section 6 : Supports d'enregistrement et stratégies de stockage

FRANÇOISE BANAT-BERGER
CLAUDE HUC

version 1

22 novembre 2011

Table des matières

Chapitre 1. Objet de la section

A. Chapitre 1. Objet de la section

En définitive, c'est toujours l'archiviste qui portera la responsabilité de la conservation de l'information. Cette section 6, consacrée aux supports et aux stratégies de stockage, doit permettre de donner à l'archiviste une vision claire des solutions possibles et des risques encourus. La mise en œuvre des solutions sera le plus souvent confiée à des informaticiens, mais l'archiviste devra être en mesure de peser les avantages et les inconvénients des solutions envisageables et de faire des choix en parfaite connaissance de cause.

On se place ici du point de vue de la préservation des séquences de bits et de l'intégrité de ces séquences au cours du temps.

Les solutions, les choix de supports d'enregistrements, les stratégies de stockage qui seront retenues s'appuient évidemment sur les technologies disponibles dans ce domaine.

Il ne faut pas croire pour autant qu'il s'agit d'une question purement technique. Les choix seront fondés sur une analyse des besoins et des contraintes présents et futurs de l'Archive et de son contexte.

Nous sommes dans un contexte d'explosion des données en volume et en importance. Le taux d'augmentation en volume varie de 40% à 100% par an suivant les domaines. De plus en plus d'utilisateurs professionnels ou particuliers ont accès à l'informatique qui est devenue un outil de travail banalisé. Ces utilisateurs ont à leur disposition de plus en plus de programmes qui génèrent des données avec des ordinateurs de plus en plus puissants.

La décroissance des coûts de stockage contribue aussi à cette explosion.

On se demande

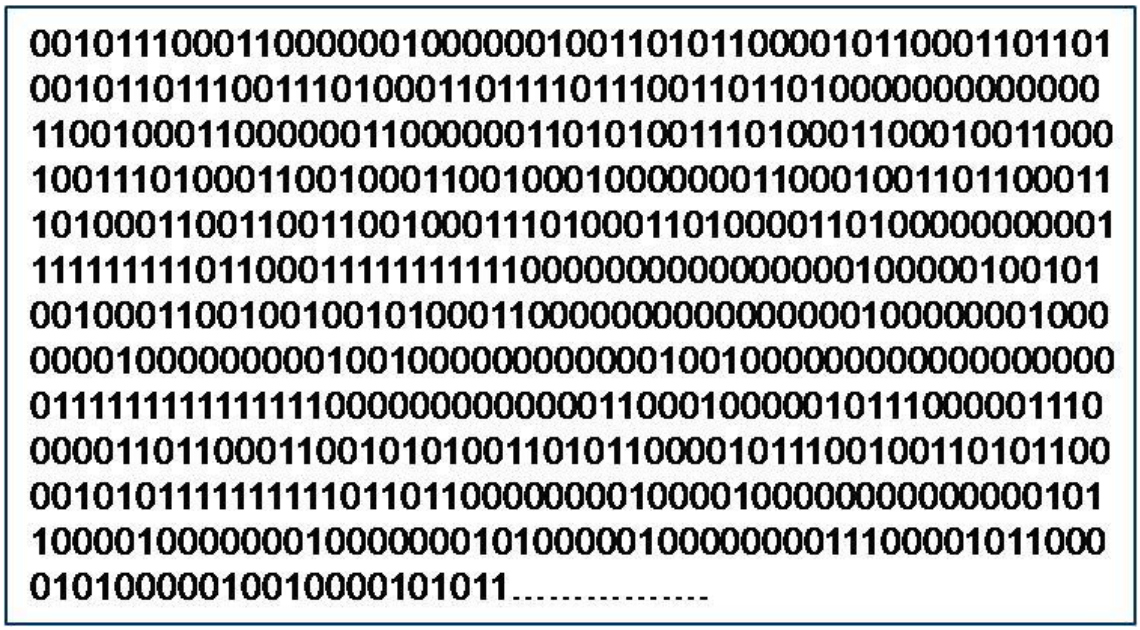
- en premier lieu, où mettre les données, sur quelle machine, quel média,
- mais aussi comment les organiser, comment les préserver dans le temps, garantir qu'elles ne seront jamais perdues tout en assurant leur intégrité.

Les données devront

- survivre aux incidents matériels et au vieillissement des supports,
- avoir leur intégrité assurée lors du passage d'un média vers un autre,
- ne subir aucune modification ni lecture non autorisée qui devront donc être empêchées.

Enfin, les infrastructures de stockage devront

- prendre en compte un certain nombre d'exigences relatives aux performances, aux temps d'accès aux documents, à la sécurité
- et en même temps respecter les contraintes en termes de limitation des coûts.



Assurer la pérennité et l'intégrité

B. Chapitre 2. Supports d'enregistrement et équipements de lecture et écriture

Les supports d'enregistrement, ce sont tous les moyens disques, bandes, cartouches, clés USB, etc. dont nous disposons pour stocker les données numériques.

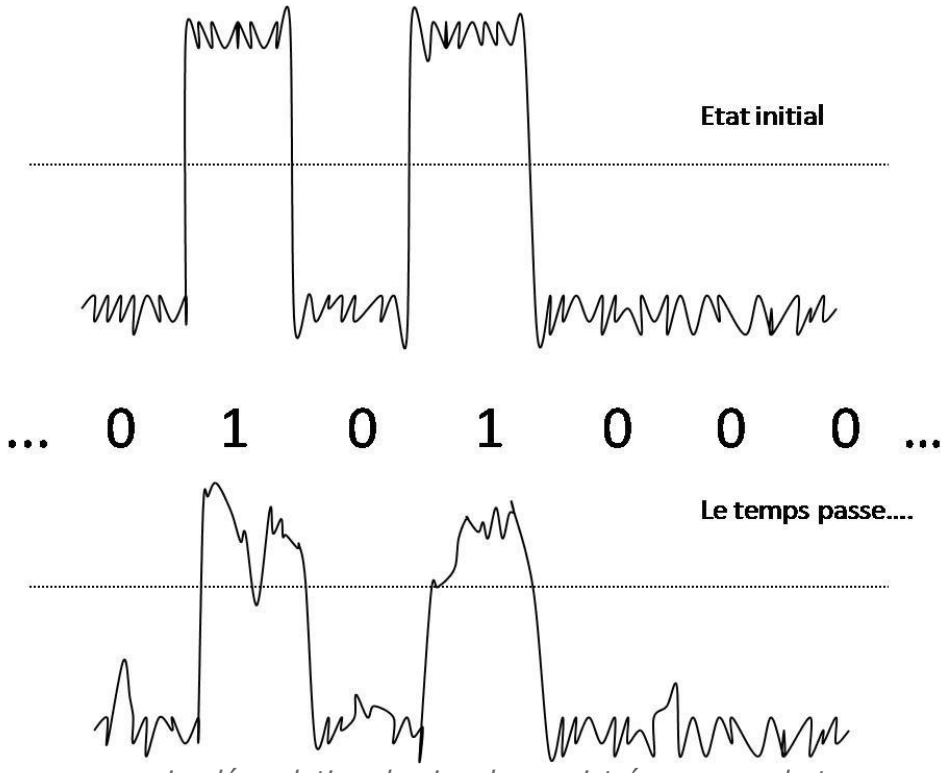
L'époque des supports de type cartes et rubans perforés est révolue depuis longtemps. L'usage des disquettes est en voie de disparition.

Nous utilisons tous maintenant des technologies optiques, magnétiques, magnéto-optiques offrant des supports à grande capacité. Naturellement, à chaque type de support sera associé un équipement de lecture.

