Smals

Evaluer et améliorer la qualité des données

Gestion intégrée des anomalies

Gestions des clients

Section Recherches

Date : mars 2011
Deliverable : 2011/TRIM1/04
Auteurs : Isabelle Boydens, Arnaud Hulstaert, Dries Van Dromme avec la collaboration de Marc Dessart et Elke Durwael

Fonsnylaan 20
1060 Brussel

Avenue Fonsny 20
1060 Bruxelles

Tel : 02/787.57.11
Fax : 02/511.12.42

Management Summary

Comme le confirment les consultances de terrain, ou encore les résultats des travaux scientifiques ou « marketing », comme ceux de Gartner en 2010, la qualité des données revêt toujours des enjeux stratégiques pour l'egovernment. Par exemple, en Belgique, la DmfA (Déclaration Multifonctionnelle – Multifunctionele Aangifte) permet le prélèvement annuel d'environ 40 milliards d'euros de cotisations et prestations sociales. Or, la DmfA est composée de données. On peut en conclure que leur qualité revêt des enjeux fondamentaux en termes de coûts-bénéfices. La question est d'autant plus sensible quand on sait que la législation et les procédures associées à la gestion de ces données sont complexes, évolutives et s'inscrivent dans des organisations hétérogènes. Face à cette réalité, Smals a créé un « data quality competency center » depuis cinq ans environ en vue de réaliser des travaux de consultance destinés à évaluer, à améliorer pratiquement la qualité des bases de données dans le domaine de l'egovernment et à réaliser des études en la matière. Le présent rapport propose plusieurs retours d'expérience originaux issus des travaux du DQ Competency Center. Ceux-ci sont présentés à travers la question des anomalies, ces valeurs déviantes par rapport au modèle attendu d'une base de données et dont on estime que le coût de traitement peut atteindre environ 15 % du revenu des entreprises dans les secteurs privé et public. Le but poursuivi dans cette étude consiste à clarifier la problématique et à présenter des solutions concrètes en vue de diminuer le nombre d'anomalies et d'en faciliter la gestion. Parmi les nouveautés de l'étude, citons les points suivants (chacun étant accompagné d'un modèle organisationnel adapté incluant des rôles métier et techniques) :

(1) La présentation d'un prototype permettant de suivre l'historique des anomalies et de leur traitement (correction, validation…) : celle-ci a été conçue sur la base de propositions originales du DQ Competency Center. Accompagnée d'un travail d'analyse, elle peut donner le jour à un développement grandeur nature adapté à un cas spécifique. Sur cette base, le rapport montre, à partir de « case studies », comment concevoir des indicateurs de suivi de la qualité des données et comment appliquer des stratégies de gestion en vue de diminuer structurellement le nombre d'anomalies en agissant sur la source de leur émergence.

(2) La présentation des aspects documentaires indispensables en vue d'accompagner le traitement des anomalies. Notamment, à partir d'une expérience de terrain, l'étude montre comment un système de gestion des connaissances généralisable permet de faciliter la correction des anomalies par les agents de l'administration.

(3) Un retour d'expérience issu de l'application des Data Quality Tools, acquis par Smals en 2009, à plusieurs bases de données de l'administration fédérale. À partir d'exemples concrets, l'étude montre comment ces outils facilitent la détection semi-automatique d'incohérences formelles et l'accompagnement de leur correction. Elle montre leur apport indéniable par rapport à un développement « home made » et les perspectives d'application.

Table des matières

But et structure du document

Suite aux derniers deliverables qu'elle a produits en 2006, 2007, 2008 et 2009, respectivement sur les « *best practices* » en matière de qualité de données[1], les « *data quality tools* »[2], la question du codage et de la conversion des caractères dans les bases de données[3] et le « *Master Data Management* »[4], la Data Quality Cel propose une mise à jour des résultats de ses travaux, initiés dès 1998[5], alors que la problématique de la qualité des données, déjà très prégnante, était encore peu reconnue.

Ces publications sont le fruit d'études menées au sein de la section « Recherches » en étroite collaboration avec les équipes de développement de Smals, à travers des travaux de consultance menés en vue de traiter les problèmes concrets se posant sur le terrain en ce qui concerne la qualité des bases de données dans le secteur de l'egovernment. Depuis lors et en parallèle, de nouveaux éléments demandent une actualisation de la problématique, sur la base des retours d'expérience :

- des consultances méthodologiques menées au cours de ces dernières années, s'agissant des bases de données de la sécurité sociale, du répertoire des entreprises ou encore du secteur des soins de santé ;
- des premières applications pratiques des Data Quality Tools que Smals a acquis en 2009 ;
- des échanges avec les développeurs, architectes et responsables des données dans le cadre des cours dispensés via le « *chain management* » chez Smals ;
- des recherches, publications[6] et mémoires d'étudiants dirigés dans le cadre du cours « Qualité de l'information et des documents numériques » dispensé à l'Université Libre de Bruxelles.

Le thème de la qualité des données comporte de multiples facettes. Ce rapport a pour objet d'appréhender la problématique à travers la question très pratique de la gestion intégrée des anomalies, ces valeurs formellement incorrectes (incohérences, doublons, valeurs incomplètes…), dont l'émergence est inévitable dans un domaine d'application empirique et dont le traitement est coûteux et pose souvent d'immenses difficultés aux gestionnaires et utilisateurs de bases de données, dans tous les domaines d'application : « *Recent works such as the*

[1] BOYDENS I., *Data Quality : Best Practices*, Deliverable, 2006/trim2/01, Smals, Section Recherches, Bruxelles, 2006.

[2] BONTEMPS Y., BOYDENS I., VAN DROMME D., *Data Quality : tools*, Deliverable, 2007/trim3/02, Smals, Section Recherches, Bruxelles, 2007.

[3] HULSTAERT A., *Préserver l'information numérique. Codage et conversion de l'information,* Deliverable, 2008/trim2/02, Smals, Section Recherches, Bruxelles, 2008.

[4] TRIGAUX J.-C., *Master Data Management - Mise en place d'un référentiel de données*, Deliverable, 2009/trim4/01, Smals, Section Recherches, Bruxelles, 2009.

[5] BOYDENS I., Evaluer et améliorer la qualité des bases de données*, Techno*, n°7, Section Recherches, Smals, 1998 ; BOYDENS I., *Informatique, normes et temps*, Bruylant, Bruxelles, 1999.

[6] Voir par exemple : BOYDENS I., « Qualité de l'information et e-administration : enjeux et perspectives » dans ASSAR S., BOUGHAZALA I., *Administration électronique : constats et perspectives*, Paris, Hermès, 2007, p. 103-120, chapitre 5.

presentation given by Simon Riggs at XML Europe 2003 or the work of Isabelle Boydens (Informatique, normes et temps, Bruxelles, Éditions E. Bruylant, 1999) 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… »[7].

La question soulevée dans cette étude est la suivante : depuis la conception d'une base de données et durant tout son cycle de vie, comment traiter, modéliser et gérer les anomalies formelles dans le cadre d'un système d'information ? Le but poursuivi consiste à en rationaliser la gestion et à en diminuer le nombre afin, *in fine*, d'améliorer les services que doivent rendre les bases de données administratives aux citoyens.

De nos jours, la qualité des données, c'est-à-dire l'adéquation d'une base de données à ses objectifs (« *fitness for use* »), continue à se révéler de plus en plus stratégique sur le plan opérationnel dans le domaine de l'administration fédérale belge. Il s'agit par ailleurs d'un domaine de recherche en constante évolution, ce que confirme Gartner dans sa « *Hype Cycle for Data Management* », publiée le 22 juillet 2010[8]. Longtemps négligé, ce domaine fait appel à un ensemble de techniques mais aussi, plus fondamentalement, à la mise en place d'une organisation en vue de maîtriser et de gérer, de manière continue, la signification des données, leur processus de création et d'exploitation et, ce, en relation avec leur impact stratégique pour le « business » des entreprises[9]. Le succès de la mise en place de technologies, telles que le SOA, dépend étroitement, selon Gartner, d'une prise en compte préalable de la qualité des données.

La structure du document, que nous présentons ci-dessous, met l'accent sur les éléments neufs par rapport aux publications antérieures de la « *Data Quality Cel* », que nous venons de citer.

Dans la première partie, nous rappelons la définition de l'approche « data quality » ainsi que ses enjeux stratégiques en fonction d'une typologie des systèmes d'information administratifs. Parmi les éléments neufs, citons :

- Les possibilités d'application d'une approche « data quality » aux systèmes documentaires[10], lorsque ceux-ci incluent une base de données relationnelle pour la gestion des métadonnées, mots-clés… associés à la description des documents[11].
- L'analyse de la notion de « source authentique », facteur stratégique de prise de décision dans un projet « data quality ».
- Une typologie approfondie des anomalies susceptibles d'affecter une base de données administrative, dont les enseignements opérationnels sont ensuite abordés dans la suite du rapport.

Dans la seconde partie, nous montrons comment les anomalies doivent être prises en considération dans le cycle de vie d'une base de données. Parmi les éléments neufs, citons :

- La modélisation de l'historique des anomalies dans le schéma conceptuel (relationnel) d'une base de données. Le prototype présenté et décrit s'inspire d'une proposition émise dans nos recherches

[7] VAN DER VLIST E., *Relax NG*, Cambridge, O'Reilly Media, 2003.

[8] BEYER M. A., FEINBERG D., FRIEDMAN T. et THOO E., *Hype Cycle for Data Management, 2010*, Gartner, 22 juillet 2010.

[9] BEYER M-A. et al., *Op. cit.*, p. 4.

[10] BOYDENS I., « Hiérarchie et anarchie : dépasser l'opposition entre organisation centralisée et distribuée ? » dans HUDON M. et El HADI W. M., éds, « Organisation des connaissances et Web 2.0 », *Les cahiers du numérique*, Paris, Hermès Sciences, 2010 (à paraître).

[11] BOYDENS I. et VAN HOOLAND S., « Hermeneutics applied to the quality of empirical databases », *Journal of documentation*, Emerald, 2010 (à paraître).

précédentes ; l'exemple présenté montre les conditions de faisabilité de l'approche. Toutefois, appliqué à une base de données « grandeur nature », ce prototype devra faire l'objet d'un développement complémentaire et surtout, s'accompagner d'un effort d'analyse et d'organisation adaptés, ces deux points constituant le socle de l'approche. Celle-ci permet ensuite la mise en place d'indicateurs de suivi continus de la qualité de la base de données et de stratégies de gestion. Ce modèle, dont nous avions présenté les fondements conceptuels en 1999, constitue une extension importante en matière de modélisation conceptuelle. Si le schéma formel d'une base de données repose nécessairement sur l'hypothèse du monde clos (*closed world assumption*), il est sujet à restructurations permanentes car les réalités empiriques qu'il permet de stocker sont intrinsèquement évolutives (notions de catégorie d'activité, d'ouvrier, d'employé…) : ainsi « *la structure d'une base de données évolue avec l'interprétation des valeurs qu'elle permet d'appréhender* »[12].

- Plusieurs stratégies de monitoring des anomalies et de gestion reposant sur le modèle d'historique des anomalies présenté plus haut et sur un retour d'expérience de consultances menées dans le secteur de la sécurité sociale. L'objectif de ces stratégies réside dans la détection rapide de la cause des mécanismes à l'origine des anomalies en vue d'en diminuer la quantité et d'obtenir des gains en termes de coûts-bénéfices.
- Une proposition de modélisation des séquences de contrôle permettant la détection des anomalies (sur la base d'une consultance préalablement menée pour la base de données DmfA de l'ONSS).
- Des recommandations en vue de documenter une base de données, incluant la gestion des données structurées et non structurées[13] et, en particulier, un système permettant la diffusion du « know how » de correction des anomalies (mis en place pour la DmfA, à travers le système Falco).

La troisième partie présente les retours d'expérience suite à l'acquisition des Data Quality Tools par Smals en 2009 : exemples d'application, conseils pour les développeurs et interactions avec le modèle de suivi des anomalies. Cette partie est entièrement neuve. Ce faisant, Smals s'inscrit dans la mouvance des conseils de Gartner : alors que beaucoup d'entreprises continuent de négliger à tort la question de la qualité des données, considérée comme facteur critique pour le business et la prise de décisions, Gartner estime que les organisations les plus matures ont acquis un « data quality tool » en 2010[14]. De plus, Gartner déconseille explicitement un développement « home made » pour remplir les fonctionnalités des DQ Tools, les meilleurs outils présents sur le marché ayant capitalisé, selon Gartner, un vaste ensemble de règles réutilisables et fournissant des bases de données d'adresses internationales difficilement accessibles par une seule organisation[15].

Les différents points présentés s'accompagnent, au fur et à mesure, d'une illustration de l'organisation à mettre en place en vue de soutenir les propositions associées. L'organisation concerne tant les rôles techniques que de type « métier » qui sont tous cruciaux, à l'instar du rôle de « *data stewardship* » dont

[12] BOYDENS I., *Informatique, normes et temps*, Bruylant, Bruxelles, 1999.

[13] BOYDENS I., « Déploiement coopératif d'un dictionnaire électronique de données administratives », *Document Numérique*, Hermes, Paris, 2001, vol. 5, n°3-4, p. 27-43.

[14] BEYER M.-A. et al., *Op. cit.*, p. 25.

[15] BEYER M.-A. et al., *Op. cit.*, p. 26.

Gartner souligne le caractère critique pour la gouvernance des données[16]. De tels rôles doivent être intégrés dans l'organisation globale de toute entreprise productrice de données et sont nécessairement transversaux.

En conclusion, plusieurs perspectives de recherches futures utiles pour la gestion de la qualité des bases de données administratives sont envisagées.

[16] BEYER M.-A. et al., *Op. cit.*, p. 18.

1. Définitions

Au seuil de l'étude, il convient de rappeler et de préciser un ensemble de définitions à portée opérationnelle. Cette étape est indispensable en vue de comprendre le cadre conceptuel dans lequel s'inscrivent les propositions opératoires associées qui seront présentées dans la suite de l'étude. Dans un premier temps, nous rappelons en quoi consiste la qualité d'une base de données (1.1). Les systèmes d'information administratifs sont ensuite envisagés plus en détail (1.2) : les propositions présentées dans cette étude s'appliquent tant aux bases de données structurées qu'aux systèmes documentaires développés dans ce domaine d'application. Par ailleurs, la notion de « source authentique » est présentée et analysée. Enfin, le concept d'anomalie est envisagé et défini plus précisément (1.3) : ce dernier point permet d'établir le raisonnement logique qui donne le jour aux résultats opérationnels présentés dans la suite de l'étude.

1.1. Qualité des données

La qualité des bases de données est aujourd'hui considérée comme un facteur stratégique. La question soulève en effet des enjeux considérables lorsque l'information est un instrument d'aide à la prise de décision, voire d'action sur le réel.

La qualité d'une base de données désigne son adéquation relative aux objectifs qui lui sont assignés. La « qualité totale » n'existe pas car le concept est relatif : sur la base d'un arbitrage de type « coûts-bénéfices », les dimensions de la qualité les plus pertinentes (fraîcheur de l'information, rapidité de transmission des données, précision…) devront être retenues dans un contexte donné. On parle de « *fitness for use* », d'adéquation aux usages.

Ces questions font l'objet de préoccupations croissantes au sein du secteur privé. Ainsi, plusieurs enquêtes menées aux États-Unis indiquent que des facteurs tels que la multiplication de sources hétérogènes partiellement redondantes, de données incomplètes ou mal documentées entraîneraient un coût s'élevant approximativement jusqu'à 15 % du revenu des entreprises. À cela s'ajoutent des coûts accrus lors de la mise en place de nouvelles technologies ainsi que des conséquences en termes de crédibilité auprès des clients ou des utilisateurs.

Alors que la problématique est tout aussi prégnante dans le contexte de l'administration électronique, elle y a trop longtemps été négligée, notamment parce que les bases de données administratives ont souvent été considérées à tort comme des systèmes d'information « simples ». Nous verrons qu'il n'en est rien et que la gestion de tels systèmes est complexe, notamment en raison des questions d'interprétation qu'ils soulèvent. De surcroît, l'information

administrative étant créatrice de droits et de devoirs, la qualité des services en ligne correspondants revêt des enjeux considérables sur les plans sociaux et financiers. En ce qui concerne les symptômes et les causes de la « non-qualité », nous renvoyons le lecteur au deliverable mentionné plus haut « *Data Quality : Best Practices* ». Arrêtons-nous toutefois sur la question des « coûts », question légitimement stratégique pour tout management.

« *Nous avons choisi la qualité parce que la chance était devenue trop chère.* »[17] À l'instar de ce que suggère l'image de « l'usine fantôme », les coûts occasionnés par la « non-qualité » des données sont souvent cachés : « *In large organizations, most of the cost of poor data quality is effectively hidden by the accounting system. Responding to customer complain is just part of customer service* ».[18] Certaines enquêtes révèlent toutefois un taux d'erreur de 5 % à 30 % dans les bases de données[19] se traduisant par une perte d'environ 10 % du revenu des entreprises examinées[20]. Par ailleurs, « *une étude gouvernementale américaine de 2002 établissait que les défauts de qualité logicielle génèrent chaque année 59,5 milliards de dollars de pertes aux États-Unis…* »[21]. C'est ainsi que la question est devenue stratégique : « *Les budgets consacrés à l'évolution des systèmes d'information sont évidemment liés à la santé des entreprises et d'une manière générale à la conjoncture économique. Mais, globalement, la tendance de l'investissement dans les technologies de l'information depuis une trentaine d'années est à la hausse. La bataille entre entreprises concurrentes se joue de plus en plus sur la qualité du système d'information.* »[22] Lors de la gestion de ces systèmes, les coûts de la « non-qualité » touchent les postes suivants :

- Vérification et correction de l'information.
- Traitement des plaintes, procès et réparation des préjudices éventuels. Sur le plan juridique[23], la qualité de l'information revêt en effet une importance certaine, l'exploitation de bases de données inadéquates aux usages pouvant provoquer des dommages d'ordre financier ou médical[24].
- Perte d'image et de crédibilité : les impacts de la « non-qualité » auprès des entreprises et administrations affectent en outre la crédibilité de ces dernières : « *les clients victimes de données incorrectes pardonnent difficilement : une livraison qui arrive trop tard suite à une erreur sur la date de livraison ou une livraison partielle sont hélas des erreurs courantes ; avec internet, la sentence est encore plus rapide ; un internaute qui s'aperçoit que les données d'un site sont peu fiables ou incorrectes, comme par exemple des prix non actualisés, ne reviendra jamais sur le site.* »[25]

[17] CARLIER A., *Management de la qualité pour la maîtrise du système d'information*, Hermès - Lavoisier, Paris, 2006, p. 25.

[18] REDMAN T., *Data Quality for the Information Age,* Artech House, Boston, 1996, p. 8.

[19] Ce taux est évalué sur la base du rapport entre le nombre d'enregistrements contenant au moins une erreur formelle (valeurs incohérentes ou incomplètes) et le nombre total d'enregistrements d'une base de données.

[20] REDMAN T., *Op. cit.* p. 6-14.

[21] CINQUIN L., « La qualité, ennemi juré de la productivité ? », *O1 Informatique*, 16 janvier 2006, n°1841, p. 26.

[22] BRASSEUR C., *Data Management. Qualité des données et compétitivité*, Hermès – Lavoisier, Paris, 2005, p. 22.

[23] WANG R. Y., LEE Y. W. et STRONG D. M., « Can you Defend Your Information in Court ? » dans WANG R. Y., éd., *Proceedings of the 1996 Conference on Information Quality*, M.I.T., Cambridge, 1996, p. 53-64. FRIEDMAN B. et NISSENBAUM H., « Bias in Computer Systems », *ACM Transactions on Information Systems*, juillet 1996, vol. 14, n° 3, p. 330-347.

[24] PADMAN R., « Quality Metrics for Healthcare Data : an Analytical Approach », dans STRONG D. M. et KAHN B. K., éds, *Proceedings of the 1997 Conference on Information Quality*, M.I.T, Cambridge, 1997, p. 19-36.

[25] BRASSEUR C., *Op. cit.*, p. 74.

- Difficulté lors du déploiement de nouvelles technologies, comme le SOA[26] : ainsi en est-il également lors de la mise en œuvre de procédures de *re-engineering* : « *In many reengineering projects the whole point is to get the right data in the right place at the right time so someone can do something for a customer. But if the data that arrive at the right time are themeselves wrong, the operation simply will not serve the customer* »[27]. Enfin, avec le développement croissant des réseaux, les impacts de la non-qualité s'en trouvent accrus : « *the advances of IT have and will continue to exacerbate the impact of poor data quality. First, technology makes data available to more people. Those who are unfamiliar with what data mean or how they were produced are especially prone to misunderstanding and are easily victimized by bad data. This is exacerbated by the natural human tendency to assume that "if it is in the computer, it must be right." »*[28]
- Erreurs de stratégie : « *définir une stratégie prend beaucoup plus de temps si les informations utilisées sont de mauvaise qualité ou tout simplement erronées ; comment réfléchir à l'avenir si la situation présente est mal maîtrisée et surtout mal connue ? Les dirigeants peuvent évidemment se fier à leur intuition mais cela n'est généralement pas suffisant…* »[29]

Notons que de façon générale, la mesure des coûts et gains informatiques est complexe car elle repose sur des jugements et des modèles subjectifs : « *l'informatique n'est pas rentable « automatiquement ». Elle est un outil individuel et collectif, social, dont l'usage dépend de nous et dont l'impact dépend de ces usages* »[30]. Ceci est d'autant plus vrai dans le domaine administratif destiné à assurer l'application du droit : *« le système juridique reste avant tout un système de valeurs et ne repose pas sur un rapport marchand : le droit est mis en œuvre sans contrepartie »*[31]. Par ailleurs, en raison de l'absence de concurrence dans le secteur public, même dans le cas d'investissements plus simples à évaluer, comme la mise en œuvre de nouvelles technologies, il est difficile de calculer les bénéfices d'ordre qualitatif correspondant, par exemple, aux coûts liés à l'achat de matériel et aux formations. En effet, on ne peut en chiffrer les avantages ultérieurs en termes des fluctuations d'assurés sociaux…[32]

Dans le monde des entreprises et des administrations, la qualité de l'information est ainsi devenue un enjeu financier et compétitif important. En dépit des progrès toujours croissants des technologies informatiques, la « qualité des données » est maintenant considérée comme une question cruciale au sein de la communauté informatique. Comme le note T. Redman : « *During the past several decades, managers have expended great effort to stay abreast of the latest information technologies (IT). Despite this, managers still do not have the accurate, timely and useful data they need to be effective. Data failures are embarrassing and costly. Recent published examples include lawsuits filed to protect consumers from incorrect credit reports, incorrect payment of municipal taxes, and rebates*

[26] NEWMAN D. et FRIEDMAN T., « Data Integration is Key to Successful Service-Oriented Architecture Implementations », *Gartner Research Note*, 12 octobre 2005.

[27] REDMAN T., *Op. cit.*, p. 10.

[28] *Idem,* p. 12.

[29] BRASSEUR C., *Op. cit.*, p. 75.

[30] PEAUCELLE J.-L., *Informatique rentable et mesure des gains*, Hermès, Paris, 1997, p. 13.

[31] NAYER A., BORRENS G. et BALTAZAR-LOPEZ S., *L'inspection du travail et la protection juridique du citoyen*, La charte, Brugge, 1995, p. 6.

[32] KRISCHAUSKY D., « Problèmes généraux posés par l'utilisation des technologies de l'information dans la sécurité sociale », dans *L'innovation dans les technologies de l'information : élément important du développement futur des systèmes de sécurité sociale. Huitième conférence internationale sur l'informatique dans la sécurité sociale*, Berlin, 22-24 octobre 1996, Editions de l'Association Internationale de la Sécurité Sociale, Genève, 1997, p. 239-252.

due to incorrect product labeling. No industry - communications, financial services, manufacturing, health care, and so on - is immune. Nor is government »[33].

Pour terminer, insistons sur le fait que toute approche destinée à améliorer la qualité des données est nécessairement pluridisciplinaire et transversale. Elle inclut à la fois des rôles éphémères liés à la durée de vie d'un projet sur le plan « métier » (par exemple, spécialiste des connaissances du domaine, juriste…) et techniques (par exemple, analyste fonctionnel et analyste technique). L'approche inclut par ailleurs des rôles transversaux inhérents à l'organisation en charge de la gestion des bases de données et des métadonnées associées et pérennes dans le temps. Ces points seront illustrés et détaillés dans la suite de l'étude.

1.2. Systèmes d'information administratifs

La dématérialisation de l'information et la mise en ligne, via Internet, de services transversaux pour les citoyens, à la base de l'administration électronique, rendent plus que jamais cruciale la question de la qualité des données. En effet, l'intégration de services requiert la mise en place de sources cohérentes et fiables. Ainsi, afin de permettre à une entreprise de remplir ses obligations fiscales et sociales ou encore, de notifier d'éventuels événements (changement d'adresse, par exemple), via un ensemble de services intégrés sur Internet, il faut disposer en « back office » de référentiels homogènes relatifs à la population cible (les entreprises, dans notre exemple). Ces référentiels doivent être pertinents, précis, à jour et documentés car ils sont susceptibles d'être exploités transversalement par différents services en fonction des besoins. Or, la mise en place de référentiels dépourvus de doublons, de redondance ou d'ambiguïté n'est pas une question simple dans la pratique. De surcroît, dans le contexte d'une exploitation partagée d'informations collectées selon un flux unique, architecture propre à l'administration électronique, un arbitrage entre les besoins des divers services exploitant cette source unique peut se présenter. On observe ce phénomène s'agissant des déclarations multifonctionnelles en ligne développées en Belgique dans le domaine de la sécurité sociale. Pour des raisons légales de paiement de prestations sociales à échéance fixe, certains organismes utilisateurs demandent de disposer de l'information très rapidement, en dépit des anomalies qui peuvent l'entacher. Par contre, d'autres institutions de la sécurité sociale préfèrent que toutes les anomalies formelles aient été traitées avant que les données ne soient diffusées. On se trouve ainsi face à un arbitrage entre la rapidité de diffusion de l'information administrative et sa fiabilité relative, les critères de qualité variant en fonction des usages.

1.2.1. Bases de données structurées

Les bases de données administratives revêtent un ensemble de caractéristiques propres que nous développons ci-dessous : fréquence et nature des modifications législatives, respect de la force probante, volume des données et des anomalies formelles à traiter, enjeux sociaux et financiers et enfin, difficultés d'interprétation.

La structure des bases de données administratives évolue au rythme de l'évolution des directives juridiques correspondantes : ainsi, dans le domaine de la sécurité sociale belge, des modifications législatives impliquant autant de versions de schéma, doivent être implémentées tous les trimestres. La question se complexifie lorsque ces modifications ont un effet rétroactif : quoique contestées,

[33] REDMAN T., « Improve Data Quality for Competitive Advantage », *Sloan Management Review*, winter 1995, p. 99.

celles-ci sont fréquentes dans la vie administrative et ce, dans divers pays européens. Par ailleurs, les versions successives doivent être conjointement maintenues, au minimum, durant la période de prescription, spécifiant la durée durant laquelle les dossiers administratifs doivent légalement être pris en compte. Dans le domaine de la sécurité sociale belge, cette période varie de cinq à trente ans selon les secteurs.

De surcroît, les informations administratives originales saisies dans les bases de données sont pour la plupart dotées du statut de force probante : c'est-à-dire qu'elles font office de preuve devant les tribunaux en cas de litige. Ainsi en est-il par exemple des déclarations trimestrielles envoyées par les employeurs et justificatives des cotisations sociales dues à l'État pour les travailleurs qu'ils emploient. En conséquence, les informations originales, même affectées d'anomalies formelles, doivent être conservées. De même, il s'agit de conserver l'historique de leur traitement (ces anomalies pouvant être corrigées ou validées suite à des inspections de terrain ou à l'interprétation des réglementations) et ce, pour plusieurs versions de schéma. *In fine*, aucune tolérance à l'erreur n'est théoriquement permise au sein de la base. En effet, les citoyens attendent légitimement une gestion équitable de leurs dossiers administratifs, qu'il s'agisse d'impôts à payer ou de droits sociaux à percevoir. Par contre, l'exploitation statistique d'une base de données suppose explicitement une tolérance à l'erreur. Prenons un exemple simple : la somme des rémunérations versées aux travailleurs salariés contribue à l'évaluation de la masse salariale nationale, laquelle permet ensuite le calcul d'agrégats statistiques. Si l'ensemble des enregistrements retenus comporte des inversions de salaires, la rémunération d'un individu A étant indûment attribuée à un individu B, l'évaluation globale n'en sera pas affectée. Par contre, de telles inversions sont dommageables sur le plan des traitements administratifs individuels. En raison de l'importance de ce point, nous reviendrons plus en détail ci-dessous sur la notion d'anomalie.

À cela s'ajoute le fait que les bases de données administratives sont généralement très volumineuses (elles répertorient potentiellement les dossiers de l'ensemble des citoyens d'un État) et que leur gestion suppose des enjeux sociaux et financiers considérables. Ainsi, les bases de données de la sécurité sociale belge répertorient chaque trimestre quelque quatre millions d'enregistrements auxquels correspondent plusieurs centaines d'attributs. On détecte trimestriellement plusieurs centaines de milliers d'anomalies formelles. Ces bases de données permettent chaque année le prélèvement et la redistribution d'environ 40 milliards d'euros.

Enfin, en tant que systèmes d'information empiriques, sujets à l'expérience humaine, les bases de données administratives soulèvent intrinsèquement des questions d'interprétation particulièrement complexes que nous proposons d'étayer dans la suite de ce document.

Au sein des bases de données administratives, on distingue deux grands types de systèmes d'information qu'il est utile de caractériser :

- les bases de données reposant sur un prélèvement régulier d'information ;
- les répertoires.

En effet, l'un et l'autre soulèvent des difficultés de conception et de gestion distinctes.

Bases de données reposant sur un prélèvement régulier

Les bases de données reposant sur un mode déclaratif (comme la DmfA, dans le domaine de la sécurité sociale[34]) sont alimentées selon un rythme régulier, connu

[34] La DmfA (*Déclaration multifonctionnelle/ multifunctionele Aangifte*) permet à l'employeur de transmettre les données de salaire et de temps de travail relatives à ses travailleurs. Cette déclaration est appelée « multifonctionnelle » parce qu'elle est

a priori. Ainsi, dans le cadre de la DmfA, les déclarations envoyées par les employeurs et justificatives des cotisations dues sont envoyées trimestriellement (elles ont en ce sens un statut de source authentique sur le plan juridique).

Du fait de ces contacts réguliers avec la population « cible », l'information est régulièrement mise à jour et relativement « fraîche » (même si des problèmes de « silence » peuvent se poser avec le travail au noir, par exemple). La difficulté majeure ne réside donc pas dans le caractère potentiellement obsolète de l'information.

Par contre, ces bases de données, répertorient un grand nombre de champs (plusieurs centaines) : elles sont destinées à appliquer une législation de plus en plus complexe et changeante (la législation évoluant tous les trimestres dans le cas de la DmfA). La difficulté principale que soulève ce type de système d'information réside donc dans la complexité et la fréquence des mises à jour du schéma de la base, les problèmes d'interprétation qui en découlent et, en corollaire, la gestion des anomalies engendrées par ces difficultés.

Répertoire

À l'inverse du cas précédent, les répertoires (appelés aussi « référentiels » ou sources authentiques), tels que la Banque Carrefour des Entreprises, en Belgique, sont alimentés sur la base de contacts irréguliers avec la population « cible ». L'alimentation de ces systèmes repose sur la communication ponctuelle d'événements : fusion d'entreprise, changement d'activité principale, changement d'adresse, autorisation d'euthanasie ou de don d'organes, etc. Notons que ces événements ne sont pas toujours simples à interpréter (il faut parfois des spécialistes pour déterminer l'activité principale d'une entreprise : s'agissant du répertoire des entreprises françaises, Sirène, un opérateur avait un jour inscrit sous « ferronnerie » une entreprise qui avait décrit son activité en mentionnant l'expression « portails d'entreprises », désignant naturellement le domaine informatique…)[35]. Mais le point le plus important est que l'information stockée dans les répertoires sera potentiellement plus obsolète du fait de cette communication ponctuelle d'événements. On trouvera dès lors davantage de cas de doublons, de « faux actifs », de silence ou d'information caduque que dans le cas précédent. Par ailleurs, une entreprise qui a fait faillite ou, pire, qui a pris la fuite à l'étranger n'aura pas pour principale préoccupation de communiquer un changement d'adresse ou de variable à l'administration.

Un répertoire, en tant que « source authentique », peut s'apparenter à une « pompe aspirante-refoulante » : elle répertorie de l'information qui fera autorité et servira à alimenter les bases de données évoquées au point précédent. À cet égard, une difficulté « type » touchant les répertoires ou « sources authentiques » réside dans leur chargement initial. Par la force des choses, avant la création d'une source authentique, il n'existait pas de source analogue. On a donc recours, pour l'alimentation initiale, à un ensemble hétérogène de fichiers de données dont les définitions ne sont pas toujours compatibles et qui incluent potentiellement des doublons. En France, il a été estimé qu'une période de dix ans fut nécessaire pour « nettoyer » le répertoire des entreprises Sirène des conséquences

utilisée par de nombreuses institutions. En effet, la DmfA ne se limite pas à la déclaration et au calcul des cotisations de sécurité sociale dues et à celui des réductions. Elle constitue aussi la source des données pour les institutions de sécurité sociale chargées de l'attribution des droits dans la sécurité sociale et du paiement des indemnités. Les secteurs suivants font usage de ces données : l'assurance maladie, le chômage, les pensions, les accidents de travail, les maladies professionnelles, les allocations familiales et les vacances annuelles.

[35] RIVIERE P., « Qualité des données et processus de recueil », Conférence présentée au CNAM, Conservatoire National des Arts et Métiers, lors des journées d'études CNAM-CSML, *La qualité des données à l'ère de l'information*, Cnam, Paris, 11-12 mars 2003.

engendrées par l'alimentation initiale[36]. En effet, à la base, on n'échappe jamais à un compromis, nécessairement imparfait, entre les besoins et les sources disponibles.

À l'inverse des bases de données reposant sur une alimentation régulière et évoquées au point précédent, les répertoires ou sources authentiques sont dotés d'un schéma plus stable. Les champs repris sont des catégories dont la définition fera l'objet de moins d'évolutions législatives (adresse, forme juridique, catégorie d'activité, etc.). Plus stables, les champs sont aussi moins nombreux. Dans le cas d'un répertoire ou d'une source authentique, l'exhaustivité des enregistrements prime sur la précision ou le détail du schéma. Toutefois au niveau des formes juridiques ou des catégories d'activité, les questions d'interprétation sont parfois complexes également.

1.2.2. Systèmes d'information documentaire

Toutes les applications d'informatique documentaire sont potentiellement concernées par la présente étude dans la mesure où :

- elles incluent nécessairement dans leur architecture un Système de Gestion de Base de Données (SGBD) en vue d'en stocker les métadonnées descriptives : mots-clés, champs d'identification, noms d'auteur, dates, types de document…[37]
- elles sont, au même titre que les bases de données structurées, soumises au principe du « fitness for use »[38], étant parfois de surcroît dotée du statut légal de force probante dans le cas du dépôt légal (par exemple le dépôt d'archives publiques soumises à contrainte réglementaire, les archives notariales ou encore le futur système d'archivage de SIGeDIS pour les contrats de travail électroniques[39]).

En raison même du concept de « fitness for use » et des contraintes de type « coûts-bénéfices » associées aux enjeux de la gestion du système d'information documentaire concerné, les mesures d'évaluation et d'amélioration de la qualité présentées dans le présent rapport pourront être d'application, selon les cas.

Notons déjà que les propositions présentées ici ont d'ores et déjà fait l'objet d'une application concrète avec succès dans le cadre des métadonnées associées à plusieurs systèmes documentaires d'envergure dans d'autres domaines d'application.

Citons l'application de procédures de « data profiling » et de documentation des données dans le cadre du système documentaire géré par

- le *Musée Royal de l'Afrique Centrale* à Bruxelles[40] ;
- le *National Archives of the Netherlands* aux Pays-Bas[41] ;

[36] GARAGNON J., « Sirène, système informatique pour le répertoire des entreprises et des établissements. Situation actuelle et développements en cours », *Courrier des statistiques*, janvier 1983, n° 25.

[37] BOYDENS I., *Documentologie*, Bruxelles, Presse de l'Université Libre de Bruxelles, 2010-2011 (syllabus, dernière édition).

[38] Voir par exemple HULSTAERT A., *Les systèmes documentaires en ligne dans le domaine de la presse écrite. Conception d'une grille d'analyse de la qualité des sources et confrontation à la mise en place d'un système*, Mémoire de fin de Master en Sciences et Technologies de l'information et de la communication, Bruxelles, ULB, 2006-2007.

[39] HULSTAERT A., *Préservation à long terme de l'information numérique. Rendre l'information accessible durablement*, Deliverable, 2010/trim1/01, Smals, Section Recherches, Bruxelles, 2010, p. 18-19 et 67-77.

[40] VAN HOOLAND S., KAUFMAN S., BONTEMPS Y., « Answering the call for more accountability: applying data-profiling to museum metadata », *Proceedings of the 2008 International conference on Dublin Core and metadata applications*, Berlin, 22- 26 Septembre 2008, p. 93-103. VAN HOOLAND S., *Metadata quality in the cultural heritage sector: stakes, problems and solutions*, Thèse de doctorat, Université Libre de Bruxelles, 2009. BOYDENS I., « Hiérarchie et anarchie : dépasser l'opposition entre organisation centralisée et distribuée ? » dans HUDON M. et El HADI W. M., éds, « Organisation des connaissances et Web 2.0 », *Les cahiers du numérique*, Paris, Hermès Sciences, 2010 (à paraître).

- Le *September 11th Memorial and Museum* de New York[42].

Dans le cas du *Musée Royal de l'Afrique Centrale* à Bruxelles, il a été possible d'identifier, à partir de ce mécanisme, les écarts formels entre les définitions structurellement attendues et les valeurs saisies dans la base de données descriptive. Une analyse de structure permet par exemple d'effectuer une typologie de tous les formats de date présents au sein du système et de quantifier chaque type : par exemple, champs vides, format de type 9999-9999 (par exemple : 1891-1912), de type AAA 9999 (par exemple : *mai 1938*), de type 99/99/9999 (par exemple : *04/08/1964*)… Cette analyse est fondamentale en raison de l'incertitude associée à la datation en histoire et histoire de l'art.

Une automatisation partielle de cette approche permet d'aider les gestionnaires du système d'information à documenter la base de données. La prise en considération de la variété des formats de dates rencontrés et de leur fréquence contribue à évaluer et à améliorer l'interprétation de l'information. L'approche peut être enrichie en fonction, par exemple, de la typologie des objets datés (statuette, panier, masque, couteau…) et de leur contexte. Une interprétation *a posteriori* des modalités de datation passées pourra alors être effectuée : pourquoi a-t-on eu recours à un intervalle, à quoi sont dus les champs vides (valeur incertaine, négligence lors de la saisie…) ? Avec l'évolution des recherches et des idéologies, les modes de description des objets ethnographiques ont évolué dans le temps. La modélisation de la base de données a pu faire l'objet de contraintes d'intégrité plus ou moins rigoureuses, en fonction des fichiers disponibles et des gestionnaires de ceux-ci. Dès lors, il est utile, à des fins scientifiques, de prendre l'historique de ces évolutions en considération dans la définition des métadonnées descriptives.

On peut citer les *National Archives of the Netherlands* aux Pays-Bas qui ont déployé pour leur fonds photographique un système de co-construction permettant aux utilisateurs de commenter et de compléter librement les métadonnées associées aux photographies. Alors que certains commentaires sont peu pertinents, d'autres sont extrêmement riches quant à l'identification de l'origine historique d'un document. Dans l'esprit de la méthode présentée dans cet article, ce fonds a fait l'objet d'une analyse critique spécifique (« data profiling » couplée à une approche d'évaluation statistique) sur la base d'une grille d'analyse qualitative objectivable. Ainsi, à partir des fichiers de transaction, une analyse critique comparative a permis de confronter le questionnement sur les métadonnées aux commentaires associés à ces dernières et d'évaluer la pertinence relative de ces commentaires. L'analyse des corrélations statistiques entre commentaires et requêtes contribue à identifier les classes de commentaires les plus pertinentes sur le plan scientifique. Près de 50 % des requêtes se centrent sur les corrections de métadonnées décrivant des photographies (corrections formelles ou de fond relatives à l'identification d'événements, à leur localisation géographique ou à leur datation) ou à des éléments narratifs contextuels. Par contre, les commentaires relatifs à des jugements de valeur individuels occupent une place marginale (moins de 3 %) dans les requêtes. Les *tags* co-construits sont ainsi dotés d'un véritable apparat critique qui permet de guider l'interprétation des métadonnées correspondantes.

Une approche analogue a été initiée en 2009 au *September 11th Memorial and Museum* de New York. Un prototype y a été développé en vue d'adapter de manière dynamique la structure du modèle de représentation aux usages. Celui-ci s'inspire des modalités de *drag and drop* des blocs d'information qu'offre l'interface *iGoogle*. Il permet aux utilisateurs d'adapter et de personnaliser (par ajout, suppression ou restructuration) les champs de métadonnées par défaut. La

[41] VAN HOOLAND S., *Idem.* BOYDENS I, *Idem.* BOYDENS I. et VAN HOOLAND S., « Hermeneutics applied to the quality of empirical databases », *Journal of documentation*, Emerald, 2010 (à paraître).

[42] VAN HOOLAND S., *Idem.* BOYDENS I, *Idem.* BOYDENS I. et VAN HOOLAND S., *Idem*.

consultation et les usages des différents champs descriptifs évoluent en effet avec la découverte de nouveaux objets (découverte de tel camion de pompiers calciné ou de tel nouveau type de débris d'armature métallique issu du World Trade Center) et la nécessité de les décrire. Le prototype a été développé en collaboration avec l'équipe du logiciel libre CollectiveAccess[43]. Le suivi des fichiers de transaction associés à l'interface de recherche incluant les champs de métadonnées descriptives permet alors aux gestionnaires du système documentaire d'évaluer l'évolution des usages et d'optimiser de manière semi-automatique l'interface de recherche. Des utilisateurs « beta » ont testé le mécanisme : le caractère extrêmement évolutif et récent du *September 11th Memorial and Museum* de New York en fait un cas d'étude idéal et le module testé sera prochainement intégré 2011 dans d'autres applications de métadonnées culturelles.

Ces approches sont applicables dans le domaine de l'egovernment. D'autres solutions opérationnelles proposées dans la suite du présent rapport pourront s'appliquer tant aux bases de données structurées qu'aux systèmes d'information non structurés selon les enjeux qu'ils soulèvent. Pour ne citer qu'un exemple, lors d'un reenginering d'une application documentaire, il peut être utile, via les techniques du « data profiling » présentées plus loin, d'effectuer un audit des métadonnées telles qu'elles ont été saisies par les utilisateurs. On pourra ainsi identifier les mots-clés trop souvent utilisés, peu utilisés ou les usages non conformes selon les besoins. Cette première étape permettra d'établir une documentation permettant un meilleur usage des données descriptives du système.

1.2.3. Notion de « source authentique », un concept pragmatique

Dans de nombreux projets relatifs à la qualité des données administratives, la notion de « source authentique » peut s'avérer stratégique car elle désigne en première instance les fichiers de référence faisant autorité et sur la base desquels les autres bases de données associées pourront être corrigées. Cette notion peut donc orienter des choix importants en termes de flux d'information, d'organisation ou de modélisation de données et de contrôles. Cela dit, la notion de « source authentique », citée dans la loi, n'est pourtant pas définie rigoureusement sur le plan juridique. Nous proposons d'exposer brièvement la situation actuelle en la matière en Belgique car il est important, lorsque l'on aborde la question de la qualité des données, d'avoir conscience de ce flou « conceptuel ».

En effet, dans la pratique, on peut être rapidement confronté à certains paradoxes importants : le « registre national » censé être une « source authentique » en matière d'identification des citoyens ou encore la « banque carrefour des entreprises », censée être une source authentique en matière d'identification des entreprises, peuvent inclure des erreurs ou présomptions d'erreurs formelles (doubles, erreurs orthographiques, adresses inadéquates…) dues à des problèmes issus des flux d'information, liés à la saisie initiale ou encore à la complexité de l'interprétation conjointe de la loi et des faits. On se trouve forcé de traiter *a posteriori* ces cas problématiques, notamment en comparant, à des fins d'investigations, ces sources « dites authentiques » avec d'autres bases de données commerciales, par exemple Graydon dans le domaine des entreprises qui, par la force des choses, n'ont pas le label « source authentique ». Le poids du label « source authentique » perd ainsi d'emblée de son aura pour des raisons pratiques inhérentes à la réalité empirique.

[43] Collective Access – The Open Source Collections Management and Cataloguing System for Museums and Archives. Site web du projet : http://www.collectiveaccess.org/

La définition de Fedict est rédigée dans le cadre de leur service « Digiflow » (sources des services publics) : « *Pour rappel, les sources authentiques sont les données enregistrées par le service public responsable de leur exactitude et offertes à la consultation d'autres acteurs. Pour chaque type de données, la source qualifiée d'authentique constitue donc la référence, garante d'une* ***information la plus exacte et la plus récente possible****. La consultation des sources de chaque administration publique est soumise à son* ***autorisation expresse selon des règles strictes de finalité et de proportionnalité.*** *En d'autres termes****,*** *la consultation ne porte que sur des informations ciblées, à l'usage d'institutions ou des départements précis, dans un but bien défini.* »[44]

Notons que cette définition prend en considération le paradoxe évoqué au seuil de ce paragraphe puisqu'elle ne parle pas de « qualité parfaite », mais d'information « la plus exacte et la plus récente possible », en référant à l'instance en charge de cette mission.

