

## **Non qualité de données & CRM : quel coût ?**

Delphine Clément  
Hewlett-Packard - 14 avenue du Général Caunègre – 40000 MONT DE MARSAN  
[delphine\\_clement@hp.com](mailto:delphine_clement@hp.com)

Brigitte Laboisie  
A.I.D. - 4 rue Henri Le Sidaner – 78000 VERSAILLES  
[blaboisse@aid.fr](mailto:blaboisse@aid.fr)

Dominique Duquennoy  
A.I.D. - 4 rue Henri Le Sidaner – 78000 VERSAILLES  
[dduquennoy@aid.fr](mailto:dduquennoy@aid.fr)

Andrea Micheaux  
A.I.D. – 4 rue Henri Le Sidaner – 78000 VERSAILLES  
[andrea.micheaux@aid.fr](mailto:andrea.micheaux@aid.fr)  
PRISM Université Paris 1 Panthéon - Sorbonne

**Résumé :** La littérature est nombreuse autour du coût de la non qualité de données : prestataires de logiciels, cabinets d'organisation, de conseil, universitaires, sociétés d'études publient des chiffres, des études, des modèles. L'objet de cet article est de faire un rapide tour d'horizon de la littérature existante sur les coûts de la non qualité dans les CRM (Customer Relationship Management), puis de présenter les travaux effectués par A.I.D. Ces travaux sont principalement à ce jour dans l'évaluation du coût de la non qualité, prenant en compte les coûts directs et indirects (opportunités manquées). Un cas opérationnel de simulation sur des campagnes de marketing direct est présenté. Enfin, une analyse critique de nos travaux actuels est faite avec en perspective les méthodes que l'on souhaite appliquer pour une évaluation plus scientifique des opportunités manquées.

### **1 Introduction**

Dans cet article, nous faisons un rapide tour d'horizon des méthodes, études existantes autour du coût de la non qualité, et ce plus particulièrement dans le domaine du CRM.

Dans une deuxième partie, nous présentons une expérimentation faite lors de l'élaboration d'un plan de campagne marketing direct multi-canal. Enfin, nous procédons à une analyse critique de cette expérimentation en proposant des pistes d'amélioration.

Non qualité de données & CRM : quel coût ?

## 2 Contexte

### 2.1 Le métier d'A.I.D.

A.I.D. est une société de services française spécialisée dans les Bases de Données marketing, la qualité de données et l'enrichissement statistique. A.I.D. travaille à l'international de part son appartenance au groupe de communication OMNICOM et les outils, référentiels développés depuis plusieurs années, et ce au niveau mondial.

Le métier d'A.I.D. est donc autour de la donnée, spécifiquement sur les données Marketing, CRM, 'Identification du client/prospect'.

## 3 Coûts de la non qualité : tour d'horizon des publications actuelles

Dans ce tour d'horizon, on peut différencier 3 types de publications :

### 3.1 Les publications scientifiques

On trouve dans cette section des articles comme Ardagna et al. (2005), sur l'influence des données de non qualité dans des algorithmes en général utilisés en manipulation, découverte de données :

- Le dédoublement ou 'record matching' consiste à retrouver dans une base de données les enregistrements correspondant à la même personne par exemple. En règle générale, il s'agit de comparer des enregistrements avec des informations similaires mais pas égales, par exemple un nom proche, une adresse similaire. Les algorithmes présentés d'une manière classique (Batini et Scannapieco (2006) – Bertiequille (2005) – Winkler (1999)) sont basés sur l'optimisation du risque. Plus précisément, il existe 2 risques : risque de regrouper à tort des enregistrements correspondant en fait à des personnes différentes ('over-kill'), risque de ne pas regrouper des enregistrements correspondant à la même personne et laisser des doubles ('under-kill'). A partir d'une population de référence, ces algorithmes optimisent en minimisant les deux types d'erreurs. En réalité, le coût de l'erreur n'est pas le même selon le type de population, l'utilisation qui va être faite des données. Agréger dans un référentiel client des contrats à tort peut être très coûteux par rapport aux opportunités ratées de ventes complémentaires, envoyer un courrier en trop aura une conséquence financière moindre. Des algorithmes tels que Vassilios et al (2001) optimisent non pas le risque mais le coût de l'erreur : l'utilisateur peut fournir une matrice de coût  $C_{ij}$ , correspondant au coût de prendre la décision  $A_i$  alors que les enregistrements devraient être dans la décision  $A_j$ .
- L'Extraction et Gestion des Connaissances, et plus particulièrement les règles d'association sont également un domaine où le coût de la non qualité a été pris en compte. Parmi les indicateurs les plus communs pour mesurer la pertinence de règles, on peut citer la confiance et le support. Mais, au-delà de ces indicateurs permettant d'évaluer la significativité des règles, il est fondamental de ne pas oublier la qualité des données elles-mêmes et leur influence sur les règles produites. Bertie-

