UN SERVICE AU CŒUR DE LA QUALITÉ DES BASES DE DONNÉES

PRÉSENTATION D'UN PROTOTYPE D'ATMS

Isabelle Boydens, Gani Hamiti** et Rudy Van Eeckhout****

Les bases de données doivent idéalement pouvoir être adaptées aux évolutions de leur environnement selon leurs usages. La prise en compte de ces changements revêt un impact stratégique pour la qualité des données administratives, et de ce fait pour les systèmes d'information statistiques qui les utilisent. Afin de gérer au mieux les transformations issues du réel observable touchant les données, il existe désormais une approche innovante, opérationnelle et généralisable à tout système de gestion de base de données relationnel.

Grâce aux avancées de la recherche en matière de qualité de données, l'étude des anomalies et de leurs traitements donne le jour à un prototype original, appelé ATMS* (Anomalies & Transactions Management System). *Ce service permet un suivi des anomalies et des traitements, en support à la méthode dite du* back tracking : *dans une approche préventive de la qualité des données, la méthode est destinée à améliorer structurellement la qualité à la source, et son implémentation révèle un retour sur investissement important. Les caractéristiques du prototype d'ATMS sont mises en relation avec le recours aux* data quality tools *en usage dans les approches curatives, offrant de nouvelles perspectives pour les systèmes d'information statistiques.

Databases must ideally evolve over time along with the change of their environment according to their uses. Taking these changes into account has a strategic impact on administrative data quality, and therefore on the statistical information systems using them. In order to help manage the transformations resulting from the observable reality affecting data, this article proposes an innovative, operational approach generalizable to any relational database management system.

Thanks to the advances of research in data quality, the study of anomalies and their management has given rise to an original prototype, called ATMS (Anomalies & Transactions Management System). This service allows the tracking of anomalies and processing, supporting the back tracking method: in a preventive approach of data quality, the method is intended to structurally improve quality at the source, and its implementation provides a significant return on investment. The characteristics of the ATMS prototype are combined with the use of data quality tools in curative approaches, offering new perspectives for statistical information systems.

* Professeur, Université libre de Bruxelles ; *Data Quality Expert*, Smals,
Isabelle.Boydens@ulb.be
** *Data Quality Analyst*, Smals,
Gani.Hamiti@smals.be
*** *Database R&D*, Smals,
Rudy.Van-Eeckhout@smals.be

« L'évolution est inséparable de la structure parce que l'ensemble lui-même est moins un système arrêté que la fixation provisoire d'un mouvement, l'ordre intelligible d'une tendance [...] *».*

Raymond Aron (La philosophie critique de l'histoire, 1969)

« L'HYPOTHÈSE DU MONDE CLOS » AU CŒUR D'UN RÉEL FLUCTUANT

Les statisticiens font de plus en plus appel à des données administratives et transactionnelles, en soi ou en complément d'autres sources (enquêtes, etc.) (Hand, 2018). L'examen de leur qualité est dès lors également stratégique pour la statistique, qu'elle soit publique ou pas.

Une base de données doit idéalement évoluer avec l'interprétation des réalités qu'elle permet d'appréhender. Les réalités normées sont en effet mouvantes. Ainsi, après les attentats des années 2015 et suivantes, les fichiers de police en la matière, tant en France (Chapuis, 2018) qu'en Belgique (Agence Belga, 2018), ont-ils connu de nombreuses anomalies[1] : doublons potentiels, « faux actifs » en dépit d'un non-lieu, effacements anticipés, difficultés d'interprétation, etc. Celles-ci se sont accumulées suite à l'émergence de catégories de menaces inédites, mais aussi l'urgence dans laquelle ces données sensibles ont dû être traitées. De telles évolutions sont constamment à l'œuvre au cœur des bases de données administratives, sources importantes pour le statisticien.

> *« Toute base de données opérationnelle bien conçue repose sur une hypothèse, celle du « monde clos ». »*

Toute base de données opérationnelle bien conçue repose sur une hypothèse, celle du « monde clos » : des domaines de définition spécifient l'ensemble des valeurs admises au sein du modèle ou du schéma de la base de données (les contraintes d'intégrité) ; les « règles métier » peuvent aussi se décliner dans le code applicatif et contribuer ainsi à la définition des données. Dès lors, une valeur non incluse dans le domaine de définition est considérée comme fausse et doit être rejetée de la base.

Or à l'échelle de millions d'enregistrements, de centaines de champs et de flux d'information, les phénomènes émergents sur le terrain ne sont pas immédiatement pris en compte au sein des bases de données. L'information s'y construit progressivement, au fil de l'interprétation humaine et en l'absence de référentiel absolu.

En dehors du domaine de définition de la base répondant à l'hypothèse du monde clos, la « réalité normée » évolue de manière continue et imprévisible, faisant fi de toute règle d'explication causale déterministe. Et quand la base de données est un instrument d'action sur le réel (dans les secteurs administratifs, médicaux, environnementaux, militaires, etc.), ces questions sont fondamentales et affectent la qualité des données.

1. Par anomalie, nous entendons ici une erreur formelle (par exemple : valeur obligatoire non complétée) mais aussi une présomption d'erreur demandant une interprétation humaine (par exemple : présomption de doublons entre enregistrements fortement similaires, émergence d'une nouvelle catégorie d'activité non prise en compte dans les tables de référence, etc.). Une typologie des anomalies est proposée plus loin.

Il est cependant possible de mieux prendre en compte ces phénomènes sur le plan opérationnel, au cœur du système d'information. Et ce, en particulier, à travers l'interprétation des anomalies et de leurs traitements.

La recherche en « qualité de données » s'intéresse à l'univers des anomalies à des fins opérationnelles. Leur analyse s'est concrétisée par la mise en place d'un service appelé **ATMS, *Anomalies & Transactions Management System.*** Ce système permet le suivi dans le temps de l'historique des anomalies et de leurs traitements sur la base d'indicateurs jugés stratégiques et variables selon le contexte d'usage. Il s'agit d'un concept innovant déjà éprouvé dans la pratique, qui s'appuie sur une méthode dite de *back tracking*.

DES ENJEUX IMPORTANTS S'AGISSANT DES BASES DE DONNÉES ADMINISTRATIVES

De nombreux systèmes d'information administratifs d'envergure sont concernés par la qualité des données, par exemple, en France, la DSN (Déclaration Sociale Nominative). Au sein de celle-ci, tout contrôle jugé essentiel est bloquant (Renne, 2018) mais certains contrôles sont « non bloquants » afin de ne pas ralentir le processus de recueil ou de collecte des déclarations. Ces derniers demandent un traitement manuel ultérieur que les gestionnaires de la base s'efforcent de rationaliser et soulève un arbitrage « coût-qualité » (Renne, 2018). L'exemple de la DSN montre également que la qualité de la base requiert que soit portée une attention particulière à la maîtrise des changements réglementaires (Humbert-Bottin, 2018)[2].

En Belgique, la base de données LATG[3] à la fin des années 1990, puis, son héritière modernisée, la DmfA[4] au début des années 2000, furent des « cas d'étude » des travaux de recherche en matière de qualité de données (Boydens, 1999 ; 2018) vu leur ampleur. La DmfA permet en effet actuellement le prélèvement et la redistribution annuels de 65 milliards d'euros de cotisations et prestations sociales à l'échelle de la Belgique. Elle fait l'objet de modifications législatives trimestrielles et constitue depuis 2001 un système d'information d'envergure intégré, doté de caractéristiques proches de la DSN sur le plan fonctionnel. Elle représente un des socles de nombreuses statistiques en matière d'emploi et de salaires en Belgique. Depuis 20 ans, les recherches ont également porté sur la qualité de nombreuses autres bases de données transactionnelles à la source de productions statistiques[5].

La qualité des données se pose aussi dans le cadre de registres administratifs transversaux tels que ceux prévus en Allemagne. Ceux-ci reposent en effet sur de nombreuses interconnexions susceptibles de soulever d'importantes questions sémantiques[6].

2. [N.D.L.R.] Ce point est régulièrement évoqué dans les articles du Courrier des statistiques, voir également l'article de Christian Sureau et Richard Merlen dans ce même numéro.
3. Base de données relatives aux salaires et aux temps de travail (*Loon en ArbeidsTijdsGegevensbank*).
4. Déclaration multifonctionnelle (*Multifunctionele Aangifte*).
5. Tous ces travaux sont effectués dans le respect de la réglementation européenne du RGPD (Règlement général pour la protection des données).
6. Voir (Bens et Schukraft, 2019). Les auteurs citent à titre d'exemple les grandeurs monétaires, comme les revenus et les chiffres d'affaires (p. 15) ou la nécessité d'un identifiant unique pour les personnes et les entreprises, abstraction faite d'un usage donné (fiscalité, Sécurité sociale, commerce, etc.) (pp. 14-15).