C. 2.1. Enregistrement numérique et dégradation au cours du temps

L'enregistrement d'une information numérique sur un support en vue de mémoriser ou conserver durablement cette information se traduit par l'inscription sur ce support de séquences de 0 et de 1. Cette inscription se réalise sous une forme codée, le codage pouvant dépendre de la technologie utilisée.

La réalité physique implique donc l'inscription de valeurs discrètes sur le support. Cette inscription va généralement se dégrader progressivement au cours du temps.



La dégradation du signal enregistré au cours du temps

A terme, la dégradation de la modulation passera un seuil fatidique à partir duquel un ou plusieurs signes ne seront plus lisibles ou seront lus avec des erreurs consistant à confondre un 1 avec un 0 et inversement.

Les causes de cette dégradation sont nombreuses :

- instabilité des matériaux utilisés,
- évolution différenciée des éléments constitutifs conduisant à leur dissociation,
- inscription de l'information sur un matériau sensible à l'environnement (photosensible par exemple),
- conditions de stockage inappropriées (température, hygrométrie, poussière, fumée de cigarette),
- usure liée à la fréquence d'utilisation,
- mauvaise manipulation (dépôt de poussière ou débris, chocs),
- procédures de maintenance non respectées,
- pannes matérielles,
- erreurs humaines,
- vandalisme et vol,
- catastrophes naturelles, sinistres, incendies.

Mais aussi :

- qualité de fabrication,
- qualité de l'enregistrement ou de l'équipement utilisé pour procéder à l'enregistrement.

D. 2.2. Technologies existantes

Schématiquement, on peut dire qu'il existe cinq grandes familles de technologies pour les supports d'enregistrement :

- Les technologies mécaniques (bandes et ruban perforées) qui ont disparu aujourd'hui.

• Les technologies magnétiques qui sont les plus anciennes et qui progressent de façon régulière depuis les années 1950. Les supports magnétiques prennent la forme de bandes (appelées aussi cartouches dans certains cas) et de disques, comme par exemple le disque dur de votre ordinateur,

• Les technologies optiques qui datent du début des années 1970 (le Laserdisc a été construit en 1972) mais qui se sont développées au moment de la mise sur le marché des premier CD audio en 1981. Les supports les plus connus sont naturellement le CD-R (disque compact enregistrable), le DVD-R (DVD enregistrable) et plus récemment le Blu-Ray (appelé aussi laser bleu mais dont l'appellation officielle est BD). Le BD a une capacité de stockage de 25 Go en simple face et 50 Go en double face. Il a aussi existé des supports optiques sous forme de bandes,

• Les technologies magnéto-optiques, combinaison des deux précédentes, la lecture étant purement optique. C'est le cas de l'UDO (Ultra Optical Disc) qui est un disque de 133 mm de diamètre. Sa capacité de stockage de 60 Go (une face) ou 120 Go (deux faces) et sa durée de vie estimée à 50 ans en font un candidat possible pour l'archivage, malgré un coût élevé et une vitesse d'enregistrement assez lente,

• Les mémoires flash qui sont des mémoires de masse à semi-conducteurs réinscriptibles. Elles sont utilisées dans les clés USB, les baladeurs numériques, les appareils photos et, depuis quelques années, sur de nouvelles générations d'ordinateurs portables.

Les technologies magnétiques, optiques ou magnéto-optiques impliquent toujours un mouvement du support pour la lecture ou l'écriture (le disque tourne, la bande magnétique doit être déroulée), ce mouvement est nécessairement un facteur de risque d'usure ou de panne, alors que ce n'est pas le cas pour les mémoires à semi-conducteurs.

Une capacité de stockage qui s'envole !

La capacité de stockage des supports d'enregistrement a augmenté de façon phénoménale en 40 ans.

Au début des années 1970, le support le plus courant était la bande magnétique 9 pistes avec une densité de 800 bpi (bits per inch) et une capacité de 20 Mo. En 2009, toujours avec une technologie magnétique, IBM produit des cartouches magnétiques beaucoup plus petites que la bande 9 pistes avec une capacité de 1 Téraoctets (To). Il y a un facteur 50 000 entre la capacité de chacun de ces deux supports. Les supports disques ont évolué dans des proportions équivalentes. Et demain, lorsque les technologies optiques ou magnétiques auront atteint leurs limites, ce sont les mémoires holographiques qui prendront probablement le relais.

Complément

CD-R et DVD-R

Le CD-R (disque compact enregistrable apparu sur le marché en 1988) et le DVD-R (Digital Versatile Disc enregistrable apparu sur le marché en 1997) constituent des supports grand public peu coûteux et potentiellement intéressants pour l'archivage.

Pour ces deux types de support, il existe des analyseurs (appelés aussi testeurs) matériels ou logiciels, qui sont capables de fournir un ensemble d'informations utiles sur l'état du support.

Les atouts :

- pas de réécriture possible (nous parlons bien des DVD-R et DVD-R et non des CD-RW et DVD-RW qui sont réinscriptibles),

- une structure physique et logique définies par la norme ISO 9660 pour le CD. Par contre, il y a plusieurs standards concurrents pour le DVD, même si aujourd'hui les lecteurs sont compatibles avec tous,

- l'existence d'outils d'analyse fiables et accessibles,

- une grande simplicité de mise en œuvre.

Mais aussi des inconvénients :

- une capacité réduite, souvent insuffisante par rapport aux besoins d'aujourd'hui, surtout pour le CD-R,

- une durée de vie nettement moindre que celle des CD audio pour lesquels le processus de fabrication est différent,

- les produits supposés être dédiés à l'archivage professionnel par leurs fabricants sont d'une fiabilité incertaine (voir ci-après, les études menées sur ce sujet par le Laboratoire national de métrologie et d'essai pour la direction des Archives de France),

- les opérations de contrôle des supports sont lourdes (pas d'automatisation du chargement dans un testeur ni de l'exploitation des résultats). De même pas de possibilité d'automatisation des migrations pour tout un lot.

Les analyseurs :

Ils permettent d'extraire un ensemble de paramètres parmi lesquels :

• l'évaluation des erreurs rencontrées avant des entrelacements : l'information « telle qu'elle a été inscrite sur le disque » :

- le BLER (BLockErrorRate) : taux de blocs comportant une erreur,

- le FBE ou BERL (Frame Burst Error ou Burst ERror Length) correspondant à la plus longue salve de blocs erronés ; ce paramètre est plus significatif que le BLER.

• l'évaluation des erreurs après désentrelacement : l'information « exploitable par la machine » :

- les erreurs E32 qui sont les erreurs non corrigibles conduisant à une perte de données,

- les erreurs E22 correspondent au nombre de blocs comportant deux erreurs corrigibles (risque élevé de E32.

Le LNE (Laboratoire national de métrologie et d'essai) recommande, pour l'archivage, un ensemble de valeurs maximales pour ces paramètres :

Titre du tableau : Les seuils acceptables proposés par le LNE

	Valeurs initiales des taux d'erreur acceptables proposées par le LNE	Valeurs maximales des taux d'erreur avant migration proposées par le LNE
FBE (BERL)	≤ 4	7
BLER (moyen)	<6	50
BLER max	<50	220
E22	≤ 7	30
E32	0	0

Etudes disponibles sur le sujet :

A la demande de la direction des Archives de France (DAF), Le LNE a conduit une série d'études sur l'utilisation des CD-R et DVD-R pour l'archivage.

Sur cette base, la DAF a émis un certain nombre de directives et de recommandations.

