication to the organization that the data quality issue is important to the business. With this solid foundation of facts and focus, rather than perceptions and apathy, organizations can truly begin to manage the quality of their data and will reap significant benefits as a result.

Recommended Reading

Some documents may not be available as part of your current Gartner subscription. "The

Gartner Business Value Model: A Framework for Measuring Business Performance"

"Findings From Primary Research Study: Data Quality Issues Create Significant Cost, Yet Often Go Unmeasured"

"Strategic Focus on Data Quality Yields Big Benefits for BT"

"Case Study: Aera Energy's Comprehensive Focus on Data Quality Generates Competitive Advantage"

"Case Study: Smith & Nephew Focuses on Data Governance as First Step in MDM Program"

Note 1 Example of Quantifying the Impact of Data Quality

By decreasing the amount of returned mail by 10%, a healthcare plan with 500,000 providers can realize $400,000 in savings over three years (source: "The importance of data quality in producing savings," Healthcare Finance News). SiriusDecisions, a sales and marketing research firm, quantifies data quality using the 1-10-100 rule, which says "It takes $1 to verify a record as it's entered, $10 to cleanse and de-dupe it and $100 if nothing is done, as the ramifications of the mistakes are felt over and over again." On the revenue side of the equation, a data quality strategy and targeted data quality improvement efforts that solve conflicts at the source can lead to a 25% increase in converting inquiries to marketing-qualified leads (source: "Data Quality Practices Boost Revenue by 66 Percent," destinationCRM.com).

Note 2 Six Sigma

M. Harry and R. Schroeder, "Six Sigma: The Breakthrough Management Strategy Revolutionizing the World's Top Corporations," Doubleday Business, January 2000.

Regional Headquarters

Corporate Headquarters

56 Top Gallant Road
Stamford, CT 06902-7700
USA
+1 203 964 0096

Japan Headquarters

Gartner Japan Ltd.
Aobadai Hills, 6F
7-7, Aobadai, 4-chome
Meguro-ku, Tokyo 153-0042
JAPAN
+81 3 3481 3670

European Headquarters

Tamesis
The Glanty
Egham
Surrey, TW20 9AW
UNITED KINGDOM
+44 1784 431611

Latin America Headquarters

Gartner do Brazil
Av. das Nações Unidas, 12551
9° andar—World Trade Center
04578-903—São Paulo SP
BRAZIL
+55 11 3443 1509

Asia/Pacific Headquarters
Gartner Australasia Pty. Ltd.
Level 9, 141 Walker Street
North Sydney
New South Wales 2060
AUSTRALIA
+61 2 9459 4600

Annexe B – Understanding the Financial Value of Data Quality Improvement

Understanding the Financial Value of Data Quality Improvement

Prepared by:

David Loshin
Knowledge Integrity, Inc.
January, 2011

Sponsored by:

Introduction

Despite the many years of lip service paid to high quality data, the difficulty in establishing quantifiable value often relegates data quality improvement to the eleventh spot on the CEO's list of top ten business imperatives. And as opposed to the technical aspects of data validation and cleansing, it appears that the biggest challenge is effectively communicating the business value of data quality improvement. Yet with a well-defined process for considering the costs and risks of low-quality data in relation to an iteratively-refined set of business impact categories not only provides a framework for putting data quality expectations into a business context, it also enables the definition of clear metrics linking data quality to business performance.

As an example, it is often suggested that data errors prevent the sales team members from up-selling and cross-selling products, and this claim is used to justify the need for a data quality improvement effort. However, a more comprehensive quantification of the number of sales impacted or of the total dollar amount for the missed opportunity is much more effective at showing the value gap, especially when it can be directly associated with specific data flaws.

The communication gap between technical analysts and business managers often impedes the articulation of the value of data quality improvement. This article looks at a hierarchical categorization of the financial dimension of business value drivers, corresponding performance measures, linking those measures to specific data issues, and a process for evaluating the relationship between acceptable performance and quality information. This article is targeted to those technical analysts who seek to understand the connection between information and optimal business performance so that the case for data quality improvement can be rooted in quantifiable measures.