Le rythme des adaptations à apporter à un système d'information varie en fonction des objectifs poursuivis : bases de données administratives ou médicales, par exemple, en tant qu'outils d'action sur le réel, d'une part ou systèmes d'information statistique, d'autre part, en tant qu'instruments d'observation ou d'aide à la prise de décision. Pour les premières, le rythme de modification sera idéalement rapide. Pour les seconds, le rythme d'évolution structurelle est beaucoup plus lent voire inexistant (dans le cas de résultats d'enquêtes, par exemple) afin d'assurer une comparaison sur le long terme alors que l'actualité et la qualité des sources qui les alimentent seront importantes.

Dans un tel contexte, le statisticien est souvent confronté à la question du meilleur moment auquel prendre « la photo » afin d'extraire les données issues d'un système d'information administratif, vis-à-vis duquel il vit une forme de perte de maîtrise (Rivière, 2018). Face à cela, les acquis des travaux en matière de qualité de données offrent des perspectives constructives.

LES TRAVAUX ACADÉMIQUES EN MATIÈRE DE QUALITÉ DE DONNÉES

« La qualité d'une base de données désigne son adéquation relative aux usages pour lesquels elle a été conçue, sous contrainte de budget. »

La qualité d'une base de données désigne son adéquation relative aux usages pour lesquels elle a été conçue, sous contrainte de budget. Les recherches dans ce domaine se sont déployées dans les années 1980 (Madnick *et alii*, 2009) avec la nécessité pour les entreprises de disposer d'adresses et de coordonnées adéquates dans leurs fichiers de clients. C'est ainsi que sont apparus les *data quality tools* ***(encadré 1)***, domaine qui s'est très vite développé au niveau international et qui reste très actif (Hamiti, 2019).

En tant qu'approche « curative » ***(figure 1)***, ceux-ci ont pour objet, sur la base de milliers d'algorithmes régulièrement enrichis, de détecter les problèmes de qualité formellement identifiables (présomptions de doubles, etc.) déjà présents dans les bases de données et d'y remédier *a posteriori* de manière semi-automatique. Ces outils permettent également la gestion automatisée des cas problématiques « en ligne », lors de la saisie dans un portail, par exemple.

Cependant, si l'on se contente d'agir en aval, on ne résout pas structurellement la cause de ces problèmes qui vont sans cesse se reproduire. Ceux-ci peuvent en effet puiser leur source dans des défauts de conception, dans l'évolution du réel représenté ou, encore, dans les flux et procédures qui alimentent les bases de données (par exemple, processus inutilement redondants générant systématiquement des doublons). Dès lors, sans action complémentaire en amont, les *data quality tools* sont destinés à être mobilisés *ad infinitum*. Ceux-ci restent néanmoins indispensables, car l'utilisateur n'a pas nécessairement accès aux flux et procédures qui ont produit les données qu'il exploite.

Dès lors, en complément des interventions « curatives », des approches « préventives » sont indispensables afin d'identifier et de résoudre structurellement à la source les causes des anomalies ***(figure 1)***.

L'Université libre de Bruxelles dédie un enseignement spécifique à la qualité des données depuis 2006, présentant les deux types d'approches (Boydens, 2021). Les travaux de recherche se situent pour une part dans l'optique des méthodes préventives (Boydens, 1999 ; 2010 ; 2012 ; 2018 ; Bade, 2011 ; Radio, 2014 ; Dierickx, 2019). Les travaux actuels en matière de *data quality research* sont également axés sur les algorithmes de type *record linkage* (Batini et Scannapieco, 2016). Ces derniers sont mobilisés par les *data quality tools* dans le cadre des opérations de « comparaison et dédoublonnage » ***(encadré 1)***.

Nombreuses sont les études qui adoptent une vision déterministe : elles envisagent l'écart de la base de données au réel en termes d'inexactitude/exactitude formelle (Srivastava *et alii*, 2019). Or, il n'existe aucune projection biunivoque nécessaire entre le réel empirique et sa représentation au sein d'une base de données.

Encadré 1. *Data Quality Tools*

En parallèle de l'approche préventive de la qualité des données, décrite dans cet article et appuyée sur l'ATMS, une approche curative existe. Celle-ci est destinée à l'amélioration semi-automatique de la qualité des données à leur entrée dans le système ou déjà présentes dans une base de données préexistante ou un ATMS. Le plus souvent, l'approche curative va mobiliser des outils gratuits ou commerciaux développés par une tierce partie. Généralement, ces outils couvrent une à trois de ces grandes familles de fonctionnalités :

- **profilage** : analyser qualitativement et quantitativement des données pour en évaluer la qualité et, souvent, débusquer des problèmes inattendus. Exemple : distribution de la longueur des valeurs d'une colonne, inférence de type, vérification ou découverte de dépendances fonctionnelles ;
- **standardisation** : conformer les données à un standard défini avec le maître d'ouvrage ou à un référentiel existant, pouvant être fourni avec l'outil. Exemple : nettoyage et uniformisation de la représentation des numéros de téléphone, correction et enrichissement d'adresses postales ;
- **comparaison et dédoublonnage** : détecter les doublons et incohérences dans les enregistrements au sein d'un jeu de données ou entre plusieurs (issus potentiellement de bases de données distinctes, en vue d'une intégration, par exemple). La comparaison se base sur des colonnes discriminantes et des algorithmes tolérants à l'erreur (mesure de la distance d'édition, comparaison de l'empreinte phonétique, etc.), déterminés avec le maître d'ouvrage qui apporte sa connaissance du métier. Les outils les plus avancés permettent ici de conserver et lier les enregistrements originaux pertinents sans les écraser et d'en construire un qui représente chaque grappe ainsi repérée. Cet enregistrement sera alors le « survivant », utilisé pour dédoublonner les jeux de données si nécessaire.

Typiquement, ces outils interviennent en *batch*, c'est-à-dire en ciblant, en différé, un ou plusieurs jeux de données déjà existants. Certains permettent cependant également d'intervenir plus en amont, en exposant ces fonctionnalités sous la forme d'une API* que l'application peut appeler au cas par cas au moment où les données entrent dans le système. Ce mode d'action permet de standardiser ou de dédoublonner les données avant leur écriture dans la base et même, si besoin, de conditionner cette écriture par la réussite des opérations qui la précèdent. L'outil implémente ainsi effectivement un pare-feu de données complémentaire au système de détection d'anomalies déjà mis en place par l'application.

** Application Programming Interface*

QUAND L'ÉTUDE DES ANOMALIES AMÉLIORE LA QUALITÉ DES DONNÉES

Outre son intérêt évident pour la qualité des données, l'étude des anomalies est importante en raison de leur pourcentage élevé qui affecte structurellement les systèmes d'information : jusqu'à 10 % selon (Boydens, 2012) et selon d'autres sources (Van Der Vlist, 2011). Or, quand les enjeux (sociaux, financiers, médicaux, etc.) le demandent, ces anomalies doivent faire l'objet d'un examen semi-automatique, voire manuel, souvent lent et fastidieux.

D'où viennent les anomalies, quelle en est la typologie et de là, comment les gérer au mieux ? Afin de répondre à ces questions, il convient de revenir préalablement sur la notion de donnée telle que nous l'envisageons ici et qui fut récemment étudiée du point de vue du statisticien (Rivière, 2020).

DONNÉES DÉTERMINISTES ET DONNÉES EMPIRIQUES

Dans le monde des bases de données (Hainaut, 2018), une donnée est un triplet (i, d, v) composé des éléments suivants :