Nous pouvons suggérer les documents suivants :

• La conservation de données sur CD-R (juillet 2004)

• Qualité des DVD disponibles sur le marché pour l'archivage des données numériques (octobre 2008)

• Qualité des CD-R disponibles sur le marché pour l'archivage des données numériques (juillet 2008)

• Guide à l'usage des services d'archives pour la réalisation de la migration de stocks de CD-R (mars 2009)

• Recommandation de la Direction des Archives de France (DAF) relative à la gravure, à la conservation, et à l'évaluation des CD-R (mars 2005)

Cette recommandation couvre l'ensemble de la chaîne en incluant le choix du CD-R, le choix du graveur, le mode de gravure, les conditions de stockage, la surveillance et le renouvellement des CD.

Autres études disponibles :

• Care and Handling of CDs and DVDs en ligne sur le site du « National Institute for Standards and Technologies »

• IASA guidelines on the production and preservation of digital audio objects, BRADLEY Kevin, 2004

Les recommandations pratiques :

• Pour la gravure :

• graver au moins 2 exemplaires

• un master d'archivage,

une copie de consultation,voire une copie de travail d'où seront tirées les copies de consultation (si elles sont très sollicitées),

• constituer un échantillon-témoin représentant tous les numéros de série, tous les graveurs, toutes les périodes de temps, et enregistrer ces métadonnées,

• Pour le stockage : stocker les supports sur de sites distants,

• Pour le contrôle du vieillissement :

• contrôler tous les 3 ans au plus à partir de la gravure,

• tester les disques échantillonnés en totalité à vitesse 1x,

• si un disque approche des seuils limite, engager la recopie du sous-ensemble.

Ces recommandations rejoignent les recommandations générales qui peuvent être proposées pour tous les types de support.

Le disque optique est support commode pour commencer, mais il y a nécessité de mettre en œuvre d'autres solutions (substitutives ou complémentaires) à moyen et long terme

E. 2.3. Comment parer au manque de fiabilité des supports et de moyens de lecture ?

Pour parer au manque de fiabilité des supports et des moyens de lecture, il existe plusieurs techniques se situant à des niveaux différents :

• Les codes correcteurs d'erreurs (CCE) qui pallient les pertes et erreurs du train de bits causés par les imperfections du support ou du lecteur. Ils permettent dans une certaine mesure de corriger une ou plusieurs erreurs au sein d'un bloc physique de données sur le support,

• L'empreinte numérique vise à contrôler l'intégrité des données. Elle permet d'apporter une garantie de non altération d'un fichier ou d'un groupe de fichiers. Elle ne permet pas de corriger les erreurs. Cependant, si on constate un écart entre un objet numérique et sa copie, l'empreinte permettra de dire quel est l'exemplaire valide.

Les techniques d'empreinte sont largement utilisées dans les processus de signature électronique qui seront abordés dans la section 10 « intégrité, authenticité et preuve » de ce module.

Complément

Les codes correcteurs d'erreurs ont généralement deux seuils :

- Un premier seuil qui définit le nombre d'erreurs pouvant être corrigées,
- un second qui définit le nombre d'erreurs pouvant être détectées.

Ces codes correcteurs d'erreurs sont principalement internes aux équipements de lecture. Ils ont un coût sous forme de séquences de bits supplémentaires qu'il faut stocker et décoder. Des algorithmes mathématiques parfois complexes sont mis en œuvre. Un exemple simple est celui des bits de parité :

Imaginons une séquence de bits sous la forme d'une suite d'octets empilés les uns après les autres. La représentation de cette suite prend la forme d'un tableau dans lequel chaque ligne contient 8 bits et représente un octet :

0	1	0	1	0	0	0	0
0	1	0	0	1	0	0	1
0	1	0	0	1	1	1	0

Suite d'octets initiale

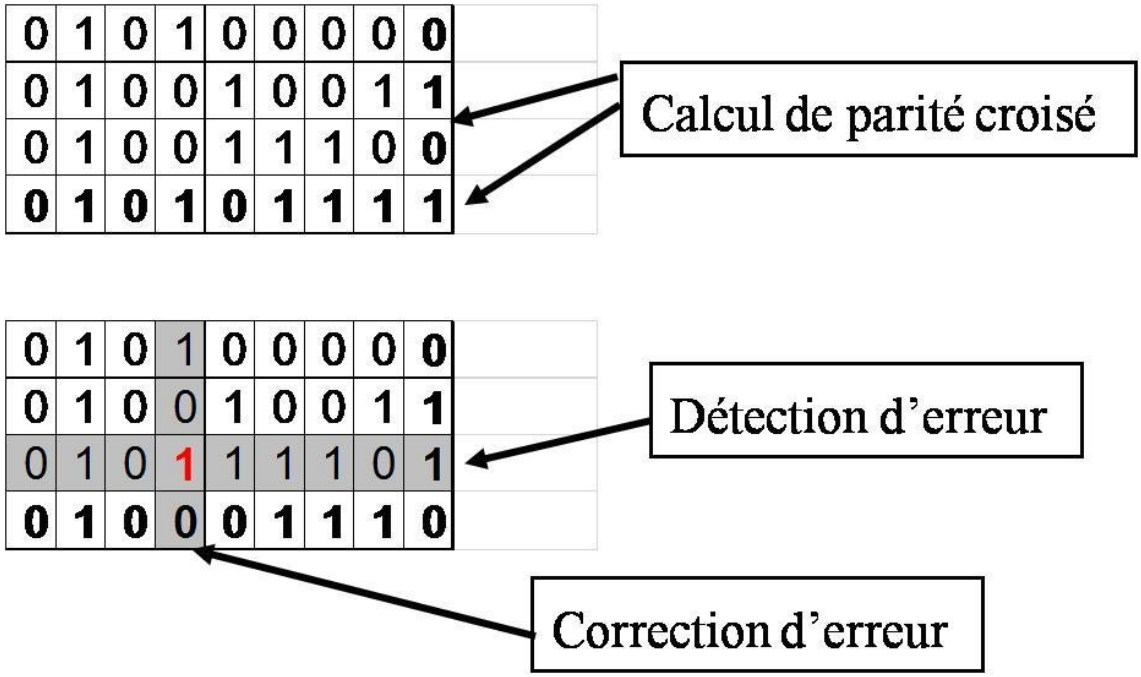
Un tableau de trois octets

Ajoutons pour chaque ligne et pour chaque colonne, ce que nous appellerons un bit de parité : ce bit prend la valeur 0 si le nombre de bits dans l'état « 1 » de la ligne ou de la colonne est pair. Sinon, ce bit de parité prend la valeur 1.

Si une erreur de lecture se produit pour un bit (remplacement d'un 0 par un 1 ou réciproquement), le lecteur va constater qu'il y a une erreur sur une ligne puisqu'il y a incohérence entre le nombre de bits dans l'état « 1 » et la valeur du bit de parité.

Les bits de parité par colonne vont nous permettre d'identifier la colonne sur laquelle il y a une erreur de bit.

Au croisement de la ligne et de la colonne identifiés, on pourra donc corriger l'erreur rencontrée.



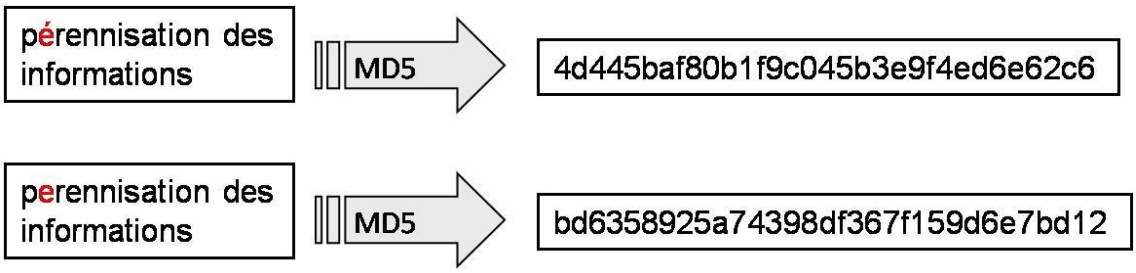
Un mécanisme simple de détection et correction d'une erreur de bit

Ce mécanisme est rudimentaire et ne pourra pas fonctionner si on rencontre plusieurs erreurs. Il s'agit ici d'une simple illustration permettant de montrer qu'en ajoutant un certain nombre de bits de contrôle, on dispose d'une capacité de corriger les erreurs jusqu'à un certain point.