The Value of High Quality Data

Both operational and analytical business applications rely on high quality data. And those organizations that lack processes for identifying and managing data quality issues introduce risks to the usability and trustworthiness of the data upon which many applications depend, leading to the types of negative financial impacts we describe in this article. This suggests that there is value in instituting processes for assessing, measuring, reporting, reacting to, and controlling different aspects of poor data quality. Data is an asset that is created or acquired and then repurposed multiple times, and process flaws introduce risks to successfully achieving business objectives. The dynamic nature of data adds to the challenges in establishing ways to assess risks as well as ways to monitor conformance to business user expectations.

Are Anecdotes Enough?

Resorting to specific anecdotes and examples of business problems linked to bad data may raise awareness temporarily, but this is no substitute for demonstrating real evidence of hard impacts. This is common in organizations that "institutionalize reactivity" by regularly engaging staff members to correct data errors when their catastrophic impacts have already occurred, instead of proactively assessing data quality assessment, performing root cause analysis, and eliminating the sources of the errors. The first step begins with developing a performance management framework that helps to identify, isolate, measure, and improve the value of data within the business contexts, which requires

- Correlating business impacts with data failures and then
- Characterizing the loss of value that is attributable to poor data quality.

This means reviewing the types of risks and costs relating to the use of information and categorizing the business impacts as a prelude to asserting data quality expectations and metrics. Business issues are directly tied to missed data quality expectations, and the framework described in this article explores categories of business impacts that may be rooted in poor data quality. By identifying and classifying business impacts and establishing the connection to reliance on high quality data, technical analysts gain the tools for effectively communicating the value of improved data quality to key stakeholders in the organization.

Financial Impacts

There are essentially two sides of the coin when evaluating financial impacts, both ultimately connected to increased profitability: increasing revenues while decreasing costs. The process of soliciting, categorizing, and measuring financial impacts involves these basic steps:

1. Clarifying and prioritizing the financial business expectations, and working with the business clients to understand where the financial expectations are not being met;
2. Finding specific examples where failed expectations has lead to known impacts;
3. Categorizing the business impact at a fine-enough level of granularity in a way that ensures that the impacts can be measured;
4. Formalizing the dependence on specific data sets and data errors;
5. Researching the history of occurrence, probability of recurrence, and cumulative impacts; and
6. Understanding and then eliminating the root causes.

Organizational Objectives

Even in organizations whose leaders have articulated well-defined high-level strategic goals, the role each staff member plays is often lost in translation. Therefore to ensure that the information technology and business function teams are aligned, it is important for the data quality analyst to establish a good rapport with business data consumers to

- Identify those business processes that are dependent on high quality data;
- Understand how those business processes use data; and
- Solicit the data consumers' expectations in a way that can be translated into defined business rules for data validity.

Communications Gap

Unfortunately, it is much easier to identify gaps in meeting financial expectations than to quantify them in measureable terms, especially in relation to their dependence on data. This is complicated even more by the communications gap that often exists between the business and technical teams. The second step can be simplified by providing a framework for categorizing business impacts at a level of granularity that can be discretely evaluated and measured. Developing a taxonomy of financial impact

classifications enables the technical analysts to map identified business issues into a specific category, which is more likely to be linked to specific data sets and their expectations for data quality.

Once the most critical impacts have been identified and linked to associated data flaws, the analysts can define data quality rules that reflect the business user expectations. Continuous inspection of conformance to formally-defined data quality rules not only provides discrete quantification of the scale of existing data problems, its results can contribute to a performance scorecard linking data quality to specific financial value.

To continue our earlier example, revenue generation may be negatively impacted through missed up-selling and cross-selling opportunities. If this is caused by inability to have complete customer visibility as a result of duplicate entries in the customer database, then measuring the number of duplicates provides a good indicator that sales opportunities are being missed. More comprehensive analysis can link the specific number of duplicates to one or more missed sales, completely closing the loop for defining business-oriented data quality rules.

