DATA TRACKING : LE "RETURN ON INVESTMENT" DE L'ANALYSE DES FLUX D'INFORMATION

ISABELLE BOYDENS

Résumé – L'egovernment repose sur la gestion de vastes bases de données dont les enjeux sont stratégiques sur les plans sociaux, financiers et juridiques. Ces bases de données sont alimentées par des flux d'information (entre citoyens, entreprises, employeurs et administrations) véhiculant un grand nombre d'anomalies formelles (valeurs déviantes par rapport aux valeurs attendues) dont la gestion est coûteuse. Le « data tracking » est une technique permettant de détecter les causes structurelles de ces anomalies et d'y remédier durablement à la source. Appliqué aux États-Unis dans les laboratoires d'AT&T, le « data tracking » a également été mis en œuvre dans le secteur de la sécurité sociale belge avec un ROI important. Ce rapport en développe les modalités d'application, les gains et les possibilités de généralisation à tout type de base de données.

Samenvatting – E-government baseert zich op het beheer van uitgebreide databases die een strategische rol spelen op sociaal, financieel en juridisch vlak. Deze databases worden verrijkt door informatiestromen (tussen burgers, bedrijven, werkgevers en besturen) die een groot aantal formele anomalieën met zich meebrengen (afwijkende waarden ten opzichte van de verwachte waarden) waarvan het beheer duur is. "Data tracking" is een techniek waarmee de structurele oorzaken van deze anomalieën kunnen opgespoord worden en die een duurzame oplossing aan de bron aanbiedt. In de VS wordt "data tracking" toegepast in de laboratoria van AT&T en het werd eveneens ingevoerd in de sector van de Belgische sociale zekerheid met een belangrijke ROI. Dit document bespreekt de toepassingsmodaliteiten, de voordelen en de mogelijkheden om de techniek te veralgemenen naar elk type database.

Table des matières

1. Introduction

À l'heure où la maîtrise des coûts est plus cruciale que jamais, nous présentons la technique du « data tracking » : appliquée à l'egovernment, cette technique demande des investissements relativement faibles et offre des gains potentiellement très élevés, en termes de qualité et de rapidité de traitement des données, des flux financiers et des avantages sociaux correspondants.

Le « data tracking » vise à évaluer quantitativement la validité formelle des valeurs introduites dans une base de données en vue d'en optimiser le traitement. L'approche permet de détecter les causes structurelles d'incohérences, d'y remédier à la source, de diminuer le nombre d'anomalies et d'en améliorer durablement le traitement. Une base de données est comparable à un fleuve : au lieu de se limiter à nettoyer ponctuellement le fond du fleuve via des contrôles de correction automatiques (comme le préconise le « data cleansing », méthode de correction automatique), il s'agit d'aller plus loin et d'en analyser structurellement les sources et les flux[1].

Éprouvée dans le secteur privé depuis une dizaine d'années, la technique s'applique également dans le cadre de l'administration fédérale où elle mériterait d'être mieux connue.

[1] BOYDENS I., L'océan des données et le canal des normes. In CARRIEU-COSTA M.-J., BRYDEN A. et COUVEINHES P. éds, Les Annales des Mines, Série "Responsabilité et Environnement" (numéro thématique : « La normalisation : principes, histoire, évolutions et perspectives »), Paris, n° 67, juillet 2012, pp. 22-29. http://www.ulb.ac.be/cours/iboydens/annales.pdf

La question du « tracking de l'information » n'est pas neuve…[2] Au Moyen-Âge, le processus de transformation de l'information, avant l'invention de l'imprimerie, se déployait de siècles en siècles avec les générations de moines copistes. Ces manuscrits nous sont souvent parvenus sous forme morcelée et partielle : il arrive que l'original soit perdu et que l'on ne dispose que d'un ensemble incomplet de copies divergentes (à la suite d'erreurs - volontaires ou non - commises par les copistes : passages transformés, omis ou ajoutés).

Afin d'établir l'appareil critique d'un manuscrit (dont on dispose souvent de multiples copies divergentes et dont on a parfois perdu l'original), l'historien construit un *stemma codicum* (généalogie des données), technique d'analyse comparative des variantes empruntée à la philologie. L'établissement du *stemma codicum*, accompagné d'un travail d'interprétation critique, offre une reconstruction conjecturale argumentée du manuscrit original.

À la fin du vingtième siècle, le géant des télécommunications américain, AT&T Laboratories, a déployé à plus grande échelle cette technique d'analyse du processus de transformation de l'information (« data tracking ») reposant en partie sur un principe analogue. La technique du « data tracking », proposée par Thomas Redman[3], vise à évaluer et à améliorer les aspects formels de la qualité des vastes bases de données contemporaines (point 2)[4]. Après en avoir présenté les grandes lignes, nous en montrons un exemple d'application pratique dans le domaine de la sécurité sociale belge, mis en œuvre pour la première fois en 2006 et réinitié en 2012 dans le cadre d'un projet récurrent et structurel (point 3). En conclusion, nous montrons comment l'approche peut être généralisée dans le secteur de l'administration publique et des soins de santé et dans de nombreux domaines au sein desquels la prise en compte de l'évolution des données dans le temps est stratégique.

L'approche présentée ici a été adaptée, enrichie et appliquée au sein de l'egovernment en Belgique par la « data quality cel » de la section Recherche de Smals. Ces travaux originaux ont fait l'objet d'échos récents au niveau international (voir : recension des travaux d'I. Boydens par D. Bade, de l'Université de Chicago, en mai 2011[5] et publication d'un

---

[2] BOYDENS I., Hiérarchie et anarchie : dépasser l'opposition entre organisation centralisée et distribuée ? In HUDON M. et EL HADI W. M., éds, Les cahiers du numérique (Numéro thématique « Organisation des connaissances et Web 2.0 »). Paris : Editions Hermès Sciences-Lavoisier, vol. 6, n° 3, 2010, pp. 77-101. http://www.cairn.info/resume.php?ID_ARTICLE=LCN_063_0077

[3] Redman T., « Data Quality : the Field Guide ». Digital Press, 2001. LOSHIN (D.), The Practicioner's Guide to Data Quality Improvment, Elsevier, Morgan-Kaufmann OMG Press, 2011. BOYDENS I., Informatique, normes et temps. Bruxelles : Éditions E. Bruylant, 1999. Cet ouvrage s'est vu décerner le prix de la Fondation "Louis Davin", conféré par l'Académie Royale des sciences, des lettres et des beaux-arts de Belgique.

[4] La qualité d'une base de données désigne son adéquation aux usages pour lesquels elle est conçue et exploitée (« fitness for use »).

[5] BADE D., It's about Time!: Temporal Aspects of Metadata Management in the Work of Isabelle Boydens. In Cataloging & Classification Quarterly (The

chapitre par I. Boydens sur ses travaux dans un livre sur l'egovernment qu'elle a coédité en 2011 aux éditions Springer à New York[6]).

2. Data Tracking et reengineering

Dans l'approche de Redman, *data tracking* et *reengineering* sont deux opérations intimement liées : elles ont pour objet l'examen et la rationalisation des flux d'information encadrant une base de données. Avec le spectaculaire développement des réseaux, certaines difficultés sont en effet exacerbées : les informations incohérentes ou incomplètes sont plus rapidement et plus massivement transmises d'un système d'information à l'autre. En corollaire, des bases de données conçues à certaines fins sont fréquemment exploitées à d'autres, l'utilisateur final se trouvant de plus en plus éloigné de la source productrice de l'information. Ainsi, pendant la première guerre du Golfe, environ 28 000 des 40 000 containers militaires américains envoyés au Moyen-Orient durent être inspectés et inventoriés manuellement, tant l'interrogation des bases de données censées en répertorier le contenu donnait lieu à des résultats incohérents. D'où cette remarque non dépourvue d'amertume : "*In general, the physical movement of material is faster than the movement of the supporting information...*"[7].