Naturellement, ces codes correcteurs d'erreur ayant des capacités limitées, il conviendra de recopier les données sur un support neuf. Cette recopie joue le rôle d'une « régénération » comme le montre la figure ci-après.

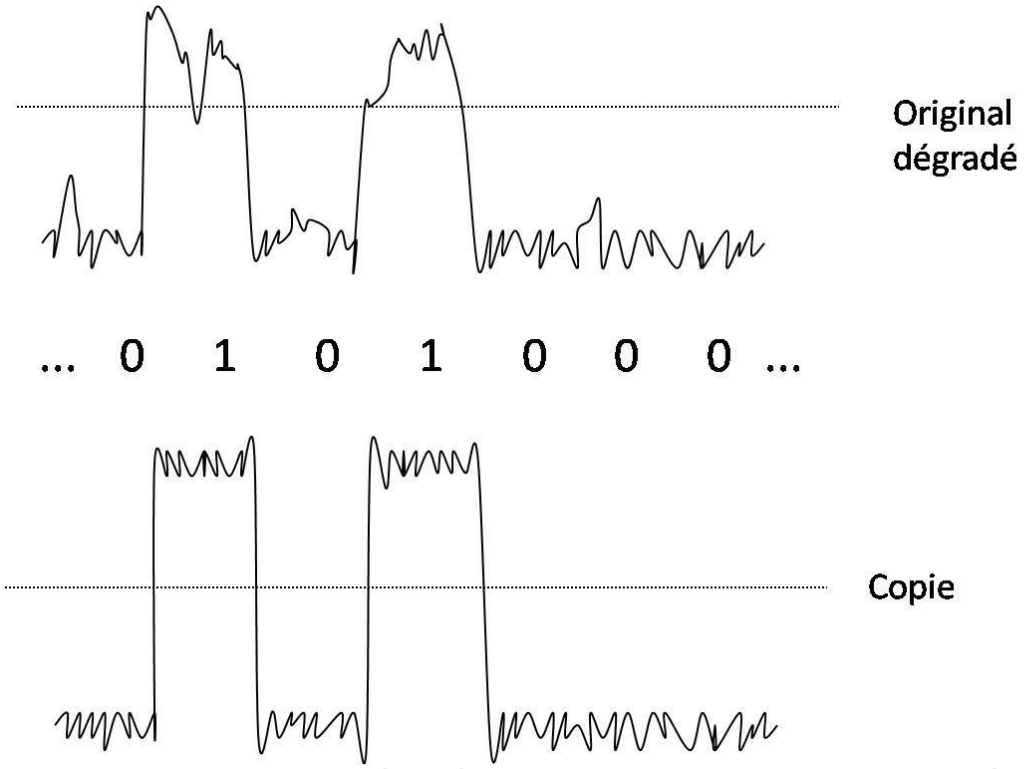
Les empreintes numériques

Les empreintes numériques sont des chaînes de caractère calculées à partir d'un algorithme ayant des propriétés mathématiques particulières, les algorithmes de hachage. Les algorithmes les plus connus sont MD5 (Message Digest 5) et SHA (Secure Hash Algorithm).



Le changement d'un seul caractère conduit à une empreinte MD5 complètement différente

Avant qu'il ne soit trop tard, on procèdera donc à une recopie des données depuis le support à risque vers un support neuf.



La recopie d'un support dégradé sur un support neuf constitue une regénération

En général, la simple lecture du support ne permet pas de déterminer le niveau de dégradation de ce support. Il est donc nécessaire de recopier les données contenues vers d'autres supports avant qu'il ne soit trop tard.

Ajoutons enfin que lorsqu'un support analogique devient partiellement dégradé, seule l'information contenue dans la partie dégradée est perdue. Ce sera le cas pour un texte dont quelques lignes deviennent illisibles. À l'inverse, compte tenu de la structure complexe des fichiers, une erreur de bit au milieu d'un fichier entraînera souvent la perte du complète du fichier. Dans le meilleur des cas, on ne perdra que la totalité de l'information qui suit l'erreur jusqu'à la fin de ce fichier. Parfois, si le fichier endommagé est la table des matières du support, c'est l'ensemble des fichiers du support qui n'est plus accessible par le système de lecture alors que ces derniers peuvent être en parfait état.

Attention

En résumé

Le manque de fiabilité et de pérennité des supports sera surmonté par un ensemble de dispositions complémentaires :

- les différentes techniques de création, gestion et traitement de codes correcteur d'erreur,
- le maintien en permanence de plusieurs copies des objets numériques, de préférence sur des supports de type différents,
- la recopie régulière des données sur d'autres supports, cela avant qu'il ne soit trop tard.

Nous disposons dans cette section de premiers éléments qui interviendront dans ce que nous appelons les stratégies de stockage.

F. 2.4. Obsolescence technologique

Les technologies évoluent sans cesse ou en remplacent d'autres, la durée de vie des technologies disque ou bande peut être courte et imprévisible. Même si le support peut

physiquement survivre pendant des dizaines d'années, la technologie qui permet de lire et d'interpréter les données peut n'exister que pendant une brève période.

Les facteurs d'obsolescence d'une technologie sont multiples. Nous retiendrons les deux principaux :

• les évolutions techniques d'une technologie par ses fabricants : elles sont relativement prévisibles ; la plupart des fabricants ont des plans d'évolutions technologiques des supports et garantissent une compatibilité avec la ou les version(s) précédente(s) ; il est alors largement possible d'anticiper correctement les migrations nécessaires ;

• les évolutions du marché : si une lecture attentive du marché permet d'anticiper ce type d'événements, ces évolutions restent, dans une certaine mesure, assez peu prévisibles ; il s'agit de l'abandon d'une technologie par un fabricant ou, pire, de la disparition de ce fabricant ; l'augmentation des coûts de maintenance liée au faible nombre ou à l'absence de fabricants de matériels de lecture est un facteur déterminant dans le choix de changement de technologie.

Exemple

Deux exemples révélateurs correspondant aux deux facteurs énoncés :

• la disparition de la bande magnétique 9 pistes 6250 bpi, d'une capacité limitée à 150 Mo, a été annoncée par les constructeurs dès les années 1990, ce qui a contraint de nombreux organismes à transférer leurs données sur bandes vers d'autres supports ; c'est le cas du CNES en France, alors même qu'il disposait de millier de bandes magnétiques dont la fiabilité était assurée jusqu'en 2015 ; la technologie des bandes 9 pistes, qui impliquait la disponibilité d'équipements de lecture/écriture lourds et coûteux, a été progressivement remplacée par d'autres supports magnétiques et optiques, à la fois moins coûteux et d'une plus grande capacité ;

• à la fin des années 1980, le constructeur canadien « Creo Products Inc » a développé une excellente technologie de bande optique d'une capacité de 1 To, ce qui était phénoménal à cette époque ; malheureusement, le marché pour une telle bande était très restreint et, en dehors du CNES et de la NASA qui utilisèrent ces bandes pour stocker des images d'observation de la terre par satellite, les clients furent peu nombreux ; ce fut un échec commercial et la fabrication prit fin au cours des années 1990.

On peut observer également que les technologies grand public, comme celle du CD-R, ont souvent une durée de vie supérieure aux technologies destinées aux usages professionnels.

