BTNG-RBHC, XXIV, 1993, 3-4, pp. 399-439.

INFORMATIQUE ET QUALITE DE L'INFORMATION
Application de la critique historique à l'étude des informations issues de banques de données

PAR

ISABELLE BOYDENS

Institut d'Etudes Socio-Historiques
Université de Liège

1. INTRODUCTION

1.1. Position du problème

> "Dans le cas précis des grands réseaux informatiques et des bouleversements révolutionnaires que l'on pourrait en attendre, la confusion [...] entre d'un côté le savoir et de l'autre l'information stockée dans les mémoires des ordinateurs est particulièrement génératrice d'illusions et de frustrations ...."[1]

Le 4 octobre 1992, les débris d'un avion de la compagnie israélienne El Al s'écrasent sur deux immeubles aux abords d'Amsterdam. Dizaines ou centaines de morts? Les autorités néerlandaises ne parviendront jamais à établir un bilan exact. Les HLM détruits abritaient en effet de nombreux illégaux qui n'étaient recensés dans aucun fichier et qu'il était par conséquent impossible d'identifier.[2] Mais très vite, une autre faille apparaît: le manque de fiabilité des données du fichier répertoriant les

1. Philippe BRETON, *Quand s'usent les idéaux. Informatique et utopie* dans *Le Monde Diplomatique*. Mai 1993, p. 32, c. 3.
2. Voir par exemple: J.-C.M., *Le drame d'Amsterdam livrera-t-il son bilan?* dans *La Libre Belgique*, 07/10/92, p. 1, c. 2.

informations d'identification personnelles, similaires à celles de notre "Registre National".[3]

Citons un autre exemple, plus concret encore. Considérons un système informatique conçu afin d'évaluer le nombre de travailleurs et le nombre d'employeurs observés dans une région et pendant une période déterminées. Les informations suivantes sont intégrées dans la banque de données:

> "Le travailleur A effectue des prestations à temps partiel et travaille respectivement pour les employeurs X et Y. L'employeur X dirige une seule entreprise relevant d'un seul secteur d'activité. L'employeur Y dirige trois entreprises relevant chacune de secteurs d'activité distincts."

Ce qui donnerait:
- Nombre de travailleurs observés: 1
- Nombre d'employeurs observés: 2

Or, en fin de parcours, les résultats obtenus à partir des informations entrantes et figurant dans la banque de données sont les suivants:
- Nombre de travailleurs observés: 2
- Nombre d'employeurs observés: 4

Nous constatons qu'un phénomène de perte d'information, donnant lieu à une déformation de la réalité observée, s'est produit au cours du processus informatique.[4] Tellement simples qu'ils semblent inimaginables, ces phénomènes sont pourtant bien réels. Nous verrons en outre que les mécanismes susceptibles de déclencher de telles pertes d'information sont plus complexes qu'il n'y paraît.

Des lignes qui précèdent, nous apprenons qu'il existe un lien entre la qualité des informations issues des banques de données et le degré de précision de notre connaissance du réel observable. Cette assertion, qui semble tomber sous le sens, s'insère dans une réflexion plus large relative au cadre de référence informatique.

Schématiquement, le fonctionnement d'une banque de données peut être défini comme suit: un premier ensemble codifié d'informations (que nous appellerons "input") est intégré et traité dans un système informa-

---

3. P.-F. VAN LOO, *La politique d'exploitation de l'informatique dans l'administration de la sécurité sociale des Pays-Bas. Communication destinée à la Conférence sur l'informatique et la sécurité sociale dans les Etats membres de la Communauté européenne*. Londres, décembre 1992, p. 10.

4. Un travailleur est comptabilisé autant de fois qu'il exerce d'emplois simultanés. Un employeur est comptabilisé autant de fois qu'il possède de sièges d'exploitation relevant de secteurs d'activité distincts.

tique dont émane un second ensemble codifié d'informations (que nous appellerons "output").

Plus précisément, l'input est un ensemble fini de données structurées destiné à représenter certains aspects du réel observable. L'output est un ensemble fini de données structurées, censé correspondre à une partie du réel observable et destiné à enrichir notre connaissance de ce réel. Entre ces deux ensembles, se situe le système informatique, que nous considérons provisoirement comme une boîte noire. La question que nous posons dans cette étude est la suivante: quelle est l'influence de cette boîte noire sur l'information? Au cours du processus qui va de la constitution de l'input à celle de l'output, dans quelle mesure un système informatique est-il susceptible d'améliorer ou de détériorer la qualité des données et de là, notre représentation du réel? Et s'il peut l'améliorer, dans quelles conditions?

Tenter de répondre à ces questions nous semble important tant sur le plan général de la transformation de l'information dans notre société que sur celui, plus spécifique, de la recherche scientifique. Les sciences humaines ont en effet de plus en plus souvent recours aux sources informatiques. Si nos outils d'analyse sont techniquement de plus en plus performants, qu'en est-il de la qualité des sources servant de base à ces analyses? Cette dernière question est particulièrement pertinente lorsqu'il s'agit d'exploiter des banques de données administratives, de plus en plus nombreuses, volumineuses et sujettes à de continuels remaniements liés à la fréquence des modifications législatives. Ces banques de données, fondamentalement conçues à des fins de gestion administrative, ne peuvent dès lors être exploitées à des fins scientifiques qu'après avoir fait l'objet d'une analyse descriptive et critique rigoureuse.

Or, pour les historiens qui s'attachent à l'étude de l'histoire contemporaine, les banques de données administratives représentent des sources d'information inestimables. Les banques de données de la sécurité sociale, les banques de données fiscales ou encore le Registre National, pour ne citer que quelques exemples, constituent en effet des sources précieuses lorsqu'il s'agit de mener des études relatives à l'histoire démographique, sociale, économique ou politique de cette seconde moitié du vingtième siècle. Dans le présent article, nous nous attachons dès lors essentiellement aux modalités d'application de la critique historique aux banques de données administratives. Les résultats que nous présentons sont destinés à être enrichis et approfondis dans le cadre d'une recherche plus large que nous menons sur le sujet.

1.2. *Méthode et sources*

Notre démarche s'inspire, quant à sa méthode, de la critique historique. De la même manière que le médiéviste construit un stemma codicum afin d'établir des filiations et de retrouver le manuscrit le plus proche de l'original, le chercheur qui a recours aux sources informatiques doit trouver une méthode appropriée afin d'interpréter et d'exploiter au mieux les informations issues de banques de données.

Une collection de données livrée à l'ordinateur est en effet sujette à de multiples et constantes transformations. Celles-ci sont liées aux opérations de formalisation, de saisie, de test, de mise à jour, de correction et d'agrégation que subissent les données. Aussi, une information saisie dans la banque de données (input), n'a-t-elle pas nécessairement la même signification que l'information censément identique obtenue en fin de parcours (output). La chose se complique lorsque l'output redevient input en vue de traitements ultérieurs ... Les altérations de sens qui pourraient survenir au cours de ces opérations seraient particulièrement graves si l'intitulé d'une donnée restait identique alors que sa signification réelle aurait changé. Dans de telles conditions, les informations finales ne sont porteuses de sens que si elles sont accompagnées d'un appareil d'interprétation critique.

Dans cette perspective, la démarche et l'esprit de la critique historique, discipline qui s'applique traditionnellement aux documents écrits, demeurent extrêmement riches et pertinents. De nouvelles règles, adaptées à l'étude spécifique des données générées et transformées au sein d'un système informatique, doivent toutefois être conçues et mises en oeuvre.

Si nous sommes ainsi amenée à mettre en lumière les caractéristiques nouvelles qu'entraîne l'apparition de l'informatique, nous ne manquerons pas de souligner, au cours de cet exposé, les similitudes existant entre les processus de génération et de transformation de l'information actuels et passés. En effet, certaines caractéristiques inhérentes à l'exploitation informatique des données ont, d'un point de vue strictement conceptuel, déjà été évoquées depuis des siècles (problèmes liés à l'agrégation des données, au passage du singulier au général, ...). Nous verrons cependant que dans le cadre de référence informatique, l'émergence et le développement de ces caractéristiques peuvent avoir des implications nouvelles.

Le présent article est le fruit d'un travail qui a consisté d'une part, à concevoir et à appliquer une méthodologie d'analyse critique de l'information traitée par ordinateur et d'autre part, à présenter des propositions concrètes afin d'améliorer la qualité de cette information. L'étude de la transformation des données au sein d'une banque de données

administrative belge en constitue la source et le point de départ.[5] Nous présentons ici les aspects de la méthode suivie et des résultats obtenus qui, en tant que tels, sont applicables à d'autres banques de données administratives.

Notre étude a exigé un travail heuristique original. Dans le cadre informatique, nous appelons "heuristique" la démarche qui consiste à rassembler et à structurer l'ensemble des renseignements indispensables afin de comprendre le fonctionnement d'une banque de données ainsi que la signification des informations qu'elle répertorie. En effet, pour maîtriser une source informatique, il ne suffit pas d'accéder aux données stockées dans la mémoire de l'ordinateur. Il faut, en plus, disposer d'une documentation très précise relative à la définition des données, aux divers traitements qu'elles subissent et à la structure des fichiers. Or, il arrive fréquemment qu'au sein d'un système informatique, les informations de ce type fixées par écrit soient très incomplètes. En effet, de nombreux renseignements relatifs à la banque de données font l'objet d'un savoir exclusivement oral partiellement détenu par divers responsables. Ces responsables, immergés dans leur travail, ont rarement une vue globale du système informatique. Aussi, le chercheur soucieux de maîtriser le fonctionnement d'une banque de données doit-il travailler simultanément dans plusieurs directions. Il s'agit de procéder par recoupement, confrontant sources écrites, sources orales et sources informatiques, relevant les incohérences éventuelles et les élucidant en recourant à de nouveaux interviews ou en effectuant des tests supplémentaires. Il faut enfin tenir compte de la multitude des intervenants (d'ordre politique, juridique, social, administratif, informatique ...) dont l'influence, nous le verrons, interagit au coeur même de l'information.

1.3. Structure de l'exposé

Après cette introduction, notre exposé compte trois chapitres:

1. Le premier chapitre définit brièvement les caractéristiques essentielles d'un système informatique;
2. Le second chapitre présente les résultats de notre démarche heuristique. Celle-ci consiste en une étude rigoureuse de l'ensemble du système informatique étudié. D'ordre descriptive, cette étape est indispensable à

---

5. I. BOYDENS, *La banque de données L.A.T.G. de l'O.N.S.S. Les flux de l'information traitée à partir d'une banque de données: étude critique.* Mémoire présenté en vue de l'obtention du grade de licenciée spéciale en science de l'information et de la documentation. Bruxelles, U.L.B., 1992.

toute étude critique ultérieure. Elle a pour but de rassembler et de structurer un ensemble homogène et complet d'informations: historique de la banque de données, diagramme des flux d'information et schéma conceptuel;
3. Le troisième chapitre est l'analyse critique proprement dite. Conformément au cadre théorique que nous avons défini, l'analyse envisage l'influence de l'informatique sur la qualité des données: au stade de l'input, au stade des traitements internes réalisés au sein de la banque de données (que nous qualifions de "boîte noire") et au stade de l'output.

