Bulletin des Archives de France sur la conservation à long terme des documents électroniques
N° 7 octobre 2001

La pérennité des documents électroniques - points de vue alarmistes ou réalistes ?

(par Claude Huc, Centre National d'Ėtudes Spatiales, Toulouse)

L'article de Jean-Luc Philip, publié dans le n° 6 du présent bulletin, nous propose dans sa conclusion de limiter l'usage des technologies numériques à la communication et à la diffusion de l'information et de ne pas envisager dans l'immédiat, d'utiliser cette technologie pour assurer la pérennité desdites informations. En d'autres termes, pour l'instant, restons prudents et poursuivons l'archivage sur support papier en attendant que la situation se clarifie.

Je souhaite par rapport à l'argumentaire proposé, soumettre au lecteur un autre éclairage qui fait apparaître des points de convergence mais aussi des différences d'appréciation réelles.

L'Étendue du problème

Le problème posé par la pérennité des informations sous forme numérique concerne tous les types d'informations qui peuvent être représentées sous cette forme, notamment :

Les documents textuels, qui peuvent être simples ou complexes, comporter des mises en page élaborées, des tableaux, des notes, des alphabets particuliers, des formules mathématiques. Notons ici que certains de ces documents (par exemple les fichiers HTML accessibles via le réseau internet) ignorent la notion même de page au sens page papier. Imprimer ces documents revient à les dénaturer,

Les plans, les graphiques, les images,

Le son,

La vidéo, le cinéma,

Les documents multimédias constitués d'une imbrication organisée des catégories précédentes (indiquons au passage que la question de la préservation des logiciels qui pose des problèmes spécifiques est en dehors du champ de cet article),

Les observations scientifiques issues des sondes spatiales, des observatoires,

Etc.

Il est utile d'insister sur le fait que certaines de ces catégories d'information n'existent et ne peuvent exister que sous forme numérique, soit parce qu'on ne connaît pas d'autre moyen fiable de les représenter, soit parce que les modes de représentations alternatifs seraient incroyablement coûteux.

La conséquence de ce premier constat est que la question de la pérennité des informations sous forme numérique doit être résolue, et ceci à court terme, sous peine de perdre une partie importante de notre patrimoine.

Un contexte mondial

Le problème est posé au niveau planétaire, il concerne toutes les nations du monde et en premier lieu les plus développées et les plus riches. Il touche une grande variété d'entreprises et

d'établissements publics ou privés :

Les institutions patrimoniales, archives nationales, bibliothèques, musées,

Les centres de recherche scientifique : à titre d'exemple, l'étude et la prévision des évolutions climatiques de la terre à moyen et long terme exigent de gigantesques bases de données qui doivent être préservées plusieurs dizaines et certainement plusieurs centaines d'années,

Les compagnies pétrolières devront, pendant encore un siècle au moins, préserver l'immense capital de données géologiques qu'elles ont constitué à grands frais.

Ce contexte global constitue finalement une chance pour tous ceux qui sont confrontés au problème et en particulier pour les Archives de France. Nous allons en effet tous pouvoir bénéficier dans ce cadre du très large effort international en cours dans ce domaine.

La question des supports de stockage masque souvent les vraies difficultés

Il existe une tendance largement répandue à centrer la question de l'archivage des informations numériques sur le choix des supports utilisés pour stocker l'information. Cette tendance est la source de multiples incompréhensions. Elle nuit souvent à la perception globale du problème à résoudre.

Quelques constats préalables peuvent être faits : les différentes générations de technologies de stockage se succèdent rapidement. Cette évolution a pour corollaire une augmentation constante de la capacité des supports et une réduction du coût de stockage par mégaoctet dans un contexte où la quantité d'information à stocker ne cesse de croître (attention, le coût de l'archivage est loin de se réduire au coût du stockage physique). A titre indicatif, la capacité des supports de stockage est passée de 30 mégaoctets (la bande magnétique 1600 bpi en 1975) à plus de 30 gigaoctets actuellement (des cartouches magnétiques à 100 gigaoctets sont déjà commercialisées, des cartouches de 200 gigaoctets sont à l'étude) soit un facteur 1000. On a même vu apparaître des technologies de bandes optiques ayant une capacité de l'ordre du téraoctet.

