L'océan des données et le canal des normes

Par Isabelle BOYDENS*

Si Xerxès fouettait la mer, les concepteurs de bases de données tentent, quant à eux, de canaliser l'océan. Le défi qu'ils doivent relever est celui de soumettre l'irréductible hétérogénéité d'observations empiriques à l'homogénéité des normes formelles. Cette interaction continue a des effets colossaux dans le domaine social, juridique, médical, militaire ou encore environnemental. Son étude nous conduit à des protocoles opérationnels qui diminuent le risque d'une inadéquation de l'information.

Introduction

En 1442, dans son *De falso credita et ementita Constantini donatione declamatio*, l'humaniste italien Lorenzo della Valle (dit Laurent Valla) démontra que la *Donation de Constantin*, cet acte par lequel l'Empereur Constantin aurait fait don de Rome et de son territoire au Pape Sylvestre Ier, était un faux antidaté de quatre ou cinq siècles [28]. En 1986, soit cinq siècles plus tard, 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 [36]. En effet, la théorie scientifique de l'époque, qui avait été modélisée dans leurs bases de données, ne permettait pas de concevoir que de telles valeurs puissent être correctes.

Quelle est la parenté entre ces deux événements ? Qu'il s'agisse de chartes médiévales ou de vastes bases de données distribuées contemporaines, la connaissance empirique se transforme sous l'effet de l'interprétation que l'on fait des informations normées qui en permettent l'appréhension. Dans ce numéro spécial de *Responsabilité & Environnement* consacré aux normes, nous nous proposons d'analyser un aspect de cette question qui fait l'objet de nos recherches depuis une quinzaine d'années : l'évaluation et l'amélioration de la qualité des bases de données au fil de leurs multiples transformations tout au long de leur cycle de vie [6, 8, 9, 10, 11].

Dérivé du nom latin *norma*, équerre, règle, le concept de norme désigne, littéralement, *« la formule abstraite ou le type concret de ce qui doit être »* (1). Comme le souligne François Ewald, une norme, en tant que principe de comparaison et d'évaluation, renvoie à la notion de commune mesure [17]. Ce faisant, elle est intrinsèquement historique et évolutive. La notion de perfection lui est donc étrangère. S'agissant des bases de données, certains proposent pourtant une approche positiviste de la norme reposant sur l'hypothèse de l'existence d'un isomorphisme entre les bases de données et le réel observable que celles-ci représentent : une donnée est jugée correcte si elle correspond « parfaitement » à la réalité observable représentée [35]. Par ailleurs, la récente norme internationale ISO 8000 (2) propose, depuis 2009, un *Master Data Vocabulary* en vue d'améliorer la qualité des données en fournissant des listes de vocabulaire standardisé universel de référence pour assurer une modélisation uniforme de l'information. L'initiative est intéressante, mais le recours à un tel lexique est délicat en raison du caractère très volatile tant du langage naturel que des processus et des réalités représentés, cela dans tous les domaines empiriques qui sont par essence sujets à interprétations dans le temps et l'espace.

Dans nos travaux, nous avons montré que le postulat d'un isomorphisme entre une base de données et le réel empirique correspondant n'était pas valide sur le plan épistémologique et qu'il était *a fortiori* infructueux sur le plan opérationnel (3). En effet, les questions que soulève la gestion des bases de données tiennent, selon nous, au décalage temporel qui sépare inéluctablement les normes informatiques déterministes (sur lesquelles repose le cadre opératoire des bases de données), les normes empiriques constitutives du domaine d'application représenté et le flux continu du réel observable sous-jacent [6]. L'évolution respective de ce double système de normes et du réel normé est nécessairement asynchrone. De ce décalage temporel découle un décalage conceptuel dont nous nous sommes attachée à analyser les incidences pratiques et épistémologiques en vue de formuler des solutions opérationnelles destinées à faciliter la gestion des bases de données.

C'est au XVe siècle, à partir de la célèbre étude philologique et comparative de Laurent Valla évoquée plus haut, que l'on date traditionnellement les débuts de la critique historique. À l'époque (l'imprimerie n'existait pas encore), le

© Sekulic/SIPA

« En mai 1999, pendant son intervention au Kosovo, l'Otan a bombardé par erreur l'ambassade de Chine à Belgrade : les bases de données cartographiques alors utilisées pour guider les missiles répertoriaient un plan de la ville obsolète et, donc, inadéquat. », *manifestation devant l'ambassade de Chine à Belgrade pendant la guerre du Kosovo.*

processus de transformation de l'information se déployait de siècle en siècle, d'une génération de moines copistes à l'autre. Afin d'établir l'appareil critique d'un manuscrit (dont on dispose souvent de multiples copies divergentes et dont on a parfois perdu l'original), l'historien construit un *stemma codicum* (une généalogie des données) selon une technique d'analyse comparative des variantes empruntée à la philologie [22]. L'établissement d'un *stemma codicum* accompagné d'un travail d'interprétation critique permet une reconstruction conjecturale argumentée du manuscrit original.