Equille (2005) propose une modélisation du coût de la non qualité sur les règles d'association découvertes. Pour ce faire, il est défini une matrice de coût  $C_{ij}$  : coût de prendre une décision  $D_i$  pour classifier une règle (intéressante, potentiellement intéressante, non intéressante) avec un vecteur de qualité  $j$  des items composants les parties de la règle.

### 3.2 Les études de composants

A la frontière entre la théorie et la mise en pratique, nous entendons par étude de composants des grilles telles que English (1999), Eppler et Herlfert (2004), Loshin (2004), Batini et Scannapieco (2006) qui fournissent une liste de critères de coûts et de gains. Destinées aux praticiens pour 'vendre' les projets qualité de données, ces études permettent de lister, évaluer les coûts et les bénéfices de la qualité d'une manière exhaustive et rigoureuse afin d'évaluer la rentabilité du projet. On peut également citer les différentes méthodes d'analyse présentées par Wang (2006), avec en particulier une méthode basée sur le concept de l'information gérée comme un produit.

L'article de Eppler et Helfert (2004) liste différentes catégories de coût : selon leur origine (perte de données par exemple par destruction abusive), selon leurs conséquences (re-saisie des informations), par dimension qualité (inexactitude, complétude, unicité,...), par règle d'évolution (linéaire, fixe, exponentielle,...). D'autres catégories sont citées comme coûts directs/indirects, court ou moyen terme,... Enfin, un arbre de classification donne une vue synthétique des catégories avec en niveau 1 la distinction coûts générés par la non qualité et coûts pour maintenir, prévenir la qualité de données. Parmi les coûts générés, la distinction suivante est faite entre coûts indirects et coûts directs. La figure 1 ci-dessous reprend en partie cet arbre.

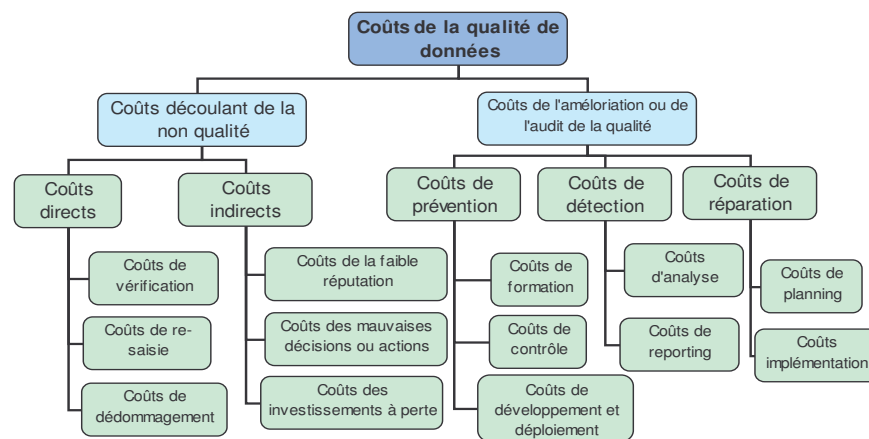


FIG.1 - Eppler et Helfert (2004) : Taxinomie

Le livre de Larry English fournit une classification des coûts ainsi que des exemples de coûts unitaires en Marketing direct. A remarquer la méthode pour prendre en compte la dimension unicité de l'information ou présence de doubles. Nous rencontrons deux écoles : compter les courriers envoyés en trop (coût direct) ou compter la perte d'opportunité sur les

Non qualité de données & CRM : quel coût ?

courriers non envoyés à des clients/prospects du fait des doubles et d'un nombre maximum d'envois atteint. Cette publication apporte également une démarche projet sur la mesure des coûts, avec la notion de segment client et de valeur client (voir la figure ci-dessous) . Ces notions sont appliquées dans le cas pratique que nous présentons plus loin.

