La conservation des bases de données

Marc LEBEL
M.b.a. Chargé de cours, Chef de la Section de la gestion de documents – Ville de Montréal

Résumé :

La pérennité des informations numériques préoccupe les archivistes. Elle pose des défis au niveau du stockage et la restitution des documents électroniques afin d'en assurer la conservation et l'intégrité. Plusieurs groupes de recherches s'intéressent de façon pertinente à ces questions. Il me semble que la problématique de la conservation permanente des bases de données n'obtient pas toute l'attention qu'elle mérite compte tenu de l'importance des bases de données dans les administrations.

Cette communication expose la compréhension et les préoccupations d'un praticien face à la problématique de la conservation des bases de données. Il aborde l'importance des bases de données dans les organisations, les pratiques actuelles et quelques considérations pour la conservation permanente des bases de données. Finalement, il envisage, comme une possible solution, l'utilisation des entrepôts de données. Bien modestement, cette communication peut dégager des pistes de recherches afin de proposer des solutions à cette problématique.

Définition

Pour ce texte, une définition restrictive d'une base de données est retenue.

- «… a database as a structured collection of data items stored, controlled, and accessed through a computer based on predefined relationships between predefined types of data item related to a specific business, situation, or problem. » (Alter, p. 512)
- « paper memos in a file cabinet are not a database because they are not accessed through a computer. Similarly, the entire World Wide Web is not a database of this type because it lacks predefined relationships between predefined types of data item » (Alter, p. 116).

L'importance des bases de données

Les bases de données sont au cœur des opérations des organisations. Elles supportent des activités comme la gestion des commandes, les relations avec les clients, la planification de la production, la gestion du personnel. À titre d'exemple, les grandes sociétés de services publics gèrent tous leurs dossiers clients grâce à des bases de données ; les documents sur support papier

deviennent accessoires. Toutes les informations opérationnelles vitales pour les organisations sont consignées dans les bases de données.

Les bases de données sont essentielles à l'automatisation des processus d'affaires, du commerce électronique, etc. Utilisées de façon intensive depuis le début des années soixante, la problématique de leur conservation se pose avec acuité.

Bien qu'omniprésentes, les bases de données seront davantage utilisées aux cours des prochaines années. Entre autres, les investissements massifs [1] pour l'implantation de systèmes intégrés de gestion (SIG ou ERP en anglais) et pour les systèmes de gestion des relations avec les clients, démontrent l'importance toujours croissante des bases de données.

Les pratiques actuelles

Peu d'organisations disposent de programmes structurés pour l'épuration de leurs bases de données et pour leur conservation à long terme. Cette situation peut s'expliquer, en partie par les coûts associés à l'épuration et la conservation des bases de données. En l'absence d'épuration des bases de données, des délais supplémentaires sont nécessaires pour :

- repérer un champ dans la base de données ;
- compléter une transaction ;
- effectuer une copie de sauvegarde ;
- accomplir une opération de maintenance.

Dans certains cas, les législations obligent l'épuration des renseignements personnels des bases de données lorsque « l'objet pour lequel un renseignement nominatif a été recueilli est accompli, l'organisme public doit le détruire » (Loi sur les archives).

Il existe plusieurs termes pour désigner l'épuration d'une base de données et la conservation d'une base de données. Quatre pratiques sont courantes mais ne constituent pas un archivage en sens archivistique du terme. La copie de sauvegarde (*backup*) est une duplication de la base, afin de la récupérer en cas de désastre ou d'altération de la base de données. Elle ne respecte aucun critère visant le recouvrement des données à long terme. L'épuration des données représente la destruction des données sans l'intention de les récupérer. L'entrepôt de données (*data wharehouse*) est le transfert d'une partie des informations vers un système distinct (l'entrepôt). Nous discuterons plus

[1.] Mentionnons la Société des Alcools du Québec et son projet VSOP de 95 millions de dollars, la Ville de Montréal et ses investissements de 50 millions, Hydro-Québec pour 42 millions. Au cours des dernières années, des firmes privées, telles Alcan, Pratt & Whitney, Bombarder, ont procédé à l'implantation de tels systèmes.

loin de cette opération qui peut être utile pour la conservation permanente des données. Finalement, les données sont transférées dans un système statique. Cette dernière solution mérite une attention particulière puisqu'elle est largement répandue.