La définition de **la plate-forme eHealth**[45] est plus générale mais limitée au secteur des soins de santé : « *les sources authentiques validées sont des bases de données de fond, gérées par les acteurs du secteur des soins de santé ou des prestataires de service ICT qu'ils se sont choisis. Ils peuvent utiliser ces sources lors de l'exercice de leur fonction dans le secteur des soins de santé* ».

En **Région wallonne, pour le Commissariat Easi-Wal, E-Administration et simplification**[46] « *Aucune définition n'a encore été juridiquement arrêtée. Voici celle qu'EASI-WAL propose : une source authentique est tout service public dépositaire de données de références instituées en vertu d'une disposition légale ou réglementaire, à qui des administrations reconnaissent le rôle de gestionnaire unique pour lesdites données dont elles ont besoin, et qui réglemente l'accès à ces données. Plusieurs sources authentiques fédérales sont déjà définies ou en cours de finalisation, comme la Banque-Carrefour des Entreprises (BCE) ou la Banque-Carrefour de la Sécurité Sociale (BCSS). Des organisations d'autres niveaux de pouvoir seront amenées également à être reconnues comme sources authentiques. Cette définition* ***pose clairement la distinction*** *à opérer entre le* ***service gestionnaire*** *et* ***la donnée*** *(ou l'ensemble de données) lorsqu'on parle de source authentique. Le principe de la source authentique des données est un élément fondamental de l'e-governement. Comme la définition ci-dessus le laisse entendre, il implique qu'il est possible d'identifier,* ***pour chaque donnée importante*** *(ex. numéro de registre national, délivrance d'un permis d'environnement, numéro de TVA, ...)* ***un et un seul service*** *administratif qui en est la* ***source*** *et qui est chargé d'en assurer* ***la gestion, à savoir le stockage et la mise à jour, en tenant compte autant que possible des besoins des autres services administratifs.*** *Les services administratifs qui ont besoin de cette donnée doivent se la procurer auprès de la source qualifiée « d'authentique »* ***plutôt que*** *de la reproduire de leur côté, d'en effectuer leur propre mise à jour et donc de* ***risquer d'introduire des incohérences et surtout des redondances d'informations. (…)*** **»**

Au niveau du Vlaams e-government[47], la définition proposée est elle aussi très pragmatique : « ***Het is zinloos dat elke dienst alle gegevens zelf wil beheren****. Dit slorpt enorm veel* ***tijd*** *en* ***middelen*** *op,* ***zonder dat de gegevens voldoende kwaliteitsvol zijn****.* ***Bovendien moet elke dienst hierdoor telkens weer dezelfde gegevens opvragen bij de klant****. En uiteraard is de klant hierover niet tevreden. Dat geldt voor elke organisatie, dus ook voor overheidsdiensten.* ***Gegevens die voor meerdere diensten belangrijk zijn, kun je toegankelijk maken via een***

[44] http://www.fedict.belgium.be/fr/binaries/AppIM_DIGIFLOW_FR_Pr%C3%A9sentationServCatal_V2.0_tcm166-63051.pdf

[45] Site web de la plateforme eHealth : https://www.ehealth.fgov.be (consulté le 9 août 2010).

[46] Site web du « Commissariat Easi-Wal, E-Administration et simplification » : http://easi.wallonie.be/easi/col_gauche niveaux_fr/easi-wal/dossiers-thematiques/sources-authentiques/index.html?LANG=fr (consulté le 11/06/2010).

[47] Site web de la « Coördinatieceel Vlaams e-government » : http://www.corve.be/ (consulté le 11/06/2010).

kruispuntbank. *Deze gegevens worden beschikbaar gemaakt voor de belanghebbende overheidsdiensten, en moeten dus* ***maar één keer ingezameld*** *en bijgehouden worden.* ***De kans op foute of achterhaalde gegevens wordt zo meteen veel kleiner***. ***Die bron, waar de oorspronkelijk ingezamelde, correcte én volledige informatie zich bevindt, noemen we de authentieke gegevensbron***. *Alle diensten moeten daar informatie ophalen. Authentieke gegevensbronnen vormen de spil van het hedendaagse databeheer.* ***Door allen deze data te gebruiken en eventuele fouten te melden, krijgen we steeds betere gegevens*** **»**.

Les définitions actuelles de la notion de « sources authentiques » sont donc plurielles, pragmatiques et sectorielles. En général, elles renvoient à **une autorité, une instance dont la mission se traduit par un service.** La notion est donc **stratégique** non seulement sur le **plan conceptuel** mais aussi sur le **plan fonctionnel des flux d'information** et des **processus.**

1.3. Concept d'anomalie

À ce stade, nous envisageons plus précisément en quoi consiste la notion d'anomalie dont l'utilité opérationnelle est rappelée au seuil de ce paragraphe. La notion est ensuite abordée à travers la réponse aux trois questions suivantes : « Qu'est-ce qu'une donnée ? », « Qu'est-ce qu'une donnée correcte ? », « Comment les données se construisent-elles progressivement ? ».

1.3.1. Utilité opérationnelle d'une définition claire

Toute démarche d'amélioration de la qualité de l'information implique la spécification préalable d'indicateurs. Ceux-ci permettront d'évaluer ultérieurement les progrès enregistrés eu égard aux objectifs poursuivis[48]. Afin d'identifier les indicateurs les plus adéquats, il convient préalablement d'examiner la nature de l'information traitée.

Cette clarification conceptuelle est d'autant plus capitale que de nombreux ouvrages en matière de qualité des données reposent sur l'hypothèse selon laquelle il existerait une relation biunivoque nécessaire entre les données et le réel observable correspondant. Cette hypothèse ne se vérifie pas dans la pratique et sa prise en compte mène à des propositions opératoires inadéquates. La mise en place d'un système d'indicateurs permettra également d'éviter certaines dérives qui sont parfois observées dans la pratique, la question de la qualité des données étant détournée à d'autres fins. Ainsi aux États-Unis, le « Data Quality Act »[49] a fait l'objet de nombreuses critiques : cette législation permet à une instance de déroger à certaines mesures si les chiffres avancés en guise de preuve sont jugés de qualité insuffisante, la notion de qualité n'étant pas définie ! C'est ainsi que certaines entreprises accusées de produire des herbicides délétères pour l'environnement et la santé publique ont pu échapper aux sanctions prévues. Afin d'examiner la nature de l'information traitée (l'information administrative dans notre cas) et d'établir ultérieurement un système d'indicateurs de suivi de la qualité adéquat, nous abordons les notions de donnée et d'anomalie. Dans de nombreux projets, qu'il s'agisse de la conception d'une nouvelle base de données ou du « reengineering » de bases existantes, ces définitions ne sont pas triviales et ont un impact direct sur le travail d'analyse soutenant le projet.

[48] RIVIERE P., « Approche coût-qualité pour l'amélioration des processus de production statistique », *Courrier des statistiques*, juin 2003, n° 105-106, p. 55-65.

[49] http://www.thecre.com/misc/20040606_worms.htm et
http://www.washingtonpost.com/wp-dyn/articles/A3733-2004Aug15.html

Nous proposons d'envisager successivement les trois questions suivantes afin d'identifier le processus le plus adéquat de constitution des indicateurs de qualité :

- Qu'est-ce qu'une donnée ?
- Qu'est-ce qu'une donnée « correcte » ? Cette seconde question nous permettra de proposer une typologie générique des anomalies.
- Comment l'information se construit-elle progressivement ?

1.3.2. Qu'est-ce qu'une donnée ?

Une donnée est un triplet (i, d, v) composé des éléments suivants : un intitulé (i), renvoyant à un concept (une *catégorie d'activité*, par exemple), un domaine de définition (d), composé d'assertions formelles spécifiant l'ensemble des valeurs admises dans la base pour ce concept (une liste contrôlée de valeurs alphabétiques, par exemple) et, enfin, une valeur (v) à un instant t (*le secteur de la chimie*, par exemple). À cela s'ajoutent les interactions entre les différentes composantes du schéma de la base de données que nous n'envisageons pas ici.

Il importe de distinguer les *données déterministes* des *données empiriques*. Les premières se caractérisent par le fait que l'on dispose à tout moment d'une théorie qui permet de décider si une valeur v est correcte ou pas. Ainsi en est-il des données algébriques : les règles de l'algèbre n'évoluant pas dans le temps, on peut savoir à tout moment si le résultat d'une somme est correct ou pas. Par contre, en ce qui concerne les données empiriques, sujettes à l'expérience humaine, la théorie évolue dans le temps avec l'interprétation des valeurs qu'elle a permis d'appréhender. Ainsi en est-il par exemple du domaine médical (où la théorie évolue au fil des expériences, comme en témoignent les recherches actuelles sur la grippe aviaire), du domaine économique (en ce qui concerne l'évaluation de la richesse nationale, par exemple) mais aussi des domaines juridique et administratif où l'interprétation des concepts légaux se transforme avec l'évolution continue de la réalité traitée et avec celle de la jurisprudence. Par exemple, quand se développèrent les « copy centers », ces boutiques mettant des photocopieuses à disposition de leurs clients, la nomenclature des activités européenne (utilisée dans les bases de données administratives pour catégoriser les entreprises) s'avéra rapidement inapte à leur recensement : elle proposait au mieux les catégories « imprimerie », « commerce de détail de livres » ou « secrétariat ». Afin de prendre en considération la catégorie « copy centers », il fallut d'abord modifier les textes réglementaires, puis adapter la structure des bases administratives en conséquence. Ceci demeure crucial dans le contexte de l'administration électronique. Ainsi, la notion d'activité principale d'une entreprise, fondamentale au sein du répertoire français des entreprises (Sirène), est une notion évolutive dont la fiabilité est difficile à évaluer.

1.3.3. Qu'est-ce qu'une donnée « correcte » ?

Pour des raisons opérationnelles évidentes, le fonctionnement d'une base de données repose sur *l'hypothèse du monde clos*[50], en vertu de laquelle toute valeur non incluse dans le domaine de définition est considérée comme fausse : on parle alors de *violation de contrainte d'intégrité*. Toutefois, s'agissant des données empiriques, si l'on sort de ce cadre formel, il se peut qu'entre le moment où la structure de la base de données a été formalisée et celui où l'information a été saisie, de nouvelles caractéristiques soient apparues au sein du domaine traité (contrairement à ce qu'affirment certaines théories postulant une relation biunivoque permanente entre les données et le réel empirique correspondant).

[50] ELMASRI R., NAVATHE S., *Fundamentals of Database Systems*, Reading, Adison Wesley, 2007.

Dans ce cas, il est impossible de vérifier la correction des valeurs de la base de données de manière automatique. Dès lors, lorsqu'une incohérence apparaît entre une valeur saisie au sein de la base et les tables de référence permettant d'en tester la validité, il peut s'avérer indispensable selon les enjeux de procéder à une vérification manuelle en contactant le citoyen ou l'entreprise concernée, par exemple (procédure appliquée dans le secteur administratif, quel que soit le pays).

On ne dispose dès lors d'aucun référentiel formel « absolu » en vue de tester la correction d'une vaste base de données empiriques. Prenons un exemple. On sait que la législation sociale est différente selon qu'elle s'applique aux ouvriers ou aux employés, les premiers et les seconds se distinguant selon la nature prépondérante de leurs activités manuelles ou intellectuelles. Dans la pratique, cette distinction n'est pas aisée à opérer, mais le flou n'a pas droit de cité dans une base de données : il faut trancher. Pour ce faire, il s'agira souvent de se rendre sur le terrain pour interpréter les « situations de fait » et d'examiner des pièces justificatives. Au fil des interprétations et de l'évolution de la jurisprudence, la signification en extension des notions d'employé et d'ouvrier évoluera dans le temps. On peut conclure de ce mécanisme que les données ne sont pas données, les données se construisent progressivement. La prise en compte de ces conclusions est a fortiori fondamentale lorsque plusieurs institutions sont concernées. Ainsi, comme nous l'avons évoqué dans l'introduction, en Belgique, les projets d'administration électronique ont donné lieu à la mise en place d'une déclaration multifonctionnelle dans le secteur social, de telle sorte que le citoyen ne doive plus déclarer en ligne qu'une seule fois l'information le concernant, cette information étant ensuite exploitée par les différents secteurs de la sécurité sociale. La question de l'interprétation des concepts prend donc elle aussi une dimension multifonctionnelle. Ainsi que nous l'avons vu dans l'introduction du chapitre, cette question peut soulever des arbitrages en fonction des besoins, notamment entre la fiabilité relative de l'information et sa rapidité de diffusion.

De ce qui précède, nous pouvons proposer une typologie générale des violations de contraintes d'intégrité ou anomalies :

- *Une erreur formelle* : 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 ou pas 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.

Ajoutons que les anomalies peuvent être détectées *ex ante*, dès 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.

1.3.4. Comment les données se construisent-elles progressivement ?

L'évolution de la jurisprudence, les transformations opérées au sein des bases de données et les catégories observables sur le terrain sont solidaires. Solidaires, mais asynchrones. Elles opèrent, suivant leur nature, au sein d'échelles de temps différentes. On distingue ainsi « le temps long » des normes juridiques, renouvelées d'un trimestre ou d'une année à l'autre, le « temps intermédiaire » de la gestion des bases de données et le « temps court » du réel observable, celui des citoyens ou des entreprises assujettis à l'administration, dont l'évolution est continue. Régulièrement, en effet, des entreprises fusionnent, se scindent, d'autres disparaissent alors que de nouvelles professions ou catégories d'activité non prises en compte dans les nomenclatures officielles voient progressivement le jour, avec la diversification des métiers de l'informatique par exemple. D'un point de vue dynamique, une base de données idéale devrait donc calquer le rythme de ses mises à jour sur la répartition - imprévisible - en « temporalités étagées » des évolutions de la réalité qu'elle appréhende. À ce qui ressemble à une gageure s'ajoute la nécessité, toujours révélée *a posteriori*, d'intégrer des observations imprévues, interdites *a priori* par l'hypothèse du monde clos.

Prenons un exemple dans le domaine des mesures en faveur de l'emploi. Suite aux directives émises par le Conseil européen de Bruxelles de décembre 1993, sur la base du *Livre blanc* de Jacques Delors pour la croissance, la compétitivité et l'emploi, des mesures d'aide à l'embauche se sont multipliées dans la plupart des pays européens en vue de lutter contre le chômage. Parmi celles-ci, les mesures de réduction de charges sociales se traduisent par une diminution des cotisations dues par les employeurs à la sécurité sociale en vue de leur permettre, en contrepartie, la réalisation d'embauches supplémentaires. Dans de nombreux pays de l'Union européenne, ces mesures se sont traduites par un foisonnement de directives législatives et d'ajustements qui rendent complexes non seulement leur mise en œuvre mais aussi l'évaluation de leur efficacité.

Ainsi, en Belgique, lors de la mise en place d'une directive administrative en faveur du secteur « non marchand », la question s'est-elle posée, 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. 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 aux nouvelles observations. En l'absence d'une telle intervention, l'écart entre la base et le réel se creuse. Nous verrons au point suivant les prolongements opérationnels de ces mécanismes lorsqu'il s'agit d'évaluer et d'améliorer la qualité des données administratives.

Quelles sont les conséquences de cette analyse en vue de spécifier des indicateurs de qualité adaptés à l'information administrative ? Les données administratives étant empiriques, on ne dispose pas de référentiel direct en vue d'en tester la correction. On ne peut appréhender leur adéquation aux besoins qu'indirectement, via une série d'indicateurs latéraux.

À cette fin, nous abordons maintenant les modalités de prise en compte des anomalies dans le cycle de vie d'une base de données. Celles-ci reposent sur une modélisation de l'historique des anomalies au cœur du schéma conceptuel, avec une prise en considération de la modélisation des séquences de contrôles et une

documentation de l'ensemble du schéma. Sur cette base, il sera possible de concevoir des indicateurs de qualité ; nous présenterons un ensemble de recommandations méthodologiques à cette fin ainsi que plusieurs stratégies opérationnelles permettant, sur cette base, d'améliorer la qualité d'une base de données de manière continue.

2. Anomalies et cycle de vie d'une base de données

Nous avons vu que les contraintes d'intégrité formelles, spécifiant les valeurs admises au sein d'une base de données, ne permettent pas nécessairement de décider si les valeurs testées sont vraies ou fausses « dans le monde réel » car ces dernières peuvent être sujettes à interprétation humaine.

Il arrive en effet fréquemment qu'une intervention humaine soit indispensable (cf. rôle des agents des services du contrôle des administrations fédérales) en vue de corriger ou valider l'information, suite à une violation de contrainte d'intégrité ou en l'absence de celle-ci.

Nous présentons ici (2.1) un modèle destiné à assurer un suivi semi-automatique du processus de détection et de traitement des anomalies et, sur cette base, de développer ultérieurement des stratégies en vue d'en diminuer le nombre et d'en rationaliser la gestion (2.2). Un dossier d'analyse de ce modèle existe et est disponible sur demande. Il est à adapter selon les projets et les enjeux y afférents. Nous envisageons ensuite la modélisation de la séquence des contrôles (2.3) et la documentation opérationnelle du système d'information (2.4).

2.1. Modélisation conceptuelle de l'historique des anomalies

En vue de poser le contexte dans lequel se situent le prototype et le modèle présenté, prenons un 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. Lorsque la gestion d'une base de données revêt des enjeux stratégiques sur les plans sociaux, juridiques et financiers, les enregistrements a priori rejetés en vertu de l'hypothèse du monde clos (anomalies formelles) doivent être sauvegardés de façon à pouvoir ultérieurement en assurer la gestion (correction de l'information ou validation de celle-ci si la valeur initiale est tout de même correcte). Nous passons donc de « l'hypothèse du monde clos » à celle d'un « monde ouvert sous contrôle ».

Même si elle demande des ressources importantes, cette intervention humaine est indispensable au niveau de l'administration fédérale dans de nombreux pays, car les anomalies à interpréter peuvent revêtir des enjeux stratégiques. Le phénomène concerne toutes les bases de données empiriques et n'est pas spécifique au

domaine de l'administration.

Dès lors, en vue de rationaliser et d'améliorer la gestion des anomalies, il sera utile de mettre en place des stratégies de gestion en vue d'en diminuer le nombre de manière à :

- obtenir au sein des organisations des gains en termes de coûts-bénéfices et de ressources humaines en charge de la gestion de ces anomalies ;
- améliorer la qualité des données et, dans notre domaine, à assurer un traitement plus précis, juste et rapide des droits sociaux des citoyens.

Afin de répondre à ces besoins, nous avons proposé dans des études antérieures un modèle en vue de sauvegarder non seulement les anomalies mais également l'historique de leur traitement[51] (ces propositions ont été validées sur la base d'expériences réelles appliquées à la base de données LATG et à la DmfA). Nous en précisons le mécanisme concrètement.

Les anomalies sont souvent traitées par des opérateurs humains au sein des services du contrôle affectés à chaque administration. Le modèle proposé ci-dessous repose sur un mécanisme d'enregistrement de la manière dont les opérateurs traitent les anomalies (correction, validation) : si la correction ou la validation résulte d'interprétations humaines, ces opérations, une fois réalisées, peuvent être enregistrées automatiquement au sein d'un système de suivi du traitement des anomalies. En cas de validation, les opérateurs en charge du traitement et de l'interprétation des anomalies au sein des services peuvent en effet « forcer » le système à accepter une violation de contrainte d'intégrité, s'ils jugent que celle-ci est tout de même valide. 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. Sur la base du modèle que nous allons présenter, un algorithme pourra alors émettre 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 en effet intéressant d'approfondir le phénomène : un cas de figure inédit (l'émergence d'une nouvelle catégorie d'activité ou l'évolution de l'interprétation d'un concept, le « secteur non marchand », par exemple) est peut-être apparu, ce qui requiert une adaptation de la structure de la base. Si l'on n'adapte pas le schéma, les anomalies correspondant à ces cas vont continuer d'apparaître en masse, nécessitant un examen manuel potentiellement conséquent et ralentissant considérablement le traitement des dossiers administratifs. Pour la sécurité sociale belge, la mise en œuvre de cette méthode concernant les déductions de cotisations a permis dans le passé 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 (cf. *infra* 2.2).

Le but du modèle présenté ici est donc de contribuer à automatiser la détection des violations de contraintes d'intégrité et d'assurer un suivi automatique du processus humain d'interprétation et de traitement des anomalies. Sur cette base, on pourra envisager un monitoring structurel des anomalies et de leur traitement de façon à mettre en oeuvre des stratégies de gestion de la base de données (data tracking, suivi des anomalies les plus fréquentes, mise en place de mesures pour le traitement des zones qui ne seraient pas corrigées, adaptation ponctuelle du schéma de la base de données, lorsque le nombre de validations formelles atteint un seuil critique…).

[51] BOYDENS I., *Informatique, normes et temps*. Bruxelles : Bruylant, 1999. BOYDENS I., « Strategic Issues Relating to Data Quality for E-government: Learning from an Approach Adopted in Belgium ». In ASSAR S., BOUGHAZALA I. et BOYDENS I., éds., *Practical Studies in E-Government : Best Practices from Around the World*, Springer, 2011, p. 113-130.

Dans le présent rapport, nous présentons un modèle générique d'historique des anomalies et de leur traitement, sous la forme d'un prototype. Plusieurs exemples de requêtes SQL représentatives sont ensuite évoquées en vue de montrer des exemples de monitoring typiques qu'il est utile de mener dans le temps en vue d'évaluer et d'améliorer la qualité d'une base de données. Ce modèle et les requêtes qu'il permet de générer sont destinés à s'appliquer à toute base de données pour autant que :

- les enjeux qu'elle soulève requièrent un suivi historique des anomalies et de leur traitement ;
- la gestion de la base, des anomalies et des indicateurs correspondants s'inscrive dans le cadre d'une organisation solide (ce point sera développé dans le présent rapport).

Pour que cette démarche porte ses fruits, il est nécessaire que la structure d'historique des anomalies et sa gestion soient reliées conceptuellement avec :

- les outils connexes : data quality tools… ;
- les applications concernées : bases de données, dictionnaires de données, bases de connaissances, datawarehouse, BI…

Dans tous les cas (conception d'une nouvelle base de données, reengineering d'une base existante ou affectation d'un historique des anomalies à une base de données en cours de gestion), les coûts en termes de gestion et d'organisation devront être envisagés, notamment en ce qui concerne l'impact sur l'existant. La situation idéale, dans laquelle on ne se retrouve pas nécessairement, est celle où l'on se situe au seuil de la structuration et de l'implémentation d'une nouvelle base de données. Dans tous les cas, la structure de gestion des anomalies doit avoir une connectivité minimum avec la base de données en production. D'une part, pour minimiser les exigences vis-à-vis de la structure applicative et, d'autre part, pour réduire l'impact sur l'existant. Les points suivants sont successivement présentés : environnement de départ et prérequis (2.1.1), positionnement de la démarche (2.1.2), modélisations conceptuelle et logique en couches (2.1.3), déroulement de l'implémentation (2.1.4) et exemple (2.1.5).

2.1.1. Environnement de départ et prérequis

Une gestion d'historique des anomalies n'a de sens que si elle s'inscrit dans le cadre d'une base de données applicative gérant préalablement elle-même son historique. En effet, à quoi bon gérer l'historique des anomalies si l'on n'a pas l'historique de l'information de base ?

Pour pouvoir gérer au mieux ces historiques et pouvoir tisser les liens entre les différentes versions de l'information de base et les versions des anomalies et des corrections, nous avons besoin d'identifiants uniques.

La conservation de l'historique des valeurs successives d'une information métier dans la même table que l'information active amène à devoir repenser la structure de l'identifiant unique. En effet, la clef métier de l'information active est très facile à isoler, il suffit de retranscrire l'identification naturelle utilisée au quotidien par ses utilisateurs. Lorsque l'on veut suivre les variations de cette information sur une période de temps, il nous faut affiner cette clef pour en différencier les étapes.

Pour ce faire, au moins deux options s'offrent à nous : la gestion purement technique, qui va réattribuer à chaque occurrence une nouvelle clef artificielle unique ou l'enrichissement de la clef fonctionnelle par un élément temporel de chronologie. Chacune de ces deux options pouvant se décliner en plusieurs solutions techniques.

La solution que nous recommandons et qui est retenue dans l'exposé sauvegarde la dimension métier à laquelle un élément chronologique est associé. Il s'agit d'apposer une estampille temporelle à la clef fonctionnelle naturelle de l'information (*Primary Key business* et *Timestamp*). Toutes les informations de la base de données applicative pour lesquelles une gestion des anomalies doit être activée se verront identifier par une concaténation de leur clef fonctionnelle et des date et heure auxquelles elles ont été émises (CREATED_TMS : Timestamp).

Afin de gérer au mieux la fin de validité d'une information métier, une date de fin (END_TMS) ou au minimum un indicateur (ACTIVE_IND) est également recommandé. Ces champs de gestion temporelle sont en principe déjà compris dans l'architecture des tables concernées puisque l'on est censé en gérer l'historique. Cependant, dans certaines implémentations, on retrouve à la place de cette gestion temporelle, une gestion par version. Celle-ci peut bien sûr également convenir pour identifier de manière unique l'enregistrement concerné par une anomalie ou une correction. L'important est avant tout que l'unicité des enregistrements soit garantie et représentée de la même manière dans toutes les tables concernées par la gestion des anomalies.

Dans tous les cas, la version reprenant les estampilles temporelles de début et de fin de période de validité reste la solution la plus confortable. Notons qu'une base de données existante faisant l'objet d'une mise en historique ne pourra pas valoriser ces estampilles pour les enregistrements du passé si aucun historique n'a été prévu lors de la conception initiale.

2.1.2. Positionnement de la démarche

La démarche se veut la plus générique possible, pouvant s'adapter aussi bien à une démarche de conception de base de données qu'à une base de données existante, gérant déjà de l'historique. Dans tous les cas, la structure de gestion des anomalies doit avoir une connectivité minimum avec la base de données en production. D'une part, pour minimiser les exigences vis-à-vis de la structure applicative et, d'autre part, pour réduire l'impact sur les performances et sur l'existant.

Dans cette optique, la seule incidence réellement liée à la mise en œuvre de la gestion d'anomalies sera limitée à l'ajout, dans chaque table visée par la gestion des anomalies, d'un champ numérique pouvant accueillir le numéro d'enregistrement en anomalie (RECORD_ANOMALY_ID).

Par ailleurs, une table de référence reprenant le catalogue des tables du schéma applicatif devra prendre place dans le schéma de gestion des anomalies, afin d'optimiser le lien entre les anomalies/corrections et les enregistrements concernés.

Outre un couplage minimum avec la base de données en production, la démarche doit permettre une gestion manuelle des anomalies aussi bien qu'automatique. Dans le cas où la règle métier associée est formellement identifiable, toute anomalie détectée une première fois de manière manuelle pourra être automatisée sur cette base (ce sera naturellement impossible si la détection relève d'une interprétation humaine non formalisable). Dans le cas de la démarche de traitement (correction ou validation), une aide semi-automatique pourra être fournie (via les data quality tools, par exemple, comme nous le verrons plus loin, mais la plupart du temps, une intervention humaine restera indispensable).

2.1.3. Modélisations conceptuelle et logique en couches

Pour soutenir cette démarche de découplage, la structure de gestion des anomalies est construite en différentes couches (Figure 1) : pratiquement, l'ampleur et la nature de ces couches devront être définies en fonction du projet concerné et des moyens disponibles. Le mécanisme de gestion de l'historique de détection des anomalies et de suivi de leur traitement présenté ici est quant à lui généralisable à tout projet, sur le plan de la logique de fonctionnement.

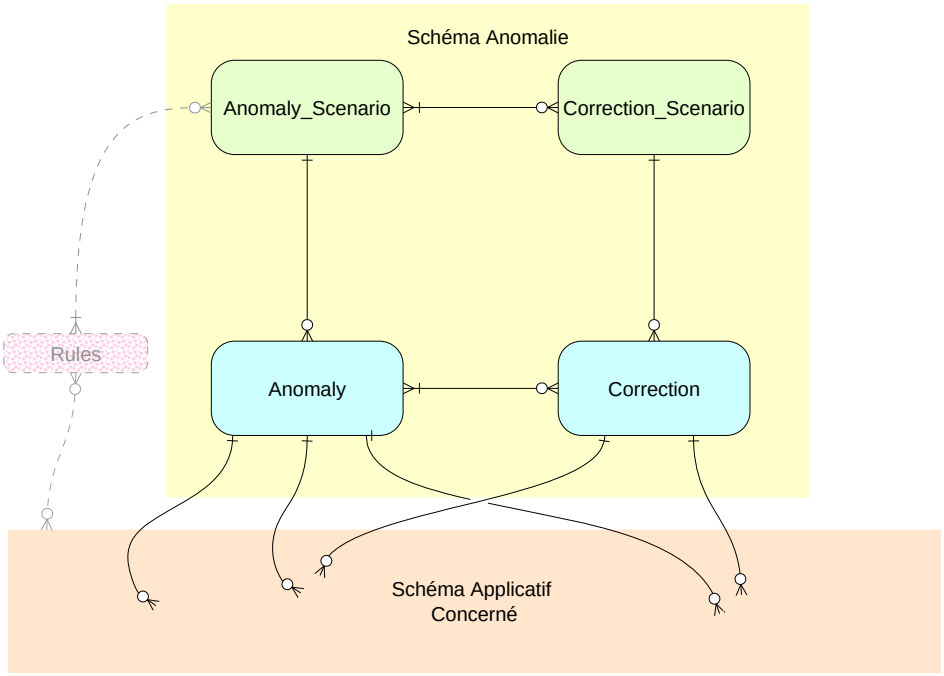


Figure 1 : Schéma conceptuel

Le schéma logique (Figure 2) comporte trois strates (représentées à titre d'exemples : tous les champs qu'elles peuvent potentiellement inclure ne sont pas repris dans chaque table sur la figure).

La première strate (en jaune dans la figure 2) reprend les données de référence techniques et métier. Il s'agit, d'une part, de la description de la structure du schéma applicatif cible (base de données) en termes de tables et de champs et, d'autre part, des types et des codes avec lesquels les informations relatives aux anomalies seront décrites. Le catalogue technique ainsi créé donne au schéma « Anomalie » une carte de la base de données qui permet de localiser le problème.

La deuxième strate (en vert dans la figure 2) s'appuie sur la première en vue d'en définir le référentiel fonctionnel lié à la base de données applicative. Il sera sage dans certains cas, de travailler à l'économie, d'alléger au maximum l'ampleur des descriptions reprises dans le schéma de la base de données structurée et de gérer dans une base de données documentaire dédiée à cette fin les informations non structurées permettant de répondre aux questions suivantes. Dans les faits, de quoi s'agit-il ? La gestion des anomalies a comme fondement deux descriptifs : les scénarios d'anomalies et les scénarios de corrections. Avant d'identifier une anomalie dans la base de données applicative, il faut savoir ce que l'on cherche. Il en faut une description précise : quelles tables sont concernées, et dans ces tables, quels champs participent à la détection des anomalies ? Ces descriptions seront catégorisées afin notamment de :

- ranger les anomalies selon les caractéristiques typées dans la première strate ; plus cette caractérisation est vaste et le typage fourni, plus on pourra sortir de statistiques en support au monitoring des anomalies.
- fournir un lexique des anomalies avec une description permettant d'en interpréter la signification (cette question est approfondie plus loin, au point 2.4).

Si tout est déjà défini dans le scénario et puisque la base de données applicative enregistre les modifications, seul un jeu de pointeurs est nécessaire pour repositionner l'information à un temps 't' et expliquer les raisons de la détection de l'anomalie.

En ce qui concerne l'autre volet de la gestion des anomalies, les traitements (corrections ou validations), on procède de la même manière, en ajoutant par ailleurs une liaison « anomalie-correction » permettant de retracer le cycle de vie des informations concernées et d'en déduire les mécanismes.

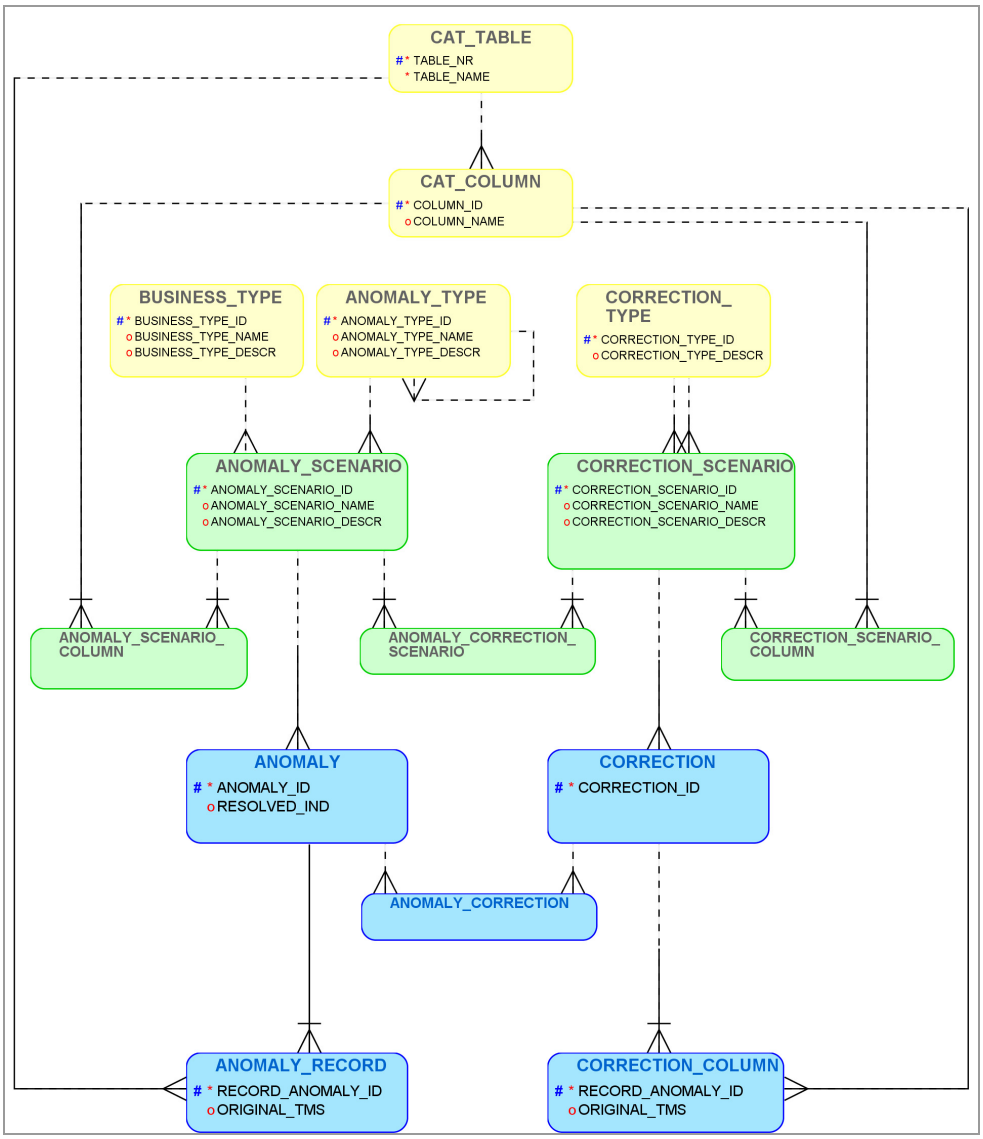


Figure 2 : Schéma logique

Dans certaines circonstances et pour autant que les impacts en termes de ressources et d'organisation soient identifiés et maîtrisés, la deuxième strate peut se construire au fur et à mesure. Aussi bien dans une nouvelle base de données

que dans une base de données en maintenance, la mise en route de la gestion des anomalies peut se faire pas à pas, scénario par scénario (avec probablement une phase de détection manuelle pendant quelque temps avant l'écriture d'une procédure de détection automatique, lorsque celle-ci est formellement possible, ainsi que nous l'avons noté plus haut).

La troisième strate (en bleu dans la Figure 2) est constituée par la machinerie de pointeurs permettant de relier l'anomalie détectée dans la base de données applicative avec son scénario dans la deuxième strate. Chaque table de la strate scénario ayant son pointeur correspondant dans la troisième strate, les corrections auront le même type de gestion que les anomalies et la liaison entre les deux sera aussi assurée.

Rappelons que cet inventaire des anomalies et des corrections identifiées dans la base de données applicative ne pourra se faire que par le biais de l'identifiant unique repris dans le champs ajouté à chaque table RECORD_ANOMALY_ID.

2.1.4. Déroulement de l'implémentation

Une fois les prérequis rencontrés et le champ RECORD_ANOMALY_ID ajouté à toutes les tables concernées par la gestion des anomalies, quelles sont les étapes suivantes ?

La première étape est l'alimentation des tables du catalogue du schéma « Anomalie » : cette étape pourra se réaliser de manière semi-automatique, elle sera plus ou moins lourde selon l'ampleur du schéma (voir plus haut). Il convient d'y référer toutes les tables du schéma applicatif faisant l'objet d'une détection d'anomalies ainsi que les colonnes concernées. Pour ce faire, le plus simple est de se référer au catalogue de la base de données et de l'enrichir d'une clef technique pour chaque niveau de la structure table-colonne.

La deuxième étape est la personnalisation des types qui seront utilisés pour décrire les scénarios d'anomalies et de corrections. Comme nous l'avons déjà évoqué auparavant, la granularité relative de ces typages relève d'un arbitrage : plus elle est riche, plus elle implique potentiellement des ressources importantes en termes de gestion, mais plus elle permet la génération de statistiques de suivi des anomalies approfondies. Cet arbitrage devra être établi au cas par cas, en fonction des projets et de leurs enjeux.

La troisième étape référence la partie métier de l'application cible. Il s'agit de décrire les scénarios d'anomalies et de corrections, et les liens qui les relient. Ces références sont à la fois descriptives et techniques. D'un côté, on explique via un texte les situations ou les actions à réaliser (comme nous l'avons mentionné plus haut, ces explications non structurées peuvent figurer dans un système d'information documentaire dédié à cette fin - cf. *infra* 2.4). De l'autre, on spécifie via le catalogue créé lors de la première étape les champs de la base de données applicative qui procèdent de l'anomalie ou, dans le cas des corrections, qui doivent être modifiés. Enfin, on lie les scénarios de corrections aux scénarios d'anomalies.

Jusque-là, on s'est contenté d'enrichir le dictionnaire technique métier et fonctionnel, rien concernant une réelle anomalie dans la base de données applicative n'a encore été référencé.

Pour chaque type d'anomalie identifié, on introduit le scénario dans la deuxième strate en le décrivant en texte libre (ou dans un système documentaire associé, éventuellement), au moyen des types et catégories de la première strate. Une fois le scénario complètement introduit, celui-ci sera référencé dans la troisième strate en même temps que l'on valorisera les pointeurs vers les enregistrements et les colonnes de la base de données applicative, cela lors de l'introduction manuelle (par un utilisateur) ou automatique (par un batch) d'une anomalie. À

cette fin, les enregistrements concernés seront associés à un RECORD_ANOMALY_ID par le moteur de la base de données.

Dans certains cas, le scénario de correction devra mûrir quelque temps avant de trouver sa place dans la gestion des anomalies et peut-être encore plus avant d'être éventuellement automatisé, comme le scénario de détection. Il faudra donc dans ce cas prévoir des écrans d'introduction manuelle des détections d'anomalies, de corrections et de liaison des unes et des autres.

Le but final étant qu'une fois identifiée, détectée automatiquement et traitée (corrigée ou validée suite à une démarche d'interprétation administrative) par un agent humain ou par un batch, l'anomalie fasse l'objet d'un historique. Celui-ci enregistre les différentes actions automatiques ou manuelles associées à ces différents processus. De la sorte, il sera possible ultérieurement d'assurer le suivi dans le temps de ces processus et de proposer des stratégies de gestion appropriées (2.2).

2.1.5. Exemple

Diagramme de la base de données applicative

La Figure 3 illustre un exemple du mécanisme de suivi des anomalies à partir d'un cas simple : une gestion d'individus salariés, avec une signalétique PERSON et ses tables de références ainsi que les paiements mensuels des salaires.

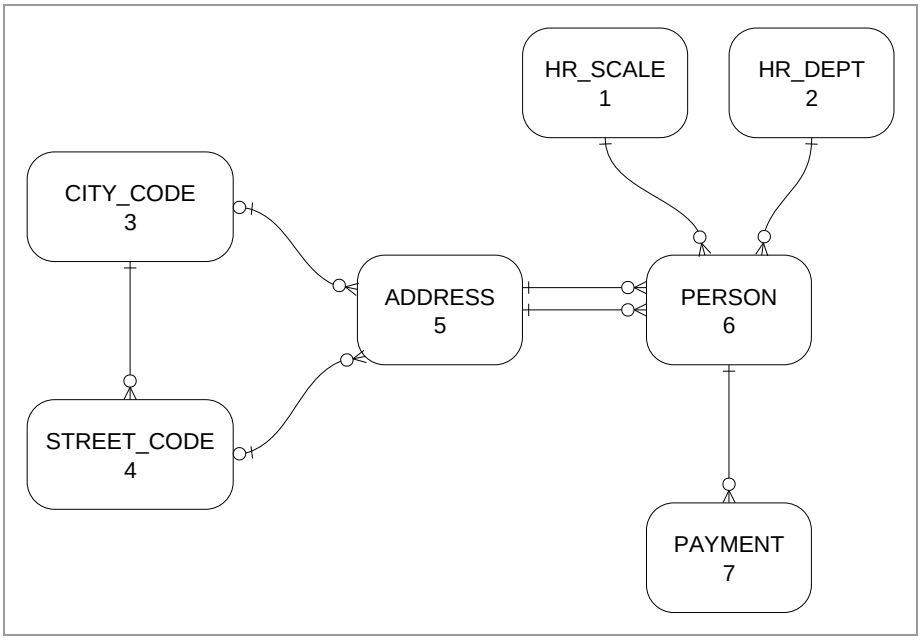


Figure 3 : Diagramme de la base de données applicative

A) État de départ dans la base de données applicative

Nous observons ci-dessous un enregistrement dans la table PERSON, identifiée dans notre catalogue comme la table numéro 6. Comme il s'agit d'une base de données historisée, un timestamp (CREATED_TMS) complète la clef métier, ici dénominée PK. La valeur de cette estampille temporelle : « *Tms 1* » correspond à la date et à l'heure auxquelles l'information contenue dans cet enregistrement a été mise à disposition dans la base de données. Aucune partie de cette information, n'a été modifiée depuis. Pour respecter les prescriptions décrites

dans ce document, lors de la phase de restructuration de la base de données applicative, un champ RECORD_ANOMALY_ID a été ajouté à la structure originelle. Comme aucune anomalie n'a encore été détectée sur cet enregistrement dans notre exemple, la valeur initiale est « null ».

PERSON (6)

PK	CREATED_ TMS	END_ TMS	yyy		...	xxx	...	RECORD_ ANOMALY_ID
536 814	*Tms 1*	NULL	yyy		...	ααα	...	

B) État de départ dans la base de gestion des anomalies

On dispose de trois compteurs pour la gestion des identifiants :

- le nombre d'anomalies détectées ;
- le nombre d'enregistrements participant aux anomalies détectées ;
- le nombre de corrections effectuées.

```
ANOMALY_ID           counter:  1433
RECORD_ANOMALY_ID    counter:  2563
CORRECTION_ID        counter:  870
```

Rappelons qu'une anomalie peut concerner plusieurs enregistrements dans plusieurs tables. Il en va de même pour les corrections.

C) Survient une détection, qui lève une anomalie

Nouvelle Anomalie ⇨ incrémentation des compteurs

ANOMALY_ID = ANOMALY_ID + 1

Pour chaque enregistrement participant au scénario anomalie

↳ RECORD_ANOMALY_ID = RECORD_ANOMALY_ID + 1

```
ANOMALY_ID           counter:  1434
RECORD_ANOMALY_ID    counter:  2564
CORRECTION_ID        counter:  870
```

PERSON (6)

PK	CREATED_ TMS	END_ TMS	yyy		...	xxx	...	RECORD_ ANOMALY_ID
536 814	*Tms 1*	NULL	yyy		...	αXα	...	2564

Au niveau de la base de données applicative, un seul changement apparaît : l'instanciation du champ RECORD_ANOMALY_ID, si celui-ci ne l'était pas encore, dans le cas où aucune modification n'aurait affecté l'enregistrement. Une fois marqué par sa première anomalie, l'information métier reçoit un identifiant « RECORD_ANOMALY » pour le reste de son existence. Chaque instanciation de l'information au travers de son historique véhiculera cet identifiant. Remarquons que cet identifiant ne représente pas une anomalie en soi, mais un mécanisme afin

de relier l'information métier à chaque anomalie détectée via un système d'identification univoque. Afin d'identifier de manière unique l'enregistrement dans le cycle de vie de l'information il faut encore lui adjoindre l'estampille temporelle CREATED_TMS.

