Smals

Rendre l'information accessible durablement

Préservation à long terme de l'information numérique

Clients et services

Section Recherches

Date : Février 2010
Deliverable : 2010/TRIM1/01
Statut : Final
Auteur : Arnaud Hulstaert

Koninklijke Prinsstraat 102
1050 Brussel

Rue du Prince Royal 102
1050 Bruxelles

Tel : 02/787.57.11
Fax : 02/511.12.42

Management Summary

L'utilisation croissante des technologies de l'information et de la communication à des fins de gestion et d'exploitation a fait de la préservation à long terme de l'information numérique un enjeu crucial pour les entreprises et les institutions (quantité croissante d'informations numériques, réglementations imposant leur conservation sur des durées relativement longues, enjeux financiers importants, nombreux types d'informations – dans certains cas très complexes – à préserver).

N'étant par nature pas auto-explicative, l'information numérique naît de l'interaction entre une séquence de bits et des éléments hardware et software, la rendant dès lors soumise à l'évolution hétérogène de ces différentes composantes. Le problème est que l'ère du numérique dans laquelle nous sommes entrés est marquée par une évolution impressionnante des technologies (les unes remplaçant les autres).

Incluse dans l'archivage, la préservation consiste à maintenir les objets archivés en état, c'est-à-dire accessibles et compréhensibles par ses utilisateurs. Du fait de la fragilité inhérente de l'information numérique, sa préservation nécessite d'appliquer de manière continue tout au long du cyle de vie de l'information des stratégies techniques et conceptuelles qui ne sont efficaces que si elles sont encadrées par une organisation financée durablement.

Au niveau de l'organisation, un préalable indispensable est un engagement fort de la direction eu égard aux budgets et compétences qui devront être rassemblés. Tout projet de ce type doit commencer par une étude du modèle conceptuel *Open Archival Information System* (OAIS) devenu progressivement une norme incontournable dans le domaine et normalisée en 2003 par l'ISO. Permettant de saisir la problématique de manière globale (du point de vue fonctionnel et au niveau des informations à rassembler), il constitue un excellent guide pour la mise en œuvre de projets d'archivage à long terme. Pour être préservée, l'information numérique doit être maintenue dans un système qualifié de fiable sur les plans de l'organisation, de la gestion et des stratégies techniques et conceptuelles mises en œuvre. Divers modèles d'audit (dont le plus élaboré est DRAMBORA) existent et offrent une aide efficace pour évaluer la capacité d'un système à préserver l'information.

Le modèle OAIS n'offrant qu'un modèle conceptuel, il reviendra à chaque organisation de traduire cette organisation en différents services, chacun assumant une partie des tâches et des responsabilités.

Une fois ce cadre organisationnel élaboré, l'organisme souhaitant préserver ces informations doit recourir à diverses stratégies techniques et conceptuelles, appliquées de manière continue. Étant donné qu'il n'existe aucune solution globale et unique, nous insistons sur l'importance de combiner ces stratégies.

Les stratégies techniques et conceptuelles opérationnelles actuellement et complémentaires sont :

- la gestion des supports de stockage, incluant le choix des supports, leur contrôle régulier et leur remplacement ;

- la gestion des formats, comprenant le choix de formats qualifiés de pérennes, leur validation et le recours à des *format viewers* ;
- la migration régulière des données vers des nouveaux formats ou des architectures matérielles et logicielles plus récentes, en veillant à la compatibilité ascendante des logiciels et en prenant soin de documenter rigoureusement le processus de migration ;
- le recours aux métadonnées, base indispensable de toute autre stratégie. À cet égard, les standards développés ces dernières années (dont METS et PREMIS) offrent une aide indéniable.

Chacune de ces stratégies permet de préserver une ou plusieurs couches (physique, binaire, logique et sémantique) de l'information numérique.

Deux autres stratégies sont parfois présentées. L'encapsulation est intéressante mais encore peu mise en oeuvre aujourd'hui. L'émulation est utilisée aujourd'hui au niveau des supports de stockage. Au niveau logiciel, elle n'est clairement pas opérationnelle à l'heure actuelle.

Le coût de la préservation demeure un problème complexe à gérer, d'autant plus que ce coût n'est que difficilement chiffrable. Dès lors, en vue de le diminuer, diverses stratégies de mutualisation peuvent être mises en œuvre.

Table des matières

Management Summary 2

But et structure du document 6

1. Contexte et enjeux 8

2. Terminologie et concepts 12

2.1. Définitions *12*
2.1.1. Stockage 12
2.1.2. Archivage 13
2.1.3. Préservation à long terme 13
2.1.4. Complémentarité entre archivage et préservation 14
2.2. Cycle de vie d'une information numérique et ILM *15*
2.3. Positionnement du système d'archivage *17*

3. Difficultés 20

3.1. Obsolescence technologique *20*
3.1.1. Hardware et supports de stockage 21
3.1.2. Composants logiciels 22
3.1.3. Formats et types d'encodage 23
3.2. Intégrité et authenticité *24*
3.3. Organisation *25*
3.4. Intelligibilité de l'information *25*

4. Stratégies organisationnelles 27

4.1. Modèle OAIS *28*
4.1.1. Le modèle fonctionnel 29
4.1.2. Le modèle de données 30
4.2. Construire un référentiel fiable *32*
4.3. Mise en place d'une organisation *33*
4.4. Coûts et stratégies de mutualisation *34*
4.4.1. Coût de la préservation 34
4.4.2. Stratégies de mutualisation 35
4.5. Synthèse *38*

5. Stratégies techniques et conceptuelles 40

5.1. Modèles en couches *40*
5.2. Typologie et positionnement des stratégies *41*
5.3. Migration et gestion des supports de stockage *42*
5.3.1. Le choix des supports de stockage 42
5.3.2. Conditions de stockage 44
5.3.3. Contrôle de l'état des supports 44

5.3.4. Migration de support 45
5.4. Gestion des formats *46*
5.4.1. Sélection de formats pour la préservation 48
5.4.2. Identification et validation 51
5.4.3. Format Viewer 52
5.5. Migration et conversion de format *53*
5.5.1. Définition et types de migration 53
5.5.2. Difficultés 54
5.5.3. Recommandations 55
5.6. Métadonnées *56*
5.6.1. Catégories de métadonnées pour la préservation 57
5.6.2. Normes et standards existants 58
5.7. Encapsulation *62*
5.8. Émulation *64*
5.9. Synthèse *65*

6. Étude de cas 67

6.1. Contexte *67*
6.2. Stratégie organisationnelle *69*
6.3. Stratégies techniques et conceptuelles *71*
6.4. La préservation de la signature digitale *75*

7. Conclusion 78

8. Bibliographie 80

9. Glossaire 82

But et structure du document

L'utilisation croissante des technologies de l'information et de la communication à des fins de gestion et d'exploitation a fait de la préservation à long terme de l'information numérique un enjeu crucial pour les entreprises et les institutions. N'étant par nature pas auto-explicative, l'information numérique naît de l'interaction entre une séquence de bits et des éléments hardware et software, la rendant dès lors soumise à l'évolution hétérogène de ces différentes composantes.

Le problème est que l'ère du numérique dans laquelle nous sommes entrés est marquée par une évolution impressionnante des technologies (les unes remplaçant les autres). Aussi peut-il arriver pour une information stockée sur un support que l'on ne dispose plus de périphériques de lecture (problème de hardware). On peut également perdre des données d'un fichier créé au moyen d'un logiciel dans une version antérieure à la version du logiciel que l'on utilise pour lire le fichier (problème de logiciel). Enfin, il arrive aussi que l'on ne dispose plus de la technologie nécessaire pour avoir accès à l'information contenue dans un fichier informatique. C'est le cas par exemple quand une institution stocke de l'information via un format propriétaire qui, au fil des années, devient illisible pour l'ensemble des programmes sur le marché (problème de format). L'information, bien que conservée, n'a pas survécu à l'obsolescence technologique et est alors inaccessible. Les efforts à déployer pour réinventer une infrastructure permettant d'y accéder peuvent s'avérer extrêmement lourds, voire vains...

Pour éviter d'en arriver là, il est impératif de mettre en place une série de stratégies techniques, conceptuelles et organisationnelles en tenant compte des bonnes pratiques en la matière. Le présent document a donc pour objectif de présenter les enjeux et les difficultés de la préservation digitale ainsi que les solutions envisageables pour y remédier.

Pour ce faire, nous exposerons tout d'abord le contexte et les enjeux de la préservation digitale (Chapitre 1). Étant donné que comme beaucoup d'autres domaines des sciences et technologies de l'information et de la communication, cette thématique n'échappe pas à un foisonnement de termes et de concepts utilisés de manières différentes par les nombreux intervenants, nous nous arrêterons ensuite sur plusieurs définitions essentielles, notamment en vue de définir le scope de la présente étude (Chapitre 2).

Un examen des difficultés et des principaux obstacles de la préservation à long terme de l'information numérique (Chapitre 3) permettra ensuite de mettre en évidence les problèmes auxquels il convient d'apporter une réponse.

Nous aborderons ensuite les stratégies organisationnelles (Chapitre 4) à mettre en œuvre en vue de répondre aux difficultés abordées dans le chapitre précédent et de soutenir les stratégies techniques et conceptuelles.

Les stratégies techniques et conceptuelles (Chapitre 5) à mettre en place seront présentées. En les positionnant par rapport aux différentes couches qui constituent une information numérique, nous verrons que ces solutions doivent être complémentaires, chacune apportant une réponse à un problème précis. Nous exposerons tout d'abord les stratégies opérationnelles (gestion des supports de

stockage, gestion des formats, migration et métadonnées) tout en prenant soin d'en exposer les difficultés inhérentes. Viendront ensuite deux stratégies souvent évoquées mais encore peu éprouvées à l'heure actuelle (encapsulation et émulation).

Enfin, nous concrétiserons les concepts et stratégies vus précédemment par une étude de cas (Chapitre 6) issue de la sécurité sociale (l'archivage à long terme des contrats de travail électroniques), avant de conclure (Chapitre 7).

Remarques liminaires

Touchant de nombreux domaines des sciences et des technologies de l'information et de la communication, la problématique de la préservation à long terme de l'information numérique est extrêmement vaste et complexe, requérant de nombreuses compétences et du savoir-faire.

Dans le cadre de cette étude, il était donc impossible de tout traiter. Le choix a donc été fait de centrer cette étude sur l'information dite non structurée[1], à savoir les documents bureautiques (Word, Excel, PDF), les présentations assistées par ordinateur, les images (TIFF essentiellement), etc. pour lesquels les solutions existantes sont davantage matures. Sont également inclus les fichiers XML. Certains points spécifiques à l'information structurée (essentiellement les bases de données) ne seront donc pas abordés dans ce délivrable mais feront l'objet d'une étude ultérieure.

De nombreux points abordés sont cependant suffisamment généraux (par exemple le modèle OAIS – cf. 4.1) pour être applicables quel que soit le type d'information.

[1] Conceptuellement, du point de vue de la nature de l'information, l'information non structurée désigne l'information dont le domaine de définition est empirique et flou par définition, c'est-à-dire que sa contrepartie dans le « monde réel » n'existe pas. Voir BOYDENS I., *Documentologie. Section de Science et technologies de l'information et de la communication*, Bruxelles : Université Libre de Bruxelles (syllabus), 2009-2010.

1. Contexte et enjeux

> *La question de l'archive n'est pas une question du passé. […] C'est une question d'avenir, la question de l'avenir même, la question d'une réponse, d'une promesse, d'une responsabilité pour demain.*
>
> Jacques Derrida

Bien que présente depuis de nombreuses années, la problématique de la préservation à long terme de l'information numérique prend aujourd'hui une dimension croissante et se trouve confrontée à cinq enjeux.

Le premier enjeu est l'augmentation continue des volumes de données numériques et de leur importance. Dans notre domaine d'application, pensons aux déclarations sociales, aux procédures d'enregistrement des médicaments, aux données utilisées pour le calcul des impôts ou des pensions, aux nombreuses informations issues du domaine de la santé…

L'exemple des e-mails est à cet égard éclairant. Une étude réalisée en 2007 par IDC[2] pour le compte d'EMC estimait que le volume d'e-mails échangés allait passer en 5 ans de 10 à 35 milliards, tandis que les entreprises allaient devoir archiver 7.000 pétaoctets d'e-mails à l'horizon 2010.[3]

Les volumes deviennent tels que la même étude estimait que, en 2008, la quantité d'informations numériques créée allait dépasser les capacités de stockage disponibles, et ceci pour la première fois dans l'histoire du numérique. Évidemment, toute information ne doit pas être gardée mais, si tel avait été notre volonté, cela n'aurait pas été possible.

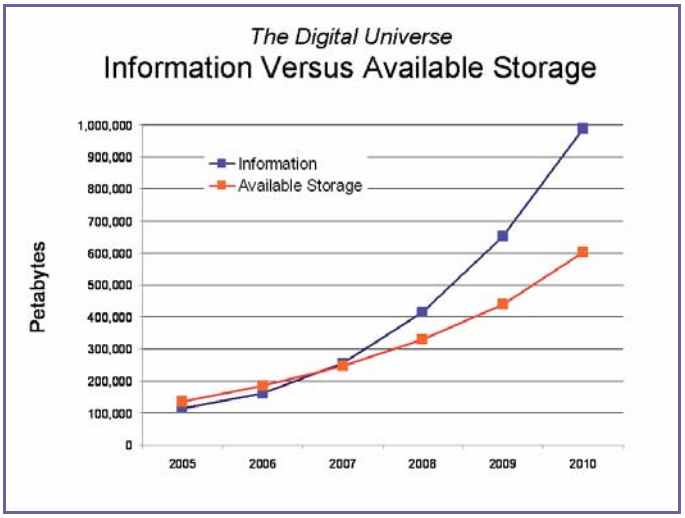


Figure 1 : Évolution de la quantité d'informations numériques créée et de la capacité de stockage disponible (source : étude IDC, 2007)

[2] GANTZ J. F. *et al.*, *The Diverse and Exploding Digital Universe. An IDC White Paper*, Mars 2007.
[3] À titre de comparaison, 2 pétaoctets correspondent à l'ensemble des fonds présents dans l'ensemble des bibliothèques universitaires américaines.

Quantité de données n'existent aujourd'hui plus que sous forme numérique, sans qu'il ne soit toujours possible de les transcrire dans le monde papier, et sont nécessaires au fonctionnement des institutions et entreprises qui ne peuvent assurer leurs missions sans ces données.

Le deuxième enjeu est la nécessité de conserver certaines de ces données sur de longues périodes, à l'instar des données « papier ».

À titre d'exemple, les réglementations imposent les périodes de conservation suivantes :

- Un dossier médical doit être conservé dans l'hôpital pendant au moins 30 ans à dater de la dernière consultation du patient[4].
- Les données nécessaires au calcul des pensions doivent être conservées jusqu'au décès du dernier ayant droit, ce qui signifie potentiellement plus de 100 ans.
- À l'heure de l'extension de la durée de vie des centrales nucléaires, il devient crucial d'en préserver l'information sur de très longues périodes afin de pouvoir permettre l'entretien, le fonctionnement mais également la démolition de ces centrales le moment venu. Par ailleurs, les informations sur les déchets radioactifs devront être conservées pendant plusieurs milliers d'années.
- Enfin, un exemple éclairant des enjeux est le monde de l'aéronautique qui a vu naître une association entre Airbus et Boeing en vue de définir un format standard pour la préservation des données tridimensionnelles nécessaires à la conception et la maintenance des avions.[5] Ces données doivent être conservées jusqu'au dernier vol de l'avion, soit plusieurs dizaines d'années.

Outre ces raisons réglementaires, la préservation des données numériques à long terme répond aussi à des besoins patrimoniaux et scientifiques. Ainsi, la conservation par la NASA des données issues des relevés satellites a permis *a posteriori* une meilleure compréhension du phénomène de diminution de la couche d'ozone, une fois ce phénomène appréhendé. En outre, toutes les expériences ne sont pas reproductibles. De même les données récoltées lors d'évènements exceptionnels (ex. éruption volcanique) ne peuvent pas être reproduites en cas de perte. Il est donc essentiel de conserver ces données pour en permettre l'étude.[6]

Un troisième enjeu découlant directement du besoin de conserver ces données est la nécessité de financer durablement la préservation de l'information. Étant donné le caractère émergent de la problématique, il est encore difficile aujourd'hui de chiffrer son coût. Une manière d'aborder ce point est de réfléchir en termes d'impact économique pour l'entreprise ou l'institution en cas de perte d'informations.

[4] AR 03/05/99 (Arrêté royal déterminant les conditions générales minimales auxquelles le dossier médical, visé à l'article 15 de la loi sur les hôpitaux, coordonnée le 7 août 1987, doit répondre) – Art 1 §3.
[5] Projet LOTAR - *Long Term Archiving and Retrieval of Digital Technical Product Data in the aerospace community* - http://www.prostep.org/en/project-groups/long-term-archiving-lotar.html (consulté le 27 octobre 2009).
[6] Exemple de perte : « *Earth observation data gathered by satellite in the 1970's were recently identified as critical for establishing a time line of changes in South America's fragile Amazon basin. The National Research Council reports the information is lost on the now-obsolete tapes on which the data were written.* », *Into the Future: On the Preservation of Knowledge in the Electronic Age. Discussion Guide*, CLIR Resources. http://www.clir.org/pubs/film/future/discussion.html#discuss (consulté le 26/01/2010).

À titre d'exemples, on peut citer les frais suivants (frais juridiques et amendes) qui résultent de l'indisponibilité de l'information :

- Pour répondre aux nombreuses demandes d'information de la justice américaine, la firme CISCO a dû régulièrement faire appel à des consultants externes pour un montant total estimé à 273 M $.[7]
- Philip Morris a été condamné à une amende de 2,75 M $ à la suite de la perte d'e-mails nécessaires dans le cadre d'un procès.[8]
- Qualcomm Attorneys a été condamné à une amende de 8 M $ pour des raisons similaires.[9]

Par ailleurs, le coût de la préservation comprend de nombreux postes de dépenses : ressources humaines, infrastructure matérielle (acquisition, maintenance, administration), logiciels (acquisition, développement et maintenance) et recours à des services tiers (horodatage, etc.).

Parallèlement, un quatrième enjeu consiste à développer et mettre en place des solutions logicielles globales permettant la gestion de la préservation de l'information vu qu'à l'heure actuelle, ces solutions n'existent pas. Les grands projets d'archivage à long terme de l'information réalisés aujourd'hui ont dû d'abord identifier les briques logicielles existantes et ont ensuite procédé à des développements et des intégrations.

Enfin, la préservation à long terme concerne plusieurs types d'information, présentant chacun des niveaux de complexité ou des contraintes différents :

- Informations créées par les entreprises et les administrations soumises à des contraintes légales et juridiques, accompagnées éventuellement d'une signature numérique.
- Documents issus de la numérisation à des fins patrimoniales (numérisation de tableaux, de documents d'archive, etc.) par des institutions culturelles. Il s'agit d'une nécessité si ces institutions veulent pérenniser leur investissement.
- Données scientifiques à structure et sémantique complexes (ex. données moléculaires).
- Sites web qui témoignent d'un phénomène sociétal. Leur conservation fournira des matériaux d'étude à destination des futurs chercheurs.

[7] When and Why Should You Start an e-Discovery Team?, *e-Discovery Team* [Blog], http://e-discoveryteam.com/2008/02/10/when-and-why-should-you-start-an-e-discovery-team/ (consulté le 14/01/2010)
[8] 2.75 Million Dollar Fine - Spoliation of E-mail, *Electronic Discovery and Evidence* [Blog], 28/07/2004, http://arkfeld.blogs.com/ede/2004/07/275_million_dol.html (consulté le 14/01/2010)
[9] Qualcomm Attorneys Hit With Multi-Million Dollar Sanctions for E-Discovery Violations, http://blawg.scottandscottllp.com/businessandtechnologylaw/2008/01/qualcomm_attorneys_hit_with_mu.html (consulté le 14/01/2010)

- Dans le cas où l'émulation est utilisée, données logicielles et leurs spécifications.
- …

Un cinquième enjeu est donc de trouver une solution en matière de préservation pour chacun de ces types d'information.

2. Terminologie et concepts

Comme d'autres domaines des sciences et technologies de l'information et de la communication, la préservation numérique n'échappe pas à un foisonnement de termes et de concepts utilisés de manière différente par les nombreux intervenants.

À titre d'exemple, ce que recouvre aujourd'hui le mot « archivage » varie, d'une part, pour les professionnels de la technique et des supports numériques qui se concentrent sur une conservation sécurisée et considèrent généralement l'archivage comme le transfert des informations sur des supports sécurisés[2] et, d'autre part, pour les professionnels de l'information qui mettent l'accent sur la gestion du cycle de vie des documents à conserver. Ceci n'est pas propre à la langue française ; le phénomène est le même avec *archiving*.[10]

Ces différences terminologiques et conceptuelles nuisent à une bonne compréhension mutuelle et rendent certaines discussions ardues.

Avant d'aller plus loin, il est donc important de définir les termes et concepts (2.1) que nous utiliserons couramment dans cette étude : stockage (2.1.1), archivage (2.1.2) et préservation à long terme (2.1.3). Ces définitions nous aideront en outre à définir le périmètre de l'étude. Nous soulignerons également la complémentarité existante entre l'archivage et la préservation (2.1.4).

Nous positionnerons ensuite les besoins en matière d'archivage et de préservation selon le cycle de vie d'un objet numérique (2.2), avant d'esquisser la manière dont un système d'archivage peut se positionner au sein d'une institution (2.3).

2.1. Définitions

2.1.1. Stockage

Le **stockage** est l'enregistrement d'une chaîne de bits sur un support physique. Il s'agit donc uniquement de traiter une information au niveau physique et binaire et de l'inscrire sur un support. Par analogie avec le monde papier, c'est une armoire dans laquelle on entasse les documents en ne les classant pas selon des critères intellectuels, mais uniquement selon des critères physiques.

Dès lors, déplacer des données d'un support *online* vers un support *off-line* (low cost type bande) relève du stockage et non de l'archivage. Le *Hierarchical*

[10] Voir la norme du Conseil international des archives, *Principes et exigences fonctionnelles pour l'archivage électronique*, 2008, www.ica.org - Module 3 - Recommandations et exigences fonctionnelles pour l'archivage des documents dans les applications métier, § 2.2.5 : « *It is essential to note that the term 'archiving' has different meanings in the records management and IT communities.* »

Storage Management (HSM) désigne un système permettant le stockage hiérarchique de données. Cela signifie qu'en fonction de divers critères (importance des données, rapidité d'accès nécessaire, etc.), les données sont réparties entre différents supports (bande magnétique, disque dur, mémoire vive, etc.) Le système transfère les données en fonction des critères établis (liés à une contrainte de coût du stockage par exemple).

2.1.2. Archivage

« Archiver consiste à prendre un objet et à le transférer sous certaines conditions dans un système qui permettra d'en assurer la préservation pendant un certain laps de temps avec toute la sécurité requise ».[11]

L'archivage se distingue donc du stockage par la notion de sécurité (la modification des documents est interdite, la destruction est également interdite sauf sous contrôle strict, toute action effectuée sur le document doit être tracée) afin que l'objet conserve sa valeur (notamment légale).

Deux conceptions différentes de l'archivage coexistent. Elles peuvent cependant toutes deux être rattachées à la définition ci-dessus et à la notion de sécurité.

1) Pour des professionnels de la technique et des supports de stockage, l'archivage consiste à transférer une information dans un système de stockage sécurisé. Une fois archivé, l'objet ne peut être ni modifié, ni supprimé (excepté via une procédure contrôlée). En outre, l'objet sera préservé aux niveaux physique et binaire (cf. 5.1).

2) En ce qui concerne les professionnels de l'information, l'archivage regroupe l'ensemble des actions, outils et méthodes mis en oeuvre pour réunir, identifier, sélectionner, classer et préserver (aux niveaux physique, binaire, logique et sémantique) des objets numériques, de manière sécurisée, dans le but de les exploiter et de les rendre accessibles dans le temps, que ce soit à titre de preuve (en cas d'obligations légales notamment ou de litiges) ou à titre informatif. À la conception précédente de l'archivage s'ajoute une organisation de l'information selon des critères métier et informationnels et non uniquement physiques. Ceci en vue d'identifier ce qui doit être archivé et de le rattacher à des règles de gestion (regroupement en dossiers, durée de conservation…). Dans ce cas-ci, l'objet doit non seulement être préservé aux niveaux physique et binaire, mais également aux niveaux logique et sémantique (cf. 5.1).

Dans le premier cas, l'objet archivé est vu sous les angles physique et binaire, tandis que dans le second, qui repose sur le premier, l'objet est appréhendé davantage pour sa valeur et son contenu informationnel.

La préservation (2.1.3) est nécessaire, aussi bien dans la première que dans la seconde conception.

2.1.3. Préservation à long terme

La **préservation** est le fait de maintenir matériellement les objets archivés en état dans le temps. Dans le monde numérique, la préservation peut se faire au niveau physique (support de stockage, chaîne de bits) et/ou au niveau logique (essentiellement format du fichier).

La norme OAIS[12] définit le **long terme** comme « *suffisamment long pour être soumis à l'impact des changements technologiques, y compris à la prise en*

[11] CHABIN M.-A., *Moreq2 et archivage sécurisé*, Fédération Nationale des Tiers de Confiance, 2009, p. 6.
[12] La norme OAIS sera abordée plus en détail au point 4.1. Il s'agit d'une norme centrale en matière de préservation à long terme.

compte de nouveaux supports et nouveaux formats de données ou à des changements de la communauté d'utilisateurs. » Dans la pratique, on parlera de **préservation à long terme dès que le délai de conservation sera supérieur à 5 ans.**

En conclusion, la préservation est incluse dans l'archivage. Il s'agira de veiller à ce que ce qui est archivé soit bien préservé.

L'objet de notre étude est la préservation numérique. Nous exposerons donc les stratégies existantes afin de préserver une information archivée aux niveaux physique, binaire et logique.

Les difficultés liées à l'interprétation de l'information (donc la couche sémantique – cf. 5.1) ne seront pas abordées car elles relèvent plus généralement des systèmes de méta-informations[13], auxquels appartiennent les métadonnées. Une étude approfondie de ces systèmes sortirait du cadre de cette étude.

2.1.4. Complémentarité entre archivage et préservation

Comme nous venons de le voir, l'archivage et la préservation sont complémentaires. À titre d'exemples, prenons deux questions fondamentales qui devront être posées lors de l'archivage de l'information :

- Que faut-il archiver ?
- Pour combien de temps ?

Ces deux questions illustrent la complémentarité entre archivage et préservation. Car même si ces questions relèvent de l'archivage, au sens des professionnels de l'information (sélection de l'information et règle de gestion), les réponses qui y seront apportées influenceront les stratégies de préservation à mettre en œuvre.

Dans le premier cas, prenons l'exemple de l'archivage à long terme d'une base de données telle que la DmfA. La *Déclaration multifonctionnelle / multifunctionele Aangifte* est une déclaration sociale que les employeurs doivent rentrer trimestriellement auprès de l'Office National de Sécurité Sociale (ONSS) afin de permettre le calcul et la perception des cotisations sociales ainsi que l'attribution de nombreux droits sociaux. Cette déclaration prend la forme d'un fichier XML composé de nombreux champs devant être remplis par les employeurs. Une fois réceptionnées par l'ONSS, les données contenues dans les fichiers sont transcrites dans une base de données en vue de permettre leur exploitation. Ces données, ainsi que toute modification apportée dans la base de données, doivent être gardées pour des raisons légales en cas de contestation devant les tribunaux ou d'évolution législative. Cependant, ces données ne sont que difficilement interprétables sans le dictionnaire trimestriel qui explicite la structure et la sémantique des termes, des champs et des valeurs. L'archivage de la base de données doit donc être pensé globalement, en y incluant le dictionnaire trimestriel. Il serait inutile d'en préserver les données s'il n'est plus possible à l'avenir de les interpréter. L'archivage devra donc veiller à la complétude de l'information archivée sous peine de rendre la préservation des données vaine.

[13] BOYDENS I., Les systèmes de méta-informations, *Techno* n°1, Section Recherches, Smals, 1997 (disponible sur http://documentation.smals.be/) ; *ead.*, *Les Dictionnaires de Données. Méthode, techniques et application pratique*, Delivrable, Section Recherches, Smals, juin 2000 (disponible sur http://documentation.smals.be/).

En ce qui concerne la seconde question, prenons l'exemple de deux documents MS Word devant respectivement être conservés pendant 3 et 50 ans (Figure 2). Les stratégies de préservation ne seront pas aussi importantes dans le premier cas que dans le second, pour lequel il faudra prévoir une conversion en un format pérenne du document, plusieurs migrations successives et un nombre plus important de métadonnées (sans oublier leur mise à jour).

