La préservation des documents numériques: Enjeux et stratégies

ALIOUALI Nadia , DAHMANE Madjid,
Centre de Recherche sur l'Information Scientifique et Technique
naliouali@mail.cerist.dz
mdahmane@wissal.dz

Introduction

Les Institutions culturelles, les centres d'archives et les bibliothèques sont, de par leur nature, leurs missions, chacune à sa manière et chacune en ce qui la concerne, les dépositaires de la mémoire des sociétés et elles prennent, de ce fait, la responsabilité de sa préservation. Cette responsabilité ne cesse de grandir face à une explosion informationnelle et un développement technologique sans cesse croissant

Par ailleurs, l'avènement technologique à engendré l'apparition de divers support, ce qui a permis aux sociétés de changer leurs mode de communication, de production, de stockage et de diffusion de l'information.

Les CD, les DVD les disques durs fixes ou transportables, et toutes les autres formes d'unité de stockage ne sont que des contenants d'informations. Par les informations qu'ils contiennent ils constituent une partie de la mémoire collective. Mais une mémoire qu'elle qu'en soit la forme, demeure toujours caractérisée par sa fragilité et toutes les caractéristiques qui l'exposent au risque de la perte partielle ou totale.

Pour le patrimoine documentaire, le danger réside dans la dégradation, la perte et la destruction sous toutes les formes possible. Les archivistes et les documentalistes ont toujours vécu le problème de la poussière, le syndrome du vinaigre, la déformation des supports, les conditions climatiques, etc. Ce qui nous amène à dire que le problème de la conservation et de préservation de la mémoire collective des sociétés attire l'attention des chercheurs, des expert et des praticiens dans ce domaine depuis bien longtemps.

L'avènement des technologies de l'information et de la communication à permis de diversifier et de multiplier les masses d'information en circulation entre les hommes. Ils ont tout aussi aidé à donner naissance à de nouvelles formes de documents et de supports (E-books, blogs, sites web….), des formats de fichiers en développement permanent. Mais toujours est-il que ces derniers sont caractérisés par un cycle de vie

assez réduit. Quelles est la durée de vie d'un site ou d'une page web ? Combien de temps peut tenir un CD ? En somme il s'agit de conserver les softwares et les hardwares qui leurs correspondent comme il a été souligné par l'UNESCO dans son rapport final lors du Sommet Mondiale sur le Société de l'Information qui s'est déroulé à TUNIS en 2005[1].

Aujourd'hui la, d'un côté, la quasi-totalité de l'information créée, gérée, transmise, stockée dans notre société est sous forme numérique et d'un autre côté, la force probante de l'écrit électronique, va permettre diminuer l'utilisation ou de se passer entièrement du support papier. Ne pas être en mesure de préserver convenablement et durablement cette information conduirait à une société sans mémoire. Que deviendrait une société qui perdrait progressivement les avancées scientifiques et technologiques qui lui auront coûté tant de ressources humaines et financières ?
Le problème est également complexe. Il réside aussi dans l'obsolescence constante des technologies: comment préserver durablement de l'information avec une technologie qui n'a pas de pérennité et qui nous conduit à changer les systèmes d'exploitation, les logiciels, les ordinateurs à peu-prêt tous les 5 ans et les médias de stockage tous les 7 ou 8 ans?
Dans un tel contexte, la préservation à long terme ne s'applique pas seulement à la conservation définitive des documents. Le long terme sera défini comme une action qui étudiera l'information contenue dans un entrepôt numérique les changements technologiques, et notamment la gestion des nouveaux médias et formats de données autrement dit étudier le cycle de vie de l'information depuis son intégration dans un système d'information jusqu'à son archivage, afin de lui définir sa traçabilité et lui permettre l'accès à long terme.

L'information étant l'élément fondamental de tout savoir. Sa préservation dés l'ors préservation des sources de savoir qui donneront lieu un jour ou l'autre à des progrès scientifiques et technologiques, car cette préservation pourra renseigner les générations futures sur nos préoccupations, notre culture, notre niveau de développement et l'état de nos connaissances.