The following sections provide some examples of this classification and categorization for two high-level categories of financial impacts (revenue growth and decreasing costs), subcategories, and descriptions of potential business measures. For each set of impact areas, examining the corresponding measures and considering their dependence on data will help the analyst to link the financial impact to high quality data.

Revenue Growth

One might say that the primary goal of any business is profitability, but it would be difficult to achieve profitability without generating revenues. There are many areas associated with growing revenue, and this section looks at four:

- Customer acquisition
- Customer retention
- Leveraging income-generating opportunities
- Cross selling and up-selling

Customer Acquisition

Companies rely on a community of customers to purchase their products and services. Table 1 provides examples of areas of impact and corresponding measures for new customer acquisition.

Area of Impact	Measures
Customer segmentation	Effectiveness of targeting individuals for customer acquisition, retention, and breadth of geographic and demographic coverage
Quality of sales leads	The degree to which sales lead data records contain complete and accurate contact information
Meeting sales targets	Determining whether sales team members are meeting or missing their sales targets
Sales channel partner effectiveness	The degree to which channel partners are successful in promoting and selling to new customers

Area of Impact	Measures
Sales channel effectiveness	Comparisons to gain of market share by competitors through the same and/or alternate channels
Sales cycle time	Monitoring how quickly leads and prospective customers progress through the stages of the sales cycle
Qualification rate	Determining the percentage of leads that can be qualified as prospective customers
Sales closure rate	Determining the percentage of prospective customer that commit to purchasing a product or service
Time spent selling	Measuring the percentage of sales team member time is spent selling

Table 1: Areas of business impact for new customer acquisition.

Customer Retention

Once the company has acquired the customer, there is a continuous drive to ensure that the customer's business is retained rather than allow the customer to purchase products or engage services from the company's competitors. Table 2 provides examples of areas of impact and corresponding measures for customer retention.

Area of Impact	Measures
Focus on high value customers	Ability to calculate customer value
Lifecycle transitions	Employing knowledge of customer life cycle events and transitions for proactive retention
Loyalty management	Enrollment in loyalty programs, allocation of loyalty rewards, and redemption of loyalty rewards
Response to customer needs	Degree of responsiveness to customer requests, inquiries, service needs
Customer portfolio management	How effective is the organization at culling out those customers that are not profitable?
Attrition	Churn rates, effectiveness in making barriers to defection
Contractual compliance	How effective is the company at monitoring adherence to contractual details
Feedback and satisfaction	Response to requests for feedback, measured customer satisfaction

Table 2: Areas of business impact for customer retention

Cross Selling and Up-Selling

Once a prospect has transitioned into being a customer, there is a desire to encourage the customer to buy additional products and services. Cross selling is a strategy for suggesting products that are complementary to the ones already purchased, such as offering french fries along with the purchased hamburger or a set of speakers with the purchase of an audio receiver. Up-selling is a strategy to entice the customer to increase the scale of size of the purchase, such as offering a larger-sized soda or an upgraded version of the audio receiver. Table 3 provides examples of areas of impact and corresponding measures for cross selling and up-selling.

Area of Impact	Measures

Area of Impact	Measures
Channel effectiveness	Measure effectiveness of advertising media channels, partner sales, and other sales channels
Customer touch points	Measure customer experience across all touch points including purchase, delivery, support, and customer service
Identifying cross sell and up-sell opportunities	Analysis that exposes opportunities for cross selling and up-selling
Instituting cross selling and up-selling processes	Developing approaches, campaigns, processes, and training for cross selling and up-selling
Process effectiveness	Measuring and improving effectiveness of cross selling and up-selling processes
Staff effectiveness	Evaluate staff effectiveness and properly incentivizing success

Table 3: Areas of business impact associated with cross selling and up-selling.