Dès lors, il est utile de disposer d'une méthode permettant de déceler les points critiques au sein d'un système d'information en vue d'en optimiser le fonctionnement[8]. À cette fin, plusieurs métriques et méthodes quantitatives, parfois peu concluantes, ont été proposées : certains ont par exemple tenté de poser les fondements formels d'un système expert intitulé *data quality reasoner*[9]. Ce dernier vise, sur la base de spécifications définies par l'utilisateur, à réaliser « automatiquement » des « modèles de jugement » de type *« if timeliness is high and source credibility is medium then the data may be of high quality* »... Tout le problème réside dans la définition préalable de ces dimensions de la qualité a priori subjectives, tâche qui revient aux utilisateurs et n'est pas

---

International Observer), volume 49, n° 4, 2011, pp. 328-338. http://catalogingandclassificationquarterly.com/ccq49nr4.html#intobs

[6] BOYDENS I., "Strategic Issues Relating to Data Quality for E-government: Learning from an Approach Adopted in Belgium". In ASSAR S., BOUGHZALA I. et BOYDENS I., éds., "Practical Studies in E-Government : Best Practices from Around the World", New York, Springer, 2011, p. 113-130.

[7] MADNICK S. E., "The Voice of the Customer: Innovative and Useful Research Directions", in AGRAWAL R., S. BAKER S. & BELL D., éds, Proceedings of the 19thConference on Very Large Databases, Dublin, VLDB, 1993, p. 702.

[8] COREY D. J., Data Quality Improvement Activities in the Military Health Services Systems and the U. S. Army Medical Department. In STRONG D. M. et KAHN B. K., éds, Proceedings of the 1997 Conference on Information Quality. Cambridge : M.I.T., 1997, p. 37-62.

[9] WANG R. Y., KON H. B. et JANG Y., A knowledge-based Approach to Assisting in Data Quality Judgement, Total Data Quality Management (T.D.Q.M.) Research Program Sloan School of Management.

envisagée par le système... L'utilité pratique de ces approches soulève dès lors de grandes réserves : "*an important point to digest is that a data quality system is not a software tool rather it is a management discipline. An effective system can not be built without any expensive or specialized package*"[10]. Parmi tous les travaux relatifs à l'analyse des processus de transformation formelle des données, l'approche de T. Redman est la plus aboutie. Elle inclut en particulier un système de mesure (contrôle statistique de la qualité), l'établissement de contrôles de conformité avec les spécifications (*data tracking)* et, enfin, la mise en œuvre de projets de *reengineering*. Nous présentons ces trois points successivement.

2.1. Le contrôle statistique de la qualité

Le contrôle statistique de la qualité (*Statistical Quality Control* - SQC) fut développé en vue d'améliorer la production industrielle. Appliquée à une base de données, la mise en œuvre d'un système de mesure statistique est destinée à assurer une évaluation continue de l'adéquation des performances aux spécifications formelles[11]. L'idée-clé de la démarche vise à prévoir et à vérifier qu'un processus stable se comportera, dans certaines limites, dont on peut fournir une estimation, comme il s'est comporté dans un passé récent. Il est dès lors préconisé de stabiliser les processus analysés ou, tout au moins, de les inscrire dans un état de « contrôle statistique ».

En effet, tous les processus sont sujets à des « variations ». Dans la mesure où la prévisibilité repose sur une hypothèse de stabilité des variations, il est nécessaire de distinguer les processus stables des processus instables, en vue, lorsque cela est possible, de rendre ces derniers « prévisibles » grâce à l'élimination des éléments instables. À cet égard, Redman distingue :

- *Les causes spéciales* (externes au processus, « accidentelles » et sources d'instabilité).
- *Les causes communes* (internes au processus, permanentes et inhérentes à celui-ci).

Redman donne l'exemple suivant : si l'on mesure pendant plusieurs jours les heures de départ et d'arrivée d'un employé depuis son domicile à son lieu de travail, les feux de signalisation tantôt plus fréquemment rouges ou verts d'un trajet à l'autre représentent des « causes communes ». Il s'agit de sources de variation internes au processus. Par contre, la crevaison d'un des pneus du véhicule ou le blocage de toutes les routes par une manifestation d'étudiants constituent autant de causes spéciales, externes au processus. Ces sources d'instabilité occasionnent, dans notre exemple, un retard anormalement élevé. La distinction entre ces deux types de causes est fondamentale, car la façon d'y remédier est radicalement distincte. Alors qu'on remédie aux « causes communes » en

[10] FIRTH C., Data Quality in Practice : Experience from the Frontline dans Wang R. Y., éd., Proceedings of the 1996 Conference on Information Quality. Cambridge : M.I.T., 1996, p. 65-71.

[11] REDMAN T., Data Quality for the Information Age. Boston : Artech House, 1996, p. 155-183.

agissant sur le processus (par exemple, les retards dus aux « feux rouges » seront minimisés en réorientant le parcours de l'employé vers un parcours comportant moins de feux de signalisation), on remédie aux « causes spéciales » en agissant sur l'environnement du processus, lorsque cela est possible (les crevaisons peuvent être minimisées par une inspection quotidienne des pneus de la voiture).

La logique est donc la suivante : si le processus est stable, il est possible de prédire qu'il se comportera dans le futur comme il s'est comporté dans le passé. Dans ce cas, si la performance d'un processus considéré comme stable n'est pas adéquate, des procédures d'amélioration sont nécessaires. Si le processus n'est pas stable, les « sources de variation spéciales » doivent être, selon Redman, identifiées et éliminées. Cependant, Redman ne précise pas comment identifier et éliminer ces sources dans le cadre d'une base de données et nous verrons plus loin que ce point ne peut pas être esquivé, lorsque nous présenterons les études de cas appliqués à l'egovernment en Belgique (point 3).

La démarche de Redman repose sur l'usage de « substituts opérationnels », nécessairement arbitraires dans l'absolu, mais ayant fait leurs preuves dans la pratique. De façon schématique, ces substituts permettent d'identifier les limites au delà ou en deçà desquelles certaines observations correspondent à des « causes » dites « spéciales ». Les formules proposées reposent sur le principe statistique de « l'intervalle de confiance »[12]. Plus spécifiquement, une observation située hors des limites définies a priori dénote la présence d'une « cause spéciale ».

Après l'analyse et l'élimination des valeurs correspondant à des causes spéciales, on peut prévoir que les performances futures seront proportionnelles à celles précédemment analysées, sur la base de l'hypothèse selon laquelle la variation demeure constante. En effet, l'approche est valide *toutes autres choses étant égales* : toute modification du processus ou toute émergence de nouvelles « causes spéciales » perturbent les résultats de l'analyse[13].

2.2. Gestion de la qualité des processus : le data tracking

Le *data tracking*, technique mise au point par T. Redman[14], est une méthode destinée à évaluer quantitativement la validité formelle des valeurs introduites dans une base de données et à en améliorer le traitement. Traditionnellement, chaque enregistrement d'une base de

---

[12] WONNACOTT T. H. et WONNACOTT R. J., Statistique. Paris : Economica, 1991, p. 285-396.

[13] Ce type de raisonnement est également mobilisé dans le domaine du "predictive analytics" auquel on a par exemple recours dans le cadre de la lutte contre la fraude sociale. Voir par exemple : MESKENS J., Predictive analytics. Session d'information. Bruxelles, Smals, Section Recherche, 29 novembre 2011. Voir aussi les blogs de la « section Recherche » de Smals « Putting predictive Analytics to Work » 1 & 2 (J. MESKENS (http://blogresearch.smalsrech.be/?p=4242) et D. VAN DROMME (http://blogresearch.smalsrech.be/?p=4586) , 2012).

[14] REDMAN T., Data Quality for the Information Age..., p.185-212.

données est assemblé au terme de plusieurs étapes (ou processus), de la même façon qu'un produit est assemblé dans une usine. La qualité des données dépend de la qualité du processus d'assemblage.

Une des caractéristiques du *data tracking* (suivi des données) réside dans le fait que l'instrument de mesure est incorporé aux processus et permet en quelque sorte une analyse continue de leur qualité[15]. Le *data tracking* repose sur une exploitation de la redondance des données que l'on retrouve dans la plupart des systèmes informatiques en vue d'en évaluer trois aspects :

- La validité formelle de données intégrées dans une seule base de données.
- La cohérence entre données intégrées dans plusieurs bases de données.
- La durée des cycles de production et de traitement de l'information.