Devant cet état de fait comment communiquer sans détériorer, comment conserver tout en en communiquant ? Pour pouvoir remplir ces deux missions dans des conditions satisfaisantes, il est nécessaire d'élaborer une politique de préservation à long terme dont l'objectif est de prévenir, d'arrêter ou de retarder la détérioration des documents et, si

[1] Sommet Mondiale sur le Société de l'Information.Tunis 2005
http://www.itu.int/wsis/docs2/tunis/off/7-fr.html

nécessaire, d'améliorer leurs conditions de conservation, ou de préserver au moins le contenu et d'en permettre l'accès à long terme au générations futures .

1-Problématique de la préservation dans les bibliothèques

Longtemps, la préservation a été limitée à la conservation et à la restauration des documents anciens, rares et précieux. Ces documents faisaient l'objet d'une attention particulière de la part de l'archiviste ou du bibliothécaire. Devant l'étendue des altérations dues à l'augmentation de la consultation des fonds, au non respect des recommandations en matière de conservation et à la mauvaise qualité des matériaux constituant les documents. Le document numérique, tout comme le document papier ou audiovisuel, n'est pas à l'abri des dommages et de la dégradation. De ce point de vue, les défis de la conservation à long terme de données numériques sont semblables à ceux qui confrontent les collections culturelles des générations antérieures.

Mais la conservation d'objets numériques a ses défis propres liés à la nature même des données numériques qui sont lisibles par des machines mais non par l'être humain,

En effet Un document numérique est un ensemble complexe de contenus d'information et de paquetage de ces derniers dans un format de données accompagné d'un programme informatique offrant les fonctionnalités de manipulation pour la lecture, la recherche. Le tout est stocké sur un support. Une meilleur préservation nécessite de connaître toutes les composantes technique pour garantir l'accès aux contenus malgré l'obsolescence des formats et des supports utilisés lors de la création du document numérique.

La conservation de données numériques sous une forme compréhensible pour l'être humain fait intervenir l'utilisation d'un ensemble complexe de techniques interdépendantes. De nombreux rapports, issues de projets de préservations, expliquent en détail pourquoi la conservation d'objets numériques représente un tel défi : Obsolescence technologique du matériel, des logiciels et des formats, vulnérabilité des supports, problèmes organisationnels et juridiques.

Dans ce sens plusieurs programmes de préservation ont été mis en place. Parmi ces programmes on peut citer

- Le programme fondamental «Préservation et Conservation» de l'IFLA (PAC) Créé lors du congrès de l'IFLA («l'International Fédération of Library Associations and Institutions) à Nairobi en 1984, le Programme fondamental

"Préservation et Conservation" est devenu effectif en 1986. Depuis 1992, le siège du PAC est à la Bibliothèque nationale de France à Paris.

- Le programme « Mémoire du Monde » initié par l'UNESCO en 1993 avec l'appui du Centre international du programme PAC, à la Bibliothèque nationale de France (Paris). Ce programme vise à sauvegarder le patrimoine documentaire mondial par les techniques les mieux appropriées, à en démocratiser l'accès et à faire mieux prendre conscience de son intérêt et de la nécessité de le préserver.
- **INTERPARES [2](International Research on Permanent Authentic Records in Electronic Systems)** un des plus importants programmes de recherche et développement consacré à la conservation à long terme des données numériques. Une vingtaine de pays y participent (par le concours de chercheurs, d'institutions d'archives, de producteurs de contenus et d'industriels). Il tend à proposer des standards, des principes d'organisation et des recommandations pour la mise en place par les gouvernements, par les institutions culturelles et patrimoniales et par les acteurs industriels, des actions concourant à la conservation des données numériques.
- **NEDLIB (Networked European Deposit Library)[3] :** Programme qui a été soutenu par la Commission Européenne de 1998 à 2001. Il regroupe huit grandes bibliothèques européenne Il vise à la constitution d'un modèle de spécifications fonctionnelles et techniques relatif aux documents électroniques publiés sur support ou diffusés sur le Web.

Garantir un accès permanent aux documents numériques malgré les changements technologiques représente le même défi (sinon un défi plus important) qu'avec des documents traditionnels. Sauf que ces dernières années, le volume d'information stockée électroniquement ne cesse de s'accroître (bibliothèques numériques, web, blogs, intranets, extranets. Avec les particularités qu'il présente, ce nouveau type de document à réorienté complètement la problématique de la préservation au niveau des bibliothèques, et centres d'archives, et a changé leurs pratiques en la matière.