En vue d'appliquer une démarche analogue dans le cadre technologique des sources informatiques [5], il importe d'en étudier, de l'intérieur, les modalités de fonctionnement. Nos travaux ont débuté avec l'étude des bases de données de la Sécurité sociale belge, qui permettent le prélèvement et la redistribution chaque année de quelque 40 milliards d'euros (4). Ces bases de données représentent des enjeux considérables, l'information administrative étant créatrice de droits et de devoirs. Nous avons été ainsi amenée à produire une méthode d'analyse généralisable à l'évaluation et à l'amélioration de la qualité de vastes systèmes administratifs d'information. Si nous nous proposons d'exposer ici les mécanismes de notre méthode, c'est parce qu'ils sont applicables à de nombreuses bases de données empiriques issues d'autres domaines. Nous développerons trois étapes successives : a) la problématique de la qualité des bases de données, b) les processus de construction de l'information et, enfin, c) les incidences opérationnelles sur la théorie de la modélisation des bases de données.

Problématique de la qualité des bases de données : définitions et exemples

Par la qualité d'une base de données, on désigne sa relative adéquation aux objectifs qui lui sont assignés. La « qualité totale » n'existe pas, le concept étant relatif : à partir d'un arbitrage du type coûts/bénéfices, les critères de qualité les plus pertinents (fraîcheur de l'information, rapidité de la transmission des données, précision, ...) devront être retenus dans un contexte donné. On parlera de *fitness for use*, d'adéquation aux usages d'une base de données [10].

La qualité des bases de données est aujourd'hui considérée comme un facteur stratégique. La question soulève en effet des enjeux considérables dès lors que l'information est un instrument d'aide à la prise de décision, voire un instrument permettant d'agir sur le réel. Ainsi, en mai 1999, pendant son intervention au Kosovo, l'Otan a bombardé par erreur l'ambassade de Chine à Belgrade : les bases de données cartographiques alors utilisées pour guider les missiles répertoriaient un plan de la ville obsolète et, donc, inadéquat. D'où cette attaque inopportune et l'incident diplomatique qui ont suivi (voir la photo ci-dessus) [8]. On trouvera d'autres exemples dans le domaine de l'environnement, en relation notamment avec la difficulté de gérer les référentiels d'adresses [13]. Si les questions liées à la normalisation

et à l'identification des adresses spatiales [27] se déclinent distinctement dans les pays développés (comme le Danemark (5)) ou en voie de développement (comme l'Afrique du Sud (6)), la qualité des systèmes d'information correspondants, en tant que référentiels, a un impact sur de nombreuses bases de données auxquelles ils sont interconnectés, cela dans des domaines stratégiques d'application les plus divers. Ainsi, il est parfois extrêmement difficile, de nos jours, de détecter rapidement les destructions illégales de bâtiments contenant de l'amiante [34] ou encore, en cas de présomption d'épidémie ou dans le contexte du contrôle de denrées, d'effectuer efficacement le suivi de la chaîne alimentaire d'un pays à l'autre, voire à l'intérieur d'un pays [27].

Avec le spectaculaire développement des réseaux, les difficultés sont exacerbées : les informations incohérentes ou incomplètes sont plus rapidement et plus massivement transmises d'un système d'information à l'autre. En corollaire, des bases de données conçues à certaines fins sont fréquemment exploitées à d'autres, l'utilisateur final se trouvant de plus en plus éloigné de la source productrice de l'information. Ainsi, pendant la première guerre du Golfe, environ 28 000 des 40 000 containers militaires américains envoyés au Moyen-Orient durent être inspectés et inventoriés manuellement tant l'interrogation des bases de données censées en répertorier le contenu donnait lieu à des résultats incohérents [21]. D'où cette remarque non dépourvue d'amertume : "*In general, the physical movement of material is faster than the movement of the supporting information...*" [Bien souvent, le flux concret des matériels est plus rapide que celui des informations logistiques les concernant] (7). Ces questions sont l'objet de préoccupations croissantes. Ainsi, plusieurs enquêtes menées aux Etats-Unis concluent que des facteurs, tels que la multiplication de sources hétérogènes partiellement redondantes, de données incomplètes ou mal documentées, entraînerait un coût pouvant représenter jusqu'à 15 % des revenus des entreprises [19, 24].

Le processus de construction de l'information normée dans les bases de données

Afin d'appréhender notre problématique sur le plan opérationnel, nous nous poserons successivement les trois questions suivantes : a) qu'est-ce qu'une donnée ? b) qu'est-ce qu'une donnée "correcte" ? et, enfin, c) comment l'information se construit-elle, progressivement ?

Qu'est-ce qu'une donnée ?

La modélisation d'une base de données repose sur le principe de l'abstraction, par lequel l'esprit humain sélectionne certaines caractéristiques et propriétés d'un ensemble d'objets et en exclut d'autres aspects considérés comme non essentiels [15]. Une base de données est donc une collection organisée de données structurées représentant certains aspects du réel observable. Une donnée est un triplet (i, d, v) composé des trois éléments suivants : un intitulé (i) renvoyant à un concept (une catégorie d'*activité*, par exemple), un domaine de définition (d) composé d'assertions formelles (contraintes d'intégrité) 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). Une fois mise en place, la base de données est intégrée dans le cadre plus large d'un système d'information incluant des flux entrants et sortants.

