Jean-Charles Hourcade, Franck Laloë et Erich Spitz

LONGÉVITÉ de L'INFORMATION NUMÉRIQUE

Les données que nous voulons garder vont-elles s'effacer ?

ACADÉMIE DES TECHNOLOGIES
POUR UN PROGRÈS RAISONNÉ, CHOISI ET PARTAGÉ

INSTITUT DE FRANCE
Académie des sciences

EDP SCIENCES

LONGÉVITÉ DE L'INFORMATION NUMÉRIQUE

Les données que nous voulons garder vont-elles s'effacer ?

Rapport du
groupe PSN (pérennité des supports numériques)
commun à l'Académie des sciences
et à l'Académie des technologies

Membres du groupe :

Erich Spitz,
Académie des sciences et Académie des technologies, président

Jean-Charles Hourcade,
Académie des technologies

Franck Laloë,
LKB/ENS, rapporteur

INSTITUT DE FRANCE
Académie des sciences

EDP SCIENCES

ACADÉMIE DES TECHNOLOGIES
POUR UN PROGRÈS RAISONNÉ, CHOISI ET PARTAGÉ

Conception de la couverture : Jérôme Lo Monaco

Maquette et mise en pages : Patrick Leleux PAO (Lisieux)

Imprimé en France

ISBN : 978-2-7598-0509-9

Résumé

Pourquoi s'intéresser à la préservation à long terme de l'information numérique, alors que les capacités de stockage numérique n'ont jamais été aussi vastes et bon marché ? C'est qu'il ne faut pas confondre deux notions très différentes, celle de **stockage** des données et celle de leur **archivage**. Les progrès spectaculaires des disques durs et la chute de leur prix permettent maintenant de stocker aisément de l'information, généralement en plusieurs exemplaires, pour s'affranchir des risques de pannes imprévisibles comme les « disk crash ». Mais **archiver** de cette façon sur des décennies ou un siècle pose un tout autre problème, du fait que les supports numériques n'ont qu'une durée de vie de 5 ou 10 ans environ. Dès qu'un disque dur arrive en fin de vie et risque de perdre définitivement les données, il est indispensable de les recopier vers un support neuf. L'évolution des supports étant difficile à prévoir, seul un suivi constant des données permet d'en assurer l'archivage, avec un coût d'organisation important.

Les disques optiques enregistrables ont quant à eux des capacités plus réduites mais sont crédités d'une meilleure durabilité,

comme le suggère le vocabulaire courant : on « grave » des données sur un disque enregistrable. Cette notion de « gravure » nous renvoie inévitablement à l'image des inscriptions antiques gravées dans la pierre et le marbre de vestiges millénaires. Ce sentiment de sécurité est malheureusement trompeur : aucun support actuellement commercialisé ne peut garantir une bonne conservation bien au-delà de 5 ou 10 ans environ !

Nos sociétés génèrent des masses toujours plus grandes d'informations, alors que la durée de vie des supports disponibles pour la conserver n'a jamais été aussi courte. Si ce problème est correctement pris en compte dans quelques organismes publics spécialisés, il est très largement ignoré du grand public ainsi que de la majorité des institutions ou entreprises. Beaucoup d'information, personnelle, médicale, scientifique, technique, administrative, etc. est en danger réel de disparition.

Le groupe PSN (Pérennité des supports numériques) a été créé à l'automne 2008 par les deux Académies, des sciences et des technologies, à la suite à la constatation de cette situation préoccupante, et avec l'ambition de faire un point sur le sujet.

Le présent rapport se donne un périmètre précis (chapitre 1), indispensable pour un sujet où les digressions possibles sont nombreuses. Il se concentre d'abord sur la fraction de l'information qui garde sa valeur à long terme : documents personnels – souvenirs familiaux, données médicales, etc. – ou documents publics – données scientifiques acquises lors d'expériences uniques, etc.

Ensuite sont discutées les stratégies possibles (chapitre 2) : « archive et oublie », dite parfois stratégie passive, la plus naturelle ; la stratégie active (migrations perpétuelles) ; la délégation à un prestataire de service ; le retour à l'analogique.

Les différents supports de stockage sont passés en revue dans un troisième chapitre (disques optiques enregistrables, bandes magnétiques, disques durs, mémoires flash, etc.), avec une brève discussion de leurs qualités et limitations.

Le quatrième chapitre évalue la possible généralisation de la stratégie active à l'ensemble des besoins de la société, qu'il s'agisse de documents personnels du grand public ou de ceux des établissements publics et des entreprises privées.