Leverage Income-Generation Opportunities

Alternatively, organizations often control existing assets that can be leveraged into generating additional income. Table 4 provides examples of areas of impact and corresponding measures for making use of existing assets to generate additional income.

Area of Impact	Measures
Exploiting existing assets	Sell existing assets whose values have appreciated, lease or sell extra capacity to other organizations
Intellectual property	Sell or license intellectual property
Cash reserves	Improve investment income from existing funds
Manage taxes	Manage and/or defer federal, state, and local tax payments

Table 4: Areas of business impact associated with leveraging income-generating opportunities.

Decreasing Costs

The other side of profitability is reducing the expenses paid for operations. Again, there are many areas in the business in which costs are incurred, but this section provides some details regarding these areas:

- Overhead and administrative costs
- Cost of goods sold
- Fees and charges

Overhead and Administrative Costs

Almost every organization utilizes a physical facility and incurs costs associated with running the business such as rent and facility maintenance. Even virtual organizations have overhead and administrative costs such as telephones, internet, furniture, hardware, and software purchase/leasing and maintenance. Table 5 provides examples of areas of impact and corresponding measures for overhead and administrative costs.

Area of Impact	Measures
Rent	Costs of rented office space

Area of Impact	Measures
Maintenance	Costs associated with building, furniture, machinery, software, and grounds maintenance
Asset purchase and licensing	Costs associated with purchasing or leasing assets and equipment, and the effort for choosing one option over the other
Utilities costs	Costs for telecommunications, energy, water, gas, etc.
Administrative staff	Number of staff members dedicated to administrative activities, percentage of time spent in overhead and administrative activities
Office supplies	Cost and use of paper, pens, notebooks, etc.
General procurement	Costs associated with spend and procurement processes

Table 5: Areas of business impact associated with overhead and administrative costs.

Cost of Goods Sold

The cost of goods sold (often referred to as COGS) comprises the costs and expenses associated with the production and sales of products, including the costs of materials, the direct labor expenses incurred in manufacturing, and the direct labor expenses incurred in selling the products. Table 6 provides examples of areas of impact and corresponding measures for the cost of goods sold.

Area of Impact	Measures
Product design	Staff and materials costs for designing new products
Raw materials	Costs of acquiring, storing, and using raw materials
Cost of production	Costs of manufacturing and finishing products
Sales staff base costs	Base salaries paid to sales account executives
Product quality	Percentage of manufactured items within acceptable quality guidelines

Table 6: Areas of business impact associated with the cost of goods sold.

Fees and Charges

For many of the services critical in running a business (such as banking , legal, and accounting) there are fees and charges incurred. Table 7 provides examples of areas of impact and corresponding measures for fees and charges.

Area of Impact	Measures
Bank fees and service charges	Bank service fees, transaction fees, interest charges, missed payment fees
Legal fees	Costs paid for legal services
Accounting fees	Costs associated with maintaining accurate financial accounts
Document fees	Fees and charges associated with application and document filing
Commissions	Payments exceeding purchase costs paid to agents or brokers
Bad debt	Costs associated with payments and debt that is difficult to collect or is uncollectable
Penalties and fines	Penalties and fines for noncompliance of regulations, penalties associated with failure to observe agreements
Merger and acquisition costs	Costs associated with process of merging operations and application systems

Table 7: Areas of business impact associated with fees and charges.

Formalizing the Dependence on Data: An Example

In order to demonstrate the process, let's look at one example from our categories, sales channel partner effectiveness. The example measure is the degree to which channel partners are successful in promoting and selling to new customers. There are at least two data sets associated with the example measure: the set of channel partners and the set of prospective customers. The analyst must consider the types of errors that may exist in those data sets and in turn, how those types of errors might lead to decreased success in channel partner success. Examples might include missing or incorrect prospect data, incomplete or in accurate partner data, or incorrect tracking of channel sales transactions.