Il est important de distinguer les *données déterministes* des *données empiriques* [6, 10]. Les premières se caractérisent par le fait que l'on dispose à tout moment d'une théorie qui permette de décider si une valeur (v) est ou non correcte. Il en est ainsi 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 s'il ne l'est 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. Il en va ainsi, par exemple, du domaine médical (dans lequel la théorie évolue au fil des expériences, comme en témoignent les recherches actuelles sur la maladie d'Alzheimer, les données dans le domaine génétique [20], dans celui de la transplantation d'organes [19] ou encore les vastes enquêtes diachroniques en vue d'identifier les facteurs à l'origine de maladies cardiaques [1]), du domaine économique (en ce qui concerne l'évaluation de la richesse nationale, par exemple [6]) ou de l'énergie (en ce qui concerne le suivi des stocks de pétrole, qui requiert le recours à des bases de données distribuées hétérogènes [18]). On retrouve la même problématique dans le secteur administratif, où l'interprétation des concepts légaux se transforme avec l'évolution continue de la réalité traitée et avec celle de la jurisprudence [6]. Ainsi, la notion d'activité principale d'une entreprise, qui est fondamentale dans le répertoire français des entreprises (Sirène), est une notion évolutive dont la fiabilité est difficile à évaluer [26].

Qu'est-ce qu'une donnée dite correcte ?

Pour des raisons opérationnelles évidentes, le fonctionnement d'une base de données repose sur *l'hypothèse d'un monde clos* en vertu de laquelle toute valeur non incluse dans le domaine de définition de la base sera considérée comme fausse. Les enregistrements d'une base de données sont substituables *salva veritate*, c'est-à-dire sans que leur valeur de vérité ne se modifie. En d'autres termes, à chaque état d'une base de données complète (toutes les valeurs dérivant logiquement d'un état donné de la base sont présentes) et cohérente (toutes les valeurs présentes sont formellement valides) est associée la valeur de vérité : "*vrai*". Toutefois, s'agissant de 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. Dans ce cas, il est impossible de vérifier l'exactitude des valeurs de la base de données de manière automatique. Dès lors, lorsqu'une incohérence apparaît entre une valeur saisie dans la base et les tables de référence permettant d'en tester la validité, il peut

s'avérer indispensable (selon l'importance de l'enjeu) de procéder à une vérification manuelle (en contactant le citoyen ou l'entreprise concerné, si l'on prend l'exemple des bases de données administratives) [6, 10].

On ne dispose en effet d'aucun référentiel formel absolu qui permettrait de tester la correction d'une vaste base constituée de données empiriques. Les bases de données sont comparables à une paire de lunettes sans lesquelles nous verrions très mal, mais que nous ne pouvons ôter pour comparer la « vraie » réalité à celle que nous voyons à travers elles. Dans un autre domaine, la question a été exposée avec une grande clarté par Jean-Louis Besson à propos de l'épistémologie des statistiques [4]. La question de l'absence de référentiel se traduit comme suit : afin de vérifier la correction de l'information répertoriée dans une base de données, il faudrait idéalement connaître *a priori* une réalité que seule cette base de données permet de connaître. 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 (manuelle ou intellectuelle). Dans la pratique, cette distinction n'est pas aisée à opérer, mais le flou n'a pas droit de citer dans une base de données : il faut donc trancher. Pour ce faire, il s'agira souvent de se rendre sur le terrain afin d'interpréter les situations de fait et d'examiner des pièces justificatives. Au fil des interprétations et de l'évolution de la jurisprudence, la signification des notions d'employé et d'ouvrier évoluera dans le temps. On peut conclure de ce mécanisme que les données ne sont jamais définitives et qu'elles se construisent progressivement [10]. C'est pourquoi, à la question « l'information est-elle correcte ? », il faut substituer celle-ci : « comment l'information se construit-elle, progressivement ? ».