Le transfert des données vers un système statique prend plusieurs formes : l'impression des données sur papier, la production de microfiche (fiche SOM) ou le transfert des données dans des fichiers plats. Dans le cas des bases de données complexes, seulement une partie des données sera reproduite puisqu'il est difficile de reproduire toutes les données dans une seule liste. Plus déplorable, ces systèmes statiques ne représentent pas les relations entre les différentes données, ce qui empêche une bonne compréhension du contexte d'utilisation des données. Les systèmes statiques réduisent considérablement le potentiel de recherche. Dans le cas des impressions sur papier et des microfiches, certaines recherches deviennent pratiquement impossibles. À titre d'exemple, pensons à une recherche dans les données d'un recensement comprenant des millions d'éléments. Le chercheur devra lire chaque enregistrement. Théoriquement, il existe des méthodes pour recharger ces informations dans une base de données. En pratique, les coûts d'une telle opération sont prohibitifs donc inabordables pour la majorité des projets de recherche. Malgré ces limites, le transfert des données vers des systèmes statiques est une pratique courante. Faute de moyens et d'alternatives viables, il s'agit de la principale stratégie d'archivage des données à la Ville de Montréal.

La valeur historique des bases de données

La valeur archivistique des bases de données ne fait pas de doutes. Des informations qui, auparavant, étaient sur support papier et conservées en permanence, ne sont maintenant créées et consignées que dans des bases de données. Par exemple à la Ville de Montréal, le nombre de documents papier conservés dans les dossiers d'employés a diminué de 80 % depuis 1980. Cette réduction s'explique essentiellement par l'utilisation des bases de données pour gérer le personnel. Un échantillonnage des dossiers d'employés était conservé en permanence. Dorénavant, il faudra donc conserver le dossier physique et les bases de données dédiées à la gestion de personnel. L'information, peu importe son support, conserve sa valeur de témoignage.

La structure des bases doit être préservée à des fins de reconstitution des pratiques de travail et comme témoignage de l'utilisation des technologies de l'information dans les organisations. Les composantes comme les tables, les champs, index, les relations entre les tables et les rapports produits, doivent donc être conservées. Toute la documentation nécessaire à la compréhension de la

base de données doit être également préservée ; cette dernière est souvent sous forme électronique et intégrée au logiciel servant au développement de la base de données.

L'évaluation des bases de données

Comme pour tout document, l'archiviste devra faire l'évaluation des bases de données. Toutes les bases de données ne méritent pas une conservation permanente.

L'évaluation devra également préciser le moment et la fréquence de versement à des fins de conservation permanente. Contrairement aux documents sur support papier, il est impossible d'identifier une version définitive d'un document puisque le contenu des bases de données est constamment modifié.

La description des bases de données

Au premier niveau, une base de données peut être décrite selon les règles de description des documents d'archives en vigueur. Cependant, la description d'une base de données demande davantage de précision puisque chaque élément (table, champ) d'une base de donnée doit être décrit. Sans description, le champ « N^{o} de téléphone » n'est pas significatif. À cette fin, la norme ISO/IEC 11179 « Technologies de l'information – Spécification et normalisation des éléments de données » doit être retenue. Cette norme devra être appliquée dès la conception des bases de données pour uniformiser leur description à toutes les étapes de leur vie.

Le format et la technologie de conservation des bases de données

La question du format de conservation des bases de données demeure entière et complexe. Le développement de solutions économiques est essentiel pour que les organisations mettent en place des programmes de conservation.