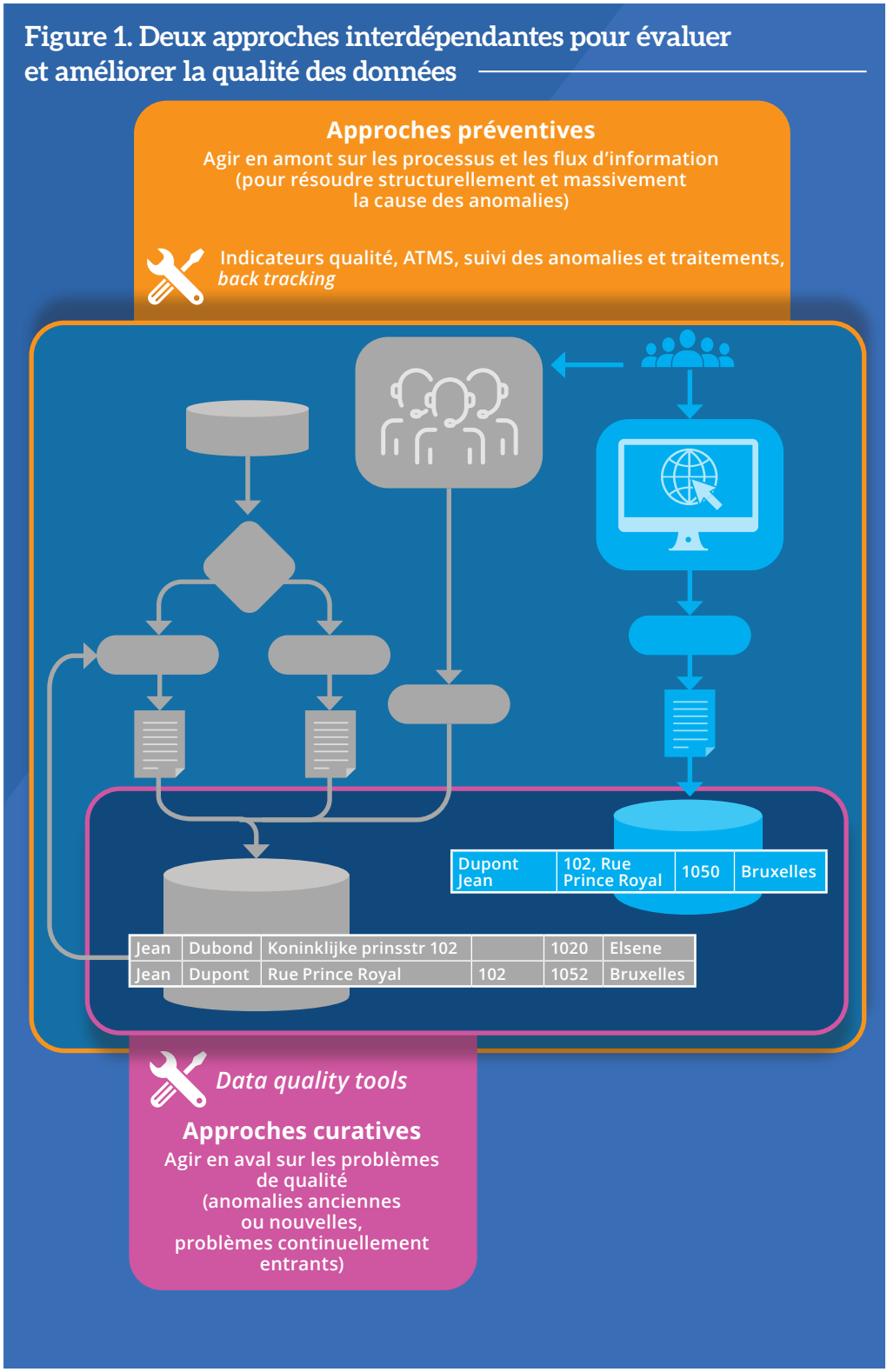
- un intitulé (i), renvoyant à un concept (une *catégorie d'activité administrative*, par exemple) ;
- un domaine de définition (d), composé d'assertions formelles spécifiant l'ensemble des valeurs admises dans la base pour ce concept (une liste contrôlée de valeurs alphabétiques d'une longueur maximale l, par exemple), complétées éventuellement de règles métier se trouvant dans le code applicatif ;
- et enfin, une valeur (v) à un instant t (le *secteur de la chimie*, par exemple).

On distingue alors les *données déterministes* des *données empiriques* (Boydens, 1999). 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 d'une opération algébrique simple portant sur un objet lui-même déterministe, comme la somme de valeurs relatives à tel champ numérique d'une base de données à un instant t. Les règles de l'algèbre pas plus que l'objet évalué n'évoluant dans le temps, on peut savoir à tout moment si le résultat d'une telle somme est correct ou pas. On dispose en effet d'un référentiel stable à cette fin.

> *« En ce qui concerne les données empiriques, sujettes à l'expérience humaine, la norme évolue dans le temps avec l'interprétation des valeurs qu'elle permet d'appréhender. »*

En revanche, en ce qui concerne les données empiriques, sujettes à l'expérience humaine, la norme évolue dans le temps avec l'interprétation des valeurs qu'elle permet d'appréhender. Ainsi en est-il par exemple du domaine médical (où la théorie évolue au fil des observations sur les patients atteints par une pathologie, comme en témoignent les recherches actuelles sur le coronavirus) mais aussi des domaines juridiques et administratifs où l'interprétation des concepts légaux se transforme avec l'évolution continue de la réalité traitée et avec celle de la jurisprudence. Comment en évaluer la validité en l'absence de référentiel absolu à cette fin ?

Figure 1. Deux approches interdépendantes pour évaluer et améliorer la qualité des données



CAPTURER ET INTERPRÉTER L'ÉVOLUTION DU RÉEL OBSERVÉ

Les données empiriques s'apparentent à des concepts « mobiles » au cœur des bases de données. En d'autres termes, la signification des concepts évolue avec l'interprétation des valeurs qu'ils permettent d'appréhender et ce, en l'absence de référentiel absolu et stable.

Approfondissant ce constat épistémologique apparemment sans issue opérationnelle, il est pourtant possible de bâtir une méthode généralisable permettant d'évaluer et d'améliorer la qualité de telles informations. La question n'est plus uniquement « les données sont-elles correctes ? » mais surtout « comment les données se construisent-elles progressivement ? ».

> *« La question n'est plus uniquement « les données sont-elles correctes ? » mais surtout « comment les données se construisent-elles progressivement ? »*

Ainsi, avec la mondialisation, de nouveaux cas non prévus initialement dans les tables de référence et dans la législation peuvent se présenter à une échelle nationale. Cela peut se produire par exemple dans le domaine de l'activité énergétique : la production d'énergie géothermique est très variable d'un endroit à l'autre du globe, certaines entreprises étrangères pourront donc utiliser une classification de leurs unités d'exploitation (« énergie renouvelable ») moins précise que celle potentiellement exigée, après examen, par la législation du pays d'exploitation (« énergie géothermique »), catégorie qui, dans notre exemple, n'est pas encore prise en compte dans la table de référence, laquelle devra faire l'objet d'une adaptation, comme expliqué ci-dessous ***(figure 2)***.

Dans ce cas, il n'est pas possible de vérifier le caractère correct des valeurs de la base de données de manière déterministe. En effet, lorsqu'une incohérence apparaît entre une telle valeur saisie au sein de la base et les tables de référence permettant d'en tester la validité, il peut s'avérer indispensable, lorsque les enjeux sont stratégiques[7], de procéder à une vérification manuelle, en contactant le citoyen ou l'entreprise concernée, par exemple. Une telle intervention peut aussi souvent être mobilisée si la catégorie attendue dans le fichier de référence pour un employeur donné ne correspond pas à la catégorie déclarée, car il se peut que l'employeur ait changé de catégorie depuis son immatriculation sans que cela n'ait été enregistré (car il ne l'a pas signalé par exemple).

C'est là, entre autres, que résidera l'intérêt d'un ATMS, en vue d'enregistrer et d'historiser les anomalies et transactions, permettant un suivi continu de celles-ci et une modification éventuelle ultérieure du domaine de définition pour l'adapter à une réalité nouvellement observée.

Illustrons ce mécanisme ***(figure 2)*** avec un autre exemple concret. En 2005, la catastrophe de l'ouragan Katrina a fait plus de 1 800 morts aux USA. Les instruments de mesure destinés à alerter les citoyens afin qu'ils quittent la zone existaient. Mais *a posteriori*, on s'est rendu compte que les bases de données qui les alimentaient n'étaient pas conçues pour intégrer l'évolution de certains phénomènes qui, alors sous-estimés, se sont révélés pourtant déterminants : il s'agissait de la montée des eaux dans les océans suite au réchauffement climatique, ainsi que de la sur-construction qui ne permet plus l'écoulement rapide de l'eau dans les sols. L'évacuation de la population fut dès lors beaucoup trop tardive. Ces mutations du réel sont encore à l'œuvre de nos jours dans les domaines hydrologiques et climatiques (Boydens, Hamiti et Van Eeckhout, 2020).

7. Par enjeux stratégiques, nous entendons des enjeux fondamentaux au regard du domaine d'application et des objectifs poursuivis : dans le domaine de la Sécurité sociale, par exemple, il peut s'agir du calcul des cotisations dues par un employeur (les taux variant avec la catégorie d'activité) ou bien des droits sociaux du travailleur (accès aux soins de santé, droit au chômage, etc.), lesquels peuvent dépendre entre autres de la validité des données signalétiques de l'employeur et de l'interprétation des anomalies associées, « stratégiques » elles aussi.

ESSAI DE TYPOLOGIE DES ANOMALIES

Une typologie des anomalies se profile alors, en fonction de leur cause potentielle et de la manière de les envisager :

- **erreur formelle certaine** due à l'intervention humaine lors de la mise à jour (champ obligatoire non complété, par exemple) ;
- **présomptions d'erreurs formelles** : présomptions de doubles ***(figure 1)*** par exemple dues à des processus redondants en amont ou incohérence avec une table de référence dont on ignore si elle a été mise à jour ;
- **erreur indétectable formellement *a priori***[8] : par exemple, omission d'une mise à jour.

Les deux derniers cas de figure peuvent quant à eux dénoter d'anomalies dues à l'évolution dans le temps du domaine empirique représenté et à l'émergence de nouveaux concepts non pris en compte ***(figure 2)***.