Dans ce cadre, la durée de vie supposée des supports n'a qu'une importance relative. Un support en parfait état devient pratiquement inutilisable s'il est issu d'une technologie obsolète. Le CNES a dû, à la fin des années 1990, migrer toutes les données spatiales stockées sur bandes magnétiques vers d'autres supports. Des milliers de bandes, enregistrées entre 1990 et 1992 auraient pu être conservées 10 à 20 ans de plus. La technologie des bandes magnétiques étant en voie de disparition, les coûts de maintenance des matériels de lecture se sont mis à croître vertigineusement, ce qui a conduit à la décision de migration. Même si aujourd'hui il est possible de créer des Compact Disc en verre d'une durée de vie d'un siècle, on peut légitimement penser que dans un siècle, ce CD relèvera d'une technologie totalement archaïque et aura une capacité de stockage ridiculement faible.

Le stockage des informations numériques doit donc se faire dans le cadre d'une véritable stratégie de préservation mettant en application un ensemble de règles visant à garantir l'intégrité de l'information dans un contexte technologique changeant : duplication de l'information sur plusieurs supports, voire sur plusieurs supports de types différents, stockage en des lieux distincts, relectures systématiques, surveillance de l'état des supports (les codes correcteurs d'erreurs sont là pour cela), renouvellement préventif à une périodicité définie (5 ans au CNES), gestion des migrations...

Dans une telle approche, on peut considérer, preuves à l'appui, qu'il est aujourd'hui possible - à condition de s'en donner les moyens techniques, financiers et organisationnels - de stocker sans perte d'information des volumes de données numériques considérables de plusieurs dizaines ou centaines de téraoctets.

A titre indicatif pour ce qui concerne le texte : une page de texte mise sous forme numérique peut nécessiter environ 10 kiloctets. Un document de 100 pages (épaisseur d'un centimètre) occupe 1 mégaoctet. Par conséquent, un téraoctet permet de stocker 1 million de documents de 100 pages qui correspondent, sous forme papier, à 10 km linéaires. Quel est le coût de construction et de maintenance des locaux de stockage correspondants ?

On doit cependant ici souligner une difficulté sérieuse : autant les grandes institutions comme les Archives de France, la BnF, les grands organismes de recherche ou d'autres ont ou auront les moyens financiers et sauront trouver les compétences pour mettre en oeuvre une telle stratégie de préservation pour le stockage, autant, pour l'instant, cette approche n'est pas à la portée des personnes privées et des institutions de petite taille.

L'analyse, la maÎtrise et la gestion des risques

La technologie du numérique nous offre des moyens pour stocker et pour exploiter des quantités extraordinaires d'informations. Elle présente aussi un certain nombre de risques majeurs. Il ne s'agit pas de les occulter, mais il ne convient pas non plus de rejeter la technologie parce que ces risques existent.

Les archives nationales actuelles sous forme papier courent de toute évidence un certain nombre de risques parmi lesquels le feu, la dégradation chimique du papier, les inondations, les imprévisibles... Il n'en reste pas moins que ces risques ont été réduits à un niveau satisfaisant par un ensemble de moyens : conception même des locaux de stockage, dispositifs de surveillance, réglementation, procédures opératoires, formation du personnel...

Il en est de même pour le fonctionnement très complexe d'un avion de ligne qui a une durée de vie proche de 50 ans. Sur une telle période, un très grand nombre de pièces sont surveillées, démontées, remplacées. Un ensemble de moyens (conception de l'avion, mise en place d'autorités internationales, certifications, contrôles, réglementation relative à l'entretien…) confèrent au transport aérien un niveau de sécurité raisonnable pour l'usage que nous en faisons.