Le dernier chapitre se concentre sur les disques optiques numériques enregistrables, pour lesquels toute une série de mesures alarmantes ont été effectuées récemment. Il souligne le caractère non fondamental des problèmes rencontrés : si le vieillissement des disques optiques enregistrables est actuellement mal contrôlé, cela tient plus aux priorités qui ont été retenues dans les choix du marché qu'à des raisons essentielles.

Le rapport propose quelques pistes qui pourraient conduire à des disques enregistrables de bien meilleure longévité.

Une série de quatre recommandations est émise en fin de rapport.

This page intentionally left blank

Table des matières

This page intentionally left blank

Introduction

Le sujet du présent rapport est la préservation à long terme de l'information numérique. Chacun sait que cette information est produite journellement en quantités énormes depuis quelques années. Le numérique a maintenant remplacé l'analogique dans presque tous les domaines : l'immense majorité des documents scientifiques, médicaux, administratifs, ou encore les souvenirs personnels (photos, vidéos, etc.), sont directement créés en numérique. L'Unesco estime la production annuelle de l'humanité à plus d'un milliard de Gigabits, soit 10^{18} (un Exaoctet)[1], chiffre qui dépasse l'imagination. Ceci s'explique par la grande commodité de l'utilisation du numérique : flexibilité d'écriture, facilité de réutilisation, stockage compact, transmission à distance aisée et presque instantanée, etc. Le fait que le numérique permette des recopies sans erreur en nombre pratiquement illimité est déjà en soi une nouveauté extraordinaire : auparavant, toutes les copies impliquaient une accumulation progressive d'erreurs et

1. http://portal.unesco.org/fr/ev.php-URL_ID=4805&URL_DO=DO_PRINTPAGE&URL_SECTION=201.html

une dégradation de l'information analogique, qui finissait par disparaître. Pour le numérique, si l'humanité s'y prend bien, en principe rien n'empêche que les informations durent bien plus longtemps que celles écrites sur papier, voire aussi longtemps que les tablettes gravées de l'antiquité ! On comprend que le mouvement vers le numérique soit maintenant irréversible.

Mais il y a loin entre possibilités théoriques et pratique. Dans les faits, la conservation numérique de l'information se heurte à de nombreuses difficultés de nature assez diverse, plus ou moins faciles à résoudre. Ce rapport se concentre sur la principale d'entre elles, la longévité des supports d'information eux-mêmes ; comme nous le verrons, c'est véritablement elle qui est la clé pour la résolution des autres. La multiplicité des problèmes à résoudre entraîne parfois, dans les discussions sur ce sujet, une certaine confusion ; toutes les questions sont abordées à la fois. Dans ce rapport, nous tenterons d'éviter cet écueil et de bien sérier les questions ; c'est pourquoi nous commençons par en fixer précisément le cadre, quitte parfois à mentionner tel ou tel sujet pour nous contenter de dire qu'il ne sera pas traité.

Une confusion courante se produit entre deux notions pourtant très différentes ; celle de **stockage** (ou de sauvegarde) des données à cout terme et celle de leur **archivage à long terme**. La première pose de moins en moins de problèmes grâce aux progrès spectaculaires des supports numériques – chacun connaît ceux des disques durs en termes de capacité et de baisse de prix. Pour stocker sur une durée de quelques années, il suffit de copier les données à sauvegarder sur quelques disques durs (un seul ne suffirait pas à cause des risques de pannes soudaines et imprévisibles de type « *disk crash* ») pour être assuré de les conserver. Mais il existe des données importantes qui doivent être gardées sur des durées bien plus longues, des décennies ou des siècles, pour pouvoir être transmises aux générations futures. On dépasse alors de beaucoup la durée de vie de tous les supports d'information numérique (5 ans environ pour les disques durs, probablement moins pour les mémoires flash). La seule méthode possible est alors de transférer les données d'un support ancien vers un support neuf, avant que la détérioration naturelle du premier ne

rende la recopie impossible ; mais comme le vieillissement des supports n'est guère prévisible, seul un suivi constant des données permet d'effectuer l'opération au bon moment. Un tel processus est coûteux, non pas du fait du prix des supports, mais des interventions humaines et de l'environnement technique qu'il nécessite. L'information numérique dont personne ne s'occupe meurt au bout de quelques années.