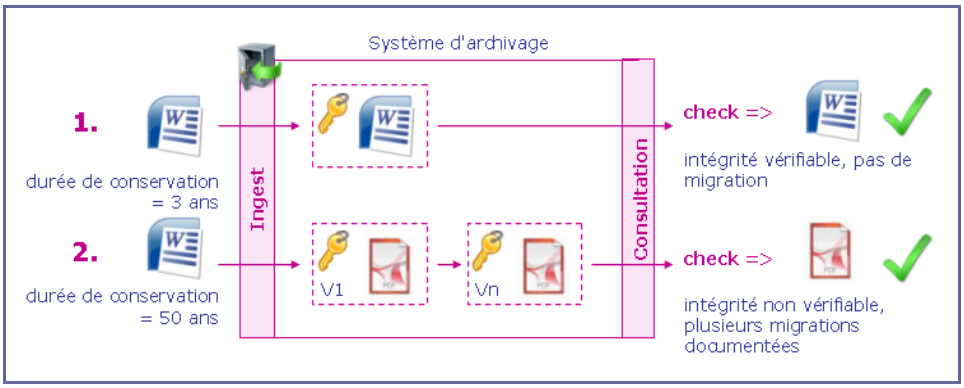


Figure 2 : Influence de la durée de conservation sur le processus de préservation

2.2. Cycle de vie d'une information numérique et ILM

Sur la base du dernier exemple (influence de la durée de conservation sur le processus de préservation), nous pouvons déduire qu'il faut idéalement appliquer les stratégies de préservation dès le début de l'archivage du document. Il n'est pas pertinent de convertir un document en un format pérenne 10 ans après qu'il ait été archivé. À ce moment-là, il sera sans doute déjà trop tard et cette transformation pourra se révéler plus complexe, voire impossible. Ensuite, il faudra appliquer les stratégies de préservation tout au long du cycle de vie de l'information numérique.

Généralement, on distingue trois états et étapes dans le cycle de vie d'une information, comme le montre la Figure 3.

Dans sa phase de production, le document est sous la responsabilité de l'utilisateur/créateur qui en a généralement une vision à court terme (répondre à une demande, entamer une procédure...). Cette information est en cours d'élaboration et doit encore être validée.[14]

Une fois validé, le *record*[15] représente une action de l'entreprise et engage donc sa responsabilité[16]. Selon la législation et la politique interne en matière d'archivage, l'information doit être gardée un certain nombre d'années avant de perdre toute valeur métier.

Elle peut alors être détruite ou conservée comme document historique et transférée vers les institutions en charge de sa conservation.

Selon ces trois états et étapes, certains critères seront adaptés. Par exemple, une demande officielle envoyée par une institution de la sécurité sociale à un employeur doit être archivée selon des critères (métier, technologiques...) précis

[14] Il s'agit par exemple des différents drafts rédigés pour une lettre à envoyer, un procès-verbal de réunion, un rapport, un contrat…
[15] La norme ISO 15489-1, *Information et documentation — «Records management» — Partie 1: Principes directeurs*, p. 9 définit les records comme des « documents créés, reçus et préservés à titre de preuve et d'information par une personne physique ou morale dans l'exercice de ses obligations légales ou la conduite de son activité ».
[16] Il s'agit ici de la version définitive archivée, envoyée au client, acceptée en réunion…

afin que l'on puisse la retrouver mais également en préserver la valeur juridique, au contraire des brouillons successifs qui pourront être éliminés.

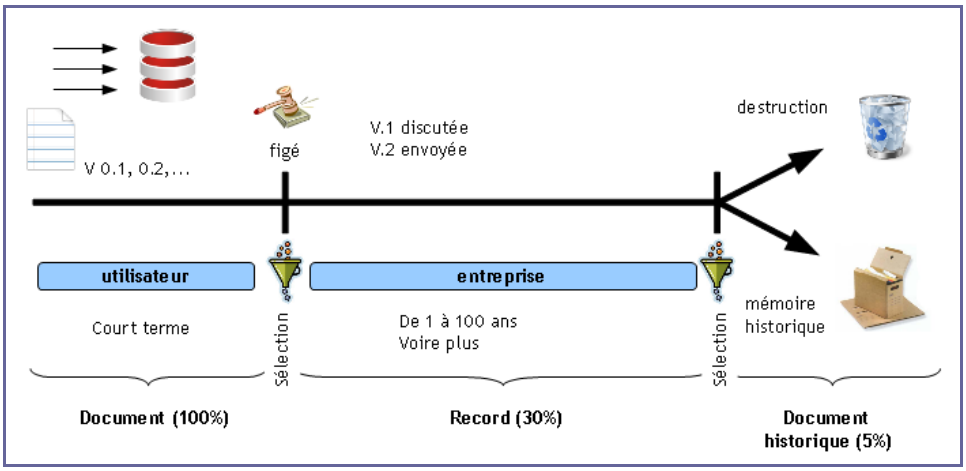


Figure 3 : Cycle de vie de l'information

Entre ces étapes, une sélection de l'information a lieu. En effet, tous les documents ne deviennent pas des *records*, et tous les *records* ne doivent pas être conservés pour des raisons patrimoniales et historiques. On estime généralement que seulement 5 à 10 % des documents seront conservés à cette fin.

Eu égard au cycle de vie de l'information, les besoins en matière d'archivage et de préservation se situent au stade du *record* et du document historique. Au cours de ces deux stades, l'information devra être archivée et préservée (Figure 4).

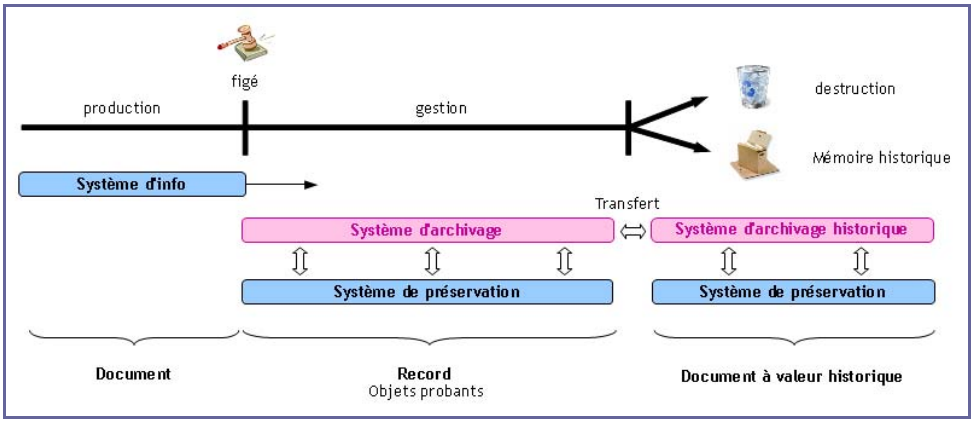


Figure 4 : Cycle de vie de l'information et besoins en matière d'archivage et de préservation

Pour cela, il est nécessaire de gérer le cycle de vie d'une information (*Information Lifecyle Management*) en tenant compte de ces trois étapes.

Issu originellement du monde du stockage, le concept d'*Information Lifecyle Management* (ILM) consistait à permettre une disponibilité optimale des informations en les stockant, en fonction de leur utilisation, sur des supports *on-line*, *near-line* ou *off-line*. Dans ce sens, la SNIA (*Storage Network Industry Association*) définit l'ILM comme « *l'ensemble des règles, processus, pratiques et outils que l'on déploie pour aligner au mieux la valeur métier des informations et l'infrastructure qui les héberge, et ce depuis leur création jusqu'à leur*

destruction ».[17] Le lien est ici direct entre la valeur métier de l'information et l'infrastructure [de stockage].

Cependant, sous l'influence de l'archivage électronique et des réglementations auxquelles sont de plus en plus soumises les données, l'ILM s'est progressivement étendu à d'autres aspects de l'information, tels que sa valeur (légale) et son contenu informationnel.

Le principe est d'avoir, dès la création de l'information, une vision sur son cycle de vie et d'aligner ainsi la valeur métier de l'information sur les besoins en termes d'archivage (stockage, sécurité, sauvegarde – *backup* –, préservation…) ainsi que de connaître le sort final réservé à l'information (destruction ou conservation historique).

2.3. Positionnement du système d'archivage

Même si les cas de figure sont légion, il nous semble intéressant de nous arrêter un moment sur la manière dont un système d'archivage se positionne au sein d'une institution. Même si la problématique reste la même, la manière de l'appréhender, de la mettre en œuvre ainsi que les systèmes informatiques et les fonctionnalités utilisées peuvent varier.

Pour cela, nous distinguerons deux cas de figure :

- un système d'archivage et de gestion de dossiers intégré à l'entreprise ;
- un système de dépôt légal, c'est-à-dire qu'en vertu d'une législation, un ensemble d'organisations doivent « déposer » leurs productions auprès d'une institution publique.[18]

Ces deux cas de figure sont ceux que l'on retrouve le plus dans les nombreuses publications traitant de la préservation à long terme.

Le premier cas présente un système d'archivage intégré au fonctionnement de l'entreprise (Figure 5) en vue d'y archiver les documents qu'elle produit et/ou reçoit. Ce système d'archivage ou de gestion de dossiers[19] déclenche des workflows selon l'arrivée de certains documents en vue de requérir leur validation et leur envoi ou la rédaction d'une réponse à un courrier reçu. Ces documents (la plupart du temps de type bureautique ou XML) sont stockés de manière sécurisée dans une base d'archives, les protégeant contre toute suppression et toute modification.

Dans de rares cas, il arrive que certains documents revêtent une valeur très importante. Ils sont alors archivés auprès d'un prestataire de services tiers qui met à la disposition de ses clients un coffre-fort sécurisé avec horodatage. Dans ce type de système, la préservation n'a lieu qu'au niveau physique. Les fonctionnalités de gestion sont limitées au transfert du document vers le coffre et à sa consultation. Les métadonnées sont limitées et consistent parfois uniquement en un index permettant de retrouver le document dans le sens où les prestataires de services ne peuvent généralement pas connaître le contenu des documents.

[17] CHABIN M.-A. *et al.*, *Dématérialisation et archivage électronique. Mise en œuvre de l'ILM*, Éd. Dunod, Paris, 2006, p. 19.
[18] Citons à ce titre l'exemple du dépôt légal de la Bibliothèque Royale de Belgique. Depuis le 1er janvier 1966, toutes les publications parues sur le territoire belge ainsi que toutes celles publiées à l'étranger par des auteurs belges doivent y être déposées à la section du Dépôt Légal.
[19] Voir HAUT H., Gestion électronique des documents, *Techno* n°15, Section Recherches, Smals, 1999 ; *id.*, Les composants d'une gestion de contenu, *Research Note* n°6, Section Recherches, Smals, 2004 ; *id.*, Workflow et gestion de dossiers, *Research Note* n°18, Section Recherches, Smals, 2008. Disponible via l'extranet de la sécurité sociale sur http://documentation.smals.be/index.htm).

Dans ce premier cas de figure, il existe de nombreux logiciels d'archivage[20] qui comportent l'ensemble des fonctionnalités nécessaires, excepté certaines fonctionnalités liées à la préservation (par exemple le contrôle systématique et automatique de l'état des supports, les métadonnées de préservation).

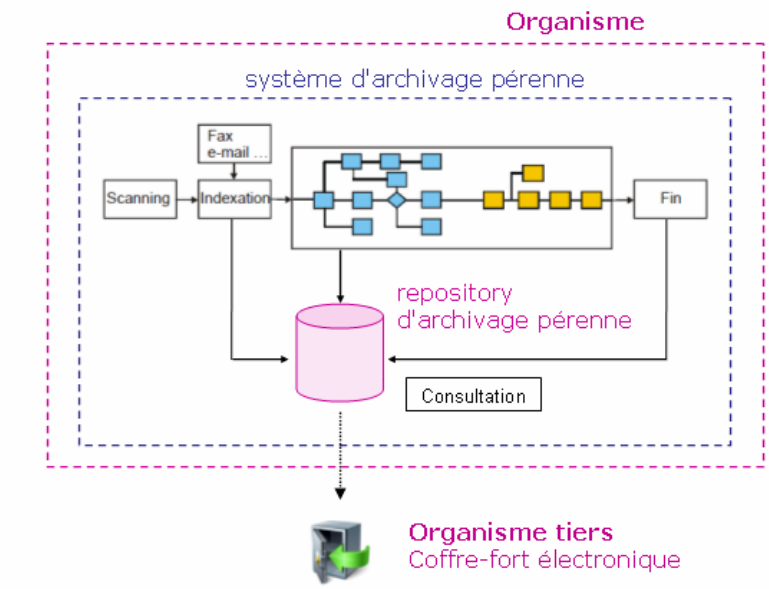


Figure 5 : Système d'archivage intégré à l'organisme

Le second cas est le dépôt légal (Figure 6). Il s'agit ici d'un dépôt d'archives, généralement public, c'est-à-dire géré par ou au nom d'une institution publique, que des producteurs d'informations, privés ou publics, utilisent pour transmettre leurs archives dans le cadre d'une contrainte légale ou règlementaire. Le système d'archivage est extérieur par rapport aux producteurs d'informations. C'est le cas des archives notariales, du futur système d'archivage de SIGeDIS pour les contrats de travail électroniques[21] (cf. chapitre 6) et du dépôt d'archives des Archives Générales du Royaume (AGR). Dans ce dernier cas, la loi sur les archives du 24 juin 1955[22] constitue le fondement de l'obligation de conservation des documents administratifs par les institutions publiques. L'État belge étant propriétaire de ces informations, ces institutions ont l'obligation de conserver leurs archives (papier et numériques) et de les transférer aux AGR pour des raisons patrimoniales et historiques.

Dans ce cas-ci, l'Archive[23] est généralement externe à l'entreprise qui produit les documents. L'Archive fixera généralement des règles à respecter (entre autres format de fichier, métadonnées obligatoires) et devra vérifier la conformité des documents qui lui sont transférés avec ces règles. Les aspects liés au transfert des fichiers sont donc plus complexes à mettre en œuvre que dans le premier cas, tandis que les fonctionnalités de workflow et d'édition sont éventuellement

[20] Ce sont les logiciels présents sur le marché de l'*Enterprise Content Management*.
[21] Nous évoquerons ce cas précis au chapitre 6. SIGeDIS a pour mission de gérer pour le compte d'autres institutions de sécurité sociale les informations relatives à la carrière des travailleurs.
[22] *Moniteur belge* du 12 août 1955.
[23] Dans la présente étude, par « Archive », nous entendons le service d'archivage, à savoir « l'organisation chargée de conserver l'information pour permettre [à ses utilisateurs] d'y accéder et de l'utiliser. » (HUC Cl. *et al.*, *L'archivage numérique à long terme. Les débuts de la maturité ?*, Éd. La documentation française, Paris, 2009, p. 263). Ce terme couvre l'ensemble des activités d'archivage (la mise en archive, la gestion des objets archivés, leur préservation et leur accès) et désigne tant le système d'archivage, que son organisation et ses gestionnaires.

moins importantes (du moins pour la production de l'information, des workflows d'indexation peuvent par contre être nécessaires).

De même, les utilisateurs de l'information archivée seront souvent externes à l'organisme. En outre, ce type de système d'archivage aura généralement des délais de conservation plus longs que dans le premier cas.

Dans ce cas-ci, il n'existe pas de logiciel paramétrable offrant l'ensemble des fonctionnalités nécessaires. Un travail plus important de développement sera donc nécessaire, sur la base des logiciels existants[24].

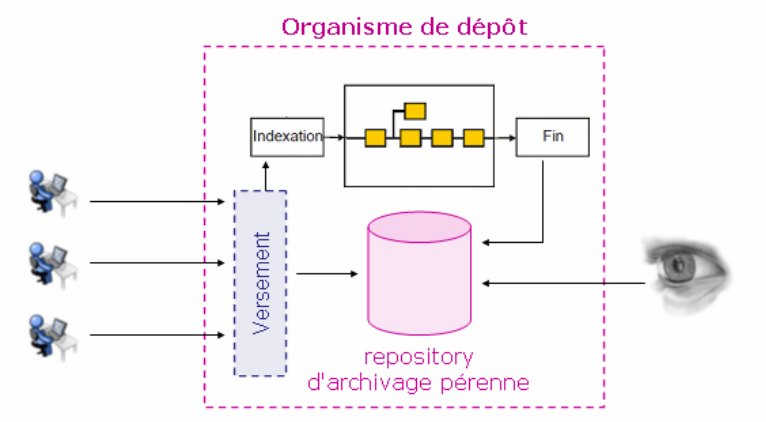


Figure 6 : Dépôt institutionnel

Au regard des publications dans le domaine, la plupart des projets de préservation numérique relèvent encore aujourd'hui du second cas de figure. L'intégration des fonctionnalités nécessaires (gestion des supports de stockage, validation de la conversion en formats pérennes, standard de métadonnées de préservation) pour la préservation dans les logiciels utilisés dans le premier cas de figure représente aujourd'hui un enjeu pour les éditeurs de logiciels comme le mentionnait un récent rapport du Butler Group[25].

[24] Deux logiciels open source sont aujourd'hui fortement utilisés pour ce type de système, principalement dans le monde des bibliothèques numériques : Fedora (http://www.fedora-commons.org/) et DSpace (http://www.dspace.org/).
[25] BURNETT S. et al, *Document and Records Management - Controlling Information Risk and Aiding Compliance*, Butler Group, 2008, p. 65-69.

3. Difficultés

Préserver l'information numérique à long terme n'est pas chose aisée (la Figure 7 met en évidence les difficultés liées à la préservation ainsi que les problématiques auxquelles elles se rattachent). Les quatre défis majeurs de la préservation à long terme sont :

- l'obsolescence technologique (3.1) ;
- la préservation de l'authenticité et de l'intégrité de l'information (3.2) ;
- l'organisation et le financement de la préservation (3.3) ;
- l'intelligibilité et l'exploitation de l'information (3.4).

Nous n'envisageons pas ici les catastrophes naturelles (incendies, inondations, tremblements de terre...) pour lesquelles il existe des précautions telles que les copies de sauvegarde et leur délocalisation, les précautions matérielles pour la construction des bâtiments et des data centers (solidité, imperméabilité, géolocalisation) en vue d'assurer la sécurité physique des supports.

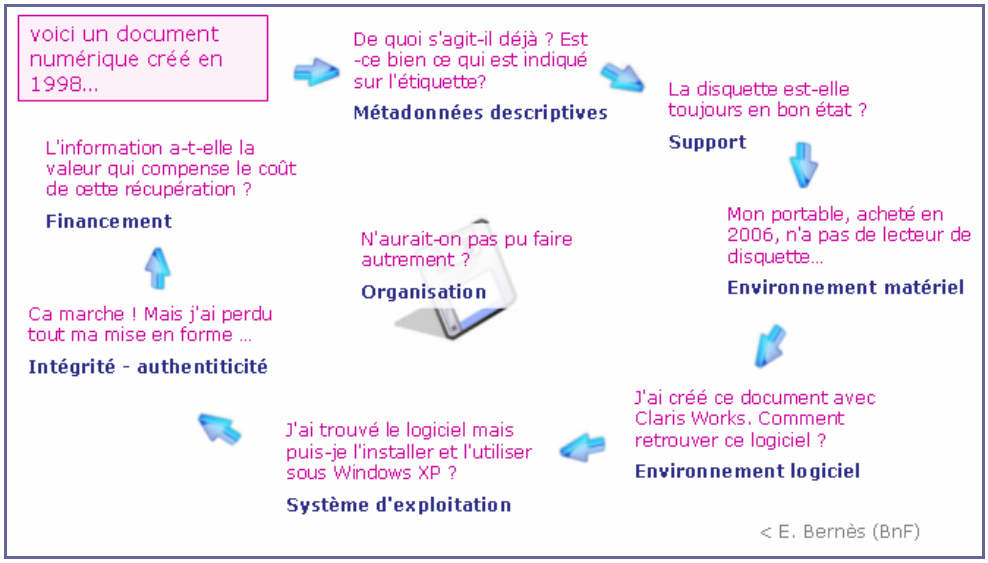


Figure 7 : Exemple de difficultés liées à la préservation

3.1. Obsolescence technologique

L'information numérique n'étant pas auto-explicative, elle n'est utilisable qu'à travers un ensemble de technologies matérielles et logicielles (Figure 8). Or, ces composants évoluent de manière asynchrone les uns par rapport aux autres,

menaçant ainsi de briser la chaîne d'accès et d'exploitation de l'information. Nous envisageons respectivement les problèmes liés au hardware et aux supports de stockage (3.1.1), aux composants logiciels (3.1.2) et aux formes et types d'encodage (3.1.3).

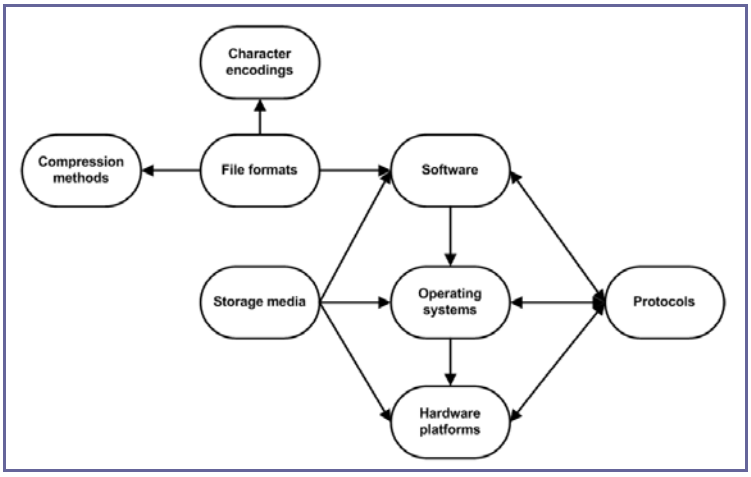


Figure 8 : Dépendances entre environnements matériel et logiciel

3.1.1. Hardware et supports de stockage

Les composants matériels d'un ordinateur se dégradent avec le temps, ce qui peut altérer l'information. De même, les supports de stockage et leurs périphériques de lecture se détériorent également, tant au niveau physique qu'au niveau technologique.

Pour être stockée, l'information doit être inscrite sous forme binaire sur un support. Or, cette inscription perd de sa qualité avec le temps. Cette perte de qualité n'est pas directement visible sur les supports numériques (au contraire de l'analogique) puisque les lecteurs sont capables de « combler » ces pertes. Cependant, à terme, la dégradation passe un seuil fatidique entraînant la perte définitive de l'information en tout ou en partie si aucune mesure n'a été prise.

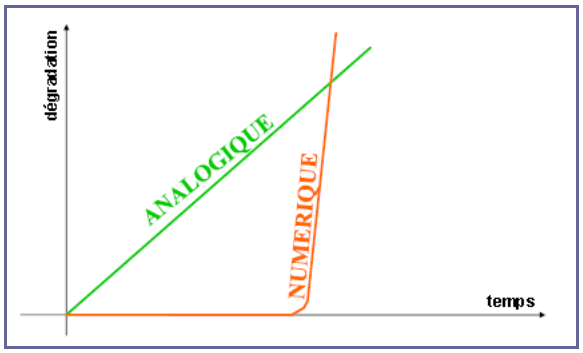


Figure 9 : Évolution comparée de la dégradation dans le temps des supports analogiques et numériques (Source : L. Duplouy – Séminaire PIN avril 09)

Les principales causes de dégradation physique des supports sont :

- une évolution chimique différente des matériaux composant le support et de leur association conduisant à leur dissociation. Cette évolution est intrinsèque au support et doit donc être prise en compte lors du choix ;
- l'inscription de l'information sur un matériau sensible à l'environnement (l'exposition répétée d'un CD à la lumière accélère sa dégénérescence) ;
- l'usure liée à une lecture et une manipulation répétées ;

- les mauvaises manipulations (dépôt de poussière, chocs).

Au niveau technologique aussi, la pérennité des supports de stockage est difficile à maintenir en raison de la nécessité de disposer d'un périphérique de lecture (associé à un driver logiciel). Or, ces éléments évoluent. Pensons aux difficultés rencontrées actuellement pour assurer la lecture des disquettes 3,5''. Certains PC ne parviennent pas à les lire bien qu'ils disposent d'un lecteur, tandis que les ordinateurs les plus récents n'en disposent même plus. Cependant, si l'évolution technologique est encore relativement prévisible, les facteurs liés au marché le sont beaucoup moins. Par exemple, pour sa solution d'archivage numérique, la Bibliothèque nationale de France avait choisi la solution de stockage de StorageTec, qui a été racheté un mois après par Sun. Il est difficile de prévoir la politique de Sun en ce qui concerne l'évolution des solutions appartenant anciennement à StorageTec.

Lors du choix d'un support de stockage, il est donc essentiel de tenir compte des risques liés au support et à la technologie, qui peuvent être contradictoires. Ainsi, d'un point de vue chimique et physique, le Minidisc est un support audio pérenne et stable dans le temps, mais il n'y a que Sony qui le maîtrise et le commercialise, ce qui représente donc un risque élevé pour la préservation à long terme.

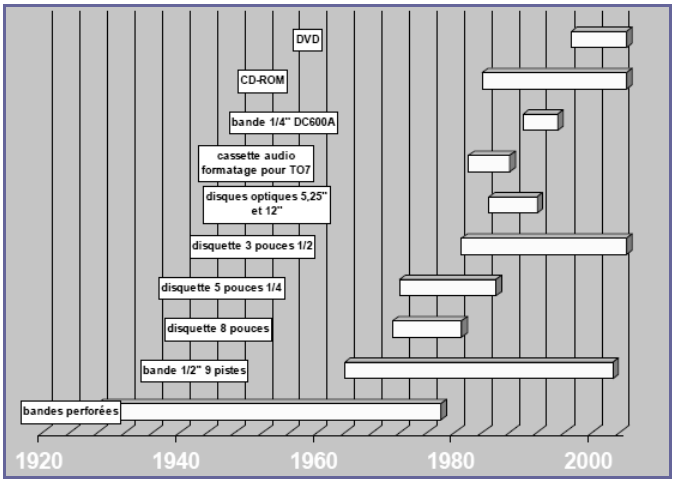


Figure 10 : Vie et mort des technologies de stockage (1920-2004) (Source : L. Duplouy – Séminaire PIN avril 09)

Par ailleurs, il se peut que tout ou partie du hardware d'un type d'ordinateur ne soit, à terme, plus supporté par la firme qui en assure la commercialisation (Figure 10).

Par exemple, au sein d'une institution, une application importante tournait sur un serveur Siemens qui n'était plus commercialisé ni supporté. En cas de crash système, il aurait donc fallu racheter ce serveur en occasion, pour autant qu'on eusse pu en trouver un.

3.1.2. Composants logiciels

Stocker l'information numérique sur un support n'est pas suffisant pour en assurer l'exploitation. Pour cela, il est nécessaire de disposer d'un logiciel capable d'accéder et de comprendre le code binaire. Or, ces composants logiciels sont eux aussi soumis à l'obsolescence pour plusieurs raisons.

La première tient à ce qu'il n'y a pas systématiquement de compatibilité ascendante[26] (*backward compatibilty*) entre les versions successives d'un même logiciel. Les informations créées avec une version x ne sont pas nécessairement exploitables sans transformation avec la version x+1.

Par exemple, le Centre national d'études spatiales (CNES) de Toulouse en France avait créé dès 1985 une abondante documentation à l'aide de la version Word de MS Office de l'époque. En raison d'un manque de compatibilité avec les versions successives du logiciel, l'ensemble de cette documentation a dû être réencodée manuellement à deux reprises, en 1990 sous Word 2 et en 1997 sous Word 95.

Aujourd'hui, l'évolution du marché entraîne une évolution rapide de l'offre logicielle. Ayant retenu certaines leçons du passé et grâce aux progrès technologiques, certains producteurs garantissent une compatibilité ascendante avec plusieurs versions antérieures. La migration des données est donc possible, mais cette opération peut induire des pertes de données ou encore la perte d'intégrité et d'authenticité, sauf en cas de compatibilité ascendante. Parallèlement, certains logiciels ne sont tout simplement plus suivis ou diffusés en raison d'une faillite ou d'un rachat de la société émettrice.

La deuxième raison de l'obsolescence des composantes logicielles est l'interaction entre les différentes couches logicielles et en particulier entre un logiciel applicatif et le système d'exploitation présent sur la machine. L'évolution des systèmes d'exploitation implique que certaines fonctionnalités du logiciel ne peuvent plus être utilisées, voire aucune puisque certains sont incompatibles avec un type de système d'exploitation bien précis.