Il doit en être de même pour le domaine numérique. Donnons ici quelques exemples de difficultés vécues :

Dégradation de supports physiques : nous avons rencontré un certain nombre de cas de dégradation des bandes magnétiques dus soit au vieillissement (surtout si la bande n'a pas été relue depuis un certain nombre d'années), soit à des défauts de fabrication. Ces dégradations ont entraîné quelques pertes de données. La stratégie de préservation définie plus haut n'était pas encore en place. Nous pensons aujourd'hui qu'elle est à même de réduire complètement le risque de perte d'information liée à cette cause.

Non disponibilité du descriptif des données : Si vous disposez d'un texte en ASCII et que la table ASCII qui vous donne les correspondances entre les bits et les caractères alphabétiques est perdue, alors votre texte est indéchiffrable. Il en va de même pour les fichiers contenant des observations

scientifiques. À chaque fichier ou à chaque collection de fichiers de même type est associé un document (appelé descriptif de fichier) qui décrit chaque champ d'information présent dans le fichier en termes de position (numéro du bit de début, longueur), de type (nombre entier, réel…), de codage (comment passer des bits contenant ce champ à la valeur du nombre), de signification (ce champ contient la température de la surface de l'océan…), d'unité physique (en degrés Celsius…). Il est donc arrivé dans le passé que parce que le descriptif des données avait été perdu, une collection complète de données devienne totalement inutilisable. La perte de ce descriptif a donc des conséquences beaucoup plus radicales que la dégradation de quelques supports.

Nous avons également rencontré un certain nombre d'autres cas qu'il est inutile de détailler ici.

La connaissance et l'analyse des risques encourus sont l'une des conditions indispensables à l'élaboration de moyens et de pratiques susceptibles de les réduire ou de les éliminer. Il est clair que dans ce processus, l'expérience pratique joue et jouera un rôle essentiel.

OÙ en sommes-nous ?

Nous pouvons raisonnablement penser aujourd'hui que les principales difficultés, pour préserver de l'information numérique à long terme ne sont pas directement d'ordre technologique mais sont plutôt liées à l'usage que nous faisons et que nous ferons de cette technologie. Les questions ouvertes sont très nombreuses :

Comment rendre l'information à préserver indépendante (dans son organisation logique) des technologies utilisées pour créer cette information ? Une partie très importante de l'information textuelle est actuellement produite à l'aide d'outils propriétaires. Ces outils produisent des formats de documents totalement fermés (c'est le cas de la plupart des logiciels de Microsoft). D'ici quelques dizaines d'années, ces documents ne seront plus lisibles ou ne le seront que très difficilement. Les orientations proposées dans l'excellent rapport au Premier Ministre de Thierry Carcenac : " Pour une administration électronique et citoyenne " publié au printemps 2001 vont dans le bon sens . Le recours à des solutions compatibles avec les standards d'internet (et notamment XML et ses standards connexes) doit selon lui constituer désormais une obligation. Il n'en reste pas moins que cette question n'est pas résolue dans tous les cas de figure : les moyens actuels permettant de passer des documents issus du monde du traitement de texte (Microsoft Word notamment) à des documents XML ouverts contenant les mêmes informations sont loin d'être simples, opérationnels et utilisables à grande échelle. Lorsque l'on veut en outre saisir et gérer un texte contenant des lettres de l'alphabet grec, des formules chimiques ou des équations mathématiques, on ajoute encore de la complexité au processus. Il y a donc sur ce point une réelle attente en outils et progiciels permettant de résoudre ces questions de façon performante, sûre et peu coûteuse.

Comment faire en sorte que tous les acteurs (producteurs d'information, archivistes, autorités de décision et de financement...) aient une compréhension globale du problème posé ? Comment en particulier convaincre les décideurs de fournir aux institutions patrimoniales, tous les moyens nécessaires pour relever ce défi ?

Comment développer, recruter et organiser les différentes compétences nécessaires ? Quelle formation nouvelle pour les archivistes ? Comment combler le fossé (parfois rempli d'incompréhensions) qui existe entre les archivistes et les informaticiens et faire en sorte qu'entre les deux, il puisse exister un domaine de compétence et de dialogue communs ?