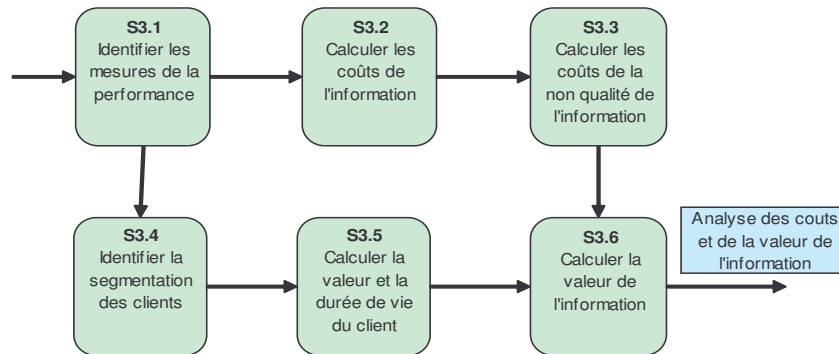


FIG. 2 - English (1999) Méthode Projet de mesure des coûts de la non qualité de l'information

### 3.3 Les études de marchés

Editeurs de logiciels, cabinets de consultants sont 'friends' d'études de marchés, d'enquêtes fournissant des valorisations monétaires sur les coûts de la non qualité.

Pour ne citer que quelques exemples :

- Une étude menée par le Datawarehousing Institute en 2002 auprès de 647 contacts au sein d'entreprises majoritairement nord-américaines (Etats-Unis et Canada) a montré que le coût de la non qualité des données client était évalué à 611 Milliards de dollars par an en Amérique du Nord en prenant en compte les coûts humains, d'impressions et d'affranchissement.
- Plus récemment, en 2006, une étude menée en Hollande auprès de 20 000 sociétés de 10 personnes et plus a montré que la non qualité génèrait auprès de ces entreprises des coût directs de 400 Millions d'euros. Ce coût prend en compte uniquement les coûts supplémentaires générés par des factures avec des adresses erronées ou des produits qui n'arrivent pas à la bonne adresse. L'étude a été menée par Ed Peelen, professeur de Marketing Direct à la Neynrode Business University, et commanditée par les sociétés Human Inference et Cendris.
- Enfin, en 2006 également, une étude a été menée par Dynamic Markets pour QAS (éditeur de logiciels) auprès de 800 professionnels (400 personnes qui se chargent d'ouvrir leur courrier) dans 8 régions : Asie-Pacifique, Benelux, France, Allemagne, Pays Nordiques, Espagne, Royaume-Uni et Etats-Unis. Citons quelques chiffres clés : en moyenne, les personnes reçoivent 25 courriers par mois destinés à des personnes qui ont quitté la société. Par ailleurs, environ 190 courriers par mois

arrivent au nom de la personne mais sont considérés par elle comme mal ciblés. L'étude montre également que seulement 5% de ces courriers sont renvoyés à l'expéditeur, avec un coût moyen observé de 637 euros par mois des courriers renvoyés, soit un coût total d'environ 150 000 euros par an.

Voir également Agosta (2003) sur une analyse du coût de l'information et des problèmes de données.

#### **4 A.I.D. : évaluation du coût de la non qualité en marketing direct et mise en œuvre opérationnelle**

L'évaluation du coût de la non qualité des données marketing, dans le CRM B2B, est importante à plus d'un titre.

Tout d'abord, accompagner la mesure de la non-qualité du coût que cela représente pour la société est un message très fort pour le management ; ce dernier, alors sensibilisé à l'importance de la qualité, est plus enclin à octroyer des budgets pour les actions d'amélioration et de prévention.

Ensuite, l'évaluation du coût de la non-qualité est essentielle pour définir des priorités d'action. En effet, il sera bénéfique pour la société d'investir en premier lieu sur la résolution des problèmes de qualité ayant le plus grand impact financier.

La non qualité des informations utilisées lors de campagnes de marketing direct, engendre , pour reprendre la classification de Eppler et Helfert, deux grandes catégories de coûts pour les sociétés : les coûts directs et les coûts indirects.

Dans la catégorie « coûts directs », nous rangeons des coûts tels que :

- Envoi en double
- Essai de contacter une personne qui a quitté la société
- poursuite judiciaire en cas de sollicitation d'un client sans opt-in
- etc...

Les coûts directs sont relativement simples à calculer et les chiffres sont peu contestables. Ils sont à décliner par canal (téléphone, email, courrier) et prennent en compte les coûts de la campagne (coûts fixes et variables).