Les conclusions qui se dégagent de l'analyse critique sont présentées à la suite de ce troisième chapitre. Ces conclusions constituent une première tentative de réponse aux questions que nous nous posions au seuil de cette étude. Un ensemble de propositions concrètes en vue d'améliorer ou de garantir la qualité de l'information répertoriée dans les banques de données est ensuite présenté. Ces propositions s'adressent d'une part, aux gestionnaires des banques de données et d'autre part, aux chercheurs qui doivent utiliser les données issues de systèmes informatiques à des fins scientifiques.

2. QU'EST-CE QU'UN SYSTEME INFORMATIQUE?

2.1. *Qu'est-ce qu'une donnée en informatique?*

> "... les résultats du travail informatique, si sophistiqués soient-ils ... ne renvoient jamais au "réel historique" mais toujours à la "métasource" et à elle seule."[6]

Du réel, infini et inaccessible dans sa totalité, se distingue la connaissance finie et partielle que nous en avons. La mise en place d'une banque de données exige l'abstraction préalable d'un sous-ensemble de ce réel tangible et sa transformation en une collection structurée de données codifiées, aptes à être intégrées dans un système informatique.

Trois notions indissociables sont à la base de chacune de ces données: son intitulé, son domaine de définition (ensemble des valeurs admises) et enfin, sa valeur, celle-ci devant obligatoirement être incluse dans le

6. J.-P. GENET, *Outils et démarche. Histoire, Informatique, Mesure* dans *Histoire et Mesure*. Paris, C.N.R.S., 1986, 1-I, p. 12.

domaine de définition.[7] Notons déjà que paradoxalement, si l'informatique est une discipline très bien conçue pour générer et gérer rapidement un grand nombre de valeurs, elle est beaucoup moins bien conçue pour générer au même rythme la documentation indispensable afin de décoder et d'interpréter ces valeurs.

Le réel appréhendé au départ subit ainsi des changements multiples, tant au stade de sa formalisation qu'à celui des opérations réalisées au sein même du système informatique. Toute altération du lien logique existant entre le référent d'une donnée et la valeur qui lui est attachée entraîne une perte d'information. Par conséquent, une donnée résultant de tels traitements n'est pleinement signifiante qu'accompagnée d'informations critiques quant à sa valeur finale: celle-ci correspond-elle bien au domaine de définition établi au départ?

2.2. *Qu'est-ce qu'une donnée "correcte"?*

2.2.1. *Adéquation avec le domaine de définition*

D'un point de vue strictement informatique, pour qu'une donnée soit correcte, il suffit que sa valeur à un instant t soit incluse dans l'ensemble des valeurs admises dans son domaine de définition. Le domaine de définition d'une donnée est lui même établi en fonction des contraintes d'intégrité, ensemble des assertions définissant les extensions possibles d'une banque de données.

On parlera dès lors, à propos d'une donnée, de *validité*, à propos d'un ensemble de données, d'*homogénéité* et de *cohérence*, à propos d'une banque de données, de *consistance* et de *complétude*.[8]

---

7. Par exemple, une donnée intitulée "sexe" peut avoir comme domaine de définition "F=femme et H=homme", avec H et F mutuellement exclusifs, code alphanumérique à une position et comme valeur à un instant t: "F". L'utilisateur qui consulte la base de données à cet instant t voit, face à l'intitulé "sexe", la valeur "F". Cette valeur n'a évidemment aucun sens si l'utilisateur ne dispose pas de la documentation permettant d'en déchiffrer le code.

8. Une donnée est *valide* si sa valeur correspond aux conditions requises définies dans le domaine de définition. Les valeurs d'une donnée sont *homogènes* si, conformément au domaine de définition, elles sont de même nature et de même type. Les valeurs de plusieurs données sont *cohérentes*, si elles n'entrent pas en contradiction avec la logique interne des données définie dans le domaine de définition des données considérées. L'ensemble des valeurs (ou l'état) d'une base de données est *consistant* à un instant t s'il satisfait à toutes les *contraintes d'intégrité* du domaine de définition des données considérées. Un système informatique est *complet* si l'ensemble des contraintes d'intégrité définies dans le domaine de définition des données permet à tout

2.2.2. *Adéquation avec le réel sous-jacent*

Nous verrons cependant au fil de l'étude que la notion de validité ou d'exactitude d'une donnée est toute relative. Par exemple, les procédures automatiques de correction peuvent assurer la cohérence logique de données intrinsèquement fausses. Considérons l'algorithme suivant: rémunération = salaire horaire * nombre d'heures prestées. Si le nombre d'heures de prestation d'un travailleur ne correspond pas à la rémunération afférente à ces prestations, les valeurs de ces deux données peuvent être automatiquement mises en cohérence ... mais correspondront-elles pour autant au réel sous-jacent? Ainsi, pour une période donnée, la rémunération déclarée pourrait être négative ou proportionnellement supérieure ou inférieure au nombre d'heures de travail en raison d'une régularisation comptable effectuée par rapport à une période précédente. Avec l'usage de programmes de correction automatique, le risque est grand de gommer des situations apparemment incohérentes mais bien réelles.

2.2.3. *Importance de la notion de temps: diachronie et synchronie*

La cohérence interne des données est intimement liée à la notion de temps. Si le domaine de définition de la donnée A (rémunération) dépend de celui de la donnée B (nombre d'heures prestées), la valeur de A peut être considérée comme exacte à un instant t et devenir fausse à l'instant t+1, parce que la valeur de B a changé.[9]

Il faut en outre distinguer les phases qui, au sein d'un système informatique, s'opèrent dans la diachronie et dans la synchronie. Nous qualifions de diachronique un processus dynamique et discontinu dans le temps traitant tout ou partie d'un même groupe informationnel à plusieurs reprises et de façon ponctuelle. C'est le cas, par exemple, de l'opération de saisie des données. Dans une banque de données administrative, les informations relatives aux prestations d'un même individu pendant une période circonscrite sont souvent intégrées en plusieurs étapes, de façon

---

moment d'affirmer et de démontrer qu'une valeur ou qu'un ensemble de valeurs sont admis ou non admis dans ce système.

9. Parmi les orientations de la recherche fondamentale en matière de banques de données, la modélisation du temps occupe une place importante. Introduire la notion "d'attribut temporel" permettrait par exemple de tenir compte du décalage entre le "temps réel" et le "temps transactionnel" (celui qui concerne les informations répertoriées dans la base de données). *Conceptual modeling databases and cases. An integrated view of information systems development.* New York, ed. by P. LOUCOPOULOS, R. ZICARI et N.J. WILEY, 1992, p. 16 et 17 et p. 87 à 116.

morcelée: sur support papier, électronique, de façon informelle, par courrier, téléphone ... Nous qualifions de synchronique une opération traitant, à un moment fixe, tout ou partie d'un même groupe informationnel. Ce second type d'opérations ignore le caractère évolutif de l'information et agit sur un état figé de la banque de données. Ainsi en est-il des opérations de consultation ou de traitement statistique.

La plupart du temps, processus diachroniques et synchroniques se superposent. Quelle en est l'incidence sur la qualité de l'information? Les traitements statistiques réalisés en fin de circuit se fondent sur un prélèvement instantané des informations au sein de la banque de données et offrent une image figée du réel. Or, cette image émane de renseignements dont la qualité et le contenu varient dans le temps. On observe dès lors un décalage temporel, et de là, conceptuel, entre le contenu des sources utilisées et celui des résultats obtenus.

2.3. *De la critique historique comme outil d'analyse de l'information traitée et transformée au sein d'une banque de données*

Nous nous proposons d'emprunter à la méthodologie de la critique historique, traditionnellement appliquée aux documents écrits, manuscrits et imprimés, les instruments permettant de dresser notre appareil critique.[10]

Les transformations successives que peuvent subir des données informatisées sont comparables aux altérations de sens observées sur les manuscrits médiévaux recopiés au fil des générations. Le médiéviste construit un stemma afin de retrouver le manuscrit le plus conforme à l'original. L'historien s'attachant à l'étude de la qualité des informations répertoriées dans une banque de données peut, quant à lui, dresser une "généalogie" des informations circulant à travers le système informatique, en reconstituer l'historique et mettre en lumière les mécanismes aboutissant à une incohérence. Une des règles de la critique historique affirme que plus le nombre des intermédiaires entre l'original et une copie est important, plus la probabilité d'erreur dans cette copie a des chances d'augmenter. Nous analyserons quant à nous le nombre et la nature des

---

10. Citons, parmi les ouvrages de référence en la matière: P. HARSIN, *Comment on écrit l'histoire*. Liège, Bibliothèque scientifique belge, G. Thone, 1964. R. MARICHAL, *La critique des textes* dans *L'histoire et ses méthodes. Encyclopédie de la Pléiade*, sous la direction de C. SAMARAN. Paris, Gallimard, 1961, p. 1247 à 1360. P. VEYNE, *Comment on écrit l'histoire. Essai d'épistémologie*. Paris, Seuil, 1971. L.-E. HALKIN, *Initiation à la critique historique*. Paris, Serge Fleury, 1982.

opérations auxquelles sont soumises les données, depuis leur génération jusqu'à leur exploitation finale (à des fins statistiques par exemple), en passant par tous les traitements intermédiaires (informatiques ou manuels) qu'elles subissent. Envisagée comme un exemple de la transformation des données dans notre société, cette étude nous mènera plus loin. Nous verrons combien le jeu des forces externes – politiques, sociales, juridiques, administratives et informatiques – en continuelle interaction, est implicitement présent au coeur même des données et rend difficile le maintien de leur intégrité.

3. LES RESULTATS DU TRAVAIL HEURISTIQUE

3.1. Préliminaires

Lorsque l'on se rend sur le terrain, il est fréquent de constater, au sein d'un système informatique, l'absence d'une documentation exhaustive et à jour relative aux banques de données gérées par celui-ci. Dans le secteur administratif, les conditions dans lesquelles se développent les grands ensembles informatiques rendent en effet très difficiles l'élaboration et la mise à jour de schémas d'ensemble fournissant une vue panoramique sur le fonctionnement de la banque de données. En effet, tant sur le plan administratif qu'informatique, il existe des pratiques informelles rarement fixées par écrit. Celles-ci sont principalement liées à la fréquence de parution d'amendements législatifs à appliquer et à la brièveté (ou l'incertitude) des délais entre l'annonce d'une nouvelle loi et la publication de l'arrêté d'exécution. La gestion quotidienne des banques de données administratives révèle en outre des particularités non prévues au départ. Or, au delà d'un certain seuil, un système informatique ne peut gérer l'exception: il faut donc trouver des "trucs administratifs" susceptibles d'assurer la cohérence du système. Sous la pression continuelle des mises à jour et des nouvelles applications à réaliser, les informaticiens doivent toutefois s'adapter à ce rythme et dans le meilleur des cas, procéder par anticipation. Or le manque de temps ne permet pas de faire précéder chaque mise à jour d'une analyse fonctionnelle donnant lieu à l'élaboration d'un schéma conceptuel idéal qu'il suffirait d'appliquer. A chaque étape vers une automatisation plus large, il s'agit de créer du "nouveau" sur de "l'ancien". On voit ainsi apparaître une multitude de fichiers parfois recopiés d'une banque de données à l'autre.