Comment faire évoluer les comportements individuels, les habitudes de service pour prendre en compte les règles de base nécessaires à la pérennisation ?

Comment développer des services et des moyens permettant aux institutions de petite taille et aux personnes privées de préserver leurs informations ?

L'information est partout. Chacun dans son domaine, dans son entreprise assiste à cette croissance constante du nombre et de la diversité des sources d'information, du nombre de messages électroniques reçus chaque jour, du nombre de documents à lire… Cette profusion tend déjà à générer des comportements nouveaux acceptant l'idée que l'information numérique est éphémère. Jusqu'où ira ce phénomène ? Comment procéder à un tri intelligent de ce qu'il convient de préserver ?

Etc.

En conclusion, on peut dire que le dispositif normatif est bientôt suffisant, que la technologie est prête (même si elle va continuer à évoluer et si certains problèmes complexes ne sont pas encore bien résolus), mais que les hommes et les femmes ne le sont pas encore.

Je rejoins complètement la conclusion de Jean-Luc Philip sur le rôle essentiel que doit jouer la communication des archives dans ce processus. La communication immédiate des archives, via le réseau Internet, devrait avoir toute une série de conséquences en cascade : augmentation du nombre des utilisateurs des archives, émergence de nouvelles catégories d'usagers, meilleure valorisation du patrimoine archivé, etc.

Enfin, parallèlement aux multiples travaux de normalisation en cours sur le sujet, il paraît nécessaire aujourd'hui de lancer des projets d'expérimentation à grande échelle. Il ne s'agit pas de jouer aux apprentis sorciers et il convient donc de conduire ces expérimentations avec un luxe de précaution. Au-delà des réflexions et analyses théoriques, au-delà des modèles et des normes, c'est bien la mise en pratique réelle de la préservation numérique qui consolidera notre expérience et qui, en définitive, nous conduira peu à peu à des pratiques sûres et reconnues, à des moyens de certification indépendants et à la maîtrise des coûts.

Appel à commentaires relatif à l'établissement d'un cadre commun d'interopérabilité des systèmes d'information des administrations

La MTIC qui est devenue depuis quelques semaines l'ATICA (Agence pour les Technologies de l'Information et de la Communication dans l'Administration), a lancé un appel à commentaires jusqu'au 15 octobre 2001, appel relatif à l'établissement d'un cadre commun d'interopérabilité entre administrations.

Il comprend cinq questions sur les standards et référentiels, le choix de la famille des XML, les

choix dans le domaine des formats de documents numériques, la définition d'un cadre commun.

Les orientations qui seront fournies à la suite de cet appel pourront avoir des conséquences importantes sur la vie à long terme des documents électroniques des administrations.

Vous êtes tous invités à répondre à cet appel sur le site internet de l'ATICA :

http://www.atica.pm.gouv.fr/interop/

Conférence à Lyon, 5 et 6 novembre 2001

Long Term Archiving of Digital Documents in Physics

L'objectif de cette conférence est de formuler des recommandations pour la conservation et la mise à disposition sur le long terme des publications électroniques en sciences physiques.

Elle se tiendra à Lyon (France), les 5 et 6 novembre prochain et est organisée par le Centre pour la Communication Scientifique Directe (CCSD CNRS).

Pour toute information complémentaire :

http://publish.aps.org/IUPAP/

Lu pour vous

Raymond A. Lorie, A project on Preservation of Digital Data dans RLG DigiNews, vol. 5, n° 3, 15 juin 2001, http://www.ohio.rlg.org/preserv/diginews/diginews5-3.html

Dans cet article le chercheur américain propose une nouvelle méthode de conservation à long terme des données et programmes. Après la migration, les recherches actuelles sur l'émulation, de nouvelles voies sont ouvertes avec le Universal Virtual Computer (UVC). Il s'agit de remédier aux inconvénients des deux précédentes approches et de permettre la conservation distincte des données et des programmes d'une part et de restaurer les unes et les autres ou l'ensemble grâce à un "UVC interpreter" d'autre part.