Complément

Titre du tableau : Quelques exemples d'évaluation des risques au niveau des supports et de la technologie

	Risque support	Risque technologie
Bande ½'' Memorex	élevé	maximum
Disquette 3''1/2	élevé	moyen à élevé
CD-R bon marché non contrôlé	élevé	très faible
ZIP, Exabyte...	faible	élevé
MiniDisc (audio)	faible	élevé

Un consortium international, l'ICPSR (Inter-University Consortium for Political and Social Research) assure la maintenance d'un très intéressant tutoriel multilingue sur la préservation numérique. On y trouve en particulier une chambre des horreurs qui dresse un

paysage impressionnant des supports bandes, disques, semi-conducteurs qui sont nés au cours des 40 dernières années et dont une bonne partie sont déjà morts ou en voie d'obsolescence ou de disparition.

G. 2.5. Conditions de stockage à respecter

La durée de vie d'un support ne va pas seulement dépendre de ses qualités intrinsèques. Cette durée de vie sera optimale si les conditions de stockage sont satisfaisantes. De mauvaises conditions environnementales sont également un facteur entraînant la dégradation prématurée des supports.

Nous pouvons retenir les recommandations suivantes pour la plupart des supports optiques ou magnétiques amovibles :

• maintenir une température entre 16°C et 23°C avec un gradient maximum de 4° C par heure,

• maintenir une humidité relative entre 30 % et 50 % avec un gradient maximum de 10 % par heure,

• utiliser des conditionnements opaques pour le stockage des médias à défaut d'utiliser le conditionnement d'origine,

• stocker les médias plutôt à la verticale,

• n'apposer aucun étiquetage ni aucune encre sur les médias,

• éviter l'exposition aux champs magnétiques,

• éviter l'exposition à la lumière du soleil et aux UV de certains systèmes d'éclairage,

• éviter l'exposition à la poussière et à la fumée,

• interdire les boissons et la nourriture dans les lieux de stockage,

• interdire de fumer dans les lieux de stockage,

• relier à la terre les librairies et les juke-boxes en métal.

Dans le cas de supports non amovibles comme des baies de disque, des recommandations spécifiques doivent être définies.

H. 2.6. Contrôle des supports en vue de déclencher des opérations de migration

Ce qui vient d'être expliqué dans cette partie du cours consacrée au stockage nous conduit à déduire qu'il est nécessaire de recopier les données sur un autre support (identique ou non), c'est-à-dire de les migrer avant le moment fatidique de la perte irrémédiable de ces données.

La difficulté est donc de pouvoir déterminer le moment où il faut mettre en œuvre cette migration de supports :

• une migration prématurée entraînera un surcoût,

• une migration tardive entraînera une perte de données.

Selon les types de supports et les moyens disponibles, il y a deux méthodes

1. Le contrôle périodique des supports

Il est toujours possible d'effectuer une relecture périodique des supports à titre de contrôle. Cette relecture pourra s'opérer sur un échantillon ou sur la totalité des supports en fonction des conditions et du type de support. Cette relecture en tant que telle ne vous apprendra

pas grand-chose sauf s'il s'avère que certains supports sont illisibles, ce qui déclenchera alors une alarme pour opérer une migration des supports dans les plus brefs délais.

La prudence incite à éviter ce type de situation.

Notons également que pour certains types de support comme les bandes magnétiques, une relecture périodique constitue un facteur de régénération du support empêchant notamment les phénomènes de collage.

Nous avons en définitive deux types de situation :

• Soit il existe des moyens matériels et logiciels capables de nous renseigner sur l'état du support AVANT qu'il devienne illisible (nous avons vu que c'était le cas pour les CD-R et les DVD-R) ;

dans ce cas, il est possible mettre en place des procédures de contrôle pour vérifier l'état des supports ; la méthode consistera :

• en premier lieu, à identifier les éléments qui sont suffisants pour caractériser l'état du support, définir les valeurs initiales à obtenir lors de l'enregistrement et les valeurs à partir desquelles le support devra être retraité ;

• ensuite, il faut déterminer un échantillon jugé représentatif de la qualité d'un sous-ensemble ou lot de supports et, pour chaque lot, déterminer le nombre d'échantillons à tester (ce qui implique d'avoir constitué des métadonnées relatives par exemple au numéro du lot, à la date de fabrication...) ; pour cela on pourra soit se baser sur la norme ISO 2859 (Procédures d'échantillonnage pour les contrôles lot par lot, indexés d'après le niveau de qualité acceptable, procédures pour l'évaluation des niveaux déclarés de qualité), soit utiliser un nombre défini par retour d'expérience ; en fonction de l'espérance de vie du support, nous pouvons déterminer une période selon laquelle l'échantillon représentatif sera vérifié.

• Soit il n'existe pas de moyens d'analyse et nous sommes contraints de mettre en œuvre des migrations planifiées décrites ci-après ; l'offre des constructeurs en matière de contrôle des médias étant pauvre, ce sera une situation courante

2. Les migrations planifiées

Pour un nombre important de supports d'enregistrement, il n'existe pas d'outils de contrôle sur lesquels nous appuyer.

Il est donc nécessaire de planifier les opérations de migrations de support, comme dans le cadre d'une maintenance préventive où les matériels, ou bien certaines pièces critiques du matériel, sont changés périodiquement pour éviter les pannes intempestives. Il s'agit de recopier les contenus des supports existants vers des supports neufs.

Pour fixer le moment de la migration :

• nous pouvons nous baser sur les indications que fournissent les fabricants et prendre des marges de sécurité raisonnables par rapport à ces indications ;

• des opérations statistiques de relecture des supports doivent néanmoins être planifiées à des fins de surveillance.

Attention

En conclusion

L'information numérique ne dure que si l'on assure un contrôle de l'état des supports et une recopie de l'information en temps utile.

Chapitre 3. Stratégies de stockage

Stocker n’est qu'une des fonctions de l'archivage et il ne faut pas les confondre. Si le stockage est un élément essentiel du processus, en aucun cas il ne constitue l’unique élément d’une stratégie de pérennisation, mais la base sur lequel repose l’archivage.

Qu’est ce qu’on attend d’un service de stockage au sein d’une Archive numérique ?

C’est une infrastructure qui est capable d’assurer la garantie d’une restitution des informations

- sans altération (intégrité, authenticité),
- sur le long terme (migrations, obsolescence),
- dans une problématique de coût contrôlé.

Revenons d’abord plus en détail sur ce que nous dit le modèle OAIS à propos de l’entité fonctionnelle « Stockage »

A. 3.1. Entité « Stockage » du modèle OAIS

L’entité « Stockage » du modèle OAIS assure les fonctionnalités suivantes :

1. Recevoir les AIP (paquets d’information)

Enregistre dans une zone de stockage approprié et envoie un acquittement signifiant que le stockage a été effectué.

2. Gérer la hiérarchie de stockage

En fonction des exigences de qualité de services des entités de versement et d’accès, en fonction de la fréquence prévue d’utilisation de l’AIP si elle est connue, on choisira les supports adéquats permettant un accès en ligne, en différé ou en léger différé. On devra s’assurer au préalable que les AIP n'ont pas été altérés durant le transfert.

Cette fonction fournit aussi des statistiques sur les supports à disposition et la capacité de stockage disponible dans les différentes couches ainsi que sur l’utilisation des AIP.

3. Assure les migrations des supports

Cette fonction offre la possibilité de reproduire les AIP dans le temps. Ce sont les migrations technologiques. Au cours de ces opérations, le Contenu d’information et l’Information de pérennisation ne doivent pas subir de modifications. Néanmoins, les données constituant l'Information d’empaquetage peuvent être changées, pour autant qu’elles continuent d’assurer la même fonction. La stratégie de migration consistera à choisir un support de stockage en tenant compte des taux d’erreurs réels et attendus caractérisant les différents

types de supports, de leurs performances, et de leur coût d'acquisition.