L'absence de références écrites et explicites précisant le fonctionnement global d'un système d'information peut être lourde de conséquences. Comment les utilisateurs futurs, ignorant les pratiques actuelles, pourront-

ils retrouver, comprendre et exploiter correctement les informations de la banque de données?

Actuellement, certaines banques de données administratives posent déjà de sérieux problèmes d'interprétation. Lorsque l'information est intégrée de façon décentralisée, des pratiques de codification locales peuvent s'installer lors de l'encodage. L'insuffisance des contrôles de cohérence hypothèque la qualité des données qui deviennent parfois indéchiffrables. La création des banques de données administratives les plus anciennes date des années septante. A l'heure actuelle, on envisage leur intégration dans des ensembles plus vastes, pensons au réseau de la Banque Carrefour de la Sécurité Sociale.[11] Rencontrer les personnalités responsables de leur conception et de leur gestion afin de réaliser une analyse rigoureuse de celles-ci est urgent et essentiel. Nous avons défini, au seuil de cette étude, la signification de la démarche heuristique dans le cadre informatique. Nous présentons ici une synthèse des résultats obtenus.

3.2. *Historique de la banque de données*

Chaque système d'information a sa propre histoire. Les concepteurs d'une banque de données sont en effet souvent confrontés à de nombreux aléas qui rendent impossible l'élaboration d'un système conforme aux modèles théoriques. L'évocation des différentes étapes du développement d'une banque de données permet d'emblée de cerner les difficultés auxquelles sont confrontés ses gestionnaires et d'appréhender leur incidence sur la qualité de l'information.

A la fin des années septante, par exemple, les liaisons entre applications se sont multipliées et des systèmes techniquement plus performants ont peu à peu remplacé les anciens. Suite à un cloisonnement entre services informatiques, au manque de communication entre personnel administratif et informaticiens ou encore à la mise en oeuvre de solutions partielles afin de répondre à des problème ponctuels, une redondance informationnelle a pu apparaître et prendre de l'ampleur au

---

11. La Banque Carrefour de la Sécurité Sociale (B.C.S.S.) est un interface permettant la transmission d'informations entre les organismes de sécurité sociale via un réseau en étoile. La B.C.S.S. ne doit pas contenir d'information mais des références aux données (répertoire des références) conservées de façon décentralisée et distribuée. Tout en garantissant la confidentialité des données, ce réseau stellaire doit supprimer toute communication latérale d'information entre les organismes concernés. La B.C.S.S. a pour objectifs principaux d'instaurer un système de collecte unique de l'information et de rationaliser l'accès à l'information.

sein d'un même système. Celle-ci s'est manifestée par un développement parallèle de fichiers, flux d'information et banques de données répertoriant ou véhiculant une information partiellement similaire. Parfois réalisées dans la précipitation, ces transformations techniques, lorsqu'elles ne sont pas précédées d'un travail conceptuel approfondi, ont ainsi introduit un manque de cohérence global au sein des systèmes informatiques.

Par ailleurs, les gestionnaires des banques de données administratives ont, dans la plupart des cas, été confrontés aux contraintes juridiques de la force probante.[12] Bien que de nouveaux textes légaux concernant l'authenticité des données sur support électronique soient actuellement à l'étude, de nombreuses informations, comportant notamment la signature des individus concernés, doivent toujours être envoyées sur un document écrit et faire l'objet d'un encodage, plus ou moins lourd suivant les cas. Ceci a entraîné la coexistence de plusieurs circuits de saisie pour un même type d'information: certaines données pouvant être envoyées sur support électronique et être automatiquement intégrées dans la banque de données alors que d'autres données doivent être encodées. Cette intégration multiple suppose d'une part, une vérification constante de la présence simultanée des différents volets d'un même document et d'autre part, de leur cohérence interne, ce qui va bien entendu alourdir le processus de traitement de l'information et accroître le risque d'apparition d'erreurs.

Depuis quelques années un nouveau phénomène apparaît: le développement de réseaux entre banques de données, comme, en Belgique, le réseau de la Banque Carrefour de la Sécurité Sociale. De telles entreprises ne se réalisent évidemment pas sans difficultés. Un important travail de mise en cohérence s'impose tant sur le plan technique que conceptuel. Les infrastructures informatiques des différentes banques de données concernées sont loin d'être homogènes et n'ont pas atteint le même niveau de développement. Sur le plan conceptuel, certaines données, dont l'intitulé est similaire et la signification apparemment identique, n'ont en réalité pas le même sens d'un système d'information à l'autre! Nous reviendrons sur ce point dans l'analyse critique.

---

12. Voir par exemple: M. ANTOINE, M. ELOY et J.-F. BRAKELAND, *Le droit de la preuve face aux nouvelles technologies de l'information* dans *Cahiers du Centre de Recherches Informatique et Droit*. Namur, Story-sciencia, 1992.

3.3. *Diagramme des flux d'information et schéma conceptuel de la banque de données*

Dans le cadre restreint de cet exposé, nous présentons ici des schémas sommaires et simplifiés. Pour plus de détails, nous renvoyons le lecteur à l'étude approfondie dont s'inspire cet article (voir note 5).

3.3.1. Diagramme des flux d'information

Faisant abstraction de la technique, le diagramme des flux d'information[13] permet de décrire un système d'information en termes de fonctions à réaliser. Alors que le langage naturel est trop vague et trop peu synthétique, ce type d'analyse structurée fait généralement office d'interface entre l'optique des utilisateurs et celle des informaticiens. Lors de la conception et de l'installation d'un système d'information, il représente l'outil adéquat permettant de rédiger un cahier des charges. Dresser un tel diagramme lorsque l'on se trouve face à une banque de données déjà opérationnelle peut également s'avérer très utile. L'aspect évolutif de tout système d'information implique que l'on fasse régulièrement "le point": en cours de fonctionnement, de nouveaux problèmes peuvent surgir et d'anciens processus peuvent devenir inadéquats ou obsolètes. Dans le feu de l'action, il arrive fréquemment que les responsables d'un système, faute de recul, ne prennent pas conscience de ces mutations. Effectuer à nouveau une analyse fonctionnelle prend alors tout son sens.

Voyons, à titre d'exemple, un diagramme simplifié qui nous permettra d'illustrer la démarche (figure 1). Le diagramme que nous présentons schématise le fonctionnement général d'un système informatique administratif "standard".[14] Concrètement, les citoyens doivent régulière-

---

13. A l'heure actuelle, les méthodes permettant de modéliser les systèmes d'information sont nombreuses. Nous nous inspirons ici de la méthode de Yourdon décrite dans l'ouvrage suivant: C. GANE et T. SARSON, *Structured system analysis. Tools and techniques*. New Jersey, Prentice-Hall, Inc., 1979.

14. Dans la figure 1, les rectangles arrondis désignent des "processus" se déroulant successivement. Les indications figurant sur un processus en désignent la fonction, l'instance responsable de sa mise en oeuvre et le numéro d'ordre. Lorsque les numéros d'ordre des processus sont suivis d'une lettre, il s'agit de processus se déroulant simultanément. Les flèches désignent des "flux d'information" (data flow). Les rectangles allongés désignent des fichiers (data store) identifiés par un nom et un numéro. Les carrés désignent des "interfaces", acteurs extérieurs au système d'information.

ment envoyer à l'administration des documents dûment complétés (déclaration d'impôts, de sécurité sociale, ...). Ces documents sont ensuite triés par les services de l'administration et intégrés dans la banque de données. Les documents écrits feront l'objet d'un encodage alors que les informations figurant sur un support électronique (disquette, bande magnétique, ...) seront automatiquement intégrées. Mais ces données peuvent contenir des erreurs liées par exemple au fait que tel individu a mal interprété les consignes permettant de compléter le document administratif ou que tel autre a volontairement transformé une situation réelle afin d'échapper au paiement d'une taxe, d'un impôt ou d'une cotisation. Les services de l'administration doivent donc préalablement tester les informations envoyées. Ces tests peuvent se réaliser via une procédure de contrôle de cohérence interne ou par comparaison avec d'autres informations relatives aux mêmes citoyens stockées sur des fichiers de contrôle ou d'aide à la saisie. Un document contenant trop d'erreurs peut être renvoyé à l'expéditeur pour correction. Les autres documents sont intégrés dans la banque de données et exploités par l'administration à des fins de gestion administrative ou de traitements statistiques. Les informations stockées dans la banque de données administrative peuvent également être communiquées à l'extérieur (universités, instances politiques, ...) et être exploitées à d'autres fins (analyses scientifiques, études prévisionnelles, ...).

A la lecture de ce diagramme, nous voyons apparaître **quatre types d'acteur:**

1) Les **"individus extérieurs"** (entité externe) qui envoient les documents dûment complétés à l'organisme administratif, chargé de les gérer, en collaboration avec les instances informatiques.

2) Les **instances informatiques** qui se chargent de la saisie des données (processus 2 A – intégration automatique – et B – encodage –), de leur contrôle (processus 3A et B) et des opérations de traitement statistique (processus 5).

3) Les **instances administratives** qui interagissent avec les instances informatiques. Lors de la réception des documents, des contrôles de fond et des traitements statistiques. Par ailleurs, celles-ci se chargent de la correction des erreurs détectées dans la banque de données (en contactant éventuellement les individus concernés) et consultent la banque de données dans le cadre de leurs tâches administratives.

4) Les **services universitaires** qui, dans le cadre de leurs recherches, effectuent également des traitements statistiques.