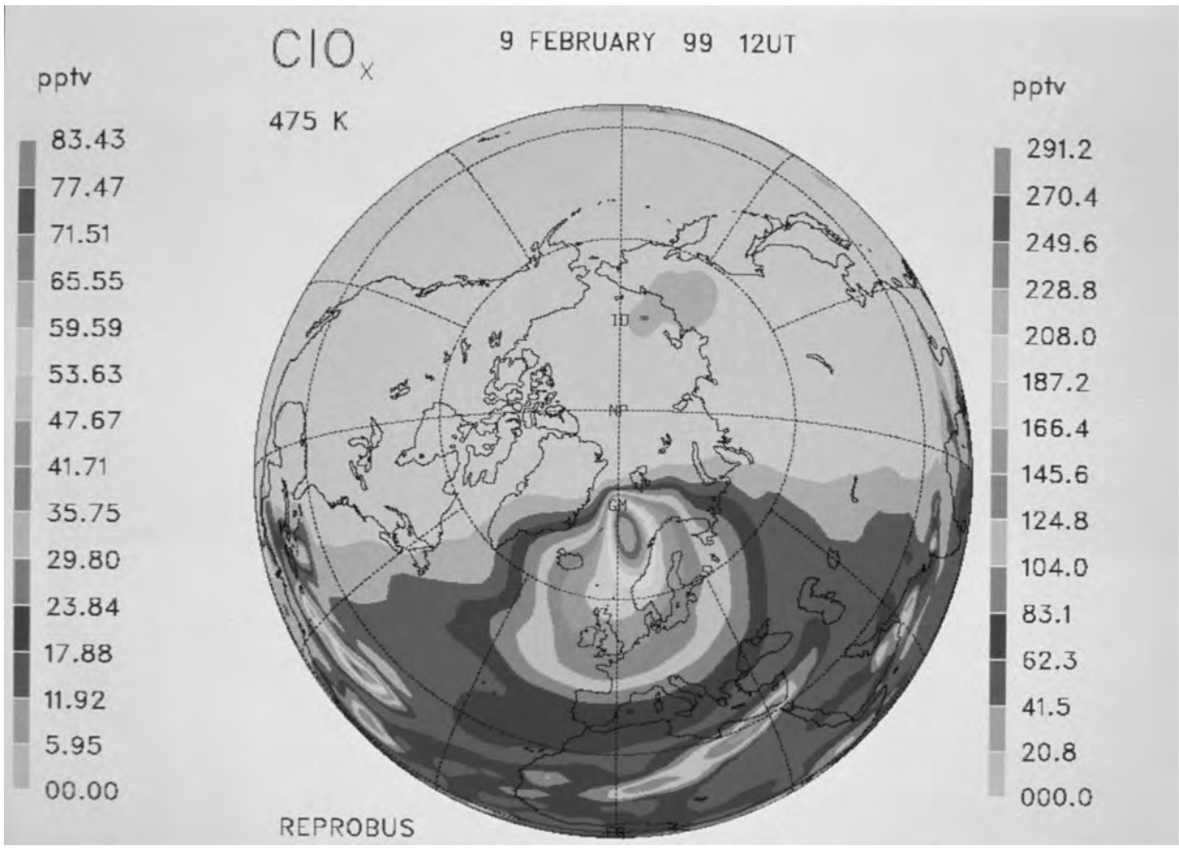
Comment les données se construisent-elles, progressivement ?

Afin d'aborder cette problématique, une approche herméneutique s'impose [6, 11]. Les faits empiriques (sujets à interprétation humaine) doivent être évalués par rapport à un principe unificateur, un horizon de similitude de sens. L'approche herméneutique consiste en une mise en relation sans cesse renouvelée entre les observations et le contexte dans lequel celles-ci s'insèrent [2, 14]. La question du temps est donc centrale. Nous recourrons à deux modèles temporels que nous offre l'herméneutique, la *temporalité étagée* de Fernand Braudel [12] et le *continuum évolutif* de Norbert Elias [16].

Le concept des temporalités étagées est une construction permettant d'identifier au sein d'un objet d'étude une hiérarchie entre plusieurs séquences de transformation coexistant. Dans le modèle de Braudel, les séquences relativement les plus rapides (les évolutions politiques, par exemple) sont conditionnées par des séquences relativement plus lentes (comme les évolutions de la géographie et celles du climat). Appliqué à un système d'information empirique, ce concept clarifie le processus de construction de l'information et permet d'identifier plusieurs échelles de transformation inter-agissantes, dont l'évolution est solidaire, mais asynchrone. Par exemple, dans un système d'information administratif, nous pouvons distinguer le temps long de la législation (qui est une théorie normée), évoluant d'un trimestre ou d'une année à l'autre, le temps intermédiaire de l'appareil de représentation administratif et informatique, dont les transformations s'opèrent d'une semaine ou d'un mois à l'autre, et, enfin, le temps court, celui du réel observable faisant l'objet de la norme et de la représentation informatique et dont l'évolution est quotidienne. Régulièrement, en effet, des entreprises fusionnent ou, au contraire, se scindent, d'autres disparaissent, alors que de nouvelles professions ou de nouvelles catégories d'activité non prises en compte par les nomenclatures officielles voient peu à peu le jour, avec, par exemple, la diversification des métiers de l'informatique. 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 *a priori* interdites par l'hypothèse d'un monde clos [6, 8, 9].

Mais la construction temporelle braudélienne doit se traduire par une stratification relative du temps au sein de laquelle les interactions entre séquences de transformation ne sont pas unidirectionnelles. Il arrive en effet que des séquences relativement plus rapides aient un impact sur des séquences de transformation relativement plus lentes, c'est-à-dire que les faits observés aient une incidence sur les théories qui en ont guidé l'observation. Le modèle de Braudel peut alors être complété par le concept du *continuum* évolutif de Norbert Elias. Celui-ci montre que le temps est une construction résultant de la mise en relation de deux ou de plusieurs séquences de transformations (chaque séquence s'apparentant à un *continuum* évolutif) dont l'une est normalisée en vue de servir d'étalon pour mesurer les autres. Ce processus de construction lui-même évolutif est intimement lié à l'état du fonds de savoirs dont dispose la société dans lequel il s'insère. En d'autres termes, plusieurs *continuums* évolutifs coexistent, chacun d'entre eux étant à la fois objet normé et référentiel normatif [6, 9].