ANOMALY

ANOMALY_ID	CREATED_ TMS	ANOMALY_ SCENARIO_ID	RESOLVED_IND
1434	Tms 2	A101-0610	N

Dans le schéma de gestion des anomalies, une fois une anomalie détectée, le compteur, automatiquement incrémenté, donne le numéro d'anomalie. L'estampille temporelle (CREATED_TMS) permettra ultérieurement de lancer des statistiques historiques quant à la gestion des anomalies. Le type de scénario que l'on va associer à l'anomalie sera choisi dans une liste préalablement définie. L'anomalie n'ayant pas encore été traitée, son statut « RESOLVED_IND » prend la valeur « N ».

Au niveau inférieur, tous les enregistrements de la base de données applicative touchés par cette anomalie seront automatiquement référencés. Pour certains scénarios, le nombre d'enregistrements pourrait ne pas être fixé a priori, mais dépendant de l'information participant à la règle métier violée. Ainsi dans le cadre de notre exemple, si la règle à vérifier est la suivante : « la somme des versements payés à un salarié durant une année doit être inférieure à 14 fois le salaire mensuel net », 14 enregistrements pour un employé pourraient être concernés, 26 pour un ouvrier, sans compter les régularisations et les primes. À cela s'ajoute l'impact sur l'enregistrement signalétique même, puisqu'il spécifie le salaire net qui entre dans la formule de calcul.

ANOMALY_RECORD

ANOMALY_ID	RECORD_ ANOMALY_ID	TABLE_NR	ORIGINAL_TMS
1434	2564	6	Tms 1

L'ANOMALY_ID est la clef étrangère (FK) qui assure un lien automatique avec l'anomalie et donc regroupe tous les enregistrements participant à la même anomalie. Le RECORD_ANOMALY_ID établit le lien avec la base de données applicative et l'information concernée. Dans notre exemple, le numéro 6 repris dans le champ TABLE_NR représente la table PERSON dans laquelle on retrouvera l'enregistrement numéro 2564. Cette information n'est pas strictement indispensable puisque le RECORD_ANOMALY_ID est unique à travers toute la base de données. Cependant, pour des raisons évidentes de performance et d'ergonomie, cette information nous a paru indispensable. Enfin l'estampille temporelle (CREATED_TMS) de l'enregistrement concerné est générée automatiquement et permet d'identifier l'instance de l'information potentiellement problématique. Comme dans le cas de la clef métier (PK), il faut également affiner temporellement l'identifiant RECORD_ANOMALY pour identifier de manière unique l'instance dans le cycle de vie de l'enregistrement concerné par l'anomalie.

D) Vient le temps des traitements (corrections ou validations)

Première étape d'une correction ⇨ l'incrémentation du compteur

↳ CORRECTION _ID = CORRECTION _ID + 1

ANOMALY_ID	counter:	1434
RECORD_ANOMALY_ID	counter:	2564
CORRECTION_ID	counter:	871

Deuxième étape : une fois les anomalies automatiquement détectées suite à une violation des règles métier ou à une interprétation humaine, les anomalies vont être progressivement interprétées et traitées (corrigées ou validées) par les agents. Puisque nous disposons d'une gestion historisée de l'information, chaque modification génère une nouvelle instance unique de cette information avec les nouvelles valeurs.

PERSON (6)

PK	CREATED_ TMS	END_ TMS	...	xxx (8)	...	RECORD_ ANOMALY_ID
536 814	*Tms 1*	*Tms 3*	...	ααα	...	2564
536 814	*Tms 3*	*NULL*		ωωω		2564

Nous avons donc ici la vue des enregistrements associés à l'information 536 814. L'enregistrement sur lequel la détection a été effectuée (CREATED_TMS = Tms1) et dont la date de fin de validité a été complétée avec l'estampille temporelle (END_TMS = Tms 3) permet de conclure qu'une nouvelle version de l'information métier est disponible.

Le nouvel enregistrement qui véhicule la nouvelle valeur « ωωω » corrigée inclut également l'estampille temporelle « Tms 3 » mais qui ici spécifie sa date de début de validité CREATED_TMS. On remarque que les deux enregistrements portent le même RECORD_ANOMALY_ID, preuve de leur instanciation de la même information métier.

Troisième étape : la gestion de la correction dans la base de données « Anomalie ».

CORRECTION

CORRECTION_ ID	CREATED_ TMS	CORRECTION_ SCENARIO_ID	ANOMALY_ID
871	*Tms 3*	C101-0610	1434

Le compteur des corrections ayant été incrémenté, on peut créer un enregistrement dans CORRECTION. Afin de pouvoir retrouver la nouvelle information via sa clef technique unique, son estampille temporelle (CREATED_TMS) doit donc être identique à celle de l'instance corrigée, soit : Tms 3.

Dans le cas où plusieurs enregistrements devraient être corrigés en même temps, leurs estampilles temporelles seraient identiques. En effet, il doit y avoir unicité transactionnelle à travers chaque étape de la démarche, que ce soit lors de la détection ou de la correction. La prise d'estampille s'effectue une seule fois au début de l'étape et est utilisée pour marquer toutes les zones date-heure de

création, de début de validité ou de fin de validité encore vierges de tous les enregistrements dépendant d'une anomalie ou d'une correction.

Pour compléter l'information présentée à ce stade, deux liens sont encore nécessaires : le type de scénario de correction utilisé (CORRECTION_SCENARIO_ID) et l'anomalie à laquelle ce scénario a été appliquée (ANOMALY_ID).

CORRECTION_COLUMN

CORRECTION_ID	RECORD_ANOMALY_ID	COLUMN_ID	ORIGINAL_TMS
871	2564	8	*Tms 1*

Pour pouvoir recréer l'état des lieux au moment de la correction, les zones suivantes associées aux enregistrements qui ont été mises à jour seront nécessaires : la clef de l'enregistrement actif au moment de la correction (RECORD_ANOMALY_ID + ORIGINAL_TMS) et de l'identité du champ concerné (COLUMN_ID). Le CORRECTION_ID permet de regrouper toutes les modifications répondant à la correction d'une anomalie.

ANOMALY

ANOMALY_ID	CREATED_TMS	ANOMALY_SCENARIO_ID	RESOLVED_IND
1434	*Tms 2*	A101-0610	Y

Une fois l'anomalie corrigée (ou validée), son statut sera mis à jour via le champ RESOLVED_IND qui dans notre exemple prend la valeur « Y » (oui).

Clef de lecture du modèle

La démarche comporte deux étapes : la détection et la correction.

Ces deux étapes sont associées à une partie bien distincte du modèle : les tables anomalies, pour l'étape de détection, les tables corrections, pour l'étape de correction (hormis le changement de statut de l'anomalie en clôture du problème).

La seule interaction avec le modèle applicatif concerne la mise à jour automatisée de l'identifiant RECORD_ANOMALY_ID lors de la détection d'un premier problème associé à cette information.

La reconstruction de l'état de l'information est possible à chaque étape du cycle de vie grâce à l'historique en natif dans la base de données applicative et grâce aux pointeurs associés aux détections et corrections du modèle présenté.

Les deux niveaux présents dans les tables du modèle ont chacun une finalité différente. Du côté détection, ANOMALY sert de regroupement et de conteneur pour les métadonnées de détection, alors que ANOMALY_RECORD sert d'aiguillage vers les enregistrements dénotant une présomption d'erreur formelle. Du côté de la correction, la table CORRECTION sert de lien vers les nouvelles valeurs et de conteneur pour les métadonnées de correction alors que CORRECTION_COLUMN sert d'aiguillage vers les colonnes à modifier (ou, une fois la correction effectuée, vers les colonnes modifiées pendant la correction).

Variations possibles

D'une part comme cela a été spécifié auparavant, la gestion par numéro de version au lieu d'estampilles temporelles est également possible. Si la notion temporelle n'est pas disponible, la séquence de détection et de modification est toujours intacte et seule indispensable. Toutefois, pour conserver les fonctions de

gestion des délais, il faudra quand même prévoir les estampilles temporelles dans les enregistrements des tables anomalies et corrections : celles-ci sont indispensables en vue d'assurer un suivi ultérieur dans le temps de l'historique des anomalies détectées et de leur traitement et de concevoir, *in fine*, des stratégies de suivi et d'amélioration de la qualité des données.

De même, la date de fin de validité dans l'enregistrement applicatif n'est pas indispensable, mais cette information est potentiellement utile : elle permet un accès direct à l'information active et donc une performance accrue au niveau du schéma applicatif. Cet inconvénient peut bien sûr être levé par l'ajout d'un indicateur d'obsolescence.

Possibilités de statistiques étendues

Déjà abordée de multiples fois dans les paragraphes précédents, la richesse des rapports (indicateurs de qualité relatifs aux anomalies et à leur traitement) pouvant être produits sera fonction de la granularité du typage de la première strate. Néanmoins, la découpe du modèle permet déjà de nombreux recoupements. Voici deux exemples concrets auxquels sont associées les requêtes SQL correspondantes (dans le point 2.2, nous évoquons, en contexte, d'autres indicateurs utiles) :

- Suivi des anomalies sur l'ensemble des périodes de référence :
```
SELECT YEAR (created_tms) as year, MONTH (created_tms) as month , anomaly_scenario_id as anomaly_scenario ,count(*) FROM anomaly GROUP BY year, month, anomaly_scenario;
```
- Nombre d'anomalies détectées mais finalement valides :
```
SELECT COUNT(*) FROM correction COR, anomaly_correction ANCO WHERE COR.correction_id = ANCO.correction_id AND COR.correction_scenario_id = 10 AND COR.created_tms BETWEEN '2010-01-01' and '2010-08-31';
```

2.2. Monitoring des anomalies et stratégies de gestion

La mise en place d'un service d'évaluation et d'amélioration de la qualité des données est nécessairement un cycle continu qui doit être soutenu par le management. Idéalement, ce service doit être mis en place dès la conception d'une base de données. Par ailleurs, il convient d'assortir la démarche d'un suivi continu pris en charge par un comité pluridisciplinaire conçu à cette fin, les actions ponctuelles prises dans l'urgence étant à proscrire. L'organisation et les rôles à mettre en place sont dès lors stratégiques. Nous les abordons ci-dessous (2.2.4).

La démarche inclut les phases suivantes :

- Établissement des responsabilités et soutien de la part du management.
- Identification des projets d'amélioration de la qualité.
- Analyse des besoins, cartographie des procédures et des données et définition des objectifs. S'agissant de l'information administrative, nous avons caractérisé les principales questions touchant les bases de données ainsi que les objectifs et enjeux correspondants (voir chapitre 1).
- Mise en œuvre d'un système d'indicateurs de qualité.

2.2.1. Indicateurs de qualité

Recommandations méthodologiques

Avant d'aborder les méthodes d'amélioration de la qualité proprement dites, évoquons quelques pistes directrices en vue de la conception des indicateurs de qualité[52]. Comme le note P. Rivière, « *le terme "indicateur" n'est pas choisi au hasard car une véritable mesure de la qualité ne serait pas atteignable. C'est une remarque désormais classique : s'il était possible de mesurer parfaitement la fiabilité d'un fichier X en le comparant à un fichier de référence Y, alors le fichier Y se substituerait au fichier X dont la production perdrait tout intérêt* […] *Comment se représenter une idée de fiabilité de l'activité principale d'une entreprise ? Dans la mesure où il n'existe pas a priori de référence externe, une possibilité simple est de prendre en compte l'ancienneté de cette variable, c'est-à-dire le nombre d'années depuis lequel elle n'a pas été modifiée. Il est clair que plus l'activité principale est ancienne et moindre sera la qualité de cette variable. C'est une information partielle et indirecte mais importante.*»[53]

Pratiquement, en vue de la conception de ces indicateurs, il est indispensable d'adopter une démarche descendante, comportant les étapes suivantes :

- cibler les besoins sur la base des objectifs de façon à établir un petit nombre d'indicateurs fondamentaux (il faut éviter une multiplicité de chiffres peu significatifs) ;
- définir en premier lieu les concepts et sur cette base, examiner les modalités de calcul opérationnel ;
- définir plusieurs niveaux d'agrégation en fonction de l'organisation en charge de l'exploitation des indicateurs ;
- documenter les indicateurs (via des méta-informations) ;
- industrialiser la production (la qualité requiert un suivi continu) ;
- pour chaque indicateur, définir des stratégies d'amélioration si les résultats témoignent d'une qualité insuffisante.

Il existe plusieurs modes de calcul des indicateurs de qualité :

- Citons en premier lieu les enquêtes sur le terrain ou inspections sur la base d'échantillons. Ce mode d'observation est incontournable lorsque les éléments visés sont indétectables formellement (nous avons envisagé ce cas plus haut : travail au noir, « faux actifs »…). Dans les autres cas, il comporte certaines limites : l'opération n'est jamais qu'un « one shot » et doit être réitérée pour offrir une photographie d'un état ultérieur de la situation. Elle est chère si elle est récurrente, notamment en raison de l'intervention humaine qu'elle requiert inévitablement et de la nécessité de traiter les « non réponses ». Par ailleurs, elle peut soulever des problèmes de crédibilité vis-à-vis de la population « cible » susceptible d'être contactée plusieurs fois d'une enquête à l'autre en fonction du principe d'échantillonnage retenu (ceci renvoie à la question de la base de sondage).
- En second lieu, il existe des méthodes qui pourront s'appuyer de manière continue sur un traitement partiellement formel de la base de données et de fichiers connexes : analyse de la cohérence interne (par exemple, cohérence dans le temps, adéquation entre une masse salariale et un effectif…) ou comparaison avec une source concurrente. Ces méthodes pourront éventuellement reposer sur le recours à des outils

[52] RIVIERE P., « Indicateurs de qualité en matière de production de données : quelques éléments de réflexion », *Courrier des statistiques*, septembre 2005, n° 115, p. 35-40.
[53] *Idem*, p. 38.

spécifiquement dédiés à l'évaluation de la qualité des données (« data quality tools »).

Exemples d'indicateurs

Les indicateurs doivent tenir compte de l'utilisation des données et des différents rôles (personnes) amenés à prendre des décisions sur cette base. À ce titre, certains préfèreront des indicateurs généraux et globaux :

- Nombre d'anomalies par secteur d'activité, ventilées sur la base d'une typologie illustrant leur « gravité » relative dans le contexte des enjeux sociaux et financiers du domaine d'application.
- Nombre d'anomalies non corrigées ayant un impact opérationnel fort.
- Durée de vie moyenne des anomalies : la durée de vie moyenne d'une anomalie est le temps écoulé entre sa présence « virtuelle » (lors de la saisie du record correspondant), sa détection et sa résolution. Cet indicateur permet d'évaluer le temps moyen de réaction des agents correcteurs (internes ou externes à la sécurité sociale) et de le comparer aux enjeux et besoins opérationnels. Sur cette base, il est possible de déterminer d'éventuelles actions à entreprendre, par exemple en vue de raccourcir ce délai sur la base d'échéances. Cette décision dépend par ailleurs d'un arbitrage de type « coûts-bénéfices ».
- Coûts de correction (en termes de « jours-hommes » et d'intervention manuelle).
- …

Tandis que d'autres privilégieront des indicateurs détaillés :

- Nombre de types d'anomalies et comparaison avec des profils de même type.
- Évolution du nombre d'anomalies dans le temps : cet indicateur permet à un expéditeur des données ou à l'organisme public qui les reçoit de vérifier l'impact des stratégies mises en œuvre en vue de diminuer le nombre d'anomalies à la source.

2.2.2. Stratégies de gestion : case studies

Nous présentons ci-dessous deux stratégies distinctes reposant sur des indicateurs de qualité. La première permet un suivi continu de la structure de la base de données (et de sa pertinence par rapport au domaine d'application) sur la base d'un monitoring d'anomalies. La seconde donne le jour à des « data trackings » ponctuels permettant de détecter dans les processus l'origine des anomalies et d'y remédier à la source.

Suivi continu de la structure de la base de données, adaptations structurelles et gains en termes de coûts-bénéfices

Ainsi que nous l'avons évoqué, 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 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, 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 : comme nous l'avons vu, un cas de figure inédit est peut-être apparu (l'émergence d'une nouvelle catégorie d'activité ou l'évolution de l'interprétation d'un concept - cf. l'exemple du « secteur non marchand » évoqué plus haut), ce qui requiert une adaptation de la structure de la base. Si l'on n'adapte pas le schéma, les anomalies correspondant à ces cas vont 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[54].

Naturellement, les difficultés rencontrées seront d'autant moins importantes 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 dans ce chapitre 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 d'une anomalie et de la rapidité de correction permet d'identifier le moment le plus opportun d'exploitation de la base de données à d'autres fins.
- Le suivi des dates de mises 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.

« Data Tracking », suivi des anomalies et mise en place de solutions structurelles d'amélioration

Le « data tracking », technique mise au point par T. Redman[55], est une méthode destinée à évaluer quantitativement la validité formelle des valeurs introduites dans une base de données et à en améliorer le traitement. Une base de données est

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

[55] REDMAN T., *Data Quality for the Information Age,* Artech House, Boston, 1996.

un lac : au lieu de nettoyer le lac (comme le préconise le data cleansing), Redman propose d'en analyser les sources et les flux. Traditionnellement, chaque enregistrement d'une base de 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. 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.

Nous renvoyons au deliverable « *Data Quality : Best Practices* », pour plus d'informations sur la méthode.

L'Office National de 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. Le but de l'opération consiste à détecter, chez l'expéditeur, les éléments à l'origine de la production d'un grand nombre d'anomalies (traitement inadéquat de certaines sources de données, interprétation inadéquate d'une directive administrative, etc.). Sur cette base, un diagnostic ainsi que des actions correctrices peuvent être posés. Contrairement à la méthode mise en place par T. Redman :

- l'échantillon d'individus et de cas retenus n'est pas aléatoire puisqu'on dispose d'une connaissance *a priori* concernant les dossiers problématiques, via l'historique des anomalies et de leur traitement ;
- il s'agit d'un « tracking arrière » (ou « back tracking ») : on part de la déclaration finale 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.

L'opération permet :

- d'obtenir des résultats durables, puisque la cause structurelle des erreurs est identifiée (qu'il s'agisse d'erreurs de programmation ou de problèmes d'interprétation de la législation en matière de temps de travail) et peut être définitivement réglée ;
- d'établir un partenariat avec les citoyens 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.

2.2.3. Workflow de correction des anomalies

Dans le cadre d'une mission de consultance pour la sécurité sociale que nous avions menée sur la gestion des anomalies, il nous avait été demandé de réfléchir à la possibilité de mettre en place un workflow de traitement des anomalies, dans un contexte où les anomalies sont un enjeu stratégique tant pour les droits des travailleurs belges que pour la gestion des ressources humaines et financières au sein des institutions en charge de leurs corrections.

L'objet de l'étude était donc d'étudier la possibilité de mettre en place une séquence de corrections des anomalies et les critères à prendre en compte pour ce

faire, tout en soulevant les difficultés possibles, aussi bien au niveau conceptuel qu'au niveau pratique, et les arbritages à effectuer.

Une séquence de corrections serait un ordre optimal dans lequel les données devraient être corrigées. Pour établir cette séquence, plusieurs critères peuvent être pris en compte, tels que l'interdépendance entre données, l'impact des données sur les droits sociaux et les échéances temporelles pour lesquelles les données doivent être disponibles sans anomalies.

Chaque séquence, quels que soient les critères pris en compte, présente des imperfections. Le choix des critères et de la séquence dépend donc des objectifs poursuivis. Ces critères ainsi que les arbitrages à effectuer sont présentés de manière plus détaillée en annexe (6.1).

Le résultat de cette étude a montré qu'une séquence de corrections automatisée et formalisée était envisageable en théorie mais difficile à mettre en œuvre dans la pratique.

Trois constats conduisent à cette conclusion :

1. La résolution d'une anomalie implique couramment de corriger un ensemble de données et non une seule en raison des dépendances entre données, ce qui fausse l'idée d'un recours à une séquence de corrections automatisée et formalisée.
2. Du fait des anomalies fictives, une donnée peut être considérée comme erronée et donc être présente dans la séquence alors que l'erreur porte *in fine* sur une autre donnée considérée comme formellement correcte.
3. Enfin, le dernier constat est d'ordre humain. Les agents en charge du traitement des anomalies sont des humains qui n'apprécient guère de devoir suivre un ordre prédéfini.

En conclusion, nous ne recommandons pas cette stratégie de gestion. La mise sur pied d'une application en vue de documenter la correction des anomalies et apporter ainsi une aide aux personnes en charge de la correction des anomalies nous semble être une stratégie plus efficace et adoptée plus facilement par les agents (2.4.2).

2.2.4. Organisation

La Figure 4 présente les grandes articulations de l'organisation associée au suivi des anomalies dans le contexte de la gestion d'une base de données. La base de règles traduisant les évolutions législatives et leurs interprétations administratives, les fournisseurs de l'information (employeurs, entreprises, citoyens…) font parvenir des données (sur un mode régulier ou non) à l'administration. Ces données sont stockées dans une base de données, gérée, à l'instar des règles associées, par les « data managers ». Ces derniers gèrent aussi l'historique de traitement des anomalies sur la base duquel les données seront traitées au fil du temps par les « spécialistes métier » (services du contrôle au sein des administrations, par exemple) qui accèdent à la base de données en lecture et en écriture. Par ailleurs, les « data quality tools » (abordés plus loin dans l'étude) permettent, on line ou en batch, d'évaluer et d'améliorer la qualité des données. En batch, ces derniers constituent également une aide à l'analyse. Dans la présente étude, les modules « gestion de l'historique des anomalies » et « data quality tools » n'ont pas le même statut : alors que les premiers se présentent sous la forme d'un prototype répondant à une évolution de la théorie en matière de modélisation conceptuelle, les seconds correspondent à un logiciel spécialisé diffusé sur le marché. L'ensemble des modules gérés par les « data managers » fait l'objet d'une documentation. Celle-ci pourra être consultée par les gestionnaires, les utilisateurs et les fournisseurs des données concernées.

En particulier, le prototype présenté plus haut, dans la pratique, devra être adapté en fonction d'une analyse approfondie de la base de données à laquelle il s'appliquera. Il constitue surtout l'illustration de la faisabilité du passage de l'hypothèse du monde clos à celle d'un monde ouvert sous contrôle prenant en considération la nécessaire interprétation des anomalies associées à un domaine d'application empirique. Les utilisateurs « métier » de l'application consultent la base de données et la traitent (corrections ou validations) : ces opérations étant stockées au fur et à mesure dans l'historique des anomalies. Sur la base de ce dernier, des indicateurs peuvent être fournis (requêtes SQL) et exploités par le management en vue d'évaluer et d'améliorer le management de la base de données. Sur la base d'une documentation régulièrement mise à jour, ces indicateurs permettront par ailleurs d'effectuer un « data tracking » des données, lequel, comme nous l'avons vu, permet potentiellement de cibler l'origine des anomalies à la source, dans les procédures des fournisseurs de l'information, d'y remédier structurellement et d'améliorer encore, sur cette base, les procédures de gestion de la base de données.

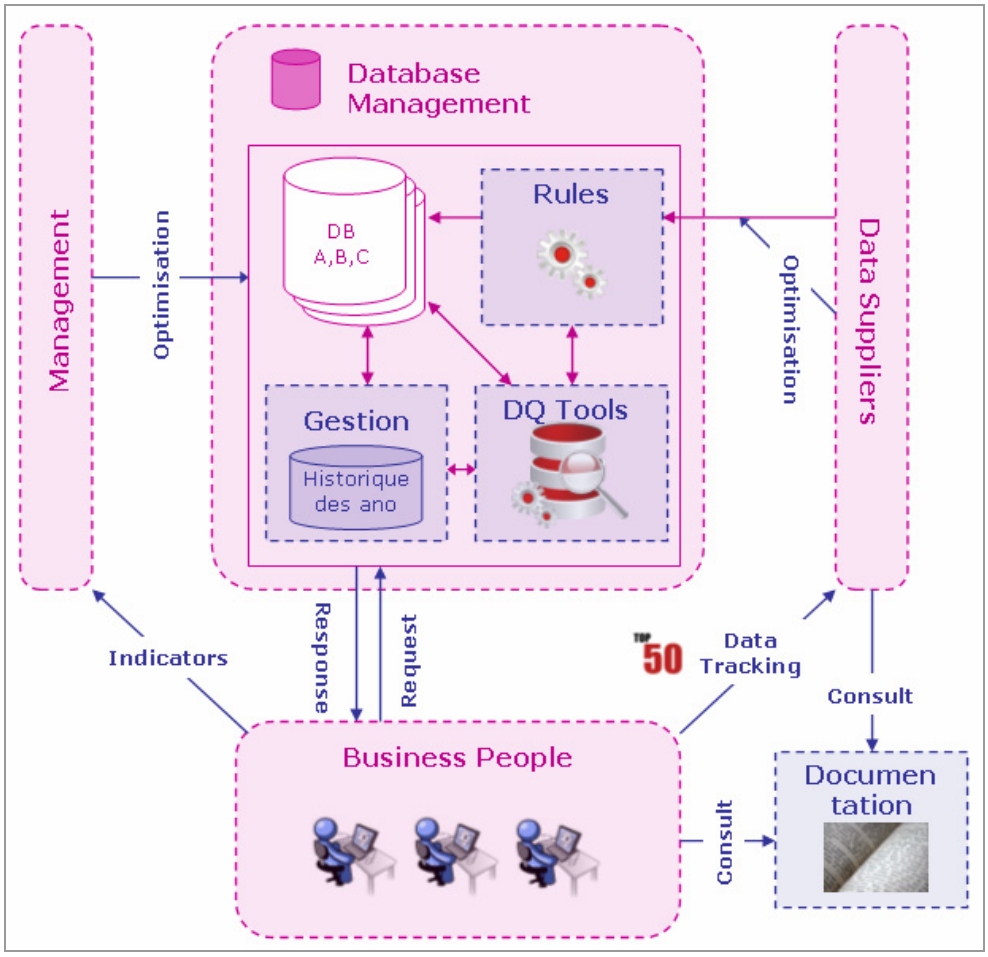


Figure 4 : Organisation en vue de gérer et d'exploiter le monitoring des anomalies et les stratégies de gestion

2.3. Modélisation de la séquence des contrôles

Lorsque des données sont introduites dans une base de données, il est nécessaire d'évaluer la qualité des données, en vérifiant qu'elles respectent les règles formelles définies préalablement, qu'il s'agisse de règles techniques, conceptuelles, métier ou issues de la législation, via l'application de contrôles automatiques.

Du point de vue conceptuel, les explications qui suivent sont également valables pour les bases de données utilisées dans les applications documentaires. Cependant, les données qui y sont stockées sont généralement moins stratégiques. De ce fait, le traitement des anomalies détectées est moins soumis à des contraintes légales et temporelles. Leur résolution ne requiert donc pas le même degré d'urgence et de rapidité.

La définition et la modélisation des contrôles sont cruciales car ils influent directement sur les anomalies qui sont générées et par conséquent sur la charge de travail nécessaire pour les gérer. La qualité des données étant un concept relatif, des choix doivent être effectués au moment de la modélisation sur la base d'arbitrages, eu égard aux besoins.

Dans cette partie, nous présentons tout d'abord une typologie des contrôles (2.3.1) et exposons la différence entre anomalies *ex ante* et *ex post* (2.3.2). Nous analysons ensuite la question de l'ordre des contrôles et des critères qui peuvent être pris en compte, des conflits potentiels et des nécessaires arbitrages à effectuer (2.3.3). Pour finir, nous formulerons plusieurs recommandations pour la modélisation de la séquence de contrôles et leur mise en œuvre (2.3.4).

2.3.1. Typologie des contrôles

Plusieurs types de contrôles peuvent être définis.

Contrôle technique

Contrôle purement formel n'impliquant aucune interprétation humaine dans son traitement ultérieur. Par exemple, le contrôle sur la forme d'une date (2010-12-31 et non 31-12-2010), contrôle détectant une somme arithmétique erronée, une erreur dans le schéma XML…

Contrôles sur le domaine de définition

Contrôle formel de la valeur d'un champ intégrant une interprétation humaine du domaine d'application (aspect « business » ou métier) associé à la base de données ou au flux de données. Par exemple, le type du contrat est 0 (temps plein) ou 1 (temps partiel).

Contrôles croisés internes

Contrôle formel (technique, portant sur le domaine de définition ou sur la présence d'un champ) émanant de la confrontation de plusieurs champs internes à une banque ou un flux de données.

Exemples :

- Incohérence entre le champ « commune » et le champ « code postal ».
- Incohérence entre le nombre déclaré de jours de travail et le nombre de jours prévu par le régime de travail.

Contrôles référentiels

Contrôle croisé entre la valeur ou la présence d'un champ (ou d'un ensemble de champs) d'une base de données (ou d'un flux de données) et une source externe de référence (souvent appelée « référentiel » ou « source authentique »)[56] (1.2.3).

En principe, le référentiel est considéré comme fiable, ce qui implique que si les données ne satisfont pas au contrôle référentiel, la faute est supposée se trouver dans les données contrôlées et non dans le référentiel.

Exemples :

- Comparaison entre le numéro d'identification d'une entreprise déclaré dans la DmfA et le numéro de cette entreprise dans le « Répertoire des employeurs de l'ONSS ».
- Comparaison entre l'adresse d'un travailleur dans une base de données de la sécurité sociale et son adresse dans le « registre national ».

Contrôles de confrontation (contrôles croisés externes)

Contrôle croisé entre la valeur ou la présence de champs issus de deux sources de données distinctes (bases et/ou flux de données) ne faisant pas office de référentiel.

Dans ce cas, il n'existe donc pas de référentiels implicitement fiables. Par conséquent, si une anomalie est détectée, il n'est pas possible de savoir *a priori* dans quelle(s) source(s) l'erreur est située.

Par exemple :

- Un travailleur fait l'objet d'une déclaration à l'embauche mais ne figure pas dans les déclarations introduites trimestriellement auprès de la sécurité sociale (ONSS).
- Une confrontation entre l'adresse d'un même travailleur dans Dimona[57] et dans la DmfA.

Contrôles de continuité

Contrôle visant à vérifier la cohérence des données dans le temps. Le contrôle peut concerner :

- **Un « item » ponctuel** : vérification de la cohérence entre une donnée (ou un ensemble de données) de la période « P » et de la période précédente « P-1 ».

  Exemple : la cohérence entre la catégorie d'activité d'un employeur au trimestre T et au trimestre T-1.

- **L'accumulation de données** : regroupement de données portant sur des périodicités différentes et vérification de la cohérence de ces données.

  Exemples :

  - Comparaison entre un total annuel de rémunérations, déclaré en fin d'année, et la somme des rémunérations déclarées chaque trimestre. Dans ce cas, il n'est pas possible de déterminer *a priori* quelles données sont erronées.
  - Bons de cotisations[58] : nombre de jours de vacances (déclarés trimestriellement dans la DmfA) supérieur au total annuel autorisé.

[56] Voir e.a. TRIGAUX J.-C., *Idem*.

[57] La déclaration immédiate de l'emploi ou DIMONA (*Déclaration Immédiate/Onmiddellijke aangifte*) est un message électronique par lequel l'employeur communique les entrées et les sorties de service de son personnel à l'ONSS. La dimension électronique du message rend celui-ci immédiat, c'est-à-dire direct et instantané.

- Étudiants : nombre de jours de travail prestés comme étudiant (déclarés trimestriellement dans la DmfA pour les différents employeurs) supérieur au total annuel autorisé.

- **La complétude** : contrôles de détection d'instances ou d'éléments de flux attendus mais manquants. Cette notion nécessite un arbitrage, une définition claire de ce que veut dire « attendu ».

  Par exemple, trimestre manquant sur une année ou diminution de la réserve acquise dans un plan de pension complémentaire.[59]

Remarques :

- Dès que plus d'une source de données est concernée, les questions suivantes deviennent cruciales :
  - rythme de mise à jour (synchrone ou pas) des sources de données impliquées dans un contrôle ;
  - domaine de définition (cohérent ou pas) des sources de données impliquées dans un contrôle.
- La typologie que nous proposons ci-dessous est indépendante du moment où le contrôle est effectué. En effet, un contrôle peut être effectué soit directement à la réception (*ex ante*), soit lorsque les données sont globalisées ou transmises à un autre organisme (*ex post*) (1.3.4).

Connaissances requises

Les contrôles présentés ci-dessus sont complémentaires. Leur implémentation et leur gestion nécessitent des connaissances tant techniques, métier que juridiques non négligeables. Ces connaissances doivent donc être pensées dès le début d'un projet : il faut réunir les connaissances requises et en assurer la pérennité tout au long de l'existence des contrôles.

Lors de la définition des contrôles, les analystes doivent définir les contrôles à effectuer, en déterminant les critères et les paramètres qui seront utilisés, ainsi que les différents cas de figure qui doivent être pris compte.

De leur côté, les personnes métier et les juristes doivent les aider à déterminer les cas de figure qui ne se posent pas ou les critères pertinents à prendre en considération en fonction de la législation.

Par exemple, dans le cadre des déclarations sociales pour les (pseudo-)prépensionnés, il faut contrôler la part patronale de l'indemnité complémentaire payée au travailleur[60], sur laquelle l'employeur paie une cotisation. En cas de difficulté (l'entreprise est qualifiée en difficulté ou en restructuration), la législation belge prévoit une réduction de cotisations afin d'aider l'entreprise. Or, cette réduction s'applique uniquement au secteur marchand puisque le secteur non-marchand bénéficie déjà de taux de cotisation relativement bas. Certains contrôles ne doivent donc pas être prévus pour le secteur non-marchand.

[58] Chaque travailleur salarié, travailleur indépendant ou assuré social disposant d'un revenu de remplacement peut, sous certaines conditions, bénéficier auprès de sa mutualité de l'assurance légale soins de santé et indemnités. Pour cela, les institutions de sécurité sociale qui perçoivent les cotisations ou effectuent des retenues, ainsi que les institutions de sécurité sociale qui accordent des revenus de remplacement communiquent, de façon électronique, aux mutualités les cotisations payées et les retenues ou les revenus de remplacement accordés. Les mutualités sont ainsi en mesure d'ouvrir le droit à l'assurance soins de santé et indemnités.

[59] Dans le cadre du deuxième pilier des pensions en Belgique, l'employeur constitue progressivement un capital pour les travailleurs de son entreprise auprès d'un organisme de pensions. Ce capital est considéré comme « acquis », ce qui implique qu'une diminution de ce capital n'est pas possible.

[60] Lorsqu'un travailleur part en prépension, il bénéficie d'une indemnité complémentaire payée par l'employeur (part patronale) et d'une allocation de chômage payée par l'ONEM.

2.3.2. Contrôles *ex ante – ex post*

Une distinction peut être faite entre

- un contrôle ***ex ante*** : contrôle effectué dès réception des données ;
- un contrôle ***ex post*** : contrôle effectué après qu'un laps de temps se soit écoulé entre la réception des données et l'exécution du contrôle. L'exécutant du contrôle peut également être un organisme différent suite à un échange de données (dans le cadre de l'e-Government).

Ces deux types de contrôles sont complémentaires, certaines incohérences ne pouvant être détectées qu'a posteriori, comme par exemple une incohérence sur un état accumulé (cf. ci-dessus) ou des cas de doublons.

Ces deux types de contrôles doivent être envisagés dès la création de l'application au risque de ne pas disposer des outils pour gérer ces anomalies détectées *ex post* de manière efficace.

Prenons un exemple. Une base de données comprend des adresses. *Ex post*, une erreur est détectée dans l'adresse et doit être corrigée. Cependant, la base de données n'a été conçue que pour contrôler les données à leur entrée et sur l'hypothèse en résultant qu'il n'y a pas de donnée erronée dans la base. Aucun historique des modifications n'existant dans la base, il faut introduire un nouveau record dans la base de données avec la bonne adresse, tout en marquant l'ancienne valeur comme inactive. Entre ces deux records, il n'est pas possible de savoir s'il s'agit d'une entreprise qui a déménagé ou s'il s'agit de la correction d'une anomalie. Par conséquent, aucune stratégie efficace de gestion des anomalies ne pourra être appliquée à cette base de données.

2.3.3. Séquence de contrôles

L'exécution des contrôles étant séquentiel, il convient d'en déterminer l'ordre, aussi bien au sein d'une même source d'informations qu'entre plusieurs sources.

La séquence de corrections doit être déterminée avec soin dans la mesure où elle influe directement sur la qualité des données et par conséquent sur la charge de travail des agents en charge de la gestion des anomalies. Les enjeux sont donc importants.

Nous présentons les types de critères à intégrer pour modéliser la séquence des contrôles. Cette modélisation est source de conflits qui ne peuvent être systématiquement écartés. Nous présentons ensuite les différents arbitrages auxquels la construction de la séquence est soumise.

Typologie des critères pour modéliser la séquence des contrôles

Cohérence des contrôles

Par souci de cohérence, certains contrôles doivent être effectués prioritairement. Lors de la réception d'un flux, il est normal que le système vérifie prioritairement si le fichier est bien formé, c'est-à-dire conforme au modèle de données attendu. De même, le système vérifiera tout d'abord la présence d'un champ avant d'en contrôler le domaine de définition.

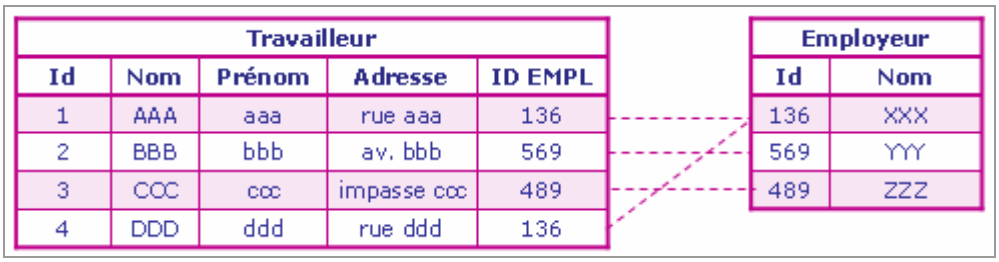
Dépendance entre données

Au sein d'une même source, les données sont souvent interdépendantes, ce qui signifie que la valeur et la validité d'une donnée (ou de plusieurs) est dépendante

de la présence ou de la valeur d'une autre donnée (ou de plusieurs) (cf. contrôle croisé interne). Trois types de dépendances sont envisagés :

- **Contrainte d'intégrité référentielle** : impose que, dans une relation R_1, la valeur d'un attribut (ou d'un groupe d'attributs) forme la clé primaire dans une autre relation R_2.

  Par exemple, lors de la déclaration des prestations d'un travailleur, l'employeur pour lequel ces prestations sont effectuées doit être mentionné. Cette mention doit référencer un employeur existant, elle est donc soumise à une contrainte d'intégrité référentielle (Figure 5).



Travailleur				
Id	Nom	Prénom	Adresse	ID EMPL
1	AAA	aaa	rue aaa	136
2	BBB	bbb	av. bbb	569
3	CCC	ccc	impasse ccc	489
4	DDD	ddd	rue ddd	136

Employeur	
Id	Nom
136	XXX
569	YYY
489	ZZZ

Figure 5 : Exemple de contraintes d'intégrité référentielle

  Dans le cadre d'une séquence de contrôles, il faut veiller à ce que la donnée référencée (R_1) soit contrôlée avant sa référence dans une autre relation (R_2).

- **Dépendance fonctionnelle** : existe entre un attribut A_1 (ou groupe d'attributs A_1, A_2,… A_n), dit attribut source, et un attribut B_2, dit attribut cible, si connaissant une valeur de A_1 (ou d'un groupes d'attributs A_1, A_2,… A_n) on ne peut lui associer qu'une seule valeur de B_2.

  Il existe par exemple une dépendance fonctionnelle entre un employeur et sa catégorie (secteur d'activité) qui lui est attribuée par l'ONSS, puisque chaque employeur ne peut avoir qu'une seule catégorie d'activité. La réciproque n'est pas vraie puisque pour une même catégorie d'activité, il peut y avoir plusieurs employeurs.

  Une dépendance fonctionnelle porte sur la sémantique des données et ne peut être déterminée automatiquement à partir d'un record. Si nous reprenons le même exemple, il se pourrait que pour une catégorie d'activité X, il n'existe qu'un seul employeur référencé. Mais il s'agit d'une situation de fait qui n'est que temporaire puisque, à tout moment, un employeur peut se voir attribuer cette catégorie. Il ne s'agit donc pas d'une dépendance fonctionnelle.

  Il est également possible que des dépendances fonctionnelles disparaissent suite à une évolution de la réalité y afférente. Par exemple, en Belgique, pendant longtemps les premiers chiffres d'un numéro de téléphone fixe indiquaient la région de l'abonné (02 = Bruxelles, 03 = Anvers…). Cependant, la possibilité donnée aujourd'hui aux abonnés de déménager sans changer de numéro de téléphone a entraîné la disparition de cette dépendance.

  Lors de l'élaboration de la séquence de contrôles, l'attribut source doit être contrôlé (contrôle technique et domaine de définition) avant sa dépendance par rapport à l'attribut cible.

- **Dépendance/règle métier** : enfin, il existe des dépendances entre données qui traduisent une logique métier. Ces dépendances peuvent être plus ou moins complexes selon les cas.

Ainsi, dans les déclarations des prestations sociales, l'employeur doit indiquer le nombre de jours prestés par un travailleur. Ce nombre de jours déclaré doit correspondre au régime de travail déclaré (temps plein vs temps partiel). En effet, si un trimestre compte 60 jours ouvrables et que le travailleur est déclaré à mi-temps, le nombre de jours déclarés pour ce travailleur ne peut excéder 30 jours. Cette règle comporte évidemment des exceptions nécessitant le recours à d'autres données pour en vérifier la validité (par exemple le nombre de jours prestés peut être supérieur à celui prévu par le régime de travail si un justificatif est présent). Une règle métier est donc définie et appliquée pour vérifier la cohérence de ces données (Figure 6).

Nb jours max	Nb jours déclarés	Régime travail	Justification	Résultat contrôle
60	30	mi-temps	-	OK
60	40	mi-temps	présent	OK
60	40	mi-temps	absent	Anomalie
60	80	tps plein	absent	Anomalie

Figure 6 : Exemple de règle métier simple impliquant plusieurs données

C'est principalement dans le cadre de ce type de dépendance que les risques de conflits entre données sont les plus importants et que des arbitrages seront nécessaires.

Rythme de mise à jour

Dans le cadre de contrôles impliquant plusieurs sources de données (contrôles référentiels et de confrontation), il est important de se poser la question du rythme de mise à jour de chaque source d'informations. Eu égard aux besoins, il peut être pertinent de contrôler les données *ex post.*

Cette situation peut être illustrée à l'aide de l'exemple suivant (Figure 7). Pour calculer et attribuer la pension des agents statutaires de la fonction publique, le Service des Pensions du Secteur Public (SdPSP) utilisera prochainement les données déclarées par les employeurs du service public dans la DmfA pour les administrations provinciales et locales (DmfA-APL). À ce titre, l'employeur devra mentionner l'échelle de traitement auquel est soumis le travailleur statutaire. Cette donnée devra donc être validée eu égard aux valeurs autorisées, c'est-à-dire que le système vérifiera que l'échelle déclarée est reprise dans le référentiel géré par le SdPSP.

Pour pouvoir effectuer ce contrôle au moment de la réception, la liste des valeurs autorisées devrait être trimestrielle à l'instar de la DmfA et ne pourrait donc évoluer que tous les trois mois.

Or, ces valeurs ne peuvent être élaborées par le SdPSP que si l'employeur l'a prévenu préalablement qu'il a instauré une nouvelle échelle. Le SdPSP doit ensuite communiquer la référence correspondante à l'employeur. Au vu du caractère asynchrone de ces flux, une anomalie pourrait être détectée si le code fourni par le SdPSP est utilisé par le déclarant, mais que le référentiel n'a pas encore été mis à jour lorsque la déclaration est envoyée (et le référentiel ne sera donc pas mis à jour avant le trimestre suivant en raison du caractère trimestriel de la DmfA).

Par conséquent, au vu de cette différence, il est préférable de ne pas effectuer le contrôle dès la réception de la déclaration mais plutôt *ex post*.

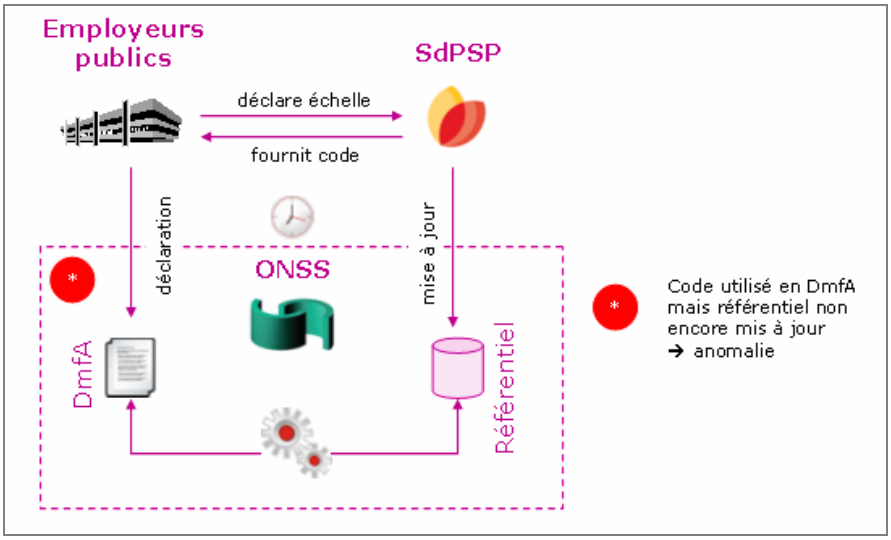


Figure 7 : Risque d'incohérence en raison du rythme de mise à jour hétérogène des données

Conflits potentiels et arbitrage

Quels que soient les critères retenus, il n'est pas possible de construire une séquence de contrôles exempte de conflits. Dès lors, l'enjeu n'est pas tant de les éviter que, partant du constat qu'ils sont inhérents à toute séquence de contrôle, de chercher à en minimiser les effets eu égard aux besoins et aux enjeux poursuivis.