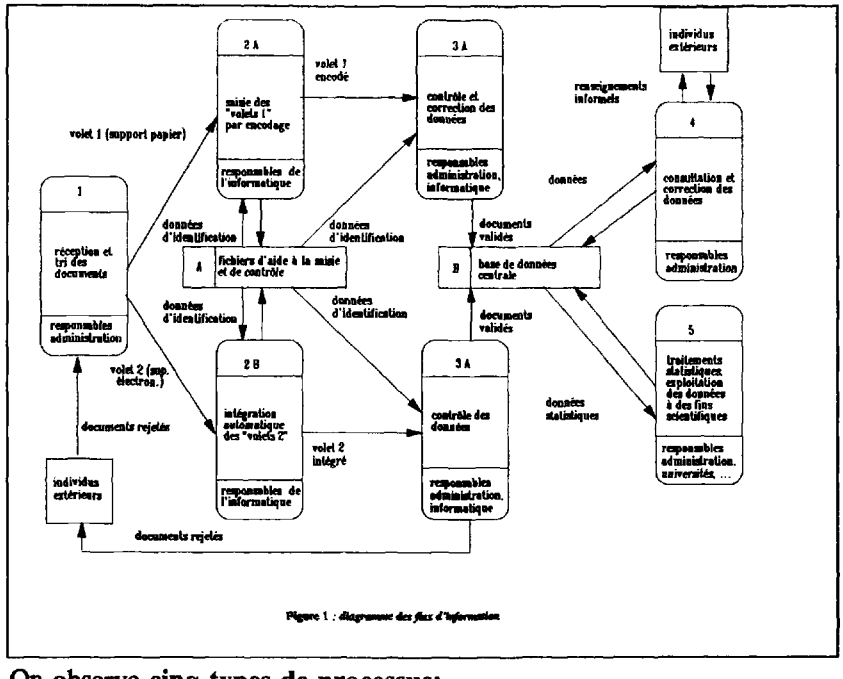

Figure 1 : *diagramme des flux d'information*

On observe cinq types de processus:

1) **Réception et tri des documents** (Processus 1);

2) **Saisie et intégration des documents** (processus 2A et B);

3) **Contrôle de fond** (processus 3A et 3B): tests portant sur la cohérence des données;

4) **Consultation et correction des données** (processus 4). Les erreurs décelées par les instances informatiques sont communiquées aux instances administratives chargées de les corriger. A ce niveau, des contacts s'établissent à nouveau avec les individus extérieurs;

5) **Traitements statistiques** (processus 5) réalisés au sein de l'administration ou à l'extérieur, dans le cadre, par exemple de l'exploitation des données à des fins scientifiques.

Deux types de fichiers (ou data store) apparaissent:

1) **Fichiers d'aide à la saisie et de contrôle** (A);

2) **Fichiers de la Banque de données centrale** (B).

Ce schéma sommaire nous offre une vue panoramique sur le fonctionnement d'un système "informatico-administratif". Voici deux observations critiques que l'on peut émettre à titre d'exemple:

1) L'information intégrée dans la banque de données émane d'une part de circuits formels (processus 1 et 2: encodage et intégration automatique des documents) mais aussi informels, lors des corrections (processus 4: communication d'informations par voie téléphonique, courrier ...). Ce

second circuit d'information est difficile à maîtriser: à tout moment de nouvelles données peuvent émaner de l'extérieur. Par conséquent, il est quasi impossible de savoir si l'information relative à un ensemble d'individus, à un instant t, est complète, correcte et homogène.

2) Comme nous l'avions évoqué dans le paragraphe précédent consacré à l'histoire des banques de données, suite à la contrainte juridique de la force probante, deux circuits de saisie formels peuvent être nécessaires afin d'assurer l'intégration d'un seul type de document (encodage et intégration automatique). C'est seulement en fin de parcours, lors des processus de correction des données que la cohérence interne entre les différents volets d'un même groupe d'informations est assurée. En fonction de la qualité de l'input, ce processus peut s'étaler sur un laps de temps plus ou moins long. Or nous voyons que simultanément un prélèvement d'information est effectué afin de réaliser des traitements statistiques alors que la qualité de la source utilisée à cette fin est aléatoire.

3.3.2. Schéma conceptuel

Le schéma conceptuel d'une banque de donnée est un outil de modélisation qui sert de support logique à l'implémentation physique.[15] Il existe de nombreux outils de modélisation de banque de données: schémas "entité-association", "relationnel", "orientés objet" ...[16] Chacun de ces schémas est doté de caractéristiques propres. Le schéma "entité-association" permet de représenter une situation concrète, proche du réel étudié. Il intervient lors de la première étape formelle du processus de modélisation. Nous en présentons un exemple sommaire et partiel afin de mettre en exergue certains problèmes que l'on peut déceler en observant les relations entre fichiers (figure 2). Alors que le schéma précédent présentait une vue globale du système "informatico-administratif", ce schéma présente une vue globale des relations conceptuelles entre données et groupes de données. Sur le schéma que nous présentons, les rectangles

---

15. R. ELMASRI et S.-B. NAVATHE, *Fundamentals of database system*. Redwood city. The benjamin/Cummings Publishing Company, Inc., 1989. C. BATINI, S. CERI et S.-B. NAVATHE, *Conceptual database design. An entity-relationship approach*. Redwood city, The benjamin/Cummings Publishing Company, Inc., 1992. F. BODART et Y. PIGNEUR, *Conception assistée des systèmes d'information. Méthode, modèles, outils*. Paris, Masson, 1993.

16. P. COAD et E. YOURDON, *Analyse orientée objets*. Paris, Masson, 1993. F. VAN ASSHE, B. MOULIN et C. ROLLAND, *Object oriented approach in information systems*. Amsterdam, IFIP, 1991.

représentent des "entités conceptuelles", caractérisant un ensemble d'informations homogènes (par exemple, les informations relatives aux travailleurs). Les ovales reliés à ces entités sont des "attributs" qui représentent les propriétés des entités (par exemple, nom, prénom et sexe d'un travailleur). Les attributs soulignés (dans notre schéma "*identifiant unique*") désignent les "clés": une clé est constituée d'un ou plusieurs attributs dont les valeurs permettent de désigner chaque occurrence d'une entité de façon univoque (par exemple, le numéro de registre national des travailleurs). Les entités sont reliées entre elles par des "associations" (losanges) dont les combinaisons sont limitées par des "cardinalités" (couples de caractères entre parenthèses). Par exemple, à une occurrence de l'entité "employeur" pourrait correspondre une ou plusieurs (1, n) occurrences de l'entité "travailleur".

Nous observons, à l'examen de la figure 2, que l'une des deux entités constitutives de la banque de données est en relation avec une entité "statistiques". Contrairement aux autres entités, cette dernière n'est pas dotée d'un attribut faisant office d'identifiant unique. De ce fait, lors du processus d'agrégation des données, un phénomène de perte d'information se produit. Une fois les données agrégées, il n'est plus possible de "faire marche arrière": le lien entre les données individuelles de base et les informations groupées (à des fins de traitement statistique) est perdu. Nous en verrons l'incidence sur la qualité des données dans l'analyse critique qui suit.

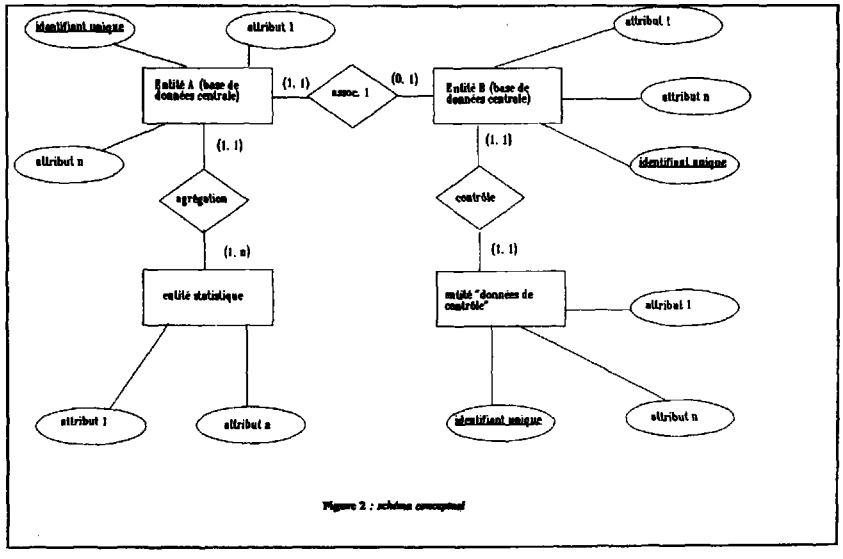

Figure 2 : schéma conceptuel

4. ANALYSE CRITIQUE

Nous nous proposons de suivre le cheminement de l'information depuis son entrée dans le système informatico-administratif décrit dans le diagramme des flux d'information jusqu'à son exploitation statistique par des services extérieurs. Ce faisant, nous mettrons l'accent sur les passages successifs du statut d'input à celui d'output et sur les risques d'apparition d'erreurs qu'impliquent de tels changements. Les conclusions que nous présentons ici sont le résultat d'une analyse approfondie dont l'objet était de tester la valeur des données en amont et en aval de chaque processus.

Conformément à la structure définie au seuil de cette étude, nous envisagerons successivement les trois stades suivants: constitution de l'input, traitements informatiques internes et constitution de l'output final.

1. Constitution de l'input - **Elaboration conceptuelle de l'input: processus de formalisation des données** - **L'input: données interprétées et complétées par les individus répertoriés dans la banque de données**
2. Traitements informatiques internes - Saisie des données - De la détection des erreurs aux corrections
3. L'output "final" - Résultat des traitements internes - Exploitation des données par des services extérieurs

4.1. Constitution de l'input

4.1.1. Elaboration conceptuelle de l'input: processus de formalisation des données

Le travail de formalisation des données constitue le premier stade du "processus de transformation". Voyons quels en sont les étapes et les protagonistes. Les informations répertoriées dans une banque de données administrative sont directement liées à la législation en vigueur. Une fois fixés par la loi, les concepts doivent être codifiés afin d'être introduits dans un système "informatico-administratif".

Deux sources d'ambiguïté, liées à la non harmonisation de la législation peuvent se présenter à ce stade. Dans le domaine de la sécurité sociale, par exemple, il existe des concepts qui, sous des intitulés

distincts, couvrent une même réalité et d'autres qui, inversement, sous un intitulé similaire, n'ont pas la même signification d'un régime à l'autre. La notion de "journée de travail", par exemple, ne signifie d'une part pas la même chose pour les organismes chargés de percevoir et de répartir les cotisations et évolue d'autre part au fil des modifications législatives.[17] Ces divergences sont d'une part à l'origine d'injustices dans la répartition des droits sociaux.[18] Et d'autre part, elles hypothèquent la mise en place d'une informatique de sécurité sociale globale et rationnelle. Car, comment formaliser des données dont la signification d'origine prête à confusion?

D'un autre ordre, mais toute aussi lourde de conséquence, la fréquence des modifications législatives constitue une entrave à la formalisation et à l'automatisation des données. Quelles en sont les implications? Lorsqu'une loi est sur le point de paraître, le texte du projet est généralement communiqué anticipativement au comité de gestion des instances administratives: en collaboration avec les services informatiques, celles-ci peuvent envisager les adaptations à mettre en oeuvre ... mais rien ne dit que le texte définitif paraîtra sous la même forme.

Les délais permettant de prendre des mesures concrètes sont donc extrêmement courts. De plus, le contenu des textes est parfois flou quand il n'a pas pour objet la modification ou l'abrogation de mesures imposées quelques mois auparavant (certaines lois peuvent avoir un effet rétroactif). C'est en catastrophe qu'il faut comprendre et interpréter la loi, traduire les concepts législatifs en concepts formalisés et intégrables dans la banque de données, concevoir de nouveaux formulaires, les faire imprimer, mettre à jour la banque de données et rédiger les fascicules explicatifs destinés aux personnes chargées de les compléter ...