Ainsi, par exemple, dans le domaine des bases de données de la Sécurité sociale, l'identification de la catégorie d'activité des employeurs est déterminante pour le calcul du taux de cotisations sociales qu'ils doivent payer à l'Etat. En Belgique, comme nous l'avons déjà mentionné, ces cotisations s'élèvent annuellement à 40 milliards d'euros environ. Les enjeux sociaux et financiers sont donc colossaux. Pour catégoriser les employeurs, la législation administrative utilise une nomenclature des activités européennes mise à jour selon une périodicité pluriannuelle. Mais entre chacune de ces mises à jour, la réalité économique ne cesse d'évoluer de manière quasi continue. Ainsi, quand se développèrent les *copy centers*, ces boutiques mettant des photocopieuses à la disposition de leurs clients, la nomenclature européenne des activités s'avéra très rapidement inadaptée à leur recensement (elle proposait, au mieux, les catégories statistiques : imprimerie, commerce de détail de livres ou secrétariat). Afin de prendre en considération la catégorie *copy centers*,



« Suite à l'interprétation d'observations empiriques inédites relatives aux réalités du temps court (des taux d'ozone anormalement bas) connues à travers les bases de données (temps intermédiaire), la théorie (ou référentiel normatif) (temps long) fut adaptée aux nouvelles connaissances pour prendre en compte le phénomène (jusqu'alors inconnu) de la baisse des taux d'ozone en certains endroits de la stratosphère. », *modélisation réalisée en 1999 lors de la campagne THESEO (campagne européenne de mesure de l'ozone stratosphérique).*

issue du temps court, il fallut tout d'abord modifier les textes réglementaires, puis adapter en conséquence la structure des bases administratives.

On observe le même phénomène de stratification temporelle dans l'exemple évoqué au début de cet article à propos des bases de données de la Nasa : suite à l'interprétation d'observations empiriques inédites relatives aux réalités du temps court (des taux d'ozone anormalement bas) connues à travers les bases de données (temps intermédiaire), la théorie (ou référentiel normatif) (temps long) fut adaptée aux nouvelles connaissances pour prendre en compte le phénomène (jusqu'alors inconnu) de la baisse des taux d'ozone en certains endroits de la stratosphère (voir la photo ci-dessus).

En soi, est infini le processus d'interprétation qui fait passer de la norme aux faits et des faits à la norme en dessinant ce que l'on appelle une boucle herméneutique. Des critères d'arrêt doivent donc être ponctuellement adoptés. Dans la présente étude, ceux-ci sont guidés par un principe d'ordre pratique sous contrainte de budget : il s'agit d'analyser la nature des arbitrages auxquels sont confrontés les gestionnaires d'un système d'information pour améliorer le processus de gestion des bases de données en fonction de leurs enjeux [6].

Développement en matière de modélisation conceptuelle : de l'hypothèse du monde clos à l'hypothèse d'un monde ouvert soumis à des contrôles automatisés

L'impact de l'évolution du réel observable (caractérisé par des séquences relativement plus rapides) sur la théorie et le système d'information associé (se caractérisant par des séquences relativement plus lentes) doit être pris en considération à des fins opérationnelles. Ce mécanisme n'est pas trivial puisqu'il implique des ressources humaines potentiellement importantes (8) tant dans le secteur administratif que dans le monde des banques [29], de l'environnement et de la médecine, comme l'illustrent les exemples évoqués plus haut.

Nous avons enrichi la théorie de la modélisation conceptuelle en distinguant les erreurs des anomalies formelles affectant les données. Les premières constituent une violation de contrainte d'intégrité certaine au sein d'une base de données : par exemple, la présence d'une valeur numérique dans un champ où sont attendues des valeurs alphabétiques. Les secondes s'apparentent à une présomption d'erreur formelle : une incohérence légale apparaît formellement (par exemple, entre la catégorie d'activité d'un

employeur et le type d'employés qu'il déclare), mais seule une interprétation humaine (avec une investigation sur le terrain, par exemple) permettra de détecter s'il y a erreur ou non et quelle est l'information pertinente. Nous proposons ainsi de passer de l'hypothèse d'un monde clos à celle d'un monde ouvert, mais sous contrôle.

Afin d'assurer une prise en compte semi-automatique de ces mécanismes, des pré-requis s'imposent. Un système de détection des anomalies (lors de la saisie, mais aussi *a posteriori*) doit être mis en place de façon à disposer d'un historique des anomalies et de leur traitement, ainsi que d'indicateurs de suivi latéraux. Des procédures claires quant à leur gestion doivent être établies, *a fortiori* lorsque la base de données s'inscrit dans un environnement fédéré. La base de données doit être structurée de telle manière que l'historique des traitements puisse être enregistré. Enfin, une procédure claire de production des indicateurs, selon une périodicité donnée, doit être définie. Le suivi quantitatif des anomalies et de leur traitement permet ensuite le déploiement de stratégies de gestion de la base. Il est par exemple possible d'évaluer la rapidité de traitement des anomalies afin de déterminer quel est le moment le plus opportun pour exploiter la base de données. Ce type d'outil est d'autant plus utile lorsque les données sont collectées en un point unique, puis exploitées de manière fédérée par différents services, à l'instar des procédures que propose l'administration électronique [6, 9].

