LES BASES DE DONNÉES SONT-ELLES SOLUBLES DANS LE TEMPS ?

Du statut administratif des individus aux flux monétaires internationaux, les bases de données, ensembles colossaux d'informations numériques organisées, répertorient et ordonnent la complexité du monde. Mais qu'advient-il quand ces grilles n'épousent plus le terrain, quand les données s'écartent de ce qu'elles représentent ? Comment ajuster les représentations enregistrées dans les bases au désordre empirique de la réalité ?

Isabelle Boydens, docteur en philosophie et lettres, est chargée de cours à l'université libre de Bruxelles, section « sciences de l'information et de la documentation ». Elle est aussi consultante à la section « recherche » de la SmalS-MvM, société informatique prestataire de services pour la Sécurité sociale belge.
iboydens@ulb.ac.be

Les bases de données se prêtent aux métaphores financières. Ne les désigne-t-on pas souvent par le terme « banques » de données ? Elles évoqueraient ainsi un « *capital d'information sur lequel on peut faire des retraits à la demande*[1] ». A condition de s'assurer que le compte est correctement approvisionné… Ce n'est hélas pas toujours le cas. L'OTAN en fit l'expérience fâcheuse en mai 1999, après avoir bombardé l'ambassade de Chine à Belgrade. Pressée d'expliquer son attaque, l'organisation incrimina les bases cartographiques utilisées pour guider ses missiles. Celles-ci répertoriaient en effet un plan de Belgrade obsolète…

Qui n'a, dans sa mémoire, un exemple d'incohérence entre la réalité et une base de données censée la représenter (voir l'encadré : « Facteurs de désordre », p. 33) ? Si l'on ne veut pas rallonger la liste de ces désordres, un préalable s'impose : identifier les mécanismes en jeu. De fait, toute construction d'une base de données implique la mise en place d'un ordre interne : il faut délimiter un « domaine de définition » hors duquel les données collectées ne seront pas pertinentes. Il ne s'agit jamais que de fixer des conventions dont on espère qu'elles refléteront efficacement la réalité, en fonction des finalités attribuées à la base. Qui songerait à l'existence d'un référentiel absolu sur lequel garantir ces choix ? Par nature, la structure rigide et prédéfinie des représentations formelles enregistrées au sein de toute base de données (l'« ordre ») ne peut jamais s'adapter totalement à l'hétérogénéité mouvante du réel (le « désordre »).

Les modèles les plus répandus de bases de données s'appuient sur trois principes pour préserver l'intégrité de leur structure au fil de leur exploitation. Le principe d'identité affirme qu'une valeur saisie dans une base peut être déplacée ou supprimée, mais pas altérée. Les deux autres principes, non-contradiction et tiers exclu, assurent quant à eux la conformité des données saisies avec les règles internes de la base. A ces principes généraux s'ajoute l'hypothèse dite du « monde clos », selon laquelle toute valeur nouvelle n'est acceptée que si elle est compatible avec le domaine de définition de la base ; dans le cas contraire, la nouvelle saisie est considérée comme une erreur formelle.

La mise en œuvre de ces principes est malmenée dans l'utilisation quotidienne d'une base de données. Dans le domaine administratif, notamment, alors que les mises à jour des bases s'effectuent en général selon une périodicité prédéfinie, les textes de lois peuvent changer au gré de la jurisprudence. Par exemple, quand se développèrent les « *copy centers* », ces


© L. Tunblork/Vu

boutiques mettant des photocopieuses à disposition de leurs clients, la nomenclature des activités européennes s'avéra rapidement inapte à leur recensement : elle proposait au mieux les catégories « imprimerie », « commerce de détail de livres » ou « secrétariat ». Il fallut d'abord modifier les textes réglementaires, puis adapter la structure des bases en conséquence.

LE FLOU N'A PAS DROIT DE CITÉ DANS LES BASES. Prenons un autre cas qui, cette fois, implique des individus et requiert donc un soin approprié : les bases de données de la Sécurité sociale belge. On sait par exemple que la législation sociale est différente selon qu'elle s'applique aux ouvriers ou aux employés, les premiers et les seconds se distinguant selon la nature prépondérante de leurs activités manuelles ou intellectuelles. Dans la pratique, cette distinction n'est pas aisée à opérer, mais le flou n'a pas droit de cité dans une base de données : il faut trancher… Une analyse détaillée de ce cas de figure paradigmatique illustre combien les transformations opérées au sein des bases de données, l'évolution de la jurisprudence et les catégories observables sur le terrain sont solidaires. Solidaires, mais asynchrones. Elles opèrent, suivant leur nature, au sein d'échelles de temps différentes. On distingue ainsi le « *temps long* » des normes juridiques, renouvelées d'un trimestre ou d'une année sur l'autre, le « *temps intermédiaire* » de la gestion des bases de données et le « *temps court* » du réel observable, celui des citoyens assujettis à l'administration, dont l'évolution est continue.