---

17. Alors que depuis 1987, l'O.N.S.S. considère toute journée entamée comme une journée de travail, jusqu'en 1989, une journée de travail devait compter au moins trois heures de prestation pour l'O.N.A.F.T.S. (Allocations familiales), jusqu'en 1990, 3 heures pour l'I.N.A.M.I. et actuellement encore, 6 heures pour l'O.N.E.M. En d'autres termes, si le champ des individus concerné est élargi au maximum quand il est question de percevoir des cotisations, il est autrement plus réduit lorsqu'il s'agit de redistribuer le revenu des perceptions. Actuellement, les instances chargées de répartir les cotisations s'efforcent peu à peu d'évaluer les montants sur une base hebdomadaire ou trimestrielle. Les promoteurs de la B.C.S.S., conscients de ces problèmes ont instauré une commission pour l'harmonisation de la législation sociale.

18. De 1987 à 1989, une personne ne travaillant pas plus de trois heures par jour pour le même employeur devait verser des cotisations à l'O.N.S.S. mais n'avait pas droit aux allocations familiales.

4.1.2. *L'input: données interprétées et complétées par les individus répertoriés dans la banque de données*

Une fois définis, les intitulés des données sont fixés sur un document: les valeurs correspondantes vont être complétées par les individus concernés.

Nous allons voir qu'avant même d'être intégrée dans la banque de données, l'information porte en germe des altérations de sens dont certaines sont invérifiables par la suite. Ainsi en est-il lorsque les personnes concernées ont mal interprété les instructions expliquant comment remplir un document administratif.[19]

Aussi sophistiqués que soient les contrôles ultérieurs, certaines de ces erreurs, présentes dans "l'input" du système d'information sont indécelables par la suite.[20] Notons en outre que le taux d'anomalies de ce type est particulièrement élevé lorsque les informations proviennent de secteurs d'activité connaissant une grande instabilité au niveau des flux de travailleurs (secteurs du bâtiment, de l'hôtellerie, du travail intérimaire, ...). Ces derniers, dont la situation professionnelle est précaire, changent souvent d'employeur et de statut. Il est dès lors difficile pour les responsables chargés de remplir les documents administratifs de comptabiliser les début et fin de contrat dans les délais imposés par l'administration. Aller sur le terrain pour interroger les employeurs est le seul moyen de connaître avec précision les pratiques en la matière.

---

19. Un responsable de l'I.N.A.M.I. affirmait à ce sujet: "*... on peut difficilement s'attendre à ce que tout un chacun puisse évaluer et interpréter correctement ces données, compte tenu de la complexité de la législation sociale. Il en résulte des malentendus, de la confusion et des erreurs qui ne peuvent être contrôlées par les instances destinataires de l'information, à savoir les mutualités. En cas de contestation, l'assuré ne dispose même pas de possibilité de recours car il a fourni lui-même les informations qui ont donné lieu à un calcul erroné ... Nous ne pouvons escompter que chaque employeur soit parfaitement familiarisé avec l'ensemble de la législation sociale, de sorte, qu'ici également, une interprétation uniforme et correcte des données ne peut plus être garantie.*" D. HERMIE, *Objectifs et conséquences de l'instauration de la Banque carrefour de la sécurité sociale* dans *Revue belge de sécurité sociale*. Numéro spécial, Bruxelles, Ministère de la Prévoyance sociale, 1989, p. 94.

20. "*Un travailleur malade ou accidenté, absent du travail depuis un certain temps, peut, malgré les instructions précises données à l'employeur, être omis sur le document de déclaration, sans qu'il soit possible de le détecter.*" *Office National de Sécurité Sociale, Rapport annuel. Exercice 1990*. Bruxelles, O.N.S.S., 1990, p. III-9.

Une dernière remarque s'impose: dans un système diachronique, des données conceptuellement interdépendantes sont physiquement dissociées et perdent leur sens. Par exemple, plusieurs documents concernant les prestations d'une même personne au cours d'une même période peuvent être envoyés à des moments différents: comment un système informatique peut-il gérer un ensemble de données fragmentaires dont l'étendue et la nature ne sont pas connues a priori?

4.2. *Traitements informatiques internes*

A ce stade, les données ont été définies et leurs valeurs complétées (dans le cas où les informations sont envoyées sur un support physique). De nombreuses données peuvent encore être intégrées dans le système de façon informelle (par courrier, téléphone, via un contact direct avec un placeur ...). Nous nous trouvons au sein du système informatico-administratif. Les mécanismes suivants vont maintenant intervenir: saisie des données, contrôle des données et correction des erreurs. En amont et en aval de chacun de ces processus internes, les données vont successivement passer du statut d'input à celui d'output.

1. Constitution de l'input - Elaboration conceptuelle de l'input: processus de formalisation des données - L'input: données interprétées et complétées par les individus répertoriés dans la banque de données
2. Traitements informatiques internes - **Saisie des données** - **De la détection des erreurs aux corrections**
3. L'output "final" - Résultat des traitements internes - Exploitation des données par des services extérieurs

4.2.1. *La saisie des données*

Citons trois sources d'erreur pouvant survenir à ce stade:
1. déformation des données lors de l'encodage;
2. introduction d'imprécisions lorsque les données sont communiquées de façon orale et sujettes à l'interprétation humaine;
3. introduction d'anomalies en provenance des fichiers d'aide à la saisie.

Ce troisième point mérite quelques explications. Afin d'éviter l'encodage répétitif et fastidieux de données dont les valeurs sont relativement stables dans le temps (nom, prénom, sexe, date de naissance, ...), une connexion automatique est assurée (via un identifiant censément unique) avec un fichier de type "registre de population", lequel fournit directement ces informations. Le problème vient du fait que, comme nous l'avions souligné dans le cas des Pays-Bas, ce "méga fichier" contient lui-même des incohérences qu'il est extrêmement difficile de corriger à posteriori. Ainsi, certaines personnes dont le nom a été mal orthographié sont reprises plusieurs fois, sous des numéros d'identification différents. Le cas se produit notamment lorsque des femmes mariées sont répertoriées tantôt sous leur nom de jeune fille, tantôt sous leur nom d'épouse ...

Ce problème, que connaissent de nombreux pays, a été évoqué lors de la conférence sur l'informatique et la sécurité sociale dans les Etats membres de la Communauté européenne qui a eu lieu à Londres en décembre 1992.

En Belgique, depuis la mise en place du réseau de la Banque Carrefour, des programmes sont mis au point afin de redresser ce type d'erreurs:[21] ces modules se basent sur le nom et sur un historique des adresses de chaque individu. Il va de soi que la réussite de l'opération passe par une coordination des traitements effectués simultanément par chaque système d'information ayant accès au Registre national. Notons que ce type de problème touche non seulement les grandes banques de données administrative mais aussi des ensembles informatiques fonctionnant dans d'autres secteurs, tel le secteur bancaire.

Signalons que des biais similaires se produisaient déjà au XVIIème siècle, dans le cadre de la tenue des Registres paroissiaux:

> "L'obligation (Saint Germain 1667) de tenir les registres en double et d'en déposer un exemplaire au greffe va centraliser le réseau tout en multipliant les inexactitudes (erreurs de recopiage). Mais surtout, cette comptabilité n'est pas fonctionnelle mais sacramentelle: on inscrit les baptêmes et non

21. H. DEJONCKHEERE et F. HOLDERBEKE, *De kruispuntbank: en stand van zaken. Gesprek met Dhr F. ROBBEN* dans: *Steunpunt. Werkgelegenheid Arbeid Vorming*. Louvain, avril 1992, p. 2 à 4.

les naissances, les enterrements et non les morts, et ceci produit des biais considérables."[22]

4.2.2. *De la détection des erreurs aux corrections*

Une fois intégrées, les données entrent dans un cycle qui va du contrôle de fond aux corrections. Mais ce cycle n'est ni régulier ni homogène. Si l'intégration des informations est une opération diachronique, les traitements ultérieurs se réalisent dans la synchronie: il n'est pas possible de déterminer a priori un instant t au terme duquel les contrôles et corrections relatifs aux déclarations d'un même trimestre seraient terminés.

Dans le meilleur des cas, les contrôles de fond portent sur l'intégralité des valeurs potentiellement présentes dans un document: les erreurs de calcul, incohérences internes, lacunes et valeurs aberrantes sont systématiquement décelées. Echappent cependant à ces contrôles toutes les anomalies n'entravant pas la logique interne d'un ensemble d'informations: inversion de valeurs relatives à des personnes distinctes, orthographe erronée d'un nom ...

Le travail de rectification est une opération délicate car la plupart des données sont interdépendantes. La correction d'une erreur peut ainsi entraîner l'apparition de nouvelles anomalies qui n'avaient pas été décelées au départ.

Lorsque des cas "hors norme" se présentent, leur gestion relève de la décision subjective des correcteurs: en dernière instance, il leur est possible de "court-circuiter" un contrôle automatique si après examen, une donnée déclarée "fausse" se révèle correcte. Laisser une place à l'intervention humaine présente des risques: lorsqu'un utilisateur décide de valider une donnée dite "erronée", le système ne réalise plus de contrôle sur les modifications ultérieures.

4.3. *L'output "final"*

Nous arrivons progressivement en fin de parcours: les données ont été formalisées et les valeurs complétées; les documents sont peu à peu intégrés et les processus de contrôle et de correction des erreurs sont

---

22. J.-L. BESSON et O. JOURNET, *Le nombre et son ombre* dans *Des Mesures*. Etudes coordonnées sous la direction de J.-L. BESSON et M. COMTE. Lyon, Presses Universitaires de Lyon, 1986, p. 22.

déclenchés. La banque de données constitue donc une source provisoirement finale.

L'output "final" peut être simplement consulté ou encore, faire l'objet de traitements statistiques. Notons qu'une banque de données administrative n'est pas fondamentalement conçue pour être exploitée à des fins statistiques mais bien pour faciliter la mise en oeuvre des tâches administratives. En effet, le caractère hétérogène de l'information répertoriée dans de telles banques de données constitue un obstacle énorme à l'élaboration de données statistiques fiables: nous avons vu que les informations relatives à une même période ne sont ni intégrées, ni corrigées simultanément.

1. Constitution de l'input - Elaboration conceptuelle de l'input: processus de formalisation des données - L'input: données interprétées et complétées par les individus répertoriés dans la banque de données
2. Traitements informatiques internes - Saisie des données - De la détection des erreurs aux corrections
3. L'output "final" **- Résultat des traitements internes** **- Exploitation des données par des services extérieurs**

4.3.1. Résultat des traitements internes

Nous citons ici, à titre d'exemple, deux situations au cours desquelles un phénomène de perte d'information se produit lors de l'élaboration de statistiques: la première donne lieu à l'obtention d'effectifs fictivement élevés et la seconde découle de l'usage d'unités statistiques inadéquates.