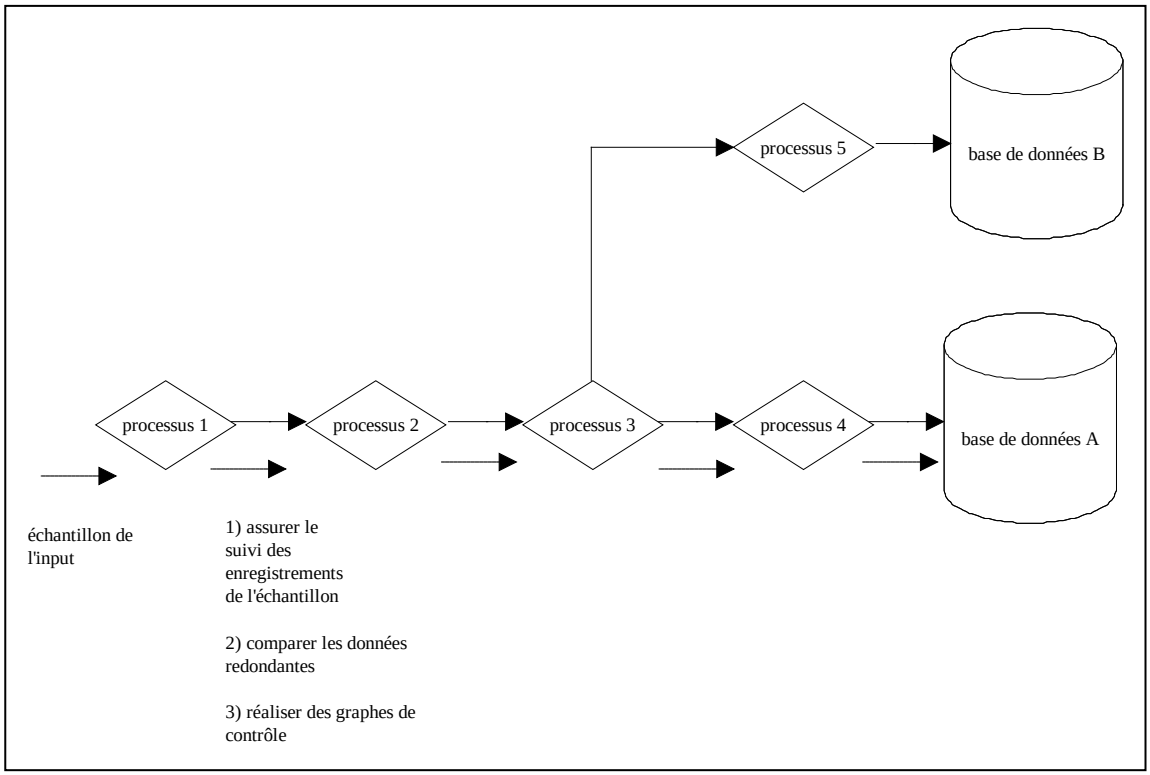
À la suite de Redman, nous présentons la méthode du *data tracking* sur la base d'un exemple fictif. La technique repose sur la mise en œuvre de six étapes successives :

Étape 1 : sur la base de l'identification préalable des chaînes d'information, sources et cycles de données (tâche d'analyse cruciale et non négligeable), prélever un échantillon aléatoire d'enregistrements introduits dans les différents processus et placer ceux-ci dans un « état de contrôle statistique » (Figure 1).

Étape 2 : « marquer » ces enregistrements de façon à en assurer le suivi.

[15] On trouve une approche similaire, à propos des systèmes d'information comptables dans : KAPLAN D., KRISHNAN R., PADMAN R. et PETERS J., Assessing Data Quality in Information Accounting Systems, Communication of the ACM, février 1998, vol. 41, n° 2, p. 72-78.

Figure 1. Identification et suivi des processus de transformation des données (T. Redman)



Étape 3 : identifier les erreurs formelles et évaluer la durée du cycle des processus. Dans la table suivante (Tableau 1), les valeurs ayant fait l'objet d'une modification sont reproduites en caractères gras. Certaines transformations affectant les valeurs d'un processus à l'autre peuvent révéler l'émergence d'erreurs formelles. Les indicateurs de contrôle incluent en outre l'étude des performances temporelles.

Tableau 1. Exemple de suivi d'un enregistrement (T. Redman)

	Processus 1	**Processus 2**	**Processus 3**	**Processus 4**	**Base de données A**
Attribut a	XYZ1	XYZ1	XYZ1	**XYZ1-001**	**XYZ1-001**
Attribut b	Oui	Oui	**Non**	Non	Non
Attribut c			K	K	K
Attribut d		1500	**5100**	5100	5100
Attribut e		Z	Z	**Z**	**1**
Attribut f					**OK**
Date entrée	01/03/89	02/03/89	20/03/89	04/04/89	04/04/89
Date sortie	01/04/89	10/03/89	01/04/89	25/04/89	
Date prévue		9/03/89	30/03/89	25/04/89	01/04/89

Étape 4 : déterminer la performance globale d'une chaîne d'information sur la base d'une classification des changements observés, incluant trois types de transformations.

- *Normalisation* : par exemple, insertion ou suppression d'espaces, délimiteurs, modification de formats. Dans notre exemple, on observe une modification de la valeur de l'attribut a à la suite du processus 3 (Tableau 1).
- *Traduction* : changements liés au passage d'un langage à l'autre, d'une codification à l'autre au fil des processus : dans notre exemple, modification de la valeur de l'attribut e à la suite du processus 4.
- *Changements faussement opérationnels* : changements de valeur témoignant la présence d'une erreur dans un processus (modification de la valeur de l'attribut b au processus 3 ou de l'attribut d suite au processus 2). Les origines de tels changements ne peuvent être décelées automatiquement et nécessitent une analyse ultérieure que Redman ne développe pas.

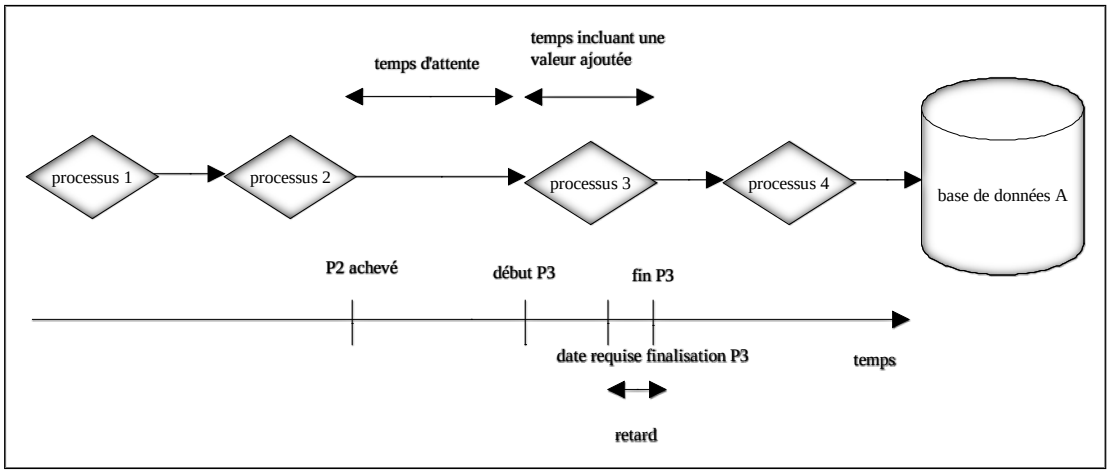
Étape 5: déterminer quelles paires de processus et quelles combinaisons de champs présentent les problèmes les plus importants[16]. Il apparaît que la plupart des changements interviennent au niveau des interfaces entre processus et organisations. L'analyse de Pareto, fréquemment utilisée dans le cadre des audits, est également appelée « principe 80/20 », car elle est basée sur l'hypothèse selon laquelle une part importante des erreurs (environ 80 %) est engendrée par seulement 20 % des causes possibles. Un ordonnancement décroissant des taux d'erreur permet d'identifier les facteurs déterminants et d'en rechercher ensuite les causes[17]. Dans notre cas, le diagramme de Pareto permet d'identifier la part des valeurs jugées problématiques par champ ou par processus. L'objectif consiste ensuite à introduire des processus minimisant les risques d'apparition d'erreur. Nous en envisageons un exemple plus loin dans l'exposé consacré au *reengineering* des processus.