Voyons maintenant comment s'opère le passage de l'hypothèse du monde clos à celle de ce monde ouvert, mais sous contrôle. Le suivi statistique des violations de contraintes d'intégrité (anomalies formelles) permet de détecter non seulement les augmentations anormales du nombre des anomalies (en fonction d'un seuil donné), mais aussi les augmentations des validations d'anomalies lors de la phase de traitement des données. 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 de saisie peut en effet « forcer » le système à accepter ladite 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 à destination du gestionnaire de la base afin que celui-ci examine si une modification structurelle de son schéma est requise. Lorsque les cas de validations d'erreurs sont importants, il est intéressant d'approfondir le phénomène : comme nous l'avons vu, un cas de figure inédit est peut-être apparu, ce qui requiert une adaptation de la structure de la base. Ainsi, en Belgique, lors de la mise en place d'une directive administrative en faveur du secteur non marchand, la question s'est posée (au regard de la réalité qui avait été progressivement appréhendée au sein de la base) de savoir s'il fallait inclure dans ce secteur les maisons de repos privées, qui en étaient *a priori* exclues du fait de leur finalité lucrative. Initialement considérées comme des cas erronés au regard du domaine de définition spécifiant le secteur non marchand, ces entreprises y ont finalement été intégrées après interprétation juridique (sur la base de la méthode présentée ici). Cela a donné lieu à une restructuration du schéma de la base de données. De manière générale, la restructuration d'une base de données résulte d'une décision humaine tendant à rendre le modèle conforme (au moins transitoirement) aux nouvelles observations faites.

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 à apparaître et devenir de plus en plus nombreuses, nécessitant un examen manuel potentiellement lourd et susceptible de ralentir considérablement le traitement des dossiers administratifs. Pour la Sécurité sociale belge, la mise en œuvre de cette méthode a permis d'améliorer la précision et la rapidité du traitement des cotisations sociales en réduisant potentiellement de 50 % le volume des anomalies formelles (qui représentaient jusqu'alors chaque trimestre de 100 000 à 300 000 occurrences nécessitant une gestion manuelle).

Conclusion

Aux côtés d'autres mesures aux fins d'évaluer et d'améliorer la qualité des bases de données (comme le recours à des techniques de traitement semi-automatique des doublons ou des incohérences [23, 33], la mise en place de projets de *reingineering* visant à améliorer structurellement les modalités d'alimentation des bases de données ou encore la conception de systèmes documentaires permettant d'interpréter les données tout au long de leur cycle de vie [7]), l'approche présentée ici a été appliquée à des domaines très divers [3] à côté de celui des bases de données administratives, telles que les classifications du département ethnographique du Musée Royal de l'Afrique Centrale à Bruxelles [31], la base du fond photographique des *National Archives of the Netherlands* aux Pays-Bas [30], celle du *September 11th Memorial and Museum* de New York [32] ou encore dans le domaine de l'apparat critique des œuvres littéraires [9].

Plus largement, les apports méthodologiques de cette étude peuvent s'avérer féconds dans tous les domaines où l'on observe, comme le notait déjà Heinrich Rickert [25], au début du XXe siècle, « [...] *une certaine relativité dans l'élaboration des concepts, non seulement dans la mesure où l'apparition d'un nouveau matériau empirique peut modifier les concepts, car cela est évident pour toutes les sciences, mais aussi dans la mesure où les points de vue recteurs des différentes sciences changent, ce dont la totale transformation de la biologie par les hypothèses de Darwin fournit un exemple* » (9). Dans toutes les matières empiriques, en effet, la théorie normée se révèle au fil de sa mise en œuvre, à travers l'interprétation conjointe de la norme couvrante et du fait couvert, qui dessine des hiérarchies enchevêtrées entre une pluralité de significations inter-agissantes et évolutives, les bases de données étant le théâtre de ces ajustements infinis.

Notes

* Docteur en Philosophie et Lettres (orientation « Sciences de l'Information »), chargée de cours à l'Université Libre de Bruxelles.

(1) Le Nouveau Petit Robert. Dictionnaire alphabétique et analogique de la langue française, Paris, Dictionnaire Le Robert, 2010, p. 1705.

(2) http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=50799

(3) Ce postulat a donné lieu à la mise en place d'algèbres de la qualité reposant, par exemple, sur le taux de correction de l'information par rapport au réel observable correspondant. À notre connaissance, cette approche n'est jamais sortie de l'enceinte des laboratoires de recherche et n'a jamais donné lieu à une quelconque application concrète dans le monde industriel ou dans celui des entreprises. D'autres auteurs (tel Thomas Redman) proposent cependant des approches du traitement de l'erreur formelle assurément très pertinentes, comme le *data tracking*, mais ils n'abordent pas la question de l'interprétation évolutive de l'information (Redman (T.), *Data Quality for the Information Age*, Boston, Artech House, 1996).