1° *Effectifs fictivement élevés*

Comme nous l'avons illustré dans le schéma conceptuel, une perte d'information peut se produire lors du processus de comptabilisation des

données et donner lieu, par exemple, à l'obtention d'effectifs fictivement élevés.[23]

Ces biais se produisent lorsque, au sein de la banque de données, le lien entre les données individuelles et les données agrégées disparaît. Une étude minutieuse du schéma conceptuel et des flux d'information permet généralement de trouver des algorithmes susceptibles d'éliminer ces biais. Bien souvent, la marge d'erreur liée à ces déformations est difficilement évaluable. Trouver des solutions à ce type d'erreur est toutefois capital dans la mesure où d'autres organismes administratifs semblent rencontrer des problèmes similaires. Dans le cas du régime de pension, par exemple, certaines personnes peuvent cumuler des avantages distribués par différents organismes. Les statistiques publiées par ces instances contiennent dès lors, un taux de "doubles emplois" qu'il est impossible d'évaluer.[24] Dans certains cas, ces doubles comptages peuvent provoquer une répartition inéquitable des avantages sociaux.[25]

---

23. Ainsi, en ce qui concerne l'O.N.S.S.: *"Pour des raisons techniques, un employeur est compté plus d'une fois dans la statistique si, en raison des modalités spéciales existant en matière de perception des cotisations ..., il appartient à plus d'une catégorie d'employeur et de ce fait renvoie à l'O.N.S.S. plus d'une formule de déclaration par trimestre. On évalue l'effectif des employeurs dans ce cas à 3380 en 1990. Cette méthode de dénombrement augmente fictivement l'effectif des employeurs lors de l'attribution d'une nouvelle catégorie d'employeur."*
"Certaines anomalies n'ont pu être évitées dans les conditions présentes: elles proviennent de la possibilité, pour un même travailleur d'occuper des emplois simultanés (soit des emplois partiels, soit un emploi principal et des emplois de courte durée) auprès d'employeurs différents, ce qui provoque inévitablement des doubles emplois." Office, ... p. III-7 et p. III-9.

24. Le rapport statistique annuel de l'O.N.P. stipule que *"en ce qui concerne le nombre de bénéficiaires, il existe entre les statistiques de l'O.N.P. et celles des autres organismes et services qui procèdent également au paiement des pensions et d'autres prestations, un grand nombre de "doubles emplois" qui ne peuvent être ni localisés, ni exprimés en chiffres." Statistique annuelle des bénéficiaires de pension.* Bruxelles O.N.P., 1990, p. 10.

25. Ainsi, dans le cadre de l'I.N.A.M.I.: *"... l'utilisation du numéro national limitera les comptages doubles lors de la fixation des effectifs, de sorte que les cotisations pourront être réparties avec plus de précision entre les unions nationales. Ces comptages doubles résultent de la double qualité que peuvent avoir certains assurés salariés/indépendants, travailleur à temps partiel/chômeur, changement de qualité pendant un trimestre, changement d'employeur dans le cours d'un trimestre, plusieurs employeurs, etc."* D. HERMIE, *Op. Cit.* p. 92.

2° *Usage d'unités statistiques inadéquates*

L'exactitude d'une donnée est relative et varie en fonction de l'usage que l'on en fait. Reprenons l'exemple de la notion de "journée de travail".

> "Une attention toute particulière doit être accordée à l'extension du champ d'application de la sécurité sociale qui englobe, depuis le 1er octobre 1987, les travailleurs dont les prestations ne dépassent pas habituellement deux heures par jour. Suite à cette modification, le nombre de journées de travail déclarées à partir du quatrième trimestre 1987 a subi une augmentation certaine. La législation prévoit en effet que toute partie de journée, pour laquelle des cotisations de sécurité sociale sont calculées, prend la valeur d'une unité au même titre qu'une journée complète. Par conséquent les gains moyens journaliers subissent une tendance à la baisse; leur publication et celle des indices d'évolution n'ont dès lors plus beaucoup de sens ..."[26]

Si du point de vue administratif et juridique, la notion de "journée de travail" est clairement définie, nous voyons qu'une fois sujette à des traitements statistiques, elle ne peut donner lieu qu'à un résultat biaisé.[27]

Comme nous l'avons signalé dans le paragraphe consacré à la formalisation des données, certains concepts n'ont pas le même sens d'un régime de la sécurité sociale à l'autre, et donc d'une banque de données à l'autre, ce qui hypothèque toute confrontation entre les résultats statistiques des organismes concernés.

4.3.2. Exploitation des données par des services extérieurs

Les informations répertoriées dans la banque de données peuvent sortir du cercle restreint que constituent les instances administratives et informatiques et être exploitées à des fins scientifiques ou politiques. Si ces données diffusées ne sont accompagnées d'aucun appareil critique, leur exploitation risque de donner lieu à des résultats biaisés qui à leur tour, peuvent être traités et transformés et ainsi de suite .... La probabilité est grande d'obtenir des informations dont la marge d'erreur est telle que les interprétations qu'on en fait n'aient finalement plus aucun sens.

---

26. *Office* ..., p. III4.-

27. Evaluer le temps de prestation en terme d'heures de travail permettrait de réaliser des traitements statistiques offrant une image plus proche de la réalité. Il convient à ce sujet de signaler le projet pilote mené par l'O.N.S.S.A.P.L.

Rappelons que l'usage de l'information répertoriée dans une banque de données administrative à des fins statistiques est une opération risquée puisque la banque de données n'a pas été conçue dans ce but. Malheureusement, il arrive fréquemment que les utilisateurs extérieurs exploitent l'information fournie avec précipitation, sans effectuer préalablement les investigations critiques qui s'imposent. Par ailleurs, l'usage d'informations issues des diverses banques de données de la sécurité sociale à des fins statistiques est compromise en raison de l'hétérogénéité des définitions adoptées et des pratiques de codification en vigueur (nous avons déjà signalé à ce sujet les significations divergentes que pouvait prendre la notion de "journée de travail").

5. CONCLUSIONS

La mise en oeuvre de l'analyse critique qui précède nous autorise à répondre – de façon partielle et provisoire – aux questions que nous soulevions au seuil de cette étude. Vu la nature de notre source de travail, les conclusions que nous présentons concernent principalement les systèmes informatiques développés dans le secteur administratif.

5.1. *Influence de l'environnement informatique sur la qualité des données*

Une remarque d'ordre général, relative à l'impact des systèmes informatiques dans notre société s'impose. Nous avons pu observer combien la tenue d'une banque de données administrative était influencée par des facteurs d'ordre juridique et politique. Ceci n'est pas nouveau. Dès le moyen âge, l'émergence progressive de l'impôt royal a peu à peu entraîné la nécessité de dénombrer précisément l'ensemble de la population qu'il s'agissait de soumettre.[28] Dans l'environnement informatique, ce phénomène peut donner lieu à une relation systématique et réciproque entre la qualité des données et certains aspects de l'exercice du pouvoir politique. Nous avons vu que lors du processus d'élaboration conceptuelle de l'input, la législation en vigueur pouvait contenir des ambiguïtés qui ont une répercussion immédiate sur la codification des données. De là peuvent naître, en output, des séries de données incohérentes et hétérogènes. Or, l'output fait l'objet de traitements statistiques donnant lieu à des prévisions conjoncturelles. Ces prévisions influenceront l'action politique dont les mesures, une fois inscrites dans

28. Voir par exemple: J.-L. BESSON et O. JOURNET, *Op. cit.* p. 16 à 23.

la loi, auront, à leur tour, un impact sur la nature de l'input. Nous observons ainsi qu'une banque de données est non seulement un instrument de gestion mais aussi un outil d'aide à la décision et qu'il existe ainsi un processus régulier de feed-back, de rétroaction de l'output vers l'input. Dès lors, une étude approfondie devra être menée sur les relations existant entre l'informatique, la qualité de l'information et certaines décisions politiques.

Dans l'introduction, nous avions défini l'output comme "un ensemble fini de données structurées, censé correspondre à une partie du réel observable et destiné à enrichir notre connaissance de ce réel." Nous observons maintenant que l'output est le résultat d'une double structuration: d'une part, de la structure définie lors de la constitution de l'input et d'autre part, de la structure qui s'est progressivement établie au fil des processus internes. Ces derniers se déroulent au sein du système informatique, ensemble clos, dont les éléments interagissent. A l'intérieur de ce système, de nouvelles données apparaissent, les unes destinées à enrichir l'input (par exemple, résultats d'opérations arithmétiques), les autres, cachées, purement informatiques (par exemple, codes de validation des tests). Voyons maintenant quel est l'impact de ce double mouvement de structuration sur la qualité des données.

Au cours du processus qui va de la constitution de l'input à celle de l'output, dans quelle mesure un système informatique est-il susceptible de détériorer la qualité des données?

Une première réflexion nous semble essentielle. Si les procédures informatiques obéissent à une logique rationnelle, la réalité qu'il s'agit de gérer est bien souvent fluctuante, insaisissable et irrationnelle. La philosophie sous-jacente aux modèles informatiques traditionnels (prônant une approche "top-down" de la réalité[29]) nous semble dès lors trop optimiste. Inadéquats, ces modèles sont cependant censés représenter intégralement une réalité qui leur échappe en partie. Nous avons montré qu'au delà d'un certain seuil, un système informatique ne pouvait gérer l'exception; certaines subtilités sont ainsi gommées par les programmes de mise en cohérence ou lors des processus d'agrégation des données. Dès lors, en raison de la complexité du réel appréhendé et des traitements à mettre en oeuvre, obtenir en output des données parfaitement correctes est illusoire. La question n'est donc pas d'espérer rencontrer un système

29. Approche ayant pour but de représenter une réalité de façon globale et exhaustive, par étapes successives, en allant progressivement du général vers le particulier.

informatique parfait mais bien d'apprendre à travailler avec des données que l'on sait imparfaites.[30] En d'autre termes, il s'agira de construire un modèle permettant de décrire les données telles qu'elles sont empiriquement et non telles qu'elles devraient être théoriquement.

Concernant le processus d'agrégation des données, un important travail de réflexion critique devra encore être mené. Clairement posée depuis le fameux paradoxe de Condorcet (1785), la question de l'agrégation des données, du passage de l'individuel au collectif, suscite de nombreuses analyses.[31] Concrètement, il s'agira de voir en quoi cette question mérite une étude spécifique lorsque l'on se situe dans un environnement informatique. Dans quelle mesure, suite aux contraintes informatiques, le passage d'un niveau d'agrégation à un autre entraîne-t-il un changement de signification? Et dans quelle mesure ce changement de sens est-il lié à la procédure d'agrégation employée? La présentation des étapes successives de la transformation des données constitue une amorce de réponse à ces questions. Rappelons trois points importants. En premier lieu, de nombreux biais liés aux procédures d'agrégation peuvent être évités lors de l'élaboration du schéma conceptuel de la banque de données (grâce à l'usage de clés primaires adéquates et de clés secondaires suffisamment précises pour maintenir un lien entre fichiers dont le niveau d'agrégation est distinct). Rappelons ensuite que la coexistence, au sein d'un système informatique, de traitements synchroniques et diachroniques peut introduire un décalage conceptuel entre les données individuelles et les résultats agrégés (voir supra, "Importance de la notion de temps: synchronie et diachronie"). Enfin, l'inévitable coexistence de traitements automatiques, dont le fonctionnement et les effets peuvent être objectivement décrits, et de procédures manuelles, empiriques et informelles, lors de la saisie ou de la correction des données, rend difficile une évaluation systématique de l'influence des traitements effectués sur les données.