Selon les besoins du métier, on décidera de considérer ces anomalies comme :

- **bloquantes** : elles sont rejetées de la base de données en vertu de l'hypothèse du monde clos précédemment évoquée ;
- **non bloquantes** : les valeurs sont tout de même intégrées selon des modalités variables au sein du système d'information avec l'enregistrement correspondant, pour deux familles de raisons :

  - les rejeter du système ralentirait le processus métier (par exemple, le prélèvement des cotisations sociales) et elles ne sont pas considérées comme « stratégiques » (voir *supra*) ;

  - les prendre en considération dans le système d'information est indispensable, car elles sont considérées comme stratégiques et sont liées à des données empiriques dont la définition est potentiellement évolutive. À partir d'un certain seuil à évaluer par les spécialistes du domaine, leur traitement demande une interprétation humaine, car elles peuvent dénoter de l'émergence de phénomènes qu'il importera de prendre en considération dans le système d'information ***(figure 2)***, moyennant une gestion de versions. En outre, elles trouvent potentiellement leur origine dans les flux alimentant la base de données, problématique qui, une fois identifiée, pourra être structurellement résolue, comme nous le verrons plus loin avec le *back tracking*.

La décision consistant à identifier les anomalies empiriques « non bloquantes » est sensible en ce qu'elle relève d'une connaissance prévisionnelle des réalités traitées à un instant t, élément lui-même évolutif susceptible de faire l'objet d'une adaptation concertée au sein du système d'information. Ceci nous renvoie à la question épistémologique de la « boucle herméneutique »[9] (Boydens, 1999 ; 2012).

Comment prendre en considération les « anomalies non bloquantes » et leurs traitements, sans affecter ni la performance, ni l'intégrité des données en production ?

Avec l'ATMS, ou *Anomalies and Transactions Management System,* on passe de « l'hypothèse du monde clos » à celle d'un « monde ouvert » sous contrôles automatisés.

8. Ces cas peuvent être uniquement cernés indirectement, *via* des moyens latéraux, dont l'ATMS (voir *infra*).
9. La démarche herméneutique consiste à envisager les phénomènes empiriques en termes d'interactions par rapport à un cadre conceptuel plus général construit en vue de leur conférer un sens. Cependant, toute démarche interprétative soulève un paradoxe : celui du « cercle herméneutique » (Aron, 1969). Chaque observation ne prend sens que confrontée à un ensemble, à une « précompréhension ». Or, la sémantique de l'ensemble repose elle-même sur l'interprétation des éléments qui le constituent. Le processus de construction que suppose l'herméneutique est par nature toujours inachevé.

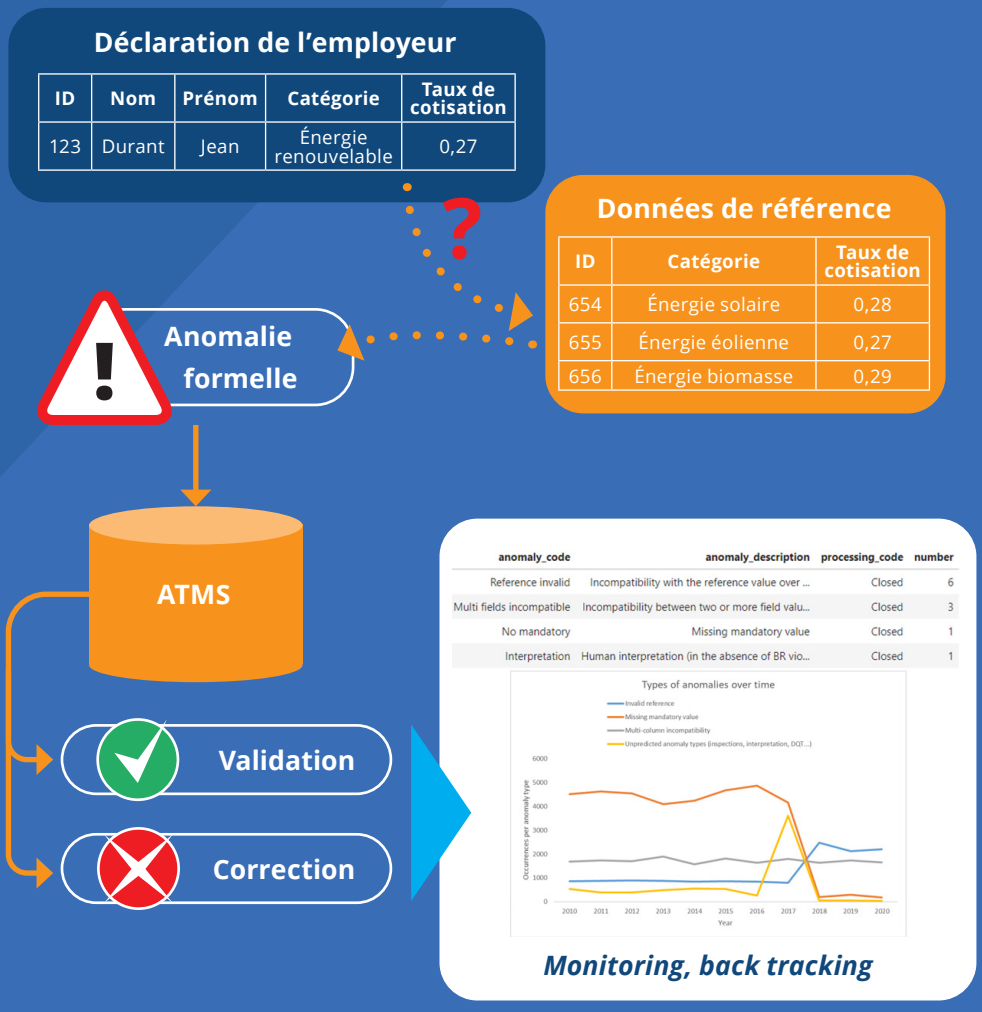
TROIS ÉCHELLES TEMPORELLES EN INTERACTION CONTINUE

« Solidaires, mais asynchrones. »

L'évolution de la norme, les transformations opérées au sein des bases de données, et la mouvance des « phénomènes » observables sur le terrain sont solidaires. Solidaires, mais asynchrones. Elles opèrent, suivant leur nature, au sein d'échelles de temps différentes.

Figure 2. Violation de « l'hypothèse du monde clos » dans un domaine empirique

Exemple : un test d'intégrité avant l'entrée des données dans la base de données principale détecte une anomalie formelle. Le traitement de l'anomalie (validation ou correction) est stocké dans l'ATMS et alimente un tableau de bord qui aidera à la prise de décision en vue d'améliorer la qualité des données.



Appliquant la notion de « temporalités étagées » de l'historien Fernand Braudel (Braudel, 1949) à l'étude des systèmes d'information administratifs en vue d'en améliorer la qualité, on peut distinguer :

- « **le temps long** » des normes (juridiques et plus largement empiriques), dont le rythme d'évolution est relativement plus lent (ainsi, dans le domaine de la Sécurité sociale belge, les modifications législatives impliquant autant de versions de schéma sont trimestrielles) ;
- le « **temps intermédiaire** » de la gestion des bases de données relativement plus rapide, ne fut-ce qu'en raison des évolutions technologiques ;
- et le « **temps court** » du réel observable, celui des citoyens ou des entreprises assujettis à l'administration, dont l'évolution est continue (Boydens, 1999).

Le concept de « temporalités étagées », théorisé par Fernand Braudel est ainsi une construction permettant d'identifier une hiérarchie entre plusieurs séquences de transformation inter-agissantes. Cette approche peut être complétée par les travaux du philosophe allemand Norbert Elias et sa notion de « continuum évolutif » (Elias, 1986). En effet, les interactions entre temporalités ne sont ni déterministes, ni unidirectionnelles, ce que laisserait entendre le modèle de Braudel seul, au sein duquel les séquences les plus lentes déterminent les plus rapides. Illustrons ces interactions par un exemple concret.

> *« Les interactions entre temporalités ne sont ni déterministes, ni unidirectionnelles. »*

En 1986, une équipe de scientifiques britanniques, spécialistes de l'étude du globe, signala la chute des taux d'ozone dans la stratosphère. Sur la base de cette observation, des chercheurs de la Nasa réexaminèrent leurs bases de données stratosphériques distribuées de par le monde ; ils découvrirent que depuis une décennie déjà, le phénomène de la baisse des taux d'ozone était resté occulté du fait que les valeurs faibles correspondantes avaient été systématiquement considérées comme des erreurs de mesure. En effet, la théorie scientifique de l'époque, modélisée dans leurs bases de données, ne leur avait pas permis de concevoir que de telles valeurs puissent être valides. Par la suite, le domaine de définition de la base a été adapté afin de considérer comme valides des taux faibles antérieurement en état d'anomalie (Boydens, 1999). De nos jours, ces phénomènes continuent d'évoluer.

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 et se révélant notamment à travers les anomalies évoquées plus haut.

ANOMALIES AND TRANSACTIONS MANAGEMENT SYSTEM : PRÉSENTATION FONCTIONNELLE

L'ATMS ou *Anomalies and Transactions Management System*[10] aide à repérer l'émergence et les augmentations de « validations » d'anomalies jugées stratégiques lors de la phase de traitement manuel. Une opération de validation signifie qu'après examen, un agent a estimé que l'anomalie correspondait dans les faits à une valeur pertinente. L'opérateur peut alors « forcer » le système à accepter la valeur sans affecter l'intégrité de la base de données principale ; le dispositif doit prévoir notamment un système de gestion de versions. Selon les droits d'accès, les agents ont accès à la fois à la base de données principale et à l'ATMS : ils peuvent donc à tout moment visualiser toutes les données, qu'elles soient ou pas en état d'anomalie potentiel, et quel que soit leur stade de traitement (correction, validation, etc.) ***(figure 3)***.

