

Communiqué du GIS-DON du 29 août 2008

Le numérique : combien utile, mais combien éphémère aussi !

1. Merveilles du numérique

Les immenses commodités offertes par le stockage numérique de l'information sont bien connues de tous : densité de l'information permettant un gain de place considérable (toute une encyclopédie volumineuse tient sur un seul disque optique numérique), facilités de classement et de recherche de toutes sortes grâce à des algorithmes toujours plus performants, accès rapide et délocalisé, circulation quasi instantanée peu onéreuse par les réseaux (envoyer un fichier à distance ne coûte pratiquement rien), souplesse de ré-utilisation, etc. L'utilisation du numérique est presque devenue une obligation : il suffit de penser pour s'en convaincre aux frais considérables qu'entraînerait la construction des locaux nécessaires à la conservation traditionnelle sur papier. Il n'est donc pas étonnant que, depuis quelques années, la presque totalité des informations produites par notre société soient stockées sur des supports numériques. Ainsi, que ce soit pour les données scientifiques (résultats des mesures des satellites par exemple), techniques, médicales (en particulier les images d'examens), administratives (gestion des carrières parmi d'autres), financières (banques), artistiques, etc.. l'utilisation du numérique est omniprésente, le phénomène est maintenant irréversible.

Pour illustrer la puissance des méthodes numériques, le plus frappant est peut-être de réfléchir à ce que chacun d'entre nous fait quand il rédige un document sur un traitement de texte; au cours de son travail, combien de fois enregistre-t-il le fichier, effaçant ainsi la version précédente du fichier pour la remplacer par une complètement nouvelle – même si c'est juste pour changer une virgule ? Si nous le faisons tous sans y penser, c'est que l'opération ne présente aucun risque d'erreur. Et pourtant, si on y pense, c'est fantastique : aucune autre méthode que le numérique ne pourrait permettre d'effectuer des milliers de fois (ou plus) la recopie sans aucune erreur. Il suffit de penser aux textes anciens qui nous sont parvenus par le travail séculaire des moines copistes ; malgré tout le soin qu'on peut y mettre, il est humainement impossible de faire des copies strictement identiques. C'est également valable pour toute technique analogique : quelle que soit sa qualité, chaque recopie introduira un peu de « bruit » qui, petit à petit, s'accumulera exponentiellement de copie à copie ; au bout d'un certain nombre de recopies, l'information initiale est noyée et devient inexploitable.

Derrière cette merveille de la recopie sans erreur se cache le travail des mathématiciens très brillants, en particulier les américains Reed et Solomon (dont chacun d'entre nous utilise le travail tous les jours sans même connaître leur existence !), qui ont mis au point des codes de correction d'erreurs numériques permettant des millions de recopies d'un fichier. Mieux encore : si nous recopions un fichier légèrement endommagé, par exemple à partir d'un CD de données un peu ancien vers un autre neuf, tant que le taux d'erreur apparu n'est pas trop grand, la copie éliminera les erreurs qui sont

apparues entretemps ! Le fichier sera restauré dans son état initial ; le numérique semble avoir les vertus magiques du Phénix renaissant de ses cendres.

Certains feront remarquer que les disques durs sont l'objet de pannes soudaines et presque imprévisibles (crashes), qu'un incendie ou une inondation peuvent faire des ravages dans un centre informatique, qu'un vol de bandes magnétiques peut s'avérer catastrophique, etc. Tout cela est vrai mais n'est, ni spécifique du numérique, ni insurmontable, précisément à cause de la facilité de la recopie du numérique. Les prix modestes des supports de stockage font maintenant que chacun peut aisément faire de multiples sauvegardes, et de les placer dans des lieux indépendants, ce qui permet de régénérer l'information en cas de catastrophe. Des systèmes comme par exemple Google délocalisent totalement l'information dans le monde : les données d'un utilisateur peuvent être un peu partout, réparties entre des centres de stockage dans des pays divers, sans même qu'il ne s'en rende compte. Il existe maintenant de nombreux prestataires de service qui proposent d'assurer la pérennité de l'information qu'on leur confie, moyennant un coût qui n'est pas excessif. Donc, ce premier problème n'est pas dramatique, puisqu'un peu de bon sens permet de le surmonter – après tout, l'information sur support papier peut elle aussi brûler et être volée ! Jusqu'ici, le numérique pourrait donc apparaître comme la solution idéale à tous les problèmes de conservation de l'information. Mais nous n'avons parlé jusqu'ici que de la sécurité à court terme, qu'en est-il à long terme ?