(4) Nous menons ces travaux au sein du centre de recherche de Smals, une société informatique prestataire de services pour l'administration fédérale belge, dont nous coordonnons le *Data Quality Competence Center* (https://www.smals.be/fr/content/data-quality) en parallèle à nos recherches menées au département Sciences de l'information et de la communication de l'Université Libre de Bruxelles (http://www.ulb.ac.be/cours/iboydens/).

(5) Le Danemark dispose d'un système d'adressage très rigoureux et complet, mais trop rigide pour prendre en compte dans ses bases de données l'hétérogénéité des adresses étrangères, dont la représentation est toutefois indispensable, en raison de la mondialisation des échanges (Sallets (J.), *La problématique des adresses spatiales dans les bases de données administratives* [mémoire de fin de Master en Sciences et Technologies de l'Information et de la Communication] sous la direction de Boydens (I.), Bruxelles, Université Libre de Bruxelles, 2011, p. 80)).

(6) Le processus d'adressage sud-africain est fortement lié à l'histoire du pays, avec, depuis la fin de l'apartheid, la prise en compte progressive des *Black townships* construites en dehors des villes, ainsi qu'avec la réattribution progressive de leurs propriétés aux citoyens dépossédés par les lois raciales. Par ailleurs, comme dans beaucoup d'autres pays du Sud, la prise en compte des *Informal Settlements* (les bidonvilles) est anarchique et dynamique. Ces mécanismes historiques en cours ont inévitablement un impact sur la qualité des bases de données qui en accompagne les développements (Sallets (J.), *La problématique des adresses spatiales dans les bases de données administratives* [mémoire de fin de Master en Sciences et Technologies de l'Information et de la Communication], sous la direction de Boydens (I.), Bruxelles, Université Libre de Bruxelles, 2011, pp. 81-90).

(7) Les difficultés qui se posèrent alors émanaient essentiellement de la confrontation entre des systèmes d'information hétérogènes dont la fusion donnait lieu à des résultats aberrants, Madnick (S. E.), *The Voice of the Customer: Innovative and Useful Research Directions, in* Agrawal (R.), Baker (S.) & Bell (D.) (éds), *Proceedings of the 19th Conference on Very Large Databases*, Dublin, VLDB, p. 702, 1993.

(8) « *Des travaux, comme la présentation donnée par Simon Riggs à XML Europe 2003 ou le travail d'Isabelle Boydens (*Informatique, normes et temps*, Bruxelles, Éditions E. Bruylant, 1999) portant sur la qualité de vastes bases de données ont montré qu'environ 10 % des documents XML (ou des valeurs associées à des données structurées) contenaient au moins une erreur* », Van Der Vlist (E.), *Relax NG*, Cambridge, O'Reilly Media, 2003.

(9) Rickert (H.), *Science de la culture et science de la nature*, Paris, Gallimard, 1997 (titre original en allemand : "Kulturwissenschaft und naturwissenschaft", 1926), p. 248.

Bibliographie

[1] ANDERSON (K. W.), DUMPHY (D. C.) & WILSON (P. W. F.), "*Management of Data Quality in a Long-term Epidemiologic Study: the Framingham Hearth Study*", *in* LIEPINS (G. E.) & UPPULURI (V. R. R.) (éd.), Data Quality Control, Theory and Pragmatics (Série "Statistics: Textbooks and Monographs"), New York, DEKKER (Marcel), Inc., vol. 112, pp. 57-68, 1990.

[2] ARON (R.), *La philosophie critique de l'histoire*, Paris, éd. Vrin, 1969.

[3] BADE (D.), "*It's about Time! Temporal Aspects of Metadata Management in the Work of Isabelle Boydens*", Cataloging & Classification Quarterly, vol. 49, n°4, pp. 328-338, 2011.

[4] BESSON (J.-L.) (éd.), *La Cité des chiffres ou l'Illusion des statistiques*, Paris, Editions Autrement, 1992.

[5] BOYDENS (I.), « Analyser le processus de transformation de l'information : du "stemma codicum" au "data tracking" », *in* ROELANTS-ABRAHAM (J.) (éd.), *Information et documentation : du réel au virtuel*, Bruxelles, Infodoc-ULB, pp. 57-70, 1998.

[6] BOYDENS (I.), *Informatique, normes et temps*, Bruxelles, Bruylant, 1999.

[7] BOYDENS (I.), « Déploiement coopératif d'un dictionnaire électronique de données administratives », *Revue Document Numérique* (« Création et gestion coopératives de documents numériques d'information et de communication »), vol. 5, n°3-4, pp. 27-43, 2001.

[8] BOYDENS (I.), « Les bases de données sont-elles solubles dans le temps ? », La Recherche, Hors-série n°9 (« Ordre et désordre »), novembre-décembre, pp. 32-34, 2002.

[9] BOYDENS (I.), « Hiérarchie et anarchie : dépasser l'opposition entre organisation centralisée et distribuée ? », *in* HUDON (M.) & EL HADI (W. M.) (éd.), *Les cahiers du numérique* (Numéro thématique « Organisation des connaissances et Web 2.0 »), Paris, Editions Hermès Sciences, vol. 6, n°3, pp. 77-101, 2010.