Si le taux de telles validations d'anomalies est élevé et récurrent ou si l'anomalie validée est stratégique, la possibilité existe que le domaine de définition de la base lui-même ne soit plus pertinent. *A priori* l'approche s'intéresse aux cas systématiques, mais elle peut également couvrir des cas peu nombreux touchant des types d'anomalies sensibles pour le métier (émergence d'une pathologie rare, par exemple).

Un algorithme peut alors émettre un « signal » destiné aux gestionnaires de la base afin qu'ils examinent si une modification structurelle de son domaine de définition, voire une révision de la norme correspondante (législation, théorie, etc.) sont requises. Les fluctuations des données environnementales ou administratives illustrées plus haut exemplifient les cas où ce mécanisme est à l'œuvre. En outre, il s'agit de conserver l'historique de leur traitement (une même anomalie pouvant être corrigée ou validée à plusieurs reprises suite à des inspections de terrain ou à l'interprétation des réglementations).

En l'absence d'une telle intervention, l'écart entre la base de données et le réel se creuserait. En effet, si l'on omet d'adapter le schéma, les anomalies correspondant à ces cas vont continuer de croître, nécessitant un examen manuel potentiellement lourd, susceptible de ralentir le traitement des dossiers et d'affecter la qualité des données avec des impacts financiers ou sociaux.

En amont, l'ATMS aide à améliorer la qualité des bases de données « source » et fournit divers indicateurs sur l'état du traitement des anomalies, aux personnes impliquées dans la gestion du système d'information (maître d'œuvre, maître d'ouvrage). Il permet par exemple :

- d'identifier les « pics » d'anomalies, de corrections et de validations (pouvant entraîner quant à elles une restructuration du schéma de la base) ;
- d'identifier les anomalies qui ne seraient jamais traitées (ni corrigées, ni validées) ;
- de déterminer le temps de stabilisation de la base de données, au fil des traitements des anomalies spécifiées, en fonction des besoins, et le moment le plus opportun en vue d'en tirer une photographie pour l'exploiter à d'autres fins.

10. Conceptuellement, logiquement et physiquement, l'ATMS est une base de données. Le terme de « système » désigne un cadre plus large, comprenant les processus, les applicatifs et l'équipe de gestion de la base, de traitement des anomalies, de même que les fournisseurs de l'information.

Certains de ces indicateurs pourraient s'avérer utiles également pour mieux exploiter les sources administratives à des fins statistiques. En support au *back tracking*, l'ATMS est un instrument plus puissant encore d'amélioration de la qualité des données, dans une perspective préventive ***(figure 1)***.

La méthode du *back tracking* est généralisable à tout domaine d'application[11]. Elle fut initiée par Thomas Redman, sous l'appellation *data tracking*.

LE *DATA TRACKING* DE THOMAS REDMAN

Le *data tracking* proposé par Thomas Redman d'AT&T Labs aux USA (Redman, 1996) vise à évaluer quantitativement la validité formelle des valeurs introduites dans une base de données et à en améliorer structurellement le traitement (Redman écarte explicitement les questions d'interprétation des données qu'il juge trop complexes).

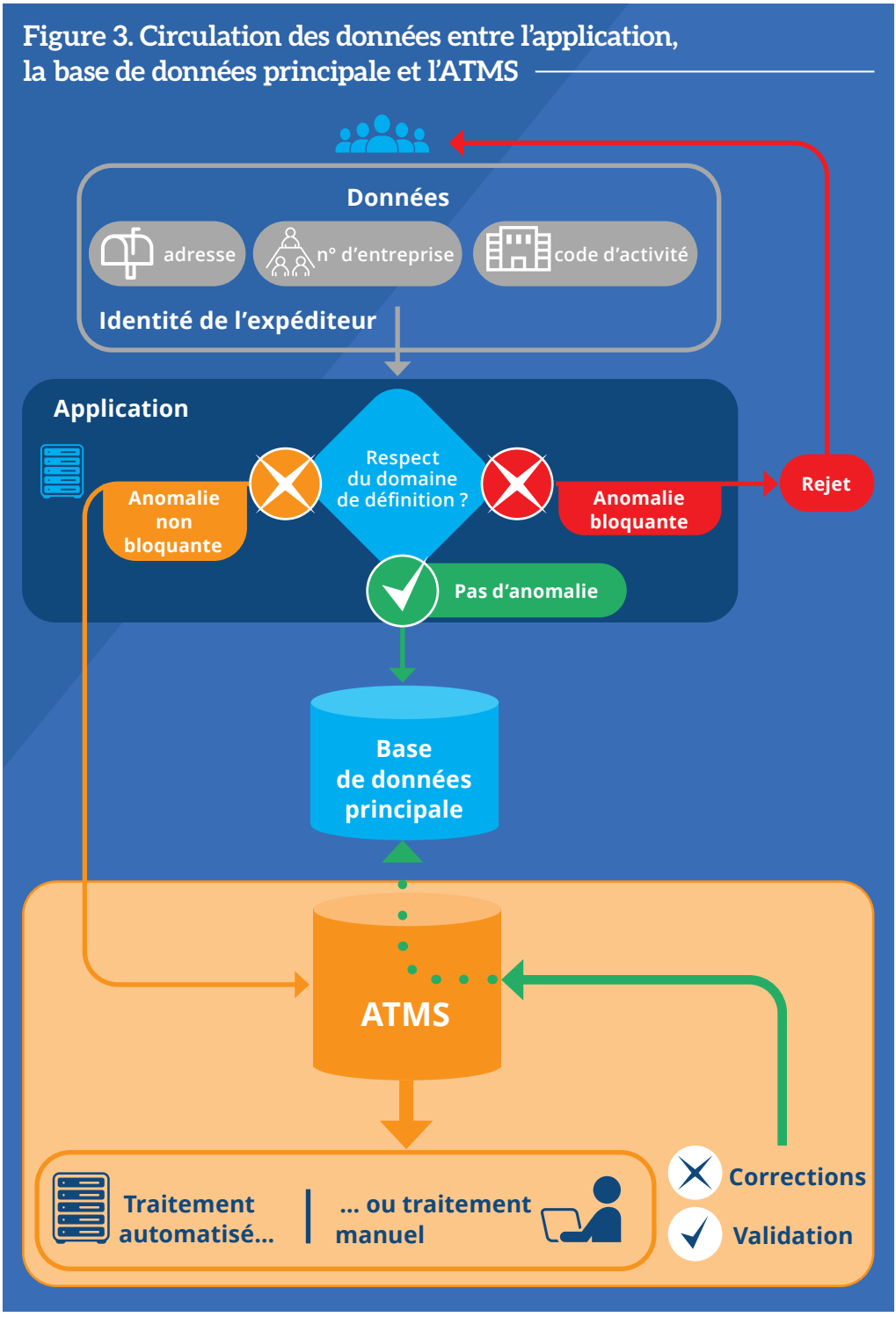
> *« Évaluer quantitativement la validité formelle des valeurs introduites dans une base de données et à en améliorer structurellement le traitement. »*

Une base de données s'apparente à un lac, selon Redman. Au lieu de nettoyer ponctuellement le fond du lac continûment alimenté par des flux et courants externes (comme le préconise le *data cleansing*, méthode de correction automatique), Redman propose, sur la base d'un échantillon aléatoire de données prélevé en entrée (amont) du système d'information, d'analyser méthodiquement les processus et les flux permettant l'assemblage des données. Le but final consiste à déterminer les causes des erreurs formelles identifiées afin d'y remédier structurellement à la source (bogues ou erreurs de programmation, par exemple).

L'opération repose sur le principe selon lequel un petit nombre de flux, processus ou pratiques sont à l'origine d'un pourcentage important d'erreurs formelles. L'approche fait référence au principe de Pareto également appelé « principe 80/20 ». Elle est ainsi basée sur l'hypothèse selon laquelle une part importante des cas problématiques (environ 80 %) est engendrée par environ 20 % des causes possibles.

11. Par domaine d'application, on entend, en modélisation de bases de données, le segment du réel observable que l'on tente de représenter à travers la base.