Enfin, certains périphériques matériels sont étroitement liés aux composants logiciels. L'évolution de ces derniers peut donc empêcher leur utilisation (stockage, impression, accès...). Par exemple, certains types d'imprimantes étaient devenus incompatibles avec le mainframe utilisé par Smals, BS 2000. De même, certaines versions de Lotus Notes ne permettaient plus d'accéder à des informations stockées auparavant sur CD-ROM.[27]

3.1.3. Formats et types d'encodage

L'information numérique n'étant qu'une succession de 0 et de 1, il est nécessaire de définir un type d'encodage (ASCII, Unicode, CCITT G4…) et un format afin de pouvoir l'interpréter (Word, PDF, TIFF…) .

La question des formats est fortement liée à celle des composantes logicielles. En effet, le développement constant de nouveaux logiciels entraîne la modification des formats existants (par exemple le cas évoqué ci-dessus au CNES) ou la création de nouveaux formats non gérés par des applications similaires.

« Par exemple, le Département de la Défense américain a dû mettre en place une coûteuse opération de « migration » en vue de restaurer les fichiers relatifs aux coordonnées géographiques des bombes lâchées durant la guerre du Vietnam. Ces données, stockées à l'époque dans une base de données « propriétaire », étaient devenues illisibles : quelques années plus tard, plus aucun logiciel du marché ne permettait de les traiter sans générer d'incohérences. L'enjeu était considérable puisqu'il s'agissait de déminer et d'identifier les bombes qui, n'ayant pas explosé pendant ou après la guerre, pourraient occasionner de nouvelles victimes. La correction des incohérences fut possible grâce aux services des « National Archives » qui avaient conservé le format propriétaire

[26] Cette compatibilité signifie qu'un logiciel d'une version x+1 permet de traiter une application créée avec la version x de ce même logiciel.
[27] Voir BOYDENS I., La préservation à long terme de l'information numérique, *Techno* n°28, Section Recherches, Smals, 2004, p. 5, n. 19.

original et l'avaient périodiquement transféré dans des environnements plus modernes, tant sur le plan logiciel qu'au niveau des formats. »[28]

Ajoutons que la question est d'autant plus difficile que la structure de l'information est complexe : « *The more complex the digital resource, the greater the potential loss is likely to be. For example, interchanging the data held in geographical information system (GIS) databases and groupware databases could involve the loss of thousands of links that have taken years of effort to create and which represent the bulk of the value of the database* ».[29] De même, le nouveau standard OOXML ne fait pas moins de 6000 pages, rendant sa maîtrise extrêmement difficile. La production de norme est souvent soumise à un arbitrage entre la précision d'un format et sa complexité.

Idéalement, il faudrait disposer de standards internationaux afin de garantir l'homogénéité des interactions entre ces différentes composantes. Dans la pratique, ces évolutions sont souvent imprévisibles et tributaires des acteurs du marché.

3.2. Intégrité et authenticité

L'**intégrité** physique d'une donnée désigne « *l'état des données qui, lors de leur traitement, de leur conservation ou de leur transmission, ne subissent aucune altération ou destruction volontaire ou accidentelle* ».[30] Il s'agit d'un concept technique qui la rend formellement vérifiable via des algorithmes d'empreinte numérique (hash).

Le principal souci est que la préservation de l'intégrité d'une donnée est contradictoire avec la plupart des stratégies de préservation qui requièrent inévitablement une modification de l'information à long terme.

À titre d'exemple, prenons un document Word converti une première fois en PDF 1.4 (considéré il y a quelques années comme la meilleure solution d'archivage à long terme) et ensuite en PDF/A-1a (considéré aujourd'hui comme le nouveau format d'archivage à long terme).

En utilisant l'algorithme de calcul d'empreinte MD5, on obtient les empreintes suivantes :

- PDF 1.4 : db61abbd93e8fcd300eb75f8744363fa
- PDF/A-1a : 2afa0860a5aaa87a785ee0139433a582

> *There is an inherent paradox in digital preservation. On the one hand, it aims to deliver the past to the future in an unaltered, authentic state. On the other hand, doing so inevitably requires some alteration.*
>
> Thibodeau, 2002, p. 28.

L'intégrité peut donc être rompue à la suite de l'application d'une stratégie de préservation recommandée (dans ce cas-ci la migration de format).

L'intégrité de l'information ne peut donc être préservée que durant une période déterminée. Pour qu'une information reste exploitable dans le temps, il est donc important de pouvoir garantir son **authenticité**.

L'authenticité d'un objet numérique est définie comme « *the quality of genuineness and thrustworthiness of some digital materials, as being what they purport to be, either as an original object or as a reliable copy derived by fully documented processes from an original.* »[31]

L'authenticité peut donc être conférée tant à un original qu'à une copie (*reliable copy*) de celui-ci. Dans le cas contraire, il serait impossible de préserver

[28] BOYDENS I., La préservation à long terme de l'information numérique, *Techno* n°28, Section Recherches, Smals, 2004, p. 2.
[29] FEENEY M. (éd.), *Digital Culture : Maximising the Nation's Investment : a Synthesis of JISCO/NPO Studies on the Preservation of Electronic Materials*, Londres, National Preservation Office, 1999, p. 45.
[30] Article « Intégrité », *Wikipedia.fr*, consulté le 26/01/2010.
[31] *Guidelines for the preservation of Digital Heritage*, UNESCO, 2003, p. 157.

l'authenticité de l'objet à long terme, en raison des modifications que les stratégies de préservation nécessitent généralement (cf. infra). Par contre, ces modifications doivent être tracées (*documented processes*).

Il n'existe aucune possibilité de vérifier cette authenticité de manière formelle. On ne peut donc disposer que d'une présomption d'authenticité, présomption qui reposera sur un ensemble d'informations annexes : documentation disponible, actions entreprises sur l'objet numérique, traçabilité de ces actions, pertinence, provenance de l'objet… Préserver une information à long terme nécessite donc de récolter l'ensemble de ces informations complémentaires.

3.3. Organisation

L'organisation est considérée par les spécialistes du domaine comme une difficulté majeure (jusqu'à 80 % du problème selon les experts).

Un exemple existant dans la sécurité sociale belge est à cet égard révélateur. Au sein d'un organisme, une application comportait plusieurs milliers de dossiers représentant le cœur même de l'activité de l'organisme. Cependant, l'application, déjà ancienne, n'était plus supportée par la firme l'ayant installée et configurée. De surcroît, elle ne pouvait tourner que sur un serveur Siemens qui n'était également plus supporté et d'ailleurs encore peu commercialisé. Par ailleurs, les CD d'installation n'étaient plus disponibles en raison de plusieurs déménagements successifs de l'organisme. Un backup régulier du contenu du système était effectué, mais il n'était jamais vérifié puisque le système de sauvegarde indiquait une réussite de l'opération. Or, après plusieurs années, l'organisme s'est rendu compte que le backup ne fonctionnait absolument pas !

Dans ce cas-ci, le risque était donc majeur : application non supportée, dépendant d'un serveur non supporté et difficilement trouvable sur le marché, impossibilité de réinstaller l'application et backup inexistant.

En mars 2007, une agence de presse rapportait un exemple similaire, à savoir que l'*Alaska Departement of Revenue* avait perdu, à la suite d'une erreur de manipulation d'un technicien, les informations relatives à un fonds d'investissement d'une valeur de 38 milliards de dollars. 800.000 images et les documents annexes ont disparu. Vérification faite, il s'est avéré que le système de backup ne fonctionnait pas et le département a donc dû rescanner l'ensemble des documents, ce qui n'a pu être réalisé qu'avec l'aide de travailleurs engagés pour l'occasion et de nombreuses heures supplémentaires, y compris pendant les week-ends, pour une valeur de 220.700 $.[32]

3.4. Intelligibilité de l'information

Enfin, il faut également veiller à pouvoir comprendre l'information préservée à long terme d'un point de vue sémantique. Préserver une information sans avoir la capacité de la comprendre et de l'interpréter n'a aucun sens.

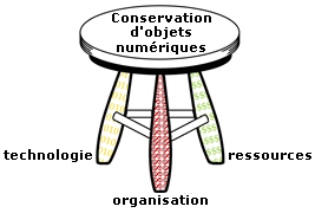
À titre d'exemple, lors d'une migration de version effectuée dans un organisme, un champ numérique d'une base de données a été migré alors que plus personne ne connaissait sa signification pour éviter tout risque. Mais on peut légitimement s'interroger sur l'utilité de préserver cette information.

Par ailleurs, lorsqu'on préserve à long terme, il convient de toujours pouvoir interpréter l'information selon sa signification originale puisque la sémantique de

[32] Alaska : The Land of Lost (Data), *CBS News*, 20 mars 2007 : http://www.cbsnews.com/stories/2007/03/20/tech/main2588756.shtml

l'information évolue dans le temps. Certes, des réinterprétations seront parfois nécessaires en raison de l'évolution du réel (par exemple une loi rétroactive sur le calcul des pensions ou la compréhension d'un phénomène climatique qui permet de réinterpréter des données antérieures), mais elles devront s'appuyer autant que possible sur le sens initial de l'information.

4. Stratégies organisationnelles



Comme nous venons de le voir, les problèmes d'organisation représentent une des principales difficultés en matière de préservation, notamment parce que les décideurs ne prennent pas toujours conscience des risques encourus. Et ceci d'autant plus que cette organisation doit perdurer dans le temps, tout en étant évidemment évolutive eu égard aux besoins et aux nécessités.

Il est donc nécessaire pour chaque organisme ou groupe d'organismes de :

- se doter d'un cadre institutionnel en matière de préservation qui traduit un engagement fort de la direction de veiller au problème et à prendre des mesures (organisationnelles[33], techniques et conceptuelles) en vue de le pallier. À titre d'exemple, pour se prémunir contre toute perte (physique ou logique) d'information, le groupe Total a mis en place une politique de conservation en août 2007 fixant des projets devant permettre une meilleure préservation des documents et un comité de préservation dirigé par le directeur juridique de Total et auxquels participent les directeurs informatique, audit et sécurité.[34]
- mettre en place une organisation incluant des services (réception, stockage, archivage…), des compétences et des responsabilités.
- définir les standards, les normes et les bonnes pratiques qui devront être utilisés et appliqués.

Dans tous les cas, avant le démarrage d'un projet d'archivage à long terme, il convient de déterminer certains éléments, dont plusieurs sont d'ailleurs communs à tout projet informatique :

- Identifier les responsables.
- Délimiter le scope : producteurs, utilisateurs et services souhaités.
- Examiner le contexte juridique et réglementaire, ce qui permettra entre autres d'identifier le niveau de sécurité à appliquer, les durées de conservation, les métadonnées métier à récolter, etc.
- Allouer les ressources.
- Comprendre le modèle OAIS (cf. 4.1).

Pour tout projet de ce type, il convient donc de commencer par l'étude de la norme OAIS (4.1), afin de pouvoir ensuite mettre en pace un référentiel fiable (4.2). Pour ce faire, il faudra mettre en place une organisation (4.3) dont nous

[33] Par exemple, désigner une personne ou un ensemble de personnes en charge de la préservation à long terme et qui sera impliqué(e) dans les projets où cette question est fondamentale.
[34] Table ronde « Archivage et conservation des bases de données » organisée par le CR2PA (Club des responsables de politiques et projets d'archivage), Paris, Total, la Défense, 14 janvier 2009.

montrerons un exemple au chapitre 6. Nous verrons ensuite différentes stratégies de mutualisation[35] en vue d'une diminution des coûts de la préservation (4.4).

4.1. Modèle OAIS

Le modèle OAIS (*Open Archival Information System*) est un modèle conceptuel destiné à la gestion, à l'archivage et à la préservation à long terme de documents numériques.[36] La mise au point de l'OAIS a été pilotée par le *Consultative Committee for Space Data Systems*, organe regroupant les grandes agences spatiales. L'OAIS est enregistré comme norme ISO sous la référence 14721:2003. Il est actuellement en cours de mise à jour.

Ce modèle constitue une référence décrivant dans les grandes lignes les fonctions, les responsabilités et l'organisation d'un système qui voudrait préserver de l'information numérique, à long terme, pour en garantir l'accès à une communauté d'utilisateurs identifiés. Le long terme est relatif et défini comme suffisamment long pour être soumis à l'impact des évolutions technologiques.

Le modèle OAIS est donc un **modèle conceptuel**. En cela :

- il ne fournit aucune indication de formats, schémas, règles ou techniques pour préserver les documents numériques ;
- il ne décrit pas les applications informatiques et techniques à mettre en œuvre, ni logicielles ni matérielles ;
- il ne donne pas de méthodologie concrète de réalisation d'un tel système (cahier des charges, workbook ou autre).

Ses objectifs sont de proposer :

- une **terminologie commune** afin d'éviter les problèmes de compréhension entre les différents acteurs impliqués dans un projet d'archivage numérique à long terme ;
- un **modèle fonctionnel** ;
- un **modèle de données**.

Bien que le modèle soit complexe et difficile à appréhender, il permet de comprendre l'ensemble de la problématique et des éléments à prendre en compte. Il convient ensuite à chaque institution d'élaborer sa propre organisation sur la base de ce modèle.

Une notion abondamment utilisée dans la norme est celle de paquet d'information (*Information Package* - IP). Le modèle repose sur l'idée que l'information constitue des paquets et que ces paquets ne sont pas les mêmes selon qu'on est en train de produire l'information, d'essayer de la préserver ou de la communiquer à un utilisateur. On a donc trois sortes de paquets :

- les **paquets de versement (Submission Information Package - SIP)** préparés par les producteurs à destination de l'Archive ;
- **les paquets d'archivage**[37] **(Archival Information Package - AIP)** transformés par l'Archive à partir du SIP dans une forme plus facile à préserver dans le temps ;

[35] La mutualisation est le fait de regrouper ou de mettre en commun des activités et des ressources en vue d'en retirer un avantage économique.
[36] Nous ne donnons ici qu'une présentation succincte du modèle. Pour une explication plus approfondie, il existe de très nombreuses présentations sur internet. Une explication pertinente se trouve entre autres dans HUC Cl. *et al.*, *L'archivage numérique à long terme. Les débuts de la maturité ?*, Éd. La documentation française, Paris, 2009.
[37] Dans les parties dédiées aux normes et aux standards, nous utilisons les termes consacrés. Si ce sens n'est pas conforme aux définitions initiales exposées au début de ce document (2.1), nous le précisons.

- les **paquets de diffusion (Dissemination Information Package - DIP)** transformés par l'Archive à partir de l'AIP dans une forme plus facile à communiquer, notamment sur le réseau.

Dans chaque paquet, à chaque stade, on trouvera des fichiers informatiques qui correspondent à l'objet ou au document qu'on veut conserver et des informations sur ce document, c'est-à-dire des métadonnées.

4.1.1. Le modèle fonctionnel

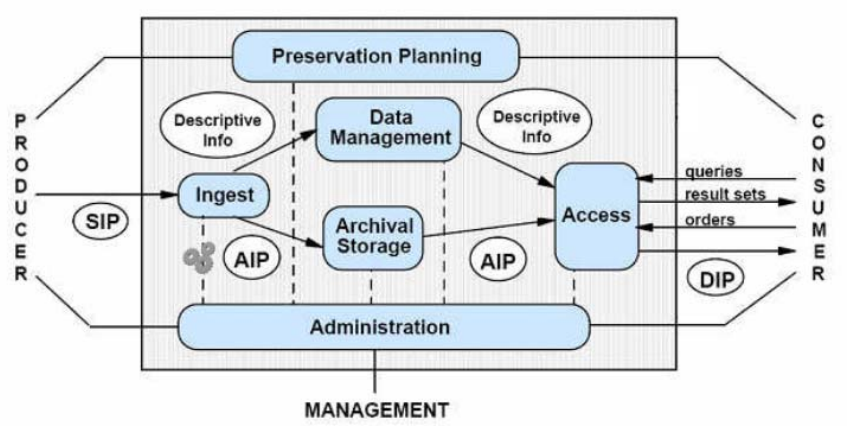


Figure 11 : OAIS - Modèle fonctionnel

Le modèle fonctionnel (Figure 11) de l'OAIS tente de couvrir, sur le plan fonctionnel, toutes les activités essentielles d'une Archive.

L'entité « **entrées** » (*Ingest*) reçoit, contrôle et valide les objets à archiver. Les objets eux-mêmes sont transmis à l'entité « stockage », tandis que les informations nécessaires à leur description et à leur gestion dans le temps sont transmises à l'entité « gestion des données ».

L'entité « **stockage** » (*Archival Storage*) assure la conservation physique des objets archivés. Elle tient les objets archivés à la disposition de l'entité « accès ». Conformément aux règles établies par l'entité « administration », elle prend en charge la réalisation des copies multiples et le renouvellement des supports anciens. Dans certains cas, l'entité « stockage » devra également être utilisée lors de l'entrée des objets numériques dans l'Archive, et ce pour des raisons légales (cf. étude de cas, chapitre 6).

L'entité « **gestion des données** » (*Data Management*) prend en charge la tenue à jour de toutes les informations internes nécessaires au système d'archivage. Elle fournit aux autres entités du système les informations descriptives des objets archivés (notamment à l'entité « accès ») et toutes les informations de gestion techniques et archivistiques nécessaires.

L'entité « **administration** » (*Administration*) assure la coordination générale du système. Elle en établit les règles internes. Elle veille à la qualité globale du service rendu et à son amélioration. Elle rend compte au management.

L'entité « **planification de la pérennisation** » (*Preservation Planning*) est la cellule de veille et de planification du système. Elle « écoute » l'environnement extérieur et émet des recommandations en vue de procéder aux évolutions nécessaires, notamment aux évolutions technologiques. Elle prépare et planifie ces évolutions. Elle est également responsable du suivi des changements qui peuvent s'opérer dans la « communauté d'utilisateurs » cible en vue de garantir que le service d'accès reste conforme aux attentes nouvelles des utilisateurs.

L'entité « **accès** » (*Access*) regroupe tous les services qui sont en interface directe avec les utilisateurs. Outre les fonctions de contrôle d'accès, il s'agit principalement de permettre aux utilisateurs de rechercher dans le « catalogue » des objets archivés et de leur fournir les objets dont ils passent commande.

4.1.2. Le modèle de données

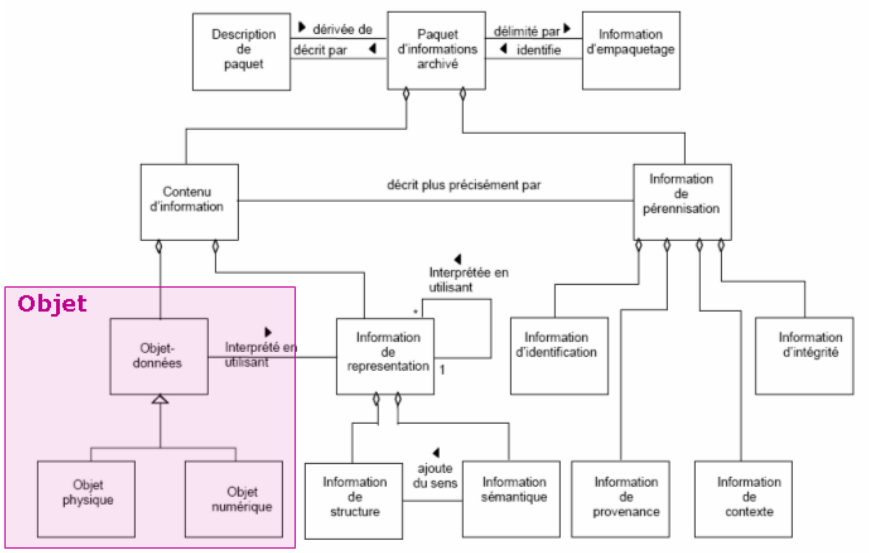


Figure 12 : OAIS - Modèle de données

Le modèle de données (Figure 12) met en évidence les différentes catégories d'informations complémentaires qui doivent être réunies en vue d'assurer l'archivage, la préservation et l'intelligibilité totale de l'objet archivé.

Les objets pris en compte par le modèle OAIS ne sont pas nécessairement numériques. Ils peuvent également être des objets physiques, même si le modèle est essentiellement tourné vers le monde numérique.

Au premier stade, *l'objet(-données)* n'est qu'une suite de caractères binaires non interprétable. Cette interprétation ne pourra se faire qu'à l'aide d'une *information de représentation*, qui permettra de passer des bits constituant l'objet numérique au contenu informationnel de cet objet.

Cette *information de représentation* est elle-même constituée de deux éléments. *L'information de structure* décrit la structure de la séquence binaire. Concrètement, elle précise le format de l'objet-donnée et la manière dont les séquences binaires doivent être interprétées (chiffre, nombre, texte, image, couleur, etc.). Le second élément est *l'information sémantique*. En cela, le modèle souligne la nécessité d'assurer l'intelligibilité de l'information. Concrètement, cette information précise que le fichier PDF conservé est un contrat de travail de Monsieur X et datant de telle année, etc. Un exemple existant d'information sémantique concerne les glossaires DmfA qui ont pour objectif de documenter la structure et la manière dont il faut interpréter les champs constituant la déclaration trimestrielle de cotisations sociales que les employeurs doivent fournir à l'Office National de Sécurité Sociale.[38]

[38] Pour de plus amples informations sur la DmfA et les glossaires, voir BOYDENS I., Edition spéciale – Un retour riche d'expériences, *Techno* n°26, Section Recherches, Smals, 2003 (accessible via l'extranet de la sécurité sociale sur http://documentation.smals.be/index.htm). Les glossaires sont accessibles sur le portail de la sécurité sociale belge : https://www.socialsecurity.be/site_fr/Applics/dmfa/index.htm

L'information de représentation étant basée sur des éléments existants (documents utilisant des chaînes de caractère, etc.), elle a sa propre information de représentation. Il y a donc une récursivité de *l'information de représentation*. Dans la pratique, il conviendra à chacun de déterminer jusqu'à quel niveau de récursivité il conviendra d'aller (cf. le cas extrême présentée dans la Figure 13).

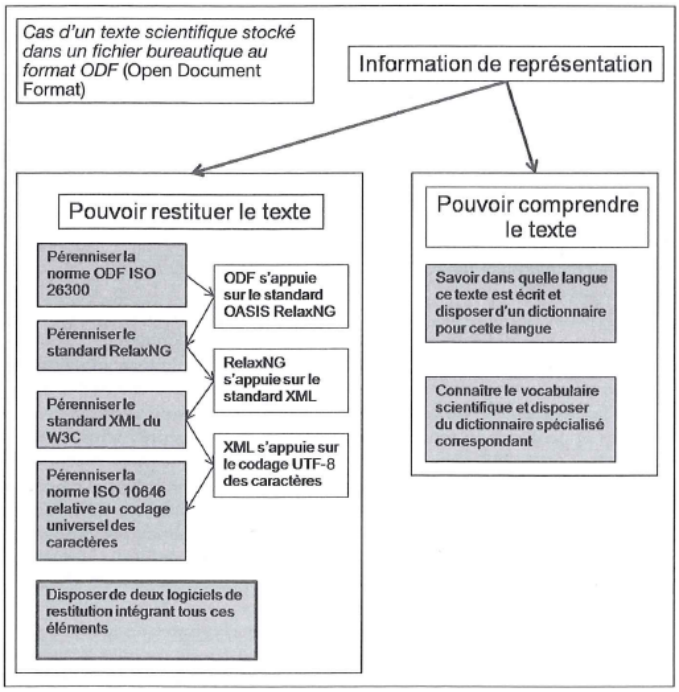


Figure 13 : Exemple de récursivité de l'information de représentation (Source : Huc Cl. et al.*, 2009)*

L'objet-donnée, associé à son *information de représentation*, en fait un *contenu d'information*, à savoir un objet numérique lisible et compréhensible.

Un *contenu d'information* seul ne suffit pas pour que l'Archive puisse remplir l'ensemble de ses missions et assurer la préservation à long terme des objets numériques. Il doit être associé à une *information de pérennisation*. Elle est requise pour préserver le contenu d'information, assurer qu'il soit clairement identifié et appréhender l'environnement de création de l'objet. *L'information de pérennisation* se subdivise en quatre catégories d'informations :

- *Provenance* : indique l'origine ou la source de l'objet, trace tous les changements intervenus depuis sa création et identifie les intervenants de ces changements.
- *Contexte* : décrit les liens entre le contenu d'information archivé et son environnement, les raisons de sa création, son rapport avec d'autres objets archivés. Elle permet de rattacher l'objet à un ou plusieurs ensembles (dossier, fond, etc.).
- *Identification* : permet d'attribuer à l'objet un identifiant unique.
- *Intégrité* : fournit un mécanisme ou un dispositif protecteur pour prémunir l'objet contre toute altération non documentée. Par exemple, il peut s'agir d'un checksum de l'objet.

Un *contenu d'information*, associé à son *information de pérennisation*, constitue un *objet numérique archivé* (AIP). Afin que les utilisateurs puissent rechercher et retrouver les objets, il faut leur adjoindre une *information de description*. Cette information n'est pas nouvelle. Elle est extraite des informations décrites dans l'AIP, elle en constitue un sous-ensemble utilisé pour l'interaction avec les

utilisateurs. Cette information ne fait pas partie de l'AIP. Cela signifie qu'elle pourra évoluer et être modifiée sans que cela n'ait un impact sur l'objet archivé.

Enfin, *l'information d'empaquetage* permettra d'identifier et de relier les différents composants d'un paquet d'information. Par exemple, un contrat de travail et ses avenants, les fiches de paie d'une année, un livre et ses différentes pages numérisées ou encore la référence de l'information de pérennisation du contenu d'information, etc.

Comme on le voit, le modèle OAIS est donc un modèle conceptuel, complexe et pouvant paraître lourd à mettre en œuvre. Il ne faut pas perdre de vue que son objectif est de donner une vue aussi complète que possible sur les éléments à mettre en œuvre pour archiver des objets numériques à long terme.

Ce modèle constitue donc un point de départ essentiel pour toute réflexion et projet sur ce domaine. Il convient ensuite de déterminer la manière concrète dont il sera mis en œuvre.

4.2. Construire un référentiel fiable

L'objecttif d'un projet d'archivage à long terme est de mettre à la disposition de ses « clients » une plateforme d'archivage numérique fiable, capable de préserver les données à long terme. Cela signifie que les utilisateurs doivent pouvoir faire confiance à l'Archive quant à sa capacité à préserver les informations de manière pérenne.

De fait, si une entreprise souhaite développer en interne ou confier à un prestataire externe l'archivage et la préservation à long terme de ses contrats électroniques, les responsables de l'entreprise doivent disposer de garanties suffisantes eu égard aux enjeux.

Afin d'offrir aux institutions des directives en matière de construction d'un référentiel fiable et dans un souhait toujours plus grand d'évoluer vers une certification des Archives, diverses méthodes d'audit ont été développées. De manière générale, ces méthodes présentent plus ou moins les mêmes éléments que l'on peut regrouper en trois grandes catégories :

- l'organisationnel, c'est-à-dire vérifier qu'on dispose de la bonne gestion, l'organisation, les compétences, la viabilité financière pour accomplir cette tâche, et qu'on garantit la transparence qui permet d'établir la confiance ;
- la gestion des objets numériques : vérifier qu'on se donne les moyens d'avoir des objets pouvant être préservés dans le temps et de collecter toutes les informations (c'est-à-dire les métadonnées) nécessaires pour leur préservation, la maintenance des accès et des performances, et la surveillance de l'environnement technologique par la veille ;
- l'infrastructure et la sécurité, bref les moyens techniques de la préservation.

Ces dernières années, divers modèles ont été élaborés et proposés :

- Mai 2002 - *Trusted Digital Repositories: Attributes and Responsibilities* (RLG et OCLC)[39]
- Août 2005 - *An Audit Checklist for the Certification of Trusted Digital Repositories* (RLG et NARA)[40]

[39] Disponible à l'adresse suivante : http://www.oclc.org/research/activities/past/rlg/trustedrep/ (consulté le 08/12/09)

- Décembre 2006 - *Catalog of Criteria for Trusted Digital Repositories* (Nestor Working Group)[41]
- Février 2007 - *TRAC – Trustworthy Repositories Audit and Certification* (OCLC)[42]
- Février 2007 - *DRAMBORA – Digital Repository Audit Method Based on Risk Assessment* (DCC et DPE)[43]

Enfin, un travail en vue d'établir une norme de certification des archives est en cours d'élaboration au CCSDS (*The Consultative Committee for Space Data Systems* – organisme à la base de l'OAIS).

DRAMBORA est actuellement le modèle le plus avancé. Il reprend les acquis des modèles précédents et propose une méthode d'autoévaluation basée sur la gestion des risques. Elle propose un audit interne en 6 étapes :

1. Définir le mandat et le domaine d'application de l'Archive.
2. Identifier les activités et les fonds de l'Archive.
3. Identifier les risques et les vulnérabilités par rapport au mandat, aux activités et aux fonds.
4. Évaluer et calculer les risques.
5. Définir les actions en vue de diminuer les risques.
6. Rédiger un rapport d'autoévaluation.