Cohérence des contrôles et dépendance entre données

En raison des dépendances entre données, il peut arriver qu'une donnée non encore contrôlée serve à en contrôler une autre, situation qui présente un double risque : tout d'abord un accroissement du nombre d'anomalies en raison de la génération d'anomalies fictives[61], ensuite l'inutilité de certaines corrections.

Pour illustrer le premier risque, prenons deux données numériques *A* et *B* soumises à la règle suivante $A < B$. Imaginons que *A* est présente mais de type alphanumérique, tandis que *B* est correcte. Deux scénarios sont envisageables comme l'illustre la Figure 8.

Dans le premier scénario, la donnée *A* est totalement contrôlée avant d'être utilisée pour le contrôle de la donnée *B*. Une anomalie étant détectée sur *A* à l'étape 2, l'étape 3 n'est pas exécutée. Une seule anomalie est détectée.

Dans le second scénario, *A* n'est pas totalement contrôlée avant d'être utilisée dans un contrôle croisé. Le système détecte donc une erreur sur *B* à l'étape 2 et détecte ensuite une anomalie sur *A* à l'étape 3. Or, *B* est correcte. L'anomalie pointée sur *B* est fictive puisqu'elle résulte de l'erreur de *A*.

[61] Une anomalie fictive est une donnée qui, bien que correcte, est pointée comme étant formellement une anomalie en raison de sa dépendance à une autre donnée incorrecte (formellement ou non).

Scénario	Etape	Contrôle	Résultat
1	1	Présence de A	OK
	2	Type de A	Anomalie
	3	Validité de B	Pas exécuté
2	1	Présence de A	OK
	2	Validité de B	Anomalie
	3	Type de A	Anomalie

Figure 8 : Ordre des contrôles et risque de génération d'anomalie fictive

Par ailleurs, la possibilité offerte de contrôler une donnée à partir d'une donnée non encore contrôlée peut engendrer des cycles d'incohérence (Figure 9). Par exemple, pour contrôler la donnée *B*, la donnée *A* est nécessaire. Cette même donnée *A* a besoin d'une donnée *C* pour être validée. Enfin, cette donnée *C* est validée à partir de la donnée *B*, entraînant un cycle d'incohérences et de corrections rendu possible parce que les données utilisées pour des contrôles croisés ne doivent pas nécessairement avoir été préalablement contrôlées.

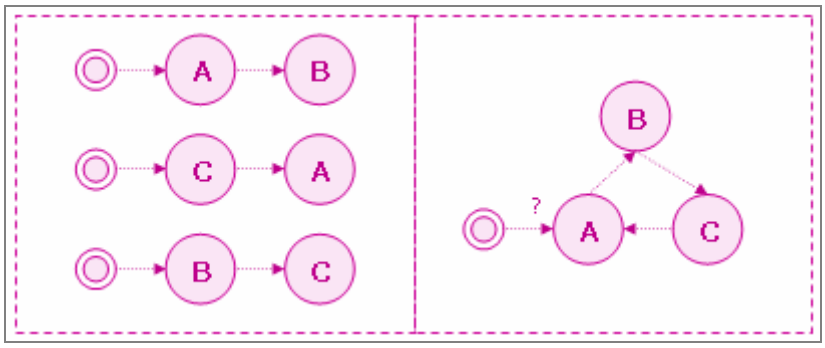


Figure 9 : Risque de cycle d'incohérence

Certains outils de *Business Rules Management System (BRMS)*[62] sont capables de détecter automatiquement ce type d'incohérences dans les séquences de corrections. Ils sont cependant relativement chers[63] et aucun n'est capable de détecter tous les types d'incohérences qui peuvent exister.

Le second risque que nous évoquions est que les corrections portant sur des anomalies découvertes à partir de contrôles croisés pourraient s'avérer inutiles. En effet, une anomalie pourrait n'apparaître que parce que le contrôle sur cette donnée se base sur une autre donnée également fautive, la correction de la première donnée impliquant automatiquement une validation de la seconde donnée, préalablement pointée comme anomalie.

Prenons un exemple illustré par la Figure 10.

1. Une anomalie est détectée sur les données A et B dans le cadre d'un contrôle croisé entre elles (la valeur « b » est incorrecte par rapport à la valeur « a »).
2. La donnée B est corrigée et sa valeur passe de « b » à « c », ce qui lève l'anomalie en B (cohérence de « c » par rapport à « a »).

[62] OGONOWSKY G., *Business Rules Technologies & web sémantique : La gestion des règles business*, Delivrable, 2008/TRIM1/01, Smals, Section Recherches, Bruxelles, 2008.
[63] À titre d'exemple, le coût d'une licence ILOG d'IBM avoisine les 500.000 €.

3. La donnée A est corrigée et sa valeur est modifiée de « a » en « d », ce qui réintroduit une anomalie sur B puisque la valeur « c » est incorrecte par rapport à « d » (mais la valeur « b » initiale était, elle, correcte).
4. La donnée B est à nouveau corrigée, sa valeur passe de « c » à « b », ce qui est correct par rapport à la valeur « d » de A.

Au final, trois corrections auront été nécessaires au lieu d'une seule.

Etape	Donnée	Valeur	Conséquence
1	A	a	Anomalie
	B	b	Anomalie
2	A	a	Anomalie
	B	b -> c	OK
3	A	a -> d	OK
	B	c	Anomalie
4	A	d	OK
	B	c -> b	OK

Figure 10 : Corrections multiples à cause des dépendances entre données

Dans la pratique, il n'est pas possible d'éviter les conflits entre dépendances. Ceux-ci doivent être identifiés et documentés en prenant soin de souligner les choix effectués et leur justification.

Comme nous l'avons expliqué précédemment (2.2.3), les interdépendances entre données et les anomalies fictives sont les principaux freins à la mise en place de séquence de corrections des anomalies.

Arbitrage : qualité vs disponibilité des données

Certaines données relèvent d'une importance stratégique pour qu'une institution puisse exercer ses missions. Il s'agit de veiller à leur qualité en y appliquant des contrôles stricts. Une erreur détectée dans ces données peut entraîner le rejet des données dans leur ensemble.

Par exemple, dans le cadre de la DmfA, la « catégorie employeur » est fondamentale pour déterminer les cotisations sociales et les droits des travailleurs (notamment selon la commission paritaire). Il a été convenu que cette donnée ne peut pas contenir d'erreur. Si une anomalie est détectée, la déclaration est rejetée et l'employeur doit corriger l'erreur et soumettre à nouveau sa déclaration sous peine d'amendes.

Toutefois, pour qu'une institution puisse garantir ses missions, il n'est pas envisageable de rejeter systématiquement toutes les données. Par conséquent, les institutions sont obligées de trouver un juste équilibre entre le fait d'autoriser la réception dans leurs bases de données de données comportant des erreurs et leur gestion, selon un arbitrage de type « coûts-bénéfices », qui dépendra des besoins.

Par exemple, les contrôles portant sur l'identification des travailleurs sont plus stricts dans le cadre des déclarations de risque social (DRS)[64] que des DmfA puisque toute anomalie détectée entraîne le rejet de la déclaration, ce qui n'est pas le cas en DmfA. Cette situation s'explique par le fait qu'une DmfA peut

[64] Un risque social se présente lorsqu'un travailleur ne peut prétendre à son salaire, par exemple en cas de maladie, d'accident du travail, de maladie professionnelle, de chômage ou de maternité. Dans de telles circonstances, la sécurité sociale prévoit un large éventail d'allocations sociales. Pour que le travailleur puisse en bénéficier, l'employeur doit introduire une DRS (déclaration de risques sociaux) auprès d'une des institutions de sécurité sociale.

comporter plusieurs milliers de travailleurs, ce qui augmente la probabilité de trouver au moins une erreur, et permet de récolter anticipativement les données de sécurité sociale du travailleur (ce qui donne le temps à l'expéditeur de corriger ses erreurs). Au contraire, une DRS ne concerne qu'un seul travailleur et est émise lorsqu'un risque social survient (chômage, invalidité, accident de travail…). Les données doivent donc être récoltées rapidement et correctement.

Cet arbitrage peut être géré en déterminant un seuil de qualité minimal auquel les données doivent satisfaire pour être utilisées, principalement les données ayant une valeur très importante et en graduant ce seuil selon l'importance des données et le type d'anomalies. Pour ce faire, des indicateurs de qualité doivent être définis afin d'offrir des mesures objectives sur lesquelles les décisions peuvent être prises.

Dans certains cas de données déclarées par le monde extérieur à une institution publique, les anomalies sont classées en différents niveaux d'importance et la déclaration est traitée différemment (rejet total, partiel ou acceptation) selon les niveaux auxquels les anomalies appartiennent (2.2.1).[65]

Selon les cas, certains éléments utilisés pour générer ces indicateurs doivent être renseignés dans la documentation des anomalies et le modèle générique d'historique des anomalies afin que le traitement puisse être automatisé.

Arbitrage : contrôle des données vs besoins et ressources disponibles

La création et la mise en œuvre des contrôles ont pour objectif de vérifier la validité et la qualité des données qui sont utilisées pour qu'une institution puisse remplir ses missions.

Cependant, plus de contrôles entraînent *de facto* potentiellement un plus grand nombre d'anomalies. Ce qui se traduit par un accroissement de la charge de travail pour l'institution qui doit les gérer.

Par conséquent, chaque institution doit trouver le juste équilibre entre le souhait de contrôler les données et ses besoins (et éventuellement ceux d'un organisme tiers pour lequel elle exerce certaines missions), tout en tenant compte des ressources disponibles.

Il est évidemment possible d'identifier des anomalies qui ne doivent pas être traitées. Dans ce cas, il est légitime de s'interroger sur leur utilité dans la mesure où ces contrôles doivent eux aussi être gérés (mises à jour métier et technique selon l'évolution de la source d'information, cohérence par rapport aux autres contrôles, ressources systèmes nécessaires pour leur exécution, etc.).

Si certaines anomalies ne doivent pas être corrigées et que l'institution souhaite maintenir les contrôles y afférents, une documentation argumentée de ces choix et sa publicité auprès des personnes concernées s'imposent.

Arbitrage : rapidité vs stabilité

Il est parfois préférable de laisser s'écouler un laps de temps avant l'exécution de certains contrôles. Les données confrontées seront à ce moment plus stables. Ce délai peut permettre de diminuer le nombre d'anomalies détectées. Ce choix requiert un arbitrage entre le souhait de disposer rapidement d'indicateurs sur la qualité des données et l'intérêt d'attendre une plus grande stabilité des données en vue de réduire le nombre d'anomalies générées.

[65] Dans le cadre de la DmfA, les anomalies sont classées en bloquantes, pourcentuelles et non pourcentuelles. Seules les anomalies bloquantes ou un nombre trop élevé d'anomalies pourcentuelles au sein d'une déclaration entraînent son rejet.

Par exemple, l'ONEM considère que les données concernant le chômage ne sont réellement stables qu'après treize mois. L'utilisation et la confrontation de ces données avec d'autres données issues de la sécurité sociale avant ce délai pourraient entraîner la génération de nombreuses anomalies qui n'apparaîtraient pas si les données ONEM étaient utilisées passé ce délai.

Dans le premier cas, l'exécution de ces contrôles permet d'obtenir rapidement des informations sur la présence d'anomalies. Dans le deuxième cas, un délai de plusieurs semaines, voire mois, sera nécessaire avant que cette information ne soit disponible.

Arrêt de contrôles

Lorsqu'une anomalie est détectée sur une donnée, il convient de déterminer si les contrôles en relation avec cette donnée doivent être effectués ou non. Pour reprendre l'exemple *supra*, si une erreur est détectée sur la donnée « âge », est-il judicieux de vérifier la validité des cotisations déclarées se basant sur l'âge ? À nouveau, ce choix dépend des besoins. L'arbitrage est le même qu'au cas précédent.

Dans le cas où les contrôles ne sont pas effectués, il faut être conscient que la correction de l'anomalie ayant stoppé les contrôles peut avoir pour conséquence l'apparition d'un plus grand nombre d'anomalies suite à l'exécution des contrôles auparavant non effectués.

2.3.4. Recommandations pour la mise en œuvre

- Réunir les connaissances et les profils (rôles) nécessaires à la gestion des contrôles et mettre en place une procédure de collaboration pour la mise à jour des contrôles (2.3.1).
- Documenter les contrôles : la documentation des contrôles et de la séquence est fondamentale pour pouvoir gérer au mieux les données et tenir compte des implications afférentes aux choix opérés. Cette documentation comprend à la fois :
  - une description humaine et métier des contrôles et des anomalies en résultant ;
  - les interdépendances entre données ;
  - la hiérarchie et l'ordre d'exécution des contrôles ;
  - la modélisation de la séquence qui permet de mettre aisément en évidence les incohérences et les conflits ;
  - une explication des arbitrages et des choix opérés.

  Cette documentation doit être maintenue et mise à jour à chaque évolution des données ou des contrôles. Une partie importante de cette documentation peut être introduite directement dans le modèle générique proposé dans le présent document.

  Les ressources pour cette mise à jour sont à prévoir dès le début.
- Pour garder la cohérence des contrôles et éviter la génération d'anomalies fictives en raison de l'interdépendance des données, il est important que les données soient d'abord validées de manière autonome (forme et domaine de définition) avant d'être utilisées pour des contrôles croisés. Si ce n'est pas possible, un choix doit être opéré et documenté. Les implications de ce choix doivent être expliquées aux personnes en charge des corrections, par exemple dans un système de gestion des connaissances (2.4.2).

- Dans un contrôle, si plusieurs sources de données sont confrontées (contrôle référentiel ou contrôle de confrontation), il est fondamental d'analyser les rythmes d'update des sources utilisées (synchrone ou pas) et leur domaine de définition au risque de générer des anomalies fictives et non pertinentes ou de confronter des données dont la sémantique est différente.
- Évolution des contrôles : les contrôles sont fortement liés aux données. Dès lors, toute évolution de celles-ci devra généralement se refléter dans la séquence de contrôles. Par ailleurs, il est souhaitable de faire évoluer les contrôles eu égard aux besoins et aux ressources disponibles. À cet égard, un suivi systématique des corrections effectuées permet de détecter les anomalies qui sont systématiquement validées par les agents et par conséquent de pointer les contrôles qui doivent éventuellement évoluer.

2.4. Documentation opérationnelle du système d'information

En vue d'accompagner les recommandations présentées dans cette étude, il est impératif, à des fins opérationnelles, de documenter le système d'information afin de permettre une interprétation univoque des données, de leur domaine de définition et de leurs modalités de traitement. Les différents points qui suivent s'inspirent d'applications concrètes mises en place sur le terrain dans le cadre des consultances « data quality » menées au sein de l'administration fédérale belge. Nous en avons extrait des recommandations généralisables à l'ensemble du domaine d'application. Dans un premier temps, les fonctionnalités à mettre en œuvre en vue de documenter une base de données sont présentées, sur la base de l'exemple des glossaires de la sécurité sociale. Les caractéristiques d'un système de gestion de la connaissance en vue d'accompagner la correction et le traitement des anomalies sont ensuite évoquées. Enfin, l'organisation globale à mettre en œuvre en vue de soutenir la documentation générale d'un système d'information est présentée.

2.4.1. Fonctionnalités à mettre en œuvre

La législation étant complexe et évolutive, un « dictionnaire électronique des données » est indispensable afin de faciliter l'interprétation de l'information administrative et des directives techniques correspondantes. À cette fin, dans le cadre des services constitutifs de l'administration électronique, un système de méta-information collaboratif multilingue a été conçu au sein de la sécurité sociale belge. Celui-ci a été déployé dans un environnement web en vue de documenter les messages XML échangés entre le citoyen ou l'entreprise et l'administration. Ce système de méta-information est en production depuis 2001[66] et s'est enrichi depuis lors[67]. Il est maintenant en cours de reengineering et permettra, au-delà des fonctionnalités de base que nous présentons ici, une ouverture aux services web et une plus grande flexibilité. L'application s'adresse à la fois aux informaticiens en charge de la gestion des bases de données et aux

[66] Pour plus d'information, voir : BOYDENS I., « E-gouvernement en Belgique. Un retour riche d'expériences », *L'informatique professionnelle (Dossier spécial "Services publics")*, Editions Gartner France, Paris, Numéro 217, octobre 2003, p. 29-35.

[67] https://www.socialsecurity.be/lambda/portail/glossaires/dmfa.nsf/web/glossary_home_fr
https://www.socialsecurity.be/lambda/portail/glossaires/dmfa.nsf/web/glossary_home_nl

instances en charge de l'envoi des messages électroniques, l'objectif étant que tous travaillent sur une base commune.

Les systèmes de méta-information comportent potentiellement trois écueils. Le premier est lié au fait que ces systèmes sont extensibles à l'infini, surtout lorsque les champs à compléter sont « libres », le langage naturel étant son propre méta-langage. Ceci implique des coûts importants en termes de gestion, lorsque les mises à jour manuelles sont nombreuses. Le second écueil tient à ce que les métadonnées peuvent être elles-mêmes erronées et incertaines. Lorsqu'elles sont d'ordre contextuel, leur validation ne peut faire l'objet de contraintes d'intégrité rigoureuses. Le troisième écueil tient au décalage temporel entre la mise à jour d'une donnée et de la métadonnée correspondante, cette dernière, surtout lorsqu'elle se présente sous une forme textuelle, n'étant généralement créée qu'au terme d'une phase d'analyse. Sur la base de ces constats, nous avons mis en place un système destiné à préserver la cohérence de l'information et à en faciliter la gestion. Il inclut les fonctions suivantes :

- gestion semi-automatique du multilinguisme (via des tables pré-contrôlées) ;
- réutilisation des définitions communes via une procédure d'héritage (les définitions génériques telles que la codification des lieux, des adresses… sont mises à jour une seule fois et ensuite propagées dans toutes les applications documentaires spécifiques où elles interviennent) ;
- gestion des versions (lorsque les définitions techniques évoluent dans le temps, le système permet le suivi des versions et spécifie pour chaque nouvelle version la liste des modifications apportées par rapport à la version immédiatement antérieure) ;
- mise en pratique du concept de WOPM (« *write once publish many* ») : l'application inclut des listes structurées (codes postaux, catégories d'activité…) qui, dans la pratique, doivent être diffusées à des fins documentaires, mais aussi en vue de tester les données saisies dans les bases de données. Afin de rencontrer les deux fonctions, l'application a été conçue de façon à générer automatiquement une même table structurée (liste de codes postaux, par exemple) sous différents formats : formats ascii, XML, word, excel et PDF. La même source peut ainsi être utilisée au sein d'applications interdépendantes ;
- système de navigation et moteur de recherche documentaire ;
- procédures de validation : en raison des enjeux légaux, sociaux et financiers liés à la gestion des bases de données et de leur documentation, chaque nouvelle version doit être validée par les responsables de l'information sur les plans technique et juridique. En vue de structurer cette validation, un système de *workflow* guide le déploiement du dictionnaire électronique. Celui-ci s'inscrit dans le cadre d'une procédure (un planning spécifie de façon rigoureuse les périodes de mise à jour, de validation, de mise en acceptation et de mise en production). Le workflow est « piloté » de façon centralisée par une équipe dédiée à cette tâche et se déploie sur un mode décentralisé dans un environnement web. Lors de la création de chaque nouvelle version, l'historique des échanges entre les différents responsables est conservé, de façon à garder un suivi du processus d'interprétation.

La mise en place d'un tel système facilite la gestion des données alimentant les services administratifs en ligne et contribue à en garantir la qualité.

2.4.2. Gestion des connaissances : modalités de traitement des anomalies

Face à la complexité des anomalies et à la nécessité qu'elles soient traitées par des agents humains, de manière aussi homogène que possible, une documentation des anomalies s'avère extrêmement importante pour permettre un partage des connaissances entre agents, d'autant plus si la rotation du personnel au sein du service en charge de la correction des anomalies est importante. Par ailleurs, l'évolution législative influence constamment la manière dont les anomalies doivent être traitées.

En conséquence, la mise en place d'une application documentaire, de type gestion des connaissances, en vue, d'une part, de décrire et d'expliciter la signification des anomalies et, d'autre part, de favoriser un transfert et un partage de connaissances entre les agents doit être envisagée comme une stratégie de gestion des anomalies.

Dans ce cadre, nous examinons tout d'abord les principaux besoins métier, informationnels et fonctionnels. Les aspects organisationnels et technologiques d'un système de gestion des connaissances seront ensuite abordés. Nous illustrerons notre propos à l'aide d'un exemple existant au sein du service du contrôle de l'ONSS avant de conclure sur quelques recommandations importantes en vue de la mise en œuvre de ce type de système.

Besoins

Nous identifierons tout d'abord les besoins métier et les objectifs de l'application tels que déterminés par le management. Ceci nous permettra d'établir la liste des informations nécessaires aux utilisateurs et ensuite les fonctionnalités requises pour éditer et consulter le contenu de l'application.

Besoins métier

Au niveau du management, l'objectif est de mettre à la disposition des personnes qui traitent les anomalies une documentation claire et complète, régulièrement mise à jour et facilement accessible et exploitable en vue :

- d'accélérer le traitement des anomalies ;
- de favoriser un travail plus homogène pour améliorer la qualité des données et éviter les problèmes d'interprétation et leurs conséquences sur la gestion des dossiers, en ce compris entre régimes linguistiques ;
- d'éviter que la correction des anomalies ne soit perçue comme une tâche technique sans valeur ajoutée, ce qui est un facteur de démotivation pour le personnel.

Besoins informationnels

Eu égard à ces objectifs, la documentation doit permettre, du point de vue de l'utilisateur, de répondre à une quadruple question :

- Qu'est-ce que je corrige ? Quelle est la signification de l'anomalie ?
- Comment dois-je la traiter ?
- Pourquoi dois-je la traiter ? Quelle en est l'utilité ?
- Quel effort est requis pour traiter cette anomalie eu égard aux bénéfices ? Cette dernière question s'adresse davantage aux managers qui peuvent ainsi estimer le rapport coûts-bénéfices du traitement des anomalies.

Pour ce faire, la documentation comprend à la fois des informations descriptives (par exemple, le numéro de l'anomalie et son impact sur l'octroi des droits sociaux) et des informations sur la manière de la traiter (causes de l'anomalie et méthodes de correction).

Cette information peut aisément être formalisée dans une « fiche anomalie » structurée en champs selon un canevas standard propre à un domaine d'application concerné.

Certains champs dont le domaine de définition est déterminé *a priori* (liste de valeurs prédéfinies et limitées telles que les codes postaux ou les codes pays) sont structurés. D'autres ont une valeur plus difficilement formalisable et contiennent de l'information libre et textuelle. Ces derniers champs doivent faire l'objet d'une attention particulière au niveau de la rédaction en vue d'uniformiser autant que possible ces champs quelle que soit la fiche anomalie. Cette uniformisation, qui peut prendre la forme de consignes de rédaction, est nécessaire pour garder un maximum de cohérence dans la documentation et faciliter sa consultation par les utilisateurs.

Besoins fonctionnels

Pour gérer cette information, plusieurs fonctionnalités sont importantes : la présence d'une interface d'édition, un workflow de validation, la gestion de champs multilingues, la gestion des versions et un moteur de recherche. Elles sont présentées successivement :

- **Interface d'édition**

  La mise à jour des informations doit être aisée et réalisée par les gens métier eux-mêmes. Elle ne doit donc requérir aucune connaissance technique. Pour ce faire, une interface d'édition doit être disponible pour les personnes en charge du contenu. Concrètement, cette interface propose à l'éditeur une série de champs à compléter sur le modèle de la fiche anomalie définie.

  La mise en forme des champs libres devra pouvoir se faire selon les habitudes des utilisateurs. Cela exclut le recours à des signes spécifiques ou aux balises de mise en forme.

- **Workflow de validation**

  Dans le cadre de la correction d'anomalies sur des données ayant un fort impact opérationnel et juridique, il est nécessaire que toute information introduite dans le système fasse l'objet d'une validation, surtout eu égard aux conséquences possibles d'une correction erronée due à la présence d'une information de mauvaise qualité et non contrôlée dans le système[68]. À cet égard, il convient de déterminer deux rôles métier, l'un avec de fortes compétences métier, l'autre avec des compétences plus techniques (structure de la déclaration, interdépendances entre champs…). Éventuellement, ces deux rôles peuvent être regroupés en un seul suivant le profil des agents.

- **Le multilinguisme**

  Au sein de la sécurité sociale et de l'administration fédérale belge, la documentation technique et métier doit être diffusée dans les différentes langues nationales. Dans le contexte d'une documentation interne, le contenu peut se limiter aux deux langues vernaculaires de l'administration fédérale : le français et le néerlandais. Dans tous les cas, l'utilisateur final aura accès à l'information dans sa langue.

[68] Dans le cas de la DmfA, chaque correction introduite par un agent de l'ONSS est transmise à l'employeur outre le fait que cette correction est utilisée pour recalculer de nombreux droits sociaux du travailleur concerné.

Pour les champs structurés, une traduction automatique peut être effectuée sur la base de tables contrôlées multilingues lors de la saisie de l'information dans une des deux langues (voir 2.4.1).

Pour les champs libres, lors de la création d'une information ou de sa mise à jour, le workflow doit inclure une étape de traduction. La cohérence de l'information (fond et forme) doit être maintenue via des recommandations pour la rédaction du contenu entre les deux langues au risque que l'information ne soit pas équivalente entre les régimes linguistiques et que le traitement de l'anomalie diffère.

Il existe principalement deux méthodes pour implémenter le multilinguisme (Figure 11).

La première consiste à créer une seule instance de la fiche comportant tous les champs quelle que soit la langue utilisée. Lors de l'édition de la fiche, une vue permet de n'afficher que les champs de la langue spécifiée.

La seconde méthode consiste à créer une version de la fiche par langue et à les relier entre elles.

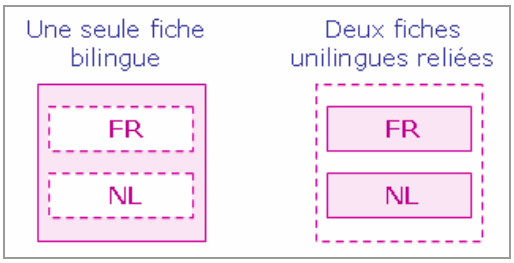


Figure 11 : Deux méthodes pour implémenter le multilinguisme

La première méthode doit être privilégiée. Tout d'abord, elle facilite la traduction car il est possible de disposer parallèlement du texte à traduire et du champ correspondant dans l'autre langue. En outre, cette méthode renforce la cohérence et la qualité de l'information car elle doit d'abord être finalisée dans une langue (c'est-à-dire création et validation) avant d'être traduite. Dans le cas contraire, les diverses versions linguistiques de l'information risqueraient d'évoluer différemment, ce qui irait à l'encontre des objectifs.

- **La gestion des versions**

  Sous l'influence de la législation et des directives administratives, le domaine d'application et par conséquent les bases de données évoluent. Cette évolution se traduit concrètement tant par la création de nouveaux champs et de nouveaux contrôles (et donc de nouvelles anomalies) que par l'évolution des anomalies existantes. Les agents étant souvent amenés à corriger un état antérieur des données, il est important que cette correction puisse se faire sur la base de la méthode corrective appliquée à ce moment. Ainsi, si les contrôles générant l'anomalie évoluent entre un instant *T* et *T+1*, l'anomalie existante en *T* ne devra pas être corrigée de la même manière qu'à l'instant *T+1*.

  Cette gestion des versions doit également permettre à l'utilisateur de visualiser facilement les changements intervenus entre deux versions différentes. Par défaut, l'utilisateur souhaite généralement obtenir la dernière version.

  L'évolution de l'information peut se faire de différentes manières. L'objectif est d'éviter une inadéquation progressive entre le contenu de la base et la documentation correspondante.

La première consiste à faire évoluer l'ensemble des fiches de manière synchrone, soit selon un planning similaire à l'évolution de la source, soit selon un planning régulier. À chaque évolution, les fiches sont adaptées et publiées sous une nouvelle version.

La seconde consiste à faire évoluer les fiches de manière asynchrone, c'est-à-dire indépendamment les unes des autres. Chaque fiche est donc modifiée au gré des besoins et une nouvelle version est publiée après validation des modifications.

La première approche (évolution synchrone) permet de disposer à tout moment d'une information mise à jour. La seconde approche (évolution asynchrone) est plus flexible mais garantit moins la cohérence de l'information. Son inconvénient majeur est qu'un doute peut survenir chez les utilisateurs sur la fraîcheur et donc la pertinence de l'information. La mention de la date de dernière modification de la fiche est dans ce cas indispensable.

Cette fonctionnalité est d'une importance relative eue égard au caractère interne de l'information, notamment par rapport à une documentation à usage externe telle que les glossaires DmfA (voir 2.4.1). De plus, l'expérience que nous en avons est encore trop récente pour pouvoir évaluer sa pertinence sur le long terme.

- **Un moteur de recherche**

  Un moteur de recherche est évidemment indispensable. Dans ce cas-ci, il peut s'avérer extrêmement simple et se limiter à deux types de recherches : une recherche sur la base du numéro (identifiant) de l'anomalie et une recherche full-text (type Google). Pour le premier type de recherche, il peut être possible, grâce à l'unicité de l'identifiant, de présenter directement la fiche à l'utilisateur sans passer par une liste de résultats (similaire à une recherche Google « J'ai de la chance »).

Système de gestion des connaissances : aspects organisationnels et technologiques

Au vu des besoins exprimés ci-dessus, la mise en place d'une application de gestion des connaissances s'avère pertinente. L'objectif d'une telle application est de permettre la création des connaissances, leur capture, leur gestion, leur partage et enfin leur application à un domaine déterminé en vue d'atteindre des objectifs spécifiques.

Organisation

Pour remplir ces objectifs, la gestion de l'information doit être organisée. Cette organisation peut prendre différentes formes. Pour schématiser, elle peut être soit centralisée soit décentralisée, chacune répondant à des besoins différents comme le montre le tableau suivant :

	Centralisée	**Décentralisée**
Source d'information	Pool d'« experts »	Tout le monde
Auteurs	Expert, rédacteur	Tout le monde
Motivation	Demande / mission	Demande / volontariat
Organisation	Forte	Moyenne
Qualité du contenu	Contrôlé	Contrôlé partiellement ou non contrôlé
Volume	Limité pour chaque pool d'experts	Potentiellement grand
Modèle de collaboration	Formel	Informel

Le mode centralisé représente davantage une approche « top-down », à savoir un pool d'experts enregistre sa connaissance dans le système en vue de la diffuser vers les personnes moins expertes. Le mode décentralisé fonctionne à l'inverse (approche « bottom-up »). Il n'y a pas d'experts désignés et chaque personne travaillant sur le sujet est en droit d'introduire de la connaissance dans le système. Il est possible de prévoir des systèmes intermédiaires. Par exemple, toute personne peut introduire de la connaissance dans le système, mais elle doit être validée par un groupe d'experts désignés. Dans le premier mode, la qualité de l'information est relativement plus élevée puisque le contenu est davantage contrôlé.

Dans le cadre de la correction d'anomalies ayant un fort impact opérationnel et juridique, il est préférable de privilégier l'approche « top-down », par laquelle le contrôle de l'information est plus important.

Dans tous les cas, l'organisation doit permettre d'éviter un écueil majeur de ce type de système, à savoir le fossé qui peut se creuser entre la quantité d'informations présente dans le système et qui doit être mise à jour et la quantité d'informations que les administrateurs du contenu peuvent gérer (Figure 12). Une différence croissante ne permet plus à l'organisation mise en place de gérer la qualité de l'information présente dans le système, entraînant *de facto* une méfiance des utilisateurs et une inadéquation progressive de l'information par rapport à leurs besoins.

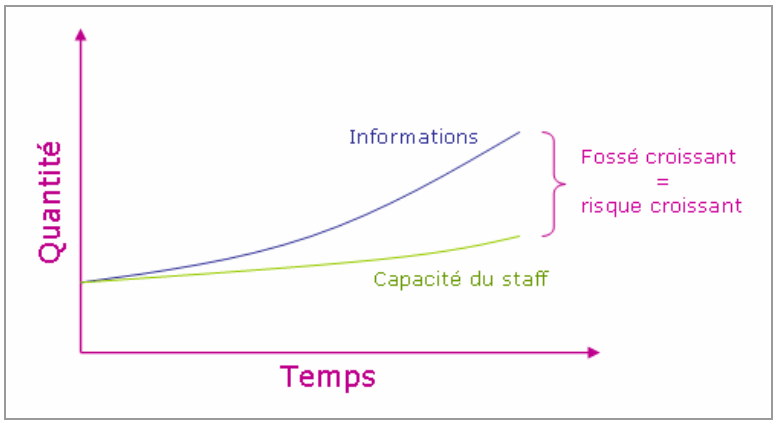


Figure 12 : Risque potentiel entre les informations existantes dans le système et les ressources humaines nécessaires pour les gérer

Une manière simple d'évaluer la qualité de l'information (principalement la partie consacrée au traitement de l'anomalie) consiste à implémenter un système de feedback binaire à l'instar de ce qui se fait au niveau de l'aide en ligne via une simple question telle que « ces informations vous ont-elles aidé ? ». Si un nombre important ou croissant d'utilisateurs répondent par la négative, il est clair que l'information ne répond pas (ou plus) aux besoins des utilisateurs. Éventuellement une enquête plus poussée permet de mettre en avant une ou plusieurs catégories de personnes pour lesquelles l'information n'est pas adaptée.

Technologies

Il existe différents types de systèmes de gestion des connaissances.

Dans le cadre d'une documentation des anomalies et des procédures de correction, l'information peut être structurée. Pour chaque fiche d'anomalie, les informations sont identiques. Certains champs sont formalisés tandis que d'autres ne le sont pas (champ libre). Chaque fiche doit respecter, de manière contraignante, un template commun imposé (au risque que ce ne soit pas le cas). La création et la validation de l'information sont contrôlées et organisées.

La recherche et la gestion de ces informations sont difficilement automatisables puisque la plupart des anomalies concernées ici demande une interprétation humaine. Cependant, même si l'aide reste manuelle, chaque anomalie porte un identifiant unique qu'il est facile d'utiliser comme critère de recherche dans l'application ou comme paramètre en vue d'une intégration au système de traitement des anomalies.

Eu égard à ces spécifications, deux technologies peuvent être envisagées : un *(Web) Content Management System* (WCMS) et un Wiki.

Un WCMS permet de définir un template commun à toutes les fiches, de rendre certains champs obligatoires, de définir des listes prédéfinies pour faciliter la saisie d'une valeur dans un champ. Ces outils intègrent « nativement » des workflows simples de validation et permettent de gérer le multilinguisme selon la recommandation évoquée précédemment.

Par nature, le wiki est un système moins contrôlé et moins formalisé. Il est difficile d'y contraindre un canevas commun pour les fiches. Un workflow de base n'est généralement pas intégré dans le système et la gestion du multilinguisme n'est que rarement prévue (il faut y créer différentes instances de la même fiche).

En conclusion, le système de gestion de contenu est davantage approprié au contexte : approche top-down, contrôle centralisé via un workflow de validation, système de templates facilitant l'encodage de l'information et son affichage homogène.

Falco - Système de know-how de correction des anomalies DmfA

Introduction

Chaque trimestre, des milliers de déclarations sociales (DmfA) sont transmises par les employeurs à l'Office National de Sécurité Sociale. Ces déclarations sont utilisées par l'ONSS(-APL) pour percevoir les cotisations sociales et, via la Banque-Carrefour de la Sécurité Sociale, par les institutions de sécurité sociale pour attribuer différents droits sociaux.

Lorsque ces déclarations sont transmises à l'ONSS(-APL), elles font l'objet de contrôles pour en vérifier la conformité. Ces contrôles génèrent des anomalies qui doivent être traitées en vue d'assurer les droits sociaux des travailleurs et une perception aussi correcte que possible des cotisations sociales (Figure 13).

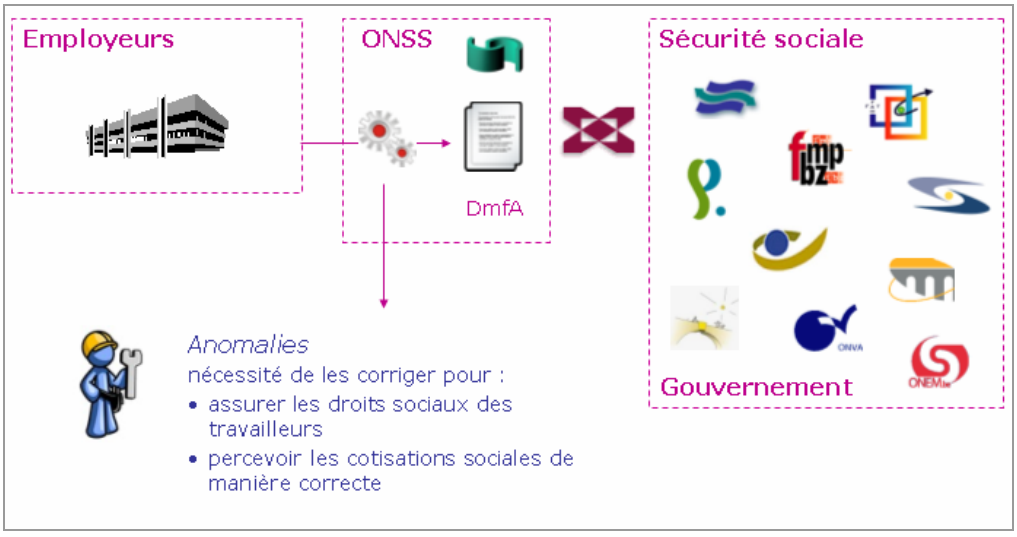


Figure 13 : Vue globale des flux d'informations DmfA et contrôles des données

Une partie de ces corrections est effectuée par les agents du service du contrôle de l'ONSS. Pour effectuer ces corrections, les agents se basent sur leur expérience et sur les documentations existantes mais éparses. Une documentation sur le traitement des anomalies existait sous forme papier, mais certaines informations n'étaient plus à jour et n'étaient pas identiques pour les deux rôles linguistiques. Le souhait était donc de disposer d'une application documentaire pour y homogénéiser les informations, y partager leurs connaissances sur la manière de corriger les anomalies, dont certaines sont complexes, et y rassembler les documentations existantes.

Prototype

Dans ce cadre, un prototype d'application documentaire de gestion des connaissances, baptisé Falco (abréviation de Fautes et Anomalies Lire et Corriger), a été réalisé à l'aide du logiciel Hippo CMS pour le service du contrôle de l'ONSS avec les objectifs suivants :

- centraliser des informations en vue de gérer les anomalies DmfA ;
- les tenir à jour via un circuit de validation ;
- favoriser un traitement plus homogène des anomalies, en vue d'améliorer la qualité et éviter les problèmes d'interprétation et leurs conséquences sur la gestion des dossiers ;
- partager des connaissances sur la manière de corriger les DmfA, ce qui s'avère particulièrement utile pour les nouveaux agents ;
- mettre à la disposition des agents une information équivalente dans les deux langues ;
- humaniser le travail des agents.

Conformément aux besoins expliqués précédemment, Falco comprend à la fois des informations descriptives et des informations sur la manière de traiter les anomalies. Chaque « fiche anomalie » contient cinq catégories d'informations (Figure 14) :

1. Définition technique : code anomalie, intitulé de la zone ou du bloc sur lequel porte l'anomalie, gravité et période de validité. Ces diverses informations sont reprises des Glossaires DmfA.
2. Traitement de l'anomalie : indication si l'anomalie doit être résolue ou non[69], indication de la nécessité de prendre contact avec l'employeur, explication de l'anomalie, ses différentes causes possibles et les méthodes de traitement. Cette catégorie d'information représente le cœur de l'application.
3. Informations complémentaires : en cas de nécessité, lien vers les Instructions aux employeurs[70] et les Glossaires DmfA.
4. Secteurs compétents : les anomalies DmfA peuvent être corrigées par trois organismes de sécurité sociale (ONSS - ONVA - Sigedis) selon une répartition des compétences existantes.
5. Impacts pour l'assuré : ces informations indiquent l'importance d'une zone pour l'octroi des droits sociaux aux travailleurs. Les données DmfA étant utilisées par de nombreux secteurs de la sécurité sociale, une anomalie non corrigée peut avoir des impacts importants sur le calcul et l'octroi de ces droits. Cette information vise essentiellement à montrer aux agents l'intérêt et l'utilité des corrections.

[69] Pour diverses raisons, certaines anomalies détectées ne doivent pas être corrigées.

[70] Dans le cadre de la DmfA, les « Instructions aux employeurs » contiennent des explications précisant qui est redevable de cotisations de sécurité sociale et dans quelle mesure. S'y trouvent également des explications relatives aux obligations des employeurs envers l'ONSS et envers les différents régimes de la sécurité sociale.

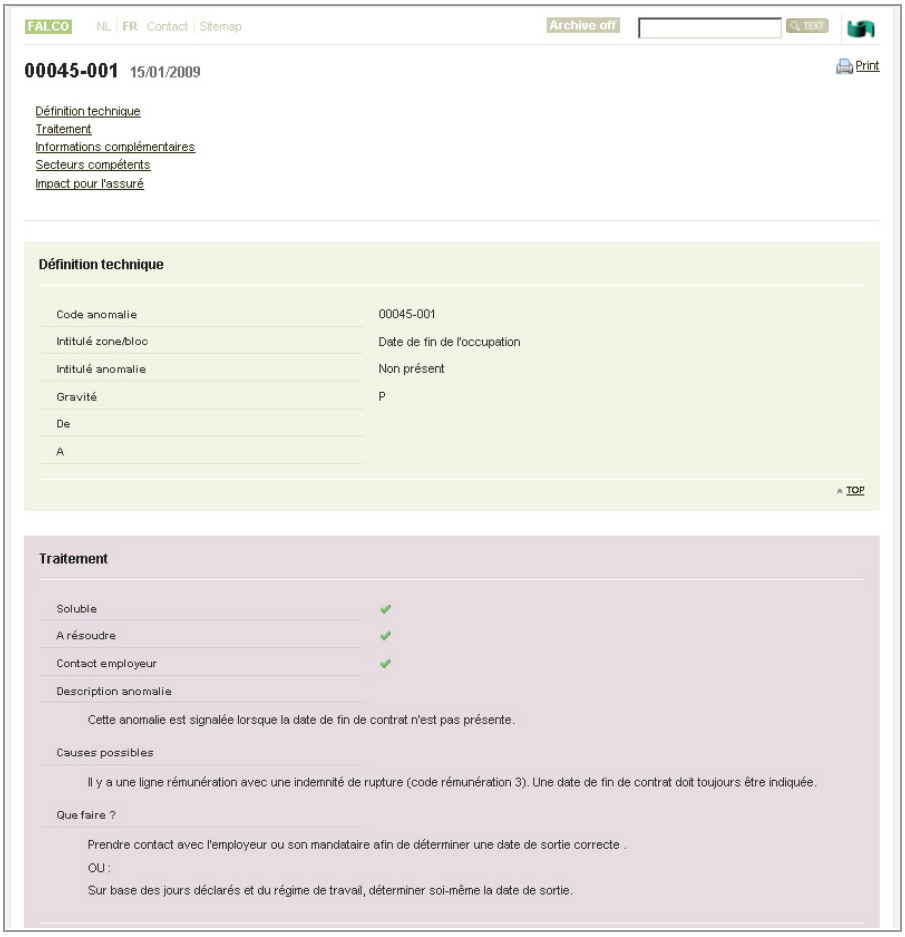


Figure 14 : Exemple de fiche anomalie du système Falco

L'édition des fiches consiste à compléter un formulaire préformaté à l'aide d'une interface d'édition (Figure 15). Pour les champs libres, des consignes de rédaction ont été mises en place pour assurer un maximum de cohérence entre les deux versions linguistiques de la fiche (fond et forme).

Chaque fiche est bilingue. Les champs complétés à l'aide d'une liste de valeurs prédéfinies sont automatiquement traduits pour alléger le travail des administrateurs et éviter les erreurs de traduction.

Les fiches sont créées et mises à jour par un groupe d'experts de la Direction des applications particulières au sein du service du contrôle (approche centralisée « top-down ») sans aucune intervention d'une personne technique.

Un circuit de validation est prévu. Chaque nouvelle fiche ou adaptation est validée par un autre membre du groupe. Actuellement, le workflow ne prévoit pas encore de circuit pour la traduction (celle-ci est effectuée au sein du groupe d'experts).

Un formulaire est disponible pour que les visiteurs puissent contacter les administrateurs du contenu.

Le système prévoit deux types de recherche : une recherche sur la base du numéro de l'anomalie et une recherche full-text (type Google). Ces deux recherches sont fournies à l'utilisateur dès la page d'accueil du site de manière à ce que le système soit utilisable le plus rapidement possible et restent accessibles directement quelle que soit la page du site.

Le système a été réalisé à l'aide d'un système de WCM. Il est donc accessible via un navigateur web standard.