Researching Impacts and the Value Gap

The fourth step of the process is researching the history of occurrence, probability of recurrence, and cumulative impacts. There is a story behind each perceived business problem, and the analyst's goal is to ask the right kinds of questions to fully assess the level of criticality. For example, for each business problem, the analyst might ask the business stakeholders these types of questions:

- What makes this a critical business problem?
- What are the measurable impacts?
- How is each impact classified?
- How is the impact measured?
- Is there a data management behavior that we are looking to influence to achieve better results?

The answers reveal the aspects of the problem that are used to determine the scope of its impact, its prevalence, its frequency, and its probability of occurrence. Together, these variables can be used to prioritize the business problems and direct the second phase of the analysis to assess the relationship to flawed data. For each business problem, the analyst can ask questions such as these:

- How is the business problem related to an application data issue?
- How often does the data issue occur?
- When the data issue occurs, how is it manifested within the business process?
- Who are the individuals who recognize the existence of the problem?
- How often is the data issue identified before the business impact is incurred?
- What rules and metrics can be defined to validate the quality of the data?
- Have we seen this happen before, either internally or in the market? If so, can we directly link the data issue to specific business impact?

Documenting the quantifiers associated with the business impact may involve some detective work – perhaps reviewing change control logs, requests for business process modifications, or requests for changes to underlying data models and application code. The requests often correlate with known business issues and will point to quantifiable measures for the impacts. Linking those measures to specific data validity metrics provides the hard link between business impact and flawed data.

Establishing the Value of Data Quality Improvement

The last step in developing the business case involves understanding the root causes of the issues and determining how those issues can be addressed. This typically means reviewing the information flow from the data creation point through the business process to determine where in the process the error was introduced. Many data quality issues are due to process failures, so correcting the process where the error is introduced will lead to greater improvement than correcting bad data downstream. With enough research into the granularity of the problem and its relationship to flawed data, one might even be able to amortize the cumulative costs related to the business impacts over the number of times an error occurs factored within the probability that the error will occur, thereby providing an estimated cost of each error.

Once the source of the introduction of the error is identified, the data analyst can consider alternatives for eliminating the root cause, instituting preventative techniques, and/or taking some corrective action. Each, or perhaps all, of these alternatives require an investment of both money and resources for the acquisition of any appropriate technologies, staffing for designing, developing, and implementing solutions, training, and ongoing maintenance of the solution. Providing a conservative estimate at this point establishes a baseline cost for remediation. At this point we have an estimate of the cost impacts associated with each specific issue, and we can call these cost impacts the value gap. This is the conservatively estimated costs that are attributable to data quality issues, and provides a quantification of the corresponding loss of value to the organization.

The result of this process is a list of financial impacts directly related to measurable data failures prioritized by the breadth of the value gap. And because the data metrics are associated with known performance measures, improving the quality of data as measured by the data metrics increases the probability that changes to the data environment will result in measurable business improvement.

About the Author

David Loshin, president of Knowledge Integrity, Inc, (www.knowledge-integrity.com), is a recognized thought leader and expert consultant in the areas of data quality, master data management, and business intelligence. David is a prolific author regarding BI best practices, via the expert channel at www.b-eye-network.com and numerous books and papers on BI and data quality. His book, "Business Intelligence: The Savvy Manager's Guide" (June 2003) has been hailed as a resource allowing readers to "gain an understanding of business intelligence, business management disciplines, data warehousing, and how all of the pieces work together." His most recent book is "The Practitioner's Guide to Data Quality Improvement," and his insights on data quality can be found at www.dataqualitybook.com. His book, "Master Data Management," has been endorsed by data management industry leaders, and his valuable MDM insights can be reviewed at www.mdmbook.com.

David can be reached at loshin@knowledge-integrity.com

About the Sponsor

Informatica Corporation (NASDAQ: INFA) is the world's number one independent provider of data integration software. Organizations around the world gain a competitive advantage in today's global information economy with trustworthy, actionable, and authoritative data for their top business imperatives. More than 4,100 enterprises worldwide rely on Informatica to access, integrate, and trust their information assets held in the traditional enterprise, off premise, and in the cloud.