Bien que ces modèles d'évaluation restent relativement théoriques, ils ont été construits sur la base d'exemples concrets. Leur prise en compte et leur application permettent de disposer d'une liste, aussi exhaustive que possible, des critères à étudier et des points d'attention.

4.3. Mise en place d'une organisation

Pour mettre en place une organisation encadrant la préservation, une méthode consiste à découper cette organisation en services indépendants.

Il n'existe évidemment pas de découpe parfaite et chaque organisation devra déterminer l'organisation qu'elle souhaite mettre en place. Cette organisation sera fonction du système mis en oeuvre et des éventuelles stratégies de coopération envisagées. Par exemple, si plusieurs organismes coopèrent en vue de développer un système commun pour le versement des objets, il sera pertinent de confier la gestion et la responsabilité de ces versements à une seule et même entité. Ce sera également le cas de projets d'archivage à long terme dans le cadre de dépôt institutionnel (archives notariales, actes légaux…). Un exemple d'organisation fondé sur une découpe en services indépendants sera présenté dans le cadre de l'étude de cas exposée au chapitre 6.

Si la préservation fait partie intégrante d'un système d'archivage d'entreprise (2.3), les aspects organisationnels liés à la préservation feront partie des tâches de l'équipe qui gère le système.

[40] Disponible à l'adresse suivante : http://www.worldcat.org/arcviewer/1/OCC/2007/08/08/0000070511/viewer/file2433.html (consulté le 08/12/09)
[41] Disponible à l'adresse suivante : http://files.d-nb.de/nestor/materialien/nestor_mat_08-eng.pdf (consulté le 08/12/09)
[42] Disponible à l'adresse suivante : http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf (consulté le 08/12/09)
[43] Site du projet : http://www.repositoryaudit.eu/

4.4. Coûts et stratégies de mutualisation

4.4.1. Coût de la préservation

Comme nous l'avons mentionné au début de ce document, le coût de la préservation est difficilement chiffrable. Un projet britannique de bibliothèque numérique, LIFE[44], tente depuis plusieurs années de créer un framework en vue d'identifier les principales catégories de coût (Figure 14).

Acquisition	Ingest	Bit-stream Preservation	Content Preservation	Access
Selection	Quality Assurance	Repository Administration	Preservation Watch	Access Provision
Submission Agreement	Metadata	Storage Provision	Preservation Planning	Access Control
IPR & Licensing	Deposit	Refreshment	Preservation Action	User Support
Ordering & Invoicing	Holdings Update	Backup	Re-ingest	
Obtaining	Reference Linking	Inspection	Disposal	
Check-in				

Figure 14 : LIFE Framework pour l'évaluation des coûts liés à la préservation

En outre, pour chacune de ces catégories, il est nécessaire de distinguer les coûts liés aux ressources humaines, à l'infrastructure à mettre en place (locaux, machines, réseaux) et au développement de logiciels.[45]

Tant ces éléments que ceux mis en avant par le projet LIFE montrent qu'il est extrêmement difficile de calculer le coût d'un projet de préservation.

Cependant, dans le cadre d'une réduction des coûts, plusieurs stratégies de coopération sont possibles. Ces stratégies reposent sur le constat que de nombreux éléments sont communs quels que soient les contextes et les domaines d'application.

À titre d'exemple, la Direction des Archives de France (DAF) a fait réaliser une étude sur les coûts d'un système d'archivage électronique pour les services publics d'archives.[46] Deux hypothèses ont été étudiées plus en détail. La première consiste à développer un système pour chaque producteur d'archives (services d'archives des ministères, archives départementales, etc.). La seconde envisage la création d'une plateforme centralisée dans laquelle chaque producteur pourrait verser ses archives. Rien qu'au niveau du coût du stockage au téraoctet,

[44] http://www.life.ac.uk/
[45] Un aperçu des coûts liés à la préservation et une évaluation de leur impact est présenté sous forme de tableau dans HUC Cl. *et al.*, *L'archivage numérique à long terme. Les débuts de la maturité ?*, Éd. La documentation française, Paris, 2009, p. 170-171.
[46] Les deux parties de l'étude (« Études de l'existant et des besoins » et « Études des scénarios ») sont disponibles sur le site de la DAF : http://www.archivesdefrance.culture.gouv.fr/gerer/archives-electroniques/platesformes/ (consulté le 18/01/2010).

la différence est significative. Le To coûte 19.560 € dans le premier cas et seulement 1.003 € dans le second.

Il en ressort également que, dès lors que les volumes archivés augmentent, les économies d'échelle sont importantes, de l'ordre d'un facteur 20 dans le cas des Archives de France.

Il est donc intéressant de voir quelles stratégies de mutualisation peuvent être mises en œuvre. Nous les illustrons à l'aide du modèle OAIS.

4.4.2. Stratégies de mutualisation

Un premier cas consisterait à mutualiser les infrastructure de stockage (Figure 15) des différentes Archives (dans ce cas-ci l'Archive 1 et l'Archive 2), comprenant le contrôle de l'état des supports, leur remplacement, un système de gestion hiérarchique des fichiers, etc. Pour rappel, l'entité « Stockage » ne gère que des séquences binaires non interprétables. Pour que le service de stockage fonctionne de manière satisfaisante, l'application de standards est requise au niveau des interfaces techniques entre le service de stockage d'une part et les entités « Ingest » ou « Accès » d'autre part.

Toutes les autres entités restent indépendantes et relèvent donc de la responsabilité individuelle des Archives. Cependant, tout changement dans ces entités qui aurait un impact sur l'entité « Stockage » ne peut se faire sans une concertation et l'accord des autres Archives.

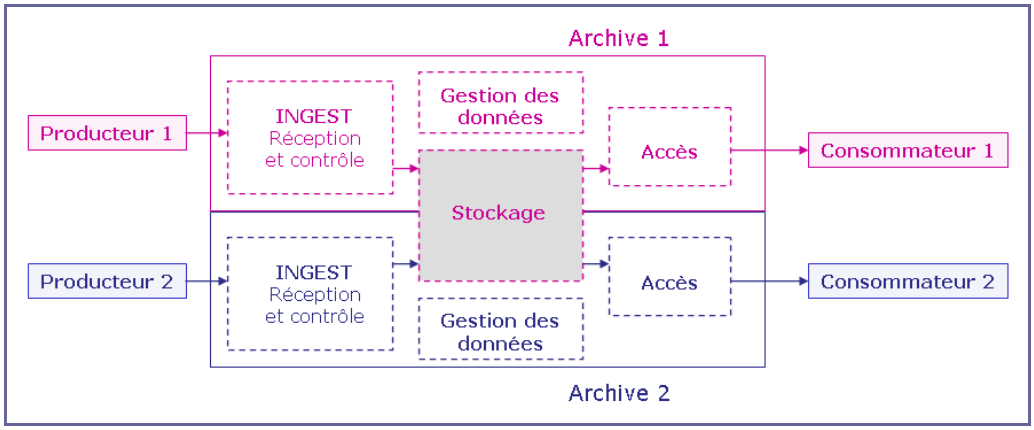


Figure 15 : Mutualisation de l'entité « Stockage »

Une seconde stratégie consisterait à mettre en commun également l'entité « Gestion des données » (Figure 16). Le service « Gestion des données » est essentiellement un service de support à la gestion des objets numériques archivés. Il administre la base de données contenant les informations de description et les informations système. Il crée et gère les schémas et les tables permettant de gérer les objets numériques. Par exemple, lorsqu'un utilisateur formule une requête via l'entité « Accès », il réceptionne la requête, effectue la recherche et renvoie les résultats à l'entité « Accès » qui les présentera à l'utilisateur.

Dans ce cas-ci, l'application de standards est requise au niveau des interfaces techniques entre les services mutualisés et les entités « Ingest » ou « Accès ». L'interdépendance entre les Archives est plus forte que dans le premier cas et une étroite concertation s'impose donc entre les Archives pour l'évolution de la plateforme.

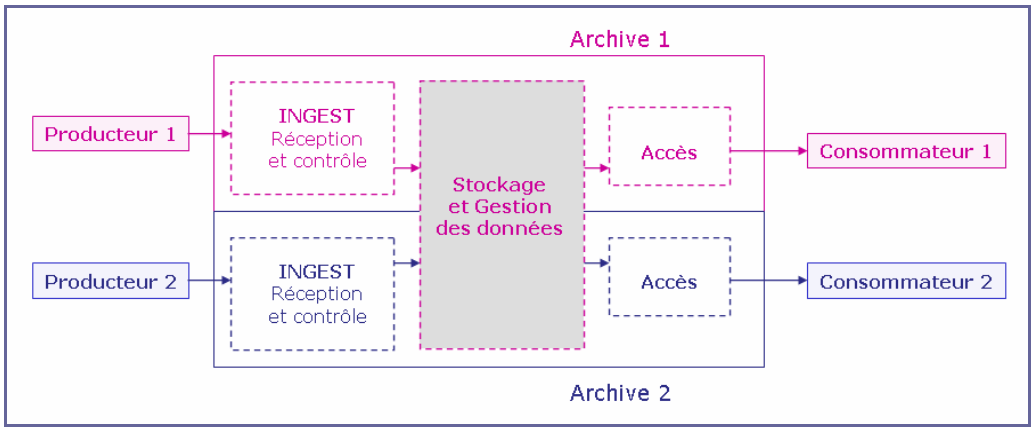


Figure 16 : Mutualisation de l'entité « Stockage » et « Gestion des données »

Enfin, une dernière solution consiste à développer un système centralisé permettant à tous les producteurs d'y archiver leurs objets numériques de manière pérenne (Figure 17).

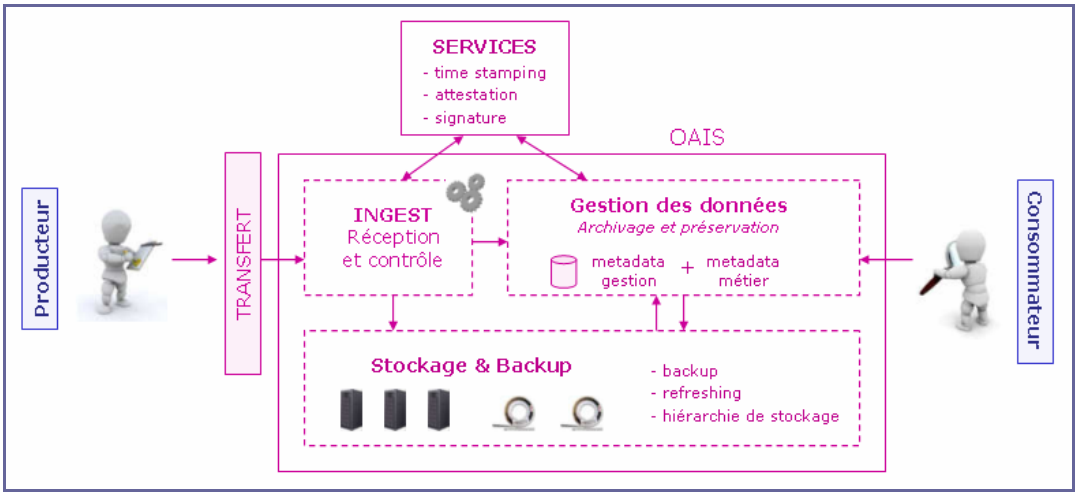


Figure 17 : Plateforme d'archivage centralisée

Les producteurs peuvent verser leurs objets numériques au moyen d'un service de versement. Celui-ci réceptionne les objets et vérifie qu'ils sont conformes aux exigences fixées par l'Archive et accompagnés des informations demandées (réception et contrôle des SIP). Un feedback est donné aux producteurs concernant la réception, le contrôle et l'acceptation/le refus des objets. Si nécessaire, le service fait appel à des services tiers en vue d'apposer un horodatage sur la transaction, de délivrer une attestation et de vérifier la conformité des signatures.

Une fois acceptés, les objets sont stockés par le service de stockage qui assure les mêmes tâches que celles décrites dans les cas précédents de mutualisation.

Enfin, un service d'archivage et de préservation gère le classement, le cycle de vie, l'accès et la préservation des objets archivés (AIP). Il veille à l'accessibilité des objets dans le temps en appliquant les stratégies de préservation adéquates. Il faut veiller à ce que l'Archive, selon la mission qu'elle a reçue, dispose des droits et autorisations adéquats. Par exemple, si la mission est de préserver au minimum la couche logique des données, elle doit disposer de l'autorisation d'effectuer des migrations (cf. point 5.5), éventuellement en gardant l'objet original. Dans la négative, l'Archive ne pourra pas remplir sa mission.

L'Archive doit définir les métadonnées communes aux différents types d'objets en vue de la préservation (cf. point 5.6). En effet, un fichier TIFF, qu'il représente un document contractuel numérisé, un tableau de peintre du XVII^{e} siècle ou issu d'un autre contexte, aura globalement une grande majorité de catégories de métadonnées de préservation communes. De même pour un document PDF signé ou non, quel que soit son contenu, il aura techniquement les mêmes catégories de métadonnées. Parallèlement cependant, il est nécessaire de laisser la possibilité à chaque producteur de définir, en accord avec l'Archive et en fonction de la communauté d'utilisateurs visée, un certain nombre de métadonnées spécifiques.

Un exemple de mutualisation est la *Florida Digital Archive*[47] aux États-Unis, lancée en novembre 2005 par le *Florida Center for Library Automation* dont l'objectif était d'offrir aux universités de l'État de Floride une plateforme de préservation afin qu'elles puissent y conserver leurs documents. Basée sur le modèle OAIS, la plateforme concentre ses activités sur des missions d'archivage et de préservation. L'accès aux documents par les utilisateurs s'effectue via les différents logiciels des universités. La plateforme est donc transparente pour les utilisateurs. À cet égard, la gestion des métadonnées descriptives (qui permettent de rechercher le document – cf. 5.6) est laissée aux institutions clientes de la plateforme.

La *Florida Digital Archive* s'engage vis-à-vis de ses « clients » à :

- mettre tout en œuvre pour maintenir à jour une plateforme fiable (un référentiel numérique fiable – cf. 4.2) ;
- recevoir les documents selon une procédure de versement négociée (métadonnées à fournir, respect des spécifications en matière de formats) et en contrôler la conformité (validation du format par exemple – cf. 5.4.2) ;
- préserver physiquement le document original sans le modifier, afin d'en garder l'intégrité et l'authenticité ;
- préserver logiquement, via migration de format, le document en maintenant à la disposition des utilisateurs une version à jour, consultable par les logiciels existant sur le marché pour les formats acceptés ;
- fournir l'original et la copie à jour sur demande.

Deux niveaux de préservation sont disponibles : « bit level », c'est-à-dire physique, et « full », c'est-à-dire physique et logique.

Le projet mutualise les infrastructures de stockage et d'archivage. Il dispose par ailleurs d'une procédure et d'un système standard pour les versements qui peut être étendu à des nouveaux besoins (principalement nouveaux formats à préserver). Tous les aspects liés à la consultation et à la recherche de l'information doivent être gérés et introduits si nécessaires dans le système (informations descriptives, durée de conservation…) par l'université cliente.

Dans le cadre de la mutualisation, diverses pistes complémentaires peuvent être mises en œuvre :

- Réutiliser des infrastructures matérielles et applicatives déjà présentes dans l'institution ou dans une institution associée. Beaucoup d'organisations ont tendance à refaire des infrastructures qui existent

[47] CAPLAN P., The Florida Digital Archive and DAITSS : a working preservation repository based on format migration, *International Journal of Digital Library*, 6, 2007, p. 305-311. Site du projet : http://fclaweb.fcla.edu/FDA_landing_page (consulté le 19/01/2010).

déjà au niveau de l'entreprise, par exemple dans un autre département, ou qu'une institution partenaire a déjà mise en œuvre.

- Le partage d'une infrastructure de stockage permet des gains significatifs.
- À l'heure actuelle, aucun logiciel ne permet de répondre à l'ensemble des besoins. Cependant, dans le domaine de l'archivage et de l'*Enterprise Content Management*, des éditeurs tels que IBM (Filenet), EMC, Global 360, Open Text, Oracle et Alfresco (open source) fournissent des solutions qui peuvent servir de point de départ. Ces logiciels traînent généralement dans leur sillage toute une série de logiciels annexes permettant de convertir certains objets en des formats pérennes, de vérifier la conformité des objets, etc.
- Par ailleurs, les expériences montrent que le versement est l'un des services les plus problématiques, principalement parce qu'il faut y gérer tous les objets venant de l'extérieur et pas toujours conformes aux exigences. Des réponses adéquates doivent être apportées dans tous les cas de figure. Dès lors, pour diminuer les tâches à remplir, il convient d'intervenir autant que faire se peut lors de la production de l'information, pour autant que cela soit possible. Cette intervention peut prendre la forme d'exigences à respecter, de recommandations, de bonnes pratiques, etc. Le but est de diminuer les erreurs à gérer. Parallèlement, il faut automatiser au maximum les contrôles, vu que les contrôles manuels sont coûteux en temps et en ressources (humaines et budgétaires).
- Enfin, le développement de standards de versement, de modèles de données, etc. est évidemment une manière de diminuer les coûts puisqu'il faut tenir compte de moins de situations différentes.

4.5. Synthèse

Comme nous l'avons vu, l'organisation à mettre en œuvre pour soutenir un projet de préservation à long terme est extrêmement importante et représente selon les spécialistes 80 % du problème.

En premier lieu et à l'instar de nombreux autres domaines, un engagement fort et clair de la direction est nécessaire pour mener à bien ce type de projet, entre autres en raison de la faible visibilité des gains que peut apporter une bonne politique de préservation à long terme.

Le modèle OAIS, norme ISO reconnue internationalement, est un modèle fondamental. Tout projet ou réflexion sur la préservation et l'archivage à long terme de données doit débuter par son étude et sa compréhension. Par ailleurs, la maîtrise de cette norme est importante pour pouvoir dialoguer avec certains fournisseurs qui se disent de plus en plus « OAIS compliant ». Une méconnaissance du modèle ne permet pas de jeter un regard critique sur cette affirmation.

Il est important que les institutions qui souhaitent préserver leurs données aient la certitude qu'elles seront conservées de manière fiable et efficace. Ceci notamment en vue de convaincre les responsables d'investir les budgets nécessaires. À cette fin, divers modèles ont été élaborés. Même s'ils ne sont en rien un gage de réussite, leur utilisation en vue de guider les réflexions et les analyses en amont et ensuite d'autoévaluer le repository en aval peut s'avérer extrêmement utile.

Il ne faut cependant pas oublier que le modèle OAIS est un modèle conceptuel visant à donner un aperçu de l'ensemble de la problématique et qu'il semble donc parfois très (trop) théorique. Chacun devra donc réfléchir à la manière de concrétiser ce modèle, via une découpe des tâches et des responsabilités au sein de différentes équipes, plus ou moins autonomes.

Enfin, la préservation implique des coûts parfois considérables. Aussi, dans la mesure du possible, est-il intéressant d'analyser les possibilités de mutualisation des efforts en vue de les diminuer. Nous avons vu que différents modèles existent

5. Stratégies techniques et conceptuelles

L'organisation mise en place doit encadrer les stratégies techniques et conceptuelles qui permettront la préservation à long terme de l'information numérique. Avant de passer à l'examen de ces stratégies, nous ferons un bref rappel des différents composants et couches d'une information numérique (5.1). Ensuite, nous verrons une typologie des stratégies et la manière dont elles se positionnent par rapport aux composants et couches de l'information numérique (5.2).

Sur cette base, nous examinerons ensuite les stratégies techniques et conceptuelles envisageables, en commençant par les stratégies reconnues aujourd'hui comme opérationnelles, à savoir la gestion des supports de stockage (5.3), la gestion des formats (5.4), la migration (5.5) et les métadonnées (5.6). Nous examinerons ensuite deux techniques moins éprouvées, i.e. l'encapsulation (5.7) et l'émulation (5.8).

5.1. Modèles en couches

Une information numérique est composée de différentes couches (Figure 18) :

- Une couche physique qui est le type de support utilisé pour enregistrer l'information. Généralement, le support est lié à une technologie d'écriture définissant un format d'enregistrement : linéaire pour les bandes LTO, hélicoïdale pour les bandes AIT, par secteur dans le cas des disques durs…

  Exemple : disque dur, bande

- Une couche binaire qui représente l'information sous forme de 0 et de 1, regroupés en octets, et ayant un sens de lecture (de gauche à droite ou de droite à gauche)…

  Exemple : suite de 0 et de 1

- Une couche logique qui interprète l'information binaire au moyen d'un format de codage pour la rendre exploitable.

  Exemple : fichier PDF, version 1.4, images de 300 dpi…

- Une couche sémantique qui décrit l'information par rapport à un domaine d'application du réel.

  Exemple : contrat de travail électronique de ABC en provenance de l'entreprise XYZ.

Cette information numérique sera ensuite exploitée au moyen de composants *hardware* et *software*.

Dès lors, la préservation à long terme de l'information numérique consistera à préserver ces différentes couches, voire, en fonction des stratégies, à préserver ces composants matériels et logiciels.

5.2. Typologie et positionnement des stratégies

Il n'existe aucune stratégie unique. Il sera donc indispensable de combiner les stratégies pour mettre en place une préservation de l'information à long terme.

Dans l'ensemble des stratégies disponibles, on peut distinguer deux grands types d'approche :

- La première, « Preserve Object », consiste à préserver l'objet numérique afin qu'il puisse être exploité dans un environnement matériel et logiciel à jour, tout en veillant à en préserver les caractéristiques essentielles.

  Stratégie représentative : la migration

- La seconde, « Preserve Technology », centre les opérations de préservation sur le matériel et les logiciels nécessaires pour reproduire l'objet numérique, sans le modifier.

  Stratégie représentative : l'émulation

La seconde approche n'est guère opérationnelle aujourd'hui. Nous aborderons cependant l'émulation, parfois utilisée dans des contextes spécifiques.

Les stratégies que nous allons étudier dans le cadre de cette étude (Figure 18) auront chacune leur rôle en matière de préservation, en préservant une ou plusieurs couches de l'information.

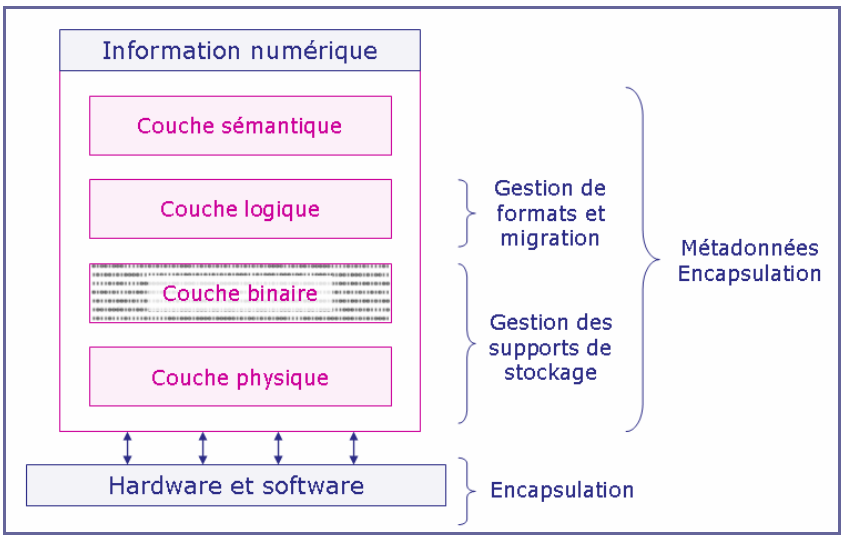


Figure 18 : Modèle en couches et positionnement des stratégies

La gestion des supports de stockage gèrera les couches physique et binaire. La gestion des formats et la migration permettront de préserver la couche logique tandis que les métadonnées permettront la préservation de l'ensemble des couches, en offrant une base nécessaire et indispensable aux autres stratégies. L'encapsulation peut offrir une aide pour ces différents couches.

L'émulation permettra de préserver les composants matériels et logiciels nécessaires pour rendre l'objet numérique. Comme indiqué ci-dessus, cette stratégie n'est pas mûre à l'heure actuelle et ne doit être utilisée que lorsqu'il n'existe aucune alternative.

5.3. Migration et gestion des supports de stockage

La gestion des supports de stockage se focalise sur la préservation des couches physique et binaire de l'information (5.1) en luttant contre la détérioration des supports de stockage dans le temps.

Cette détérioration, due tant à l'évolution intrinsèque des composants chimiques du support qu'à sa manipulation et à ses conditions de stockage, entraîne un affaiblissement du signal binaire enregistré. Au-delà d'une certaine mesure, il n'est plus possible de distinguer les 0 et les 1. Autrement dit, l'information est perdue.

Pour pallier ce problème, il est nécessaire de migrer périodiquement les informations d'un support de stockage à un autre, plus jeune ou plus robuste. Cette migration peut également résulter d'une migration périodique ou du souhait d'évoluer vers une autre technologie.

La principale difficulté sera de déterminer le moment opportun pour réaliser cette migration. Une migration trop précoce entraînera un surcoût, tandis qu'une migration trop tardive impliquera une perte d'informations.

Trois méthodes peuvent être mises en oeuvre :

1. **choisir des supports de stockage pérennes**, plus adéquats pour la préservation à long terme ;
2. agir sur les **conditions de stockage et de manipulation des supports** afin d'allonger leur durée de vie ;
3. **contrôler régulièrement l'état des supports** pour déterminer le moment de migration.

5.3.1. Le choix des supports de stockage

Certains supports présentent des qualités qui les rendent plus favorables pour une préservation à long terme de l'information.

Pour évaluer la capacité d'un support eu égard à la préservation à long terme, il faut prendre en compte plusieurs critères :

- La **stabilité intrinsèque** du support (inertie chimique du support) et sa robustesse (résistance aux agressions) : tout support de stockage est composé de différentes couches chimiques qui se dissocient avec le temps et mettent en péril l'information qui y est inscrite.
- Une **large diffusion de la technologie** implique un marché commercialement plus étendu et donc suivi avec davantage d'attention. Il est préférable que cette technologie soit maîtrisée et commercialisée par plusieurs constructeurs pour limiter le risque en cas de faillite ou d'abandon de la technologie à la suite d'une décision stratégique du constructeur ou d'un rachat. La technologie doit être basée sur des normes publiques.
- **L'existence d'outils de contrôle du support** permet de suivre sa dégradation progressive et d'intervenir à temps (cf. 5.3.3).

- La **simplicité des opérations de recopie** est également un facteur qui permet de limiter les risques d'erreurs de manipulation ou de mise en œuvre.
- La **protection contre l'effacement accidentel** est indispensable en vue d'éviter qu'une personne n'efface inconsciemment ou non des informations stratégiques pour l'institution.

Notons toutefois que le choix de ces supports dépend fortement de l'environnement dans lequel on se trouve. Dans le cas de données devant être disponibles dans des délais très courts, les supports magnétiques, tels que les disques durs via des système RAID[48], restent les plus utilisés, notamment dans les *data centers*.

Pour des raisons de coût, ce type de support et ce type d'environnement ne sont pas toujours possibles. Dès lors, d'autres types de support near ou off-line[49] peuvent être utilisés, tels que les bandes magnétiques qui sont encore abondamment utilisées aujourd'hui ou encore les CD et les DVD en cas de volume peu élevé.

	CD-R	DVD-R	disque dur	bande 9940	bande LTO
stabilité	mauvaise	mauvaise	moyenne	bonne	bonne
diffusion	bonne	bonne	moyenne	mauvaise	bonne
outils de contrôle	oui	oui mais cher	oui	oui mais cher	oui mais cher
opérations de recopie	facile	facile	facile	facile	facile

Figure 19 : Exemple d'évaluation de divers types de support de stockage

Comme on peut le voir (Figure 19), dans le cadre d'une préservation de l'information à long terme, l'usage des CD et DVD n'est guère recommandé. Cependant, ils peuvent être utiles. Une étude réalisée par la Direction des Archives de France[50] montre qu'il faut choisir de préférence des CD possédant les caractéristiques suivantes :

- film métallique réfléchissant en or de préférence ;
- couche de colorant organique phtalocyanine ou azoïque ;
- capacité de 74 ou 80 minutes ;
- modèle dédié à la conservation de longue durée ;
- conditionnement en boîtiers rigides.

En ce qui concerne les bandes, la technologie *Linear-Tape Open* (LTO) est plus répandue que la technologie des bandes 9940 de Sun/StorageTek.