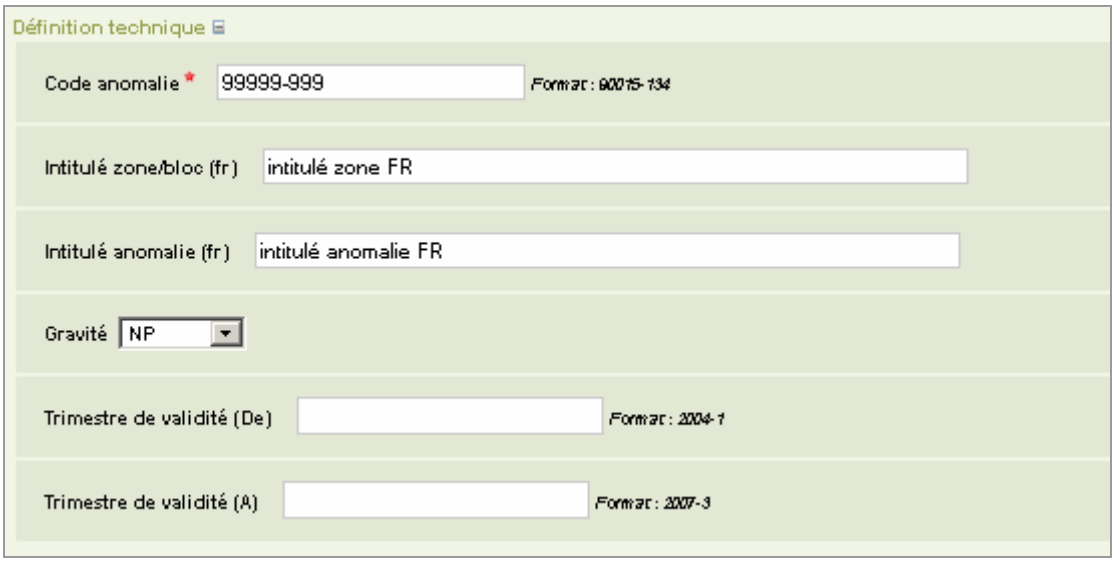


Figure 15 : Interface d'édition d'une fiche anomalie au sein de Falco

Extensions prévues

Suite à la mise en place du prototype et aux tests qui se sont déroulés pendant plusieurs mois, le système a largement prouvé son utilité et est pleinement accepté par les utilisateurs. Diverses évolutions et extensions sont envisagées :

- Intégration de cette documentation avec Odysseus, l'outil de traitement des anomalies, afin d'augmenter le confort d'utilisation pour les agents en leur donnant accès à l'ensemble des informations nécessaires en limitant les manipulations à effectuer (changement d'application, d'environnement, etc.). Cette intégration peut être minimale : les URL générées par le CMS sont standardisées et mentionnent l'identifiant de l'anomalie. Au sein du système de traitement des anomalies, un bouton situé à côté des anomalies permettrait d'afficher directement la fiche correspondante en récupérant comme paramètres l'identifiant de l'anomalie et la langue de l'utilisateur.
- Automatisation accrue afin d'alléger la charge de travail des administrateurs et en même temps d'éviter les erreurs d'encodage via un « import » automatique des modifications effectuées dans les Glossaires DmfA.
- Extension de Falco à d'autres domaines d'application et d'autres services au sein de l'ONSS, par la création de nouveaux modèles de fiche et l'adaptation de l'organisation. Ces domaines et services présentent des caractéristiques similaires telles que la gestion d'informations (semi-)structurées, la gestion d'anomalies, des instructions sur la manière de traiter ces anomalies ou certains types de dossiers… À titre d'information, nous pouvons citer les systèmes Dimona, Limosa[71] et les Déclarations de risques sociaux (DRS) ainsi que le service de la perception et celui du recouvrement.
- Intégration des documents à envoyer à l'employeur pour lui demander des justificatifs ou des informations complémentaires. Pour faciliter le

[71] La déclaration Limosa, du nom d'un oiseau migrateur, permet à tout travailleur salarié, indépendant ou stagiaire étranger qui vient travailler temporairement en Belgique de le communiquer à la sécurité sociale belge.

travail des agents, les formulaires pourraient être attachés à la fiche anomalie. Une partie des champs pourraient être complétés automatiquement via des paramètres récupérables dans l'application de traitement des anomalies.

Recommandations pour la mise en œuvre

Recommandations générales

- Opter pour la simplicité. L'expérience que nous avons menée au sein du service du contrôle de l'ONSS a montré que l'acceptation et l'utilisation du système par les administrateurs du contenu autant que par les utilisateurs reposent essentiellement sur la simplicité de l'application. L'information y est présentée de manière claire et orientée vers son utilisation pratique. Le moteur de recherche permet rapidement, sur la base de l'identifiant des anomalies, de trouver l'information nécessaire au traitement de l'anomalie.
- Se positionner du côté de l'utilisateur pour structurer le contenu. Celui-ci doit trouver le plus facilement et le plus aisément possible l'information qu'il recherche. Elle doit être exploitable et claire afin que l'utilisateur comprenne clairement ce qu'il doit faire pour traiter l'anomalie.
- Par ailleurs, l'utilisateur est fortement en demande de connaître l'utilité des corrections qu'il effectue. Cette information doit donc être présente dans la mesure du possible bien qu'elle ne soit pas toujours aisée à créer (sa création peut nécessiter des ressources importantes). Si nécessaire, cette information peut être créée progressivement en commençant par les anomalies les plus courantes.
- Accorder une attention particulière à la rédaction et à l'encodage des informations, principalement pour les champs libres, en vue de garder une cohérence de fond et de forme entre les différentes versions linguistiques. À cet égard, il est préférable que le workflow de traduction soit postérieur au workflow de validation. Des consignes de rédaction doivent absolument être fixées au sein de l'équipe de rédaction et transmises aux personnes en charge de la traduction.
- Veiller à la qualité de l'information contenue dans le système eu égard aux besoins des utilisateurs. Nous avons vu qu'un simple système de feedback binaire permet déjà d'y veiller.
- Intégrer, même de manière minimale, le système de traitement des anomalies et la documentation nécessaire pour effectuer ce traitement afin de faciliter le travail des utilisateurs et permettre une utilisation plus confortable de l'application.

Gestion du projet - Acteurs et répartition des rôles

Différents acteurs doivent être réunis (Figure 16).

Dès le départ, il est indispensable de constituer un groupe de travail de six-huit « utilisateurs » (gestionnaires de l'information et utilisateurs) aux profils variés (des personnes ayant une expérience forte du domaine et des personnes arrivées plus récemment, des personnes ayant une affinité avec les technologies informatiques et d'autres moins, etc.).

Ce groupe de travail est fondamental car c'est avec lui que l'analyse des besoins (informationnels et fonctionnels) sera effectuée avec l'aide du chef de projet.

Dans un second temps, si nécessaire, un échantillon plus large est constitué (utilisateurs pilotes) en vue d'effectuer un test plus global et plus étendu dans le temps. Ces utilisateurs doivent eux aussi être représentatifs de l'ensemble des utilisateurs et donc avoir un profil varié.

Le chef de projet gère quant à lui le déroulement de la mission et tous les contacts avec le groupe de travail, les utilisateurs, l'équipe de développement. Il peut également assurer les formations.

L'équipe de développement est chargée de développer un prototype qui sera rapidement soumis au groupe de travail. Elle développe ensuite l'application et en assure le suivi technique. Elle prodigue également des recommandations au chef de projet sur les différentes manières de mettre en œuvre certaines fonctionnalités et des conséquences y afférentes. Elle fournit également des informations sur le temps et les budgets nécessaires pour le développement afin que le chef de projet puisse, en concertation avec le groupe de travail, opérer les arbitrages coûts-bénéfices qui s'imposent.

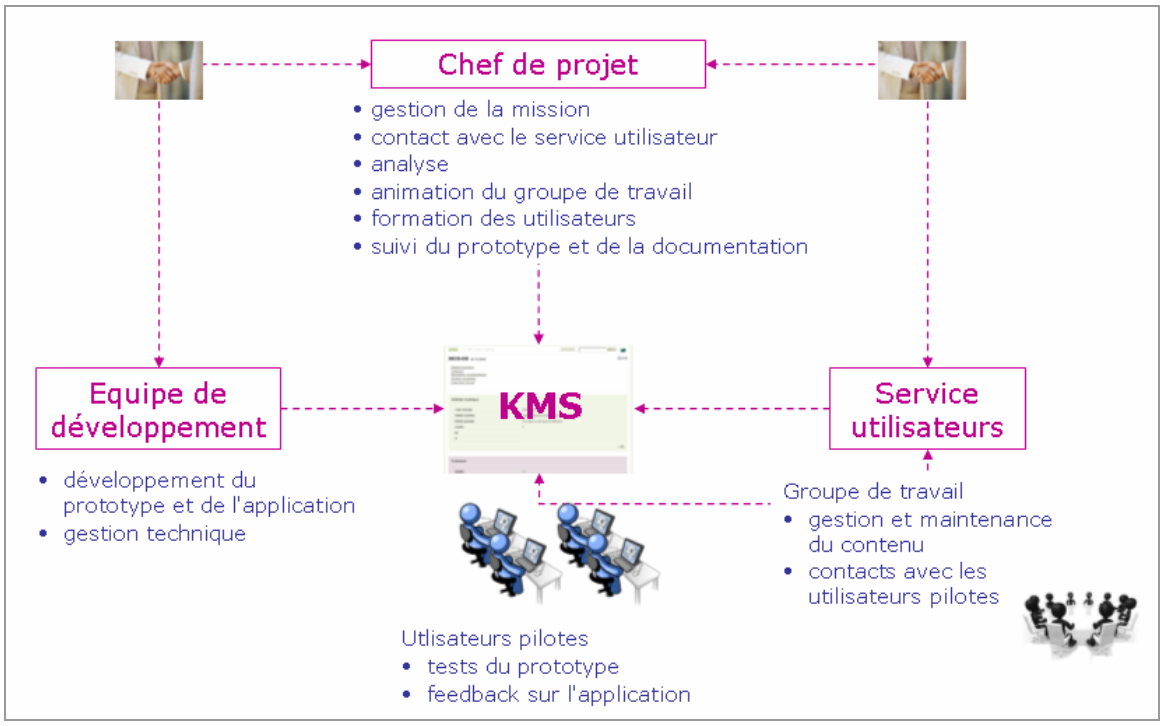


Figure 16 : Acteurs du projet et répartition des rôles

Étapes pour la mise en œuvre

Comme tout projet, la mise en place d'un tel système passe par diverses étapes (Figure 17) : analyse, choix du logiciel, conception, validation, mise en place/test/adaptation et mise en production.

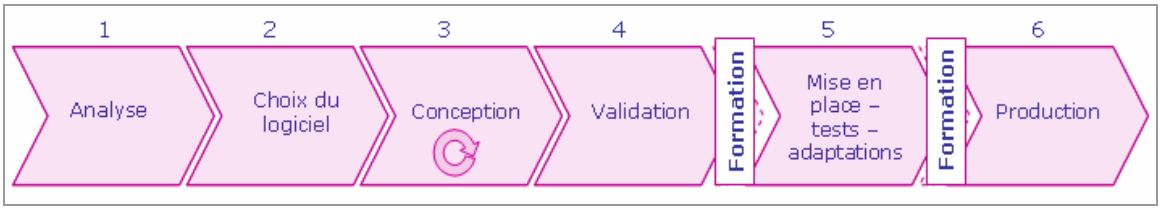


Figure 17 : Étapes de la mise en place d'un système de gestion des connaissances pour la correction des anomalies.

Après avoir rassemblé les acteurs, la première étape consiste à analyser les besoins, aussi bien informationnels, fonctionnels qu'organisationnels. Cette analyse se déroule en plusieurs réunions avec le groupe de travail (et éventuellement des interviews plus personnalisées), en commençant par les besoins informationnels. Pour celles-ci, la démarche consiste à identifier :

- les informations descriptives nécessaires au traitement (par exemple identifiant et intitulé de l'anomalie) ;
- les informations qui fournissent des indications sur la manière de résoudre l'anomalie ;
- les informations qui permettent de faire comprendre aux utilisateurs l'intérêt de leur correction.

Il faut également déterminer l'ordre d'apparition des informations à l'écran et leur mise en pages ainsi que la traduction des différents champs.

La seconde étape consiste à définir les besoins fonctionnels en se basant sur ceux que nous avons expliqués ci-dessus. Il s'agit également à cette étape de définir les différents rôles qui interviendront dans le workflow et les droits qui y sont associés (édition, publication, suppression).

Enfin, il faut déterminer l'organisation à mettre en place pour assurer la qualité de l'information contenue dans le système.

Chaque réunion doit faire l'objet d'un procès-verbal validé lors de la réunion suivante. Le résultat final, comprenant l'analyse des besoins, la fiche type des anomalies et la traduction des champs, doit être validé par le groupe de travail.

L'équipe de développement réalise une première maquette qui sera soumise au groupe de travail. Celle-ci peut être réalisée parallèlement à l'analyse, sur la base des premières discussions, afin de permettre aux membres du groupe de travail de réagir et de préciser leurs besoins et souhaits.

Une phase facultative peut être menée en vue de tester l'acceptation de l'application par les utilisateurs. Un risque de rejet existe, principalement dans un environnement où les tâches sont répétitives, complexes et fastidieuses. Pour ce faire, un prototype incluant les informations souhaitées et les fonctionnalités principales est réalisé et soumis à un groupe d'utilisateurs représentatifs. Si l'expérience s'avère concluante, la phase de développement est poursuivie en vue de finaliser l'application.

En parallèle, les gestionnaires de l'information prépare le contenu sur la base des décisions prises lors de l'analyse afin de pouvoir l'introduire dès la mise à disposition de l'application et avant que celle-ci ne soit mise à la disposition des utilisateurs.

Enfin, il nous semble important de mentionner qu'une bonne communication est nécessaire pour assurer la réussite du projet, notamment en raison des réticences que ce type d'application est à même de susciter.

2.4.3. Organisation

La création et la gestion de la documentation opérationnelle du système d'information nécessitent la mise en place d'une organisation et d'acteurs dédiés à cette fin (Figure 18) :

- Un comité de pilotage (« Council » sur la figure ci-dessous) permet de réunir des représentants des différentes parties concernées[72]. Les membres de ce conseil doivent avoir des profils aussi bien métier que IT. Ce conseil est chargé de définir :
  - les données ;
  - les anomalies ;
  - les règles métier qui seront appliquées sur les données, aussi bien à travers des systèmes de business rules que des outils de Data Quality ;
  - la manière dont les anomalies devront être traitées ;
  - la documentation associée à ces éléments.
- Un service de gestion de l'information chargé de coordonner le travail du conseil au niveau des aspects documentaires et de mettre la documentation à jour sur cette base. Cette documentation comprend des informations descriptives et des informations « correctives » (à savoir des informations sur la manière dont les anomalies doivent être traitées). Dans certains cas, cette documentation peut reposer sur une base commune, héritée et adaptée pour des bases spécifiques.

Et enfin, naturellement, des gestionnaires de bases de données, chargés de conseiller et d'appliquer les décisions du conseil, au niveau des données, des anomalies, des règles de validation des données et de Data Quality Tools.

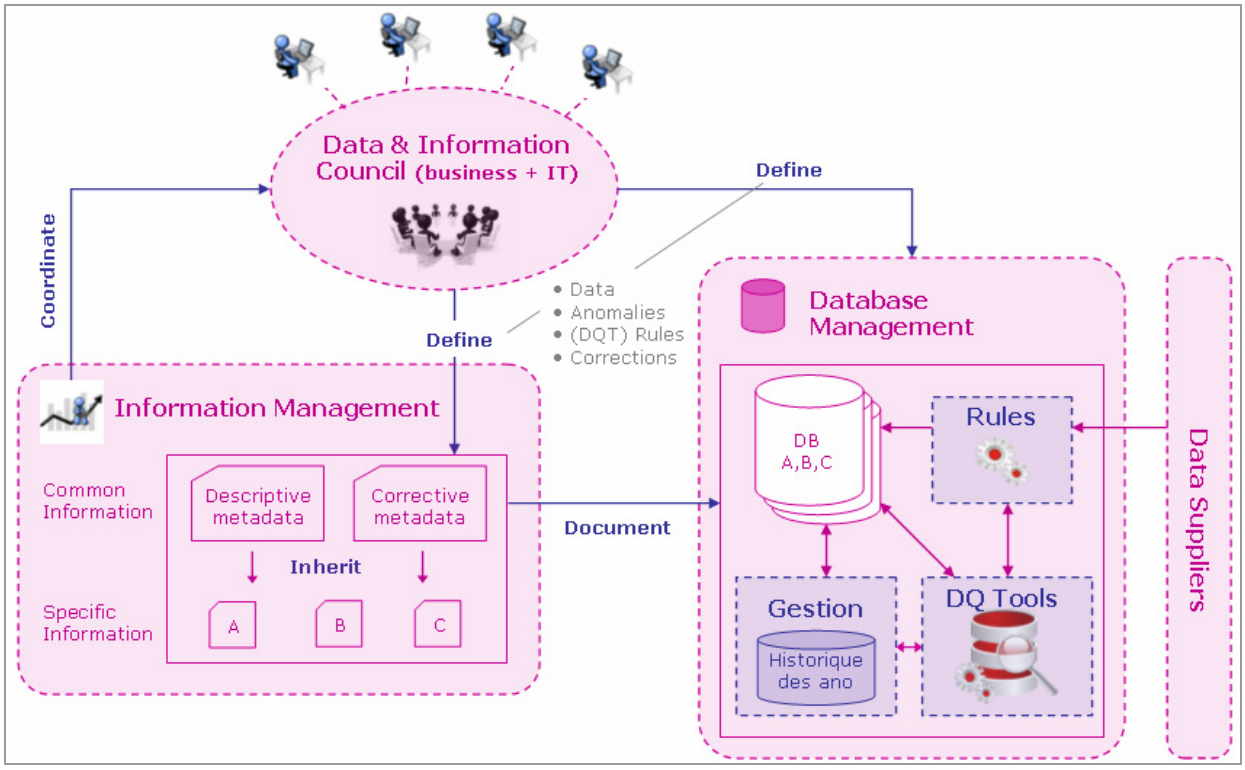


Figure 18 : Organisation pour la gestion de la documentation

[72] Pour de plus amples informations sur la gouvernance des données, voir TRIGAUX J.-C., *op. cit.*

3. Gestion des anomalies et Data Quality Tools

Ce chapitre a pour objet de montrer en quoi les outils de data quality peuvent aider à gérer les anomalies. Grâce à notre expérience, acquise dans des projets concrets, nous sommes en mesure de proposer des exemples d'application sur des données réelles dans ce contexte.

Les outils de qualité des données autorisent un traitement efficace de plusieurs types importants d'anomalies. Sans outils, cette opération serait en effet ardue, voire impossible. Dans ce chapitre, nous indiquerons quelles fonctionnalités des outils conviennent à quelles tâches dans la gestion des anomalies (détection et traitement) et expliquerons comment elles peuvent s'intégrer dans les processus globaux de gestion des anomalies.

Dans une certaine mesure, l'utilisation des data quality tools impose de suivre une méthodologie, où le *profiling*, la *standardisation*, le *matching* et le *monitoring* occupent une place bien déterminée dans le cycle continu d'amélioration de la qualité des données, introduit dans une autre étude de la section Recherches[73]. Les fonctionnalités des outils peuvent aussi être classées suivant ces axes principaux.

Par conséquent, les trois premiers paragraphes de ce chapitre s'intitulent successivement « Data Profiling », « Standardisation des données » et « Data Matching ». Chacun de ces paragraphes débute par une définition générale du concept, ainsi que de la place qu'y occupent les data quality tools, et se clôture par des conseils pratiques destinés aux analystes et aux développeurs. Toutes les fonctionnalités sont illustrées à l'aide d'exemples anonymisés, issus de projets concrets. Lorsqu'il est pertinent, le lien avec la gestion historique des anomalies (2.1) est établi.

Dans cette étude, le monitoring est considéré au-delà du seul cadre des data quality tools et est traité dans (2.1). Le Data Monitoring assisté par les outils de data quality a déjà été abordé de manière générale dans l'étude précitée.

[73] BONTEMPS Y., BOYDENS I., VAN DROMME D., *Data Quality : tools*, Deliverable, 2007/trim3/02, Smals, Section Recherches, Bruxelles, 2007.

3.1. Data Profiling

3.1.1. Définition

Le Data Profiling est l'utilisation de techniques analytiques dans le but de découvrir la structure, le contenu et la qualité **réels** d'une collection de données[74]. Le Data Profiling se distingue des techniques d'analyse de données destinées à obtenir des informations métier à partir des données. En effet, le Data Profiling est utilisé pour dégager des informations factuelles à propos des données.

À cette fin, le Data Profiling emploie comme input tant les données elles-mêmes que toutes les métadonnées connues correspondant aux données. L'output est constitué de métadonnées formellement précises (corrigées et exhaustives) et d'informations supplémentaires sur les données imprécises[75] (Figure 19).

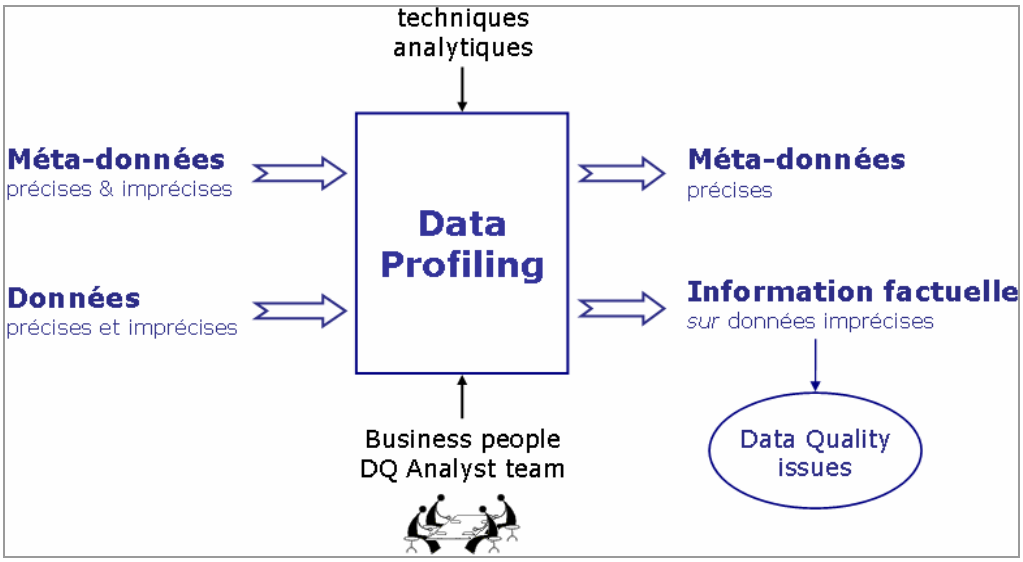


Figure 19 : Data Profiling

Le Data Profiling est un **processus** qui débute par une phase de découverte, au cours de laquelle sont dressées les caractéristiques des données et leur cohésion structurelle.

S'ensuit une phase de validation où les spécialistes métier indiquent, suivant les critères « Fitness for Use », quelles caractéristiques sont réellement des règles auxquelles devraient satisfaire les données et leur cohésion structurelle. Une comparaison est alors possible avec la documentation et les métadonnées existantes, qui peuvent ainsi être complétées et corrigées.

Une fois la définition des données réelles, « fit for use » réalisée, les règles peuvent être implémentées sous forme de contrôles pour détecter de façon automatisée les données qui s'opposent aux métadonnées, c'est-à-dire détecter les anomalies.

Ce n'est donc qu'une fois que les problèmes liés à la qualité des données ont traversé une phase de validation qu'ils peuvent formellement être confirmés comme étant des anomalies. Ceci est conforme à la définition des anomalies (1.3.3). Tous les problèmes de qualité des données ne débouchent pas sur des anomalies formelles.

[74] OLSON J., *Data Quality : the Accuracy Dimension*, Elsevier, Burlington, 2002.

[75] Dans le contexte de la qualité des données, « **précis** » veut dire « adéquat par rapport à l'utilisation envisagée » (cf. la notion « Fitness for Use » - cf. 1.1).

Exemple : si l'analyse des données dans la phase de découverte révèle que deux encodages différents apparaissent dans une zone de données réservée à des codes pays (d'une part, la partie composée de deux lettres de la norme ISO 3166[76], comme BE, FR, DE… et, d'autre part, la partie numérique, soit respectivement 150, 250, 276…), il s'agira plutôt d'un problème de Data Governance ou de Master Data Management. Le business doit clairement décider, pour toutes les applications qui alimentent la base de données analysée, quel encodage sera autorisé dorénavant. Ensuite, une anomalie formelle pourra être implémentée à titre de contrôle sur le domaine de définition pour le champ dédié au code pays.

Pour plus d'informations sur la méthodologie qui peut être suivie pour exécuter un Data Profiling complet, voir l'étude de la section Recherches précitée et son inspiration, l'oeuvre de référence dans ce domaine[77].

Implications

Comme seuls les utilisateurs métier sont capables d'évaluer l'adéquation aux usages, un projet de Data Profiling doit être réalisé par une équipe constituée tant d'analystes de données que de spécialistes métier (3.5).

Logiquement, le Data Profiling est le début du cycle d'amélioration de la qualité et **devrait se trouver à la base de tout développement** de nouvelles applications qui se serviront des données, **de toute migration** des données vers d'autres applications, voire **de toute nouvelle exploitation** des données.

Limitations

Le Data Profiling ne garantit pas de trouver toutes les données incorrectes, mais permet uniquement de trouver les règles et désignations de règles qui ne sont pas respectées. Parfois, il est possible d'indiquer clairement quelles valeurs de quelles zones de données sont incorrectes. Parfois, on sait seulement qu'une combinaison de données est invalide, sans connaître les enregistrements et/ou valeurs concrets incriminés. En outre, il se peut que des données qui respectent toutes les règles soient quand même incorrectes (1.3.3).

3.1.2. Data Profiling à l'aide d'outils de qualité des données

Les data quality tools procurent une interface utilisateur graphique dans laquelle, partant des données complètes réelles, il est dressé un aperçu efficace de bon nombre des problèmes de qualité des données qu'il faut spécifiquement examiner dans le cadre d'un Data Profiling. Cet aperçu offre des résumés et des chiffres pratiques, mais aussi la possibilité d'effectuer des *drill-downs* vers les enregistrements et données réels sous-jacents.

Sur la base des fonctionnalités de Data Profiling propres aux outils de data quality et à l'appui d'exemples concrets sur des données réelles, ce paragraphe indique les types de problèmes de qualité des données (et donc, par extension, anomalies) qui peuvent être détectés et traités efficacement.

Analyse des données par colonne (Column Property Analysis)

Tout d'abord, les bases de données sont analysées par colonne (champ, attribut), pour chaque table (entité). Les valeurs non conformes aux métadonnées (définitions, documentation, conditions, restrictions) pouvant être définies au niveau de l'attribut même sont dites *invalides*.

[76] http://www.iso.org/iso/country_codes.htm.

[77] OLSON J., *Data Quality : the Accuracy Dimension*, Elsevier, Burlington, 2002.

Lors du chargement de chaque entité dans l'environnement des data quality tools, des métadonnées spécifiques sont ajoutées à tous les attributs, qui permettent très facilement à un analyste des données de :

- vérifier dans quelle mesure toutes les valeurs présentes dans les données correspondent à la documentation pour chaque champ ;
- détecter toutes les anomalies qui peuvent être définies au niveau d'un champ.

Autrement dit, il devient très facile de détecter toutes les valeurs *invalides*.

La Figure 20 montre quelles informations sont résumées pour chaque champ dans les outils de qualité des données et de Data Profiling :

- type (*Type*, *Inferred Type*)
- valeurs présentes et distribution (*Unique Values*, *Distribution*)
- modèles (*Patterns*, *Masks*)
- valeur minimale et valeur maximale (*Min*, *Max*)
- longueur de la valeur la plus courte et de la valeur la plus longue (*Min Len*, *Max Len*)
- présence de *null values*

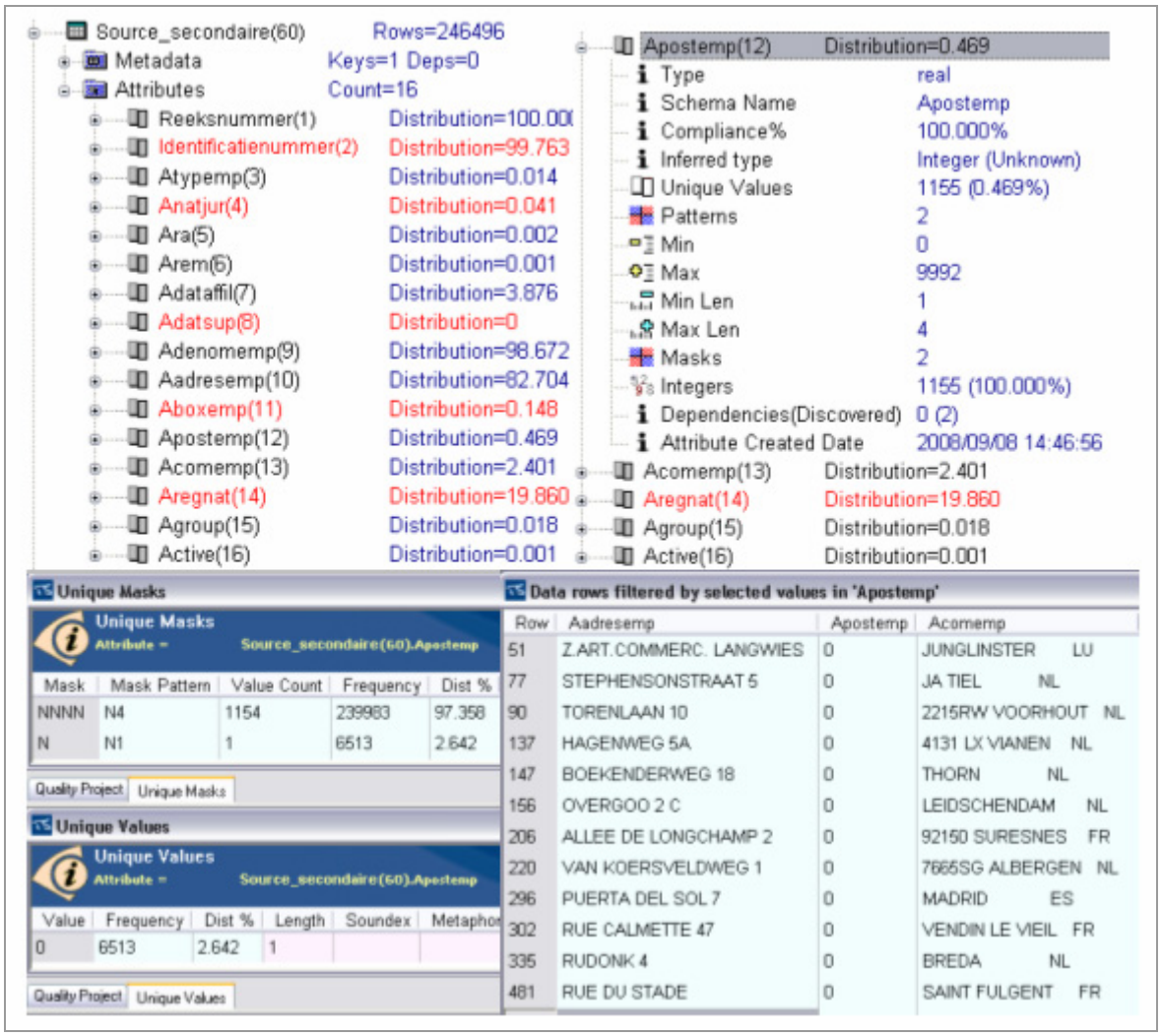


Figure 20 : Exemple de Data Profiling

En haut : tous les attributs d'une entité ont été analysés ; les informations sont résumées et disponibles pour un drill-down à l'aide de métadonnées (à droite : un analyste examine les détails de l'attribut Apostemp : le fait que deux « masks » soient présents attire l'attention).

En bas : drill-down vers les modèles et valeurs présents (à gauche : le modèle « NNNN » est normal pour la Belgique, mais pas « N » ; il s'agit vraisemblablement d'une même valeur « 0 » qui apparaît dans 6513 enregistrements). À droite : drill-down vers les enregistrements afférents ; il s'agit vraisemblablement d'adresses étrangères.

L'exemple ci-dessus (Figure 20) montre comment, dans un projet concret, un problème a été découvert dans un champ réservé au code postal (*Apostemp*). Un *drill-down* via les modèles présents (*Masks*) montre que la valeur incorrecte « 0 » apparaît, plus particulièrement (grâce à un *drill-down* approfondi) pour les adresses étrangères. Celles-ci présentent souvent, pour leur code postal, un autre modèle que quatre caractères numériques. Une analyse plus approfondie a révélé que la base de données a été conçue initialement pour autoriser quatre caractères numériques seulement. Les codes postaux étrangers non conformes apparaissent *overloaded* dans un autre champ (*Acomemp*, prévu pour le nom de la commune).

Drill-down

Les aperçus élaborés à partir de tout élément d'information offrent la possibilité de naviguer vers toutes les valeurs concernées et la distribution de leur apparition, puis vers une liste de tous les enregistrements dans lesquels apparaissent les valeurs sélectionnées.

Clés primaires

Si un attribut a été désigné comme clé primaire, toutes les valeurs qui apparaissent doivent être uniques et il ne peut y avoir de *null values*. Les outils de qualité des données permettent de contrôler cela très facilement grâce aux métadonnées *Unique Values* et *Null Count*, établies de manière standard par attribut. En outre, durant le chargement des entités, certains outils de qualité des données vérifieront automatiquement si chaque attribut convient comme clé primaire.

La simplicité avec laquelle ceci peut être analysé dans l'environnement graphique d'un data quality tools est illustrée ci-dessous (Figure 21 et Figure 22).

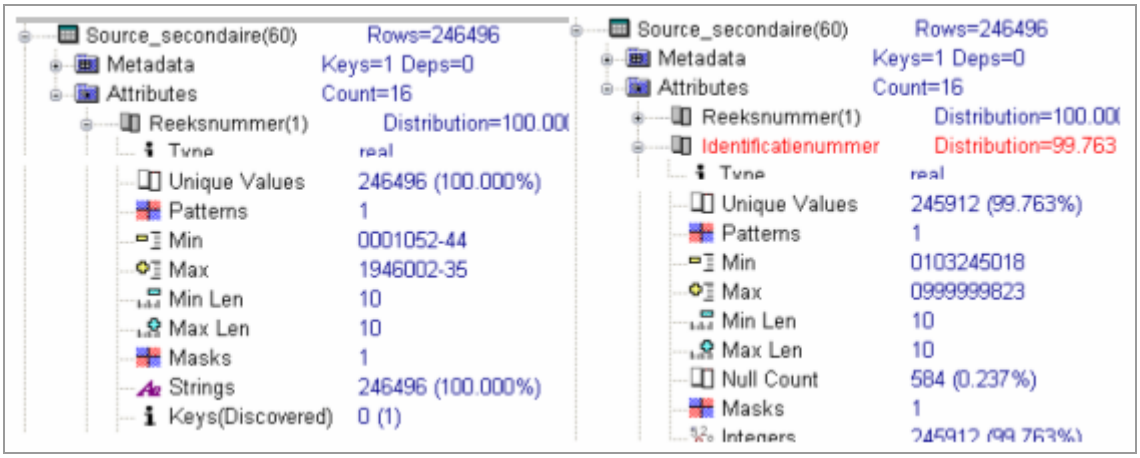


Figure 21 : Analyse des clés primaires.

À l'aide des métadonnées affichées, on voit aisément dans quels attributs sont respectées les caractéristiques de clés primaires (à gauche : Reeksnummer affiche une Distribution de 100 % dans les Unique Values et peut donc être une

clé primaire ; à droite : Identificatienummer ne présente pas une distribution parfaite et affiche un Null Count de 584). Notez que dans les métadonnées de Reeksnummer (Keys(Discovered)), l'outil de qualité des données affirme déjà reconnaître cet attribut comme clé probable.

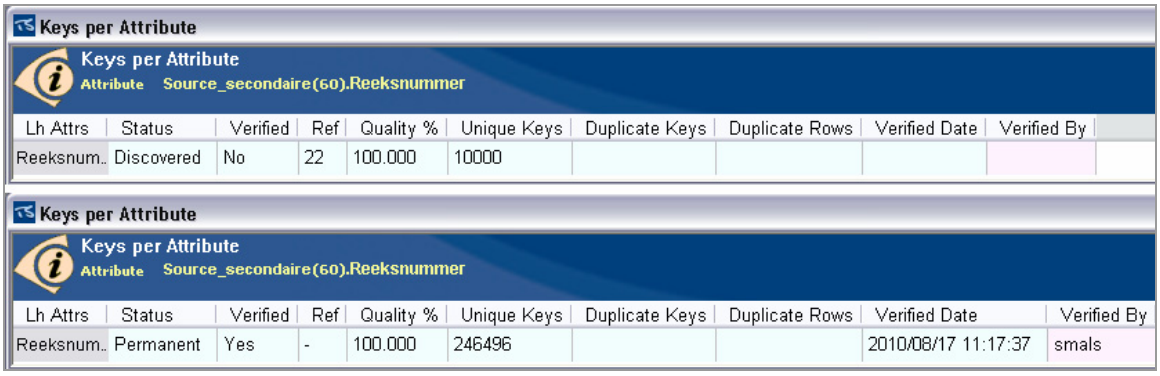


Figure 22 : Analyse des clés primaires.

Sur la base d'un échantillon de 10.000 enregistrements, l'outil de qualité des données indique que Reeksnummer est une clé probable (statut : Discovered) ; à tout moment, un analyste des données peut vérifier cela entièrement par la suite (statut : Permanent).

Une fois que toutes les entités ont été étudiées par attribut, la cohésion structurelle des données (tant au sein d'une entité qu'entre les entités, voire entre plusieurs bases de données) peut être analysée. Cet ordre dans la manière de procéder correspond à l'ordre recommandé pour l'exécution des contrôles (2.3.4).

Relations, intégrité référentielle et cohésion structurelle

La cohésion structurelle des données peut être documentée explicitement. La cohésion au sein d'une source unique de données est même généralement documentée de façon explicite. Dans le cas des systèmes DBMS relationnels par excellence, la cohésion structurelle peut explicitement être implémentée et imposée par le schéma des bases de données et les *structural contraints*[78] (ratios de cardinalité des relations binaires, *participation constraints*, *existence dependency*…). Ce n'est pas le cas avec tous les systèmes DBMS et, dans la majorité des cas s'avère-t-il, indépendamment des systèmes DBMS et des applications connexes, l'intégrité structurelle n'est *pas complètement* imposée.

Ci-après, nous abordons les aspects majeurs de la cohésion structurelle et illustrons la manière dont les outils de data quality aident à analyser celle-ci.

Clés étrangères, join analyses

Si un attribut a été désigné comme clé étrangère (selon la documentation) dans une table A, deux conditions doivent être remplies :

- l'attribut est une clé primaire dans une autre table B (cf. supra, Figure 17) ;
- la table A ne peut contenir une valeur de clé étrangère qui n'apparaît pas dans la table B.

Pour contrôler ce point à l'aide des data quality tools, les *join analyses* représentent la solution la plus commode. L'analyste désigne les deux tables entre lesquelles une relation devrait exister ainsi que les attributs respectifs qui devraient constituer la relation de clé. Pour chaque *join analysis*, l'outil de qualité

[78] ELMASRI R., NAVATHE S., *Fundamentals of Database Systems (fifth edition)*, Addison-Wesly, Boston, 2007.

des données crée un diagramme de Venn ainsi qu'une large série de *join metadata* (Figure 23 et Figure 24).

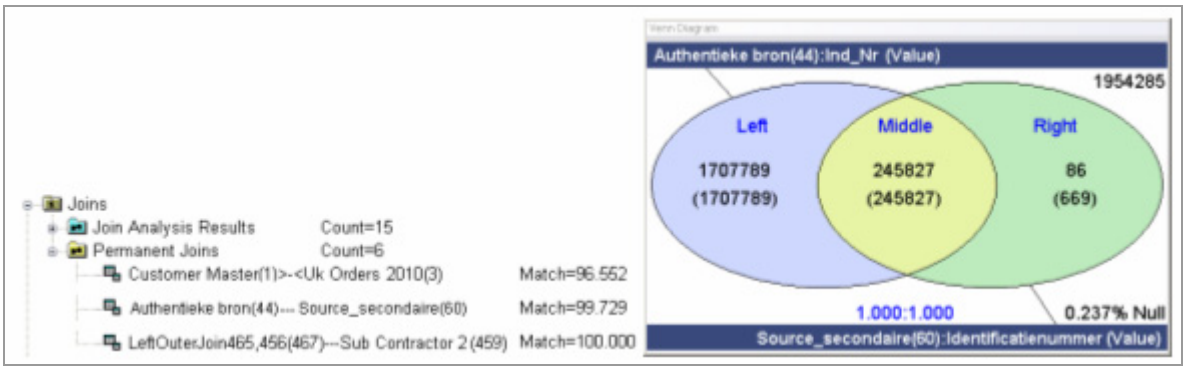


Figure 23 : Analyse des join metadata et diagramme de Venn.

À gauche : aperçu des join analyses déjà réalisées.

À droite : diagramme de Venn indiquant dans quelle mesure les données présentes respectent la relation de clé étrangère. Il s'agit ici de 246.496 numéros d'identification d'une source Source_secondaire : toutes les valeurs devraient être retrouvées comme un Ind_Nr d'une (copie de) source authentique Authentieke bron. Dans la plupart des cas, cela est correct (Middle : 245.827 enregistrements et valeurs), excepté pour une petite partie (Right : 86 valeurs différentes dans 669 enregistrements ne sont pas retrouvées dans la source authentique).

Join Metadata
Join = Authentieke bron(44):Ind_Nr --- Source_secondaire(60):Identificatienummer

Metadata	Value	Description
Matching Values	245827	The number of unique joined values
Inner joined rows	245827	The number of rows in the inner join - joining rows only
Outer joined rows	1954285	The number of rows in the outer join - joining rows and non-joining rows from both sides
Left Non-Matching Values	1707789	The number of unique values on the left-hand side that did not join
Left Non-Matching Rows	1707789	The number of rows on the left-hand side that did not join
Left outer joined rows	1953616	The number of rows in the left outer join - joining rows and non-joining left rows
Right Non-Matching Values	86	The number of unique values on the right-hand side that did not join
Right Non-Matching Rows	669	The number of rows on the right-hand side that did not join
Right outer joined rows	246496	The number of rows in the right outer join - - joining rows and non-joining right rows
Left Loaded Rows	1953616	Number of rows in the left hand source entity
Left Filter		The filter used to select the rows from the left hand entity
Left Selected Rows	1953616	The number of rows selected by the Left hand filter (or loaded rows if not filtered)
Right Loaded Rows	246496	Number of rows in the right hand source entity
Right Filter		The filter used to select the rows from the right hand entity
Right Selected Rows	246496	The number of rows selected by the right hand filter (or loaded rows if not filtered)

Figure 24 : Join metadata.

Tant depuis l'aperçu des join analyses que depuis le diagramme de Venn, des drill-downs sont possibles vers les métadonnées et les enregistrements afférents. Ici, un sous-ensemble des possibilités les plus importantes est présenté.

Cette analyse peut être réalisée tant parmi les tables d'une même base de données qu'entre les tables de deux sources de données distinctes. La seconde option est surtout importante pour les contrôles référentiels par rapport à une (copie d'une) source authentique (1.2.3). Ce type d'anomalies peut donc être détecté très facilement avec des outils de qualité des données.

Clés composées

Une clé est dite composée lorsqu'elle est formée par la combinaison de plusieurs attributs. Ceci aussi peut être détecté et vérifié par un outil de data quality (Figure 25).

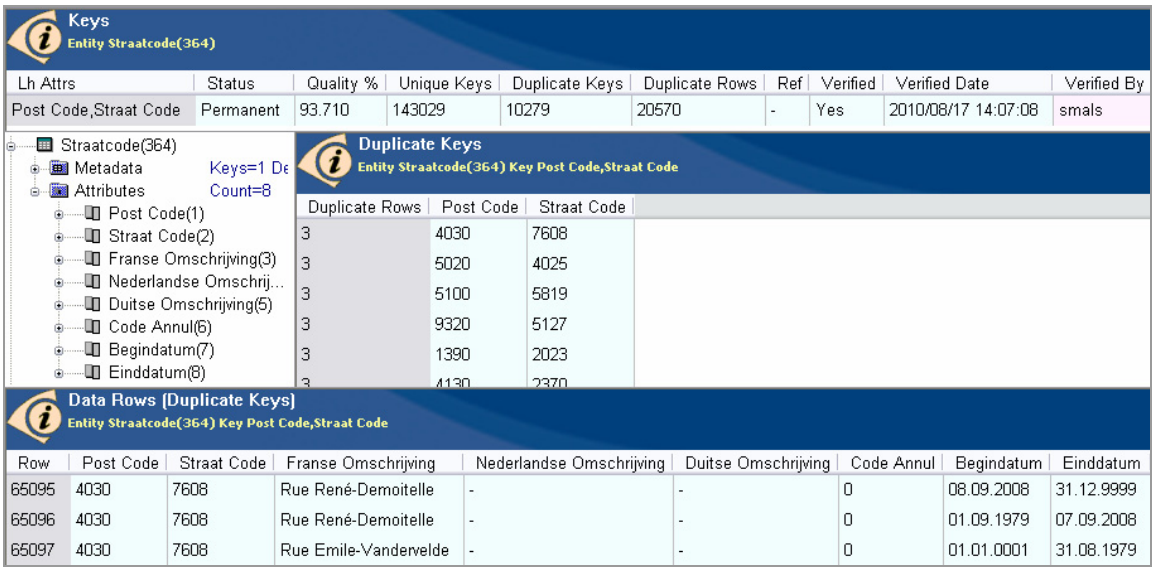


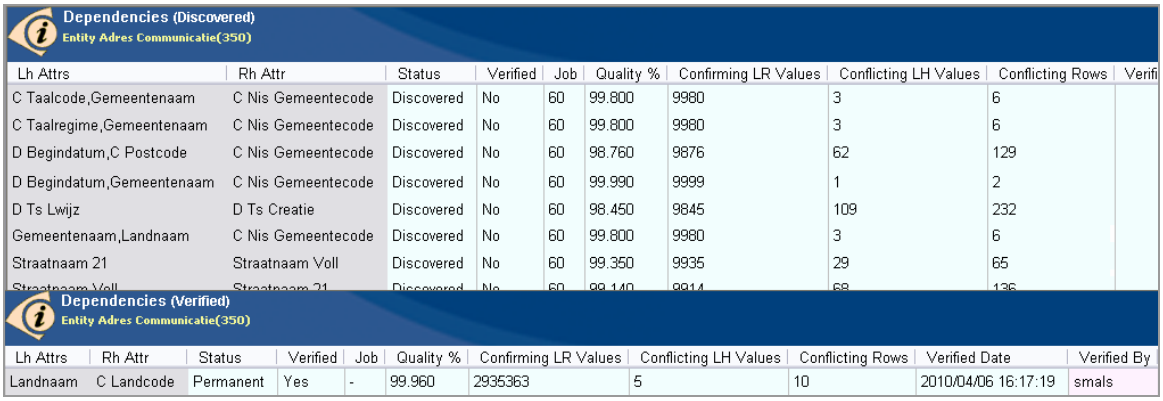
Figure 25 : Clés composées.