Figure 3. Circulation des données entre l'application, la base de données principale et l'ATMS



LE *BACK TRACKING* : DE L'ORTHOGONALITÉ À L'INTERACTION

Dans le courant des années 2000, l'application à grande échelle de la méthode de Redman (Boydens, 2010) a abouti à une méthode originale dénommée *back tracking* (Boydens, 2018). En se basant toujours sur le principe de Pareto posé par Redman, le *back tracking* enrichit sa méthode sur cinq aspects importants :

- le modèle de la base de données est étendu et relié à un historique des anomalies et de leurs traitements selon les principes exposés *supra* ;
- un suivi continu des cas jugés les plus stratégiques est mis en place à partir de l'ATMS, de façon à faciliter la gestion de la qualité de la base de données. Le suivi du traitement des anomalies permet de détecter, dans les domaines d'application fortement évolutifs, l'émergence de nouveaux phénomènes observables demandant une adaptation ponctuelle du domaine de définition de la base de données, voire des normes associées, en vue de diminuer le nombre d'anomalies fictives à traiter ;
- l'échantillon de cas retenus n'est pas aléatoire comme chez Redman, puisque l'on dispose d'une connaissance *a priori* concernant la totalité des données jugées problématiques (*via* l'historique exhaustif des anomalies et de leurs traitements). L'approche permet une sélection plus précise et représentative des cas à investiguer dès le début de l'opération, réduisant ainsi l'inévitable marge d'erreur d'un échantillon aléatoire ;
- au-delà de l'erreur formelle (déterministe et détectable *via* un algorithme, cas uniquement pris en compte par Redman), les questions d'interprétation des données au fil de l'évolution de la législation (ou de toute théorie empirique mobilisée) et des réalités appréhendées sont également abordées ;
- il s'agit d'un *tracking* arrière (ou *back tracking*) : on part, en synergie avec les fournisseurs de la base de données, de la situation finale (base de données principale et ATMS) pour revenir, étape par étape, à chaque source et processus qui en a permis l'élaboration, jusqu'à l'identification des causes à l'origine de cas problématiques. L'objectif est d'éviter le traitement de données ou de flux inutiles et de travailler de manière plus économe, renforçant les gains de l'opération. En effet, la recherche des origines structurelles des anomalies prend fin dès que toutes leurs causes par type[12] ont été détectées, sans que tous les flux ne soient inutilement parcourus (dans le cas du *data tracking* de Redman, tous les flux doivent être parcourus, ce qui constitue une perte de temps, ceux-ci pouvant inclure des dizaines, voire des centaines de processus).

> *« L'opération de back tracking repose ainsi sur un suivi préalable des anomalies et transactions, lui-même mis en place après spécification des indicateurs de qualité stratégiques. »*

L'opération de *back tracking* repose ainsi sur un suivi préalable des anomalies et transactions, lui-même mis en place après spécification des indicateurs de qualité stratégiques. Elle permet ensuite d'identifier, au sein des processus et des flux de données, en partenariat avec le fournisseur de l'information et le gestionnaire de la base[13], les éléments à l'origine

12. Comme évoqué *supra*, les anomalies sont spécifiées à un instant t par référence à un domaine de définition : valeur absente, valeur incohérente par rapport à une autre, etc. Ces types d'anomalies sont susceptibles d'évoluer dans le temps, si nécessaire, *via* une gestion de versions du domaine de définition.
13. Il s'agit du maître d'œuvre, du maître d'ouvrage ou de tout gestionnaire ayant une prise sur la base de données au moment de l'identification des éléments problématiques.

de la production d'un grand nombre d'anomalies systématiques ou jugées stratégiques : traitement inapproprié de certaines sources de données, émergence de situations nouvelles non encore prises en compte dans le domaine de définition de la base, interprétation inadéquate de la législation, concept mal documenté, erreurs de programmation, etc. Sur cette base, un diagnostic ainsi que des actions correctrices durables et structurelles peuvent être posés (correction de code formel dans les programmes, restructuration de processus, adaptation de l'interprétation d'une loi, clarification de la documentation, etc.).

Dans le domaine de la Sécurité sociale belge (en l'occurrence, à partir de la base de données DmfA mentionnée *supra*), les tests « grandeur nature » réalisés en synergie avec les développeurs, les spécialistes du domaine d'application ainsi que les expéditeurs de l'information furent concluants. Les apports de la méthode furent démontrés à maintes reprises dans les années deux mille (diminution structurelle du nombre d'anomalies de 50 % à 80 %, gain de temps grâce à une réduction du travail intellectuel fastidieux de vérification, meilleure interprétation de la loi, perception et redistribution financières plus rapides, etc.). Vu son caractère généralisable, la méthode fut actée dans la législation belge sous la forme d'un arrêté royal contraignant[14] en 2017 (Boydens, 2018). Cette législation s'inscrit dans le cadre de « baromètres de qualité » appliqués à la Sécurité sociale. Si les effets de la méthode sont durables et structurels, celle-ci doit être appliquée de manière récurrente afin de prendre en compte d'éventuels nouveaux phénomènes, ce qui demande toutefois un effort dégressif dans le temps. Elle repose sur une organisation, des procédures et un système d'information rigoureusement documentés (Boydens, 2010)[15].

Le mécanisme d'ATMS qui a soutenu les opérations de *back tracking* fut à ce jour déployé à grande échelle *via* un Système de gestion de base de données (SGBD) hiérarchique et du code externalisé spécifique associé à un moteur de règles générique. Désormais, un nouveau développement d'ATMS générique est applicable à tout SGBD relationnel et aux technologies actuelles.

L'ATMS RELATIONNEL : UNE DYNAMIQUE ENTRE DONNÉES EN PRODUCTION ET GESTION DES ANOMALIES

Le *Centre de Compétence en Qualité de Données* de Smals[16], qui repose sur une synergie entre la section *Databases* et la section Recherche, a entrepris la mise en place d'un prototype appliqué aux SGBD relationnels et aux standards associés. Le prototype repose sur la version *open data* complète de la Banque Carrefour des Entreprises, correspondant belge du répertoire des entreprises Sirène en France.

Dans ce nouveau modèle, la base de données principale et l'ATMS sont séparés, ce qui nécessite de convenir du routage des données entre les deux systèmes et de définir les principes permettant de stocker les anomalies et transactions associées.

14. Voir références juridiques en fin d'article.
15. Concrètement, au-delà d'un certain seuil d'anomalies fixé par l'administration, les fournisseurs de déclarations sociales sont contraints d'en diminuer le nombre dans un délai donné, en participant à une opération de *back tracking*.
16. Smals est une société informatique créée en 1939 prestataire de services pour l'administration fédérale et régionale belge.

BASE DE DONNÉES PRINCIPALE ET ATMS : LES DEUX HÉMISPHÈRES DU MONDE REPRÉSENTÉ

L'un des traits majeurs de la spécification de l'ATMS relationnel est la séparation induite entre la base de données principale, d'une part, et la base de données des anomalies et de leurs traitements, d'autre part. La ***figure 4*** illustre cette séparation, qui repose sur les deux principes suivants :

- la base de données principale représente les concepts du domaine d'activité qu'elle sert. À ce titre, elle est modélisée en vue de traiter des données qui respectent strictement le domaine de définition, en vertu de l'hypothèse du monde clos (voir *supra*) ;
- il est possible de factoriser l'enregistrement et le traitement des données en anomalie dans un système dédié distinct de la base de données principale, à condition que celui-ci soit conçu de façon extensible.

Au-delà de la satisfaction d'hypothèses de conception théoriques, la séparation entre base de données principale et ATMS ouvre la porte à un certain nombre d'avantages non négligeables[17] :

- la conception de la base de données principale est simplifiée, puisqu'elle s'affranchit de la prise en compte des anomalies ou de leur traitement ;
- le contenu de la base de données principale reste en permanence conforme au domaine de définition, sans en empêcher l'évolution ;
- la gestion des anomalies peut faire l'objet de processus et outils standardisés, réutilisables à l'échelle d'une équipe, d'un projet ou d'une organisation entière ;
- l'existence d'un ATMS dédié encourage une prise en compte accrue du domaine de définition au plus tôt dans la conception du système informatique, qui devra se charger de la détection des anomalies et du routage des données en conséquence.