Étape 6 : construire à intervalles réguliers des graphes de contrôle en vue de synthétiser et d'évaluer la conformité aux performances à atteindre et de fournir les bases d'une amélioration. Enfin, le *data tracking* inclut l'étude des écarts temporels entre processus. L'objectif consiste, sur cette base, à déceler les temps d'attente excessifs et à y remédier en révisant les processus concernés et l'organisation correspondante (Figure 2).

[16] PIERCE E., Using P-charts to track data quality (some Observations Based on a Simulation Study). In STRONG D. M. et KAHN B. K., éds, Proceedings of the 1997 Conference on Information Quality. Cambridge : M.I.T., 1997, p. 170-186.

[17] LIEPINS G. E., Reflections on Validation and Quality Assessment of FPC Form 4 Data. In LIEPINS G. E. et UPPULURI V. R. R., éds, Data Quality Control. Theory and Pragmatics (Série "Statistics : Textbooks and Monographs"). New York : Marcel Dekker, Inc., vol. 112, 1990, p. 36-39.

Figure 2. Analyse du temps et des cycles de finalisation des processus (T. Redman)



2.3. Reengineering des processus

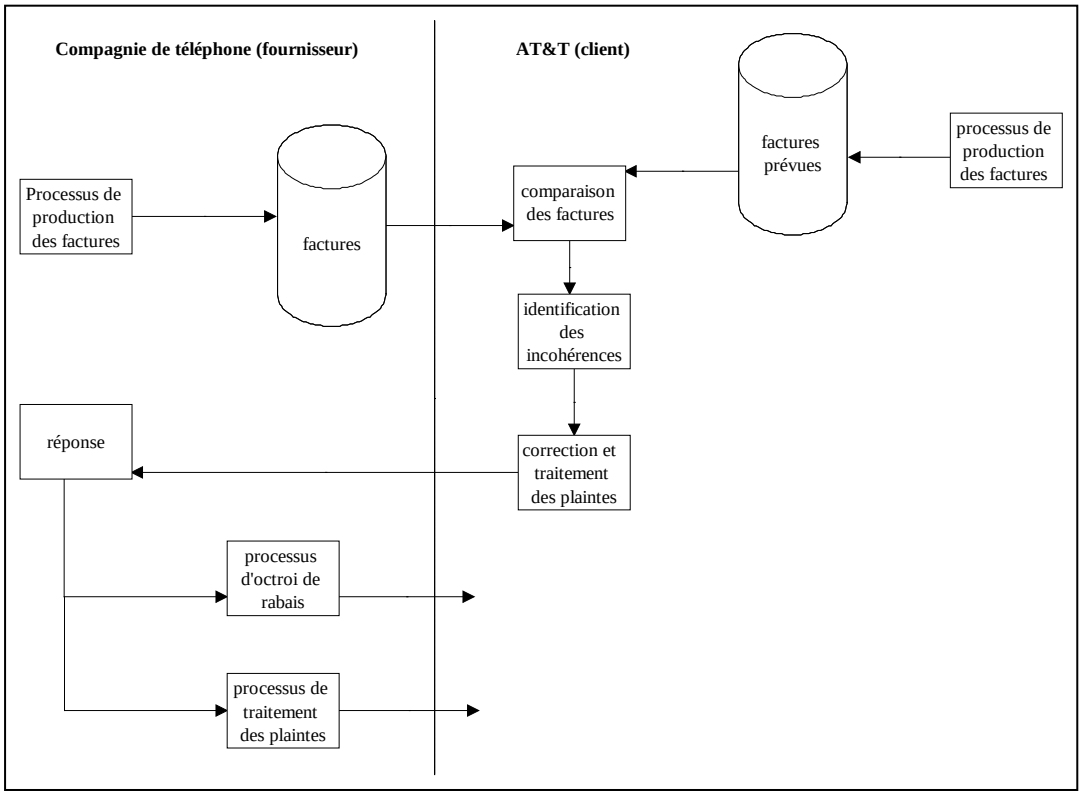
Sur la base du *data tracking*, une démarche complémentaire consiste à concevoir une restructuration fonctionnelle en vue d'améliorer les processus[18]. Cette démarche s'apparente aux techniques du *Business Process Reengineering* (BPR)[19]. Ainsi, un indicateur de l'émergence des problèmes de qualité réside dans l'intensité du travail d'inspection et de correction des données quotidiennement mené au sein d'une organisation. La révision du processus de facturation d'AT&T est extrêmement éclairante sur ce point. Dans les années quatre-vingt, AT&T s'est en effet séparé des quelques centaines de compagnies fournissant l'accès aux services locaux de téléphonie. Dans un premier temps, cette vaste restructuration fut laborieuse, l'identification des connexions entre les lignes et leurs propriétaires n'étant pas toujours aisée. Dès lors, le traitement du processus de facturation entre AT&T et les fournisseurs d'accès locaux fut jugé marginal. AT&T traite en effet régulièrement des factures relatives à l'achat d'accès locaux à des compagnies de téléphonie. Simultanément AT&T vérifie l'exactitude de chaque facture en construisant, à partir de ses propres bases de données, une « facture prévue » (Figure 3). De cette comparaison découlent d'éventuelles incohérences, lesquelles donnent lieu à des plaintes et à des demandes de correction. À leur tour, ces dernières peuvent donner lieu à de nouvelles vérifications (l'origine de l'incohérence n'étant pas toujours aisée à déterminer), au calcul de rabais ou à des « contre-plaintes ».

[18] REDMAN T., Improve Data Quality for Competitive Advantage. Sloan Management Review, winter 1995, p. 99-106. REDMAN T., Data Quality for the Information Age... p. 85-96.

[19] HAMMER M. et *CHAMPY* J., Le reengineering. Paris : Dunod, 1993. Nurcan S., L'apport du Workflow dans une démarche qualité. Ingénierie des systèmes d'information, 1996, vol. 4, n° 4, p. 470-473.

Le ROI est potentiellement important : alors que les montants facturés s'élèvent alors annuellement à environ quinze milliards de dollars, les coûts liés aux opérations de traitement de la redondance et à la vérification des factures, quoique difficiles à évaluer, ne sont pas négligeables : les responsables d'AT&T estiment que 6 % des montants facturés sont dus à la correction des données.

Figure 3. La vérification des factures à AT&T avant réorganisation (T. Redman)



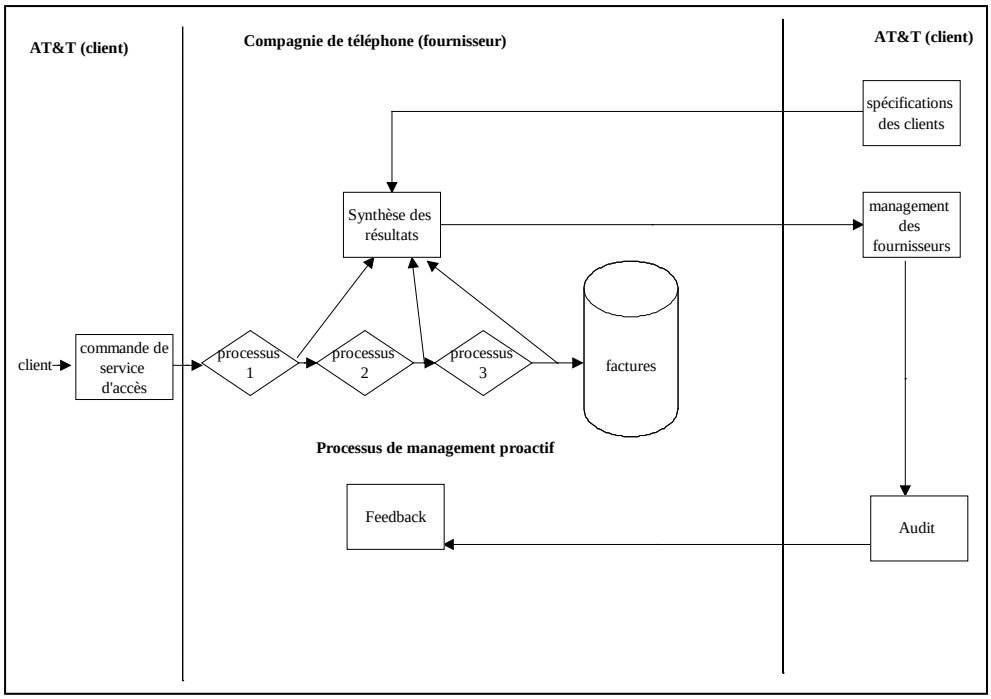
Afin d'améliorer la qualité des données et de minimiser les coûts liés à leur correction, une restructuration des processus fut mise en œuvre sur une période de deux ans en vue de fusionner les différentes bases de données en une seule, rationalisée et contrôlée. La méthode inclut quatre points :

- Un partenariat entre clients et fournisseurs de l'information et un partage de la responsabilité des données et des processus.
- Une gestion proactive des processus impliquant l'usage du *data tracking* et le contrôle des données en amont, au niveau du fournisseur de l'information, avant envoi des factures.
- Une documentation plus claire et plus complète des modes d'établissement des montants.
- Des audits réguliers, en vue d'assurer une validation continue des traitements.