This page intentionally left blank

Chapitre 1

Périmètre du rapport

1.1 ■ *Une sélection de l'information par le contenu*

Il ne sera pas question dans ce rapport de chercher à conserver à tout prix toutes les informations que chacun d'entre nous produit ou consulte, y compris des données d'intérêt momentané. Nous partirons du principe qu'une sélection est nécessaire dans le choix de celles qui ont réellement besoin d'être préservées à long terme. De façon très générale, on peut en simplifiant distinguer deux catégories d'informations :

– celle qui prend un intérêt croissant dans le temps, ou du moins garde un intérêt constant ;
– celle qui peut être importante pendant quelques temps, mais dont l'intérêt diminue et va s'effacer progressivement (ou même rapidement) au cours du temps.

Dans la première catégorie, chacun d'entre nous range évidemment les documents de sa mémoire familiale, en particulier tous les documents, photographies et souvenirs légués par nos ancêtres, ainsi que ceux que nous désirons transmettre à nos enfants. Cette catégorie inclut également les documents médicaux (en particulier les images médicales) qui doivent être conservés pendant 30 ans, les documents juridiques et légaux (actes de propriété par exemple) dont la valeur peut durer des siècles, les documents administratifs (nécessaires par exemple au calcul de la pension de retraite). En plus de ces informations plutôt personnelles, il existe également toute une catégorie d'informations publiques qui ont une grande valeur et sont évidemment à préserver : résultats scientifiques (données des satellites et sondes spatiales, des grands accélérateurs, bases de données biologiques, sociologiques, etc.). De plus, les sociétés privées doivent garantir l'accès à l'information concernant des produits complexes qu'elles fournissent, matériels et immatériels (plans de bâtiments, centrales, etc.) ; Dassault Aviation doit conserver les plans de ses avions pendant 70 ans ; les entreprises pétrolières doivent conserver des informations géologiques concernant l'exploitation des sites, qui peut parfois s'interrompre pendant des décennies. Ainsi, dans ce document, nous ne prendrons en compte que cette première catégorie d'information, celle dont la valeur dure longtemps, et sa conservation à long terme (plusieurs décennies, voire plus).

Nous supposerons également que, dès la création de l'information, un minimum de précautions élémentaires a été respecté ; c'est en principe le cas si, dès la saisie, l'utilité de la préservation à long terme était déjà claire. Nous ne nous intéresserons donc pas à la récupération de documents produits dans n'importe quelles conditions, avec un format et un logiciel pris selon les hasards du moment, et sans documentation puisque seul un objectif de court terme est visé. Par exemple, la récupération de l'intégralité des données du disque dur donné par une personnalité, à des fins d'archives historiques, pose parfois des problèmes redoutables « d'archéologie informatique ». La Direction des Archives de France sait à quel point les problèmes de classement de fichiers, en formats fluctuants, de systèmes d'exploitation changeants, etc.

peuvent être délicats, mais ils sortent du cadre de ce rapport. Heureusement il existe également des formats de stockage de l'information qui offrent de bonnes perspectives de pérennité, et nous supposons qu'un choix raisonnable sur ce plan a été effectué.

Dans la même veine, ce rapport ne traitera donc pas non plus des questions relatives aux projets comme la « photographie du web », la conservation de tous les messages électroniques, du dépôt légal des jeux électroniques, etc., non parce que ces questions sont sans intérêt, mais parce qu'elles sortent du périmètre défini.

1.2 ■ *Les supports, les formats, les logiciels*

Ce rapport se concentre sur la longévité des supports physiques sur lesquels est écrite l'information, et donc sur leur évolution dans le temps et la maîtrise des processus de vieillissement intrinsèques, même sans utilisation. Ainsi il ne traite pas :

- des problèmes liés à une mauvaise utilisation ou à une conservation négligente (CD laissé au soleil, rayures ou bris, etc.) ;
- des accidents (incendies, inondations...) et pannes soudaines ; leur prévention relève des techniques classiques de sauvegarde, alors que ce rapport se concentre sur celles de l'archivage à long terme ;
- de l'usure éventuelle lors de la lecture (usure des bandes magnétiques contre les têtes de lecture, blanchiment du colorant d'un CD-R par le faisceau laser, etc.) ;
- des questions de matériel de gravure ou de lecture[2] ;
- des questions de formats et de logiciels, dont la rapide évolution peut rapidement rendre illisibles certains documents. C'est l'aspect qui est le plus généralement saisi par tous, à tel point que c'est presque devenu une idée reçue, chacun

2. Nous disons quelques mots de la question de l'adéquation du couple support-graveur dans l'appendice 7.

> s'empressant d'évoquer l'évolution rapide des matériels et des logiciels dès que l'on évoque la conservation des documents numériques. Le problème est certes très réel, mais relativement bien circonscrit dans la mesure où il est possible, à des fin d'archivage, de choisir des formats de fichiers de façon raisonnable ; cet unique aspect du problème ne doit donc pas en occulter d'autres, plus difficiles à résoudre.