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.

A 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. Par exemple, avant la découverte par des chercheurs britanniques de la chute des taux d'ozone, dans les années 1980, les valeurs faibles correspondantes

FACTEURS DE DÉSORDRE

Il est fréquent que l'exploitation des bases de données soit la source de situations au mieux absurdes, parfois dramatiques. Sait-on que, pendant la guerre du Golfe, environ 28 000 des 40 000 containers militaires américains envoyés au Moyen-Orient durent être inspectés et inventoriés manuellement ? L'interrogation des bases de données correspondantes, censées en répertorier le contenu, donnait lieu à des résultats trop incohérents pour qu'on leur fasse confiance.
Qui se souvient de la pression que supporta le gouvernement de Jimmy Carter de la part de l'opinion publique aux Etats-Unis, à la suite de l'embargo pétrolier imposé par les pays de l'OPEP en 1973 ? Le Département de l'énergie fut soupçonné de collusion avec les industries pétrolières, car il affirmait que les réserves fédérales en carburant étaient suffisantes pour répondre aux besoins de la population. En réalité, les informations à la source des décisions du gouvernement, extraites des multiples bases de données relatives aux ressources et aux flux commerciaux pétroliers, s'avérèrent erronées. Un audit mené pendant cinq ans montra, en auscultant près de 2 200 bases de données, que l'agrégation des valeurs recueillies dans des bases différentes confondait notamment les quantités d'énergie disponibles et distribuées. Les données, fiables séparément, avaient reçu des dénominations si ambiguës d'une base à l'autre qu'il était impossible de les croiser sans obtenir des résultats aberrants. ◇

TROIS STRATÉGIES

Les données s'accumulent dans une base comme l'eau dans un bassin d'épuration. Il faut éliminer ce qui flotte en surface, analyser la composition pour vérifier qu'elle est conforme à une norme, repérer la source des affluents perturbateurs. Et, de la même façon qu'on traite différemment l'eau destinée à l'irrigation et celle alimentant un réseau domestique, la méthode retenue doit s'adapter au type de base de données dont on veut améliorer la qualité.

Les bases statistiques agrégées, qui rassemblent de grandes quantités de données indifférenciées pour en extraire des indicateurs statistiques (PIB, taux de natalité), sont traitées par *data cleansing*. Il s'agit d'éliminer les valeurs « aberrantes » au sein de vastes collections, à la manière du « lissage statistique » de longues séries numériques.

Les chercheurs du programme « Total Data Quality Management » du MIT[5] proposent une autre approche. Leur méthode, baptisée « *data tagging* », associe aux données de la base des indicateurs de qualité (exactitude, précision, obsolescence...) considérés comme objectifs. La mise à jour des indicateurs est manuelle, et relativement lourde dans le cas d'une base volumineuse. Elle repose sur l'hypothèse fragile que le réel et sa représentation évoluent de concert, et qu'il est possible d'établir un référentiel absolu des indicateurs de qualité.

Thomas Redman[6] est, lui, le concepteur du « *data tracking* ». Le système d'information est comparé à un fleuve : les tests du type *data cleansing* permettent de nettoyer ponctuellement le fond du fleuve mais n'endiguent pas les affluents d'information de qualité douteuse. Ce sont donc ces flux qui sont étudiés et, si nécessaire, restructurés, en vérifiant notamment la fiabilité des transformations opérées sur les données (validité des algorithmes de calculs de taux bancaires ou fiscaux, harmonisation d'une base à une autre du codage formel du sexe, du titre ou du nom d'individus, etc.). ◇

étaient systématiquement considérées comme des anomalies dans les bases de données de la NASA[2]. En effet, la théorie d'alors, modélisée dans la base, ne permettait pas de penser que de telles valeurs puissent être correctes. La NASA a ensuite adapté la structure de la base, ajoutant les « anomalies » à l'ensemble des valeurs admises. Dans de tels cas, la restructuration de la base résulte d'une décision humaine tendant à rendre le modèle provisoirement conforme aux nouvelles observations. Ce phénomène de transformation correspond au mécanisme dit de « boucle étrange[3] ». En l'absence d'une telle intervention, l'écart entre la base et le réel se creuse inexorablement. Et la base de données, en tant qu'instrument d'action sur le monde, devient facteur de désordre.