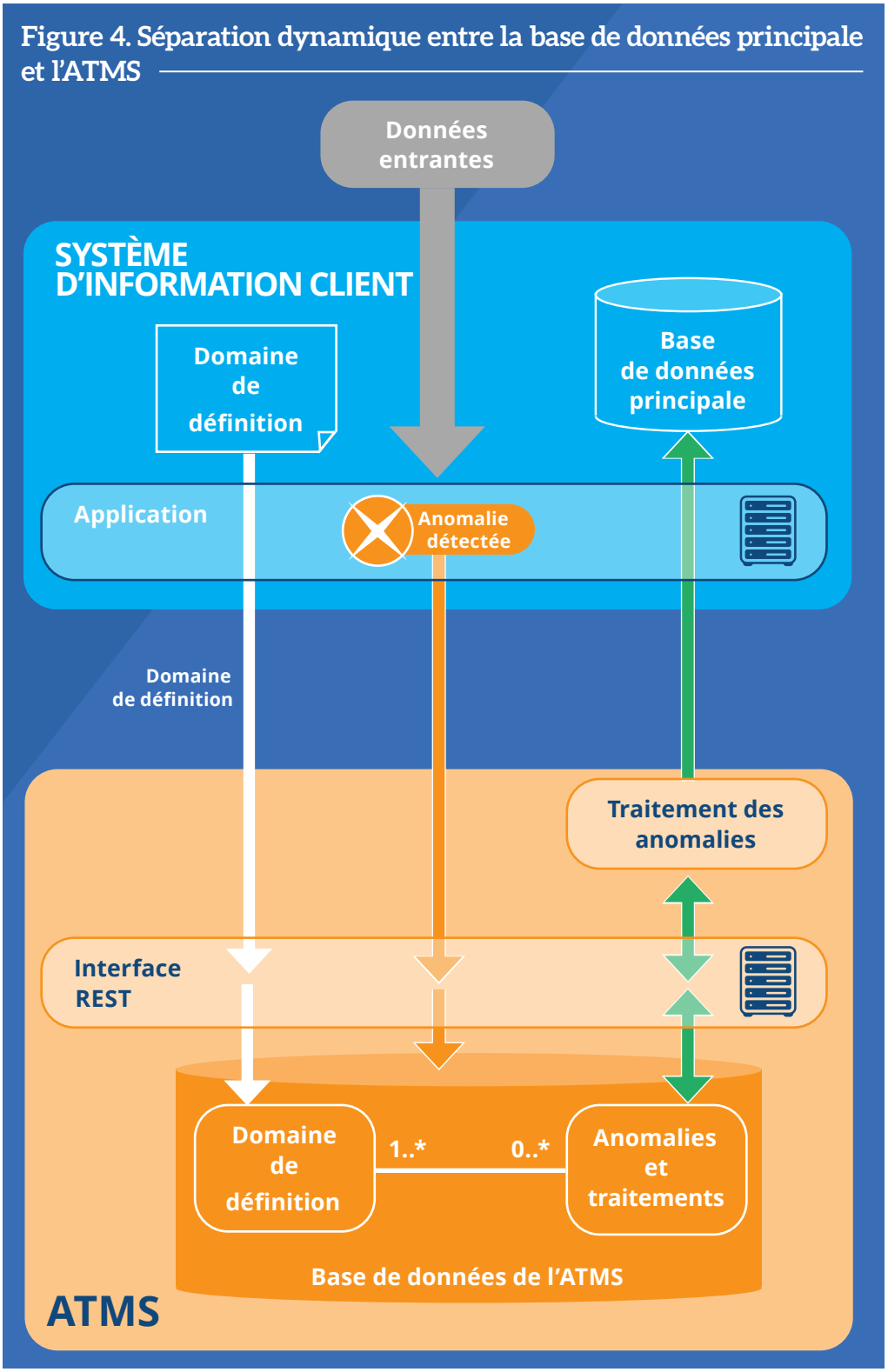
ROUTAGE DES DONNÉES

Lorsque des données entrent dans le système d'information, leur conformité au domaine de définition (voir *supra*) est vérifiée. Si des anomalies non bloquantes sont détectées, les données correspondantes sont envoyées dans l'ATMS ***(figure 3)***. Ce n'est qu'après avoir été sujettes à vérification et traitement intellectuel, historisé automatiquement, que les données seront intégrées à la base de données principale[18]. Dans un cas de validation, tel qu'évoqué précédemment, il se peut que le domaine de définition doive être adapté avant que les enregistrements ne puissent être acceptés dans la base de données principale.

Ce routage automatique n'exclut aucunement la possibilité qu'un agent déclenche manuellement une anomalie sur des données conformes au domaine de définition et déjà présentes dans la base de données principale. Ceci pourrait se justifier, par exemple, par une inspection de terrain qui révèle qu'un enregistrement est frauduleux ou obsolète.

17. En corollaire, la resynchronisation de l'état des deux bases de données en cas de catastrophe doit être pensée en amont. Diverses options plus ou moins sophistiquées sont possibles, allant de l'arrêt pur et simple de l'un des deux systèmes si l'autre ne répond plus, au recours à une file de messages au sein d'un système tiers dédié.
18. L'état complet du système d'information se compose donc de l'union – indifféremment inclusive ou exclusive, l'intersection n'étant de toute façon pas acceptée – entre les données valides et les anomalies non traitées, distribuées respectivement dans la base de données principale d'une part et dans l'ATMS d'autre part.

Figure 4. Séparation dynamique entre la base de données principale et l'ATMS



La circulation de données entre la base principale et l'ATMS doit être fluide et standardisée ; elle est donc implémentée dans des procédures automatisées, dont le principe a déjà été illustré une première fois par la ***figure 3***. La nature bidirectionnelle de cette circulation requiert de pouvoir préserver ou reconstruire l'état original de l'information à partir de l'ATMS. À cette fin, deux parties constitutives permettent de stocker respectivement le domaine de définition ainsi que les anomalies et métadonnées associées ***(figure 4)*** ; nous ne détaillerons pas ce fonctionnement ici mais plus de références sont disponibles en ligne (Boydens, Hamiti et Van Eeckhout, 2020).

IMPLÉMENTATION DU PROTOTYPE ET PERSPECTIVES

Le prototype implémente intégralement la base de données de l'ATMS. Celui-ci a été développé et testé de façon itérative en l'exposant, très tôt, à un système d'information simulé par une application rudimentaire et une base de données principale réaliste, les *open data* du Répertoire des entreprises belges ***(encadré 2)***. Dans ce cadre, quatre scénarios d'utilisation ont été implémentés à titre d'exemple :

- la correction d'une valeur simple en anomalie ;
- la correction d'une anomalie déclenchée par l'incompatibilité de données entrantes avec un enregistrement déjà existant dans la base de données principale ;
- la validation d'anomalies par lots : le traitement d'une anomalie (ici la validation) est propagé automatiquement à une série d'autres anomalies désignées comme similaires par l'agent ;

Encadré 2. Quelques détails techniques sur l'ATMS

Le prototype décrit dans cet article a été développé sur le SGBD PostgreSQL ; il est néanmoins transposable à n'importe quel système permettant la manipulation de données au format JSON. Cette notation « sans schéma » permet de stocker les anomalies les plus variées au sein d'une table relationnelle classique, ainsi que de les échanger sous forme de messages entre le système d'information principal et l'ATMS. Les traitements produisant et consommant ces messages JSON sont implémentés de part et d'autre sous la forme de procédures stockées accessibles en requêtant directement les bases de données ; en pratique, dans un projet de production, cette logique serait typiquement exposée *via* des interfaces REST.

Dans le cadre de ce prototype, le volume de la base de données principale est d'environ 4,4 gibibits (GiB) pour un total de 30,2 millions d'enregistrements répartis essentiellement sur 9 tables. L'exécution des trois premiers scénarios de la façon la plus exigeante possible (exécution en boucle de 100 000 itérations sans aucune optimisation explicite*) a permis d'observer :

- un temps d'exécution constant, souvent de l'ordre du millième de seconde ou moins, pour la plupart des opérations de calcul et d'insertion légère (par exemple, générer un message de création d'anomalie complet, allouer une anomalie pour traitement par un agent, marquer une anomalie comme traitée) ;
- une croissance linéaire du temps d'exécution pour les opérations d'écriture plus importantes (enregistrement de la version corrigée d'une anomalie) ;
- une croissance linéaire de la consommation d'espace de stockage.

** Par exemple, PostgreSQL permet de déclencher explicitement des opérations comme « VACUUM ANALYZE » afin d'optimiser le stockage et les plans d'exécution des requêtes. D'autres SGBD relationnels offrent des fonctionnalités analogues.*

- enfin, la création de plusieurs vues permettant de suivre les anomalies et leurs traitements dans le temps à un niveau plus ou moins agrégé[19]. Ce quatrième scénario est fondamental pour orienter les opérations de *back tracking* telles que décrites précédemment.

Le prototype ayant rencontré un succès technique et suscité l'intérêt de la maîtrise d'ouvrage, un projet pilote est en cours. Cette initiative pourra ultérieurement être appliquée grandeur nature à des bases de données de l'administration belge. Elle est par ailleurs généralisable à tout système d'information. Plusieurs cas de figure sont possibles pour la mise en œuvre d'un ATMS. Idéalement, celui-ci est envisagé dès la conception du système d'information, en lien avec la base de données principale et les outils de qualité de données ***(encadré 1)***, si l'on en dispose. Un projet de réingénierie représente également un moment opportun pour l'intégration d'un ATMS à un système existant[20].

Les apports récurrents du *back tracking* évoqués plus haut, méthode nécessairement supportée par un ATMS, appliqué à la base de données DmfA constituent un précédent encourageant fortement la généralisation de ce prototype adapté à des technologies récentes : diminution structurelle du nombre d'anomalies de 50 % à 80 %, gain de temps grâce à une réduction du travail intellectuel fastidieux de vérification, meilleure interprétation de la norme, perception et redistribution financières plus rapides, de manière plus générale, amélioration de la qualité de toute base empirique, mise en place d'un partenariat entre les gestionnaires de la base et les fournisseurs de l'information, etc. Comme nous l'avons vu, le service peut bénéficier tant aux gestionnaires des bases de données administratives dont la qualité est améliorée qu'aux statisticiens souhaitant les exploiter à d'autres fins.