4. Contrôle régulièrement l'intégrité des informations confiées

Cette fonction garantit, avec une probabilité statistiquement acceptable, qu'aucun AIP n'a été corrompu lors d'un quelconque transfert interne des données de l'Entité « Stockage ». L'Information d'intégrité assure, dans une certaine mesure que le Contenu d'information n'a pas subi au fil du temps, de modification, qu'il y ait ou non déplacement de l'AIP d'un média à un autre, qu'il y ait ou non accès à cet AIP.

5. Régénère les données en cas de sinistre.

Le « plan de reprise d'activité » fournit un mécanisme pour dupliquer les contenus numériques de l'Archive et stocker la copie dans une installation géographiquement distante.

6. Fournit les AIP à l'entité d'accès

Cette fonction transmet à l'Entité « Accès » les copies des AIP stockés. Cette fonction reçoit une demande d'AIP précisant les AIP demandés et les livre sur le type de support demandé ou les transfère vers un espace de stockage provisoire.

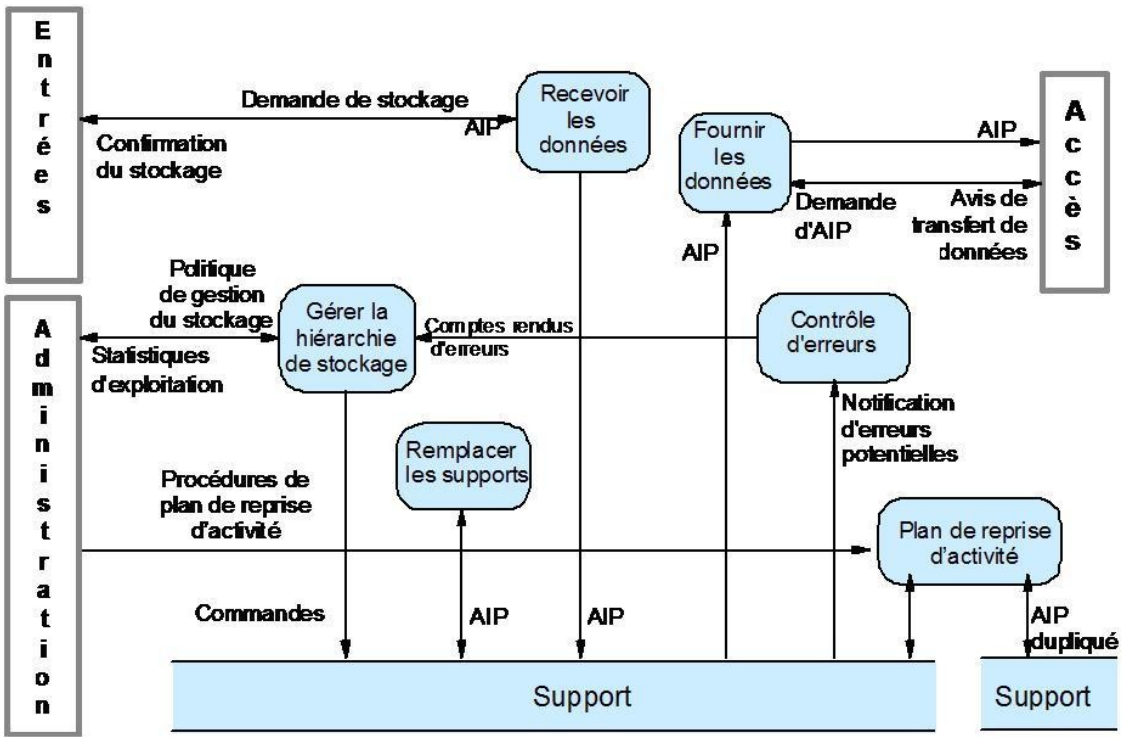


Schéma fonctionnel détaillé de l'entité « Stockage »

Nous allons donc retrouver l'ensemble de ces fonctionnalités dans la mise en œuvre d'un service de stockage.

B. 3.2. Abstraction de la plate-forme matérielle

Le stockage prend la forme d'un ensemble de moyens organisés en système. Un « système de stockage » est un ensemble de logiciels et de matériels qui rendent un service pleinement défini.

Sachant que cette plate-forme va évoluer régulièrement puisqu'elle est constituée de matériels et de logiciels qui vont changer au cours du temps, sachant qu'elle va aussi devoir gérer des opérations de surveillance et de renouvellement des médias de stockage, il apparaît qu'il y a un grand intérêt à constituer cette plate-forme de façon autonome par rapport au reste du système d'archivage numérique.

Ce point de vue pragmatique repose essentiellement sur les retours d'expérience des plates-formes existantes.

Avec une telle approche :

- les évolutions de la plate-forme de stockage doivent être sans conséquence sur l'organisation logique de l'archivage : le producteur des données mais aussi l'Archive gèrent une organisation logique des données et la capacité de transformer ces données en une information intelligible. Cette organisation logique n'a aucune correspondance avec l'organisation physique des données sur les supports d'enregistrements, organisation qui changera au cours du temps en fonction des technologies et des supports disponibles,
- les changements de politique de la hiérarchie du stockage (classes de service) et les migrations des supports doivent être transparents, c'est-à-dire sans impact, sur les entités utilisatrices du service de stockage.

Il s'ensuit que les utilisateurs (appelés aussi clients) du service de stockage n'ont pas besoin de savoir sur quels supports d'enregistrement leurs documents sont stockés à un moment donné. Ils ont besoin de disposer des garanties de service adéquates en termes de pérennisation, intégrité, accès…

Il s'agit ici d'une approche en termes de service. C'est par exemple ce que nous attendons du courrier postal, à savoir que notre courrier parvienne à son destinataire en bon état et dans les délais prévus. Peu nous importe les moyens de transport que La Poste utilise pour acheminer le courrier.

Cette conception facilitera également la mutualisation, c'est-à-dire l'utilisation du service de stockage par différents groupes, départements, projets au sein de l'organisme, voire chez des organismes partenaires.

Complément

Service de stockage au Centre national d'études spatiales (CNES)

Cet exemple est présenté plus en détail dans la section 12 consacrée aux études de cas. Il est ici résumé pour illustrer l'abstraction du stockage.

En 1990, le CNES disposait de plusieurs dizaines de milliers de bandes magnétiques 9 pistes contenant des données scientifiques.

En raison de plusieurs facteurs, notamment de la disparition annoncée des technologies de stockage sur bandes magnétiques 9 pistes et de la volonté de rendre les données scientifiques accessibles et utilisables par la communauté la plus large, le CNES a fait le choix, dès 1992, de la mise en place d'un service central de stockage.

La vocation de ce service est d'apporter une véritable garantie de conservation à long terme des bits, quelle que soit la technologie de stockage utilisée.

Ce service spécialisé en charge de pérenniser les fichiers est le STAF (Service de transfert et d'archivage des fichiers). On peut noter ici une confusion habituelle et perverse entre archivage et stockage).

Il est opérationnel depuis 1994 et se présente comme une entité indépendante des projets ou des services d'archive. Ces derniers sont les clients du STAF et s'adressent à lui au moyen d'un ensemble de commandes de base permettant notamment de demander le stockage ou la restitution d'un fichier ou d'un ensemble de fichiers. Ces communications passent par le réseau interne du CNES. Le STAF a donc une mission très simple :

- recevoir des fichiers sans avoir à connaître leur format ni leur contenu informationnel,
- assurer la conservation à long terme de ces fichiers,
- garantir leur intégrité,
- garantir leur confidentialité,
- les restituer à la demande.