Ces constats peuvent-ils servir à élaborer une stratégie pour améliorer la qualité des bases ? La majorité des équipes de recherche concernées tendent à se préoccuper de la cohérence formelle des systèmes de représentation. Or, on l'a vu, la fidélité d'une base à la réalité dépend moins des données elles-mêmes que d'une adéquation judicieusement établie entre les modalités de leur collecte et l'hétérogénéité fluctuante du réel. Dans le cadre de mes travaux sur la Sécurité sociale belge, j'ai proposé une stratégie de gestion[4] consistant à déterminer le moment où la structure de la base n'est plus performante, relativement à la manière dont s'organisent les faits sur le terrain. Pour cela, un algorithme intègre les interprétations humaines des données que le système considère comme des erreurs formelles. D'ordinaire, lorsque le format ou la valeur d'une nouvelle entrée n'est pas conforme au domaine de définition de la base, l'opérateur peut « forcer » le système à accepter la donnée. 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. L'algorithme propose alors au gestionnaire de la base une modification structurelle de son schéma.

Les bases de données sur lesquelles s'appuie le système des cotisations sociales en Belgique régulent la perception annuelle de quelque 30 milliards d'euros. Au-delà de l'enjeu financier, c'est l'accès à l'égalité sociale qui est concerné. Chaque donnée compte : tout individu doit disposer de données personnelles de qualité égale à celles de ses concitoyens. Autant d'enregistrements erronés, autant de procès potentiels à l'encontre de l'administration. Pour la Sécurité sociale belge, la nouvelle méthode a permis d'améliorer la précision et la rapidité de traitement des cotisations sociales, réduisant potentiellement de 50 % le volume d'anomalies formelles, qui représentent chaque trimestre 100 000 à 300 000 occurrences à traiter manuellement.

UNE ILLUSION D'ORDRE. Si l'algorithme proposé est particulièrement adapté aux bases de données individuelles, il concernerait peu un autre type de bases, les bases statistiques agrégées. Celles-ci regroupent, au service de l'Insee ou d'Eurostat par exemple, des données considérées dans leur ensemble, comme le PIB, le taux de croissance, les volumes des stocks nationaux de matières premières… Elles se prêtent à d'autres stratégies d'amélioration, proches des méthodes en cours dans la production industrielle (voir l'encadré « Trois stratégies », ci-dessus).

Quel que soit leur type, les bases de données peuvent créer une illusion d'ordre, de transparence et d'instantanéité puissante, car le virtuel s'impose de plus en plus comme un mode de gestion du réel. Elles doivent pourtant être considérées comme un facteur potentiel de désordre face à la complexité mouvante de la réalité, puisqu'il n'existe aucun référentiel absolu à l'aune duquel les valider. Il est alors indispensable, pour garantir leurs performances, de dépasser la dichotomie du vrai et du faux et de violer l'hypothèse du monde clos qui les sous-tend. En se donnant les moyens d'ajuster continûment les modèles des bases de données aux observations empiriques de terrain, il devient ainsi raisonnable de légitimer leur rôle premier d'instrument d'action sur le réel, producteur d'un ordre cohérent et intelligible. **I.B.** ◆

✻Le concept de **TEMPORALITÉS ÉTAGÉES**, adapté ici à l'information administrative, a été forgé par l'historien Fernand Braudel. Celui-ci distinguait le *« temps long »* des structures géographiques, le *« temps intermédiaire »* des conjonctures économiques, et le *« temps court »* de l'événement politique : F. Braudel, *La Méditerranée et le monde méditerranéen à l'époque de Philippe II*, Armand Colin, 1976.

RÉFÉRENCES

(1) R. Escarpit, *L'Information et la communication. Théorie générale*, Hachette, 1991, p.152.
(2) L. R. Wiener, *Les Avatars du logiciel*, Addison-Wesley France, 1994, p. 37.
(3) D. R. Hofstadter, *Gödel, Escher, Bach, les brins d'une guirlande éternelle*, InterEdition, 1985.
(4) I. Boydens, *Informatique, normes et temps*, Bruylant, 1999, p. 363.
(5) R. Y. Wang, M. Ziad et Y. W. Lee, *Data Quality*, Kluwer academic Publishers, 2000 ; K.-T. Huang, Y. W. Lee et R. Y. Wang, *Quality Information and Knowledge Management*, Prentice Hall, 1999.
(6) T. Redman, *Data Quality for the Information Age*, Artech House, 1996 ; *Data Quality*, The Field Guide, Digital Press, 2000.

POUR EN SAVOIR PLUS

- I. Boydens, *Informatique, normes et temps,* Bruylant, 1999.
- R. Elmasri, S. Navathe, *Fundamentals of Database Systems,* The Benjamin/Cummings Publishing Company, Inc., 1999.
- W. Kent, *Data and Reality. Basic Assumption in Data Processing Reconsidered,* Elsevier, 1981.

www.larecherche.fr