2. Problèmes du numérique

Il n'est évidemment pas question de conserver à long terme toute l'information numérique qui est produite à chaque instant, tous les SMS ou les messages échangés par courrier électronique dans le monde par exemple ; une sélection s'impose. Mais il existe des données que l'humanité a un besoin essentiel de conserver pendant des décennies, voire pendant des siècles ; nous avons déjà mentionné plus haut les données scientifiques (satellites, accélérateurs de particules, etc.), les données administratives (dossiers de retraite par exemple), et les exemples sont variés et nombreux. Ce qui importe alors est que, pour ces informations sélectionnées, il soit effectivement possible de faire en sorte qu'elles soient conservées longtemps, un siècle par exemple. Ainsi les photos et vidéos seront transmises du grand-père au petit-fils, les informations médicales pourront suivre les patients toute leur vie, les données des satellites auront largement le temps d'être exploitées par les chercheurs.

Qu'en est-il donc du stockage numérique de l'information à long terme ? Hélas, un rapide examen révèle rapidement l'ampleur du problème : dans les faits, la conservation de l'information sur support numérique est très loin d'aller de soi, même pendant une durée relativement courte, dix ans par exemple. Il y a trois raisons principales à cela. Les deux premières sont relativement bien connues de tous, nombre d'entre nous y ayant déjà été confrontés ; la dernière l'est beaucoup moins, alors qu'elle est plutôt plus grave, ce qui la rend d'autant plus dangereuse. Commençons par les difficultés les plus bénignes.

La première faiblesse du numérique est l'obsolescence rapide des techniques d'enregistrement et de lecture, avec l'évolution constante des appareils qu'elle entraîne. Les exemples sont nombreux, par exemple le format DAT (Digital Audio Tape) qui, après avoir été très répandu dans l'enregistrement musical, a maintenant presque totalement disparu ; il est devenu difficile, parfois impossible, de relire des données enregistrées dans ce format, par manque de lecteur ou même tout simplement par impossibilité de faire réparer celui qu'on possède, auquel plus aucun fabricant ne s'intéresse. La seconde faiblesse, liée à la précédente, est l'évolution rapide des formats de fichier informatiques et des logiciels qui servent à les créer et les manipuler, sans parler des systèmes d'exploitation des ordinateurs, dans certains cas. Il est naturel que la pression commerciale encourage cette évolution, car il est évidemment intéressant pour les vendeurs de logiciels et de matériel de pousser les clients à renouveler fréquemment leurs licences et leurs appareils.

Ces deux difficultés sont bien réelles, mais pas insurmontables dès lors qu'on les a identifiées. En cas de besoin spécifique de conservation numérique à long terme, l'utilisateur privilégiera un support et un format les plus courants possible, augmentant ainsi les chances que le corpus de données correspondant stimule le maintien de la méthode. De ce point de vue, un support aussi courant et commode que le CD/DVD semblerait très bien placé, du fait de son utilisation très large. Son chiffre de production est tellement énorme¹ qu'il restera certainement très longtemps un marché suffisamment rentable pour que les fabricants maintiennent une production d'appareils de lecture². Quant aux formats de fichiers, certaines règles simples permettent de procéder raisonnablement : éviter les formats dits « propriétaires », favoriser les formats (comme certaines bases de données, LaTeX, etc.) qui se basent uniquement sur des fichiers texte (dont la migration est aisée et dont les chances de relectures à l'avenir sont fortes), etc.. Prenons l'exemple du format pdf : bien qu'effectivement « propriétaire », il est ouvert, et de plus il existe maintenant tant de documents dans le monde qui utilisent ce format qu'il semble inconcevable que leur relecture dans 50 ans soit impossible ; la nécessité créera la fonction. Le format Pdf/A, conçu dans cette optique et reconnu par l'ISO, a de bonnes chances d'être pérenne.

3. Le véritable problème

On pourrait donc penser que les deux problèmes précédents soient rapidement résolus par l'apparition de normes de conservation digitale ; là aussi, au fond, c'est une simple question de bon sens. Pourquoi alors cette évolution tant souhaitable ne se produit-elle pas ? Parce que, par derrière, existe un problème bien plus sérieux, qui bloque l'apparition des solutions aux deux premiers : les

¹ La production annuelle totale actuelle de disques enregistrables dans le monde dépasse 10 Milliards d'exemplaires.

² Dans le domaine de l'analogique, on constate en ce moment une résurgence de l'offre de lecteurs de disques « microsillons », ce qui illustre bien comment l'existence d'un grand nombre de supports d'information maintient le marché des lecteurs.

supports physiques eux-mêmes sur lesquels sont écrites les informations numériques se dégradent constamment dans le temps. C'est le problème de la stabilité physique des supports qui contient la clé de autres problèmes : si nous disposions de supports stables pour stocker l'information, cela créerait automatiquement un marché pour des logiciels et des matériels durables permettant leur relecture ; mais, tant que la base du système elle-même est instable, il n'y a aucune raison d'espérer voir émerger par miracle cette durabilité.