Tous ces problèmes sont, certes, très importants, mais ils se situent en aval du problème principal, celui de la longévité de l'information physique sur le support lui-même. Si une information importante est toujours présente sur un support numérique, on saura toujours la relire dans 50 ans ; mieux, l'existence même de cette information sur des supports pérennes suscitera le maintien ou la ré-apparition sur le marché des appareils de lecture nécessaires, avec les formats adaptés. Mais, bien sûr, si l'information elle-même disparaît du support physique, il n'y a plus aucune raison que cette continuité technique soit assurée. C'est bien le support physique de l'information qui est la clé de tout le processus de conservation !

1.3 ■ *Les métadonnées, les normes, la certification*

Nous n'aborderons pas non plus la question du classement et de l'identification des informations (métadonnées) et normes associées. Ces questions sont essentielles pour l'organisation de la communication des données aux utilisateurs, ainsi que de leur collecte, qui font partie de la mission d'organismes tels que le CNES ou la BNF. Elles ne sont cependant pas spécifiques de l'information numérique, mais se posent également pour toute autre forme d'information : des données, quelle que soit leur forme, perdent rapidement tout intérêt si personne ne sait à quoi elles correspondent, ni comment y accéder.

Il existe également des normes de caractère organisationnel plus que technique ; le modèle conceptuel de gestion OAIS (*Open Archival Information System*) sera brièvement mentionné au paragraphe 2.2.

Enfin, il existe des normes techniques, par exemple celles définissant les conditions de vieillissement artificiel en étuve (température, humidité) des disques optiques numériques[3] (ISO, AFNOR). En tant que point de repère permettant de comparer entre elles les mesures différentes, elles sont fort utiles. L'étape délicate est l'extrapolation (d'un grand facteur) pour obtenir une estimation de durée de vie en conditions normales d'utilisation. Cette extrapolation est généralement basée sur une formule d'Arrhenius ou d'Eyring contenant une exponentielle et une seule énergie d'activation, avec tous les risques que cela comporte : dans un système complexe mettant en jeu des processus multiples avec des constantes de temps probablement très différentes, un tel calcul peut conduire à des résultats totalement erronés – nous y reviendrons plus en détail au paragraphe 5.2.

Un autre problème important, mais qui sort du cadre de ce rapport, est la certification : il est important de pouvoir garantir que le fichier restitué au bout de quelques décennies est strictement identique au fichier original. La certification d'intégrité du fichier permet de garantir l'absence de modifications dues à des erreurs techniques, la certification d'authenticité qu'il n'a pas subi de modifications intentionnelles. Cette dernière est, en histoire par exemple, absolument essentielle pour des raisons évidentes. Différents outils sont utilisés pour associer à des fichiers des empreintes numériques (« *hash coding* ») qui permettent de garantir leur intégrité.

1.4 ■ *Une information personnalisable*

Ce rapport ne prend en compte que l'information écrite sur des supports enregistrables, que chacun peut utiliser en quelques exemplaires pour préserver ses données personnelles. Il ne sera

3. La norme ISO/IEC 10995 est la plus récente : http://www.iso.org/iso/catalogue_detail?csnumber=46554 ; elle se limite aux effets de la température et de l'humidité, sans prendre en compte les contaminants chimiques, l'exposition à la lumière, etc.

28 janvier 2009 : Madame M. Campana et Monsieur J. M. Besse (ministère de l'Industrie)

9 mars 2009 : Monsieur S. Deleonibus, CEA-LETI (Grenoble)

9 mars 2009 : Monsieur T. Ihashi (Bifrostec, Japon)

9 mars 2009 : Messieurs J. P. Gleyzes, M. Buégué, et Madame D. Boucon (CNES)

11 mars 2009 : Madame G. Pinson, Messieurs J. B. Henniart et C. Girard, Secrétariat d'État au Développement de la Région Capitale

2 avril 2009 : Monsieur L. Ranno, Département nanosciences, Institut Néel, UJF, Grenoble

20 mai 2009 : Messieurs F. Daumas, directeur du CINES (Montpellier) ; M. Auffret et O. Rouchon, département archivage et diffusion du CINES

25 juin 2009 : Monsieur Christian Amatore, membre de l'Académie des sciences

25 juin 2009 : Monsieur Gilles Lachaud, directeur de l'Institut de Mathématiques de Luminy

29 juin 2009 : Messieurs T. Ihashi (Bifrostec) et A. Inoué (Mitsubishi Chemicals)

22 octobre 2009 : Madame E. Dion et Monsieur D. Reizine (APHP, radiologie)