Une troisième réflexion est liée à la nature même de l'information traitée au sein de la banque de données. En informatique, une donnée dépourvue de documentation est indéchiffrable et devient inaccessible. Or

---

30. Le souci d'apprendre à travailler avec des données imparfaites va évidemment de pair avec celui de créer des systèmes d'information conceptuellement plus proches de la réalité et par là-même, plus performants.

31. Voir par exemple O. ARKHIPOFF, *Importance et diversité des problèmes d'agrégation en comptabilité nationale. Esquisse d'une théorie générale de l'agrégation* dans *La comptabilité nationale face au défi international*, E. ARCHAMBAULT et O. ARKHIPOFF éds. Paris, Economica, 1990, p. 365 à 388.

nous avons vu que souvent cette documentation n'est pas diffusée par écrit et relève d'un savoir exclusivement oral. Il est évident que le recours aux pratiques informelles transmises de façon orale existait bien avant le développement massif des banques de données. Mais l'automatisation de l'information offre des possibilités de stockage, de traitement et de diffusion telles que le moindre point obscur, la moindre confusion, la moindre donnée inintelligible peuvent provoquer des pertes d'information en cascade dont l'ampleur est difficile à mesurer.

Dans le même ordre d'idée, une autre source de détérioration de l'input est liée à l'effet multiplicateur que peut entraîner, au sein d'un réseau, la propagation des erreurs d'une banque de données à l'autre.

Voyons maintenant en quoi les traitements informatiques sont susceptibles d'améliorer la qualité des données.

Les procédures de contrôle automatique des données représentent un outil extraordinairement puissant afin de tester la cohérence interne et la validité des données. Certes, ces tests, considérant tantôt comme erronées des données exceptionnelles mais correctes, négligeant tantôt des anomalies manifestes, ne sont pas absolus. Mais ne perdons pas de vue qu'ils font place à une absence presque totale de contrôle. Il y a à peine dix ans, avant la mise en place généralisée de l'informatique dans l'administration, les fonctionnaires étaient débordés par une masse de documents écrits qu'il était impossible de gérer systématiquement. Les contrôles effectués portaient alors sur quelques documents prélevés au hasard parmi des dizaines de milliers d'autres.

Dans cet esprit, il convient de rappeler une évidence: sans l'apport de l'informatique, la gestion, la diffusion et la confrontation rapides de millions de données sont impensables: à une masse de données dont la marge d'erreur est parfois incertaine, ferait place l'absence totale ou la carence d'information.

Si l'on ne peut éviter l'apparition d'erreurs au sein des banques de données, savoir qu'elles existent, pourquoi et comment elles se déclenchent constitue déjà un grand progrès. Et c'est sur ce plan que l'esprit et la démarche de la critique historique se révèlent efficaces. La mise en lumière du cheminement progressif des données au sein du système informatique permet d'une part, de cerner à tout moment l'origine et la cause d'une éventuelle perte de sens et d'autre part, d'imaginer des méthodes originales permettant d'accroître le degré de fiabilité des données. Dans cette perspective, nous présentons un ensemble de propositions afin d'améliorer la qualité des informations traitées et transformées au sein d'une banque de données.

5.2. *Propositions en vue d'améliorer la qualité des informations issues de banques de données*

Les propositions que nous émettons s'adressent d'une part, aux gestionnaires de banques de données et d'autre part, aux chercheurs qui exploitent ces données à des fins scientifiques. Nous verrons cependant que cette division sur base des tâches et des compétences est artificielle: un travail efficace afin d'améliorer la qualité des données devrait résulter d'une étroite collaboration entre chaque groupe concerné.

5.2.1. *Propositions destinées aux gestionnaires de banques de données*

En amont du système d'information, nous avons vu que la formalisation de données dont la signification d'origine est ambiguë provoquait de nombreuses incohérences. Il est donc impératif d'harmoniser les concepts susceptibles d'être codifiés dans une banque de données. Au besoin, cette harmonisation devrait faire l'objet de modifications législatives.

Au sein du système d'information, pour les raisons évoquées plus haut, il importe de substituer au savoir oral une documentation écrite, systématique et à jour. La tenue systématique de documents écrits décrivant de façon exhaustive chaque information et chaque traitement permet d'éviter que de nombreux problèmes préexistants se développent à une allure exponentielle.[32] Cette première mesure permettrait ensuite de rationaliser les flux d'information, fichiers et banques de données afin d'obtenir une structure informatique plus cohérente.

Des procédures automatiques de correction (par exemple, afin de remédier aux biais et incohérences dans les résultats statistiques) pourraient ensuite être mises en oeuvre. Citons, à titre d'exemple l'opération SUSE.[33]

---

32. Cette démarche s'insère dans le cadre d'une acquisition des connaissances par "reverse engineering". Lorsqu'une base de données a été mal documentée (schéma conceptuel incomplet ou inexistant) ou lorsqu'une perte d'information se produit en raison de la disparition des concepteurs du système d'information, ce domaine de la recherche fondamentale a pour objet de mettre au point une méthode susceptible de capter à nouveau les "méta-informations" manquantes.

33. L'opération S.U.S.E. (Système Unifié de Statistiques d'Entreprises), menée en France sous l'égide de l'I.N.S.E.E., a pour objectif de centraliser l'information sur les entreprises grâce à la confrontation et la fusion de deux sources statistiques: les fichiers fiscaux B.I.C. (Bénéfices Industriels et Commerciaux) et les enquêtes annuelles d'entreprises (E.A.E.). Opération de grande envergure (près de 500.000 entreprises sont répertoriées), S.U.S.E. représente une expérience intéressante de mise en cohérence systématique et exhaustive de données issues de systèmes informatiques.

Par ailleurs, les recherches en matière de modélisation de banques de données offrent de nombreuses perspectives: le développement de systèmes "orientés-objet" et de "systèmes experts" permettront peut-être, à terme, la gestion de données "incertaines"[34] dont la valeur est sujette à une constante évolution. Cela dit, si l'un des objectifs de l'intelligence artificielle réside dans une prise en compte du contexte afin de résoudre l'ambiguïté de certaines situations, de nombreuses questions se posent encore, tant sur le plan conceptuel que pratique. L'expertise humaine est-elle simulable? Peut-on la réduire à un ensemble fini de règles analytiques?[35] De nombreux aspects de l'intelligence artificielle et, plus précisément, des systèmes orientés-objet méritent toutefois notre attention (nous avons évoqué, dans le corps de l'étude, les recherches relatives à la modélisation du temps).

En aval et au sein du système d'information, il faudrait assortir les informations initiales (input), intermédiaires et finales (output) d'un appareil de documentation et d'interprétation critique (via, par exemple un dictionnaire des données automatisé[36]). Nous avons défini les modalités de conception d'un tel dictionnaire. Quelles en sont les caractéristiques essentielles? Chaque donnée devrait être clairement insérée dans un schéma d'ensemble et assortie d'une série de définitions (domaine de définition initial et définitions successives se rapportant aux modifications de la donnée sous l'effet des processus de transformation interne et

---

En tant que telle, elle peut servir de support méthodologique à d'autres initiatives. B. CAMUS, T. FERRE, M. ROUSSET et M.-H. TAMISIER, *SUSE, système unifié de statistiques d'entreprises (sources, méthodes, apports).* Série E, n° 86, Paris, I.N.S.E.E., septembre 1983.

34. L'ouvrage suivant donne un excellent aperçu de l'état de la question en matière de gestion des données "incertaines" dans les bases de données: E. ZIMANYI et A. PIROTTE, *Imperfect knowledge in databases.* Louvain-La-Neuve, Unité d'informatique de l'U.C.L., Research Report RR 92-36, octobre 1992.

35. On trouve une brillante approche critique du sujet dans: H. L. DREYFUS, S.E. DREYFUS et T. ATHANASIOU, *Mind over machine. The power of human intuiton and expertise in the Era of the computer.* Oxford, Basil Blackwell, 1986.

36. La conception d'un tel dictionnaire rencontre plusieurs grandes préoccupations de la recherche fondamentale en matière de bases de données et notamment, la conception de "dictionnaires des données" appelés "repositories". L'originalité du dictionnaire évoqué dans le présent article réside toutefois dans sa fonction "d'interprétation critique". P. LOUCOPOULOS, *Conceptual Modeling dans Conceptuel modeling, databases and cases. An integrated view of information systems development.* Edited by Pericles LOUCOPOULOS and Roberto ZICARI, N. JOHN WILEY and sons, inc., New York, 1992, p. 15.

externe) et d'un appareil d'interprétation critique (évaluation de la fiabilité de la donnée). Le dictionnaire des données devrait en outre être doté d'un historique. En effet, dans le domaine administratif, la signification d'une même donnée peut varier dans le temps, au gré des modifications législatives.

5.2.2. *Règles de la critique historique à appliquer lors de l'exploitation des informations issues de banques de données à des fins scientifiques*

Les historiens qui ont recours aux sources informatiques reçoivent généralement très peu d'informations, d'ordre descriptif ou méthodologique, relatives au fichier de données qu'ils vont exploiter. Ils se trouvent dès lors dans une situation comparable à celle du médiéviste qui doit établir, critiquer et interpréter un manuscrit censément ancien dont, a priori, il ne sait rien. Mais nous pourrions presque affirmer que la comparaison s'arrête là. En effet, si l'usage de la critique se révèle indispensable dans le domaine de l'exploitation des banques de données, les règles traditionnelles de la critique historique sont mal adaptées au cadre de référence informatique. Illustrons ce propos en citant trois caractéristiques propres aux systèmes informatiques. Premièrement, la notion de donnée informatique, telle que nous l'avons définie dans cet article, est très éloignée de celle de document ou de témoignage, traditionnellement envisagée en histoire.[37] En second lieu, contrairement à l'historien qui peut effectuer de longues recherches afin d'établir, de critiquer et d'interpréter un seul manuscrit, nous nous trouvons ici face à une information de masse: certaines banques de données administratives répertorient en effet des dizaines de millions d'enregistrements. A cela s'ajoute une troisième spécificité liée à la constante évolution, dans le temps et dans l'espace, que subissent ces grands ensembles d'information.