Dans la catégorie « coûts indirects », plus difficile à mesurer, nous rangeons des coûts tels que :

- opportunités manquées
- impact sur la satisfaction client
- prise de décisions stratégiques erronées
- etc...

Le calcul du coût de la non qualité de données dans un CRM a comme point de départ une mesure objective de la non qualité.

Ainsi, les dimensions de complétude et de validité de l'information client sont mesurées, tout comme le taux d'enregistrements client en double. Cette mesure s'effectue sur l'ensemble de la base CRM, il s'agit d'indicateurs qualité « a priori ». Voici ci-dessous un exemple de publication d'indicateurs qualité « a priori ».

Non qualité de données & CRM : quel coût ?

The screenshot shows the 'BDQS PUBLICATION' interface. At the top, there are navigation tabs: 'MAJOR EVOLUTIONS', 'BENCHMARK', 'DETAILED REPORT', and 'REPORTS'. Below this, a summary table provides key statistics:

	Total Population	Population of Sample	Accuracy Perimeter	% Completeness LN
Sites	2,399,159	2,399,159	2,193,320	99 %
Contacts		915,005	915,005	100 %

Below the summary table, there are filter options: 'DISPLAY ALL' or select: 'COMPLETENESS', 'ACCURACY', 'UNIQUENESS', 'OVERLAP', 'SYNCHRONIZATION', 'PRIVACY FOCUS'. The 'COMPLETENESS' section is expanded, showing two tables of data quality metrics:

	% Completeness	% Fake Value	% No Compliance length
Company Name	100 %	2.15 %	0.03 %
Company Division	0 %		
Company Address Line 1	95.46 %	1.83 %	0.86 %
Company Address Line 2	12.82 %	0.23 %	0.01 %
Company Address Line 3			
City Area			

	% Completeness	% Fake Value	% No Compliance length
Personal Title	0.51 %	0 %	0 %
Contact First Name	97.77 %	0.08 %	0 %
Middle Initial	0.78 %	0 %	0 %
Contact Last Name	100 %	0.28 %	0 %
Business function			
Job Role			

FIG. 3 – Exemple d'indicateurs qualité de données a priori

Par ailleurs, il est intéressant de prendre également en compte des indicateurs qualité « a posteriori », c'est-à-dire des mesures de qualité ciblées sur les informations utilisées lors de campagnes marketing. Ces indicateurs sont le taux de NPAI (campagne de mailing), le taux d'emails non aboutis (campagne d'Emailing), le taux de faux téléphones (campagne de télémarketing) et le taux de contacts obsolètes (tout type de campagne)..

Dans une seconde étape, nous avons documenté les coûts directs par campagne, par variable et par problème qualité.

**Mailings**

Variable	Complétude	Pollution	Obsolescence	Normalisation	Consistance	Syntaxe	Double
contact							3,00 €
société			3,00 €				3,00 €
adresse				0,24 €			

**E-mailings**

Variable	Complétude	Pollution	Obsolescence	Normalisation	Consistance	Syntaxe	Double
contact							0,20 €
société			0,20 €				0,20 €

**Telemarketing**

Variable	Complétude	Pollution	Obsolescence	Normalisation	Consistance	Syntaxe	Double
contact			2,00 €				4,00 €
société			4,00 €				4,00 €

TAB. 1 – Coûts unitaires directs par dimension qualité

Dans cet exemple, 0,24 € correspond au coût supplémentaire de l'affranchissement postal lorsque l'adresse n'est pas normalisée (à moduler par pays, type d'envoi). 2,00 € est le coût supplémentaire de télémarketing lorsque le contact a quitté la société et qu'il est nécessaire de rechercher son successeur.

Enfin, il convient d'évaluer le poids du problème de qualité sur les coûts indirects de la campagne marketing. Par exemple, un téléphone manquant ou erroné dans le cadre d'une campagne de télémarketing aura un impact de 100% ; par contre, dans le cadre d'une campagne d'Emailing, l'impact sera de 0%. C'est pourquoi il est important de créer une matrice d'évaluation comprenant d'une part le type d'information ou la variable (nom de la société, adresse, téléphone, etc...), d'autre part, le problème qualité (complétude, doubles, validité, etc...) et enfin, le pourcentage d'impact selon le type de campagne effectuée.