Cette mission simple est en même temps une responsabilité très lourde puisque toute perte d'intégrité des fichiers conduit à la perte des informations contenues.

Aujourd'hui, les migrations de support sont réalisées de façon continue par le STAF et ne sont pas visibles des clients du service. En quinze ans d'existence et avec une volumétrie qui s'approche à grands pas du pétaoctet, le STAF n'a pas perdu une seule donnée, il a démontré l'intérêt et l'efficacité des principes sur lesquels il a été construit, à savoir une totale indépendance de la fonction de stockage par rapport aux autres entités fonctionnelles d'un service d'archivage long terme.

L'approche, retenue aujourd'hui pour nombre de grands sites d'archivage numérique repose sur les mêmes principes.

C. 3.3. Modes de stockage et hiérarchie de stockage

On distingue trois modes de stockage.

Stockage en différé (appelé off line en anglais)

• C'est le mode de stockage qui s'apparente le plus au magasinage de documents physiques. Il s'agit d'entreposer les supports « sur étagères ».

• Il nécessite tout de même de prendre les précautions inhérentes à la nature des supports, notamment le respect d'un ensemble de conditions environnementales. Les migrations des supports sont opérées manuellement et sont déclenchées selon des procédures non-automatisées de contrôle de l'état de support. La fiabilité de ce mode de stockage repose entièrement sur le respect des procédures mises en place. Elles doivent donc faire l'objet d'une étude et d'une documentation détaillées. L'enjeu sera pour l'organisation de les respecter sur le long terme.

• Ce mode se prête particulièrement aux fonds pour lesquels le taux de consultation est faible.

• Du fait de son apparente simplicité de mise en œuvre, la tentation est grande de vouloir accepter de plus en plus de supports hétérogènes. La nécessité de disposer de compétences permettant de maîtriser l'ensemble des supports conduit à une dispersion de l'effort porté sur leur suivi et notamment le contrôle de leur état. Cette dispersion nuit à la fiabilité du stockage. Pour éviter ce travers, ce mode de stockage doit se limiter à un nombre restreint de types de supports.

Stockage en ligne (appelé « on line » en anglais) :

• Ce mode de stockage est le plus performant au niveau des temps d'accès mais aussi le plus cher. Les données sont directement accessibles. Elles sont enregistrées sur des disques magnétiques.

Stockage en léger différé (appelé « near line » en anglais) :

• Ce mode de stockage désigne les modes de stockage qui utilisent des systèmes robotiques pour manipuler les supports appelés juke-boxes ou librairies. Les supports acceptés par ce type d'équipement sont divers : bandes, disques optiques.

• Les données ne sont pas directement accessibles. Le support doit être acheminé depuis son emplacement de stockage jusqu'au lecteur par un bras manipulateur. Dans le cas des bandes, il est nécessaire de dérouler la bande jusqu'à l'emplacement du ou des fichiers à lire ou à écrire. De ce fait, les temps d'accès à l'information sont très variables selon le type de média utilisé et les performances de l'équipement, de quelques secondes à plusieurs minutes.

• Ce type de système est souvent destiné à gérer de gros volumes de données à partir de quelques téraoctets jusqu'à des dizaines de pétaoctets.

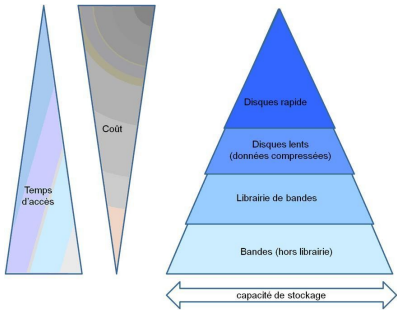
Complément

La gestion de la hiérarchie du stockage

Dans cette gestion (appelée hierarchical storage management en anglais, ou HSM en

abrégé), la donnée va être prise en charge par le système qui, en fonction de paramètres prédéfinis, va appliquer une politique de stockage. Ainsi, les données peu utilisées vont passer de disques ultra-rapides à des disques classiques pour finir sur des bandes en fonction du niveau du taux de consultation et de la politique de hiérarchisation des données préétablie.

Les données vont descendre petit à petit vers la base de la pyramide, passant vers des supports moins coûteux qui ont plus de capacité , la contrepartie étant une dégradation des performances, principalement en termes de temps d'accès aux données. Le but est d'optimiser les coûts de stockage automatiquement.



D. 3.4. Classes de service

La classe de service définit pour un (ou plusieurs) type(s) de données, le niveau de qualité attendu du service de stockage, en termes de :

- garantie de performance,
- garantie de stockage (nombre de copies = sécurité),
- garantie de disponibilité,

Pour les performances :

- le temps minimal, moyen et maximal de stockage d'un objet,
- le temps minimal, moyen et maximal de récupération d'un objet.

Mais on pourra aussi se préoccuper d'autres fonctions pour lesquelles des exigences de performances sont utiles.

Pour la garantie de stockage, la classe de service peut permettre de paramétrer les éléments suivants :

- nombre de copies,
- capacité totale (en octets),
- taille maximale d'un objet (en octets),
- nombre maximal d'objets,
- compression matérielle (active ou inactive),
- cryptage (actif ou inactif),
- type de cryptage (algorithme utilisé, longueur de codage),
- type d'écriture (de type Worm ou non),
- type de support (disque dur, bande, disque optique, disque magnéto-optique, etc.).

Pour la disponibilité, les engagements du service de stockage peuvent prendre différentes formes :

- disponibilité du service 24h sur 24 avec une durée moyenne d'indisponibilité inférieure à 30 minutes par jour,
- la durée d'indisponibilité peut aussi être fixée par mois.

La disponibilité du service est à fixer avec précision car le réseau qui permet d'accéder au service de stockage peut ne pas dépendre de ce service. Il peut s'agir du réseau d'entreprise par exemple. Nous pouvons donc avoir des situations dans laquelle le service de stockage est opérationnel mais le réseau permettant de l'utiliser ne fonctionne pas.

Exemple

- le temps de mise à disposition maximal d'un fichier d'une taille inférieure à 100 Mo ne doit pas dépasser 1 minute pour une base de 1.000.000 fichiers,
- au moins 2 copies du fichier sont enregistrées (le fichier est donc stocké en 3 exemplaires),
- le fichier est disponible 24 h/24 et 7 j/7.

Chapitre 4. Politiques de stockage

Trois politiques de stockage peuvent être envisagées :

- la mise en œuvre interne,
- la mutualisation,
- l'externalisation.

A. 4.1. Mise en œuvre interne

La mise en œuvre interne offre à l'organisation (une entreprise, une institution…) la maîtrise complète de la plateforme de stockage et de son évolution. Cette liberté de décision permet d'adapter l'ensemble des dispositions et des choix aux stricts besoins de cette organisation.

En contrepartie, cette mise en œuvre nécessite d'importantes ressources humaines et financières et le développement de compétences spécialisées.

è Cette mise en œuvre interne n'est souhaitable que pour des entreprises ou des institutions de taille importante.

B. 4.2. Mutualisation

La mutualisation d'un service de stockage pérenne entre plusieurs partenaires offre des avantages importants :

- un partage des coûts,
- des économies d'échelle importantes : le coût moyen de stockage annuel d'un Go décroît de façon importante en fonction du volume stocké,
- une mise en commun des compétences permettant d'accroître la sécurité et la fiabilité du stockage.

Il est possible de mettre en place un stockage mutualisé sur des critères géographiques ou institutionnels, indépendamment des logiques métier et des contenus.