19. Types d'anomalies les plus fréquents, anomalies validées par qui et quand, anomalies non traitées, etc.
20. Si le système original dispose déjà d'une forme de gestion des anomalies, la transposition de ce contenu à l'ATMS peut cependant représenter un certain effort en fonction du degré auquel les anomalies et leurs métadonnées de traitement sont fondues et éclatées dans le système.

BIBLIOGRAPHIE

AGENCE BELGA, 2018. Des lacunes dans la base de données belge sur les terroristes. In : *La Libre Belgique*. [en ligne]. 1er mars 2018. [Consulté le 31 mai 2021]. Disponible à l'adresse : *https://www.lalibre.be/belgique/des-lacunes-dans-la-base-de-donnees-belge-sur-les-terroristes-5a97a5f4cd700399f72087da*.

ARON, Raymond, 1969. *La philosophie critique de l'histoire.* 1969. Édition Librairie philosophique J. Vrin. Collection Points – Sciences humaines. ISBN 2560848158182.

BADE, David, 2011. It's about Time!: Temporal Aspects of Metadata Management in the Work of Isabelle Boydens. In : *Cataloging & Classification Quarterly (The International Observer).* 16 mai 2011. Volume 49, n° 4, pp. 328-338.

BATINI, Carlo et SCANNAPIECO, Monica, 2016. *Data and Information Quality. Dimensions, Principles and Techniques*. Springer, New York. ISBN 978-3-319-24106-7.

BENS, Arno et SCHUKRAFT, Stefan, 2019. Modernisation des registres administratifs en Allemagne – Développements actuels et enjeux pour la statistique publique. In : *Courrier des statistiques*. [en ligne]. 27 juin 2019. Insee. N° N2, pp. 10-20. [Consulté le 31 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4168390/courstat-2-3.pdf*.

BOYDENS, Isabelle, 1999. *Informatique, normes et temps*. Bruylant, Bruxelles. ISBN 2-8027-1268-3.

BOYDENS, Isabelle, 2010. Strategic Issues Relating to Data Quality for E-government: Learning from an Approach Adopted in Belgium. In : *Practical Studies in E-Government. Best Practices from Around the World.* 17 novembre 2010. Springer. New York. pp. 113-130 (chapitre 7). ISBN 978-1489981899.

BOYDENS, Isabelle, 2012. L'océan des données et le canal des normes. In : *La normalisation : principes, histoire, évolutions et perspectives.* [en ligne]. Juillet 2012. Annales des Mines, Responsabilité et Environnement. Édition FFE. Paris. N° 2012/3 (67), pp. 22-29. [Consulté le 31 mai 2021]. Disponible à l'adresse : *http://www.ulb.ac.be/cours/iboydens/annales.pdf*.

BOYDENS, Isabelle, 2018. *Data Quality & « back tracking » : depuis les premières expérimentations à la parution d'un Arrêté Royal.* [en ligne]. 14 mai 2018. Smals Research. [Consulté le 31 mai 2021]. Disponible à l'adresse : *https://www.smalsresearch.be/data-quality-back-tracking-depuis-les-premieres-experimentations-a-la-parution-dun-arrete-royal/*.

BOYDENS, Isabelle, 2021. *Qualité de l'information et des documents numériques*. Cours dispensé à l'Université libre de Bruxelles, Master en sciences et technologies de l'information et de la communication.

BOYDENS, Isabelle, HAMITI, Gani et VAN EECKHOUT, Rudy, 2020. *Data Quality: "Anomalies & Transactions Management System" (ATMS), prototype and "work in progress"*. [en ligne]. 8 décembre 2020. Smals Research. [Consulté le 31 mai 2021]. Disponible à l'adresse : *https://www.smalsresearch.be/data-quality-anomalies-transactions-management-system-atms-prototype-work-in-progress/*.

BRAUDEL, Fernand, 1949. *La Méditerranée et le monde méditerranéen à l'époque de Philippe II.* Armand Colin, Paris.

BYRNES, Nanette, 2016. Why we should expect algorithms to be biased. In : *MIT Technology Review.* [en ligne]. 24 juin 2016. [Consulté le 31 mai 2021]. Disponible à l'adresse : *https://www.technologyreview.com/2016/06/24/159118/why-we-should-expect-algorithms-to-be-biased/*.

CHAPUIS, Nicolas, 2018. Des fichiers de police mal organisés et trop complexes. In : *Le Monde.* 17 octobre 2018.

DIERICKX, Laurence, 2019. *Why News Automation Fails*. [en ligne]. Février 2019. Computation & Journalism Symposium, Miami, USA. [Consulté le 31 mai 2021]. Disponible à l'adresse : *http://mastic.ulb.ac.be/wp-content/uploads/2019/02/Why_news_automation_fails.pdf*.

ELIAS, Norbert, 1986. *Du temps.* Édition Fayard. Paris. ISBN 978-2818503454.

HAINAUT, Jean-Luc, 2018. *Bases de données – Concepts, utilisation et développement.* Octobre 2018. Édition Dunod, Paris, collection InfoSup. 4e édition. ISBN 978-2100790685.

HAMITI, Gani, 2019. *Data Quality Tools: concepts and practical lessons from a vast operational environment*. [en ligne]. 13 mars 2019. Université libre de Bruxelles. Cours-conférence. [Consulté le 31 mai 2021]. Disponible à l'adresse : *https://mastic.ulb.ac.be/wp-content/uploads/2019/03/Data_Quality_Tools_ULB_2019.pdf*.

HAND, David J., 2018. Statistical challenges of administrative and transaction data. In : *Journal of the Royal Statistical Society.* [en ligne]. Series A, 181, Part 3, pp. 555-605. [Consulté le 31 mai 2021]. Disponible à l'adresse : *https://spiral.imperial.ac.uk/bitstream/10044/1/61527/2/Statistical%20challenges%20of%20administrative%20and%20transaction%20data%20FINAL%20version.pdf*.

HUMBERT-BOTTIN, Élisabeth, 2018. La déclaration sociale nominative. Nouvelle référence pour les échanges de données sociales des entreprises vers les administrations. In : *Courrier des statistiques.* [en ligne]. 6 décembre 2018. Insee. N° N1, pp. 25-34. [Consulté le 31 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/3647025/courstat-1-6.pdf*.

MADNICK, Stuart E., WANG, Richard Y., LEE, Yang W. et ZHU, Hongwei, 2009. Overview and Framework for Data and Information Quality Research. In : *Journal of Data and Information Quality.* [en ligne]. 1er juin 2009. Volume 1, n° 1, pp 1–22. [Consulté le 31 mai 2021]. Disponible à l'adresse : *https://dl.acm.org/doi/10.1145/1515693.1516680*.

RADIO, Eric, 2014. Information Continuity: A Temporal Approach to Assessing Metadata and Organizational Quality in an Institutional Repository. In : *Metadata and Semantics Research*. 27-29 novembre 2014. 8th Research Conference, MTSR 2014, Karlsruhe. Springer, Cham. Communications in Computer and Information Science, vol 478. ISBN 978-3-319-13673-8.

REDMAN, Thomas C., 1996. *Data Quality for the Information Age*. Artech House Computer Science Library. ISBN 978-0890068830.

RENNE, Catherine, 2018. Bien comprendre la déclaration sociale nominative pour mieux mesurer. In : *Courrier des statistiques*. [en ligne]. 6 décembre 2018. Insee. N° N1, pp. 35-44. [Consulté le 31 mai 2021]. *https://www.insee.fr/fr/statistiques/fichier/3647029/courstat-1-7.pdf*.

RIVIÈRE, Pascal, 2018. Utiliser les déclarations administratives à des fins statistiques. In : *Courrier des statistiques*. [en ligne]. 6 décembre 2018. Insee. N° N1, pp. 14-24. [Consulté le 31 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/3647013/courstat-1-5.pdf*.

RIVIÈRE, Pascal, 2020. Qu'est-ce qu'une donnée ? Impact des données externes sur la statistique publique. In : *Courrier des statistiques*. [en ligne]. 31 décembre 2020. N° N5. [Consulté le 31 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/5008707/courstat-5-8.pdf*.

SRIVASTAVA, Divesh, SCANNAPIECO, Monica et REDMAN, Thomas C., 2019. Ensuring High-Quality Private Data for Responsible Data Science: Vision and Challenges. In : *Journal of Information and Data Quality.* 4 janvier 2019. Volume 1, n° 11, pp. 1-9.

VAN DER VLIST, Eric, 2011. *Relax NG*. Mai 2011. Édition O'Reilly Media. ISBN: 0596004214.

FONDEMENTS JURIDIQUES

Arrêté royal du 2 février 2017 modifiant le chapitre IV de l'arrêté royal du 28 novembre 1969 pris en exécution de la loi du 27 juin 1969 révisant l'arrêté-loi du 28 décembre 1944 concernant la sécurité sociale des travailleurs. In : Moniteur belge. [en ligne]. [Consulté le 31 mai 2021]. Disponible à l'adresse : *http://www.ejustice.just.fgov.be/mopdf/2017/02/20_2.pdf#Page113*.