2-Les enjeux de la préservation des documents électroniques

2.1-La structure du document numérique

Contrairement au document papier dont le support et le contenu sont intrinsèques, le document numérique comporte plusieurs « couches »4. La couche physique qui

[2] http://www.interpares.org/
[3] nedlib.kb.nl/
[4] Catherine Lupovici. Les besoins et les données techniques de préservation. in 16 th IFLA council and general conference. August 16th-25th, 2001 http://www.ifla.org/IV/IFLA67/163-168f.pdf

correspond aux données d'information qui facilite l'accessibilité et la compréhension du document, la couche binaire qui est une suite de code binaire de 0 et 1 et la couche structure qui est en fait un langage de programmation qui permet l'assemblage des données binaires et leurs interprétation, la couche application à son tour permet de transformer les données structurées en objet signifiant à travers un format JPEG, HTML, XML,…etc. Les objets résultant sont prêts à être manipulés par des logiciels adéquats pour leur lecture et leur manipulation.

Donc le contenu d'un document numérique est constitué de tout ce paquet de données codées, la perte d'une partie implique la perte de l'intégrité et de la valeur du document, ce qui implique aussi la préservation de la totalité dans le cas d'un processus de conservation, car préserver une partie ne garantie pas un accès ultérieur au document lui-même.

2.2-Le vieillissement du support

La fragilité et la durée de vie[5]; des supports d'information est une question qui est très ancienne au niveau des bibliothèques mais qui demeure d'actualité quand il s'agit de la conservation des documents électronique.

Avec le numérique la famille des supports s'agrandit. Les institutions chargés de collecter et de diffuser on dû intégrer au sein de leurs collections ces nouveaux supports d'édition et de stockage. Mais chacun sait qu'aucun support n'est éternel. Selon l'expertise de certaines bibliothèques comme la Library of Congrès ou la BNF, il apparaît que le disque en plastique utilisé pour les CD audio, les CDI, les CD ROM pressés et dupliqués pourraient n'avoir qu'une durée de vie de l'ordre de 10 à 25 ans dans les conditions moyennes de conservation et d'utilisation. Le disque enregistrable n'aurait qu'une durée de vie de l'ordre de 3 ans avant d'être gravé (en raison du vieillissement de la couche sensible) et une durée de vie sans altération de l'ordre de 5 à 10 ans une fois gravé. Plus fragiles que les supports utilisés jusqu'alors, ayant une espérance de durée de vie encore plus courte, les nouveaux supports utilisés pour les documents électroniques imposent des actions de conservation, de restauration, et de rafraîchissements périodiques.

2.3-L'obsolescence des techniques

Plus l'information est dense en termes de surface (physique, binaire, structure et application), plus compliqués sont les dispositifs permettant de la rétablir sous une forme perceptible à un être humain. Entre le fichier informatique, qui représente le

[5] Digital Preservation Coalition. Media and format.
Http://www.dpcoline.org/praphics/medfor/media.html

document numérique, et la machine s'intercalent plusieurs couches logicielles, comme les systèmes de codage, les systèmes d'exploitation, les programmes, qui toutes doivent être compatibles à la fois avec le document et avec le type d'ordinateur utilisé. Cette dépendance technique de l'utilisateur qui constitue une contrainte importante, devient une difficulté majeure dans le cadre d'une politique de préservation des documents numériques. En effet, on estime que le cycle de validité des programmes et des périphériques est de l'ordre de 2 à 56 ans.

Passée cette période, un certain nombre de documents ne seront plus accessibles. En vue d'illustrer ce problème, Julia Martin et David Coleman7 ont développé l'exemple suivant : Un chercheur commence son travail en 1988 sur un ordinateur IBM 286. Il sauvegarde ses fichiers sur des disquettes de 5 pouces. Dans les années 1990, ce même chercheur fait l'acquisition d'un IMAC, ordinateur qui ne comporte qu'un lecteur de cédéroms. Le chercheur souhaitant consulter ses archives de 1988 sera donc dans l'impossibilité de le faire.

Mais le problème se pose également au niveau d'une couche plus profonde de la lecture du document, il s'agit des logiciels et des applications nécessaires. Ainsi, la compatibilité ascendante entre les versions successives d'un logiciel peut être assuré, mais ce n'est pas toujours le cas.