[10] BOYDENS (I.), "Strategic Issues Relating to Data Quality for E-government: Learning from an Approach Adopted in Belgium", *in* ASSAR (S.), BOUGHZALA (I.) & BOYDENS (I.), éd., P*ractical Studies in E-Government: Best Practices from Around the World*, New York, Springer, pp. 113-130 (chapitre 7), 2011.

[11] BOYDENS (I.) & VAN HOOLAND (S.), "Hermeneutics applied to the quality of empirical databases", *Journal of documentation*, vol. 67, issue 2, pp. 279-289, 2011.

[12] BRAUDEL (F.), *La Méditerranée et le monde méditerranéen à l'époque de Philippe II*, Paris, Armand Colin, 1976.

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

[14] COLLINGWOOD (R.G.), *The Philosophy of History*, cité dans PROST (A.), *Douze leçons sur l'histoire*, Paris, Seuil, p. 85, 1996.

[15] ELMASRI (R.) & NAVATHE (S.B.), *Fundamentals of Database Systems*, 6th Edition, Boston, Addison Wesley, 2011.

[16] ELIAS (N.), *Du temps*, Paris, Fayard, 1996.

[17] F. EWALD (F.), « La mesure et la norme », *Catalogue de l'exposition « Mesure et démesure »*, Paris, Cité des sciences et de l'industrie, 1995, pp. 71-85.

[18] LOEBL (A.S), *Accuracy and Relevance and the Quality of Data*, *in* LIEPINS (G.E.) & UPPULURI (V. R. R.) (éd.), Data Quality Control. Theory and Pragmatics (Série "Statistics: Textbooks and Monographs"), New York, DEKKER (Marcel), Inc., vol. 112, pp. 103-141, 1990.

[19] LOSHIN (D.), *The Practicioner's Guide to Data Quality Improvment*, Elsevier, Morgan-Kaufmann OMG Press, 2011.

[20] MACKAY (P.), GIRARD (N.) & DEMERS (D.), « La dérive des finalités dans l'interprétation : les cas des impacts de l'information génétique sur les droits économiques et sociaux », *in* THOMASSERT (C.) & BOURCIER (D.) (éd.), *Interpréter le droit : le sens, l'interprète, la machine*, Bruxelles, Bruylant, pp. 247-288.

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

[22] MARICHAL (R.), « La critique des textes », *in* SAMARAN (C.) (éd.), *L'histoire et ses méthodes*, Encyclopédie de La Pléiade, Paris, Gallimard, pp. 1247-1360, 1961.

[23] OLSON (J.), *Data Quality: The Accuracy Dimension*, Burlington, Elsevier, 2002.

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

[25] RICKERT (H.), *Science de la culture et science de la nature*, Paris, Gallimard, 1997 (titre original en allemand : "Kulturwissenschaft und naturwissenschaft", 1926).

[26] RIVIÈRE (P.), « Indicateurs de qualité en matière de production de données : quelques éléments de réflexion », *Courrier des statistiques*, Paris, INSEE, septembre 2005, n°115, pp. 35-40.

[27] SALLETS (J.), *La problématique des adresses spatiales dans les bases de données administratives*, mémoire de fin de Master en Sciences et Technologies de l'Information et de la Communication, sous la direction de Boydens (I.), Bruxelles, Université Libre de Bruxelles, 2011.

[28] VALLA (L.), *La Donation de Constantin* (*sur la donation de Constantin à lui faussement attribuée et mensongère*), traduit et commenté par GLARD (J.-B.), Paris, Les Belles Lettres, 1993.

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

[30] VAN HOOLAND (S.), *Spectator becomes annotator: possibilities offered by user-generated metadata for image databases*, Paper presented at Immaculate Catalogues: Taxonomy, Metadata and Resource Discovery in the 21st Century, 13-15 september 2006, University of East Anglia, UK.

[31] VAN HOOLAND (S.), KAUFMAN (S.) & BONTEMPS (Y.), "Answering the call for more accountability: applying data-profiling to museum metadata", *Proceedings* of *the International conference on Dublin Core and metadata applications*, 22-26 september 2008, Berlin, Dublin Core Metadata Initiative, Berlin, pp. 93-103.

[32] VAN HOOLAND (S.), *Metadata quality in the cultural heritage sector: stakes, problems and solutions,* Thèse de doctorat, Département Sciences de l'Information et de la Communication, sous la direction d'Isabelle Boydens, Université Libre de Bruxelles, 2009.

[33] VAN HOOLAND (S.), VERBORGH (S.), De WILDE (M.), HERCHER (J.), MANNENS (E.) & VAN DE WALLE (R.), *Free your Metadata: Integrating cultural heritage collections through Google Refine reconciliation*, 2011, Pre-submission paper available on Free your Metadata website : http://freeyourmetadata.org/publications/freeyourmetadata.pdf

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

[35] WAND (Y.) & WANG (R. Y.), *Anchoring Data Quality Dimensions in Ontological Foundations*, Communications of the ACM, vol. 39, n°11, pp. 86-95, novembre 1996.

[36] WIENER (L.R.), *Les avatars du logiciel*, Paris, Editions Addison-Wesley France, 1994.