La restructuration (Figure 4) eut quatre conséquences positives.

- Une amélioration de la qualité des factures (liée à la suppression de la redondance initiale).
- La suppression de la procédure de test réalisée par AT&T.
- Une baisse significative des coûts liés à la correction de l'information (gains en personnel et en matériel) et à la gestion des plaintes et litiges : les gains financiers de l'opération furent évalués aux deux tiers du montant antérieurement consacré au traitement de l'information (gains en temps de traitement, en personnel et en montant facturé).
- Un renforcement de la crédibilité de l'information et de la confiance mutuelle entre partenaires.

Figure 4. La vérification des factures à AT&T après réorganisation (T. Redman)



Si la mise en œuvre d'une telle collaboration rencontre des obstacles d'ordre psychologique et organisationnel gommant la distance traditionnelle entre clients et fournisseurs, elle a contribué à responsabiliser chaque partenaire et elle peut être généralisée à d'autres systèmes d'information.

2.4. Évaluation critique

Les points forts de l'approche de Redman résident dans la cohérence entre l'approche théorique et les propositions opérationnelles formulées dans le cadre du *data tracking*. Ces dernières ont mis en lumière l'importance des processus de transformation de l'information et la nécessité d'un instrument d'évaluation continu. En particulier, cet instrument permet de réduire les files d'attente dans les traitements et de diminuer la redondance au sein des bases de données. Les précédents opérationnels qui ont permis de valider la méthode dans le cadre des activités d'AT&T témoignent de son efficacité.

Toutefois, l'analyse des processus laisse ouvertes trois questions. En premier lieu, l'approche est limitée par le fait que *la correction de l'information n'est mesurable que dans le cadre interne d'un langage formel*. Dès lors, la formulation, en termes de taux de correction, des objectifs à atteindre n'inclut pas les éléments informels de l'information : qu'en est-il du processus d'interprétation de valeurs associées à des concepts empiriques complexes, comme la durée du temps de travail, par exemple ? En second lieu et en corollaire, *l'analyse est purement horizontale*, n'envisageant que la suite séquentielle des processus. Or, les critères d'ordre « vertical » relatifs à la sémantique du domaine d'application doivent également être pris en compte. Enfin, l'interprétation, au sein des bases de données, des violations de contraintes d'intégrité[20], n'est pas abordée. Nous avons envisagé ces trois questions dans nos travaux, en complément de l'approche de Redman, et en montrons (point 3) une application pratique généralisable adaptée à l'egovernment.

3. Applications dans le domaine de l'administration fédérale: deux stratégies de gestion

Nous présentons ici deux types de stratégies de gestion originales, inspirées partiellement de l'approche de T. Redman et appliqués au domaine de la sécurité sociale : une application spécifique du « data tracking » (3.1.) et une approche reposant sur l'interprétation des « validations » de violations de contraintes d'intégrité (« anomalies formelles », 3.2.). Les deux approches offrent un ROI important (3.3.) et sont généralisables, non seulement au secteur de l'egovernment, mais aussi à l'ensemble des domaines d'application empiriques (point 4).

[20] Une violation de contrainte d'intégrité désigne une déviance par rapport au domaine de définition d'une base de données, spécifiant l'ensemble des valeurs admises au sein de la base. ELMASRI R. et NAVATHE S. B., Fundamentals of Database Systems, Addison Wesley, 2011 (6e éd.). On parle aussi d'anomalie. Une typologie des cas de figure possible est présentée dans le tableau 2.

3.1. Data tracking & stratégie de traitement des erreurs formelles

Une approche de type « data tracking » est particulièrement indiquée dans le cadre de l'egovernment, qu'il s'agisse des répertoires d'entreprises, des bases de données fiscales ou sociales et ce, dans de nombreux pays[21]. Ainsi, par exemple, en Belgique, la DmfA (*Déclaration Multifonctionnelle - Multifunctionele Aangifte)* permet le prélèvement annuel d'environ 45 milliards d'euros de cotisations et prestations sociales (redistribuées ensuite aux assurés sociaux). Composée de plusieurs centaines de champs, cette base de données, dont les définitions évoluent avec la législation, gère trimestriellement plus de quatre millions d'enregistrements, au sein desquels il n'est pas rare de déceler 10 % d'anomalies formelles. On trouve de telles proportions dans d'autres secteurs, comme le secteur bancaire[22]. Le traitement de ces anomalies formelles (semi-automatique et demandant l'intervention d'agents qualifiés) soulève des enjeux sociaux et de gouvernance colossaux en raison de l'importance des montants en jeu et de la complexité des flux d'information associés.

Dès 2006, le secteur de la sécurité sociale belge a appliqué la méthode du *data tracking* afin d'assurer le suivi des processus au niveau du « top 50 » des employeurs commettant le plus d'anomalies (violations de contraintes d'intégrité) dans les déclarations sociales envoyées à l'administration. L'hypothèse de l'opération repose sur le principe selon lequel un petit nombre d'expéditeurs (employeurs) sont à l'origine d'un pourcentage important d'erreurs formelles (ou anomalies), comme l'illustre la figure 5 (faisant référence au principe de Pareto défini plus haut au point 2).

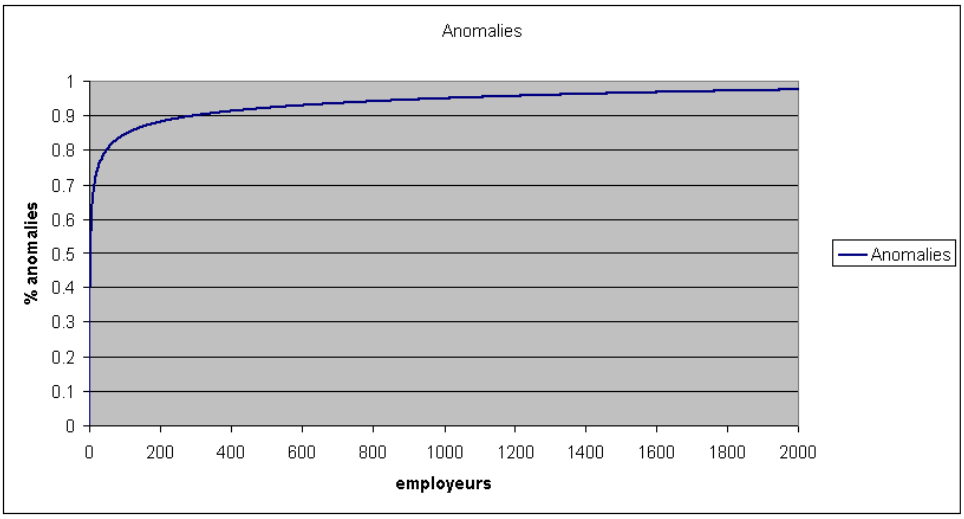
Le « hit parade » (ou « top ») des expéditeurs retenus pour l'analyse peut être adapté selon le type d'expéditeur ou de secteur (public ou privé, par exemple), en fonction de l'évolution des observations, de la législation, des procédures internes et de la population concernée.

---

[21] BOYDENS I., "Strategic Issues Relating to Data Quality for E-government: Learning from an Approach Adopted in Belgium". In ASSAR S., BOUGHZALA I. et BOYDENS I., éds., "Practical Studies in E-Government : Best Practices from Around the World", Springer, 2011, p. 113-130.

[22] «*Recent works about the quality of large databases have shown that about 10% of XML documents (or data records) contain at least one error. This level of quality is unacceptable for many applications*".VAN DER VLIST (E.), Relax NG, Cambridge, O'Reilly Media, 2003.