Tenant compte des contraintes nouvelles qui viennent d'être évoquées, une méthodologie spécifique en matière de critique des données informatiques peut être établie. Un important travail d'analyse critique a été mené dans ce sens par M. Laffut et C. Ruyters dans le cadre d'une recherche relative à l'examen du travail intérimaire en Région wallonne. Cette étude a en effet nécessité l'exploitation, à des fins scientifiques, de la banque de données des T-Services de la Région.[38] Nous ne pouvons

---

37. P. HARSIN, *Op. Cit.* p. 22 à 24.

38. Voir la partie III: *"les données des T-services"*, et plus spécifiquement, le chapitre II: *"analyse critique de la base de données"* dans M. LAFFUT et C. RUYTERS, *Le travail intérimaire: impasse ou transition vers un emploi stable?* Convention de

développer ici l'ensemble des étapes particulières et des subtilités qu'implique une telle analyse. Nous nous limiterons à l'exposé des orientations générales de la méthode.[39]

Rappelons que les historiens distinguent la critique externe (ou de provenance), dont l'objet consiste à étudier un document indépendamment de son contenu et la critique interne (ou de crédibilité), proche de l'herméneutique, qui consiste à interpréter le document. Ces deux démarches, quoique distinctes, sont dans la pratique intimement liées. Nous nous inspirons ici de cette double approche.

D'emblée, une règle générale, qui concerne l'ensemble de la démarche critique, doit être formulée. L'analyse des données informatiques exige notamment la mise en oeuvre de tests de forme, de validité et de cohérence. Lorsque ceux-ci mettent en lumière la présence d'anomalies, le chercheur ne doit jamais corriger directement le fichier concerné sans conserver une copie de la situation précédente. Nous avons montré en effet que la présence d'incohérences au sein d'une banque de données pouvait découler d'une situation exceptionnelle, non prévue par le programme, mais bien réelle. La démarche idéale dans ce cas consiste à retourner aux sources (lorsqu'elles sont accessibles). Rencontrer les informaticiens et les responsables de l'administration chargés de la gestion de la banque de données ou encore les individus qui ont communiqué les informations posant problème permet d'acquérir un ensemble unique et précieux de renseignements relatifs à la qualité des données.

Critique externe:

Concernant l'aspect physique de la source, une série de tests informatiques, relatifs au format des données (longueur, type, format d'édition, ...) doivent être mis en oeuvre. La donnée, en tant qu'information codifiée est-elle valide? L'ensemble des données d'un fichier forme-t-il un tout cohérent quant à son format?

Concernant la provenance de la source, l'acquisition de renseignements précis passe par un travail heuristique rigoureux, tel que nous l'avons défini au seuil de cette étude (établissement de l'historique, du diagramme des flux de l'information et du schéma conceptuel). La critique

---

recherche entre le Ministère de l'Emploi de la Région wallonne et le Service d'histoire quantitative et de développement de l'Université de Liège, rapport final, janvier 1993, p. 105 à 182.

39. Nous préparons un ouvrage approfondi relatif aux modalités d'application de la critique à l'analyse des données issues de systèmes informatiques.

externe traditionnelle envisage les points suivants: date, origine, auteur. Transposés à l'étude des sources informatique, ont-ils encore un sens?
- *identification de la date:* nous avons évoqué l'importance que revêt la notion de temps dans une banque de données. Conformément aux concepts que nous avons définis, l'exploitation d'un fichier à des fins scientifiques est une opération synchronique. Le chercheur dispose dès lors d'une image figée et partielle d'une banque de données au sein de laquelle les informations sont sujettes à une évolution constante. Dans le domaine administratif, connaître la date de création des données que l'on consulte prend tout son sens. En effet, suite aux fréquentes modifications législatives que nous avons mentionnées, la signification d'une donnée dont l'intitulé reste identique peut varier dans le temps.
- *identification de l'origine:* dans le même ordre d'idée, nous avons montré que la définition d'une donnée pouvait varier en fonction de l'institution chargée de la gérer. Par exemple, si l'on interroge plusieurs banques de données afin de connaître le nombre total d'heures de travail prestées par telle catégorie de travailleurs pendant un intervalle de temps défini, le résultat ne sera pas le même d'un régime de sécurité sociale à l'autre. Il est donc essentiel de connaître l'origine des données consultées.
- *identification de l'auteur:* nous avons observé qu'en informatique, la création d'une donnée était le résultat d'une multitude d'interventions. Résumons les étapes définies dans notre étude: juristes et politiciens définissent un concept, celui-ci est interprété par l'administration et codifié par les informaticiens. La valeur de la donnée est ensuite fournie par les instances extérieures concernées, testée et éventuellement transformée et corrigée tour à tour par des informaticiens et des responsables de l'administration. Vu le nombre des intervenants, parler d'auteur n'a plus beaucoup de sens. Cela dit, nous avons constaté que le taux d'erreur observé dans certaines plages d'une banque de données variait fortement en fonction de la provenance des informations. Il est donc utile de rassembler des renseignements critiques relatifs aux groupes d'individus qui fournissent l'information et à la qualité des données transmises par ceux-ci.

Critique interne:

Concernant la cohérence interne des données, des tests de validité et de cohérence[40] permettent de contrôler la vraisemblance de l'information.[41] Ces tests doivent être conçus de façon à contrôler le domaine de définition de chaque donnée et les contraintes d'intégrité de l'ensemble de la banque de données. La richesse de la documentation relative à la banque de données dont dispose le chercheur détermine l'aisance avec laquelle il pourra programmer ces tests. Dans la plupart des cas, cette documentation est pauvre, voire inexistante. Il faut dès lors procéder par approximations successives, émettre des hypothèses relatives aux anomalies décelées et les vérifier. Idéalement, ces hypothèses devraient être soumises à l'épreuve d'une investigation sur le terrain: in fine, seules les personnes responsables de l'intégration des données sont susceptibles de fournir la signification précise de certaines informations.

Une confrontation avec d'autres sources informatiques répertoriant des données similaires peut s'avérer fructueuse à condition d'agir avec prudence. Nous avons montré qu'un même intitulé pouvait véhiculer des significations distinctes. Il faut dès lors s'assurer que les données que l'on compare sont effectivement comparables, ce qui implique une connaissance approfondie des pratiques informatiques et administratives relatives à chaque source.

La méthode critique que nous présentons se caractérise par la coexistence de tests informatiques et d'une investigation directe sur le terrain. Au terme des opérations, le chercheur acquiert une connaissance unique et originale relative à la banque de données. Il dispose en effet d'informations précises concernant la qualité des données et l'origine des erreurs décelées. Aussi, un dialogue constructif pourrait-il s'installer entre les gestionnaires des banques de données et les utilisateurs de celles-ci. Un tel échange faciliterait la conception et la diffusion de dictionnaires de documentation et d'interprétation critique des données que nous évoquions plus haut. Il est en effet essentiel d'éviter que se creuse davantage le

---

40. La définition de ces concepts figure dans le chapitre I du présent article.

41. A propos d'un examen du Registre de population, Mme S. Pasleau a énuméré en 1988 un certain nombre de règles élémentaires concernant ce type de test. Contrairement à notre démarche, son approche est principalement centrée sur la mise en oeuvre de procédures automatiques. Voir: S. PASLEAU, *Tests de cohérence et de validité des données informatisées. Essai de critique interne automatisée des documents historiques* dans *Standardisation et échange des bases de données historiques.* Paris, CNRS, 1988, p. 165 à 173.

gouffre entre les potentialités technologiques croissantes de l'informatique et la difficulté non moins croissante d'accéder à une information réellement pourvue de sens.

Informatica en kwaliteit van de informatie
Toepassing van de historische kritiek op de studie van informatie geleverd door databanken

DOOR
ISABELLE BOYDENS

Samenvatting

De historici die zich toeleggen op de studie van de nieuwste geschiedenis maken meer en meer gebruik van informaticabronnen, onder andere van administratieve databanken. Die laatste vormen echter zelden een homogeen en coherent geheel. Bijgevolg zijn de informaties die voortkomen uit administratieve gegevensbanken pas echt betekenisvol, wanneer ze vergezeld worden van een apparaat tot kritische interpretatie.

Teneinde zo'n kritisch apparaat tot stand te brengen, blijven de geest en de werkwijze van de historische kritiek uiterst afdoend. Niettemin dienen nieuwe regels, specifiek aangepast aan de studie van gegevens voortgebracht binnen een informaticasysteem, te worden aangewend. Dit artikel is de vrucht van een omstandige kritische studie van de informatiestromen die binnen een Belgische administratieve gegevensbank worden verwerkt.

Het artikel behelst vijf delen. In de inleiding stellen we het vertrekpunt van onze studie, de methode en de bronnen voor. In het eerste hoofdstuk worden de kenmerken van een informaticasysteem kort gedefinieerd. In het tweede hoofdstuk worden de resultaten van een origineel heuristisch werk met betrekking tot het geheel van het informaticasysteem (historiek, diagram van de informatiestromen en conceptueel schema) voorgesteld. Het derde hoofdstuk omvat de eigenlijke kritische analyse: die baseert zich op het concipiëren en op de inwerkingstelling van een kritische analysemethodologie aangepast aan de studie van de door een computer behandelde informatie. De ontwikkelde methode laat enerzijds toe om de mechanismen, zowel endogene als exogene, in staat om incoherenties te veroorzaken binnen een gegevensbank, te ontdekken. Anderzijds maakt

ze het mogelijk een diepgaande overweging van epistemologische aard met betrekking tot het informatica-referentiekader aan te vangen. In het besluit geven we een geheel van concrete voorstellen op teneinde de kwaliteit van de geïnformatiseerde gegevens te verbeteren. Die voorstellen zijn gericht enerzijds tot de beheerders van informaticasystemen, anderzijds tot de historici die gebruik maken van administratieve databanken met wetenschappelijke doeleinden.

Computing and quality of information
Historical critic applied to the study of information from databases

BY
ISABELLE BOYDENS

Summary

Historians occupied with the study of contemporary history more and more have to rely upon computer sources and more specifically administrative databases. The latter rarely make out a homogeneous and coherent collection. Therefore data from databases can only be fully fruitful when accompanied by a method for critical interpretation.

In order to establish such a method, the spirit and the ideas of the historical critic remain extremely pertinent. New rules however, adapted for the study of computer data, have yet to be worked out. This article is the result of a huge critical study of currents of information processed by a Belgian administrative database.

The article consists of five parts. In the introduction, the point of departure of our study, its method and sources are being presented. The first chapter defines in short the characteristics of a computer system. The second chapter presents the results of an original heuristic work covering the whole of the computer system (history, diagrams of information currents and conceptual schedule). The third chapter is the critical analysis itself: it is based upon the conception and creation of a critical methodology of analysis adapted to the study of computer processed information. This method allows us on the one hand to reveal the mechanisms – endogeneous and exogeneous – that are bound to bring forth incoherences in the database. On the other hand, it allows to set up a profound epistomological reflexion on computer issues. In our conclusion we present a collection of concrete proposals for improving the quality of computer data. These proposals are adressed both to managers

of computer systems and to historians who wish to explore administrative databases in their scientific work.