D'un point de vue général, il est donc possible de dire que la mise en place d'une politique d'archivage de documents numériques, quels qu'ils soient, nécessite de prendre en considération les aspects techniques liés à l'environnement de chaque document. Mais aussi il faut étudier le contexte économique général dans lequel les équipements et les formats informatiques sont produits.

2.4-Des contenus hypertextuels

Les documents numériques dont les sites web, sont des objets particulièrement complexes. Lors de leur production, ils intègrent plusieurs types de documents : des textes, des images, des fonctionnalités, parfois du son et des animations. L'ensemble est structuré par des liens hypertextuels, généralement en HTML. Chaque type de document inclus dans un site Web intègre donc des formats différents. Par exemple un site Web peut comporter des images en format graphique GIF, des documents textuels en .DOC et des animations au format Flash le tout structuré en HTML.

---

[6] Chiffres donnés dans la plupart des articles cités en bibliographie.

[7] **MARTIN, Julia et COLEMAN, David.** The archive as an ecosystem. In *Michigan University*. http://www.press.umich.edu.jep/07-03/martin.html

Selon Peter Lyman8, une page Web contiendrait en moyenne quinze liens à d'autres pages et cinq objets différents (images, sons ou autres).
Ainsi, les problèmes posés par la dépendance technique aux formats pour tout type de document numérique se multiplient dans le cas des sites Web du fait même de leur complexité et de l'intégration de plusieurs formats.

2.5-Un environnement instable

Un site Web est un document très instable. Il naît à un instant T et disparaît ensuite sans laisser de trace, ou encore il est mis à jour constamment en donnant naissance à plusieurs versions du même site. Ce qui nous amène à nous demander a partir de quel moment peut-on considérer qu'une modification est significative pour qu'elle puisse donner lieu à une nouvelle version du site, et combien de version qu'il faut conserver d'un même site Web. Outre cette variabilité, un site Web est aussi un document éphémère. Selon une étude commandée par l'OCLC9, la durée de vie moyenne d'un site Web serait de six semaines.

2.6-Des documents sans limites

Nous l'avons vu précédemment, le document numérique est formé de plusieurs couches correspondant chacune à un certain niveau du document.
De ce fait, la question de l'archivage des sites Web ou de toute forme de documents numériques correspond à un choix fondateur : à quel niveau d'abstraction du document décide-t-on de commencer à préserver et plus concrètement, quelle partie du document doit-on conserver ? Pour une application informatique, doit-on conserver la présentation ou faut-il conserver l'ensemble des fonctionnalités. Il est plus intéressant de conserver touts ces aspects. Cependant, il faut savoir que, plus l'on se place à un haut niveau d'abstraction du document, intégrant ainsi toutes les couches logiques (binaire, structure, objet, application), plus les conditions de sa conservation sont complexes et surtout dans le cadre des sites web
En effet, les sites Web, pour la plupart, renvoient vers d'autres sites par le biais de liens hypertextes. Le fait de renvoyer l'utilisateur à un autre site ce dernier constitue donc un élément du site de renvoie. Par ailleurs, ces liens hypertextes participent de la navigabilité d'Internet et font partie. Donc la question que se posent les spécialistes

[8] **LYMAN , Peter.** Archiving the World Wide Web. In *The journal of electronic publishing* http://quod.lib.umich.edu/cgi/t/text/text-idx?c=jep;view=text;rgn=main;idno=3336451.0007.301

[9] **OCLC (Online Computer Library Center).** Web characterization. In *OCLC's Website.* http://wcp.oclc.org

dans le domaine de la conservation et la préservation est faut-il alors archiver à la fois les sites Web eux-mêmes et ceux vers lesquels ils pointent un lien.

2.7-L'authentification sur Internet.

L'authentification des documents numérique sur internet est un grand problème à considérer lorsqu'il s'agit de mettre en place une politique de conservation et de préservation à long terme. En effet, chacun de nous a été confronté, à un moment ou à un autre, à des difficultés pour évaluer le site qu'il visualise. L'auteur ou le créateur du site n'est pas toujours clairement identifié, ainsi que la date de la dernière modification, le titre de la page Web n'est pas toujours significatif et n'est parfois formé que par les premières lignes du texte de la page. La date de modification et la date de création des sites Web ne sont pas toujours bien indiquées. De ce fait, dans le cas des sites Web, il y a toujours une série de confusions entre la date de création du site, la date de sa dernière modification et la date de consultation du site.