[48] La technologie RAID (acronyme de *Redundant Array of Independent Disks*, soit « Ensemble redondant de disques indépendants ») permet de constituer une unité de stockage à partir de plusieurs disques durs. L'unité ainsi créée (appelée **grappe**) a donc une grande tolérance aux pannes (haute disponibilité) ou une plus grande capacité/vitesse d'écriture. La répartition des données sur plusieurs disques durs permet donc d'en augmenter la sécurité et de fiabiliser les services associés.

[49] Un support off-line est par exemple un bande de stockage, tandis qu'un support near-line désigne par exemple des cassettes magnétiques ou des disques optiques regroupés dans des juke-boxes.

[50] Direction des Archives de France, *Note d'information DITN/RES/2006/008 - Objet : résultats de l'étude sur des CD-R et des graveurs du marché*, 2006. Disponible à l'adresse suivante : http://www.archivesdefrance.culture. gouv.fr/gerer/archives-electroniques/stockage/ (consulté le 18/01/2010)

5.3.2. Conditions de stockage

Les conditions de stockage influent grandement sur la longévité des supports. Il est donc indispensable d'appliquer les recommandations existantes, entre autres celles relatives au taux d'humidité et à la température optimale pour les différents supports utilisés. L'application de ces recommandations permet d'augmenter la durée de vie des supports et ainsi de limiter leur remplacement ainsi que les pertes d'informations.

Media	*Access storage (allows immediate access and playback)*		*Long-term storage (preserves the media as long as possible)*	
	Temperature (°C)	*Relative Humidity (%)*	*Temperature (°C)*	*Relative Humidity (%)*
Magnetic tape cassettes 12.7 mm	18 to 24	45 to 55	18 to 22	35 to 45
Magnetic tape cartridges	10 to 45	20 to 80	18 to 22	35 to 45
Magnetic tape 4 & 8mm helical scan	5 to 45	20 to 80	5 to 32	20 to 60
Magnetic tape	Room ambient (15 to 23), maximum variation 4°	Room ambient (25-75), maximum variation 20%	As low as 5°, maximum variation 4°	As low as 20%, maximum variation 10 %
CD-ROM	10 to 50	10 to 80	18 to 22	35 to 45
DVD-ROM			4 to 20	20 to 50

Source : Harvey 2005, p. 121

Ces conditions sont généralement disponibles sur les sites des constructeurs ou sur demande. Par ailleurs, la plupart des institutions d'archives publient des recommandations et font régulièrement des communications sur les conditions de stockage à respecter.

5.3.3. Contrôle de l'état des supports

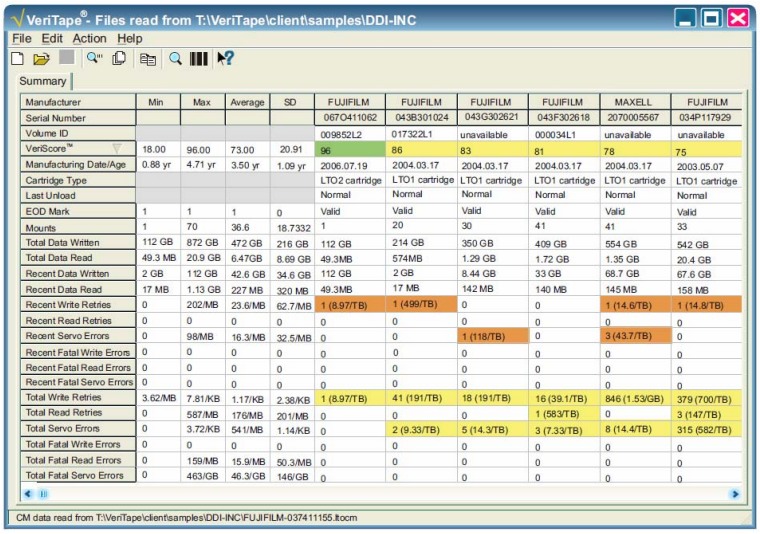
Comme nous l'avons dit précédemment, la dégradation progressive des supports nécessite leur remplacement périodique, la difficulté étant d'opérer ce remplacement au moment opportun. Une migration trop précoce entraînera des coûts trop élevés, tandis qu'une migration trop tardive impliquera une perte d'information. Afin d'évaluer au mieux le moment opportun, il est recommandé de contrôler périodiquement l'état des supports.

La méthode consiste à :

1. identifier les éléments qui sont suffisants pour évaluer l'état du support et seront donc analysés ;
2. déterminer pour chaque élément les valeurs qui seront obtenues lors de l'enregistrement de l'information sur le support et les valeurs seuil à partir desquelles une procédure de remplacement devra être exécutée ;
3. déterminer un échantillon à contrôler : il est impossible de contrôler tous les supports en cas de grande quantité. Il convient donc de regrouper les supports de chaque type en lots selon des caractéristiques à définir telles que le constructeur, leur date d'achat, leur lot, leur fréquence d'utilisation… et de sélectionner un échantillon représentatif de l'ensemble ;
4. déterminer la périodicité du contrôle. À titre d'exemple, dans le cas de CD, on estime que la périodicité peut être comprise entre un an et demi

et cinq ans, selon la qualité du CD, de la gravure, des conditions de conservation…

Les méthodes d'évaluation selon les types de support évoluent dans le temps et il convient donc de se tenir informé de ces évolutions. À titre d'exemple, on considère généralement que le calcul du taux d'erreurs à la lecture est suffisant pour juger de l'état d'un CD. Ce principe avait d'ailleurs été repris dans la norme AFNOR Z 42-011-2 définissant la qualité des CD-R à la gravure et durant la conservation. Ce principe est actuellement remis en cause.[51] Il n'existe cependant pas encore de modèle alternatif.



Manufacturer	Min	Max	Average	SD	FUJIFILM	FUJIFILM	FUJIFILM	FUJIFILM	MAXELL	FUJIFILM
Serial Number					067O411062	043B301024	043G302621	043F302618	2070005567	034P117929
Volume ID					009852L2	017322L1	unavailable	000034L1	unavailable	unavailable
VeriScore™	18.00	96.00	73.00	20.91	96	86	83	81	78	75
Manufacturing Date/Age	0.88 yr	4.71 yr	3.50 yr	1.09 yr	2006.07.19	2004.03.17	2004.03.17	2004.03.17	2004.03.17	2003.05.07
Cartridge Type					LTO2 cartridge	LTO1 cartridge	LTO1 cartridge	LTO1 cartridge	LTO1 cartridge	LTO1 cartridge
Last Unload					Normal	Normal	Normal	Normal	Normal	Normal
EOD Mark	1	1	1	0	Valid	Valid	Valid	Valid	Valid	Valid
Mounts	1	70	36.6	18.7332	1	20	30	41	41	33
Total Data Written	112 GB	872 GB	472 GB	216 GB	112 GB	214 GB	350 GB	409 GB	554 GB	542 GB
Total Data Read	49.3 MB	20.9 GB	6.47GB	8.69 GB	49.3MB	574MB	1.29 GB	1.72 GB	1.35 GB	20.4 GB
Recent Data Written	2 GB	112 GB	42.6 GB	34.6 GB	112 GB	2 GB	8.44 GB	33 GB	68.7 GB	67.6 GB
Recent Data Read	17 MB	1.13 GB	227 MB	320 MB	49.3MB	17 MB	142 MB	140 MB	145 MB	158 MB
Recent Write Retries	0	202/MB	23.6/MB	62.7/MB	1 (8.97/TB)	1 (499/TB)	0	0	1 (14.6/TB)	1 (14.8/TB)
Recent Read Retries	0	0	0	0	0	0	0	0	0	0
Recent Servo Errors	0	98/MB	16.3/MB	32.5/MB	0	0	1 (118/TB)	0	3 (43.7/TB)	0
Recent Fatal Write Errors	0	0	0	0	0	0	0	0	0	0
Recent Fatal Read Errors	0	0	0	0	0	0	0	0	0	0
Recent Fatal Servo Errors	0	0	0	0	0	0	0	0	0	0
Total Write Retries	3.62/MB	7.81/KB	1.17/KB	2.38/KB	1 (8.97/TB)	41 (191/TB)	18 (191/TB)	16 (39.1/TB)	846 (1.53/GB)	379 (700/TB)
Total Read Retries	0	587/MB	176/MB	201/MB	0	0	0	1 (583/TB)	0	3 (147/TB)
Total Servo Errors	0	3.72/KB	541/MB	1.14/KB	0	2 (9.33/TB)	5 (14.3/TB)	3 (7.33/TB)	8 (14.4/TB)	315 (582/TB)
Total Fatal Write Errors	0	0	0	0	0	0	0	0	0	0
Total Fatal Read Errors	0	159/MB	15.9/MB	50.3/MB	0	0	0	0	0	0
Total Fatal Servo Errors	0	463/GB	46.3/GB	146/GB	0	0	0	0	0	0

Figure 20 : VeriTape - exemple d'outil pour le contrôle de l'état des bandes LTO

Selon les types de support, il existe des matériels et logiciels capables d'analyser l'état du support et de présenter des rapports d'analyse (Figure 20).

5.3.4. Migration de support

La conséquence logique de la dégradation des supports est qu'il faut pourvoir à leur remplacement. Ceci est une condition *sine qua non* sans laquelle l'information ne pourrait pas faire l'objet d'autres traitements en raison de sa disparition physique.

On distingue plusieurs types de migration de support, selon que la manière d'adresser physiquement l'information sur le support est modifiée ou non (Figure 21) :

- **Refreshing** : migration de l'information vers un support de même type.
- **Duplication** : migration vers un support de même type ou un autre. La manière d'adresser le contenu reste identique.
- **Ré-empaquetage** : migration de l'information vers un support d'un autre type. La manière d'adresser le contenu est modifiée.

[51] Voir par exemple le « Pôle de recherche sur la conservation des données sur disques optiques numériques (GIS-DON) » (France) qui étudie les supports optiques et organise régulièrement des conférences : http://www.lne.fr/fr/r_et_d/gis-don/conservation-donnees-numeriques-gis-don.asp (consulté le 30 octobre 2009).

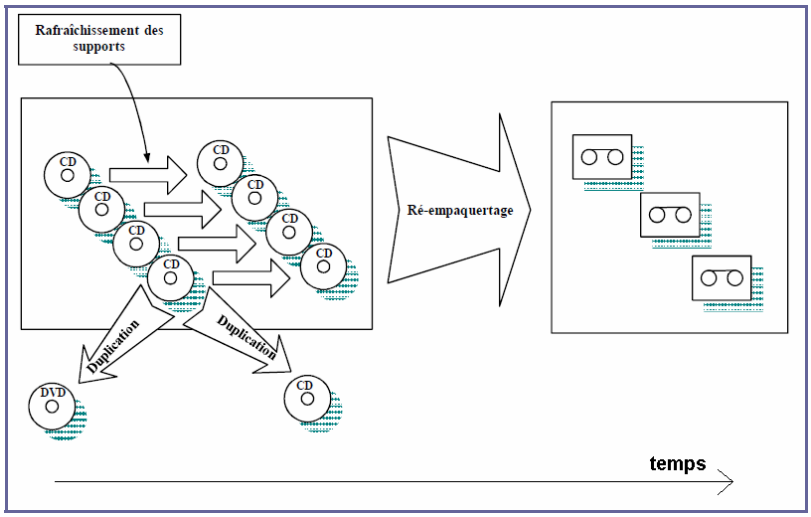


Figure 21 : Les différents types de migration
(source : Huc Cl., Séminaire PIN avril 09)

Dans la mesure du possible, cette migration ne doit pas influer sur les applications utilisant le système de stockage.

Une migration de support peut demander énormément de temps. À titre d'exemple, une migration de bande prend approximativement 30 minutes (manipulation humaine et copie à l'aide d'un appareil adéquat). Si on ne dispose que d'un seul appareil, pour un total de 10.000 bandes, 5.000 heures de travail seront nécessaires.

L'automatisation de ce processus doit être privilégiée. Par exemple, le Centre National d'Études Spatiales (CNES) à Toulouse (France) a mis en place en 1992 un service centralisé de stockage (STAF) pour l'ensemble des missions spatiales et des centres de recherches menées dans l'institution. En dix ans, le nombre de clients a augmenté d'un facteur 10 et le volume de données d'un facteur 50. Le STAF est cependant toujours composé de la même équipe de quatre personnes, grâce à l'automatisation opérée via des développements logiciels internes et l'augmentation des capacités de stockage des supports.

cnes
CENTRE NATIONAL D'ÉTUDES SPATIALES

Quels que soient les mesures mises en œuvre ou les critères étudiés au moment du choix du support, il ne faut pas oublier que cette stratégie n'est pas suffisante et « have the potential for endangering content by providing a false sense of security *».*[52]

5.4. Gestion des formats

Sans définition de format, un fichier n'est qu'une suite de zéros et de uns dépourvue de signification, d'autant plus qu'une même séquence binaire peut avoir de nombreuses interprétations (Figure 22). Un format est un ensemble de règles et d'algorithmes permettant d'organiser l'information dans un objet numérique, sous forme de bits.

La définition d'un format indique les subdivisions, le codage, les séquences, l'organisation, la taille et les relations internes qui définissent le format de

[52] KENNEY *et al.*, 2003 dans HARVEY R., *Preserving Digital Materials*, Éd. Saur, Munich, 2005, p. 123

manière unique et qui en permettent l'interprétation et la restitution. À titre d'exemple, une définition de format doit indiquer l'emplacement des séparations significatives à l'intérieur de la chaîne de bits et dire si un sous-ensemble de cette chaîne doit être interprété comme un caractère ASCII, une valeur numérique, une instruction machine, une sélection de couleur ou autre chose.

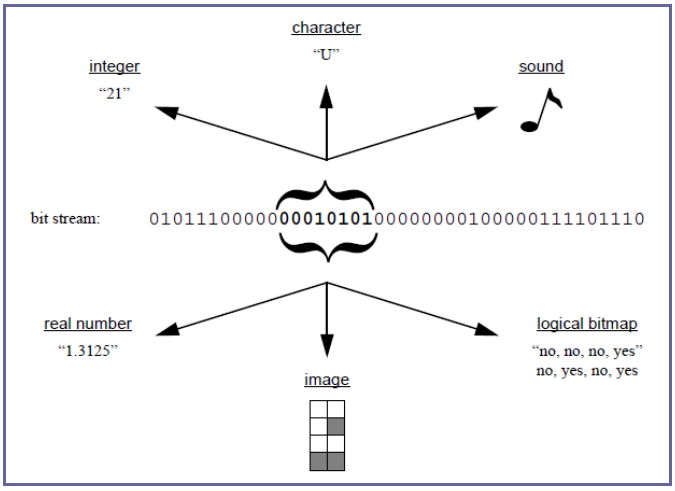
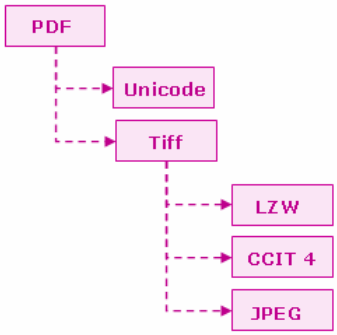


Figure 22 : Multiples interprétations d'une même séquence binaire (source : Rothenberg, 1999)



Interdépendance des formats

Il existe de nombreux types de formats différents qui sont parfois reliés ou imbriqués. Par exemple, le format XML est composé de caractères textuels basés sur les standards ASCII ou Unicode.[53] Un fichier PDF peut contenir du texte, des images (qui sont alors au format TIFF, GIF, JPEG ou PNG), voire des vidéos pour lesquels il existe également plusieurs normes et standards, déclinés généralement en plusieurs versions, chacune ne disposant pas des mêmes possibilités et fonctionnalités (cf. image ci-contre). La correspondance « 1 objet numérique = 1 fichier = 1 format » n'est donc pas toujours valable. Par exemple, un fichier TIFF contenant des métadonnées au format IPTC ou XMP a la correspondance « 1 objet numérique = 1 fichier = 3 formats ».

Dès lors, la compréhension et la maîtrise des formats nécessitent une prise en compte globale de chacun d'eux et des relations qu'ils entretiennent les uns avec les autres.

Par ailleurs, certains formats sont propriétaires et donc difficilement maîtrisables en raison de la non-publication de leur documentation.

Par exemple, le CNES a dû réaliser une triple opération semi-manuelle de ses documents bureautiques. Un premier système bureautique propriétaire avait été utilisé dès 1985. Au début des années 1990, vu l'évolution de la micro-informatique et du marché des logiciels bureautiques essentiellement dominé par Microsoft, une migration de DOS vers MS Word s'est imposée. Devant l'impossibilité technique d'opérer cette migration, il a fallu réencoder les documents manuellement. Six ans plus tard, le CNES s'est rendu compte que les documents enregistrés en MS Word n'étaient que partiellement compatibles avec MS Word 97 pour Windows. Dès lors, le texte a pu être récupéré, mais la mise en pages complexe de plusieurs milliers de tableaux a dû, elle, être refaite.

D'autres formats sont liés à des systèmes d'exploitation spécifique. En 1999, le CNES a procédé au remplacement de ses machines de calcul basées sur le système d'exploitation NOS/VE par des machines basées sur le système UNIX.

[53] HULSTAERT A., *Préserver l'information numérique. Codage et conversion de l'information*, Delivrable, Section Recherches, Smals, 2008 (accessible via l'extranet de la sécurité sociale sur http://documentation.smals.be/index.htm).

Cette migration a révélé que la plupart des données enregistrées présentaient des structures logiques et des encodages propres aux systèmes d'exploitation qui avaient été utilisés pour créer ces données, systèmes eux aussi en voie de disparition. Ces systèmes d'exploitation ajoutaient, au sein même des fichiers, des informations propres au système afin de faciliter leur gestion. Ces fichiers étaient donc totalement dépendants du système d'exploitation et leur migration a nécessité un travail de « nettoyage ».

Enfin, le nombre de formats ou de versions de formats ne cesse de croître. Chaque format comporte son lot de spécificités qu'il est nécessaire de maîtriser. Ainsi, la dernière version 1.7 de PDF permet d'incorporer des objets tridimensionnels via des méthodes de compression (voir image ci-contre).

Vue tridimensionnelle d'une douille dans un fichier PDF 1.7

Dès lors, une gestion des formats consistera à :

- sélectionner des formats pour la préservation (
- identifier et valider les formats des objets numériques (5.4.2) ;
- développer et utiliser des format viewers (5.4.3).

5.4.1. Sélection de formats pour la préservation

Certains formats sont plus adaptés à la préservation à long terme, en ce qu'ils facilitent le processus de préservation en simplifiant les opérations de préservation et en réduisant le nombre d'opérations à exécuter puisque ces formats sont moins soumis à l'obsolescence.

Pour évaluer si un format est pérenne, il faut tenir compte de plusieurs critères qui influent sur la faisabilité et sur le coût de la préservation (Figure 23). Ces critères sont importants quelle que soit la stratégie adoptée comme base pour de futures actions de préservation : migration vers de nouveaux formats, émulation de logiciel ou approche hybride.

Les deux critères fondamentaux sont l'ouverture et l'indépendance. D'autres critères annexes (diffusion, volume, complexité, transparence) peuvent également être appliqués.

Ouverture

L'ouverture désigne la mesure dans laquelle les spécifications complètes et les outils de validation de l'intégrité technique du format existent et sont accessibles.

La préservation de l'information dans un format numérique est impossible sans une compréhension approfondie de la manière dont l'information est codée et structurée. Il faut donc préférer des formats largement adoptés et dont la maintenance est assurée par un organisme de normalisation international ou supranational (comme l'Union européenne par exemple). Ce critère n'est cependant pas exclusif puisque le format TIFF est propriétaire de la firme Adobe et considéré comme un format pérenne pour les images.

Indépendance

L'indépendance se réfère à la mesure dans laquelle un format particulier dépend notamment d'autres formats, du matériel, du système d'exploitation ou de logiciels spécifiques et à la complexité de traiter ces dépendances dans le futur. Par exemple, des données scientifiques récoltées à l'aide de senseurs peuvent devenir inutiles sans le logiciel permettant leur visualisation ou leur analyse, logiciel qui peut s'avérer lui-même difficile à maintenir, même avec le code source disponible. D'autres données, par exemple certains types de fichiers audio, ne peuvent être utilisés qu'à l'aide de plateformes ou de logiciels spécifiques car leur format inclut des métadonnées visant à protéger les droits

intellectuels de l'information. Ces mécanismes empêchent toute reproduction ou modification (par exemple lors d'une migration de format) de l'information, opérations nécessaires pour assurer la préservation de l'information à long terme.

Diffusion

La diffusion consiste à voir si le format est utilisé par les principaux créateurs, diffuseurs et les utilisateurs de l'information. Cela comprend tant son utilisation en tant que format maître pour la livraison aux utilisateurs finaux que comme moyen d'échange entre les systèmes. Une diffusion large entraîne généralement une obsolescence moins rapide. Ces outils pour la visualisation et la migration sont plus susceptibles d'être produits par les fournisseurs sans que des investissements lourds, car très spécifiques, ne soient nécessaires.

Le critère de diffusion n'est cependant pas prioritaire. À titre d'exemple, le format MS Word est largement répandu, mais ne constitue en rien un format susceptible de favoriser une préservation de l'information.

Volume

Certains formats engendrent des fichiers volumineux. À titre d'exemple, un PDF/A nécessite l'encapsulation dans le fichier de toutes les polices de caractères utilisées. De même, TIFF sera souvent utilisé pour l'archivage des images sans aucun mode de compression, ce qui rend ces fichiers très volumineux.

Complexité

La complexité d'un format est un frein à sa maîtrise et donc complexifie les opérations de préservation qui devront être menées. Par exemple, la norme OOXML fait 6.000 pages tandis que la norme ODF n'en fait que 800 !

Transparence

La transparence réfère à la mesure dans laquelle la représentation numérique est ouverte à l'analyse directe avec les outils de base, y compris la lisibilité de l'homme en utilisant un éditeur de texte seulement. Les formats numériques dans lesquels les informations sont représentées simplement et directement seront plus faciles à migrer vers de nouveaux formats et plus sensibles à l'archéologie numérique.[54]

De nombreux formats numériques utilisés pour diffuser l'information emploient des techniques de cryptage ou de compression. Le cryptage est incompatible avec la transparence, tandis que la compression n'est qu'un obstacle à la transparence. Cependant, pour des raisons pratiques, certains fichiers (audio, image ou vidéo) ne peuvent jamais être stockés dans une forme non compressée. Dans ce cas, et pour autant que cela soit possible, il est préférable d'utiliser des méthodes de compression réversibles (tout en veillant à conserver l'algorithme de compression).

[54] L'archéologie numérique consiste à partir d'objets numériques devenus inexploitables (physiquement et/ou logiquement) et à procéder à du *reverse engineering* afin de récupérer les données.

	PDF	PDF/A	ODF	OOXML	TIFF	JPEG 2000
Ouverture	Standard ouvert	Standard ouvert	Standard ouvert	Standard ouvert	Standard ouvert	Standard ouvert
Indépendance	moyen**	bon	très bon	bon	bon	bon
Diffusion	très bon	moyen	faible	faible	très bon	faible
Volume	moyen	élevé	faible	faible	élevé	faible
Complexité	moyen**	moyen	faible	complexe	faible**	moyen
Transparence	faible**	moyen	bon	moyen	bon	moyen

** dépend des fonctionnalités utilisées

Figure 23 : Critères de sélection des formats appliqués à quelques formats courants

Enfin, il faut également tenir compte des **besoins métier et réglementaires**. Prenons par exemple un document Word à préserver comportant quelques phrases écrites avec des polices de caractères spécifiques (*24 Ko*).

Si le look & feel et la mise en pages du document sont importants et que le document peut être figé, une conversion en PDF/A sera recommandée (*116 Ko*).

Dans le cas où la préservation de la mise en forme du document n'est pas requise mais uniquement son contenu informationnel, une conversion en format ODF peut s'avérer plus judicieuse (*10 Ko*). Il s'agit d'un format ouvert, XML, permettant l'édition et relativement simple.

Si le souci principal est la simplicité du format pour augmenter autant que possible la préservation de son contenu informationnel, une transformation en fichier texte simple (*1 Ko*) pourrait également être envisagée.

Sur cette base, les formats recommandés pour la préservation à long terme sont :

Types d'objet	Format de production	Format de préservation
Document textuel	DOC, ODF, RTF, TXT	XML, ODF, OOXML, TXT, PDF(1.4), PDF/A
Document de présentation	PPT, ODP	ODP, PDF(1.4), PDF/A
Tableur	XLS, ODS, CSV	XML, ODS, CSV, PDF(1.4), PDF/A
Image	TIFF, PNG, JPEG, GIF	TIFF, JPEG2000

Une distinction entre formats d'entrée et formats de préservation est souvent opérée dans les systèmes d'archivage. Les formats d'entrée sont ceux pour lesquels le système est capable de réaliser une conversion dans les formats retenus pour la préservation à long terme.

Cette stratégie consiste donc généralement à :

- déterminer un nombre restreint de formats de préservation ;
- déterminer un faible nombre de formats acceptés en entrée (ils doivent cependant permettre de couvrir une majorité des besoins des fournisseurs d'informations) ;
- effectuer des tests des formats en entrée sur la base d'un logiciel testeur afin de s'assurer de la conformité de ces formats avec leurs spécifications ;

- convertir les formats d'entrée dans les formats de préservation dès l'entrée du fichier dans le système d'archivage ;
- archiver dans un journal les opérations de test et de conversion.

Cette stratégie nécessite parfois que certaines fonctionnalités offertes dans les formats d'entrée soient restreintes en vue de permettre la conversion dans les formats de préservation. À titre d'exemple, nous avons vu que le format PDF 1.7 permet d'inclure des objets tridimensionnels. Ces objets ne sont pas autorisés dans la norme PDF/A.

Dans le cas d'un système de dépôt légal (cf. 2.3), l'institution d'archivage n'a aucune prise sur la production d'information et l'information mise en dépôt est généralement soumise à des contraintes légales (cf. entre autres le chapitre 6). Ce type de système devra donc mettre en place des contrôles à l'entrée plus importants ainsi qu'une procédure de traitement et de notification au producteur des résultats du dépôt (rejet ou acceptation).

5.4.2. Identification et validation

Toute action appliquée sur un objet numérique (ouverture, modification, migration, etc.) nécessite d'en connaître le(s) format(s). Pour ce faire, l'extension du fichier n'est pas suffisante car il peut exister plusieurs formats ayant la même extension, tels que le « .pdf » qui recoupe les formats suivants :

- Systems Management Server (SMS) Package Description File (Microsoft Corporation) ;
- ArcView Preferences Definition File (ESRI) ;
- Netware Printer Definition File ;
- Acrobat Portable Document Format (Adobe Systems Inc.) ;
- Ventura Publisher EPS-variation Page (Corel Corporation) ;
- P-CAD Database Interchange Format (Altium Limited)
- …

Cette méthode ne permet pas non plus d'identifier les formats imbriqués, ni leurs propriétés.

Pour identifier un fichier, il faut donc se baser sur sa structure interne, par exemple les balises incorporées dans le fichier pour mentionner son format, mais aussi la manière dont est agencée l'information.[55]

Le format identifié, il convient d'en vérifier la conformité avec ses spécifications. Il est souvent estimé, à tort, que les outils et systèmes utilisés pour créer des objets numériques les produisent correctement. Une analyse du *Harvard Digital Repository Service*[56] menée en 2004 a montré que 4 % des fichiers XML, 1,4 % des fichiers TIFF (soit 6.323 fichiers) et 1 % des fichiers JPEG n'étaient pas conformes aux spécifications de ces formats, ce qui peut affecter le contenu du fichier comme l'illustrent les exemples de fichiers TIFF corrompus présentés dans la Figure 24. Un document PDF/A peut être invalide car seule la balise indicative est présente, mais le fichier ne respecte pas la structure définie par la

[55] Divers outils, de qualité variable, permettent de procéder à l'identification. Un exemple est fourni dans HULSTAERT A., *Digital Record Object Identification (DROID) - File Format Identification Tool*, Quick Review n° 22, 2009 ; *id., JSTOR/Harvard Object Validation Environment (JHOVE) 1.5 - File Format Identification and Validation Tool*, Quick Review n°25, 2010. Disponibles sur http://documentation.smals.be/index.htm

[56] Il s'agit d'un repository numérique que l'université de Harvard met à la disposition de ses membres pour le stockage, l'archivage et la préservation d'objets numériques. Plus d'informations : http://hul.harvard.edu/ois/systems/drs/ (consulté le 11 janvier 2010)

norme. De même, les documents Word produits par les versions actuelles de MS Office 2007 ne sont pas conformes à la norme Office Open XML (ISO).