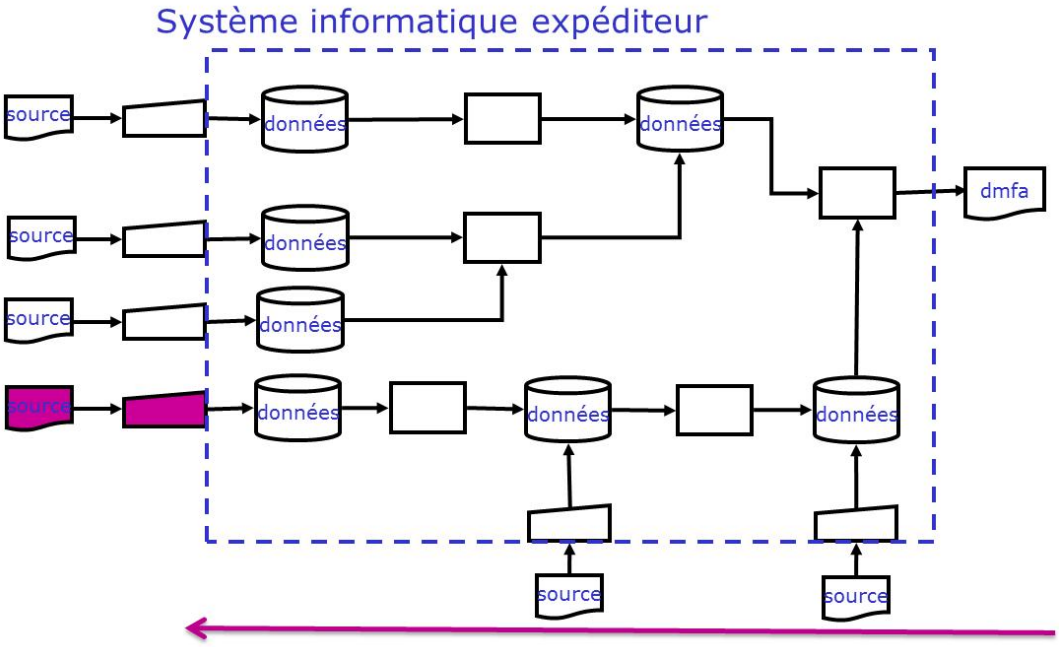
Figure 5. Principe de « Pareto » appliqué aux bases de données administratives



Le but de l'opération consiste à détecter, chez l'expéditeur et en partenariat avec l'administration, les éléments à l'origine de la production d'un grand nombre d'anomalies systématiques (traitement inadéquat de certaines sources de données, interprétation inadéquate de la législation, erreurs de programmation, etc.). Sur cette base, un diagnostic ainsi que des actions correctrices peuvent être posés (correction de code formel dans les programmes, adaptation de l'interprétation d'une loi…). Notons qu'en raison du contexte, l'application présentée ici inclut deux adaptations par rapport à la méthode de Redman :

- L'échantillon d'individus et de cas retenus n'est pas aléatoire puisque l'on dispose d'une connaissance *a priori* concernant les dossiers problématiques, via un historique des anomalies et de leur traitement.
- Il s'agit d'un « tracking arrière » (ou « back tracking ») : on part de la situation finale (base de données) pour revenir, étape par étape, à chaque source et processus qui en a permis l'élaboration. L'objectif est d'éviter le traitement de données ou de flux inutiles pour l'analyse (Figure 6).

Figure 6. « Back tracking »



L'opération permet :

- d'établir un partenariat avec les fournisseurs de l'information en vue d'en améliorer la qualité dans l'intérêt de tous ;
- de mettre en place des solutions structurelles d'amélioration peu coûteuses, ne nécessitant aucun développement logiciel : l'opération lancée en 2006 (« top 50 ») a en effet donné lieu à des résultats très positifs ; la plupart des expéditeurs concernés n'étant, en 2012, plus repris dans le même échantillon du « top 50 » et produisant un taux d'erreurs formelles relativement faible pour la plupart ;
- d'obtenir des résultats potentiellement durables, puisque la cause structurelle des erreurs systématiques est pratiquement identifiée (qu'il s'agisse d'erreurs de programmation ou de problèmes d'interprétation de la législation en matière de temps de travail, par exemple) et peut être théoriquement définitivement réglée.

Toutefois, ce dernier point n'est valable que tant que les conditions « externes » demeurent constantes (voir point 2.4, évaluation critique de l'approche de Redman). Or, toute base de données empirique s'inscrit nécessairement dans un environnement ouvert et changeant au sein duquel l'interprétation de la base évolue avec le traitement des valeurs qu'elle permet d'appréhender. Pour cette raison, il est conseillé de relancer de manière régulière l'opération de « data tracking » en vue de :

- s'assurer de la permanence des résultats obtenus ;
- détecter d'éventuelles nouvelles sources d'erreur ;

- mettre en place un processus continu d'évaluation et d'amélioration de la qualité des données, avec un ROI important.

3.2. Data tracking & interprétation continue des concepts administratifs

L'opération évoquée au point précédent a été relancée en 2012 dans le cadre d'un projet récurrent. Elle inclut également un suivi des « anomalies validées » (c'est-à-dire jugées tout de même correctes au terme d'un examen intellectuel), permettant, au-delà de l'erreur formelle, une adaptation ponctuelle des règles métier à l'interprétation des concepts administratifs.

À cette fin, nous avons enrichi la théorie de la modélisation conceptuelle en distinguant les erreurs des anomalies formelles affectant les données (voir aussi le tableau 2). Les premières constituent une violation de contrainte d'intégrité certaine au sein d'une base de données : par exemple, la présence d'une valeur numérique dans un champ où sont attendues des valeurs alphabétiques. Les secondes s'apparentent à une présomption d'erreur formelle : une incohérence légale apparaît formellement (par exemple, entre la catégorie d'activité d'un employeur et le type d'employés qu'il déclare), mais seule une interprétation humaine (avec une investigation sur le terrain par exemple) permettra de détecter s'il y a erreur ou non et quelle est l'information pertinente. Nous proposons ainsi de passer de l'hypothèse d'un monde clos[23] à celle d'un monde ouvert, sous contrôles automatisés. Afin d'assurer une prise en compte semi-automatique de ces mécanismes, des pré-requis s'imposent.

[23] L'hypothèse du monde clos désigne le fait que toute violation de contrainte d'intégrité est considérée comme une valeur fausse au sein de la base de données. ELMASRI R. et NAVATHE S. B., Fundamentals of Database Systems, Addison Wesley, 2011 (6eme éd.).

Tableau 2. Typologie des violations de contraintes d'intégrité

Une typologie générale des violations de contraintes d'intégrité ou (anomalies formelles) permet de distinguer :

- *Une erreur formelle* « certaine » : par exemple, une valeur numérique apparaît alors que le domaine de définition spécifie des valeurs alphabétiques uniquement.
- *Une présomption formelle d'erreur (que l'on appelle aussi « anomalie »)* : par exemple, si la catégorie d'activité d'un employeur déclarée à un moment t ne correspond pas à la catégorie initialement enregistrée lors de l'immatriculation de l'employeur, on observe une violation formelle de contrainte d'intégrité (anomalie) qui doit ensuite faire l'objet d'une investigation humaine en vue de voir quelle est la catégorie réelle de l'employeur, celle-ci ayant pu évoluer dans le temps. Une analyse intellectuelle est indispensable afin de savoir s'il s'agit d'une erreur et d'élucider le cas.
- *Une erreur indétectable formellement au sein de la base de données* : c'est le cas, chaque fois qu'on est confronté à du « silence », des enregistrements qui devraient figurer dans la base et qui ne s'y trouvent pas (dans le cas du travail au noir, par exemple) ou à des cas de « faux actifs » (entreprises qui ne sont plus actives mais qui figurent encore dans la base sans qu'aucun signe formel n'en fournisse le moindre indice). Seules des inspections sur le terrain permettent la détection de tels cas.

C'est dans ces deux derniers cas de figure que les stratégies de gestion relatives au suivi des validations d'anomalies (point 3.2.) s'appliquent particulièrement.

Ajoutons que les anomalies peuvent être détectées *ex ante*, lors de la saisie des données au sein de la base de données ou *ex post*, après la saisie des données, par exemple lorsqu'il s'agit de détecter des doublons en batch ou de confronter des sources concurrentes.