#### E-mailings

Variable	Complétude	Pollution	Obsolescence	Normalisation	Consistance	Syntaxe
contact	50%	25%	100%		40%	40%
société	100%	50%				
email	100%	100%	100%		50%	100%

#### Mailings

Variable	Complétude	Pollution	Obsolescence	Normalisation	Consistance	Syntaxe
contact	30%	15%	50%		40%	40%
société	100%	50%				
address			100%			100%

#### Telemarketing

Variable	Complétude	Pollution	Obsolescence	Normalisation	Consistance	Syntaxe
contact	30%	15%	50%		20%	20%
société	100%	50%				
tel	100%	100%	100%			100%

TAB. 2 – Coûts unitaires indirects par dimension qualité

Pour appliquer sur la population, nous avons attribué à chaque client une moyenne d'intention d'achat ainsi qu'une valeur moyenne d'achat (ces moyennes peuvent varier selon le canal marketing utilisé). La table ci-dessous fournit un exemple de valeur d'un contact selon le canal d'adressage. Cette valeur est obtenue en multipliant l'intention d'achat par la valeur moyenne d'un achat pour ce type de campagne. Il est à noter que ces valeurs sont à moduler selon le segment de client : en BtoB par exemple, cette matrice sera en général déclinée par taille de l'entreprise ou par croisement secteur d'activité, taille de l'entreprise.

Segment	valeur TMK	valeur Email	valeur Mail
Segment de clients à fort potentiel	€ 538.00	€ 244.00	€ 125.00

Données fictives

1.5 – Valeur de contact

## Non qualité de données & CRM : quel coût ?

A partir de ces hypothèses, nous avons effectué l'exercice de simuler, sur un trimestre de campagnes de marketing direct, les coûts de la non qualité. Le plan de campagne prévu a été décliné par canal, permettant ainsi d'avoir les cibles prévues et d'analyser leur niveau qualité de données. Cette approche très opérationnelle évite le biais d'extrapolations trop théoriques sur l'ensemble de la base de données, avec des probabilités d'utilisation.

Nous avons pu ainsi mesurer les coûts directs et indirects sur le plan de campagne, utilisant les coûts de fabrication des différentes campagnes ainsi que les rentabilités prévisionnelles. Le tableau 4 ci-dessous présente un récapitulatif des coûts obtenus, splité par segment de clients. Le segment 1 correspond à de grandes entreprises, avec une valeur importante, le segment 2 des entreprises de taille moins importante. On notera que le coût de la non qualité représente au global 15% des ventes estimées, le coût le plus fort provenant des opportunités ratées. Si les chiffres présentés ci-dessous ne sont pas les chiffres réels, la proportion a été respectée. Ces faibles coûts directs s'expliquent par le fait que des travaux qualité de maintenance sont appliqués en permanence sur le CRM, en particulier sur les facteurs générateurs de coûts directs : doubles, obsolescence, normalisation d'adresses.

	Segment 1	Segment 2	Total
<b>Nombre de Messages à envoyer</b>	6 376	98 995	105 371
<b>Ventes Estimées</b>	2 359 K€	12 176 K€	14 536 K€
<b>Coûts Directs de la non Qualité</b>	2 K€	31 K€	33 K€
<b>Coûts Indirects : Manque à Gagner</b>	214 K€	1 871 K€	2 085 K€
<b>Coût Total de la Non Qualité</b>	216 K€	1 902 K€	2 118 K€

TAB. 4 – Coûts Globaux Evalués

## 5 Critique et perspectives

Cette première expérimentation avait le mérite de recenser les critères de la non qualité, de réfléchir à leur évolution. Les coûts directs ont été faciles à mesurer, la partie 'opportunité ratée' beaucoup plus subjective. On arrive là aux limites du système car, pour que les chiffres soient reconnus, ils ne doivent pas être contestables ou le moins possible. Pourquoi une valeur erronée dans le prénom fait baisser de 25% l'efficacité de l'emailing, de 30% le mailing ? Ces chiffres, fournis par l'expert marketing, laissent un peu 'sur notre faim' et blessent l'esprit scientifique. Des campagnes marketing avec des échantillons témoins 'propres' versus des échantillons témoins de moins bonne qualité, l'analyse de la rentabilité d'actions passées selon leur niveau qualité de données (phénomène plus difficile à isoler des autres facteurs d'influence sur des campagnes initialement non prévues pour ce type d'expérimentation), sont quelques pistes que l'on peut envisager. La mesure s'avère plus complexe lorsque les variables de ciblage sont erronées par exemple.