Figure 24 : Exemples de fichiers corrompus

Il est donc indispensable de vérifier si les objets numériques préservés sont conformes à la norme dès leur entrée dans le système d'archivage, et ce pour trois raisons :

1. La conformité des objets numériques facilite les migrations de format. Les convertisseurs transforment la structure du document en se basant sur la structure existante. Si l'objet est non conforme, il y a donc un risque accru d'altération importante de l'objet lors de la migration.
2. Sur la base de la norme, il est possible de créer une application permettant la manipulation des objets numériques. En cas de non-conformité, aucune documentation sur ces erreurs ne permettra de les prendre en compte pour créer cette application.
3. Lorsque la préservation de l'objet est confiée à un tiers, elle permettra d'établir les responsabilités en cas de problème.

5.4.3. Format Viewer

Un *viewer* est un logiciel qui permet d'afficher rapidement des fichiers numériques sans avoir à passer par le logiciel utilisé pour les créer. Généralement, ils ne permettent pas d'éditer les fichiers. Un exemple bien connu est le logiciel Adobe Reader qui permet d'afficher les fichiers PDF. Cette stratégie peut être appliquée à des fichiers « anciens », c'est-à-dire des fichiers pour lesquels le format n'est plus utilisé couramment.

Cette méthode permet de ne pas devoir absolument migrer les fichiers dans un nouveau format plus récent, ces opérations de migration étant coûteuses (en investissement et en temps).

Par exemple, une banque belge conserve environ 50 To de fichiers TIFF 3 et 4. Une migration de ces fichiers n'est pas envisageable étant donné le temps que cette migration prendrait. Dans ce cas-ci, le fournisseur du système (SAP) s'est engagé à fournir pendant le temps nécessaire un viewer pour la visualisation des fichiers.

Cette stratégie présente cependant plusieurs inconvénients :

- Le *viewer* dépend lui aussi de l'environnement hardware et software, et devra donc lui aussi être maintenu, soit via une migration (développement et adaptation d'une nouvelle version en cas de changement de son environnement), soit par émulation.
- De nombreux *viewers* ne sont pas fiables à 100 % et contiennent eux aussi des erreurs. Dès lors, si l'original n'est plus disponible ou accessible, comment vérifier la qualité du *viewer* ?

- Certains formats autorisent la création de balises privées comme le format TIFF. Aucune garantie n'est offerte par les *viewers* commerciaux quant à l'accessibilité de ces balises.
- Plus le format est complexe, plus la création d'un *viewer* est difficile et ne permet pas toujours de rendre la complexité du format.

La stratégie *format viewer* est donc fortement liée à la gestion des formats. Le choix d'un format pérenne (ouverture, diffusion...), standard et normalisé facilitera le choix, l'obtention, le fonctionnement et/ou le développement d'un viewer.

5.5. Migration et conversion de format

5.5.1. Définition et types de migration

La migration est un *« ensemble de tâches organisées, conçues pour effectuer le transfert périodique d'objets numériques d'une configuration matérielle et logicielle à une autre, ou d'une génération de technologie informatique à une autre plus récente et qui implique une modification de l'information »*.[57] Il s'agit d'une stratégie de préservation pleinement opérationnelle qui présente l'avantage de maintenir les objets numériques dans un environnement informatique à jour, ce qui facilite l'accès et l'exploitation des informations en fonction des nouveaux besoins.

En outre, bien que complexe par divers aspects, cette stratégie est appliquée depuis de nombreuses années dans les entreprises, qui ont de ce fait acquis une solide expérience en la matière.

Outre la migration de support que nous avons déjà abordée, on distingue les migrations de formats et les migrations logicielles (que ce soit une migration vers une nouvelle version du logiciel ou vers un logiciel différent).

Dans une étude sur les risques liés à la migration de format[58], les auteurs identifiaient cinq raisons de migrer les informations vers un **nouveau format** (nouvelle version d'un même format ou un autre format tout à fait différent) :

1. l'obsolescence d'un format ou sa faible expansion sur le marché ;
2. la dépendance du format par rapport à une configuration hardware, software, un système d'exploitation spécifique qui est amené à évoluer ;
3. le fait qu'un format soit propriétaire et que le vendeur ne compte pas rendre ses spécifications publiques ;
4. une rationalisation des formats à gérer pour diminuer les coûts, par exemple il n'est pas rare de disposer de fichiers TIFF en version 4, 5 et 6 ; dans ce cas, la migration permet de rationaliser les coûts liés à la gestion de ces différentes versions.
5. le besoin de disposer d'une gestion des métadonnées intégrées plus complexes et plus précises.

Quant aux **migrations logicielles**, elles sont généralement la conséquence :

- dans le cas d'une migration de version, d'une évolution du logiciel qui permet à l'utilisateur de bénéficier de plus de fonctionnalités[59] ;

[57] *Digital Preservation Management*, Didactiel en ligne de la *Cornell University*, Partie 2 – Stratégies, 2004. Disponible à l'adresse suivante : http://www.icpsr.umich.edu/dpm/dpm-french/terminology/strategies.html (consulté le 26/01/2010) ; LAWRENCE G. W. *et al.*, *Risk Management of Digital Information: A File Format Investigation*, Washington, 2000

[58] *Ibidem*

[59] « *Une cause de migration peut être liée à l'accroissement de la taille autorisée des fichiers de données. On est dans certains cas passé d'un nombre maximum d'enregistrements par fichier de 16 millions à 2 milliards, ce qui a nécessité un accroissement*

- d'une incompatibilité entre une version logicielle et l'évolution de l'environnement *hardware* et *software* ;
- de l'évolution du marché logiciel ;
- …

5.5.2. Difficultés

La migration présente cependant un certain nombre de difficultés qu'il faudra prendre en compte.

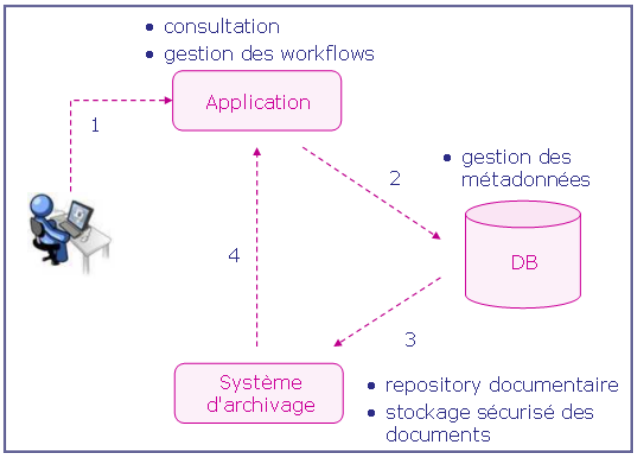
1) Comme indiqué plus haut, la migration entraîne inévitablement une modification de l'information. Dès lors, l'intégrité de l'information ne peut être vérifiée de manière formelle et un input intellectuel sera nécessaire pour valider le processus de migration de manière à ce que l'authenticité des informations puisse être garantie.

Conséquence logique de ce qui précède, il y a un risque de modification en « cascade » de l'information, de sorte que l'information dans sa dernière version migrée peut être très différente de l'information originale. Ceci illustre qu'il est important que les migrations successives soient fiables, documentées et validées.

2) Il n'est pas toujours possible d'avoir une correspondance biunivoque entre l'information originale et l'information migrée. Par exemple, la migration d'une base de données relationnelle vers une base de données XML ou la migration de feuilles de calcul autorisant les chiffres à virgule flottante à 16 unités vers d'autres ne les autorisant qu'avec 8…

3) Une migration est une opération complexe.

Par exemple, la conversion d'un fichier PDF 1.6 en un fichier PDF/A-1a peut s'avérer difficile, voire impossible. En effet, la création d'un fichier PDF/A nécessite d'encapsuler, à l'intérieur du fichier, l'ensemble des polices de caractères utilisées dans le fichier. Si cette conversion est réalisée longtemps après la création du fichier PDF 1.6, il peut s'avérer impossible de disposer des polices de caractères.



Un autre exemple est la migration d'une application de workflow. Après migration, il faut évidemment que les étapes du workflow soient respectées afin que les utilisateurs ne doivent pas redémarrer le processus depuis son état initial. Or, dans certains cas, y compris en cas de migration de version, un développement sera nécessaire pour remettre le workflow dans le même état.

Dans une application de gestion de dossiers existante, la consultation d'un document passe par trois logiciels différents. La première application sert d'interface avec l'utilisateur et gère le workflow. Pour atteindre un document, il faut y accéder via ses métadonnées stockées dans une base de données qui renvoie au document stocké dans le repository documentaire (cf. figure ci-contre).

Dans la situation TO BE, l'ensemble de ces tâches sera assuré par un même logiciel d'archivage électronique. Cette plateforme nécessite de réinjecter tous les documents stockés dans l'ancien système (500 Go) regroupés avec leurs

de la taille des entrées d'index (répertoriant les clés primaires) : la migration est destinée à transformer les valeurs des anciennes entrées d'index en vue de les rendre compatibles avec les nouvelles. », BOYDENS I., La préservation à long terme de l'information numérique, *Techno* n°28, Section Recherches, Smals, 2004, p. 7, n. 23.

métadonnées (2 Go dont certaines tables très complexes contiennent 40 millions de records).

Dès lors, une possibilité serait d'effectuer une « migration on demand ». Cela signifie que la migration est effectuée automatiquement par le système lorsque le document est consulté. Après un certain laps de temps, on procéderait ensuite à la migration de tous les documents restants. Cette solution demande moins de ressources en une seule fois, mais nécessite le maintien de certains ressources de l'ancien système aussi longtemps que la migration n'est pas totalement terminée. Il faut donc, dans ce cas aussi, prêter attention à l'obsolescence de ces ressources

5.5.3. Recommandations

Il convient, dès la création des informations, de suivre les modifications matérielles et logicielles ainsi que les interactions tolérées entre ces composants. Le suivi de ces versions permet d'identifier les migrations à effectuer.

De manière générale, il est conseillé de ne pas sauter de version majeure dans la mesure où la compatibilité ascendante[60] (*backward compatibilty*) n'est garantie que pour un nombre déterminé de versions. Cependant, certaines versions majeures n'ont pas d'impact sur la structure des données et n'apportent que des fonctionnalités supplémentaires, ce qui simplifie le processus de migration.

Dans certaines institutions, il est de coutume de sauter une version majeure, passant par exemple de la version 2.x à la version 4.x. L'expérience montre cependant qu'il faudra parfois recourir à la version « sautée », c'est-à-dire la version 3.x dans notre exemple, pour effectuer la migration.

Dans tous les cas, l'opération peut être délicate comme le reconnaissent les producteurs de logiciels eux-mêmes : *« ... but you also know that upgrading your databases and applications from currently installed Oracle products can be a complex and nerve-wracking job »*.[61]

Il faut donc considérer une migration comme un projet en soi et donc appliquer toutes les procédures établies dans ce cadre (étude préliminaire, analyse, tests, production, validation).

Il est recommandé d'éviter les couches de complexité qui ne sont pas indispensables. Les logiciels ou versions de logiciels n'utilisent pas nécessairement les mêmes modes de compression, ce qui implique des opérations de décompression-recompression supplémentaires. En outre, un risque existe de ne plus pouvoir accéder aux données si l'algorithme de compression n'est plus connu ou accessible. L'enjeu est identique au niveau du cryptage des données.

Selon l'importance des données et le contexte, il peut être intéressant d'archiver les différentes versions d'une même information produites au fur et à mesure des migrations. En cas de dépôt légal par exemple, la version originale doit généralement être conservée pour preuve du dépôt. En cas de longues durées de conservation, l'objet numérique passera par plusieurs versions pour en assurer l'accessibilité. La dernière doit naturellement être gardée et, selon les besoins, l'avant-dernière qui peut éventuellement servir en cas de mauvaise migration constatée ultérieurement.

La migration impliquant une modification de l'information, il est opportun de s'interroger et de déterminer les caractéristiques de l'information qui devront être préservées durant tout le délai de conservation afin de permettre à l'information

[60] Cette compatibilité signifie qu'un logiciel d'une version v+1 permet de traiter une application créée avec la version v de ce même logiciel.
[61] BURKE B., *Inside Oracle database 10g. The Great Migration Week experiment*, décembre 2003, cité dans BOYDENS I., La préservation à long terme de l'information numérique, *Techno* n°28, Section Recherches Smals, 2004, p. 7.

de garder sa valeur (*significant properties*). Cette analyse doit intégrer les besoins règlementaires et métier, et être continue puisqu'elle dépend de l'évolution des besoins.

Enfin, il est impératif de documenter rigoureusement la procédure pour permettre une traçabilité maximale, mais aussi de s'assurer de pouvoir continuer à exploiter les données.

En 2007, le groupe Total a établi une politique de conservation des données. Celle-ci stipule que les données structurées non gérées par SAP doivent être archivées via une mise à plat. Cependant, il est nécessaire de pouvoir continuer à les exploiter pour répondre aux demandes de l'administration fiscale. Total a donc pratiqué un test. Parallèlement à leur maintien dans le système, les données ont été archivées via une mise à plat et cette procédure a été documentée. Le groupe a ensuite fait appel à un informaticien extérieur, sans connaissance particulière du dossier. Celui-ci a pu traiter les données et obtenir les mêmes résultats que via le système où les données avaient été maintenues, en une semaine grâce à la documentation. Il s'agissait également pour Total de s'assurer de la possibilité pour le groupe de répondre à toute demande judiciaire ou fiscale, même en cas de départ des personnes ayant procédé à la mise à plat.[62]

5.6. Métadonnées

Une métadonnée est « *une donnée destinée à décrire une information et ce faisant à l'interpréter en vue d'en faciliter la gestion.* »[63] La gestion des métadonnées est nécessaire pour :

- restituer l'objet ;
- documenter les actions sur les objets ;
- organiser la structure des archives (par exemple la gestion des différentes versions d'un objet) ;
- comprendre l'objet ;
- retrouver l'objet.

Les métadonnées sont donc indispensables pour toute autre stratégie de préservation.

Les difficultés liées à l'interprétation de l'information ne seront pas abordées car elles relèvent plus généralement des systèmes de méta-informations[64], auxquelles appartiennent les métadonnées. Une étude approfondie de ces systèmes sortirait du cadre de cette étude.

Il est impossible de prévoir les besoins futurs en termes de métadonnées. Dès lors, assurer la pérennité de l'information, c'est s'engager à collecter, générer et extraire une quantité importante d'informations complémentaires.

[62] Table ronde « Archivage et conservation des bases de données » organisée par le CR2PA (Club des responsables de politiques et projets d'archivage), Paris, Total, la Défense, 14 janvier 2009.
[63] BOYDENS I., *Automatique documentaire. Section de Science et technologies de l'information et de la communication*, 5e édition, Bruxelles : Université Libre de Bruxelles (syllabus), 2006-2007, chapitre 1, p. 30.
[64] BOYDENS I., Les systèmes de méta-informations, *Techno* n°1, Section Recherches, Smals, 1997 (disponible sur http://documentation.smals.be/) ; *ead.*, *Les Dictionnaires de Données. Méthode, techniques et application pratique*, Delivrable, Section Recherches, Smals, juin 2000 (disponible sur http://documentation.smals.be/).

Pour cela, il faut :

- identifier les besoins généraux et spécifiques ;
- identifier et étudier les normes et standards disponibles ;
- comprendre leurs agencements, leurs possibilités et leurs limites.

Le recours aux métadonnées présente plusieurs difficultés[65] :

- Les métadonnées sont extensibles à l'infini. À chaque niveau méta, on pourrait ajouter un nouveau niveau méta, et ce à l'infini. Cela se traduit par la lourdeur et le coût de leur gestion, surtout lorsqu'elle repose sur une mise à jour manuelle.
- Les métadonnées peuvent être elles-mêmes erronées et incertaines : leur validation ne peut systémiquement faire l'objet de tests d'intégrité rigoureux.
- Un décalage temporel peut exister entre la mise à jour d'une donnée et la mise à jour de la métadonnée correspondante, vu que la mise à jour de cette dernière requiert éventuellement une phase d'analyse. Ceci est particulièrement valable pour les métadonnées descriptives.
- Enfin, les métadonnées sont elles-mêmes des données et doivent donc elles aussi être préservées.

« Si les métadonnées sont indispensables, il convient d'y recourir avec parcimonie et de privilégier les métadonnées générées semi-automatiquement. En outre, il convient d'agir parallèlement en « amont », lorsque l'on dispose d'une « prise » sur la source, à savoir lors de la création de l'objet numérique à préserver. »[66]

La définition d'un modèle de métadonnées (*metadata model*) qui s'appuie sur des normes et des standards reconnus peut répondre en partie à ces difficultés. À ce titre, un travail universitaire[67] a été réalisé en collaboration avec la section Recherches de Smals dans le cadre de la présente étude. Ce travail a permis de mettre en lumière les difficultés liées à la constitution d'un noyau de métadonnées en raison de l'hétérogénéité des normes et des standards étudiés (OAIS, PREMIS, Dublin Core, MoReq, METS, EAD) qui proviennent généralement de sphères de recherches et d'activités différentes (académique, scientifique, bibliothécaire, records management…), ce qui implique des différences dans les concepts à la base de ces normes et standards. Par ailleurs, ce travail a permis d'établir les bases d'un noyau commun de métadonnées applicable à un ensemble d'informations (principalement non structurées) de la sécurité sociale et du secteur des soins de santé en vue de la préservation à long terme.

5.6.1. Catégories de métadonnées pour la préservation

Sur la base du modèle OAIS vu précédemment (4.1), quatre catégories de métadonnées peuvent être distinguées :

- Les métadonnées **descriptives** qui permettent essentiellement de retrouver l'objet numérique. Il s'agit par exemple des mots-clés et index

[65] Ce passage est basé sur BOYDENS I., La préservation à long terme de l'information numérique, *Techno* n°28, Section Recherches, Smals, 2004, p. 9.
[66] *ibidem*
[67] RECTEM C., *La pérennisation digitale dans le secteur public : étude critique de plusieurs modèles ouverts de méta-information*, mémoire réalisé dans le cadre du Master en sciences et technologies de l'information et de la communication de l'ULB, Université Libre de Bruxelles, 2009.

de l'objet. Cette catégorie relève prioritairement de la théorie en langages documentaires.

- Les métadonnées de **structure** qui expliquent le contenu d'un *Archival Information Package* (AIP).
- Les métadonnées **techniques** qui permettent de restituer l'objet numérique.
- Les métadonnées **administratives** qui gèrent et documentent le cycle de vie de l'objet :
  - Les métadonnées ***d'identification*** identifient de manière unique l'objet au sein du système d'archivage. Il existe des techniques, essentiellement pour les environnements web, qui permettent d'attribuer un « identifiant pérenne » aux objets, c'est-à-dire un identifiant qui reste valable malgré les transformations successives des environnements matériel et logiciel.
  - Les métadonnées de ***provenance*** et de ***contexte*** qui documentent le cycle de vie de l'objet, à savoir sa provenance, les actions qu'il a subies (test de validité, migrations…). Elles établissent également le lien avec l'information de représentation.
  - Les métadonnées ***d'intégrité*** qui permettent de vérifier que l'objet n'est pas altéré.
  - Les métadonnées de ***droit*** (si d'application, elles ne figurent pas dans le modèle OAIS) qui déterminent entre autres le propriétaire de l'objet, l'institution en charge de sa conservation, les actions autorisées (consultation, reproduction, etc.)…

Ces métadonnées devront tant que faire se peut être collectées à l'entrée de l'objet dans le système d'archivage.

5.6.2. Normes et standards existants[68]

Ces dernières années ont vu apparaître un grand nombre de normes et de standards liés directement ou indirectement à la préservation et à l'archivage. Ce foisonnement n'est pas aussi anarchique qu'il pourrait paraître au premier regard, il est possible d'organiser et de classer ces standards les uns par rapport aux autres (Figure 25).

La plupart sont au format XML, qui présente l'avantage d'une séparation entre les valeurs et leur structure logique quelle que soit la plateforme logicielle ou hardware, et est donc de plus en plus préconisé.

Avant de passer à l'examen de quelques-uns des standards les plus importants, voyons quels liens ceux-ci entretiennent entre eux.

La norme centrale est le modèle OAIS. Comme nous l'avons vu, il s'agit d'un modèle conceptuel qui ne spécifie pas une méthode d'implémentation. Elle sert cependant de référentiel à un grand nombre d'initiatives plus opérationnelles.

Ainsi, PREMIS a été directement conçu sur la base de l'OAIS et en respecte les recommandations. METS et XFDU sont deux normes pour l'empaquetage au sens de l'OAIS. Leur utilisation en parallèle de PREMIS est fortement conseillée.

[68] Ce chapitre s'appuie sur le résultat du travail issu de la collaboration entre la section Recherches de Smals et le département des sciences et technologies de l'information et de la communication de l'ULB dans le cadre d'un mémoire universitaire. Voir note 67.

MoReq est un ensemble de spécifications méthodologiques et pratiques en vue de l'archivage des objets numériques. Cependant, pour tous les aspects liés à la préservation, MoReq renvoie à l'OAIS.

Enfin Dublin Core est une norme de métadonnées descriptives qui est reconnue par PREMIS et qui présente l'avantage d'être bien connue et répandue.

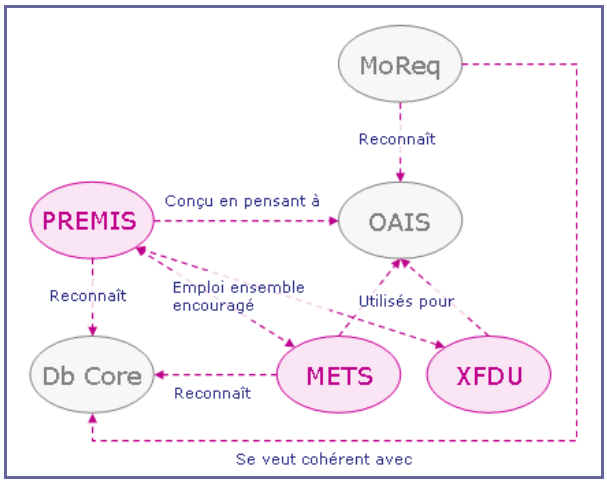


Figure 25 : Lien et relation entre les normes et standards de métadonnées existants

METS

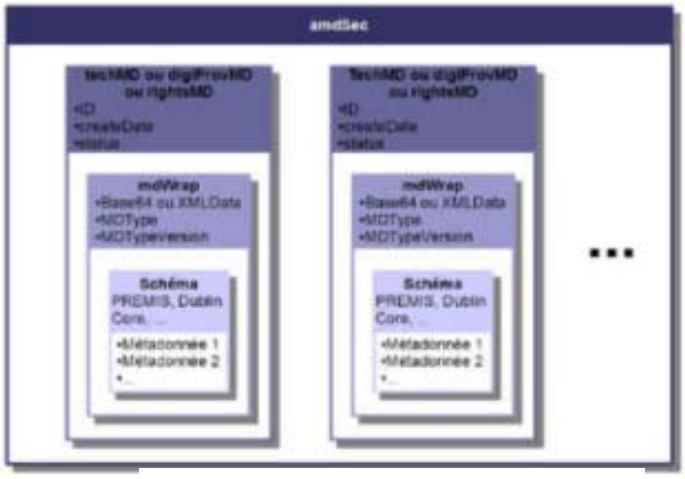


Figure 26 : Fichier METS - encapsulation

Le *Metadata Encoding and Transmission Standard* (METS) est un schéma XML développé à l'initiative de la *Digital Library Federation* (organisme regroupant les bibliothèques américaines) et maintenu par la *Library of Congress*. Sa première version date de 2001 ; en avril 2009 est sortie la version 1.9.

METS est une spécification d'encodage et de transmission de métadonnées qui fournit les moyens nécessaires à la gestion d'objets numériques ainsi que l'échange de ces objets (entre systèmes ou avec les utilisateurs).

METS est la norme la plus utilisée actuellement pour la constitution des paquets d'information tels que définis dans le modèle OAIS. Au format XML, il présente l'avantage d'être relativement mûr et présente une structure extrêmement souple permettant d'encapsuler au sein d'un fichier METS un ensemble de métadonnées provenant d'un autre format tel que PREMIS, selon le principe des poupées russes (Figure 26).

Par contre, le format s'avère relativement complexe en cas de grande quantité d'objets au sein du même fichier METS. En outre, étant donné qu'il ne « possède » pas de modèle conceptuel, une migration de ou vers METS en provenance d'un autre schéma de métadonnées peut s'avérer difficile.

Moins qu'un dictionnaire de métadonnées, METS est un container offrant une structure souple permettant de contenir et d'organiser l'ensemble des métadonnées nécessaires à la gestion d'un objet numérique et à son échange, objet qui peut être simple ou complexe.

Un document METS est composé de 7 sections (Figure 27) :

- `metsHdr` : l'en-tête METS permet d'encoder un minimum de métadonnées descriptives à propos du document METS lui-même.
- `dmdSec` : section des métadonnées qui rassemble les métadonnées descriptives de l'objet ou des objets (METS ne fournit aucune métadonnée descriptive (ni vocabulaire, ni syntaxe) et renvoie vers les standards en vigueur. Notons que le comité éditorial de METS reconnaît « officiellement » certains standards).
- `amdSec` : section des métadonnées administratives qui se subdivisent en :
  - `techMD` : les métadonnées techniques documentant le format, ...
  - `rightsMD` : les métadonnées de droit documentant le copyright, droits d'accès...
  - `sourceMD` : les métadonnées sources documentant le document à la source de la création de l'objet.
  - `digiProvMD` : les métadonnées de provenance digitale contenant les données sur les opérations de préservation comme les migrations par exemple.
- `structLink` : les liens structurels permettent de faire état d'hyperliens entre différents objets repris dans le document METS.
- `behaviorSec` : section qui donne des informations sur comment rendre le(s) fichier(s) informatique(s) intelligible(s) pour l'utilisateur.
- `fileSec` : section des fichiers qui contient l'inventaire des fichiers informatiques décrits dans le document METS.
- `structMap` : la carte structurelle donne la structure de l'objet et lie les différentes sections du document.

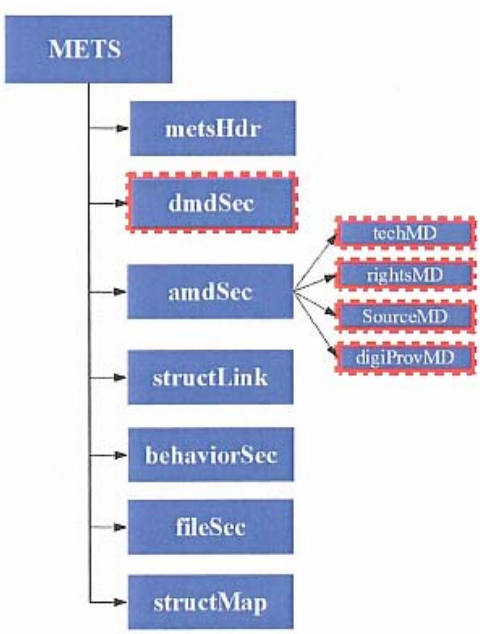


Figure 27 : Structure d'un fichier METS

Toutes les sections peuvent être répétées.

Les sections entourées dans la figure indiquent les possibilités offertes par METS pour insérer des métadonnées en les y « emballant » ou de pointer vers un ensemble de métadonnées extérieur ou de faire les deux. On peut donc insérer dans un document METS des métadonnées encodées en PREMIS, Dublin Core... Un document METS est donc extensible à souhait, ce qui en fait le format actuellement privilégié pour l'archivage à long terme dans le cadre des systèmes de dépôt légal (2.3) et pour l'implémentation des paquets d'informations définis dans le modèle OAIS.

XFDU

Créé par le *CCSDS* pour le même usage que METS, le format XFDU est plus récent. Bien qu'il présente des avantages pour remplacer METS (notamment un modèle conceptuel), il n'est pas encore fortement répandu et manque de maturité.

PREMIS

Un groupe de travail regroupant plus de trente experts de cinq pays différents représentant les bibliothèques, les musées, les archives, les administrations et le secteur privé a été mis en place en 2003 dans le but de définir un noyau de métadonnées de préservation utilisable par la communauté de la préservation numérique. Par métadonnées de préservation, le groupe entend : « *the information a repository uses to support the digital preservation process* ».

Publié en 2005 pour sa 1ère version, en 2008 pour la 2e, le modèle PREMIS (*Preservation metadata : implementation strategies*), basé sur le modèle OAIS, est composé de trois parties :

- un **modèle de données** PREMIS (modèle conceptuel du schéma de métadonnées) ;
- un **data dictionary** regroupant l'ensemble des balises, leurs règles d'utilisation et leur signification ;
- **des exemples concrets** d'utilisation.

Des considérations sur l'implémentation ainsi qu'un glossaire viennent compléter le tout. Techniquement, PREMIS est un schéma XML qui offre le moyen de décrire toutes les informations relatives à des objets numériques en vue de leur conservation à long terme.