Les bandes magnétiques vieillissent en une dizaine d'années, parfois moins, et la seule façon de conserver leur information est de les recopier sur une bande plus neuve, etc. ad infinitum. C'est ce que font les grandes bibliothèques pour conserver leur données ; mais si la recopie est oubliée pendant un certain temps, tout est perdu. Compte-tenu des coûts d'une telle opération fastidieuse, sommes-nous sûrs que toutes les informations scientifiques récoltées par exemple par la NASA sont ainsi sauvegardées, même si beaucoup n'ont jamais été encore utilisées et peuvent un jour se révéler précieuses? Pour les disques durs, ils reposent sur un principe physique un peu analogue aux bandes magnétiques; malgré leurs avantages bien connus en termes de rapidité d'accès, pour la longévité ils ne font guère mieux. De fait, jamais dans toute son histoire l'humanité n'a utilisé de techniques aussi instables pour enregistrer ses données

Qu'en est-il des CD enregistrables ? N'avons-nous pas vu il y a quelques années une publicité dans Le Monde sur le thème « enregistrez sur CD-R vos photos pour l'éternité » ? Publicité mensongère s'il en est, car dans la réalité beaucoup de CD enregistrables ont une durée de vie de quelques années - on cite parfois une moyenne d'environ 5 ans, avec d'énormes fluctuations : certains CD-R deviennent illisibles en un an ou deux, d'autres dépassent dix ans³. Les disques optiques enregistrables s'effacent lentement, mais constamment, et ceci de façon aléatoire, même si on les garde à l'abri dans un tiroir sans les utiliser. Où est la longévité des photos de nos grands-parents sur papier ? Pour les DVD enregistrables, ce n'est pas mieux. Quant à la nouvelle génération des DVD haute définition (Blu-Ray), le fait que toute l'information soit stockée aussi près de la surface du disque, particulièrement vulnérable à toutes sortes d'agressions physiques et chimiques, montre qu'elle n'a été conçue particulièrement pour résister au temps. Sous la pression commerciale, la seule chose qui prime est la densité d'information ; comme toujours, la longévité a été oubliée.

Et pourtant, les faits sont là : tout patrimoine numérique abandonné à lui-même, ne fut-ce que 5 ou 10 ans, risque d'être définitivement perdu. Certes il est très louable de rétro-numériser des documents de toute nature, mais à quoi bon si le résultat de cet investissement risque de disparaître rapidement ? Le domaine médical est concerné : quand les médecins stockent des images d'exams importants, quand les dentistes font des radios dentaires afin de pouvoir suivre l'évolution de leurs patients à long terme, savent-ils que ces photos sont enregistrées sur des CD qui risquent de n'être plus lisibles dans quelques années ? Quant à tous ceux d'entre nous qui numérisent leurs souvenirs de

³ Nous ne parlons ici que des disques optiques enregistrables ; les disques « pressés », vendus tout enregistrés dans le commerce, sont conçus sur un principe physique différent et ont une durée de vie beaucoup plus grande.

famille dans l'espoir de les transmettre à leurs enfants et aux générations futures, pensons-nous que ce patrimoine familial se décompose un peu plus chaque année ? On peut craindre dans quelque temps de grandes difficultés dans bien des branches de l'activité humaine si le problème n'est pas pris à bras-le-corps pendant qu'il est encore temps. L'humanité perd la mémoire.

Rien sur le plan technique n'empêche pourtant d'imaginer des disques optiques numériques qui aient une bonne stabilité, pourvu qu'ils soient conçus dans cette optique. On peut même rêver de l'apparition d'un « standard longue durée » de support numérique optique dont, on peut le penser, les clients publics et privés seraient nombreux. Il faut juste pour cela un standard suffisamment ouvert et des prix raisonnables, ce qui n'est pas le cas des quelques tentatives isolées en ce sens (disques UDO par exemple). Les rares CD-R « archivage » proposés sur le marché n'ont pas fait l'objet de contrôles indépendants, et les premiers tests les concernant sont plutôt alarmants, même s'il est trop tôt pour conclure. Quoi qu'il en soit, non seulement ce standard n'existe pas pour le moment mais, pire, rien ne nous assure qu'il soit en train d'émerger.

4. Perspectives

Comment en sommes nous arrivés à une telle régression, par comparaison avec les parchemins du moyen-âge, ou les tablettes cunéiformes assyriennes qui ont traversé des dizaines de siècles? C'est que notre civilisation de consommation n'est pas bien armée pour traiter ce type de problème. Ce qui paraît vendeur aux fabricants de supports d'enregistrement, de graveurs, et lecteurs, semble avoir été jusqu'à maintenant uniquement la capacité de stockage ainsi que la vitesse d'enregistrement.