Pour une entité « Code de rue », une clé composée (« code postal », « code de rue ») a été détectée et vérifiée. Plusieurs niveaux successifs de drill-down *sont présentés : depuis la clé vérifiée (Keys) vers tous les conflits (Duplicate Keys : aperçu des combinaisons qui n'apparaissent pas de façon unique), puis vers les enregistrements individuels (Data Rows). La raison est tout de suite évidente : l'historique de cette rue unique a été inclus dans l'analyse.*

Dépendances fonctionnelles

Une dépendance fonctionnelle entre deux ensembles d'attributs *X* et *Y*, notée *X* → *Y*, est une sorte de contrainte structurelle qui implique que les valeurs présentes dans les attributs de l'ensemble *Y* soient déterminées par les valeurs présentes dans les attributs de l'ensemble *X* ; alternativement, les valeurs du composant *X* d'un enregistrement déterminent de façon unique les valeurs du composant *Y*.

Ces dépendances fonctionnelles peuvent être détectées par des outils de data quality pour toutes les combinaisons possibles d'attributs et vérifiées par l'analyste le cas échéant (Figure 26 et Figure 27).



Dependencies (Discovered)
Entity Adres Communicatie(350)

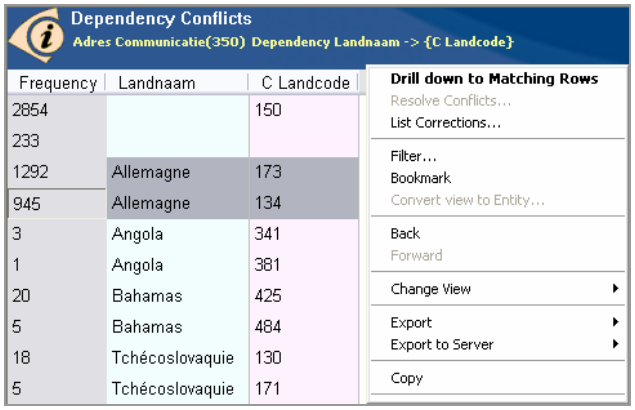
Lh Attrs	Rh Attr	Status	Verified	Job	Quality %	Confirming LR Values	Conflicting LH Values	Conflicting Rows	Verifi
C Taalcode,Gemeentenaam	C Nis Gemeentecode	Discovered	No	60	99.800	9980	3	6	
C Taalregime,Gemeentenaam	C Nis Gemeentecode	Discovered	No	60	99.800	9980	3	6	
D Begindatum,C Postcode	C Nis Gemeentecode	Discovered	No	60	98.760	9876	62	129	
D Begindatum,Gemeentenaam	C Nis Gemeentecode	Discovered	No	60	99.990	9999	1	2	
D Ts Lwijz	D Ts Creatie	Discovered	No	60	98.450	9845	109	232	
Gemeentenaam,Landnaam	C Nis Gemeentecode	Discovered	No	60	99.800	9980	3	6	
Straatnaam 21	Straatnaam Voll	Discovered	No	60	99.350	9935	29	65	

Dependencies (Verified)
Entity Adres Communicatie(350)

Lh Attrs	Rh Attr	Status	Verified	Job	Quality %	Confirming LR Values	Conflicting LH Values	Conflicting Rows	Verified Date	Verified By
Landnaam	C Landcode	Permanent	Yes	-	99.960	2935363	5	10	2010/04/06 16:17:19	smals

Figure 26 : Dépendances fonctionnelles.

Toutes les dépendances possibles sont détectées par l'outil sur la base d'un échantillon de 10.000 enregistrements. En bas : un analyste a fait contrôler une dépendance de manière exhaustive : un nom de pays détermine un code pays unique (C Lancode).



Dependency Conflicts
Adres Communicatie(350) Dependency Landnaam -> {C Landcode}

Frequency	Landnaam	C Landcode
2854		150
233		
1292	Allemagne	173
945	Allemagne	134
3	Angola	341
1	Angola	381
20	Bahamas	425
5	Bahamas	484
18	Tchécoslovaquie	130
5	Tchécoslovaquie	171

Drill down to Matching Rows
Resolve Conflicts...
List Corrections...
Filter...
Bookmark
Convert view to Entity...
Back
Forward
Change View
Export
Export to Server
Copy

Figure 27 : Drill-down vers les conflits de dépendances.

Ensuite, l'analyste a la possibilité d'effectuer un drill-down vers les conflits de dépendance (anomalies de ce type). À partir de cet aperçu, l'analyste a d'autres options pour effectuer des drill-downs vers les enregistrements en infraction, à exporter, etc.

Les conflits qui peuvent être ainsi révélés correspondent aux anomalies détectables à l'aide de contrôles croisés et/ou contrôles référentiels (2.3.1).

Règles métier

Outre les limites qui peuvent être définies au niveau d'un seul champ et les limites structurelles, d'autres liens existent entre les attributs d'une ou plusieurs entités. Généralement, ils peuvent être exprimés formellement sous forme de règles qui imposent que les valeurs qui apparaissent conjointement dans un ensemble précis d'attributs soient des **combinaisons acceptables**.

Nous nous limitons dans ce document à ce sous-ensemble de règles métier qui sont également des « Data Rules ». Une « Data Rule » est une règle exprimant une condition à propos des données d'une ou plusieurs colonnes, qui *doit toujours être vraie*.

En raison du caractère formel des règles métier, on dira, en accord avec la définition des anomalies (1.3.3), que les écarts par rapport aux règles métier sont toujours des anomalies.

Les outils de qualité des données permettent de formuler explicitement ces règles métier au niveau des métadonnées à chaque entité ainsi que de détecter dans quelle mesure les données présentes respectent ces règles. Si le concept métier à propos duquel se prononce la règle recouvre plusieurs entités, on créera d'abord dans l'environnement de l'outil une nouvelle entité composée à l'aide d'un *join* (cf. supra : *join analysis*).

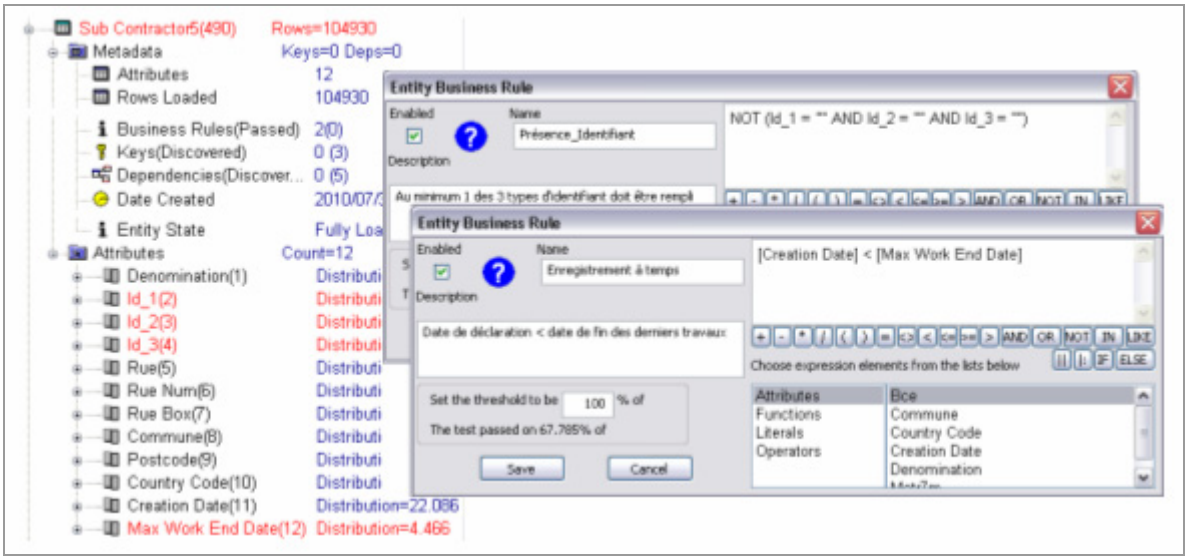


Figure 28 : Définition de règles métier.

À gauche, nous voyons une entité Sub Contractor *pour laquelle des règles métier ont été définies. À droite, nous voyons comment l'interface graphique permet de fixer des règles métier.*

Business Rules
Entity = Sub Contractor5(490)

Name	Description	Threshold	Result	Fail Count	Passing Fraction	Pass Count
Enregistrement à temps	Date de déclaration < date de fin des derniers travaux	100	failed	33803	67.785	71127
Présence_Identifiant	Au minimum 1 des 3 types d'identifiant doit être rempli	100	failed	3063	97.081	101867

Figure 29 : Aperçu des règles métier définies.

L'aperçu montre le nom et la description ainsi que le résultat d' exécution (« Passing Fraction », « Fail Count », etc.). Un drill-down est possible, tant vers les enregistrements qui respectent les règles (matching rows) que vers ceux qui les violent (failing rows).

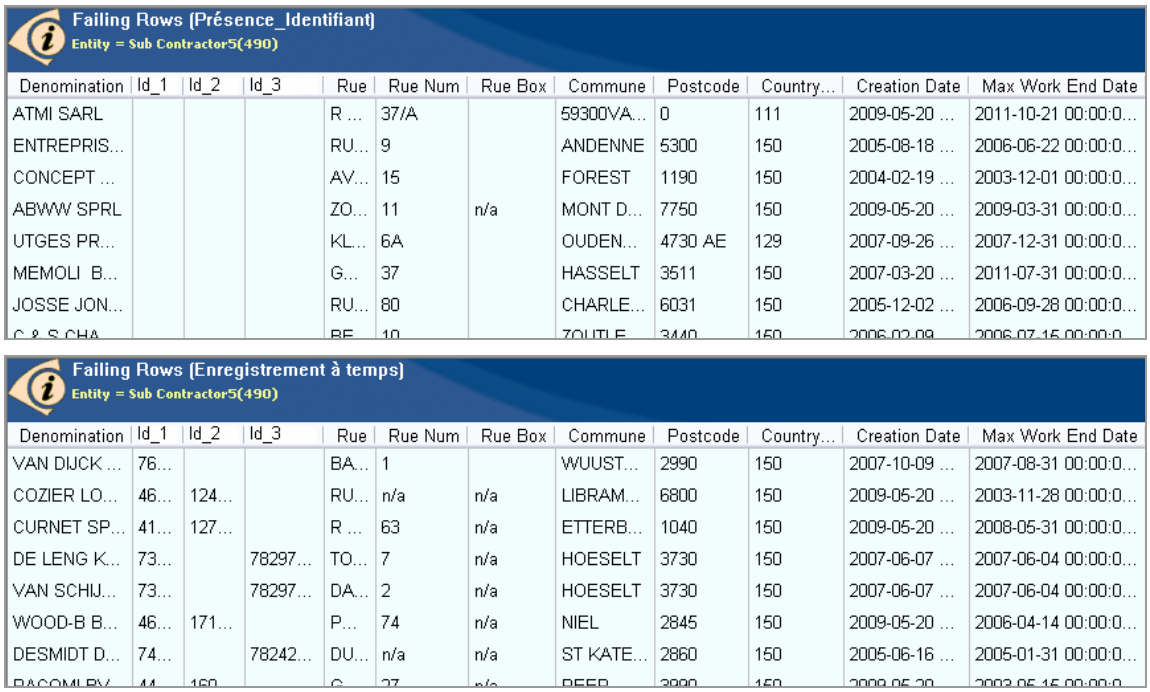
L'exemple ci-dessus (Figure 28 et Figure 29) montre comment des règles métier ont été définies et suivies dans un projet concret. L'entité (*Sub Contractor*) a été composée à l'aide d'un *join* pour également ajouter l'attribut *Max Work End Date*, de sorte que des règles métier puissent être suivies à ce niveau également. Pour des raisons didactiques, les règles métier ont été légèrement adaptées ici.

Une première règle métier, dénommée « Présence_Identifiant », vérifie si au moins une des trois possibilités de données d'identification est présente. Une seconde, dénommée « Enregistrement à temps », vérifie si la déclaration des sous-traitants concernés (*Sub Contractor*) a bien été remplie à temps, en l'occurrence au plus tard avant la fin des travaux.

Exemples

Le tableau ci-dessous permet au lecteur d'aisément vérifier que les enregistrements montrés dans la Figure 30 sont à tort.

Nom	Description	Expression
Présence_Identifiant	Au minimum 1 des 3 types d'identifiant doit être rempli	NOT(Bce = "" AND Noss = "" AND Matr7m = "")
Enregistrement_à_temps	Date de déclaration < date de fin des derniers travaux	[Creation Date] < [Max Work End Date]



Failing Rows (Présence_Identifiant)
Entity = Sub Contractor5(490)

Denomination	Id_1	Id_2	Id_3	Rue	Rue Num	Rue Box	Commune	Postcode	Country...	Creation Date	Max Work End Date
ATMI SARL				R ...	37/A		59300VA...	0	111	2009-05-20 ...	2011-10-21 00:00:0...
ENTREPRIS...				RU...	9		ANDENNE	5300	150	2005-08-18 ...	2006-06-22 00:00:0...
CONCEPT ...				AV...	15		FOREST	1190	150	2004-02-19 ...	2003-12-01 00:00:0...
ABWW SPRL				ZO...	11	n/a	MONT D...	7750	150	2009-05-20 ...	2009-03-31 00:00:0...
UTGES PR...				KL...	6A		OUDEN...	4730 AE	129	2007-09-26 ...	2007-12-31 00:00:0...
MEMOLI B...				G...	37		HASSELT	3511	150	2007-03-20 ...	2011-07-31 00:00:0...
JOSSE JON...				RU...	80		CHARLE...	6031	150	2005-12-02 ...	2006-09-28 00:00:0...
C & S CHA...				BE...	10		ZOUTLE...	3440	150	2006-02-09 ...	2006-07-15 00:00:0...

Failing Rows (Enregistrement à temps)
Entity = Sub Contractor5(490)

Denomination	Id_1	Id_2	Id_3	Rue	Rue Num	Rue Box	Commune	Postcode	Country...	Creation Date	Max Work End Date
VAN DIJCK ...	76...			BA...	1		WUUST...	2990	150	2007-10-09 ...	2007-08-31 00:00:0...
COZIER LO...	46...	124...		RU...	n/a	n/a	LIBRAM...	6800	150	2009-05-20 ...	2003-11-28 00:00:0...
CURNET SP...	41...	127...		R ...	63	n/a	ETTERB...	1040	150	2009-05-20 ...	2008-05-31 00:00:0...
DE LENG K...	73...		78297...	TO...	7	n/a	HOESELT	3730	150	2007-06-07 ...	2007-06-04 00:00:0...
VAN SCHIJ...	73...		78297...	DA...	2	n/a	HOESELT	3730	150	2007-06-07 ...	2007-06-04 00:00:0...
WOOD-B B...	46...	171...		P...	74	n/a	NIEL	2845	150	2009-05-20 ...	2006-04-14 00:00:0...
DESMIDT D...	74...		78242...	DU...	n/a	n/a	ST KATE...	2860	150	2005-06-16 ...	2005-01-31 00:00:0...
RACOMI BV...	44...	160...		G...	27	n/a	BEER...	3990	150	2009-05-20 ...	2003-05-15 00:00:0...

Figure 30 : Règles métier.

Possibilité de drill-down vers les enregistrements qui violent les règles métier. En haut : « Présence_Identifiant ». En bas : « Enregistrement_à_temps ».

3.1.3. Conseils pour les analystes, les développeurs et les chefs de projets

Détection d'anomalies

Les outils de data quality s'avèrent donc parfaitement appropriés pour effectuer tant les contrôles au niveau d'un seul champ que les contrôles liés à l'intégrité référentielle (cohésion structurelle) ainsi que - à l'aide de règles métier et d'une analyse des dépendances - les contrôles croisés.

Data Profiling et les formes normales

Pour l'exécution d'un Data Profiling exhaustif, le plus simple est que les données soient en troisième forme normale[79], mais si tel n'est pas le cas, un outil de data quality offre les fonctionnalités permettant, d'une part, d'établir la forme normale et, d'autre part, de dévoiler les défauts dans les données où l'intégrité référentielle n'est pas respectée. Ces défauts se trouvent souvent à la base d'échecs de migrations de données vers un autre modèle ou vers un autre système SGBD.

Migrations de données, re-engineerings d'applications

En établissant les problèmes qui peuvent se produire avec les données, les modèles et les contraintes existantes, les outils de data quality permettent

[79] ELMASRI R., NAVATHE S., *Fundamentals of Database Systems (fifth edition)*, Addison-Wesly, Boston, 2007.

d'encadrer ou préparer le transfert de données vers un nouveau modèle ou un nouveau système SGBD.

Dans le cadre de re-engineering d'applications, il s'agit entre autres des défis de migration, de choisir :

1) quel niveau de contrôle de qualité sera renforcé par l'application, outre les contrôles d'intégrité référentielle supportés par le système SGBD ;

2) quels standards utiliser, en cas de manque de standardisation ou en cas de conflits entre différents standards (voir 3.2.2).

Toute analyse nécessaire pour opérer les bons choix, peut être supportée par les outils de data quality, sur la base de l'ensemble des données existantes et en concertation avec le business.

Analyse de données non structurées

Parmi les fonctionnalités de Column Property Analysis, certains outils de data quality commencent à supporter des analyses axées sur un contenu de type non structuré (souvent présent dans des champs textuels), par exemple à l'aide de l'analyse de phrases (Phrase Analysis). Ceci permet d'établir la distribution de toutes les phrases (de longueur 1 à N) qui se présentent dans les textes libres du champ analysé et d'en extraire certaines informations ou de standardiser davantage le contenu non standardisé (Figure 31).

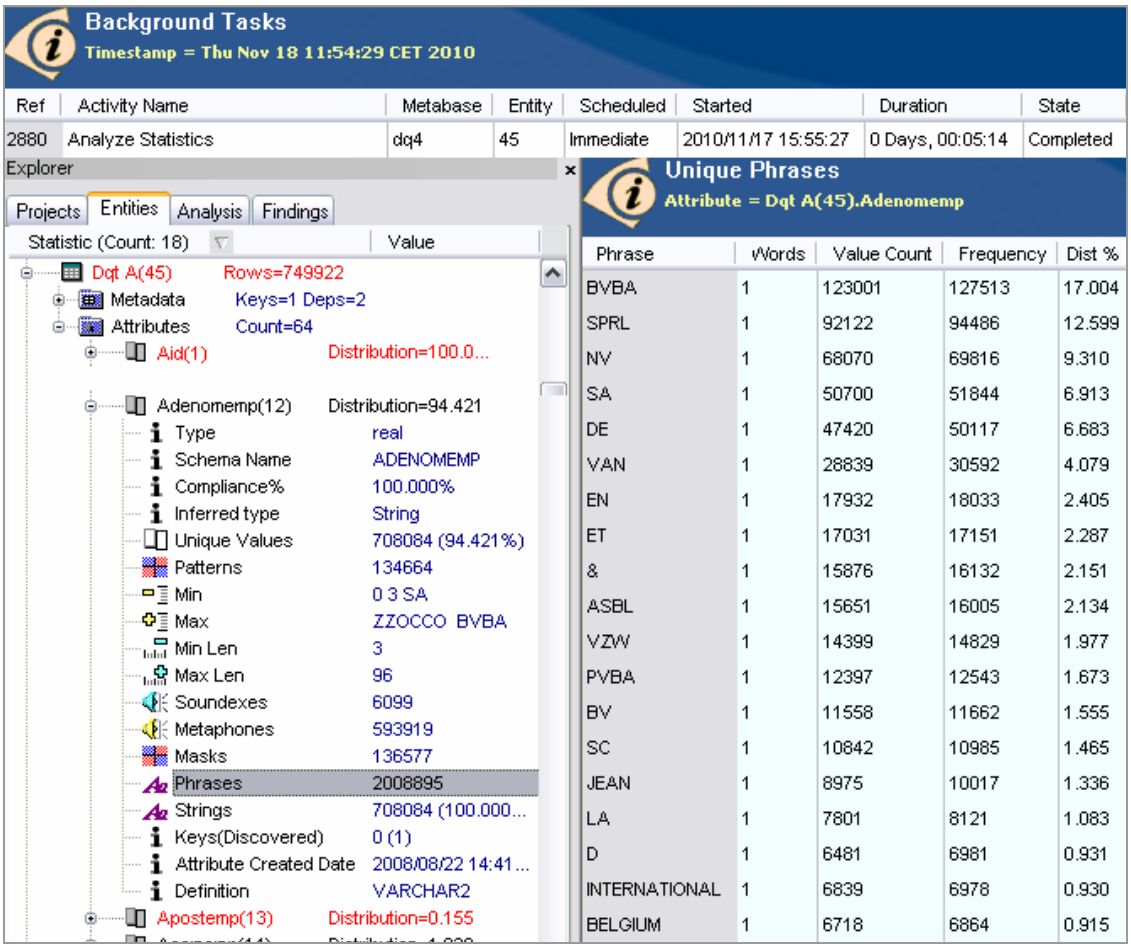


Figure 31 : Occurrence et distribution de toutes les phrases de longueur 1 à 5 présentes dans le champ de dénomination d'entreprises (Adenomemp).
Analyse réalisée en cinq minutes, pour une entité de 750.000 records. On voit apparaître les différentes formes juridiques courant en Belgique, apparemment surchargées dans le champ de la dénomination.

3.2. Standardisation des données

3.2.1. Définition

Une deuxième fonctionnalité dans le cycle de l'amélioration de la qualité des données est la standardisation des données, que l'on peut définir au moyen des objectifs suivants :

- disposer de conventions univoques pour une représentation correcte des données (= standards) pour tous les attributs ;
- corriger la représentation des données, afin qu'elles suivent le standard (les diagnostics du Data Profiling indiquent dans une grande mesure là où la correction est nécessaire) ;
- corriger certains problèmes, identifiés par les activités de Data Profiling ;
- parsing et enrichment (enrichissement) des données :
  - la (re)structuration des champs non structurés, structurés de manière erronée, ou surchargés[80], au moyen de la subdivision en unités significatives (*parsing*) ;
  - l'ajout de connaissances avec une règle métier - comme des calculs, des annotations (dans le cas où une adresse est invalide, un numéro de maison inexistant...), l'information provenant des *tableaux de connaissance* (codes postaux de tables annexes, données géographiques et démographiques, données de logiciels commerciaux, Enterprise Reference Data ou Master Data) ;
- former la base d'un Data Matching (cette technique est détaillée en 3.3) plus performant en facilitant le *matching* champ par champ pour les champs standardisés.

La standardisation des données peut donc être interprétée à trois niveaux :

- comme la résolution des problèmes de qualité des données dus au manque de standardisation (recensés à l'aide d'un Data Profiling ou non) ;
- comme l'ensemble des actions (manipulations des données) exécutées en préparation à un Data Matching (corrections automatiques, en interne, dans la mémoire…) ;
- comme un projet de nettoyage en lots destiné à améliorer la qualité des données, après avoir convenu du standard visé avec le business (conformément au « Fitness for Use ») (= aspect Data Governance de certains problèmes de qualité des données en rapport avec le manque de standardisation).

3.2.2. Standardisation des données à l'aide d'outils de qualité des données

Selon la complexité du domaine concerné par le manque de standardisation, plusieurs fonctionnalités peuvent apporter une aide.

[80] Un champ est dit être surchargé s'il contient, outre l'information prévue ou censée être présente, d'avantage d'information. Par exemple, un champ de rue qui contient le nom de la rue plus le numéro et la boîte.

Standardisation dans le cas de domaines restreints

Un domaine restreint peut se décrire à l'aide d'une liste finie et exhaustive de valeurs autorisées univoques. Si l'on impose un standard pour la représentation des données dans un domaine restreint, la transformation requise pour passer des représentations alternatives à la représentation standard serait facilement définissable et exécutable. Ceci peut se faire aisément dans un outil de Data Quality.

Une liste fermée de codes constitue un exemple typique de domaine restreint. Plusieurs données de la DmfA ont ainsi un domaine restreint (code travailleur, code prestation, activité par rapport au risque…), comme en témoignent les « annexes structurées ».

La Figure 32 offre un exemple typique. Le « code pays » est interprété de plusieurs manières lors de l'encodage des données. Il en résulte un manque évident de standardisation dans les valeurs de ce champ dans la banque de données (voir Orig Landcd). Les quatre dernières colonnes montrent comment un outil de data quality a standardisé cela en quatre restitutions possibles de la norme ISO-3166 pour les codes pays.

Data Rows (Dynamic)
Entity = pl tranfrmr p26(513)

Tsq Denom	Adres	Boite	Tsq Postcd	Tsq Commune	Orig Landcd	Iso3166 Cd	Iso3166 2l	Iso3166 3l	Iso3166 Nl	Ins Cd	Ins Nl
GREEFS ...					150	056	BE	BEL	België	150	België
JACOBS ...	MAH...			BLACKROC...	116	372	IE	IRL	Ierland	116	Ierland /Eire/
SUIR EN...	OLD...			WATERSFO...	IRL	372	IE	IRL	Ierland		
BONCIK J...	268		067 82	MODRA NAD...	141	703	SK	SVK	Slowakije	141	Slovaakse Republiek
BONCIK J...	HIA...		067 82	MODRA NAD...	SK	703	SK	SVK	Slowakije		
BELROO...	Vizi...		1031	BUDAPEST	115	348	HU	HUN	Hongarije	115	Hongarije (Rep.)
BELROO...	VIZI...		1031	BUDAPEST	H	348	HU	HUN	Hongarije		
VOORBIJ ...	SICI...		1045 AX	AMSTERDAM	129	528	NL	NLD	Nederland	129	Nederland
VOORBIJ ...	SICI...		1045	AMSTERDAM	NL	528	NL	NLD	Nederland		
DELAT KFT	MAD...		1131	BUDAPEST	H	348	HU	HUN	Hongarije		
DE WAAL...	POS...		1440	AP PURMER...	NL	528	NL	NLD	Nederland		
FUNDERI...	POS...		1440	AP PURMER...	NL	528	NL	NLD	Nederland		
SANDMA...	czer...	4	20 349	LUBLIN	122	616	PL	POL	Polen	122	Polen (Rep.)
SANDRA...	CZE...		20 349	LUBLIN	PL	616	PL	POL	Polen		
TPA EDIL ...	Via ...		24129	BERGAMO	128	380	IT	ITA	Italië	128	Italië
TPA GRO...	VIA ...		24129	BERGAMO	I	380	IT	ITA	Italië		
VLASMA...	Stee...		2407 BD	ALPHEN AA...	129	528	NL	NLD	Nederland	129	Nederland
VLASMAN	STE...		2407	BD ALPHEN ...	NL	528	NL	NLD	Nederland		
VLASMAN	STE...		2407	BD ALPHEN ...	NL	528	NL	NLD	Nederland		

Figure 32 : Tables des codes-pays.

Standardisation de modèles distincts utilisés pour déclarer les codes pays (Orig Landcd → Iso3166, INS). Étant donné qu'à partir des codes ISO 3166, il n'est pas toujours possible d'établir un lien univoque avec les codes INS, une standardisation sur la base des codes INS n'est pas toujours possible. L'inverse est toutefois vrai.

Standardisation dans le cas de domaines complexes

La standardisation est dans ce cas bien plus difficile et doit être supportée par des fonctionnalités supplémentaires comme le parsing et l'enrichissement, avec des informations spécifiques aux domaines et des données référentielles externes.

Parsing & Enrichment

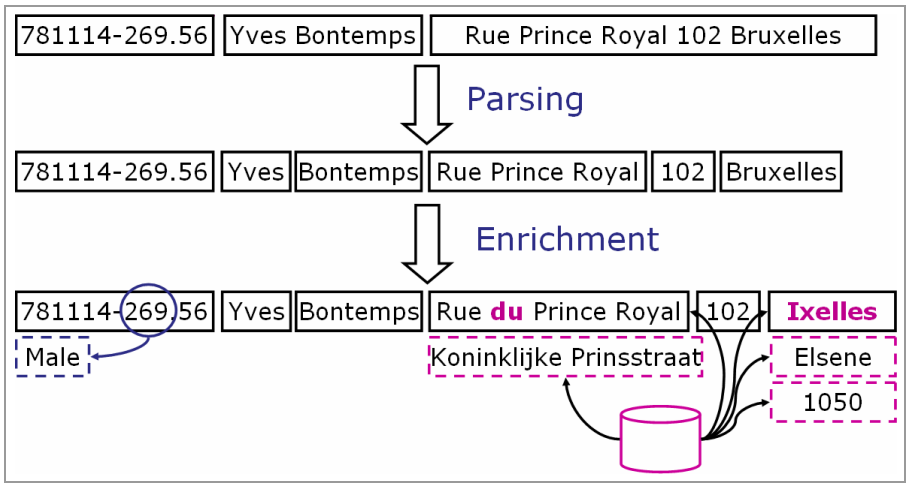


Figure 33 : Exemple du résultat d'une standardisation des données, à l'aide des techniques du parsing et de l'enrichissement.

Dans l'exemple ci-dessus (Figure 33), un champ d'adresse composé est subdivisé grâce au parsing en parties significatives (nom de la rue, numéro, commune), puis enrichi avec le nom de la rue dans l'autre langue nationale, avec le code postal et la commune. Le champ composé du nom est subdivisé en nom de famille et prénom.

Grâce à une connaissance spécifique du domaine, un numéro de registre doté d'une structure connue (*pattern*) peut être subdivisé afin d'en retirer des informations pouvant servir à l'enrichissement :

Pattern :	99.99.99-999.99
Sémantique :	YY.MM.DD-CJN.CD (Compteur Journalier des Naissances, Check Digit)
Enrichissement :	CJN mod 2 = 0 → sexe = 'M' CJN mod 2 = 1 → sexe = 'F' Date de Naissance = DD '/' MM '/' YY

Le développement d'une logique exécutable en état d'effectuer la standardisation des données revient donc à l'encodage d'une multitude de règles et de connaissances métier, indispensables à un parsing et un enrichissement corrects. Ceci représente un effort considérable.

Afin de soutenir les opérations de standardisation des données, les logiciels de Data Quality incorporent des bases de connaissances spécialisées avec :

- une abondance de règles applicables (au-delà de 50.000) ;
- des dizaines d'années/homme d'expérience ;
- des *lookup tables* (tables associatives) avec des informations géométriques et démographiques ;
- un support (bases de connaissances) dans des contextes et régions spécifiques (pays, langue, culture) ;
- la reconnaissance de *patterns*, des expressions régulières, grammaire…

Vous trouverez un exemple plus élaboré, présentant directement les avantages de la standardisation *infra* (3.4), ceci en guise d'illustration des résultats atteints dans le cadre d'un projet concret de nettoyage (standardisation et *matching*) d'adresses.

3.2.3. Conseils pour les analystes et les développeurs

Standardisation des données et Master Data Management

Avertissement : la standardisation des données est souvent liée à des tables de codes, qui sont toujours des Master Data et sont souvent utilisées dans plusieurs bases de données, applications et environnements (intra-institutionnels ou interinstitutionnels). La résolution de ce sous-ensemble de problèmes de qualité des données doit donc tenir compte des règles particulières valables pour la gestion des Master Data. Vous trouverez plus d'informations à propos des Master Data et du Master Data Management dans une autre étude de la section Recherches[81].

Standardisation des données en support du Data Matching

Les techniques de Data Matching (3.3) bénéficient de la standardisation préalable des données à comparer : les résultats seront en effet bien plus précis. Ce point sera illustré davantage dans 3.4.1.

Champs surchargés (*overloaded fields*)

Dans la pratique, les *champs surchargés* sont très fréquents, certainement dans le cas des données d'adresses. Sans outils de qualité des données, il est difficile de traiter cet aspect ou de déterminer la gravité de la situation actuelle.

3.3. Data Matching

3.3.1. Définition

Le Data Matching ou Record Linkage fait référence à la recherche d'enregistrements relatifs à une même entité réelle.

Comme les bases de données (et les applications qui les alimentent) se servent de *modèles* de la réalité, elles sont rarement totalement précises et complètes, comme l'illustre la Figure 34 :

- une même entité réelle peut être représentée de différentes manières imparfaites ;
- elle peut donc à la fois être présente dans plusieurs enregistrements d'une même table ou base de données et absente dans d'autres tables ou bases de données.

On parle alors de doublons et d'incohérences dans une ou plusieurs bases de données.

[81] TRIGAUX J.-C., *Master Data Management - Mise en place d'un référentiel de données*, Deliverable, 2009/trim4/01, Smals, Section Recherches, Bruxelles, 2009.

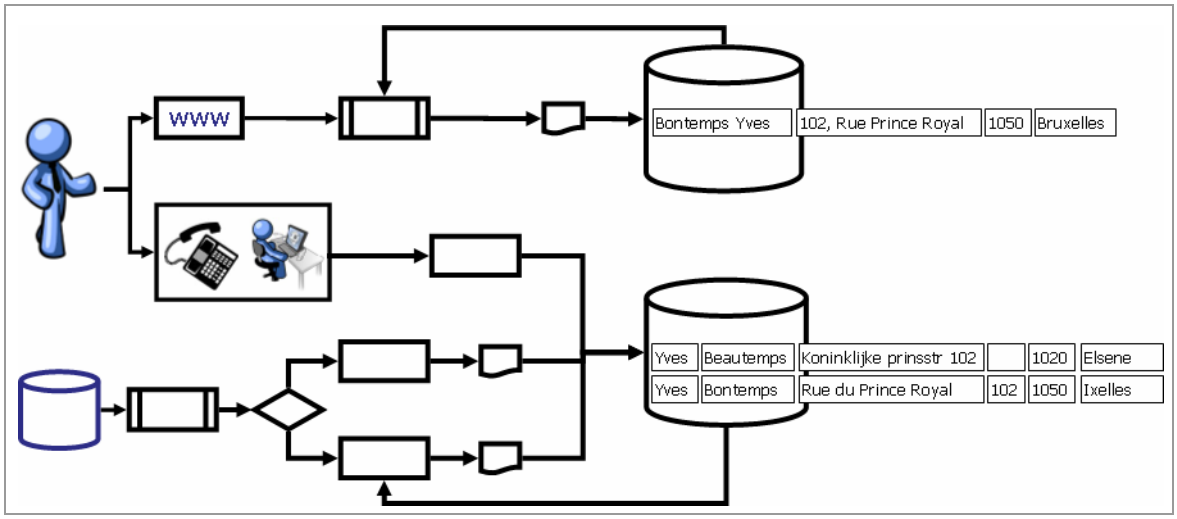


Figure 34 : La réalité et comment elle peut être enregistrée. Origine de doublons et d'incohérences.

Les objectifs du Data Matching sont donc les suivants :

- Déduplication d'une base de données, avec détection des doublons.
- Data Integration : relier plusieurs bases de données, qui ont leur propre modèle, pour détecter et traiter les incohérences afin de trouver l'information adéquate dans un but précis (y inclus : détection des doublons, détection de la fraude, migration des données vers un nouveau modèle et système…).
- Éviter l'introduction de nouveaux doublons ou incohérences (*online matching* à l'entrée des données).

Vu le manque de standardisation et de contrôles de qualité qui existe souvent dans les applications et les processus qui alimentent les bases de données, les techniques de Data Matching sont censées être capables de traiter le manque de standardisation, la présence de fautes de frappe et autres imprécisions, le multilinguisme, etc. ; on parle alors aussi de **Fuzzy Matching**.

En principe, le Data Matching consiste à comparer deux par deux chaque enregistrement avec tous les autres enregistrements (tant dans une seule source qu'entre plusieurs sources), afin de décider à chaque comparaison s'il s'agit d'un *match* ou d'un *non match*. Certaines techniques permettent également de décider que le résultat de la comparaison soit un *suspect match* (match potentiel).

La comparaison entre les deux enregistrements se fait généralement à deux niveaux :

- Dans une première phase, la similarité entre les enregistrements est définie champ par champ.
- Dans une seconde phase, les similarités champ par champ sont agrégées pour parvenir à un résultat global pour chaque paire d'enregistrements.

Comparaisons champ par champ

De nombreuses méthodes permettent d'opérer des comparaisons champ par champ existent. Nous en avons décrit certaines dans une étude précédente[82]. Ces méthodes peuvent globalement être subdivisées en deux familles.

[82] BONTEMPS Y., BOYDENS I., VAN DROMME D., *Data Quality : tools*, Deliverable, 2007/trim3/02, Smals, Section Recherches, Bruxelles, 2007.

- D'une part, il y a les méthodes **booléennes**, qui en résultat d'une comparaison débouchent toujours sur un *match* ou un *non match* (1 ou 0). Il s'agit soit d'algorithmes phonétiques, soit de méthodes consistant en une combinaison de prédicats (logiques) et de règles.
- D'autre part, il existe les méthodes comparatives dont le résultat se présente sous la forme d'un score de **similarité**, autorisant une appréciation *match* ou *non match* plus subtile. Ces méthodes peuvent être *word-based* (comparaison entre des chaînes de mots individuels) ou *token-based* (comparaison entre des chaînes de mots multiples). La difficulté réside parfois dans la définition d'une valeur seuil (rayon *k* - Figure 35) à partir de laquelle la similarité est interprétée comme *match* ou *non match*.

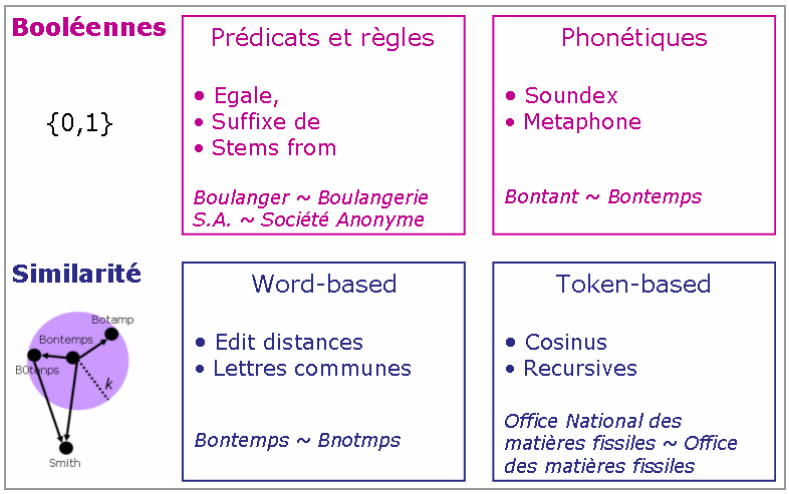


Figure 35 : Data Matching - Familles de méthodes de comparaisons champ par champ.

Vous trouverez plus de détails encore dans notre étude précédente sur les data quality tools[83] ainsi que dans une étude comparative de méthodes toujours courantes[84].

Comparaisons par enregistrement (Record Linkage)

Une fois les comparaisons champ par champ accomplies, les résultats sur les enregistrements complets doivent être combinés pour déboucher sur une évaluation *match* ou *non match* (également *suspect match* dans certains cas). Globalement, deux approches peuvent être distinguées : l'approche déterministe, et l'approche probabliste.

- Dans l'approche **déterministe**, l'appréciation de *match* peut être décrite au moyen d'une table de vérité : on énumère de manière explicite les combinaisons de comparaisons champ par champ qui, selon les utilisateurs métier, débouchent respectivement sur *match*, *non match* et *suspect match* (selon les critères « Fitness for Use »). L'ordre des règles dans la table de vérité peut avoir de l'importance.

  Ceci peut être interprété et géré comme un ensemble de règles métier qui expliquent pourquoi il s'agit ou non d'un *match*. Le caractère

[83] BONTEMPS Y., BOYDENS I., VAN DROMME D., *Data Quality : tools*, Deliverable, 2007/trim3/02, Smals, Section Recherches, Bruxelles, 2007.

[84] COHEN W., RAVIKUMAR P. et FIENBERG S., « A comparison of string distance metrics for name-matching tasks », dans *The IJCAI Workshop on Information Integration on the Web (IIWeb)*, Acapulco, 2003. http://www.cs.cmu.edu/~wcohen/postscript/ijcai-ws-2003.pdf

explicite de cette approche permet d'itérer avec les utilisateurs métier afin de trouver les bonnes règles qui correspondent au « Fitness for Use », ceci afin de trouver les bonnes définitions pour les doublons et les incohérences dans un contexte donné.

- Dans l'approche probabiliste, l'un ou l'autre algorithme attribue à chacun des sous-critères champ par champ (souvent des scores de similarité, donc exprimés en variables continues) un poids relatif. Le résultat est donc un score continu, ce qui laisse de la place à l'interprétation. Dans la pratique, on travaille à deux seuils sur l'échelle continue (Figure 36).

  On a ainsi généralement moins de règles à gérer, mais pour chacune d'elles, le bon choix des poids relatifs et le bon positionnement de chaque seuil s'avèrent difficiles : les résultats sont très sensibles aux changements. Le caractère « black-box » permet moins facilement d'itérer avec des utilisateurs métier et d'expliquer pourquoi l'algorithme apprécie le score agrégé comme un *match*, un *non match* ou un *suspect match*.

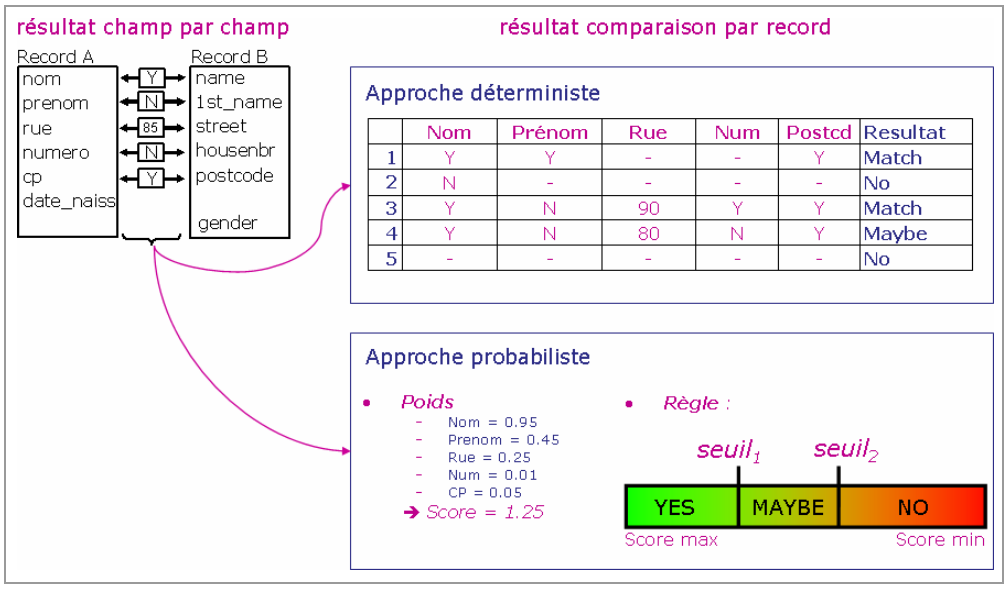


Figure 36 : Data Matching - Illustration des principes de base des approches déterministes et probabilistes à afin d'arriver à une appréciation globale de match, en fonction des résultats de comparaisons champ par champ.

3.3.2. Data/Fuzzy Matching à l'aide d'outils de Data Quality

Lors de l'exécution d'un Fuzzy Data Matching à l'aide d'outils de qualité des données, deux aspects sont essentiels :

- **haute performance** lors de la mise en œuvre des stratégies de *matching* choisies ;
- **haute flexibilité** lors du paramétrage des stratégies de *matching*.

Tous deux sont nécessaires, d'une part pour que les analystes puissent rapidement y voir clair, mieux affiner et davantage itérer avec les spécialistes métier, d'autre part pour permettre l'intégration en ligne dans des applications avec la garantie d'une performance suffisante de l'ensemble tout en préservant la possibilité de réutilisation et d'adaptation aux besoins spécifiques.