Ainsi que nous l'avons évoqué en effet, toute procédure d'amélioration, et *a fortiori* toute stratégie de gestion, repose sur un système d'indicateurs de qualité. Celui-ci repose à son tour sur un système de détection d'anomalies « *ex ante* », lors de la saisie, et « *ex post* », après la saisie, en vue de détecter des présomptions de doublons par exemple. Afin de traiter ces anomalies, surtout si l'on se trouve face à un système d'information fédéré, des procédures, validées par toutes les parties, doivent être mises en place (qui traite / quoi / quand et comment). Cette question est souvent délicate dans la pratique, car elle relève de la responsabilité politique de chaque institution concernée. Enfin, un historique des anomalies (par type) et de leurs corrections ou validations est indispensable. Nous avons montré dans une précédente étude publiée

en 2011 comment modéliser un tel historique lors de la conception d'une base de données[24].

Nous présentons un exemple d'exploitation opératoire de tels indicateurs. Le suivi statistique des violations de contraintes d'intégrité (« anomalies formelles ») permet de détecter non seulement les augmentations « anormales » (en fonction d'un seuil donné) d'anomalies, mais aussi les augmentations de « validations » d'anomalies lors de la phase de traitement. Une opération de validation signifie qu'après examen, un agent a estimé que l'anomalie, qui est une présomption d'erreur formelle, correspondait à une valeur pertinente. L'opérateur peut en effet « forcer » le système à accepter la valeur. Si le taux de telles validations d'anomalies est élevé et récurrent, la probabilité est grande que la structure de la base elle-même ne soit plus pertinente. Un algorithme émet alors un « signal » destiné au gestionnaire de la base afin qu'il examine si une modification structurelle de son schéma est requise. Lorsque les cas de validations sont importants, il est intéressant d'approfondir le phénomène : un cas de figure inédit mais représentatif et récurrent est peut-être apparu, ce qui requiert une adaptation de la structure de la base.

Ainsi, par exemple, dans le domaine des bases de données de la sécurité sociale, l'identification de la catégorie d'activité des employeurs est déterminante pour le calcul du taux de cotisations sociales qu'ils doivent payer à l'État. En Belgique, ces cotisations s'élèvent annuellement à 45 milliards d'euros environ. Les enjeux sociaux et financiers sont donc colossaux. Pour catégoriser les employeurs, la législation administrative utilise une nomenclature des activités européennes mise à jour selon une périodicité pluriannuelle. Mais entre chacune de ces mises à jour, la réalité économique ne cesse d'évoluer. Ainsi, quand se développèrent les copy centers, ces boutiques mettant des photocopieuses à la disposition de leurs clients, la nomenclature européenne des activités s'avéra très rapidement inadaptée à leur recensement (elle proposait, au mieux, les catégories statistiques : imprimerie, commerce de détail, de livres ou secrétariat). Afin de prendre en considération la catégorie copy centers, il fallut tout d'abord modifier les textes réglementaires, puis adapter en conséquence la structure des bases administratives.

Citons un autre exemple : en Belgique, la mise en place d'une directive administrative en faveur du secteur « non marchand » a posé la question, au regard de la réalité progressivement appréhendée au sein de la base, de savoir s'il fallait inclure dans ce secteur les maisons de repos privées, a priori exclues car poursuivant des finalités lucratives. Initialement considérées comme des cas « erronés » au regard du domaine de définition spécifiant le secteur « non marchand », ces entreprises y ont finalement été intégrées, après interprétation juridique, sur la base de la méthode présentée ici. Ceci a donné lieu à une restructuration du schéma de la base de données. La restructuration de la base résulte d'une décision humaine tendant à rendre le modèle provisoirement conforme

[24] HULSTAERT A., BOYDENS I., VAN DROMME D., Gestion intégrée des anomalies - Evaluer et améliorer la qualité des données, Delivrable, Section Recherche, Smals, 2011.

aux nouvelles observations. En l'absence d'une telle intervention, l'écart entre la base et le réel se creuse.

Si l'on n'adapte pas le schéma (le schéma incluant non seulement la structure de la base de données et les règles métier mais aussi les listes de contrôle de valeurs admises), les anomalies correspondant à ces cas vont en effet continuer d'apparaître en masse, nécessitant un examen manuel potentiellement conséquent et ralentissant considérablement le traitement des dossiers administratifs. Pour la sécurité sociale belge (s'agissant des déductions de cotisations), la mise en œuvre de cette méthode a permis d'améliorer la précision et la rapidité de traitement des cotisations sociales, réduisant potentiellement de 50 % le volume d'anomalies formelles qui représentaient alors chaque trimestre de 100.000 à 300.000 occurrences à gérer manuellement[25].

Naturellement, les difficultés rencontrées seront d'autant moins difficiles à maîtriser si la base de données a été conçue selon les règles de l'art en matière de « data modelling » et avec une prise en compte des « changements potentiels ». Cela dit, même dans ce cas, comme on ne peut pas « tout prévoir », un suivi de la structure de la base tel qu'évoqué ici reste indispensable.

Sur la base de la stratégie présentée, d'autres indicateurs de suivi des anomalies peuvent être produits en vue de la mise en place de stratégies d'amélioration :

- Le suivi des valeurs nulles (non complétées) permet par exemple d'en détecter via une enquête le motif et d'examiner si les données non complétées sont encore utiles.
- Le suivi de la durée de vie des anomalies et de la rapidité de correction de celles-ci permet d'identifier le moment le plus opportun d'exploitation de la base de données à d'autres fins que celles pour lesquelles elle a initialement été conçue (exploitation statistique, calcul spécifique d'un avantage social ponctuel pour les travailleurs…), par exemple, en fonction du pourcentage toléré d'anomalies résiduelles non traitées d'un type donné.
- Le suivi des dates de mise à jour des valeurs associées aux données (via les « timestamps ») permet d'évaluer la fraîcheur relative de l'information. Dans certains cas, l'absence de mise à jour depuis un laps de temps jugé « long » (plusieurs années par exemple) est un indice d'obsolescence de l'information. Une enquête sur le terrain est alors utile, si la donnée est stratégique, en vue d'en évaluer la validité.

[25] BOYDENS I., « Les bases de données sont-elles solubles dans le temps ? », La Recherche, Sophia Publications, Paris, novembre-décembre 2002, p. 32-34.

3.3. Return On Investment : composantes et évaluations (synthèse)

Dans la présentation du « data tracking » proposée par Redman (point 2), mais aussi plus spécifiquement dans le domaine de l'egovernment, tout au long des études de cas présentées aux points 3.1. et 3.2., nous avons exemplifié et chiffré précisément les éléments concrets du ROI d'une approche de type « data tracking ». Nous en rappelons ici les composantes principales.

Les coûts de l'opération sont peu élevés et ne demandent aucun investissement en logiciel ou en développement. Ces coûts portent sur du « manpower » demandant à une petite équipe spécialisée, tant dans le domaine « métier » de la base de données que sur ses aspects conceptuels et logiques, de préparer l'opération et d'en assurer le suivi, ce qui peut se chiffrer pour une base de données de l'ampleur de la DmfA à maximum trois ou quatre mois/homme. Lorsque l'opération est récurrente, ce qui est conseillé, les coûts sont dégressifs et portent sur l'analyse de l'évolution de la législation et de la population concernée.

Les bénéfices des opérations et stratégies de gestion présentées aux points 3.1. et 3.2. incluent des éléments quantifiables :

- Gains en manpower pour le traitement des anomalies dont le nombre diminue structurellement (chiffrable sur la base des indicateurs générés à partir de la base).
- Gains en manpower pour les expéditeurs de l'information (chiffrable sur la base des indicateurs générés à partir des systèmes d'information externes des expéditeurs).
- Gains en termes de rapidité et de précision de traitement et de prélèvement financier des cotisations sociales et de redistributions de celles-ci aux citoyens (chiffrable sur la base des indicateurs générés à partir de la base).

Ces bénéfices incluent également des éléments qualitatifs non quantifiables :

- Gain en qualité de l'information et du service offert aux citoyens, employeurs et entreprises.
- Gain en crédibilité de l'administration fédérale et renforcement du partenariat entre administrations, entreprises et assurés sociaux.
- Gain en motivation au sein de l'administration fédérale pour les agents en charge de la gestion des bases de données et du traitement - a priori complexe et fastidieux - des anomalies : le « data tracking » conduit en effet à une réflexion « high level » et motivante relative à l'interprétation de la base de données et de son environnement et s'accompagne d'une diminution du travail répétitif que constitue le traitement des anomalies.