Certes, une solution pour contourner le problème serait de créer dans chaque administration, chaque hôpital, chaque entreprise, etc. des services spécialisés uniquement dans la sauvegarde et surtout dans une constante surveillance des données, avec migration régulière sur des supports neufs. Certains organismes, peu nombreux, ont de tels services. Pour les particuliers, ils pourraient s'adresser à des sociétés commerciales spécialisées. Pourquoi pas ? Mais, à l'échelon national, un calcul simple montre qu'il faudrait une bonne dizaine de milliers de « curateurs numériques » pour faire ce travail de Sisyphe, voire plus. On imagine facilement les dépenses correspondantes, en salaires et en matériel ! Ne serait-il pas plus simple et économique de produire des supports enregistrables de bonne longévité, que chacun puisse enregistrer et conserver à sa guise pendant des décennies ou plus ?

Les forces du marché paraissent à elles seules incapables de développer un support d'enregistrement qui prenne en compte la nécessité impérieuse d'une conservation à long terme, par exemple comparable à celle d'un document écrit sur papier (quelques siècles). La seule chance de progresser dans ce domaine semble donc être que les pouvoirs publics prennent l'initiative de stimuler une action. Le sujet n'est pas particulièrement difficile sur le plan technique, surtout si on le compare à d'autres défis technologiques sociétaux actuels. Les quelques laboratoires publics qui s'intéressent au problème se heurtent au fait qu'il n'est pas considéré comme prioritaire. Si l'on veut que les forces du

public et du privé s'allient enfin sur ce dossier urgent, le seul espoir semble être une prise de conscience du grand public. Il faut que chacun comprenne les enjeux, et qu'on n'attende pas 10 ou 20 ans pour constater que les témoignages d'une génération ont disparu, alors que ceux des générations précédentes existent toujours. Il faut également que les chercheurs eux-mêmes réalisent les enjeux, eux qui accumulent tant de données qu'il convient de conserver, par exemple celles des grands accélérateurs de particules dont le prix est énorme et que personne n'entend perdre. Espérons que le problème ne sera pas mis sous le boisseau pendant plus longtemps !

Anecdotes

1. Une astrophysicienne de l'Observatoire de Meudon a fait autrefois un travail théorique basé sur des données scientifiques fournies par un satellite de la NASA, qu'elle a publié dans une grande revue. Plusieurs années plus tard apparaissent de nouvelles idées théoriques ; elle désire reprendre son travail d'analyse pour en tenir compte. Elle demande donc à nouveau à la NASA de lui fournir les données brutes. Réponse de la NASA : nous sommes désolés, les bandes sur lesquelles les données de ce satellite étaient enregistrées ne sont plus lisibles.

2. Un patient consulte son ophtalmologiste qui, comme chaque année, lui fait faire une photographie du fond de l'œil. Il lui demande si le but est de suivre l'évolution de ces photos sur le long terme, afin de déceler des évolutions lentes possibles de la rétine ; réponse affirmative du praticien. Le patient demande alors : ainsi, pour moi, vous avez les photos depuis dix ans ? Réponse : « justement, la semaine dernière, je me suis aperçu que nous ne pouvions plus relire les CD sur lesquels sont enregistrés ces photos, il faudra que je demande à notre informaticien pourquoi... ».

3. Autre exemple scientifique, cette fois. A l'observatoire astronomique de Tenerife sont stockées de nombreuses photos argentiques prises au cours des années. Depuis un certain nombre d'années, c'est le stockage numérique qui est utilisé : alors que toutes les photos argentiques sont toujours disponibles, les photos numériques les plus anciennes ne sont plus lisibles.

4. Un père de famille a hérité de son propre père des films de cinéma pris pendant son enfance et sa jeunesse. Il les fait numériser pour pouvoir les distribuer sur des disques optiques à tous ses enfants, alimentant ainsi la mémoire familiale. C'est une opération délicate et relativement chère, mais qui se justifie dans la perspective d'une facile distribution de ces souvenirs et de leur conservation à long terme. Mais, si personne ne se soucie de suivre l'état de ces disques optiques, ce patrimoine sera perdu dans dix ans, alors que les photos jaunies transmises par des générations plus anciennes sont toujours utilisables !

Franck Laloë, Président du conseil scientifique

GIS-DON : Groupement d'intérêt scientifique sur les disques optiques numériques

Adresse internet : http://www.lne.fr/fr/r_et_d/gis-don/conservation-donnees-numeriques-gis-don.asp ; Contact : gisdon@lne.fr