PREMIS ne définit que des métadonnées de préservation. Pour d'autres métadonnées, il renvoie à d'autres normes et/ou standards.

Cinq domaines pertinents pour les métadonnées de la préservation ont été identifiés par le groupe de travail. Ces cinq domaines ont guidé le développement de PREMIS.

- La **provenance** : informations ayant trait à l'historique de la préservation de l'objet numérique, depuis sa création et en suivant les changements successifs dans la préservation physique et/ou logique.
- L'**authenticité** : informations nécessaires pour attester que l'objet numérique conservé est bien ce qu'il prétend être, qu'il n'a pas été altéré, soit intentionnellement, soit par inadvertance, sans documenter le fait.
- Les **activités de préservation** : documentation des actions entreprises pour la préservation de l'objet numérique, et garde la trace des conséquences de ces actions sur la présentation, le rendu ou la fonctionnalité de cet objet.
- L'**environnement technique** : description du matériel, des systèmes d'exploitation et des progiciels nécessaires à l'activation et à l'utilisation de l'objet en l'état dans lequel il est préservé.
- La **gestion des droits** : conservation de tous les droits restrictifs de la propriété intellectuelle qui limitent le pouvoir du dépôt de conservation de prendre des dispositions pour la préservation de l'objet numérique et sa dissémination vers les utilisateurs présents ou futurs.

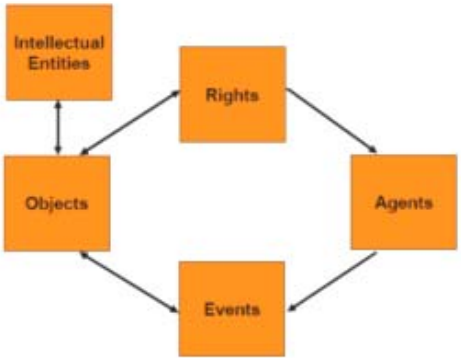


Figure 28 : Modèle de données de PREMIS

Le modèle de données de PREMIS identifie 5 entités différentes dans son modèle de données (Figure 28) :

- **L'entité intellectuelle** : ensemble cohérent de contenu raisonnablement décrit comme une unité, par exemple un livre, une carte, une photo ou une base de données. Parce que cette entité est bien décrite dans les métadonnées descriptives, elle est considérée comme en dehors du champ du dictionnaire des données.
- **Objet numérique** : unité discrète d'information dans la forme numérique.
- **Evènement** : action qui implique au moins un objet ou un agent connu du dépôt de préservation.
- **Agent** : personne, organisme ou progiciel associé aux évènements de préservation durant la durée de vie d'un objet.
- Les **droits** : assertions liées à un ou plusieurs droits ou permissions appartenant à un objet et/ou un agent.

et trois types de relations :

- Les relations de **structure** montrent les relations entre les objets et leurs parties.
- Les relations de **dérivation** résultent de la réplication ou de la transformation d'un objet : le contenu intellectuel de l'objet résultant est le même, mais l'instanciation de l'objet et son format, sans doute, sont différents.
- La relation de **dépendance** existe lorsque un objet en requiert un autre pour renforcer sa fonction, ses résultats ou la cohérence de son contenu.

Même si aujourd'hui, on ne dispose pas encore du recul nécessaire pour juger PREMIS, il est de plus en plus adopté dans le monde pour les programmes de préservation et présente des avantages indéniables (format XML, flexibilité, communauté d'utilisateurs en expansion).

Il semble donc progressivement se positionner comme une spécification essentielle dans le domaine.[69]

5.7. Encapsulation

Bien que moins éprouvée, l'encapsulation présente des caractéristiques intéressantes et est généralement associée à la préservation à long terme de l'information numérique. Nous avons donc choisi de l'aborder afin d'en montrer les avantages et les limites.

L'encapsulation consiste à réunir dans une seule et même entité physique l'objet numérique à conserver et l'ensemble des informations nécessaires à sa gestion et son accès.

Au départ, cette stratégie était surtout liée à l'émulation[70] (5.8), dans la perspective de constituer un objet autodescriptif pouvant s'exécuter sur un émulateur. Pour ce faire, il fallait réunir en un seul objet :

- le fichier et son environnement logiciel ;
- des informations sur l'émulateur devant être utilisé pour exécuter le fichier ;
- des informations sur le type de matériel nécessaire…

Clairement, ce principe était trop lourd et complexe à gérer. Une expérience fut menée à l'Université de Leeds en Angleterre dans le cadre du projet CAMiLEON, qui visait à émuler un ordinateur ICL1700 datant des années 1970 et qui a permis de mieux comprendre les informations nécessaires à ce genre d'exercice. Les résultats ne furent guère concluants sur la faisabilité de cette approche : « *the only test of an encapsulated specification is at the point it is used to implement a rendering tool.* » Dès lors, « *the risk of missing vital information in the specification seems to invalidate this approach.* »

Depuis lors, l'encapsulation a également été adoptée en dehors de l'émulation, comme base pour d'autres stratégies. En Australie, le projet *Victorian Electronic Records Strategy* (VERS) fut lancé en 1999 dans le but de pérenniser les documents issus de l'administration australienne, en rendant les objets autodescriptifs. La méthode consiste à créer des packages d'information (*VERS Encapsulated Object – VEO*) regroupant toutes les informations nécessaires dans une même entité physique afin de faciliter la gestion des objets (y compris le

[69] Il est intéressant de noter que plusieurs logiciels d'ECM supportent ce standard, cf. BURNETT S. et al, *Document and Records Management - Controlling Information Risk and Aiding Compliance*, Butler Group, 2008, p. 129-131.
[70] L'émulation consiste à reproduire le fonctionnement d'un matériel et/ou d'une couche logicielle spécifique(s), généralement sur une machine plus récente (machine hôte).

contrôle d'intégrité même si l'objet ne se trouve plus dans le système d'archivage) et le reverse engineering[71] si nécessaire.[72]

Chaque VEO Information package (VEO IP) contient l'« objet à préserver » (*payload*), les métadonnées du VEO et une (ou plusieurs) signature(s) numérique(s) qui garantissent l'authenticité et l'intégrité du VEO (Figure 29).

Chaque payload (*record*) est accompagné de métadonnées et d'un ou plusieurs documents. Chaque document est accompagné de ces métadonnées et peut exister en un ou plusieurs formats (*encoding*). Par exemple, sur la Figure 29, le record est composé de deux documents dont le premier est disponible en Word et en PDF.

La méthode ainsi développée permettra théoriquement d'inclure aisément les documents migrés tout en conservant l'original de manière à pouvoir conserver la capacité à vérifier la signature digitale et donc leur authenticité et leur intégrité.

En combinant cette stratégie avec une gestion des formats (standardisation et limitation du nombre de formats – actuellement PDF, XML et TIFF), l'objectif affiché de VERS est de rendre les objets aussi durables que possibles, pour minimiser les efforts nécessaires au développement de viewers spécifiques ou à une opération de migration.

Cette stratégie présente l'avantage de créer des objets autodescriptifs et diminue ainsi le risque de pertes d'informations en raison de l'évolution asynchrone des systèmes utilisés. Elle constitue en outre une base solide pour l'application d'autres stratégies expliquées précédemment. Cependant, elle présente un risque de multiplication du nombre d'informations inclues, ce qui augmente *de facto* la complexité du modèle et des coûts liés à sa gestion.

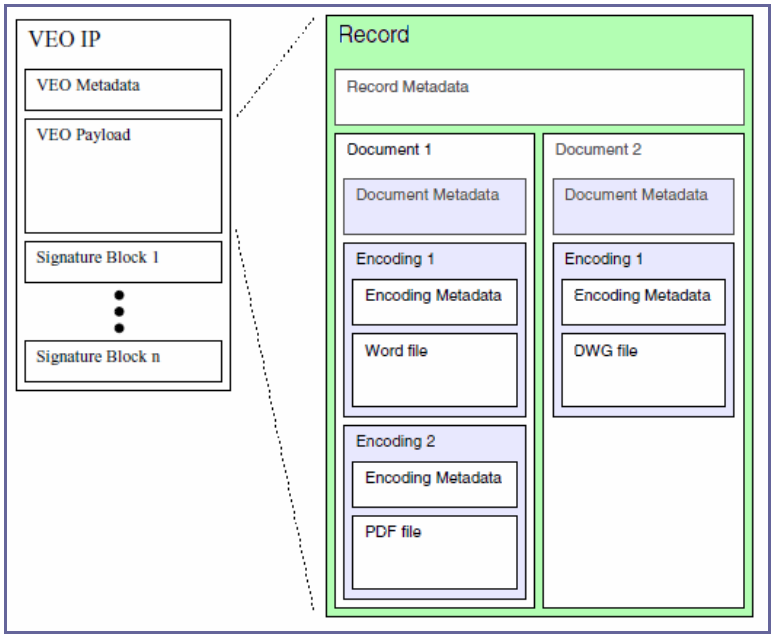


Figure 29 : Exemple de fichier VERS

[71] Activité qui consiste à étudier un objet pour en déterminer le fonctionnement interne ou sa méthode de fabrication. Article « Rétro-ingénierie », *Wikipedia.fr*, consulté le 26/01/2010

[72] QUENAULT H.S., VERS: Practical Digital Preservation, *Document numérique*, 8/2, 2004, P. 23-35 ; WAUGH A., The design of the VERS encapsulated object experience with an archival information package, *International Journal on Digital Libraries*, 6/2, 2006, p. 184–191.

5.8. Émulation

Enfin, une dernière stratégie de préservation souvent évoquée est l'émulation. Parfois confondue en certaines circonstances avec la virtualisation[73], l'émulation consiste à reproduire le fonctionnement d'un matériel et/ou d'une couche logicielle spécifique(s), généralement sur une machine plus récente (machine hôte).

Le principe de fonctionnement d'un émulateur n'est pas compliqué en soi : chaque instruction de l'environnement émulé est interceptée par l'émulateur et traduite en une ou plusieurs instructions exécutables par la machine hôte.

L'émulation peut porter sur différents niveaux présentant une complexité croissante : matériel, systèmes d'exploitation et applications.

Lancé à la fin des années 80, le projet CAMiLEON consista en la réalisation d'un émulateur pour permettre la lecture d'une version multimédia du *Domesday Book*[74] produite en 1986 par la BBC pour en célébrer le 900e anniversaire. Stocké sur des supports devenus progressivement illisibles par les ordinateurs plus modernes, la construction de l'émulateur était devenue indispensable pour permettre d'accéder à nouveau au contenu des supports. Trois années furent nécessaires pour y arriver. Les rapports des concepteurs de l'émulateur témoignent des difficultés à construire ce type d'application (dans ce cas-ci, sa réalisation fut possible grâce au recours à des ingénieurs ayant participé au projet de conception) et mentionnent « *[that] emulation is not necessarily superior to migration for preserving the original look and feel of complex digital objects* ». Et cela d'autant plus si l'information originale est peu documentée.

À ce dernier point, les tenants de l'émulation répondent par la nécessité d'encapsuler dans l'objet numérique l'ensemble des informations nécessaires à son interprétation (logique et sémantique – cf. 5.7). Il serait donc nécessaire d'inclure dans l'objet l'ensemble du programme nécessaire à son exécution ou du moins de le conserver pour pouvoir l'utiliser en cas de nécessité. À l'extrême, cela consisterait, par exemple dans le cas d'un programme java, à préserver non seulement le code mais également l'ensemble des librairies utilisées ainsi que les éventuelles applications annexes utilisées par le programme principal. Outre cela, il faut encore que cette préservation soit légalement possible et que l'institution bénéficie des droits de reproduction du logiciel.

D'autres obstacles sérieux doivent être pris en compte :

- Les ressources (budgétaires et techniques) nécessaires au développement des émulateurs. Certaines compétences techniques disparaissent avec l'évolution des technologies. Par exemple, à l'heure actuelle, il est extrêmement difficile de trouver des développeurs COBOL.
- Cette stratégie permettra-t-elle de prendre en compte les changements de paradigme dans l'évolution des technologies, telle l'interaction de l'utilisateur avec la machine (disparition éventuelle des souris et des claviers) ?
- Un émulateur dépend lui aussi de l'environnement hardware et software et devra donc lui aussi être maintenu.

[73] Pour de plus amples explications, OGONOWSKI G., *Virtualisation de serveur. Une technologie bien réelle*, Delivrable de la section Recherches, Smals, 2008, p. 19-21 (accessible via l'extranet de la sécurité sociale sur http://documentation.smals.be/index.htm).
[74] Le *Domesday Book* (ou simplement *Domesday*), en français « Livre du Jugement Dernier », est un grand inventaire de l'Angleterre terminé en 1086, réalisé pour Guillaume le Conquérant, l'équivalent de nos jours d'un recensement national.

- L'émulation peut s'avérer coûteuse en ressources système, ce qui implique que l'hôte doit être plus puissant que le système émulé ; en pratique, il faut compter un facteur 10 minimum, parfois même 100, voire plus. Cette baisse de performance est le prix à payer pour une totale indépendance vis-à-vis de l'architecture matérielle du système hôte.

Actuellement, l'**émulation logicielle** est considérée par la plupart des spécialistes de la préservation comme une technologie immature et n'offrant que de faibles garanties en matière de préservation à long terme. Elle peut cependant s'avérer utile dans certaines circonstances particulières, par exemple pour des formats ou applications très spécifiques comme les anciens jeux vidéo.

L'**émulation de stupport de stockage** est par contre plus généralement maîtrisée et certains constructeurs la proposent comme solution pour pallier l'obsolescence de certaines technologies (de stockage notamment). Ainsi Smals utilise l'émulation pour simuler un support de stockage. Construit par des spécialistes (dans ce cas-ci le fournisseur), l'émulateur permet de reproduire le fonctionnement des cassettes de bandes magnétiques. Cette méthode a permis de conserver la compatibilité du système de stockage avec les applications existantes. Dans le cas d'une émulation de ce type, il est nécessaire de s'assurer de la disponibilité de l'émulateur dans le temps par rapport au temps requis et de la possibilité de disposer d'un outil de migration (d'une stratégie de sortie) dans le cas où le fournisseur décide de mettre fin à la maintenance de l'émulateur.

5.9. Synthèse

Il n'existe pas de solutions techniques uniques. Préserver l'information à long terme consiste donc à combiner les différentes stratégies présentées ci-dessus, chacune pouvant intervenir sur une ou plusieurs couches de l'information (5.2).

Ces stratégies techniques et conceptuelles doivent être appliquée le plus tôt possible dans le cycle de vie de l'information, si possible dès leur création. Que ce soit pour le choix du support de stockage, du format de création et de préservation ou l'extraction de métadonnées, une intervention en amont facilite l'application de ces stratégies.

Bien que les stratégies puissent être regroupées en deux grandes catégories (5.2), seules les stratégies *« preserve object »* peuvent être considérées comme opérationnelles à l'heure actuelle, excepté dans certains cas bien spécifiques.

Pour préserver l'information, quatre stratégies complémentaires sont aujourd'hui pleinement opérationnelles et doivent être mises en œuvre :

1. La **gestion des supports de stockage** (5.3) via le choix, le contrôle de l'état des supports et leur remplacement.
2. La **gestion des formats** (5.4) de fichier en vue de sélectionner dès le début un format adéquat pour assurer une préservation à long terme. Ce format devra autant que possible assurer l'indépendance des données par rapport aux matériel et logiciels, afin d'éviter par exemple les problèmes rencontrés par le CNES lors de la migration (de format et de logiciel) de ses données et de sa documentation spatiales. À cette fin, il est fortement recommandé d'utiliser des formats ouverts, c'est-à-dire dont la documentation est publiquement disponible.

   Nous avons également vu qu'il était indispensable d'identifier de manière précise le format des données et de vérifier leur conformité à la norme, sous peine de rendre les migrations plus difficiles et risquées (risque de déformation non contrôlée de l'information).

3. Même si elle présente des difficultés inhérentes, la **migration périodique** (de format et de logiciel) (5.5) des informations vers des systèmes et formats plus récents est nécessaire. Pour cela, il faut veiller au maximum à la compatibilité ascendante entre les différentes versions d'un même logiciel. Dans tous les cas, une migration est un projet en soi et les procédures définies dans ce cadre doivent donc être respectées.
4. Le recours aux **métadonnées** (5.6) afin de documenter l'objet et son cycle de vie. Diverses normes et standards sont apparus ces dernières années et procurent une base solide pour la construction d'un noyau standard de métadonnées devant être recueillies en vue de la préservation à long terme.

L'encapsulation (5.7) est compatible avec les stratégies précédentes et peut servir de base pour faciliter la migration et éviter que le lien entre un objet numérique et ses métadonnées ne soit rompu.

L'émulation logicielle (5.8), souvent présentée comme une solution potentielle, n'est aujourd'hui pas opérationnelle, excepté pour des formats complexes et spécifiques et pour lesquels aucune autre solution n'existe.

Enfin, il est indispensable de documenter au maximum l'ensemble des opérations effectuées.

Bien que lourde à créer, la documentation permet d'assurer une traçabilité des actions entreprises et, de ce fait, de veiller à l'authenticité des informations en vue de leur réutilisation. À l'instar des données, cette documentation doit elle aussi être préservée à long terme.

6. Étude de cas

Afin de concrétiser les stratégies organisationnelles, techniques et conceptuelles évoquées ci-dessus, nous allons les illustrer au moyen d'un case study. Pour ce faire, nous avons sélectionné un projet de l'a.s.b.l. SIGeDIS consistant à archiver les documents électroniques échangés entre un travailleur et un employeur dans le cadre d'une relation de travail.

Après une brève explication du contexte de l'étude de cas (6.1), nous verrons un exemple d'organisation pouvant être mise en place dans le cadre de ce projet (6.2), et les différentes stratégies techniques et conceptuelles qui pourront être appliquées et comment (6.3). Enfin, pour terminer, nous expliquerons brièvement la problématique de la préservation à long terme de la signature numérique et la manière d'y répondre (6.4).

6.1. Contexte

Créé à la suite du Pacte entre les générations voté en 2005 par le gouvernement, SIGeDIS a pour mission de gérer pour le compte d'autres institutions de sécurité sociale les informations relatives à la carrière des travailleurs. À ce titre, la loi du 3 juin 2007 portant sur des dispositions diverses relatives au travail a stipulé qu'un contrat de travail[75] signé électroniquement[76] équivaut à un contrat de travail signé manuellement. La loi confie à SIGeDIS la mission d'archiver ces documents afin que le travailleur détienne une copie authentique de son contrat et puisse de cette manière garder une trace des avantages que cette relation de travail confère pour la pension.

La loi stipule que le document signé devra être archivé auprès d'un prestataire de services (ci-après dénommé tiers archiveur) qui devra en assurer l'archivage jusqu'à l'expiration d'un délai de cinq ans à compter de la fin du contrat de travail (Figure 30).

Ce délai expiré, si le travailleur le souhaite, le document doit être transmis pour dépôt légal (cf. 2.3) à SIGeDIS sous une forme lisible et exploitable qui en assurera la conservation jusqu'à l'expiration d'un délai de dix ans à compter du décès du travailleur. Ce délai peut donc potentiellement être supérieur à 100 ans dans le cas extrême d'un travailleur âgé de 20 ans qui décèderait à l'âge de 110 ans.

[75] Et plus généralement tout document électroniquement signé échangé entre le travailleur et l'employeur dans le cadre d'une relation de travail.
[76] La signature numérique doit répondre à un ensemble de critères afin qu'elle puisse être considérée comme valide.

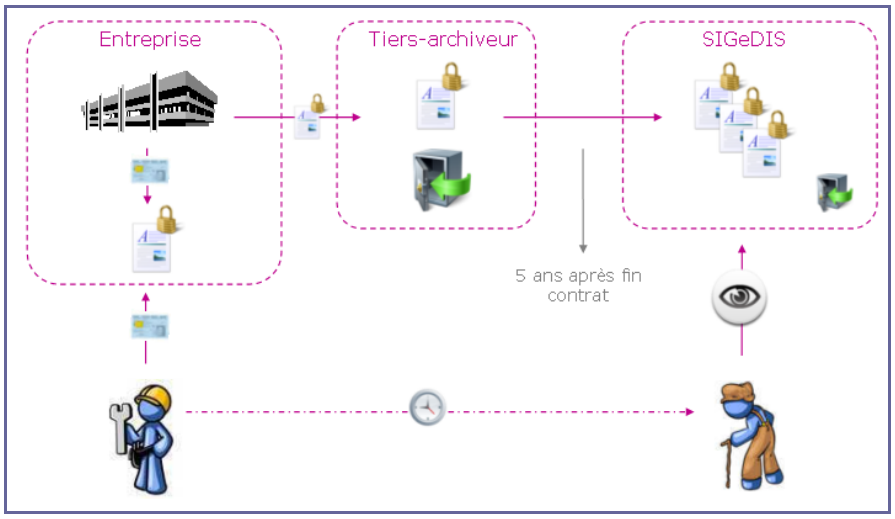


Figure 30 : Schéma du contexte de l'étude de cas

Dans le cadre de ce projet, il est donc nécessaire de préserver :

- les documents signés ;
- les preuves de la validité de la signature électronique du prestataire de services qui transférera le moment venu les documents vers SIGeDIS.

L'application du modèle fonctionnel de l'OAIS dans le cadre de cette étude de cas (Figure 31) montre les acteurs suivants :

- Les producteurs avec lesquels il faudra négocier le protocole de versement sont les employeurs (qui produisent réellement le document) via les tiers-archiveurs[77] (dans ce cas-ci, cela pourrait être les secrétariats sociaux) qui assureront dans la pratique la transmission du document à SIGeDIS.
- SIGeDIS qui devra assurer l'archivage et la préservation des documents aussi longtemps que nécessaire. Cette institution devra également transformer les SIP en AIP via un enrichissement des métadonnées descriptives, techniques et administratives. Elle devra également mettre en place un système de consultation.
- Les consommateurs sont principalement les travailleurs mais aussi, dans une moindre mesure, les entreprises.

[77] La mention « tiers-archiveur » désigné toute personne physique ou morale qui se charge pour le compte de tiers d'assurer et de garantir la conservation et l'intégrité des documents électroniques.

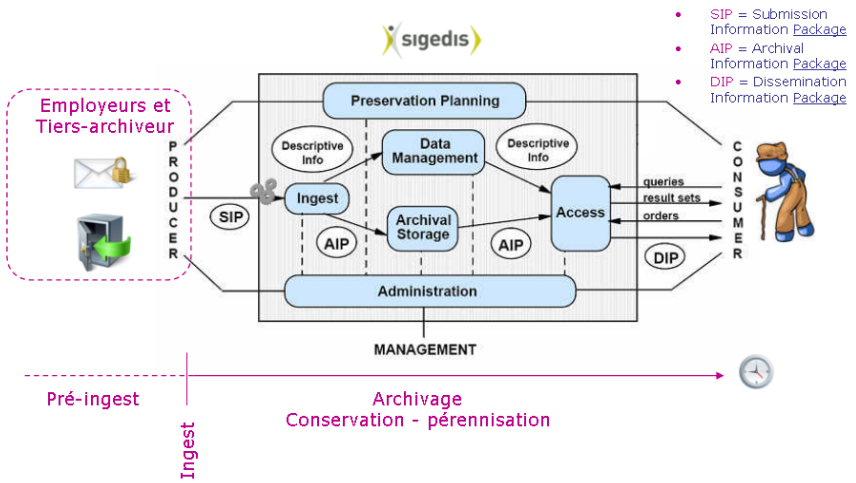


Figure 31 : Modèle fonctionnel de l'OAIS appliqué au cas de SIGeDIS

L'« ingest » (c'est-à-dire le versement des documents dans le système de dépôt légal de SIGeDIS) se situe au moment du versement des documents vers SIGeDIS. Nous pouvons cependant identifier une période de « pré-ingest » qui sera la période pendant laquelle le document sera créé, signé et enfin archivé auprès du prestataire de services. Pour assurer la préservation de ces documents, SIGeDIS devra intervenir dès cette phase de « pré-ingest », notamment en déterminant les conditions minimales pour l'acceptation du document par SIGeDIS.

6.2. Stratégie organisationnelle

Comme préconisé *supra* (4.3), l'organisation à mettre en place peut être découpée en services indépendants (« Entrée », « Stockage », « Gestion des données et accès ») et coordonnés (Figure 32).

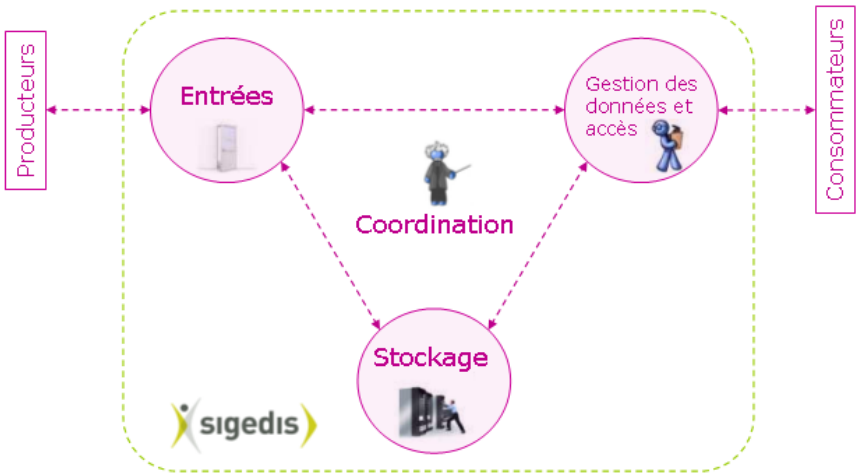


Figure 32 : Organisation possible pour l'archivage des documents électroniques chez SIGeDIS.

Coordinateur

Le rôle du coordinateur est surtout un rôle de gestion de projet, à savoir :

- responsabilité globale du système d'archivage ;
- relation avec le management ;
- relation avec les entités externes (producteurs et consommateurs) ;
- gestionnaire des ressources et des moyens ;
- organisation et planification des activités :
  - organisation de la répartition des tâches et responsabilités entre les différents services ;
  - mise en place des structures de concertation (réunion, groupe de travail, etc.).
- …

Service « Entrée »

Le service « Entrée » devra gérer les relations avec les producteurs. Pour cela, il mettra en œuvre la norme PAIMAS (*Producer-Archive Interface Methodology Abstract Standard* – ISO 20652) qui est un guide méthodologique pour la conduite formalisée des échanges entre des producteurs et une Archive. Les expériences réelles d'archives numériques ont mis en évidence que l'une des grandes difficultés de la mise en place d'un service d'archivage se situe dans la non-conformité des objets numériques fournis en entrée au système d'archivage. Ces difficultés trouvent elles-mêmes leur origine dans l'insuffisance du travail préparatoire entre producteur et Archive. L'objectif de PAIMAS est de pallier cette insuffisance en proposant une démarche structurée de dialogue préalable. Cette norme fait l'objet d'une définition d'un standard de mise en œuvre au sein du standard PAIS (*Producer Archive Interface Specification*).

La mise en œuvre de ce standard doit permettre de définir un protocole de versement incluant les aspects pratiques du versement ainsi que les exigences techniques que tout SIP doit respecter (par exemple, le NISS du travailleur doit être indiqué, le numéro BCE de l'entreprise, la conformité du document avec le format standard défini, etc.).

Une fois ces contrôles effectués et le document accepté, le SIP doit être complété via la génération de métadonnées descriptives et techniques supplémentaires afin de constituer un AIP complet pouvant être archivé.

L'AIP constitué, il est transmis pour stockage à l'entité « Stockage » et les informations complémentaires sont également transmises à l'entité « Gestion des données et accès ».

Les compétences requises pour ce service sont :

- celles de l'archiviste/gestionnaire d'informations afin de gérer les relations avec les producteurs, de définir les métadonnées à récolter pour le SIP et le AIP, d'organiser ces informations au sein d'un ensemble structuré… ;
- celles de l'informaticien afin de définir les formats de fichiers et de métadonnées, de vérifier la conformité des spécifications (contrôle de qualité sur les formats et les métadonnées), de développer et d'automatiser la procédure de versement…

Les retours d'expérience montrent qu'il s'agit du service le plus problématique en raison de la présence d'acteurs externes au système d'archivage.

Service « Stockage »

Le service « Stockage » doit gérer les objets au niveau binaire. Il doit en assurer l'intégrité tout au long de sa conservation. Il procède donc à une surveillance de l'état de dégradation des supports tel qu'expliqué ci-dessus et procède au remplacement des supports détériorés.