Ce sont précisément ces aspects qui justifient l'utilisation d'outils de qualité des données et en expliquent le succès.

Optimisation de performance : « Blocking »

Exécuter un *matching* consiste généralement à comparer chaque enregistrement d'une source avec chaque enregistrement d'une autre source (ou à comparer tous les enregistrements avec chaque autre enregistrement en cas de détection de doublons dans un seul répertoire). Par exemple, si les répertoires à comparer comportent un nombre d'enregistrement de l'ordre de 10^6 (millions), le nombre de comparaisons à réaliser sera de l'ordre de 10^{12} (mille milliards).

Or, même avec l'infrastructure hardware la plus moderne, une telle opération est **loin d'être négligeable.** Surtout au niveau du temps de traitement, cela peut occasionner un problème insurmontable. L'une des possibilités d'**optimisation** les plus répandues dans les logiciels modernes de Data Matching est généralement appelé « **Blocking** ». Il en existe un grand nombre de variantes, d'affinements et d'alternatives[85].

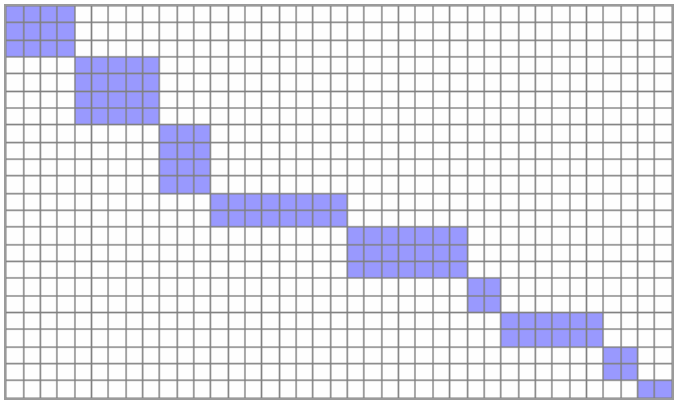


Figure 37 : Principe de base de la technique d'optimisation « Blocking ». L'espace de recherche pour les enregistrements similaires est réduit par une fragmentation en sous-espaces de même valeur pour une colonne critique déterminée.

La technique d'optimisation « Blocking » repose sur les trois **principes de base** suivants :

1. Définition de bloc sur la base d'une colonne critique (code postal, par exemple) :
   - pour chacune des valeurs présentes dans la colonne critique, les enregistrements qui possèdent la même valeur pour cette colonne sont réunis dans des blocs (voir figure ci-dessus) ;
   - une indexation et un tri de tous les enregistrements sont donc nécessaires.
2. Comparaison d'enregistrements deux par deux dans chaque bloc :
   - seuls les enregistrements d'un même bloc (donc avec une même valeur pour la colonne critique) sont comparés deux par deux, et ce pour chaque bloc. Dans le cas des codes postaux, seules les entreprises auxquelles correspond le même code postal seront comparées deux par deux.

[85] WINKLER W. E., *Overview of Record Linkage and Current Research Directions*, 2006. http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf

3. Identification ou définition d'une ou plusieurs colonnes critiques :
   - à l'aide de techniques automatisées ou sur la base de connaissances du domaine, les données critiques du Blocking doivent donc être choisies de telle manière que l'on puisse espérer qu'aucun *match* ne soit possible entre les enregistrements qui revêtent différentes valeurs pour la donnée critique ;
   - on choisira donc de préférence les données déjà standardisées ou au moins relativement stables, fiables et précises. Cependant, étant donné qu'une certitude est impossible, il y aura un « **trade-off » entre la précision et la performance**.

Remarque. Les outils de qualité des données peuvent **en grande partie compenser ce trade-off** en offrant la possibilité de définir la colonne critique comme une colonne de métadonnées basée sur plusieurs données critiques ainsi qu'en autorisant plusieurs définitions de telles colonnes critiques en même temps et en en combinant les résultats.

L'exemple proposé dans la Figure 38 et la Figure 39 montre comment créer une ou plusieurs colonnes de métadonnées (*Window Key 01* et *Window Key 02*) à l'aide d'outils de qualité des données via l'interface utilisateur graphique pour servir de colonne critique pour le « Blocking ».

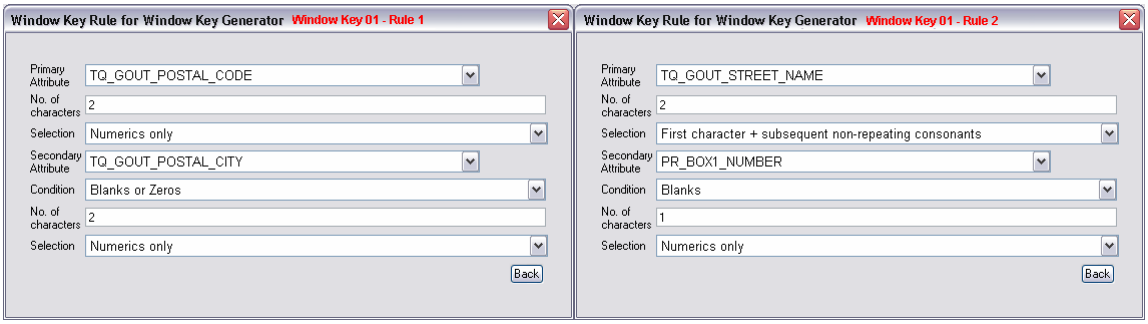


Figure 38 : constitution du Window Key 01.

Window Key 01 *est élaboré comme la composition de deux caractères (Numerics only) du code postal (à gauche), suivis de deux caractères (premier caractère + consonnes non répétées suivantes) du nom de la rue (à droite). Des alternatives sont prévues sur la base de la commune (à gauche) et du numéro de boîte (à droite) si l'un des deux champs composants est vide ou ne contient que des zéros.*

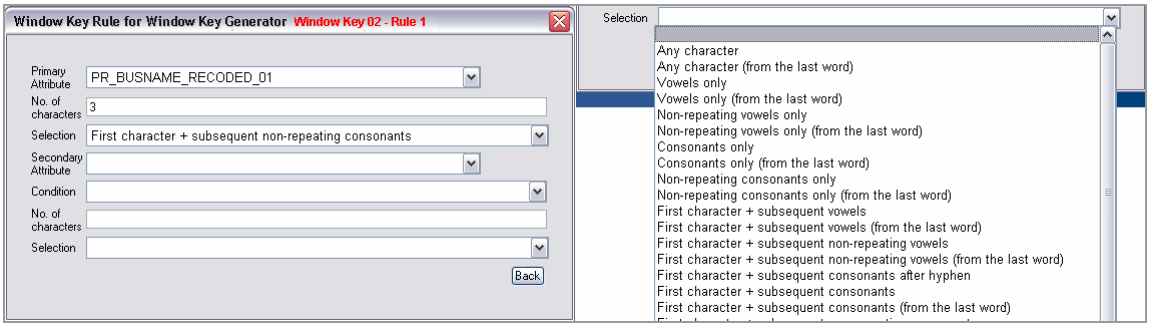


Figure 39 : constitution du Window Key 02.

À gauche : une alternative, Window Key 02*, est élaborée sur la base du nom de l'entreprise (premier caractère + consonnes non répétées suivantes).*
À droite : un extrait des nombreuses options dont dispose un analyste pour sélectionner des sous-chaînes lors de la création d'une colonne critique composée (« window key ») pour le « blocking ».

Une longue liste de méthodes permettant de choisir des caractères et des sous-chaînes pour élaborer des *window keys* est proposée en annexe (6.2).

Flexibilité et réutilisation du paramétrage des stratégies de matching

Les data quality tools permettent d'implémenter des stratégies de *matching* avec une très grande souplesse. Ceci est essentiel, car un Data Matching réussi requiert **énormément d'ajustement fin.** En effet, à l'instar du Data Profiling, le Data Matching traverse plusieurs phases où l'adaptation et la réutilisation doivent chaque fois être possibles : une phase de découverte ou d'analyse, une phase de validation et une phase d'exploitation.

Dans la **phase de découverte**, sur la base de l'expérience, des connaissances métier et de la connaissance des données (souvent acquises grâce à un Data Profiling préalable), des choix initiaux sont opérés pour une stratégie de *matching* et un analyste adapte les stratégies le mieux possible aux besoins initialement connus via un *trial-and-error*.

Dans la **phase de validation**, une concertation intensive a lieu avec les spécialistes métier et les propriétaires métier afin d'accorder les stratégies de *matching* aux critères « Fitness for Use ». Ainsi, par exemple, il est possible de chercher collectivement la bonne définition de ce qui est « double » ou « incohérent » lors d'une détection *fuzzy* de doublons, d'une détection de fraude, etc. Il est aussi possible de subdiviser les modèles de *matching* en types, de sorte qu'une exploitation distincte soit possible pour chaque type (voir aussi la Figure 42).

Enfin, dès que le business a validé le choix des stratégies de Data Matching à mettre en œuvre, les définitions résultantes peuvent être implémentées pour exécuter le Data Matching, le Fuzzy Matching ainsi que la détection des doublons et des incohérences dans l'environnement de production. Dans cette **phase d'exploitation** également, il s'agit de pouvoir continuer à suivre les besoins en évolution du business d'une part et à apporter les adaptations qui s'imposent d'autre part.

Si le business souhaite détecter et traiter certains de ces doublons et incohérences **comme des anomalies**, il suffit d'établir un lien direct avec les systèmes de suivi de l'historique des anomalies (2.1).

Meilleures pratiques

Un bon outil de qualité des données procure donc **une interface utilisateur graphique permettant de paramétrer** des stratégies, de sorte qu'il ne faille pas développer de code dans les phases d'analyse et de validation. En outre, les stratégies de matching validées peuvent être **réutilisées** dans d'autres projets similaires ou servir de point de départ pour des contextes analogues, mais aux besoins différents.

Ceci a déjà été illustré en partie ci-dessus dans l'exemple de la création de *window keys* (Figure 38 et Figure 39). Nous allons maintenant illustrer cela à l'aide du paramétrage de modèles de *matching* (*match types*). La Figure 40 montre comment un analyste peut aisément fixer des critères de comparaison champ par champ (plus de détails en annexe - 6.3), indiquer à cette fin des niveaux de score et les utiliser dans une table de vérité pour un score de similarité agrégé, selon l'approche déterministe.

Field List Editor - C:\Documents and Settings\DRV\Application Data\Trillium Software\TssUI\11.5\cache\dq4\metabase\project15\settings\p427_8_bebus1fl

Description	Score A	Score B	Score C	Score D	Score E	Comparison Routine	Propagation R...	Field Name 1	F...	F...	Routine Modifier
business_name	100	90	80			busname		PR_BUSNAME_RECODED_01			
bus_sort	95					BUSNAME		Pr_busname_recoded_01			SORT
bus_squish	95	75				SPELLING		Pr_busname_recoded_01			SQUISH
street_name	95	87	82			streets		TQ_GOUT_STREET_NAME			
house_number	100	98	90			houseno		TQ_GOUT_HOUSE_NUMBER			
box_number	100					absolute		PR_BOX1_NUMBER			
postal_code	100	95	70			POSTCODE		TQ_GOUT_POSTAL_CODE			
city	100	90	72			SOUNDEX2		TQ_GOUT_POSTAL_CITY			

Figure 40 : Définition de modèles de matching selon l'approche déterministe.

Dans une première phase, l'analyste configure des scores de similarité champ par champ à l'aide du choix de l'algorithme pour la comparaison champ par champ (Comparison Routine, Modifier), du champ concerné (Filed Name) et des seuils pour des niveaux de score significatifs par routine de comparaison (Score A ≥ 95, Score B ≥ 87…). Notez que plusieurs critères sont possibles par champ.

Grade Pattern Editor - C:\Documents and Settings\DRV\Application Data\Trillium Software\TssUI\11.5\cache\dq4\metabase\project15\settings\p427_8_beb

Category	Pattern ID	business_name	bus_sort	bus_squish	street_name	house_number	box_number	postal_code	city
P	110	A	-	-	A	A	A	-	-
P	115	A	-	-	A	B	A	-	-
P	120	A	-	-	A	C	A	-	-
P	125	A	-	-	B	A	A	-	-
P	130	A	-	-	C	A	A	-	-
P	400	A	-	-	A	-	-	-	-
P	401	-	A	-	A	-	-	-	-
P	402	-	-	A	A	-	-	-	-
P	135	B	-	-	A	A	A	-	-
P	140	A	-	-	B	B	A	-	-
P	145	B	-	-	B	A	A	-	-
P	403	A	-	-	B	-	-	-	-
P	404	-	A	-	B	-	-	-	-
P	405	-	-	A	B	-	-	-	-
P	406	B	-	-	B	-	-	-	-
S	150	A	-	-	A	-	A	-	-
P	155	B	-	-	A	B	A	-	-
S	160	B	-	-	A	C	A	-	-

Figure 41 : Définition de modèles de matching déterministes - deuxième phase.

Dans une deuxième phase, l'analyste dresse une table de vérité selon l'approche déterministe, où la combinaison de scores de similarité sur plusieurs champs donne lieu à une appréciation agrégée de la similarité. L'ordre est important. Les plus fiables se trouvent en tête.

Le score de similarité résultant est une appréciation qualitative (Category) :

- P (Pass) est considéré comme *match*.
- S (Suspect) est considéré comme *suspect match* et peut être exploité autrement.
- F (Fail) est interprété comme « certainement pas un match ».

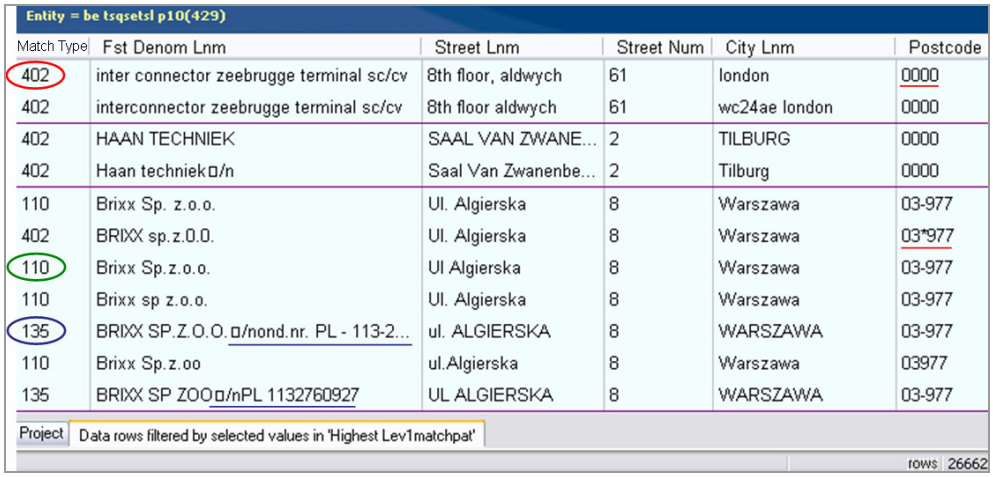
Dans chaque cas, cette catégorie est encore accompagnée de la notion Pattern ID (*match type*), permettant d'accorder à chaque *match type* une fiabilité propre et un traitement spécifique. Il est ainsi possible de faire correspondre chaque *match type* à un ANOMALY_TYPE_ID unique et d'y associer un ou plusieurs CORRECTION_SCENARIO_ID (2.1).

Flexibilité dans la concertation avec le business (validation)

Grâce à l'interface utilisateur graphique, aux possibilités *drill-down*, à la haute performance et à la flexibilité du paramétrage des stratégies de *matching*, il est facile de soumettre des résultats à la validation des spécialistes métier et d'aboutir à un résultat « fit for use » en un certain nombre d'itérations.

La Figure 42 montre comment les *match types* peuvent aider les analystes lors de l'ajustement précis des routines de *matching*. La métadonnée « Match Type »

permet à l'analyste de rassembler tous les résultats liés à un *match type* déterminé et d'évaluer si des valeurs seuil de scores de similarité sont trop élevées ou, au contraire, trop basses. Si les résultats sont organisés par grappe de doublons, telle que dans la Figure 42, on peut facilement déterminer avec les personnes du business quels modèles de *match* peuvent être considérés comme fiables et quels modèles de *match* doivent idéalement être traités avec prudence.


Entity = be tsqsetsl p10(429)

Match Type	Fst Denom Lnm	Street Lnm	Street Num	City Lnm	Postcode
402	inter connector zeebrugge terminal sc/cv	8th floor, aldwych	61	london	0000
402	interconnector zeebrugge terminal sc/cv	8th floor aldwych	61	wc24ae london	0000
402	HAAN TECHNIEK	SAAL VAN ZWANE...	2	TILBURG	0000
402	Haan techniek□/n	Saal Van Zwanenbe...	2	Tilburg	0000
110	Brixx Sp. z.o.o.	Ul. Algierska	8	Warszawa	03-977
402	BRIXX sp.z.0.0.	Ul. Algierska	8	Warszawa	03*977
110	Brixx Sp.z.o.o.	Ul Algierska	8	Warszawa	03-977
110	Brixx sp z.o.o.	Ul. Algierska	8	Warszawa	03-977
135	BRIXX SP.Z.O.O.□/nond.nr. PL - 113-2...	ul. ALGIERSKA	8	WARSZAWA	03-977
110	Brixx Sp.z.oo	ul.Algierska	8	Warszawa	03977
135	BRIXX SP ZOO□/nPL 1132760927	UL ALGIERSKA	8	WARSZAWA	03-977

Project | Data rows filtered by selected values in 'Highest Lev1matchpat' | rows 26662

Figure 42 : Présentation des résultats d'une détection de doublons.

L'exemple montre les résultats d'une détection de doublons effectuée sur la base des valeurs de dénomination et d'adresse. Les résultats sont représentés en groupes. Le « Match Type » indique le critère (déterministe) dont le logiciel Data Quality s'est servi pour désigner des enregistrements comme des « matches ».

Le business est en mesure d'associer un degré de fiabilité à chaque « Match Type », et ensuite accepter ou corriger par « Match Type ». Dans l'exemple ci-dessus, le type 110 semble très fiable, le type 135 présente une grande différence dans la dénomination et le type 402 affiche une différence dans la dénomination ainsi qu'un problème au niveau du code postal.

Il est ainsi possible de se concerter avec le business pour déboucher sur des **stratégies de *matching* validées** et des *match types* dont la fiabilité est connue.

En ce qui concerne la détection des doublons, on peut dire que l'on est réellement parvenu à la **définition des doublons** (définition des éléments à considérer comme des « doublons » selon le business) et donc de la manière dont ils doivent être détectés. En outre, tant la forme du résultat que le processus itératif de découverte et de validation même soutiennent les décisions à prendre quant à la façon dont chaque type de doublon doit être traité.

Lien avec (les systèmes de) la gestion historique des anomalies

Dans le cadre de la gestion des anomalies et de leur historique, **chaque match type validé peut être considéré comme une anomalie** et lié à un ANOMALY_TYPE_ID unique. Le fait de pouvoir disposer d'une telle **typologie** grâce aux outils de qualité des données présente un avantage considérable, car chaque type (d'anomalie détectée) devra plus que probablement être **traité d'une manière propre**. Ainsi le lien avec les **scénarios de correction** possibles (2.4.2) peut-il être établi aisément.

3.3.3. Conseils pour les analystes, les développeurs et les chefs de projets

Avertissement - *Null values*

Certains calculs sont sensibles aux *null values*. Lors du choix des méthodes et des algorithmes de Data Matching, il est primordial d'examiner l'impact de la présence de *null values*. Vérifiez bien à chaque fois l'impact sur le score résultant et convenez avec le business de l'éventuel sort à réserver aux *null values* présentes dans les attributs qui sont importants pour décider si un élément est double ou incohérent.

Suggestions pour les Window Keys

1. Lors de la configuration de Window Keys, il faut toujours tenir à l'œil la taille des *windows* (sous-espaces) résultantes. Si l'on veut préserver le gain en performance, celles-ci contiendront de préférence moins de 500 enregistrements en moyenne et peu d'exceptions à taille élevée. On peut utiliser les fonctionnalités de l'outil de data quality soi-même pour demander les histogrammes des Unique Values des *windows keys* résultantes.[86]

2. Le *trade-off* entre précision et performance, inhérent à la technique du *blocking*, peut être mitigé par la définition de plusieurs *windows keys* et ensuite la combinaison des *matchings* résultants. Ceci constitue une meilleure pratique. Dans le cas d'un nettoyage de noms et d'adresses, nous conseillons de toujours configurer sur la base de la dénomination (3 à 4 caractères : premier caractère + consonnes non répétées suivantes) et sur la base des éléments de l'adresse (ex. 4 caractères : 2 caractères numériques du code postal suivis des 2 premiers caractères du nom de la commune).

Plusieurs routines de comparaison champ par champ pour le même champ

L'approche déterministe avec les tables de vérité permet de faire intervenir plusieurs critères sur le même champ pour décider si la comparaison par paire entre deux enregistrements est un *match*, un *suspect match* ou un *non match*. Cette approche est très performante et fait partie des atouts des outils de qualité des données.

Les Figure 43 jusqu'à Figure 46 illustrent clairement cet aspect, à l'aide d'un exemple concernant des noms d'entreprises. Il s'avère ici très important aussi de disposer d'une routine de comparaison alternative qui puisse composer avec des successions de mots. Dans la pratique, cela peut être utile lors d'une recherche, d'un *clustering* et/ou d'une comparaison de noms d'entreprises qui sont des compositions ou des dénominations de compositions d'entreprises, telles qu'une Tijdelijke Handelsvennootschap (TH) / Société Momentanée (SM).

[86] Un exemple de tailles de *windows* et du temps de traitement y afférent se trouve en annexe 6.4.

Grade Pattern Editor - C:\Documents and Settings\DRV\Application Data\Trillium Software\TssUI\1

Category	Pattern ID	business_name	bus_sort	bus_squish	street_name	house_number
P	110	A	-	-	A	A
P	115	A	-	-	A	B
P	120	A	-	-	A	C
P	125	A	-	-	B	A
P	130	A	-	-	C	A
P	400	A	-	-	A	-
P	401	-	A	-	A	-
P	402	-	-	A	A	-
P	135	B	-	-	A	A
P	140	-	-	B	A	B
P	145	B	-	-	B	A

Field List Editor - C:\Documents and Settings\DRV\Application Data\Trillium Software\TssUI\11.5\c

Description	Score A	Sc B	Sc C	D	E	Comparison R...	Field Name 1	Routine M...
business_name	100	90	80			busname	PR_BUSNAME_RECODED_01	
bus_sort	95					BUSNAME	Pr_busname_recoded_01	SORT
bus_squish	95	75				SPELLING	Pr_busname_recoded_01	SQUISH
street_name	95	87	82			streets	TQ_GOUT_STREET_NAME	
house_number	100	98	90			houseno	TQ_GOUT_HOUSE_NUMBER	
box_number	100					absolute	PR_BOX1_NUMBER	
[illegible]	100	95	70			BUSNAME	TQ_GOUT_POSTAL_CODE	

Figure 43 : Trois critères de similarité alternatifs (business_name, bus_sort et bus_squish) définis pour un même champ (Pr_busname_recoded_01).

Ce champ est une forme déjà standardisée de la dénomination originale (Fst Denom Lnm) : les signes de ponctuation et autres ont été supprimés. Faites surtout attention à l'alternative avec Pattern_ID égale à « 401 ».

Lev1 Matched	Lev1 ...	Fst Denom Lnm	Tsq Street	Tsq Hsno	Tsq Box	Tsq City	Postcode
00001655	401	Konstiak Matus	DRAZKOVCE	165		MARTIN	038 02
00001655	401	Matus Konstiak	DRAZKOVCE	165		MARTIN	038 02
00001954	110	Cizmar Jarislav	FEKISOVCE	29		SOBRANCE	072 33
00001954	110	Cizmar Jarislav	FEKISOVCE	29		SOBRANCE	072 33
00001954	135	Cizmar Jarislav□/nIdentificatienr. 1410001151...	FEKISOVCE	29		SOBRANCE	072 33
00001954	401	JARISLAV CIZMAR	FEKISOVCE	29		SOBRANCE	07233
00002271	110	FIRMA INTREX	NOWA	6	35	GOSTYNIN	09-500
00002271	110	Firma Intrex	NOWA	6	35	GOSTYNIN	09500
00002271	135	FIRMA INTREX□/n□/nBTW : PL 9 710 270 624	NOWA	6	35	GOSTYNIN	09500
00002271	401	intrex firma	NOWA	6	35	BOSTYNIN	09-500
00003506	401	F.HANSSENS-ENSCH AVOCAT	AVENUE LOUISE	349	17	IXELLES	1050
00003506	401	HANSSENS F. - ENSCH AVOCAT	AVENUE LOUISE	349	17	ELSENE	1050
00005043	135	csd go sport sa	CHAUSSEE DE B...	6		IXELLES	1050
00005043	135	*CSD GO SPORT SA = 7244587-58	CHEE DE BOOND...	6		IXELLES	1050
00005043	401	* GO SPORT CSD * ***...	BOONDAALSEST...	6		BRUSSEL	1050
00005209	401	LOUIS DE WAELE SA - VALENS SA SM...	AVENUE BRUGM...	27		ST GILLES	1060
00005209	401	VALENS SA-LOUIS DE WAELE SA SM ...	AVENUE BRUGM...	27		ST GILLES	1060
00008770	135	SIESA SA - SOCIETE IMMOBILIERE ET D'E...	AV FRANKLIN RO...	160		BRUXELLES	1050
00008770	135	*SOCIETE IMMOBILIERE ET D'ENTREPRIS...	AV FR ROOSEVE...	160		BRUXELLES	1050
00008770	401	SOCIETE IMMOBILIERE ET D ENTREPRIS...	AV FR ROOSEV...	160		IXELLES	1050
00010713	401	BOSCH ROBERT NV ...	RUE HENRI GENE...	1		ANDERLECHT	1070
00010713	401	ROBERT BOSCH SA ...	RUE H-JOSEPH G...	1		ANDERLECHT	1070

Figure 44 : Importance du match type « 401 ».

Il est en mesure de trouver des matches lorsque l'ordre des mots n'est pas identique.

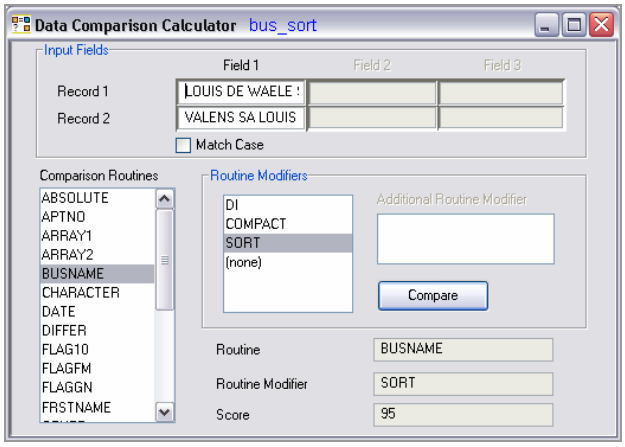


Figure 45 : Routine faisant partie du critère « bus_sort » du match type « 401 ».

Le critère de comparaison « bus_sort » utilise la routine BUSNAME (modifiant : SORT) et génère un score de 95 lors de la comparaison entre « LOUIS DE WAELE SA - VALENS SA SM » et « VALENS SA-LOUIS DE WAELE SA SM ». Ceci permet d'inclure un match type *fiable dans la table de vérité ci-dessus.*

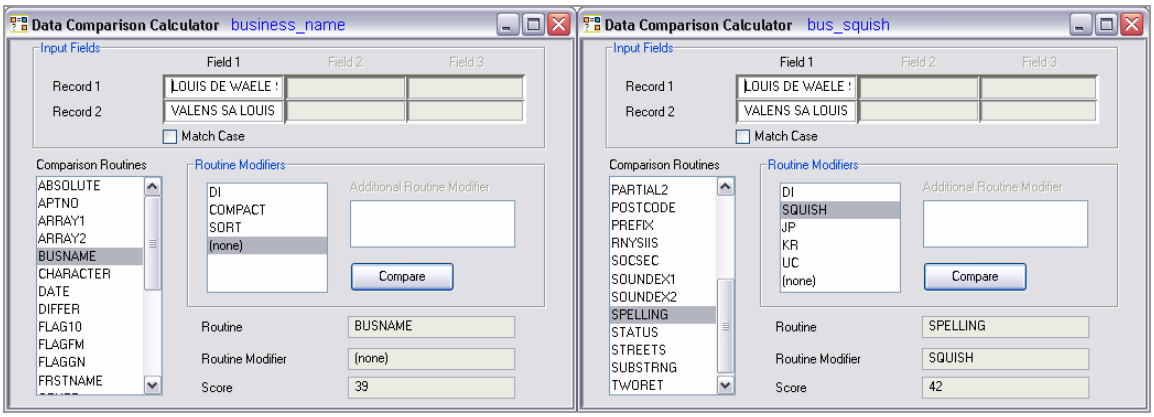


Figure 46 : Routines alternatives défaillantes en face de ce type d'imprécision.

Lles routines de comparaison alternatives ne sont pas à même de trouver les matches du type « 401 ». Les scores de similarité atteints avec les autres routines sont en effet bien trop faibles (39 et 42). Même avec les seuils de score moins élevés B et C sur les critères business_name et bus_squish (voir par exemple les Pattern_ID 135 et 140), on ne trouverait pas ces matches.[87] *En outre, lorsque les seuils ne sont pas suffisamment élevés, on obtient un trop grand nombre de faux positifs.*

Développement interne

Vu la **flexibilité et la performance** qu'offrent les outils de qualité des données, nous déconseillons d'essayer de développer soi-même toutes ces fonctionnalités. En effet, Gartner[88] **déconseille explicitement** un développement « home made » pour remplir les fonctionnalités des outils de qualité des données, les meilleurs outils présents sur le marché ayant capitalisé un vaste ensemble de règles réutilisables et fournissant des bases de données d'adresses internationales difficilement accessibles par une seule organisation. Gartner estime que les organisations les plus matures ont acquis un outil de qualité des données en 2010.

[87] En définissant trop bas les seuils, on obtient en outre un trop grand nombre de faux positifs.

[88] BEYER M. A., FEINBERG D., FRIEDMAN T. ET THOO E., *Hype Cycle for Data Management, 2010*, Gartner, 22 juillet 2010.

3.4. Exemples d'application des data quality tools

Dans ce paragraphe, nous présenterons quelques exemples d'application typiques, qui montrent dans quel contexte les outils de qualité des données offrent réellement une plus-value.

3.4.1. Standardisation et matching d'adresses ; *cleansing*

Dans un premier exemple, nous illustrerons **l'amélioration de la qualité des adresses**. En effet, en plus des nombreux algorithmes et méthodes de standardisation et de *matching* présentés aux section 3.2 et 3.3, les outils de qualité des données disposent aussi généralement de riches **bases de connaissances** avec des milliers de règles pour le nettoyage de noms et d'adresses, **axées sur des régions spécifiques**. Dans l'illustration ci-dessous, on peut voir comment la qualité de données *legacy* d'une source authentique a pu être améliorée sensiblement.

C Postcode	Tq Gout Postal Code	Straatnaam Voll	Tq Gout Street Name	Huisnummer	Pr House N...	Gemeentenaam	Tq Gout Postal City
1020	1020	RUE E VANDER AA	RUE ERNEST VANDER AA	1	1	Brussel	BRUSSEL
1020	1020	rue Vander Aa	RUE ERNEST VANDER AA	3	3	Bruxelles	BRUXELLES
1050	1050	91 R VAN AA	RUE VAN AA	_	91	Elsene	ELSENE
1050	1050	27 R.VAN AA	RUE VAN AA	_	27	Elsene	ELSENE
1050	1050	RUE VAN AA	RUE VAN AA	94	94	Ixelles	IXELLES
1050	1050	RUE VAN AA	RUE VAN AA	94	94	Elsene	ELSENE
1020	1050	rue Van Aa	RUE VAN AA	2	2	Bruxelles	IXELLES
1050	1050	2 R VAN AA	RUE VAN AA		2	Ixelles	BRUXELLES
1000	1000	R JOSEPH II 40	RUE JOSEPH II		40	Bruxelles	BRUXELLES
1000	1000	rue Joseph II 71 (...	RUE JOSEPH II		71	Bruxelles	BRUXELLES
1040	1000	Rue Joseph II	RUE JOSEPH II	71	71	Brussel	BRUSSEL
1040	1000	Rue Joseph II 5-7	RUE JOSEPH II		5-7	Bruxelles	BRUXELLES
1040	1000	Rue Joseph II 67A	RUE JOSEPH II		67A	Bruxelles	BRUXELLES
1030	1000	rue JOSEPH II, 114 -	RUE JOSEPH II	116	114 - 116	Schaarbeek	BRUXELLES

Figure 47 : Nettoyage d'adresses (standardisation et matching), montré sur des données réelles.

*A l'aide des bases de connaissance du logiciel, le code postal est corrigé,le nom de rue est standardisé, les éléments de l'adresse sont correctement répartis (*parsing*), le nom de commune est corrigé, les doublons sont détectés et organisés en grappes. On voit tant les données originales que les suggestions de corrections émises par l'outil de data quality.*

L'exemple de la Figure 47 illustre clairement l'importance de la standardisation du contenu des champs à comparer dans le cadre d'un fuzzy matching.

L'utilisation d'outils de qualité des données a permis dans ce projet d'améliorer la qualité **sans le moindre développement**. En outre, les outils de qualité des données ont fourni les résultats (suggestions de corrections) sans toucher aux données originales et ont **ajouté une série de métadonnées** indiquant :

- dans quels cas la suggestion de correction diffère des données originales ;
- la fiabilité de la suggestion de correction ;
- une typologie des problèmes si les règles de l'outil n'autorisent pas une correction univoque.

De telles métadonnées permettent d'**exploiter aisément** les résultats des outils de qualité des données et de les **intégrer** dans des systèmes de gestion de l'historique des anomalies ainsi que dans des processus métier.

3.4.2. Matching et détection des incohérences entre deux sources

Un deuxième exemple mérite une attention toute particulière, car il illustre comment les outils de qualité des données peuvent offrir une solution lors de la recherche de **relations entre des banques de données entre lesquelles il n'existe pas de relation de clé (fiable)**. Pensez à la **détection de fraude**, aux **contrôles d'exhaustivité**, à la « **identity resolution** », … Un large spectre d'anomalies de types divers peut utiliser des outils de qualité des données pour s'attaquer à ce genre de problèmes (Figure 48).

Source	Dénomination	Adres	Boite	Postcd	Commune	cdpays	Match Type
L	PROJEKTSERWIS WANDA LITTY	BOHATEROW MODLINA 63	43	05-100	NOWY DWOR MAZ	122	106
L	PROJEKTSERWIS WANDA LITTY	BOHATEROW MODLINA 63	43	05-100	NOWY DWOR MAZ	122	106
R	PROJEKT SERWIS (LUTY WANDA)	UL. BOHATEROW MODLINA 63	42	05-100	NOWY DROW MZAOWIE	PL	115
R	PROJEKT SERWIS LUTY WANDA NOWY DWOR	UL BOHATEROW MODLINA 63	43	05-100	NOWY DWOR MAZOWIE	PL	115
R	PROJEKT SERWIS LUTY WANDA	BOHATEROW MODLINA 63/43		05-100	NOWY DWOR MAZ	PL	137
R	PROJEKT SERWIS WANDA LUTY	BOHATREROW 63LOK	43	05-100	NOKY DWOR MAZ	PL	106
R	PROJEKT SERWIS WANDA LUTY	BOHATREROW MODLINA 63LOK	43	05-100	NOKY DWOR MAZ	PL	138
R	PROJEKT SERWIS WANDA LUTY	BOHATREROW MODLINA 63LOK	43	05-100	NOKY DWOR MAZ	PL	138
R	PROJEKT SERWIS	NOWY DWOR MAZOWIECKI 63		05-100	NOWY DWOR MAZOWIE	PL	137

Figure 48 : Le Fuzzy Matching et la détection des incohérences (entre deux sources : L et R), montrés sur des données réelles.

Pour illustrer le principe, une grappe de doublons est affichée. Elle démontre le lien qui peut être établi, sans clé, entre deux registres.

Étant donné que les deux sources confrontées n'ont pas le même schéma de banque de données, des routines pour la transformation des sources ont également été configurées dans l'outil de qualité des données. Ceci illustre la capacité **d'intégration de données** des outils de qualité des données. De la même manière, il est possible de transformer les résultats des outils de qualité des données en un schéma souhaité avant de les exporter, de manière à pouvoir largement soutenir les **migrations de données**.

3.4.3. Outil utilisé – Trillium Software

Depuis fin 2009, Smals dispose d'outils de Data Quality. En 2008, le cahier des charges a été publié en deux phases puis testé de manière extensive. La solution qui a été choisie est le Trillium Software System (TS Discovery et TS Quality) de Trillium Software (voir également Gartner Magic Quadrants pour les outils de Data Quality des dernières années).

C'est avec cet outil que tous les projets, tous les cas de figure, et tous les exemples mentionnés dans ce rapport, ont été réalisés.

3.5. Organisation

La Figure 49 montre comment peut se présenter l'organisation associée aux Data Quality Tools dans le cadre de la gestion des anomalies. L'organisation proposée est **générique** : elle est donc valable tant pour des projets locaux que pour de plus vastes programmes liés à une institution, tant pour le reengineering d'applications que pour les projets d'intégration et de migration de données, de détection de doubles et d'incohérences, de lutte contre la fraude, des initiatives de standardisation et de la gestion des anomalies. Et ce, tant pour les **contrôles de qualité on-line** que pour les **projets batch**.

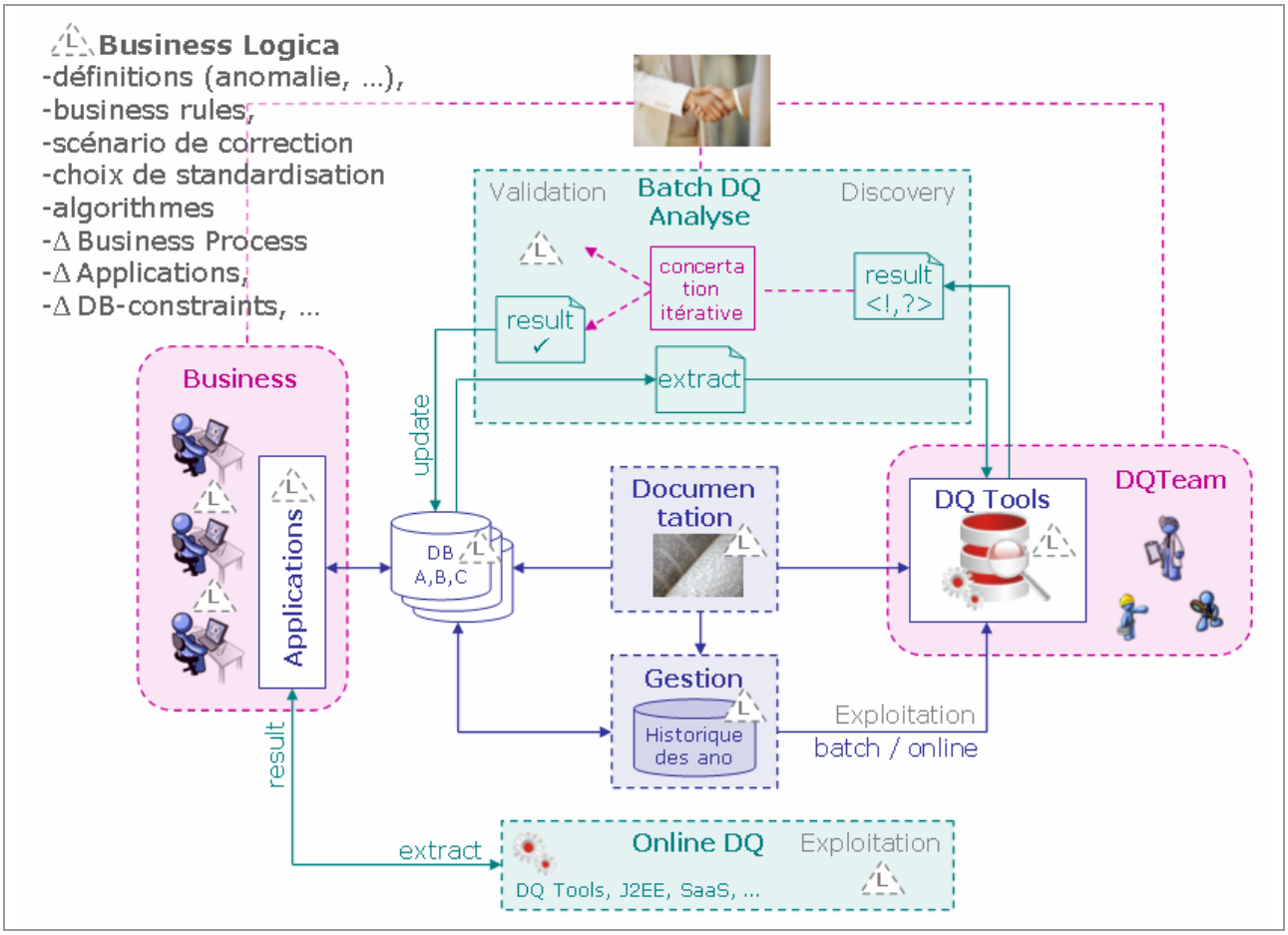


Figure 49 : Organisation pour l'exploitation des Data Quality Tools

La Figure 49 sert aussi à démontrer le **processus collaboratif** en vue d'améliorer la qualité d'une base de données ou de définir, de détecter, ou de traiter les anomalies.

Processus découpé en phases

Discovery

En premier lieu, les Data Quality Tools pourront jouer un rôle de support durant les phases **d'analyse** et **d'inception** des projets. Ceci rejoint ce que nous avons appelé dans la section sur le Data Profiling (3.1.2) la « phase de découverte » (voir Figure 49 : « Discovery »).

L'équipe Data Quality reçoit un extract, qui est le plus complet et le plus récent possible. Celui-ci est analysé, avec l'aide des Data Quality Tools sur le plan du

Data Profiling, de la standardisation des données et du data matching. Ceci peut être réalisé de manière exhaustive ou être orienté vers des problématiques « business » spécifiques liées à la qualité des données.

Le résultat consiste initialement en un ensemble de découvertes quant à la qualité des données, comme le manque de standardisation, le non-respect de contraintes d'intégrité, la présence de valeurs et de patrons non documentés, le non-respect de l'intégrité référentielle, la détection de doubles ou d'incohérences avec les types initiaux de match (Figure 40), etc. Ceci ne représente que quelques exemples d'anomalies. Dans certains cas, les Data Quality Tools permettent également de proposer des solutions immédiates de corrections relatives aux problèmes détectés.

Validation

Dans la phase de validation, les analystes montrent les résultats provisoires aux spécialistes de la connaissance « business », lesquels donnent un feedback. Ce feedback peut consister en :

- une validation des cas découverts initialement ;
- une adaptation orientée des découvertes initiales.

Les phases de « discovery » et de validation peuvent être itératives, dans le cas où le feedback des spécialistes du domaine donne lieu à de nouvelles phases de « discovery », dans une sorte de « ronde de validation ». De cette façon, le processus de réflexion itératif entre les spécialistes du domaine et les analystes est idéal en vue de soutenir la phase de définition des spécifications des projets.

Le résultat final de la phase de réflexion itérative consiste toujours en deux points :

- d'un côté, des résultats validés, qui peuvent être potentiellement immédiatement exploités ;
- d'un autre côté, des décisions relatives à la connaissance et au business, telles que : définition des anomalies et des contraintes d'intégrité, choix sur le plan de la standardisation, algorithmes et « match types » fiables et adaptés, scénarios de correction, processus business, applications et schémas de banques de données à adapter.

Le but est donc l'explicitation de la logique business (qui souvent existe seulement sous une forme implicite) et l'obtention d'estimations, de décisions stables concernant les problèmes business (relatifs à la qualité des données et à d'autres points).

Exploitation

Les *résultats validés* peuvent être immédiatement exploités, par exemple, pour traiter des données de production (Figure 49 : « update »), afin d'en améliorer la qualité. Ceci se produit typiquement dans des projets de « qualité de données » purs.

La *connaissance* et *la logique business validées* peuvent être considérées comme les résultats de l'analyse fonctionnelle relative aux spécifications. La connaissance obtenue est en première instance exploitée dans les phases suivantes du développement. L'exploitation en production peut se réaliser tant en batch qu'on-line, et ceci tant en faisant usage de Data Quality Tools qu'en utilisant un développement propre (tant que celui-ci respecte la logique business validée). Bien entendu, il s'agit de la même logique qui sera illustrée dans le système documentaire.

Dans le cas spécifique et particulier d'une gestion des anomalies, les définitions validées d'anomalies et de scénarios de correction sont exploitées dans le système de gestion de l'historique des anomalies, pour la détection et le

traitement des anomalies. Là également, on a le choix d'utiliser des Data Quality Tools ou de procéder à un développement propre en batch ou on-line.

Les rôles

Comme annoncé, l'approche intégrée proposée de gestion des anomalies est interdisciplinaire et la bonne collaboration des différentes équipes et fonctions est cruciale pour le succès du projet.

Equipe Data Quality

L'équipe Data Quality (Figure 49 : « DQTeam ») doit comprendre le mieux possible la problématique « business », analyser les sources de données requises et prendre en charge l'étude de leur qualité. L'équipe doit aussi gérer les règles, les algorithmes et les match types nécessaires ainsi que les adapter aux besoins du business selon l'input des spécialistes du business. Ceci se réalise en concertation itérative avec les spécialistes du business.