Enfin, une autre évolution de cette expérimentation reste bien entendu la mesure du coût de la maintenance, nettoyage versus le coût de la non qualité, facette non prise en compte dans cette première évaluation.

## Conclusion

Cette expérimentation, effectuée dans le cadre d'un travail avec une grande multinationale, a attiré l'attention du numéro 2 mondial, là où les questions CRM habituelles restent à un niveau inférieur. C'est un moyen d'élever le niveau du débat et de permettre aux équipes CRM de débloquer les budgets nécessaires pour remédier à ces problèmes. Le challenge consiste maintenant à rendre cette expérimentation plus scientifique.

## Lexique

B2B : Business to Business : Vente aux entreprises  
B2C : Business to Consumer : Vente au grand public  
Benchmark : Comparaison CRM : Customer Relationship Management  
Customer Data Integrity : Intégrité des données Client  
Customer Knowledge Management and Data Stewardship : Gestion de la connaissance client et services sur les données  
Emailing : Campagne marketing par envoi d'email  
NPAI : N'habite Pas à l'Adresse Indiquée  
Opt-in : Terme marketing ou légal qualifiant une adresse courriel. Une adresse courriel Opt-in signifie que l'utilisateur de cette adresse a eu préalablement un accord de la part du propriétaire de l'adresse pour l'utilisation de cette adresse dans un cadre précis.  
Over kill : Lors d'un dédoublement, rapprochement à tort de deux enregistrements  
Record Matching : Fusion de deux enregistrements  
ROI : Retour sur investissement  
Under kill : Lors d'un dédoublement, non rapprochement à tort de deux enregistrements

## Références

- Agosta L., (2003), *The Costs of Information and Data Quality Defects – The Data Strategy Advisor*, DM Review Magazine, [www.dmreview.com](http://www.dmreview.com)
- Batini C., Scannapieco M., (2006). *Data quality: concepts, methodologies and techniques* Springer
- Berti-Equille, L. (2005). *Qualité des données multi-sources : un aperçu des techniques issues du monde académique*. Journées CRM & Qualité des Données au CNAM
- Berti-Equille L., (2005). *Cost of Low-Quality Data over Association Rules Discovery*. Proceedings of International Symposium on Applied Stochastic Models and Data Analysis (AMSDA 2005) Brest, France.
- Ardagna D., Cappiello C., Comuzzi M, Francalanci C., Pernici B., (2005). *A broker for selecting and provisioning high quality syndicated data* pp.262-279. Proceedings of the 10<sup>th</sup> International Conference on Information Quality. Boston
- English L. P. , (1999) *Improving Data Warehouse and Business Information Quality*, Wiley

Non qualité de données & CRM : quel coût ?

- Eppler Martin J, Helfert M. (2004) *A framework for the classification of data quality costs and analysis of their progression*, MIT Conference on information quality.
- Loshin D., (2004) *Enterprise Knowledge Management - The Data Quality Approach*. Morgan Kaufmann Series in Data Management Systems
- Turney P., (2000) *Types of Cost in Inductive Concept Learning*, Proceedings of the cost-sensitive learning workshop at the 17<sup>th</sup> ICML-2000 Conference Stanford
- Vassilios S. Verykios, George V. Moustakides, Mohamed G. Elfeky, (2001). *A Bayesian decision model for cost optimal record matching*. Springer-Verlag
- Wang Y. R., Lee W. L., Pipino L.L., Funk J. D., (2006), *Journey to Data Quality*, MIT Press
- Winkler WE., (1999). *The State of Record Linkage and Current Research Problems*. Statistics of Income Division, Internal Revenue Service Publication R99/04

## Summary

We have a lot of publications around the cost of non quality data: software companies, consulting firms in Management, Academics, and market research companies publishing surveys. The goal of this article is to provide an outline on the existing publications on the costs of non quality in CRM systems, and in a 2<sup>nd</sup> part to present the works done by AID. These works are principally today in the evaluation of the costs of non quality data. They are taking in account direct costs but also indirect costs: missed opportunities typically. An operational case based on a simulation of these costs on direct marketing campaigns is also presented. Finally, a critical analysis of our actual works is done within mind the methodologies we wish to apply to evaluate more scientifically the missed opportunities.