Les compétences requises pour ce service sont essentiellement des compétences en relation avec la gestion des supports de stockage : connaissance des supports, des conditions de stockage, des technologies de stockage (robots, baie de disque, etc.), méthodes pour surveiller l'état des supports, méthodes pour gérer les migrations…

Service « Gestion des données et accès »

Enfin, le dernier service « Gestion des données et accès » assume la responsabilité de la lisibilité et l'intelligibilité des objets dans le temps ainsi que de leur communication aux utilisateurs. Pour cela, il gère le cycle de vie des objets en planifiant les opérations de préservation sur la base des recommandations établies.

L'archivage des objets (ici les documents signés) répond aux besoins d'une communauté d'utilisateurs (dans ce cas-ci les travailleurs et les employeurs), l'entité « Gestion des données et accès » doit également formuler des propositions d'amélioration du système d'archivage en fonction de l'évolution de ces besoins ou des utilisateurs (ajout de métadonnées descriptives par exemple). Vu qu'elle gère la base de données permettant aux utilisateurs de mener des recherches au sein de l'Archive, elle effectue les adaptations nécessaires pour répondre aux besoins des utilisateurs.

Ce service doit également participer à la définition des propriétés signifiantes (c'est-à-dire les caractéristiques de l'objet qui doivent perdurer tout au long de la période de conservation). Étant responsable de la conservation et en contact direct avec les utilisateurs, le service est bien positionné pour assurer ce rôle en collaboration avec les gens du métier.

6.3. Stratégies techniques et conceptuelles

Une fois l'organisation définie, il convient d'identifier les stratégies techniques et conceptuelles qui devront être mises en œuvre.

Stockage pérenne

Le temps d'accès n'est pas fondamental dans le cadre de ce projet. Les objets peuvent donc être stockés sur des supports off ou near-line (cf. 5.3.1). En outre, SIGeDIS n'a pas d'obligation de destruction des documents. Cette suppression, une fois le délai de conservation écoulé, sera donc effectuée lors du remplacement des supports.

Le service « Stockage » devra veiller à stocker ces bandes dans une salle respectant les conditions de conservation recommandées (cf. 5.3.2), à contrôler l'état des supports (cf. 5.3.3) et à les remplacer si nécessaire (cf. 5.3.4).

Au moment du remplacement, les objets marqués comme n'ayant plus d'utilité ne devront pas être recopiés.

Gestion des formats

Étant donné que

- SIGeDIS devra assurer la préservation des documents sur des périodes très longues
- et que les documents seront d'abord archivés auprès d'un prestataire de services privé avant de parvenir chez SIGeDIS qui ne possède donc aucune emprise directe sur la production de l'information

il est impératif que SIGeDIS fixe des exigences en matière de format que les producteurs d'informations devront respecter s'ils veulent voir leurs documents acceptés par l'institution au moment du transfert.

Ces exigences doivent porter autant sur le format que sur d'éventuelles restrictions au sein même de ceux-ci.

Le plus simple semble être l'adoption du format PDF/A. Dans sa version 1a (la plus restrictive) si les documents ne sont pas produits à la suite d'une numérisation, dans sa version 1b si c'est le cas.

Les avantages de ce format sont qu'il inclut directement une série de limitations alors que l'utilisation d'un fichier PDF simple nécessiterait que le producteur respecte des restrictions, ce qui ne manquerait pas de générer au moment du transfert des documents vers SIGeDIS un grand nombre de problèmes en raison du non-respect de ces restrictions.

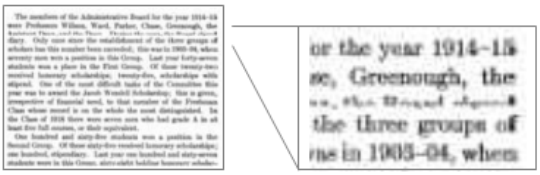


Figure 33 : Exemple de problème dû à la non-conformité d'un objet numérique

Des objets dont le format ne serait pas conforme à la norme pourraient engendrer des problèmes à long terme (Figure 33). Il est donc impératif que SIGeDIS mette en place une procédure de validation des documents dès leur réception, accompagné d'un système de feedback vers le déposant.

Certaines questions restent ouvertes aujourd'hui, entre autres les conséquences au cas où le document ne respecte pas les exigences fixées. Le document peut-il être refusé par SIGeDIS ? Ces questions devront être résolues lors de la mise en place du système d'archivage.

Métadonnées

Comme nous l'avons vu précédemment, pour assurer l'archivage à long terme des documents récoltés par SIGeDIS, il convient de définir les métadonnées qui devront être collectées. Ces métadonnées devront permettre tant la gestion et la consultation des documents que leur préservation, ce qui explique la présence de métadonnées « métiers » telles que le numéro d'identification du travailleur, le numéro d'identification de l'employeur, la durée de conservation… et la présence des métadonnées de préservation (métadonnées techniques par exemple issues de PREMIS).

Il faut distinguer les métadonnées qui devront être fournies par le tiers-archiveur (l'employeur) lors du dépôt des documents et ceux qui seront générées par SIGeDIS au moment du dépôt. Il s'agit de la distinction opérée dans le modèle OAIS entre les *Submission Information Package* (SIP) et les *Archival Information Package* (AIP).

Il est évident que lors du dépôt d'un document, l'employeur doit indiquer le NISS du travailleur afin que SIGeDIS puisse rattacher ce document à un travailleur. Il s'agira donc d'une métadonnée obligatoire dans le SIP. À l'inverse, il n'est pas nécessaire que l'employeur mentionne les métadonnées techniques (formatName, formatVersion) puisque SIGeDIS peut aisément générer ces métadonnées, notamment sur la base du contrôle de la conformité du format de fichier avec les spécifications établies par SIGeDIS. Via ces opérations de contrôle et de génération de métadonnées complémentaires, le SIP pourra recevoir le statut d'AIP.

Pour constituer ces *information packages*, les différents normes et standards de métadonnées évoqués ci-dessus peuvent être utilisés. Pour constituer le paquet en tant que tel, METS est tout à fait adéquat. Il permet d'insérer dans le schéma XML des métadonnées de préservation en provenance de PREMIS, celles liées à la signature digitale (norme XAdES-A[78]) ainsi que les métadonnées métier.[79]

Ce fichier XML METS pourrait être conservé parallèlement au document en question.

La durée de conservation étant de 10 ans après le décès du travailleur, il conviendra de prévoir un processus permettant à SIGeDIS d'encoder la date du décès afin que la période de conservation puisse démarrer. Dans le sens où SIGeDIS assure l'identification des travailleurs au sein de la sécurité sociale, cette institution sera avertie automatiquement du décès d'une personne.

Enfin, au terme de la période de conservation, il est préférable de confier à une personne la confirmation de la suppression. Une suppression automatique ne permettrait pas de pallier les éventualités liées à une nécessité de conserver l'information plus longtemps (affaire judiciaire, contentieux avec les héritiers, etc.).

```
<?xml version="1.0" encoding="UTF-8"?>
<mets:mets>
<objID>
<type>
<profile>
<mets:hdr>
        <createdate>
        <lastModDate>
        <agent>
</mets:hdr>
<mets:dmdSec ID="dmd002">
        <mets:mdWrap MDTYPE="Sg" LABEL="SIGEDIS Metadata">
        <mets:xmlData>
                <worker>
                        <Sg:NISS>
                        <Sg:name>
                        <Sg:first name>
                <employer>
                        <Sg:numBCE>
                        <Sg:location>
                <...>
        </mets:xmlData>
        </mets:mdWrap>
       </mets:dmdSec>
       <mets:amdSec>
        <mets:techMD ID="object1">
                <mets:mdWrap MDTYPE="PREMIS:OBJECT">
                <mets:xmlData>
```

[78] XAdES (*XML Advanced Electronic Signature*) : spécification du W3C prolongeant XML Signature (XMLDSIG) sur la non-répudiation des messages signés électroniquement. La signature XAdES satisfait aux obligations légales pour les signatures électroniques évoluées, telles que définies dans la Directive européenne pour les signatures électroniques. Elle offre une authentification et une protection de l'intégrité minimales.

[79] Pour rappel, un noyau opérationnel qui pourrait servir dans ce cas-ci a été mis au point dans le travail réalisé par C. Rectem en collaboration avec la section Recherches de Smals – voir note 67.

```
                <premis:object>
                        <...>
                        <premis:format>
                                <premis:formatName>
                                <premis:formatVersion>
                        </premis:format>
                        <premis:software>
                                <premis:swName>
                                <premis:swVersion>
                                <premis:swType>
                        </premis:software>
                        <...>
                </premis:object>
                </mets:xmlData>
                </mets:mdWrap>
        </mets:techMD>
        <...>
</mets:amdSec>
```

Exemple simplifié de fichier METS incluant des métadonnées descriptives (worker, employer) et techniques (< PREMIS)

Encapsulation

Vu la longue période de conservation, il pourrait être pertinent de constituer l'AIP comme une seule entité comprenant l'objet numérique et ses métadonnées, à l'instar de ce qui a été fait dans le cadre du projet VERS (cf. 5.7). Pour constituer cet AIP, un fichier METS serait utilisé (cf. l'exemple ci-dessus).

Une partie de ces métadonnées serait copiée dans un système permettant aux utilisateurs d'émettre des requêtes sur le système afin de consulter les fichiers (essentiellement les métadonnées descriptives) ou aux gestionnaires du système de gérer les données dans le cadre des processus de préservation (formats de fichier, versions…).

Ce mécanisme éviterait que le lien entre l'objet et ses métadonnées ne soit perdu en raison de l'évolution asynchrone des systèmes d'archivage et de gestion des métadonnées.

Migration

Afin d'éviter les risques d'obsolescence des formats, il faudra migrer périodiquement les fichiers vers un nouveau format. Les fichiers étant au format PDF/A, le rythme des migrations de format sera sans doute assez faible. Pour faciliter leur migration d'une application à l'autre, il conviendra également de ne pas utiliser de techniques de compression.

Une première migration pourrait même se produire dès l'entrée du document. Par exemple si le format n'est pas PDF/A, mais que SIGeDIS est capable de le générer valablement à partir du document original.

Pour l'utilisateur, ce sont essentiellement le contenu et l'authenticité du document qui importent (plus que la mise en forme). Dès lors, une migration du fichier vers un format plus simple (type ODF) pourrait être envisagée, tout en veillant à garantir l'authenticité du document lors de cette migration.

uelle que soit la migration effectuée, il faudra au minimum que SIGeDIS conserve :

- la version originale du document envoyée par le producteur afin de pouvoir prouver l'intégrité et l'authenticité du document reçu ;
- la dernière version migrée du document.

Enfin, par souci de sécurité, SIGeDIS pourrait également conserver l'avant-dernière version migrée du document afin de pouvoir en repartir en cas de constatation *a posteriori* d'anomalies générées par la dernière migration.

De même, SIGeDIS devra veiller à exécuter régulièrement une migration logicielle (que ce soit vers une nouvelle version ou vers un logiciel différent), en veillant à la compatibilité ascendante (notamment en ne sautant pas de version majeure). Il est conseillé de ne pas compresser ou crypter les données.

6.4. La préservation de la signature digitale

La préservation de la validité de la signature numérique sur des périodes très longues est difficile étant donné que la durée de validité des certificats est limitée dans le temps. La solution consiste donc à créer une chaîne de confiance continue entre le moment de la signature et le moment présent afin qu'aucun doute ne puisse survenir sur la validité de la signature.

Pour cela, il faut horodater périodiquement l'objet signé ainsi que sa signature. L'horodatage est un processus qui appose sur un objet numérique un *timestamp* qui est une séquence de caractères contenant suffisamment d'informations, la plupart du temps une date et une heure, pour situer un événement dans le temps. De ce fait, l'horodatage certifie que l'objet numérique est aujourd'hui tel qu'au moment où le *timestamp* a été apposé. Donc, si un *timestamp* est apposé à un instant *t*, il prolonge la validité de la signature jusqu'à la date d'expiration du certificat du serveur d'horodatage, qui vraisemblablement est plus lointaine que celle du signataire.

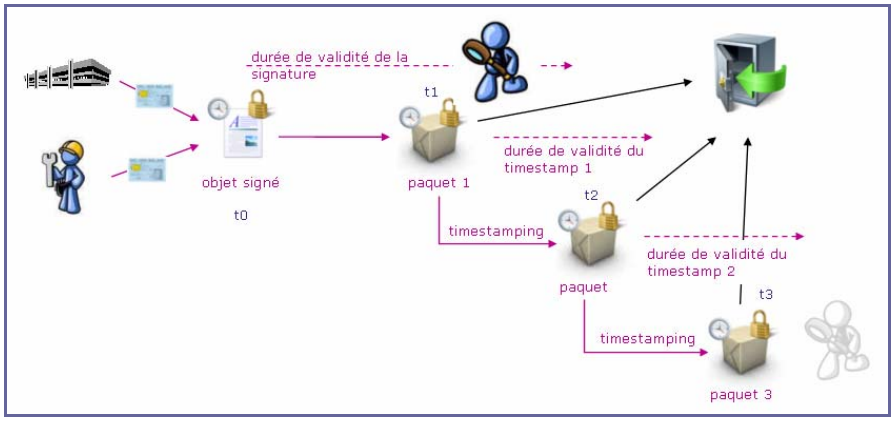
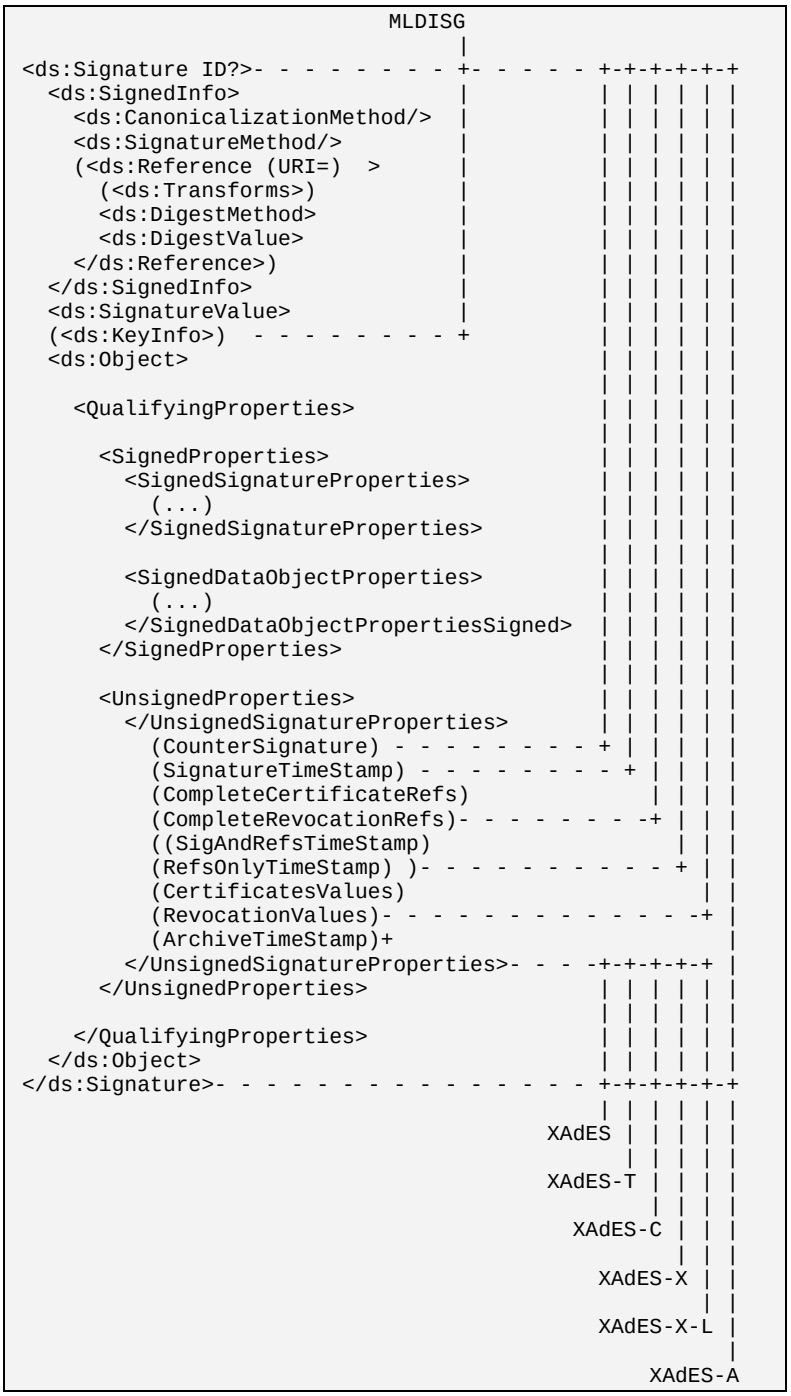


Figure 34 : Création d'une chaîne de confiance en vue de préserver à long terme la validité d'une signature numérique

Comme l'illustre la Figure 34, supposons que la signature apposée au temps t0 est valide au temps t1 (mais plus si le *timestamp* apposé en t1 est apposé après l'expiration du certificat de la signature utilisée en t0), date à laquelle on appose un nouveau *timestamp*. Ce dernier prolonge la validité de la signature jusqu'à la date d'expiration du certificat du serveur d'horodatage utilisé en t1. Et ainsi de suite. Dès lors, une chaîne de confiance peut être établie entre *t3* et *t1*, ce qui signifie que si la signature utilisée en *t3* est valide, celle apposée en *t0* l'était aussi.

Concrètement, il faut donc apposer un nouveau *timestamp* de manière régulière. La norme XAdES-A(rchiving) (cf. ci-dessous) permet cette opération par l'ajout successif d'un nouveau bloc XML « `ArchiveTimeStamp` ».

Pour obtenir une signature XAdES-A, il faut respecter la signature XAdES-X-L qui inclut les certificats. De cette manière, si l'autorité de certification disparaît, les certificats sont toujours disponibles pour vérification.



Structure d'une signature respectant la norme XAdES et éléments présents dans chaque forme de la norme

Pour déterminer le moment où il faut re-horodater le document signé et ses signatures, deux solutions sont possibles :

1. On se base sur la fin de validité du dernier certificat utilisé (c'est-à-dire celui utilisé dans la dernière balise « `ArchiveTimeStamp` ») et on appose un nouveau *timestamp* avant la fin de validité du dernier

certificat. Cette solution ne permet cependant pas de se prémunir contre les failles ou l'évolution rapide de la cryptographie.

2. On appose un nouveau *timestamp* périodiquement, par exemple tous les ans.

Dans le cadre de l'archivage des contrats de travail chez SIGeDIS, le mécanisme décrit ici devra être mis en œuvre pour la préservation de la signature digitale des documents entrants afin que SIGeDIS puisse certifier que le document original en sa possession est bien le document envoyé par le tiers-archiveur.

Chez celui-ci, vu que le document pourra y être archivé potentiellement sur de longues périodes, le tiers-archiveur devra pouvoir lui aussi appliquer ce mécanisme d'horodatage régulier. Si le tiers-archiveur respecte la norme XAdES-A, lors de la réception, SIGeDIS devra contrôler la validité du dernier « `ArchiveTimeStamp` » puisque c'est l'élément qui permettra de vérifier la chaîne de confiance. Si cet élément n'est pas valide (c'est-à-dire si les signatures et certificats qui s'y trouvent ne sont plus valides), SIGeDIS ne peut pas être sûr de disposer du document original et devra donc refuser le document.

7. Conclusion

Au vu de la place de plus en plus prépondérante que les technologies de l'information et de la communication prennent au sein des sociétés et des institutions publiques, la préservation à long terme de l'information numérique est devenu un enjeu crucial (quantité croissante d'informations numériques, réglementations imposant leur conservation sur des durées relativement longues, enjeux financiers importants, nombreux types d'information – dans certains cas très complexes – à préserver).

N'étant par nature pas auto-explicative, l'information numérique naît de l'interaction entre une séquence de bits et des éléments hardware et software, la rendant dès lors soumise à l'évolution hétérogène de ces différentes composantes. Le problème est que l'ère du numérique dans laquelle nous sommes entrés est marquée par une évolution impressionnante des technologies (les unes remplaçant les autres).

Inclus dans l'archivage, la préservation consiste à maintenir les objets archivés en état, c'est-à-dire accessibles et compréhensibles par ses utilisateurs. Du fait de la fragilité inhérente de l'information numérique, sa préservation nécessite d'appliquer de manière continue tout au long du cyle de vie de l'information des stratégies techniques et conceptuelles qui ne sont efficaces que si elles sont encadrées par une organisation financée durablement.

Au niveau de l'organisation, un préalable indispensable est un engagement fort de la direction eu égard aux budgets et compétences qui devront être rassemblés. Tout projet de ce type doit commencer par une étude du modèle conceptuel *Open Archival Information System* (OAIS) qui est devenu progressivement une norme incontournable dans le domaine, qui fut normalisée en 2003 par l'ISO. Permettant de saisir la problématique de manière globale, il constitue un excellent guide pour la mise en œuvre de projets d'archivage à long terme. Pour être préservée, l'information numérique doit être maintenue dans un système qualifié de fiable sur les plans de l'organisation, de la gestion et des stratégies techniques et conceptuelles mises en œuvre. Divers modèles d'audit (dont le plus élaboré est DRAMBORA) existent et offrent une aide efficace pour évaluer la capacité d'un système à préserver l'information.

Le modèle OAIS n'offrant qu'un modèle conceptuel, il reviendra à chaque organisation de traduire cette organisation en différents services, chacun assumant une partie des tâches et des responsabilités.

Une fois ce cadre organisationnel élaboré, l'organisme souhaitant préserver ces informations doit recourir à diverses stratégies techniques et conceptuelles, appliquées de manière continue. Étant donné qu'il n'existe aucune solution globale et unique, nous insistons sur l'importance de combiner ces stratégies.

Quatre stratégies complémentaires, techniques et conceptuelles sont actuellement opérationnelles :

- la **gestion des supports de stockage**, incluant le choix des supports, leur contrôle régulier et leur remplacement ;
- la **gestion des formats**, comprenant le choix de formats qualifiés de pérennes, leur validation et le recours à des *format viewers* ;
- la **migration régulière** des données vers des architectures matérielles et logicielles plus récentes, en veillant à la compatibilité ascendante des logiciels et en prenant soin de documenter rigoureusement le processus de migration ;
- le **recours aux métadonnées**, base indispensable de toute autre stratégie. À cet égard, les standards développés ces dernières années (dont METS et PREMIS) offrent une aide indéniable.

Chacune de ces stratégies permet de préserver une ou plusieurs couches (physique, binaire, logique et sémantique) de l'information numérique (Figure 35).

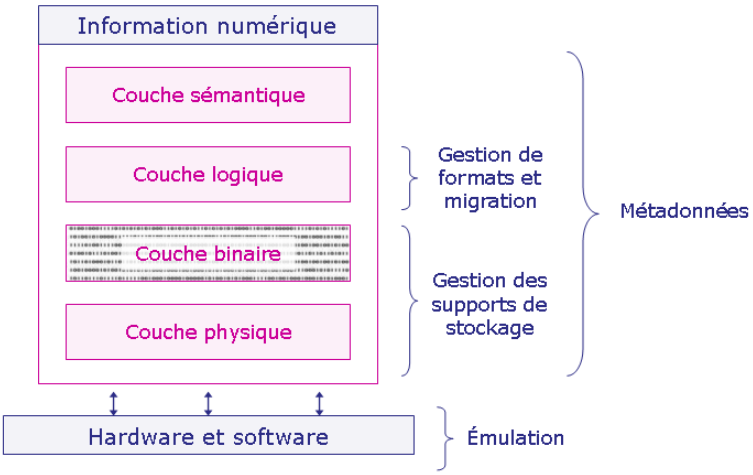


Figure 35 : Modèle en couches et positionnement des stratégies

Deux autres stratégies sont parfois présentées. Dans le cas de l'encapsulation, la stratégie est intéressante mais encore peu implémentée aujourd'hui, tandis que l'émulation n'est clairement pas opérationnelle à l'heure actuelle et ne doit être utilisée que dans des cas bien spécifiques pour lesquels il n'existe aucune autre solution.

L'application de ces stratégies à un cas concret issu de la sécurité sociale a montré leur complémentarité et leur faisabilité.

Le coût de la préservation demeure cependant un problème complexe à gérer, d'autant plus que ce coût n'est que difficilement chiffrable. Dès lors, en vue de le diminuer, diverses stratégies de mutualisation peuvent être mises en œuvre.

Enfin, au niveau de l'offre logicielle, aucun système commercial actuel n'est en mesure de faire de l'archivage *pérenne*. L'intégration de cette problématique dans les systèmes d'archivage représente donc un enjeu de taille aussi bien pour les éditeurs de logiciels que pour les intégrateurs et développeurs de solutions sur mesure. Affaire à suivre certainement !

8. Bibliographie

Lectures recommandées

BORGHOFF U.M. *et al.*, *Long-Term Preservation of Digital Documents. Principes and Practices*, Éd. Springler, New York, 2003.

BOYDENS I., La préservation à long terme de l'information numérique, *Techno* n°28, Smals, 2004.

CHABIN M.-A. *et al.*, *Dématérialisation et archivage électronique. Mise en œuvre de l'ILM*, Éd. Dunod, Paris, 2006.

HARVEY R., *Preserving Digital Materials*, Éd. Saur, Munich, 2005.

HUC Cl. *et al.*, *L'archivage numérique à long terme. Les débuts de la maturité ?*, Éd. La documentation française, Paris, 2009.

HULSTAERT A., *Préserver l'information numérique. Codage et conversion de l'information*, Delivrable, Section Recherches, Smals, 2008.

HULSTAERT A., *Digital Record Object Identification (DROID) - File Format Identification Tool*, Quick Review n°22, Section Recherches, Smals, 2009.

HULSTAERT A., *JSTOR/Harvard Object Validation Environment (JHOVE) 1.5 - File Format Identification and Validation Tool*, Quick Review n°25, Section Recherches, Smals, 2010.

RECTEM C., *La pérennisation digitale dans le secteur public : étude critique de plusieurs modèles de méta-information*, Mémoire de la section STIC de l'Université Libre de Bruxelles, Bruxelles, 2009.

ROTHENBERG J., *Ensuring the Longevity of Digital Information*, 1999.

THIBODEAU K., Overview of technological approaches to digital preservation and challenges in coming years, *The State of digital preservation : an international perspective*, Washington, 2002.

Webographie recommandée

Norme OAIS : http://vds.cnes.fr/pin/documents/projet_norme_oais_version_francaise.pdf

Norme PAIMAS : http://vds.cnes.fr/pin/documents/030115_CCSDS_651_R1.pdf

METS : http://www.loc.gov/standards/mets/

PREMIS : http://www.loc.gov/standards/premis/

Moreq2 : http://www.moreq2.eu/panellists.htm

Site du groupe PIN (Pérennisation de l'information numérique) : http://www.aristote.asso.fr/PIN/

Digital Preservation Tutorial, Cornwell University, 2005 : http://www.icpsr.umich.edu/dpm/

Expertise Centrum eDavid : http://www.edavid.be/

Serveur PADI – Preserving Access to Digital Information : http://www.nla.gov.au/padi/

Digital Preservation Europe : http://www.digitalpreservationeurope.eu/

Digital Curation Centre : http://www.dcc.ac.uk/

Site des Archives Générales du Royaume (Belgique) : http://www.arch.be/

DLM Forum : http://www.dlmforum.eu/

9. Glossaire

AIP	Archival Information Package (OAIS)
AIT (bande)	Advanced Intelligent Tape
BCE	Banque-Carrefour des entreprises
CCSDS	Consultative Committee for Space Data Systems
DIP	Dissemination Information Package
DRAMBORA	Digital Repository Audit Method Based on Risk Assessment
EAD	Encoded Archival Description
HSM	Hierarchical Storage Management
ILM	Information Lifecycle Management
ISO	International Organization for Standardization
JPEG 2000	Joint Photographic Experts Group 2000
LOTAR	Long Term Archiving and Retrieval of Digital Technical Product Data
LTO (bande)	Linear Tape-Open
METS	Metadata Encoding and Transmission Standard
MoReq	Model Requirements for the Management of Electronic Records
NISS	Numéro d'identification de sécurité sociale
OAIS	Open Archival Information System
ODF	Open Document Format
OOXML	Office Open XML
PAIMAS	Producer-Archive Interface Methodology Abstract Standard – ISO 20652
PAIS	Producer-Archive Interface Specification
PDF/A	Portable Document Format / Archiving
PREMIS	Preservation metadata : implementation strategies
RM	Records Management
SIGeDIS	Sociale individuele gegevens - données individuelles sociales (parastatal)
SIP	Submission Information Package (OAIS)
VERS	Victorian Electronic Records Strategy
XadES-A	XML Advanced Electronic Signature - Archiving
XFDU	XML Formatted Data Unit
XMP	Extensible Metadata Platform (Adobe)