L'équipe doit au moins inclure les rôles d'analyste fonctionnel, d'analyste technique et de spécialiste « data quality tools ».

Equipe Business

Du côté du business sont au moins impliqués les rôles et responsabilités suivants :

- le management qui définit stratégiquement quelles problématiques méritent d'être traitées, avec quelle priorité et par qui ;
- les spécialistes du business, qui sont désignés dans le cadre de la concertation itérative avec l'équipe Data Quality, lesquels traduisent les spécifications stratégiques à destination des analystes et qui donnent à ces derniers le feedback requis de sorte qu'ils puissent adapter les règles, algorithmes et match types requis vers les vrais besoins du business. Si une interprétation de la loi est nécessaire, des juristes doivent également intervenir. Les spécialistes du business désignés décident, en accord avec le management, comment les choix doivent être orientés sur le plan de la standardisation.

Autres fonctions de support

La liste ci-dessus n'est pas exhaustive : en fonction du type de projet, d'autres rôles peuvent être nécessaires, comme : administrateur de base de données, gestionnaire d'information, chef de projet, architecte et développeur, spécialiste middleware et helpdesk.

4. Conclusions

Les enjeux de l'analyse critique de l'information sont stratégiques pour une entreprise ou une administration. L'étude a montré comment une telle analyse pouvait contribuer à en améliorer la qualité à l'aide d'une organisation globale structurée à cette fin, d'une méthode fonctionnelle et conceptuelle et de techniques, tels les Data Quality Tools.

Par ailleurs, parmi les facteurs à l'origine des échecs en matière d'intégration de données, on trouve la mise en place d'approches trop ambitieuses, pas assez contrôlées sur le plan organisationnel et trop peu soutenues par le management. À travers le prisme de la question ciblée du traitement des anomalies, nous nous sommes dès lors efforcés de proposer des pistes concrètes, bien ancrées dans une architecture pratique. Ces pistes s'inscrivent dans la continuité des recommandations du DQ Competency Center (consultances et études). Parmi les enseignements novateurs et opérationnels de l'étude, rappelons les points suivants (chacun étant accompagné d'un modèle organisationnel adapté) :

- La modélisation générique de l'historique des anomalies en vue de déployer une stratégie de gestion permettant d'en diminuer le nombre. À partir d'une méthode conceptuelle originale dont nous avons rappelé les principes, il s'agit de passer de « l'hypothèse du monde clos » à celle du « monde ouvert sous contrôle ». Cette dernière permet le traitement d'une réalité empirique dont les valeurs doivent faire l'objet d'une interprétation humaine en vue d'être validées. Un exemple de prototype opérationnel a été présenté en vue de montrer que ce modèle d'historique pouvait faire l'objet d'un développement adapté à nos bases de données administratives. En fonction du contexte, il devra être complété dans la pratique par un travail d'analyse spécifique. Par ailleurs, l'impact fonctionnel et organisationnel, sur le plan des ressources disponibles par exemple, devra être pris en compte en vue d'envisager comment un modèle sera implémenté dans une situation spécifique.
- Les apports des « data quality tools », sur la base d'une expérience pratique, ont été présentés. Ces outils constituent un software spécialisé applicable « en batch » et « on line » en vue de la détection et de la correction des anomalies. « En batch » par ailleurs, ces outils sont très performants en tant qu'instruments d'aide à l'analyse, quelle que soit la fréquence des « change requests » émis par les utilisateurs en vue de définir les notions délicates et complexes de « doubles », « d'incohérence », etc.
- Un modèle de gestion des connaissances en vue d'accompagner et de faciliter la correction des anomalies par les agents de l'administration, sur la base de définitions homogènes et pratiques. Ce modèle a fait ses

preuves sur le terrain et peut être généralisé à une grande variété de situations dans le domaine de l'egovernment.

Parmi les perspectives à investiguer dans le futur, citons :

- Le suivi du marché «Data Quality Tools » en ce qui concerne :
  - l'approche « Data Quality as a Service », même si cette architecture soulève des problèmes de sécurité et de confidentialité, éléments sensibles dans le cadre de l'administration électronique.
  - les outils « open source » et initiatives, comme *Google Refine*.
- L'application d'une approche « data quality », non seulement aux bases de données relationnelles, mais aussi aux modèles émergents :
  - les « *NoSQL Databases* »[89] dont la structure répond à des critères de performance et s'applique essentiellement à de vastes bases de données non structurées, diffusées sur le web. Elles incluent la notion de « Large Object » (LOB) pour le stockage des documents et s'inspirent des OODBMS des années 1990, qui n'ont, alors, pas connu de réel succès. Dans le cadre de bases de données administratives « multi-utilisateurs », telles qu'envisagées dans notre domaine d'application, ce nouveau type de bases de données pose un problème de consistance car il ne respecte pas le modèle « ACID » (Atomicity, Consistency, Isolation, Durability) permettant de maintenir l'intégrité des données via la gestion des transactions. Ce type de base de données peut ainsi avoir un impact négatif sur les efforts en matière d'intégration de données. Il mérite toutefois notre attention.
  - Les « *Column-Store Database Management Systems* », dont les fonctionnalités de compression répondent à des critères de performance et qui sont notamment conseillées dans le cadre des « Data Marts » et des stratégies d'archivage en général.[90]
- Un suivi des développements en matière d'ontologies appliquées aux métadonnées, en tant que support à l'approche « *data quality* », la documentation des bases de données étant stratégique en vue d'en assurer la qualité. Nous l'avons vu dans ce rapport avec l'exemple des glossaires de la sécurité sociale ou de l'application Falco. Une première étude a été menée en ce qui concerne les normes du web sémantique, incluant les ontologies et les standards RDF et OWL[91]. L'étude avait soulevé la richesse des ontologies mais aussi, en corollaire, leur complexité et la lourdeur en termes de ressources humaines et d'input intellectuel que requièrent leur conception et leur maintenance. Toutefois, si cet arbitrage demeure intact, le domaine mérite d'être suivi. Depuis lors, la norme SKOS du W3C (reposant sur la syntaxe d'une ontologie) est apparue en vue d'intégrer en un seul langage des langages documentaires préexistants hétérogènes.[92] En effet, la tendance actuelle va dans le sens d'une atténuation de la différence entre données structurées et données non structurées[93] et les ontologies sont conçues

[89] BEYER M-A. et al., *Op. cit*., p. 7-8.
[90] BEYER M-A. et al., *Op. cit*., p. 25.
[91] BOYDENS I., Du Web sémantique au web pragmatique, *Research Note*, n° 4, Section Recherches, Smals, Bruxelles, 2004.
[92] DRAMAIX C., *SKOS (Simple Knowledge Organization System) : apports et limites pour la conversion des thésaurus dans le cadre du Web sémantique*, Mémoire de fin de Master en Sciences et technologies de l'information et de la communication, Bruxelles, ULB, 2008-2009.
[93] BEYER M-A. et al., *Op. cit.*, p. 15.

dans ce sens. De manière générale, les « *Metadata Repositories* » sont stratégiques en vue d'accompagner la gestion des systèmes d'information (à de nombreux niveaux : description sémantique des données, description des composantes logicielles…). Mais les échecs dans ce domaine furent fréquents à ce jour, notamment en raison d'un manque d'attention accordée à la gestion de projets et à l'organisation. Parfois, l'échec est dû à la mise en place de projets trop ambitieux, comme l'avait déjà relaté la NASA, dans son retour d'expérience datant de la fin des années 1990, « *the meta-data myth* ».[94]

- Une approche novatrice de type « *collaborative data quality* », dont nous exemplifions les grandes lignes. L'environnement de l'administration fédérale inclut plusieurs sources authentiques fondamentales pour l'identification des personnes physiques et morales (par exemple, le Registre National, la BCE...). Ces sources comportent des problèmes de qualité (adresses erronées, doublons...). Elles sont elles-mêmes alimentées par différents sources ou initiateurs et sont exploitées, en tant que sources authentiques, par un grand nombre d'institutions fédérales (SPF, parastataux...). À l'heure actuelle, chaque institution traite la qualité de ces sources de manière partielle et isolée. L'objectif de l'approche serait de proposer une organisation et une structure d'échange des anomalies et de leurs corrections collaboratives et « partagées » entre tous les interlocuteurs afin de rendre plus efficaces l'évaluation et l'amélioration de la qualité des données. L'étude aurait pour objet d'examiner la faisabilité d'une telle approche à travers un prototype (sur les composantes de la donnée « adresse » de la BCE, par exemple, en raison de son impact stratégique), permettant :
  - l'échange partagé entre les interlocuteurs concernés (source authentique et échantillon d'utilisateurs) des anomalies, des propositions de corrections associées et les métadonnées qui en permettent l'interprétation (via un modèle structuré d'historique des anomalies) ;
  - la communication de décisions validées par l'institution « source authentique » concernant la correction des anomalies détectées par la source authentique ou par les institutions utilisatrices (certaines corrections que ces dernières effectuent ne seront pas validées officiellement car elles seront trop « ad hoc ») ;
  - la détection, par les « data quality tools », de patterns d'anomalies récurrents ou spécifiques, et la communication de ceux-ci, de manière à faciliter les investigations qui permettent d'y remédier à la source.

  Sur la base de ce prototype et de ses résultats, un ensemble de recommandations générales pourraient être identifiées et présentées sur les plans organisationnel, conceptuel et technique.

[94] BOYDENS I., *Documentologie*, Bruxelles, Presse de l'Université Libre de Bruxelles, 2010-2011 (syllabus, dernière édition).

5. Bibliographie

BEYER M. A., FEINBERG D., FRIEDMAN T. et THOO E., *Hype Cycle for Data Management, 2010*, Gartner, 22 juillet 2010.

BONTEMPS Y., BOYDENS I., VAN DROMME D., *Data Quality : tools*, Deliverable, 2007/trim3/02, Smals, Section Recherches, Bruxelles, 2007.

BOYDENS I., Evaluer et améliorer la qualité des bases de données*, Techno*, n°7, Section Recherches, Smals, 1998.

BOYDENS I., *Informatique, normes et temps*, Bruylant, Bruxelles, 1999.

BOYDENS I., « Déploiement coopératif d'un dictionnaire électronique de données administratives », *Document Numérique*, Hermes, Paris, 2001, , vol. 5, n°3-4, p. 27-43.

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

BOYDENS I., « E-gouvernement en Belgique. Un retour riche d'expériences », *L'informatique professionnelle (Dossier spécial "Services publics")*, Editions Gartner France, Paris, Numéro 217, octobre 2003, p. 29-35.

Boydens I., La préservation à long terme de l'information numérique. Techno 28 – 09/2004, Bruxelles : Smals, Sections Recherches, 2004.

BOYDENS I., Du Web sémantique au web pragmatique, *Research Note*, n° 4, Section Recherches, Smals, Bruxelles, 2004.

BOYDENS I., *Data Quality : Best Practices*, Deliverable, 2006/trim2/01, Smals, Section Recherches, Bruxelles, 2006.

BOYDENS I., « Qualité de l'information et e-administration : enjeux et perspectives » dans ASSAR S., BOUGHAZALA I., *Administration électronique : constats et perspectives*, Paris, Hermès, 2007, p. 103-120, chapitre 5.

BOYDENS I., *Documentologie*, Bruxelles, Presse de l'Université Libre de Bruxelles, 2010-2011 (syllabus, dernière édition).

BOYDENS I., « Strategic Issues Relating to Data Quality for E-government: Learning from an Approach Adopted in Belgium », dans ASSAR S., BOUGHAZALA I., BOYDENS I., eds., *Practical Studies in E-Government*, Springer, 2011, p. 113-130 (chapitre 7).

BOYDENS I., « Hiérarchie et anarchie : dépasser l'opposition entre organisation centralisée et distribuée ? » dans HUDON M. et El HADI W. M., éds, « Organisation des connaissances et Web 2.0 », *Les cahiers du numérique*, Paris, Hermès Sciences, vol. 6, n° 3, 2010, p. 77-101.

BOYDENS I. et VAN HOOLAND S., « Hermeneutics applied to the quality of empirical databases », *Journal of documentation*, Emerald, vol. 67, issue 2, 2011, p. 279-289.

BRASSEUR C., *Data Management. Qualité des données et compétitivité*, Hermès – Lavoisier, Paris, 2005.

CARLIER A., *Management de la qualité pour la maîtrise du système d'information*, Hermès - Lavoisier, Paris, 2006.

CINQUIN L., « La qualité, ennemi juré de la productivité ? », *O1 Informatique*, 16 janvier 2006, n°1841.

COHEN W., RAVIKUMAR P. et FIENBERG S., « A comparison of string distance metrics for name-matching tasks ». In *The IJCAI Workshop on Information Integration on the Web (IIWeb)*, Acapulco, 2003.

DRAMAIX C., *SKOS (Simple Knowledge Organization System) : apports et limites pour la conversion des thésaurus dans le cadre du Web sémantique*, Mémoire de fin de Master en Sciences et technologies de l'information et de la communication, Bruxelles, ULB, 2008-2009.

ELMASRI R., NAVATHE S., *Fundamentals of Database Systems (fifth edition)*, Adison-Wesley, Boston, 2007.

FRIEDMAN B. et NISSENBAUM H., « Bias in Computer Systems », *ACM Transactions on Information Systems*, juillet 1996, vol. 14, n° 3, p. 330-347.

GARAGNON J., « Sirène, système informatique pour le répertoire des entreprises et des établissements. Situation actuelle et développements en cours », *Courrier des statistiques*, janvier 1983, n° 25.

HULSTAERT A., *Les systèmes documentaires en ligne dans le domaine de la presse écrite. Conception d'une grille d'analyse de la qualité des sources et confrontation à la mise en place d'un système*, Mémoire de fin de Master en Sciences et Technologies de l'information et de la communication, Bruxelles, ULB, 2006-2007.

HULSTAERT A., *Préserver l'information numérique. Codage et conversion de l'information,* Deliverable, 2008/trim2/02, Smals, Section Recherches, Bruxelles, 2008.

HULSTAERT A., *Préservation à long terme de l'information numérique. Rendre l'information accessible durablement*, Deliverable, 2010/trim1/01, Smals, Section Recherches, Bruxelles, 2010.

KRISCHAUSKY D., « Problèmes généraux posés par l'utilisation des technologies de l'information dans la sécurité sociale », dans *L'innovation dans les technologies de l'information : élément important du développement futur des systèmes de sécurité sociale. Huitième conférence internationale sur l'informatique dans la sécurité sociale*, Berlin, 22-24 octobre 1996, Editions de l'Association Internationale de la Sécurité Sociale, Genève, 1997, p. 239-252.

MADNICK S. E., WANG R.-Y., YANG W.-L, HONGWEI Z., « Overview and Framework for Data and Information Quality Research », *Journal of Data and Information Quality*, Vol. 1, No. 1, 2009, p. 2-22.

NAYER A., BORRENS G. et BALTAZAR-LOPEZ S., *L'inspection du travail et la protection juridique du citoyen*, La charte, Brugge, 1995.

NEWMAN D. et FRIEDMAN T., Data Integration is Key to Successful Service-Oriented Architecture Implementations, *Gartner Research Note*, 12 octobre 2005.

OGONOWSKY G., *Business Rules Technologies & web sémantique : La gestion des règles business*, Delivrable, 2008/trim1/01, Smals, Section Recherches, Bruxelles, 2008.

OLSON J., *Data Quality: The Accuracy Dimension*, Elsevier, Burlington, 2002.

PADMAN R., « Quality Metrics for Healthcare Data : an Analytical Approach », dans STRONG D. M. et KAHN B. K., éds, *Proceedings of the 1997 Conference on Information Quality*, M.I.T, Cambridge, 1997, p. 19-36.

PEAUCELLE J.-L., *Informatique rentable et mesure des gains*, Hermès, Paris, 1997.

REDMAN T., *Data Quality for the Information Age,* Artech House, Boston, 1996.

REDMAN T., *Data Quality : The Field Guide*, Digital Press, Boston, 2001.

REDMAN T., « Improve Data Quality for Competitive Advantage », *Sloan Management Review*, winter 1995.

RIVIERE P., « Qualité des données et processus de recueil », Conférence présentée au CNAM, Conservatoire National des Arts et Métiers, lors des journées d'études CNAM-CSML, *La qualité des données à l'ère de l'information*, Cnam, Paris, 11-12 mars 2003.

RIVIERE P., « Approche coût-qualité pour l'amélioration des processus de production statistique », *Courrier des statistiques*, juin 2003, n° 105-106.

RIVIERE P., « Indicateurs de qualité en matière de production de données : quelques éléments de réflexion », *Courrier des statistiques*, septembre 2005, n° 115, p. 35-40.

TRIGAUX J.-C., *Master Data Management - Mise en place d'un référentiel de données*, Deliverable, 2009/trim4/01, Smals, Section Recherches, Bruxelles, 2009.

VAN DER VLIST E., *Relax NG*, Cambridge, O'Reilly Media, 2003.

VAN HOOLAND S., KAUFMAN S., BONTEMPS Y., « Answering the call for more accountability: applying data-profiling to museum metadata », *Proceedings of the 2008 International conference on Dublin Core and metadata applications*, Berlin, 22- 26 Septembre 2008, p. 93-103.

VAN HOOLAND S., *Metadata quality in the cultural heritage sector: stakes, problems and solutions*, Thèse de doctorat, Université Libre de Bruxelles, 2009.

WANG R. Y., LEE Y. W. et STRONG D. M., « Can you Defend Your Information in Court ? » dans WANG R. Y., éd., *Proceedings of the 1996 Conference on Information Quality*, M.I.T., Cambridge, 1996.

WINKLER W. E., Overview of Record Linkage and Current Research Directions, 2006. http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf

6. Annexes

6.1. Workflow de corrections des anomalies

Dans le cadre d'une mission de consultance pour la sécurité sociale que nous avions menée sur la gestion des anomalies, il nous avait été demandé de réfléchir à la possibilité de mettre en place un workflow de traitement des anomalies.

Les données visées par l'étude sont utilisées par la sécurité sociale pour le calcul et l'attribution des droits sociaux des travailleurs. Le nombre d'anomalies (plusieurs centaines de milliers chaque trimestre) nécessite de nombreuses corrections (tant par le citoyen que par la sécurité sociale) consommatrices de temps et de ressources. Une diminution des anomalies ou une optimisation des corrections permettraient de réduire fortement les ressources mobilisées.

L'enjeu est donc stratégique, aussi bien en termes de droits des travailleurs qu'en termes de gestion des ressources humaines et financières au sein des institutions en charge de la correction des anomalies.

L'étude avait donc pour objet d'examiner la possibilité de mettre en place une séquence de corrections des anomalies et les critères à prendre en compte pour ce faire, tout en soulevant les difficultés possibles, à la fois aux niveaux conceptuel et pratique, ainsi que les arbitrages à effectuer.

Une séquence de corrections serait un ordre optimal dans lequel les données devraient être corrigées. Pour établir cette séquence, il faut déterminer les critères à prendre en compte eu égard aux besoins et procéder à des arbitrages, comme nous le verrons ci-dessous.

Enfin, cette séquence est évolutive puisqu'elle doit suivre l'évolution de la source de données, ce qui pose la question de l'historique des séquences de corrections pour que les états antérieurs des données puissent être corrigés (par exemple, si une donnée disparaît à un moment *T*, les données en *T-1* ne pourront pas être corrigées à l'aide de la séquence correspondante en *T*). Cette historicisation est complexe et pose de nombreuses questions non élucidées.

Dans la pratique, la résolution d'une anomalie implique la correction (modification) d'une ou plusieurs données. Il est donc préférable de centrer la séquence de corrections sur les données et non sur les anomalies. Ceci permet de mettre en évidence les données qui gagneraient à être corrigées en priorité en raison de leur importance (définie par les critères retenus pour élaborer la séquence). Le postulat est donc que, quel que soit le type d'anomalie détecté sur cette donnée, l'impact de cette anomalie est identique (que ce soit un problème de domaine de définition ou une incohérence par rapport à une autre donnée).

Objectifs et critères pour modéliser la séquence de corrections

Une séquence de corrections aurait pour objectifs :

- d'accélérer le traitement des anomalies (notamment en évitant les risques de corrections multiples) et de diminuer leur nombre[95] ;
- d'améliorer le traitement des données ayant un impact métier important[96] ;
- de permettre un accès plus rapide à l'information pertinente.[97]

Eu égard à ces objectifs, trois critères avaient été étudiés :

- les dépendances entre données[98] ;
- l'impact des données sur les droits sociaux des travailleurs dans le cadre de la sécurité sociale[99] ;
- les échéances temporelles pour lesquelles les données devaient être disponibles sans anomalies.[100]

Chaque donnée se voit attribuer une place dans la séquence qui reflète son importance. Des combinaisons de ces critères ont également été étudiées.

Prenons un exemple qui combine les deux critères suivants : dépendance entre données et impact des données sur les droits sociaux. L'objectif est d'obtenir un bon compromis entre :

- cohérence des données (nécessité de corriger en priorité les données les plus utilisées pour vérifier la validité des autres données) ;
- besoins business (nécessité de corriger en priorité les données dont l'impact sur les droits sociaux est le plus élevé).

Ces deux critères combinés permettent d'élaborer la séquence de la Figure 50. Sont donc corrigées en premier lieu les données qui présentent un degré de dépendance élevé et un impact important sur les droits sociaux.

[95] À titre d'exemple, dans la DmfA, la catégorie employeur détermine les droits de l'employeur (déduction de cotisations, réduction de charges) et des travailleurs (droits sociaux). Par conséquent, une erreur dans cette zone entraîne inévitablement la détection de nombreuses incohérences au sein de la déclaration en raison des dépendances entre données (2.3.3). La correction prioritaire de cette donnée aurait alors pour conséquence la disparition de nombreuses anomalies puisque les données déclarées seraient formellement cohérentes vis-à-vis de la catégorie employeur.

[96] Dans le cadre de la sécurité sociale, certaines données doivent être valides pour que le travailleur puisse bénéficier de ses droits sociaux tels que le remboursement de ses soins de santé ou l'attribution de son pécule de vacances.

[97] Les anomalies étant résolues plus vite, les dossiers concernés pourraient être traités plus rapidement.

[98] L'objectif visé est de corriger en priorité les données sur lesquelles se basent les contrôles croisés internes (2.3.1) et donc les données dont dépend la validité des autres données. En raison des dépendances entre données, une anomalie sur ces données peut potentiellement générer beaucoup d'anomalies fictives. La correction de ces données pourrait impliquer la disparition de ces anomalies fictives et par conséquent une baisse du nombre d'anomalies à traiter.

[99] L'objectif poursuivi avec ce critère est de corriger prioritairement les données ayant un fort impact sur l'attribution des droits sociaux. De cette manière, si un travailleur doit bénéficier d'un droit social (chômage, assurance maladie-invalidité, revenu de remplacement suite à un accident du travail, etc.), ce droit peut lui être attribué, et ce le plus correctement possible. L'intérêt métier (et politique) de ce critère est donc important.

[100] Dans ce cas-ci, l'objectif est de corriger en premier lieu les données nécessaires à l'attribution d'un droit social à une date fixe. Par exemple, en Belgique, les pécules de vacances sont payés entre le 2 mai et le 30 juin. Pour que ces pécules puissent être calculés, les données doivent être disponibles à partir du 15 avril. À ce moment, si une donnée n'est pas correcte, le pécule ne pourra pas être calculé. Si la correction de la donnée n'est pas effectuée dans des délais raisonnables, la sécurité sociale n'est pas en mesure de respecter la législation et les conventions collectives de travail.

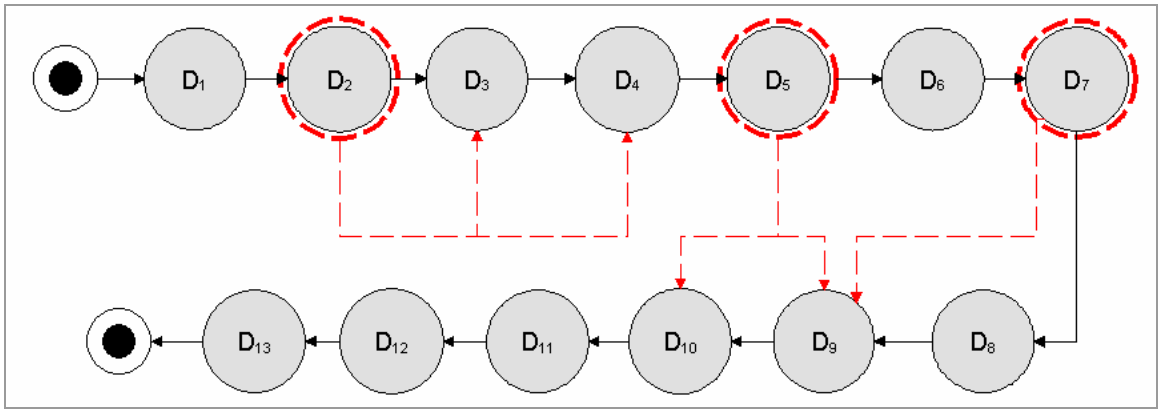


Figure 50 : Exemple de séquence de corrections

Comme on peut le constater, des incohérences (indiquées à l'aide des lignes tiretées) subsistent dans la séquence. Elles sont inévitables. Les flèches en pointillé indiquent les zones utilisées pour contrôler la donnée entourée mais corrigée après celle-ci. Par exemple, la donnée D_2 est corrigée avant les données D_3 et D_4 alors que la validité de D_2 dépend de D_3 et D_4.

Par conséquent, D_2 étant corrigé avant D_3 et D_4, lorsque ces deux dernières données seront corrigés, il se pourrait qu'une anomalie soit à nouveau détectée sur D_2, ce qui renvoie au problème des corrections multiples que nous avions évoqué précédemment (voir 2.3.3). La problématique est la même avec D_5 par rapport à D_9 et D_{10} et avec D_7 par rapport à D_9.

Chaque séquence, quels que soient les critères pris en compte, présente des imperfections. Le choix des critères et de la séquence dépend donc des objectifs poursuivis.

Le critère temporel (échéance) est plus complexe à gérer et pose une difficulté majeure : l'échéance est une information dynamique (plus l'échéance se rapproche, plus la correction des données devient urgente) tandis que les séquences de corrections n'évoluent que périodiquement étant donné qu'elles se basent sur un état de la source à un moment *t*.

Ce critère implique des opérations de maintenance et de mise à jour de la séquence relativement complexes, ce qui le rend difficile à appliquer dans la pratique.

Arbitrages

Dans le cadre d'une séquence de corrections, les arbitrages suivants se posent :

- *Imperfections* : comme nous l'avons vu, une séquence de corrections contient toujours des imperfections. Des choix doivent être opérés pour déterminer celles qui peuvent être acceptées eu égard aux objectifs poursuivis.
- *Qualité vs disponibilité* : il est difficile, voire impossible, de concilier des visions de travail différentes si les données sont utilisées à des fins différentes. Ainsi, il se peut que pour une même donnée, un secteur souhaite qu'elle soit disponible (mais éventuellement non corrigée définitivement) de manière à pouvoir octroyer le droit, quitte à faire une rectification plus tard tandis qu'un autre souhaite que la donnée soit corrigée de manière sûre et définitive avant de l'utiliser.
- *Complexité vs gestion des séquences* : la prise en compte de plusieurs critères enrichit certes la séquence et son adéquation aux objectifs poursuivis, mais la rend aussi plus complexe à gérer. Il convient de

déterminer dans quelle mesure cette complexité est pertinente dans le cadre d'un arbitrage « coûts-bénéfices ».

- *Rigidité vs liberté de l'utilisateur* : faut-il contraindre les agents correcteurs à suivre cette séquence ou faut-il leur laisser la liberté d'éventuellement corriger d'une manière différente ?

Obstacles à la mise en œuvre de séquences de corrections

Le résultat de cette étude a montré qu'une séquence de corrections automatisée et formalisée était envisageable en théorie, mais difficile à mettre en œuvre dans la pratique.

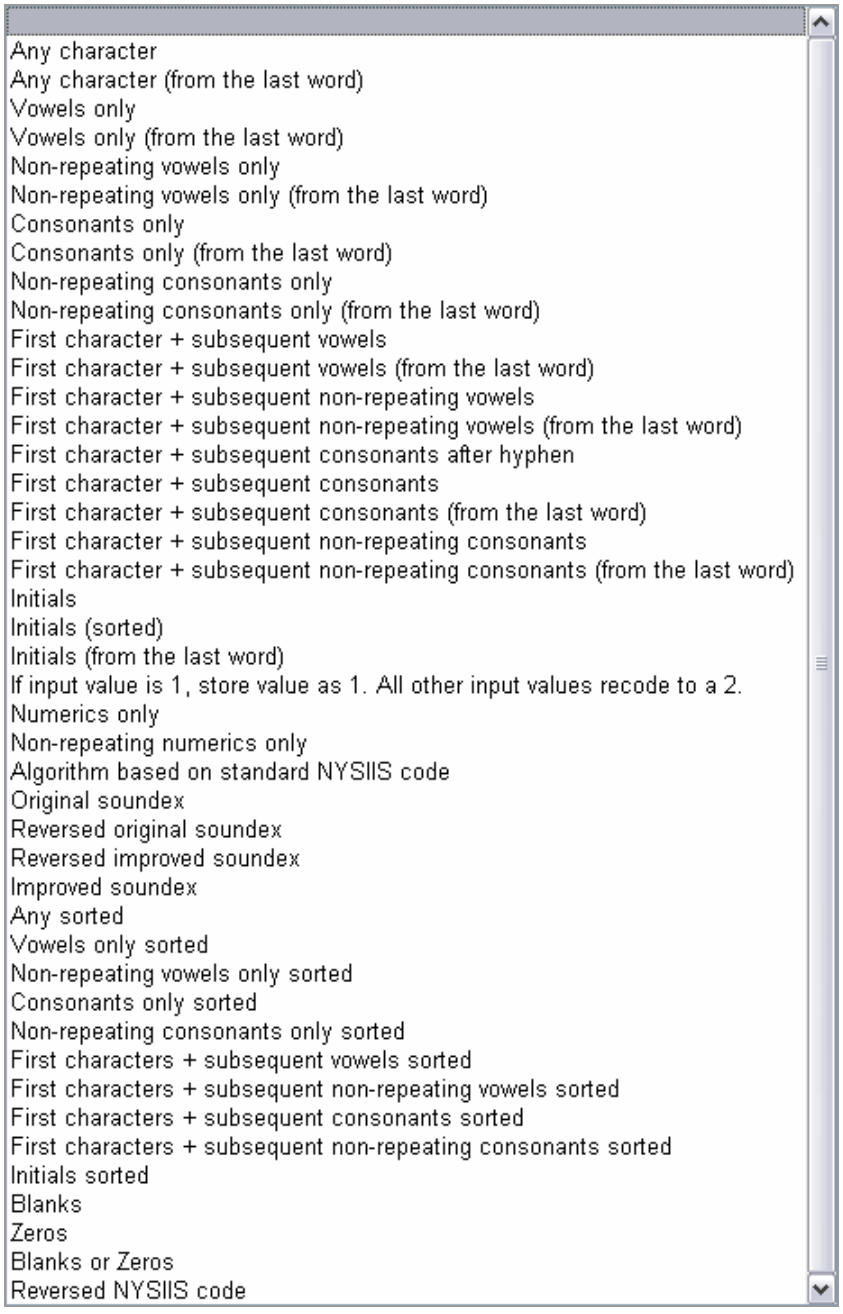
Trois constats conduisent à cette conclusion :

1. La résolution d'une anomalie implique couramment de corriger un ensemble de données et non une seule en raison des dépendances entre données. De ce fait, les personnes qui corrigent les anomalies modifient plusieurs autres données avant de valider leur correction, ce qui fausse l'idée d'un recours à une séquence de corrections automatisée et formalisée.
2. Du fait des anomalies fictives, une donnée peut être considérée comme erronée et donc être présente dans la séquence avec une place quelconque alors que l'erreur porte *in fine* sur une autre donnée considérée comme formellement correcte et qui, de ce fait, n'est pas présente dans la séquence.
3. Enfin, le dernier constat est d'ordre humain. Les agents en charge du traitement des anomalies sont des humains qui n'apprécient guère de devoir suivre un ordre prédéfini qui leur ôte toute initiative sur la base de leurs connaissances et expérience.

En conclusion, nous ne recommandons pas cette stratégie de gestion. La mise sur pied d'une application en vue de documenter la correction des anomalies et apporter ainsi une aide aux personnes en charge de la correction des anomalies nous semble être une stratégie plus efficace et adoptée plus facilement par les agents (voir 2.4.2).

6.2. Routines pour la sélection de caractères lors la création de *window keys*

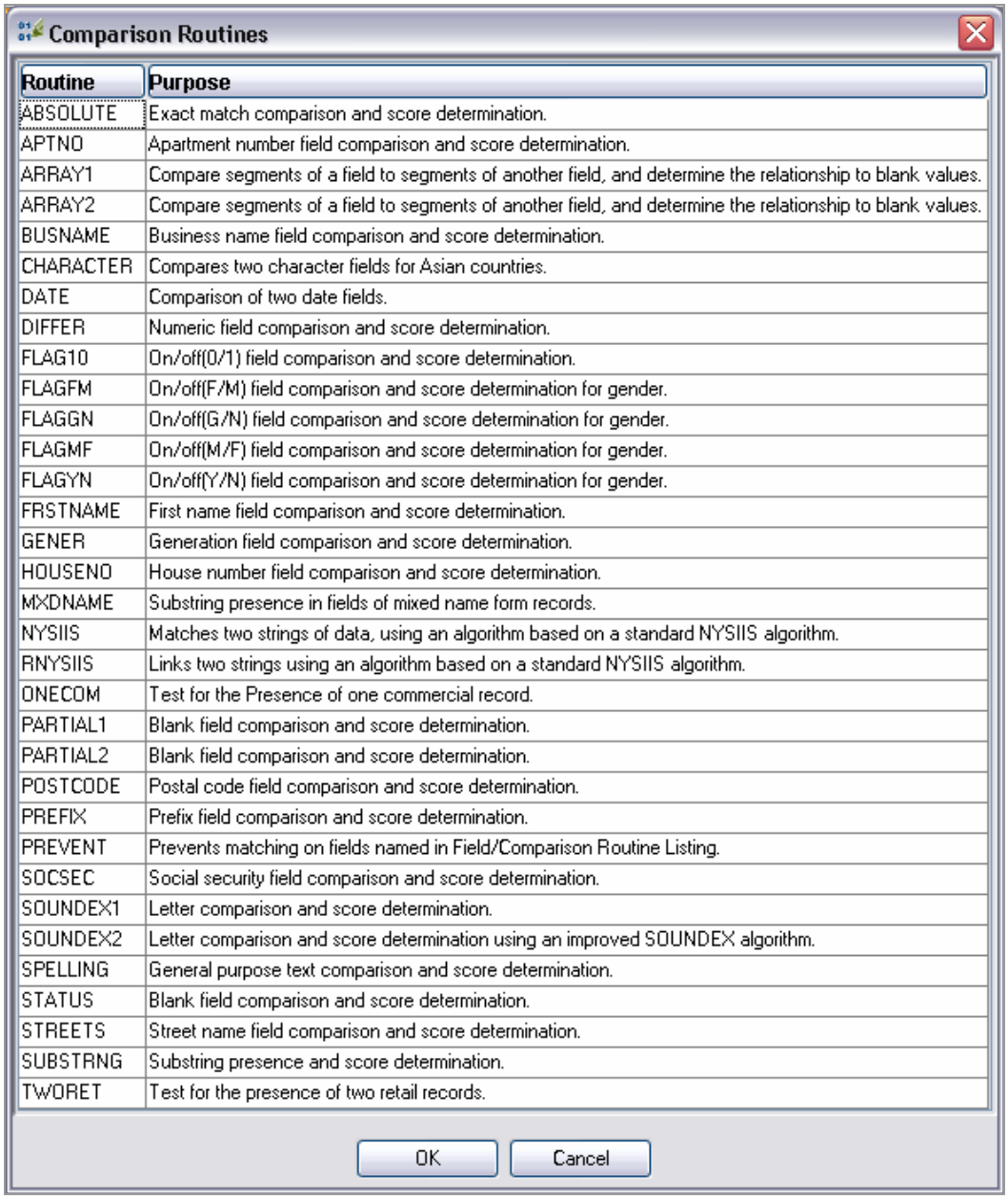
La liste ci-après reprend les meilleures pratiques quant à la manière dont des caractères (sous-chaînes) peuvent être sélectionnés, axées sur la définition de *window keys* (métadonnées composées pour servir de colonne critique dans la technique d'optimisation « blocking »).



La plupart de ces méthodes sont évidentes ou suffisamment connues dans la littérature librement accessible. Si vous avez des questions, vous pouvez en outre toujours vous adresser au Centre de compétences Data Quality de Smals.

6.3. Routines de comparaison champ par champ

Pour la création de modèles de matching, les data quality tools viennent avec toute une série de routines de comparaison champ par champ, qui sont en fait des « best practices ».



Comparison Routines

Routine	Purpose
ABSOLUTE	Exact match comparison and score determination.
APTNO	Apartment number field comparison and score determination.
ARRAY1	Compare segments of a field to segments of another field, and determine the relationship to blank values.
ARRAY2	Compare segments of a field to segments of another field, and determine the relationship to blank values.
BUSNAME	Business name field comparison and score determination.
CHARACTER	Compares two character fields for Asian countries.
DATE	Comparison of two date fields.
DIFFER	Numeric field comparison and score determination.
FLAG10	On/off(0/1) field comparison and score determination.
FLAGFM	On/off(F/M) field comparison and score determination for gender.
FLAGGN	On/off(G/N) field comparison and score determination for gender.
FLAGMF	On/off(M/F) field comparison and score determination for gender.
FLAGYN	On/off(Y/N) field comparison and score determination for gender.
FRSTNAME	First name field comparison and score determination.
GENER	Generation field comparison and score determination.
HOUSENO	House number field comparison and score determination.
MXDNAME	Substring presence in fields of mixed name form records.
NYSIIS	Matches two strings of data, using an algorithm based on a standard NYSIIS algorithm.
RNYSIIS	Links two strings using an algorithm based on a standard NYSIIS algorithm.
ONECOM	Test for the Presence of one commercial record.
PARTIAL1	Blank field comparison and score determination.
PARTIAL2	Blank field comparison and score determination.
POSTCODE	Postal code field comparison and score determination.
PREFIX	Prefix field comparison and score determination.
PREVENT	Prevents matching on fields named in Field/Comparison Routine Listing.
SOCSEC	Social security field comparison and score determination.
SOUNDEX1	Letter comparison and score determination.
SOUNDEX2	Letter comparison and score determination using an improved SOUNDEX algorithm.
SPELLING	General purpose text comparison and score determination.
STATUS	Blank field comparison and score determination.
STREETS	Street name field comparison and score determination.
SUBSTRNG	Substring presence and score determination.
TWORET	Test for the presence of two retail records.

OK Cancel

Toutes ces routines sont documentées dans la documentation des outils de qualité des données dont Smals dispose. Si vous avez des questions, vous pouvez toujours vous adresser au Centre de compétences Data Quality de Smals.

6.4. Tailles des *windows* et temps de traitement afférent

Ci-après figure un exemple de tailles de *windows* dans le cadre d'une détection de doublons. Il s'agissait d'un Data Matching complexe de données internationales de dénominations et d'adresses d'entreprises (plus de 532.000 enregistrements), où plus de 35 *match types* ont été définis sur la base de 8 critères sur 6 champs. Le temps de traitement s'est élevé à 1h 10min 24sec sur une machine virtuelle (2-CPU, 3Ghz, 4Gb, Solaris 10). La taille moyenne des *windows* s'élève environ à 30, mais la distribution est loin d'être uniforme ou

normale (Figure 51). Cela est prévisible : étant donné que les *window keys* ont été définies sur la base de caractères du code postal et du nom de la commune, la taille des *window keys* résultantes suit dans une certaine mesure également la répartition géographique des entreprises.

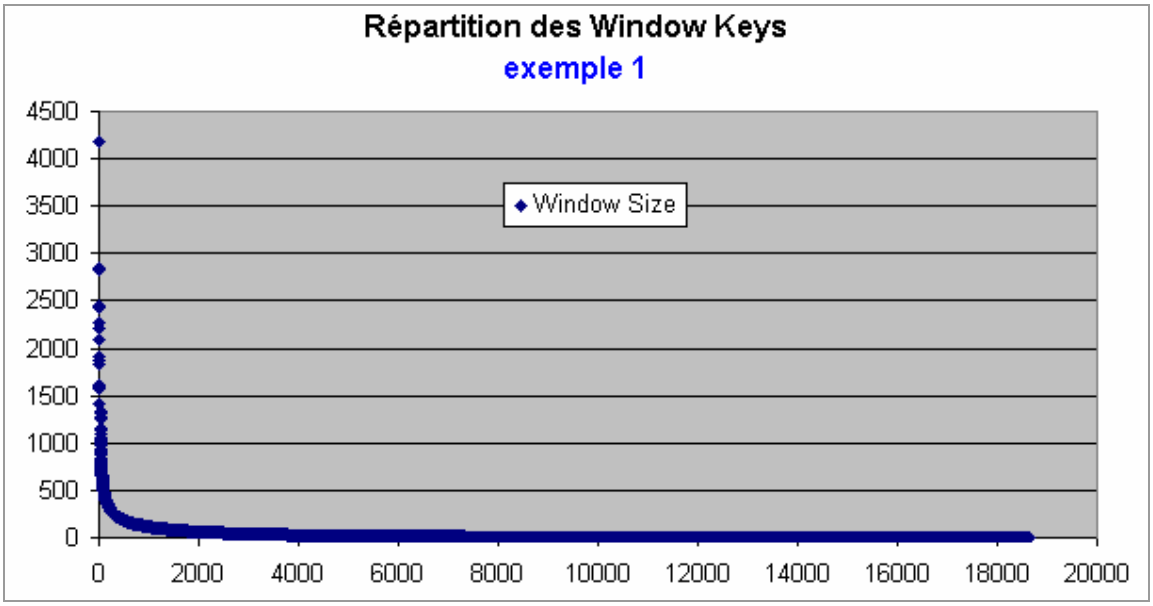


Figure 51 : Distribution de 532.584 enregistrements sur 18.642 windows. Plus grande window : 4183 enregistrements. 90 windows contiennent plus de 500 enregistrements, 217 windows contiennent plus de 300 enregistrements.

Bien entendu, des optimisations sont encore possibles ici. Pour plus d'informations, vous pouvez toujours vous adresser au Centre de compétences Data Quality de Smals.

Pour démontrer l'influence des *window keys*, nous vous proposons encore l'exemple suivant. Dans un projet comportant des données d'adresses belges (environ 275.000 enregistrements), un Data Matching complexe a été réalisé, où 20 *match types* ont été définis sur la base de 5 critères sur 4 champs. Le temps de traitement s'est élevé à seulement 5min 41sec sur la même machine virtuelle (2-CPU, 3Ghz, 4Gb, Solaris 10).

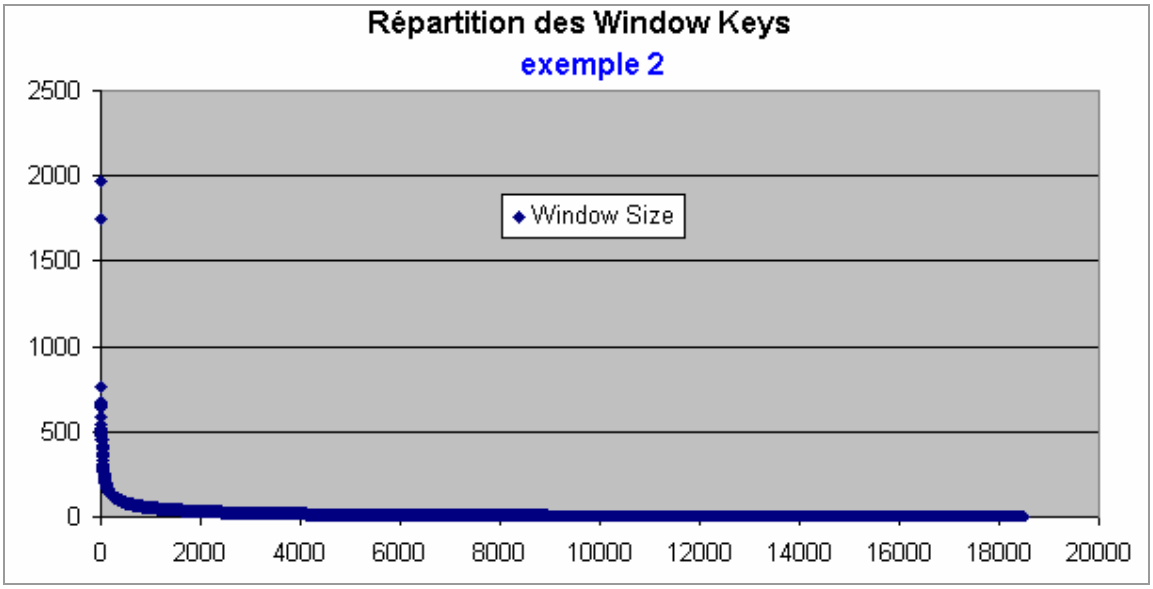


Figure 52 : Distribution de 274.906 enregistrements sur 18.495 windows. Plus grande window : 1973 enregistrements. 11 windows contiennent plus de 500 enregistrements, 40 windows contiennent plus de 300 enregistrements.