4. Conclusions et généralisation

En guise de synthèse, retenons que la pertinence de l'analyse statistique des processus varie en fonction du domaine d'application envisagé et des objectifs de l'analyse. Certains processus relatifs à des éléments répétitifs et dont l'identification formelle est simple s'y prêtent bien. Sur cette base, le *data tracking* est une méthode efficace (point 2) en vue de détecter et de corriger les erreurs formelles affectant les programmes d'application (point 3.1.).

D'autres processus, relatifs au traitement de données dont l'interprétation est problématique, demandent une étude des validations d'anomalies (ou violations de contraintes d'intégrité formelles) présentée au point 3.2. Certaines difficultés sont en effet latentes en l'absence de toute modification de valeur et de toute erreur formellement décelable (comme l'interprétation des catégories d'activité ou de la durée du temps de travail, par exemple). Ces difficultés tiennent à l'interprétation des concepts empiriques dont la définition n'est jamais totalement donnée a priori mais se construit progressivement au fil des traitements[26].

Bien appliqué, le « data tracking » donne lieu à un ROI important, s'il est mis en œuvre dans le cadre d'une organisation et d'une gouvernance adaptées (point 3.3.). Dans certains cas, l'opération demandera en outre de recourir, en complément, aux « Data Quality Tools » en vue d'une aide semi-automatique à la détection et à la correction d'incohérences ou de doublons par exemple[27].

La technique du « data tracking » s'applique ***dans de multiples domaines stratégiques***, les enjeux étant considérables dès lors que l'information est un instrument d'aide à la prise de décision, voire un instrument permettant d'agir sur le réel.

Ainsi, à propos des bases de données de la NASA, en 1986, une équipe de scientifiques britanniques, spécialistes de l'étude du globe, signala la chute des taux d'ozone dans la stratosphère. Sur la base de cette observation, des chercheurs de la NASA réexaminèrent leurs bases de données stratosphériques distribuées de par le monde ; ils découvrirent que depuis une décennie déjà, le phénomène de la baisse des taux d'ozone était resté occulté du fait que les valeurs faibles correspondantes avaient été systématiquement considérées comme des erreurs de mesure. En effet, la théorie scientifique de l'époque, qui avait été modélisée dans leurs bases de données, ne permettait pas de concevoir que de telles valeurs puissent être correctes. À la suite de l'interprétation d'observations empiriques inédites relatives aux réalités du temps court (des taux d'ozone anormalement bas) connues à travers les bases de

[26] BOYDENS I. et VAN HOOLAND S., Hermeneutics applied to the quality of empirical databases. In Journal of documentation, volume 67, issue 2, 2011, pp. 279-289.
http://www.emeraldinsight.com/journals.htm?articleid=1911713&show=abstract

[27] A ce propos voir le "Data Quatilty Competence Center" de Smals : https://www.smals.be/fr/content/data-quality
https://www.smals.be/fnl/content/data-quality

données, la théorie (ou référentiel normatif) fut adaptée aux nouvelles connaissances pour prendre en compte le phénomène (jusqu'alors inconnu) de la baisse des taux d'ozone en certains endroits de la stratosphère.

Le « data tracking » peut s'appliquer également dans le domaine de l'environnement et de l'administration fédérale, en relation notamment avec la difficulté de gérer les référentiels d'adresses[28]. Si les questions liées à la normalisation et à l'identification des adresses spatiales[29] se déclinent distinctement dans des pays dont l'histoire et la culture sont distinctes, comme le Danemark[30] ou l'Afrique du Sud[31] par exemple, la qualité des systèmes d'information correspondants, en tant que référentiels, a un impact sur de nombreuses bases de données auxquelles ils sont interconnectés, dans des domaines stratégiques d'application les plus divers. Ainsi, il est de nos jours parfois extrêmement difficile de détecter rapidement les destructions illégales de bâtiments contenant de l'amiante[32] ou encore, en cas de présomption d'épidémie ou dans le contexte du contrôle de denrées, d'effectuer efficacement le suivi de la chaîne alimentaire d'un pays à l'autre, voire à l'intérieur d'un pays : autant de domaines d'application stratégiques pour lesquels une approche de type « data tracking » est potentiellement bénéfique.

---

[28] COETZEE S., COOPER A., PIOTROWSKI P., LIND M., WELLS M., WELLS E., GRIFFITHS N., NICHOLSON M., KUMAR R., LUBENOW J. & al., What address standards tell us about addresses. SOFocus+Online, June 2010. http://www.iso.org/iso/ru/bonus_ article_addressstandards_biblio.pdf

[29] SALLETS J., La problématique des adresses spatiales dans les bases de données administratives, mémoire de fin de Master en Sciences et Technologies de l'Information et de la Communication, sous la direction de Boydens (I.), Bruxelles, Université Libre de Bruxelles, 2011 ; Prix de l'Association Belge de Documentation, 2012 (http://mastic.ulb.ac.be/2012/05/une-etudiante-du-mastic-primee-par-labd/)

[30] Le Danemark dispose d'un système d'adressage très rigoureux et complet, mais trop rigide pour prendre en compte dans ses bases de données l'hétérogénéité des adresses étrangères, dont la représentation est toutefois indispensable, en raison de la mondialisation des échanges (SALLETS J., Op. cit., p. 80).

[31] Le processus d'adressage sud-africain est fortement lié à l'histoire du pays, avec, depuis la fin de l'apartheid, la prise en compte progressive des Black townships construites en dehors des villes, ainsi qu'avec la réattribution progressive de leurs propriétés aux citoyens dépossédés par les lois raciales. Par ailleurs, comme dans beaucoup d'autres pays du Sud, la prise en compte des Informal Settlements (les bidonvilles) est anarchique et dynamique. Ces mécanismes historiques en cours ont inévitablement un impact sur la qualité des bases de données qui en accompagne les développements (SALLETS J., Op. Cit., pp. 81-90).

[32] VAN VEENSTRA A. & JANSSEN M., "Architectural principles for orchestration of cross-organizational service delivery: Case studies from the Netherlands", in ASSAR S., BOUGHZALA I. & BOYDENS I., éd., Practical Studies in E-Government: Best Practices from Around the world, New York, Springer, 2011, p. 167-185 (chapitre 10).

Le centre de compétences Data Quality fait partie de la section Recherche de Smals. Le centre de compétences peut se targuer d'une **expérience intensive sur le terrain depuis 2004**. Pour la plupart des projets, les membres de la cellule Data Quality travaillent main dans la main avec diverses divisions de Smals, comme la section Développement des applications & Projets, Traitement de l'information ainsi que la section Statistiques ou avec les services de clients et d'institutions membres. Les différentes tâches sont ensuite réparties en concertation avec chacun.

En parallèle avec les missions de consultance autour de la qualité des bases de données administratives des institutions membres, les collaborateurs du centre de compétences donnent aussi des formations et mènent des recherches actives dans ce domaine.

Voir site web de Smals: https://www.smals.be/fr/content/data-quality

Het competentiecentrum Data Quality maakt deel uit van de sectie Onderzoek. Het competentiecentrum heeft een **intensieve ervaring op het terrein sinds 2004**. De leden van de cel Data Quality werken voor de meeste projecten samen met diverse afdelingen van Smals, zoals de sectie Toepassingsontwikkeling & Projecten, Informatieverwerking en de sectie Statistieken, of met diensten van klanten en lidinstellingen. De verschillende taken worden dan in onderling overleg verdeeld.

Parallel met de consultancyopdrachten omtrent de kwaliteit van de administratieve databases van de lidinstellingen geven de medewerkers van het competentiecentrum ook opleidingen en verrichten zij actief onderzoek in dit domein.

Zie website Smals: https://www.smals.be/nl/content/data-quality

La section Recherche de Smals produit régulièrement des publications couvrant de nombreux domaines du marché IT actuel. Vous pouvez obtenir ces publications soit via l'extranet:

http://documentation.smals.be

soit en prenant contact avec le secrétariat de la division « Clients & Services » au 02 787 58 88.