Les grandes institutions d'archives ont développé des modèles et des applications informatiques pour la conservation des bases de données. Mentionnons, entre autres, Constance des Archives Nationales de France, AERIC de la U.S. National Archives and Records Administration, Ericson à la Bibliothèque et Archives Canada. Les solutions choisies par ces initiatives peuvent être regroupées en quatre grandes catégories :

- préservation de la technologie originale pour conserver les informations ;
- émuler la technologie originale sur de nouvelle plate-forme ;
- migrer les logiciels et les données vers les formats plus récents ;
- convertir les informations vers des formats plus standards.

À première vue, et des recherches pourraient le vérifier, la conversion des informations vers des formats plus standards semble l'alternative la plus économique. Les coûts seraient minimisés par :

- l'entretien d'une seule plate-forme technologique ;
- la diffusion des bases de données archivées (par internet ou autrement) à partir d'un seul format ;
- la conversion des données vers un seul format permettant le développement d'une expertise en la matière ;
- la migration pour suivre les changements technologiques d'une seule plate-forme et d'un seul format.

Un consensus dans le milieu archivistique à propos du format standard permettrait de maximiser ces avantages.

Les normes de conservation

Les normes de conservation des données devraient s'inspirer de celles des entrepôts de données, qui regroupent l'ensemble des données d'une organisation dans un seul système. Leur élaboration ne découle pas de besoins archivistiques mais des demandes des gestionnaires pour effectuer des recherches croisées. Les entrepôts de données sont également utilisés pour une conservation intermédiaire après l'épuration des bases actives. Dans un entrepôt, les données ont les caractéristiques suivantes, qui répondent aux préoccupations des archivistes :

- elles sont intégrées ou normalisées. Par exemple, la longueur du champ « prénom » aura toujours la même forme ;
- elles sont datées et non-modifiables ;
- elles ne doivent pas être volatiles afin de pouvoir répéter les recherches.

De plus, les données, les informations peuvent être détaillées ou bien agrégées. Ce choix découlera de l'évaluation archivistique.

Grâce à la centralisation des données, des recherches croisées sont possibles. Par exemple, il sera possible de faire une corrélation entre l'augmentation du budget dédié au marketing et les résultats des ventes.

L'intervention de l'archiviste

Si l'entrepôt de données est retenu pour la conservation des bases de données, l'archiviste doit intervenir dès leur création au stade actif. Une intervention tardive entraînerait des coûts importants si des modifications sont demandées et des informations pourraient être irrémédiablement détruites. Idéalement, l'entrepôt de données utilisé à des fins administratives contiendrait toutes les informations à conserver en permanence. Au besoin, une épuration additionnelle serait effectuée lors du versement dans l'entrepôt définitif. Un parallèle avec la théorie des trois âges peut être établi :

- au stade actif, les données sont conservées dans les bases courantes ;
- au stade semi-actif, les données sont conservées dans les entrepôts de données administratives ;
- pour la conservation permanente, les données sont conservées dans les entrepôts de données définitives.

Conclusion

Les entrepôts de données semblent, à première vue, une solution prometteuse et économique pour la conservation permanente des bases de données pour les moyennes et grandes organisations, puisqu'il s'agit de récupérer des applications développées pour répondre aux besoins administratifs. Cette solution est probablement trop onéreuse pour des petites organisations où les entrepôts de données sont inexistants. Des moyens plus adaptés sont à prévoir.

Avant de généraliser cette pratique, il faut cependant vérifier si elle permet de conserver toutes les données utiles sans altérer leurs caractéristiques originales, ni perdre le contexte de leur création. Les praticiens et les chercheurs devront répondre à ces questions puisque la nécessité de conserver des bases de données est incontournable.

Bibliographie

ALTER, S., *Information systems, a management Perceptive*, Addison-wesley, Reading Mass, 1999, 523 p.

BOUSSAID, O., LALLICH, S., *Entreposage et fouille des données,* Toulouse, Cépaduès, 2003, 282 p.
Gouvernement du Québec, Loi sur les archives, L.R.Q., c. A-2.1.

MENSCHING, J., CORBITT, G., « ERP data archiving – a critical analyst », *The Journal of Enterprise Information Management*, Volume 17, Number 2, p. 131-141.