Exemple

- possibilité d'envisager au niveau d'un département français, la gestion de l'ensemble des

données numériques par une seule entité (dossiers vivants et archives courantes traitées par les services administratifs et dossiers archivés gérés par les services d'archives),

• autre possibilité au niveau national : stockage sécurisé et centralisé des données et documents de la recherche et des universités,

• existence aux Etats-Unis du San Diego Supercomputer Center (SDSC) qui fédère le stockage des données de quarante universités dont l'université de Californie,

• En Europe, les travaux de l'« Alliance for a Permanent Access » ou encore le projet CLARIN (Common Language Resources and Technology Infrastructure) pour les ressources langagières vont dans cette direction mais avec un spectre qui dépasse largement la question du stockage.

Le risque principal réside dans la difficulté de trouver des partenaires.

Complément

Un service de stockage mutualisé peut être installé chez l'un des partenaires ou construit comme une entité totalement et uniquement dédiée au stockage pérenne comme le montre la figure suivante.

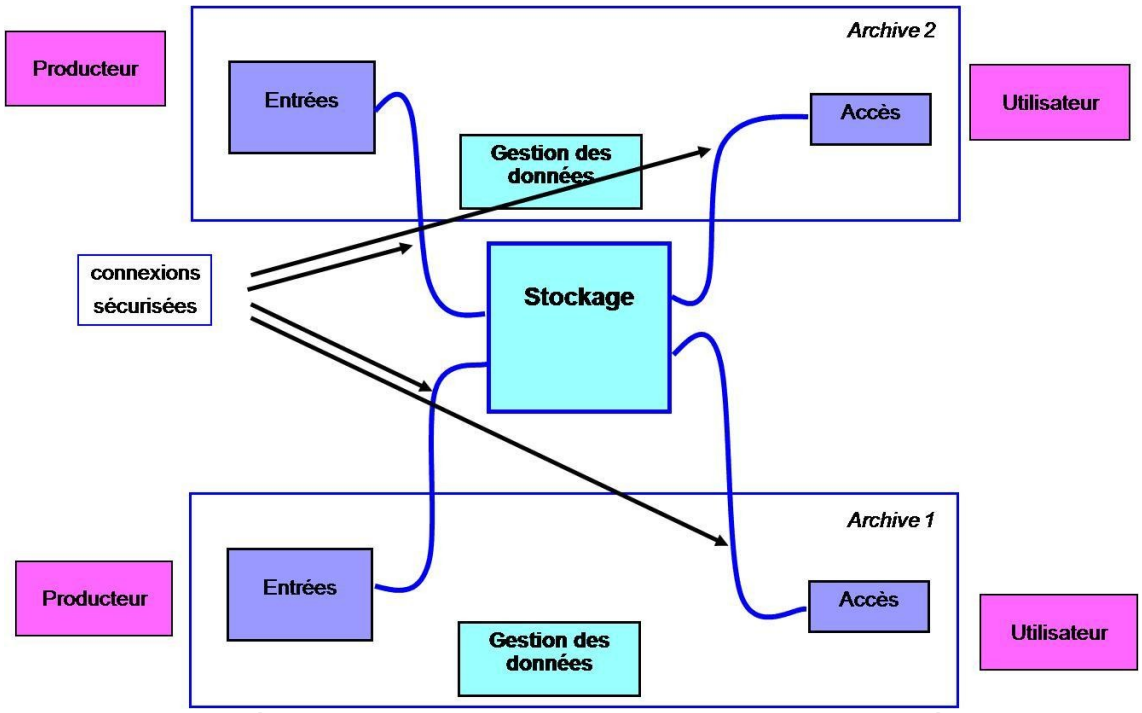


Schéma de principe d'un service de stockage mutualisé

C. 4.3. Externalisation (tiers archiveurs)

Cette voie permet le recours à des acteurs disposant des compétences et des moyens. Elle facilite les économies d'échelle, compte tenu de l'importance des coûts fixes (personnel, matériel).On doit néanmoins se poser la question de la pérennité de l'opérateur et de la sécurisation du réseau permettant de transmettre ou de récupérer les données.Il s'agit là d'un marché émergent. Sa crédibilité n'est pas encore consolidée mais il devrait à terme offrir certains avantages comparables à ceux de la mutualisation.La norme Afnor NF Z42-013, examinée dans la section 5 de ce module, contient un chapitre consacré aux tiers archiveurs et aux types de contrats que l'on peut passer avec eux. Des clauses types sont proposées pour ces contrats.

Chapitre 5. Quelques cas d'école

Ces quelques cas visent à illustrer la diversité des réponses organisationnelles et techniques qui peuvent être apportées par une Archive, en fonction du volume de données à archiver.

A. 5.1. Une petite Archive (100 Go)

Nous pouvons rencontrer le cas d'une numérisation de documents patrimoniaux pour lesquels se pose le problème du stockage des fichiers de conservation. Ces documents sont consultés régulièrement, mais à partir de fichiers de consultation avec des formats compressés et à partir de l'application de consultation. Le problème de pérennité se pose pour l'exemplaire de conservation : stockage sur disque optique en 2 ou 3 exemplaires dont un stocké à distance).

On fera en sorte de ne pas utiliser le même support ou en tout cas le même fabriquant pour les différentes copies.

Un contrôle régulier (annuel) est à opérer.

Une recopie sous 5 à 7 ans est à prévoir.

Il convient d'explorer les voies d'un archivage de sécurité dans une archive partenaire ou chez un prestataire.

Nous attirons également l'attention sur les risques possibles au niveau des petites structures et sur les moyens de pallier ces risques :

- compétences insuffisantes pour déterminer les paramètres optimums pour le contrôle,
- ne pas respecter parfaitement les procédures de gravage, de contrôle et de migration,
- consacrer trop de ressources internes au stockage des données et au contrôle des supports, aux dépens de la mission première de la structure : ce sera le cas dès qu'on aura à manipuler 50 ou 100 CD-R ou plus ; on peut recommander à ce stade, c'est bien préférable, d'avoir 1 Blu-Ray de 25 Go en simple couche plutôt que 50 CD à graver, contrôler, migrer…

B. 5.2. Une Archive de taille moyenne

Exemple d'un archivage de documents administratifs produits sous forme électronique.

Nous supposons que ces documents sont livrés sur CD-R ou DVD-R contrôlés.

Le stockage sera organisé de façon mixte sur disque et sur bande :

• la totalité des fonds sur un serveur doté de disques RAID (Redundant array of inexpensive disks),

• une copie sur bandes magnétiques LTO (Linear Tape Open) stockées à distance via une architecture permettant le transfert de données par réseau.

C. 5.3. Une grande Archive (500 To ou plus)

Ce sera le cas pour un archivage massif de publications en ligne de données scientifiques.

L'entrée des données dans le système d'archivage implique un ensemble de contrôle et un stockage momentané sur disque avant leur transfert au service de stockage.

Le stockage sera organisé sur bandothèque (bibliothèque de bandes magnétiques) sécurisée, répliquée en permanence sur deux sites,

Un stockage hiérarchique sera par ailleurs mis en place avec plusieurs niveaux de service (temps d'accès) selon la fréquence des demandes et les besoins des usagers.

Glossaire

Archive

Organisation chargée de conserver l'information pour permettre à une communauté d'utilisateurs cible d'y accéder et de l'utiliser (glossaire du modèle de référence OAIS).

Empreinte

Empreinte (empreinte numérique ou condensat ou hash) : Résultat d'une fonction de hachage appliquée sur une chaîne de caractères de longueur quelconque visant à réduire celle-ci en une donnée de longueur fixe représentative de cette chaîne de caractères. L'empreinte est l'un des éléments permettant de vérifier l'intégrité d'un document, d'un flux, d'un lot, d'une transmission,... (comparaison d'empreintes).