D'un autre côté, le nom du domaine joue un grand rôle dans l'identification du type de site, (.org, .net…) de sa localisation (.fr, .dz…) et de son objectif (commercial .com ou non).

Par ailleurs, les sites Web ne comportent pas d'identificateur unique comme l'est par exemple l'ISBN pour les documents papiers. Dans le cadre d'une politique de conservation cette absence est problématique, car on peut avoir le même site qui est archivé dans deux serveurs différents

L'utilisation d'un identificateur unique pour chaque site, page ou autres sur Internet est une nécessité dans le cadre d'une politique de conservation et préservation à long terme, tout d'abord pour identifier à tout moment et le plus rapidement possible un site et en vue de faciliter la gestion des différentes versions d'un même site,

2.8-Problème juridique

Nous savons que toute conservation d'un document numérique entraîne la modification sur celui-ci surtout dans le cas d'une migration tout en sachant que cela touche à l'intégrité du document. Particulièrement les sites web sont reconnus comme des œuvres et sont soumis à une réglementation du droit d'auteur qui protège l'œuvre et son auteur de toute tentative de changement, ce qui n'est pas inévitable dans le cadre d'une politique de conservation.

Les droits d'auteur se composent à la fois de droits moraux et de droits patrimoniaux. Les droits moraux protègent l'auteur et son œuvre de toute dénaturation. Or les procédures de conservation des sites Web peuvent entraîner la modification de ceux-ci, et donc une perte d'intégrité contraire aux droits moraux et plus exactement au droit au

respect de l'oeuvre110. De la même façon, dans le cadre du référencement des sites Web, une bibliothèque peut faire le choix d'intégrer des métadonnées dans le fichier informatique du site Web. Cette intégration pourrait être considérée comme une dégradation du site et risque même d'être considérée comme une atteinte aux droits d'auteur. A ce moment là les données d'ordre juridique risquent de devenir une contrainte importante pour les bibliothèques, ne serait-ce que parce que cette contrainte pourrait infléchir les conditions de valorisation des fonds archivés et empêcher, par exemple, la mise en ligne de ces archives.

3-Quelques mesures de préservation

Un document numérique dépend d'un environnement technologique donné et celui-ci sera inévitablement obsolète après un certain nombre d'années ce qui implique la perte du document lui-même. A cet effet, un certain nombre de stratégies ont été mises au point pour résoudre ce problème et sauver à temps le patrimoine numérique.

Ces stratégies s'inspirent de deux approches:

La première approche consiste à préserver intégralement l'environnement technologique d'origine afin de pouvoir reproduire l'objet numérique à l'avenir également. Dans ce cadre, on distingue deux possibilités : la conservation de la technologie et l'émulation de la technologie. La conservation de la technologie implique la conservation de répliques en bon état de fonctionnement de l'ensemble du matériel et des logiciels. L'émulation consiste à programmer les ordinateurs actuels et futurs pour qu'ils puissent émuler les systèmes et plates-formes vieillis et devenus obsolètes.

Ces deux techniques ne sont guère envisageable car la première implique de maintenir en fonction un matériel périmé tandis que la deuxième, outre sa complexité et sont coût, s'avère impossible à mettre en œuvre surtout dans le cas ou les logiciels étaient couplés à du matériels spécifiques

La deuxième approche vise à empêcher de manière directe le vieillissement des formats numériques. On peut à nouveau distinguer deux techniques : la migration des informations et l'encapsulation. La migration consiste à transférer périodiquement des données numériques d'une configuration matérielle et logicielle à une autre ou d'une génération d'ordinateurs à la suivante. L'objectif de la migration est de conserver l'intégrité des documents numériques et de perpétuer la capacité des usagers à les

[10] **BENSOUSSAN, Alain (dir.).** Internet : Aspects juridiques. Paris : Hermès, 1998, 2° éd. revue

retrouver, les afficher et les utiliser alors même que la technologie évolue. Il est bien sûr plus facile de copier des fichiers dont les formats sont normalisés.
L'encapsulation, désignée aussi sous l'appellation « voie de référence » est une technique issue du modèle OAIS (Open Archive Information System). Elle se présente donc comme un moyen de regrouper autour du contenu lui-même, les informations contenant les instructions nécessaires au décodage de ses bits dans le futur par n'importe quel système. Elle propose une succession de couches d'instruction. Une couche externe sur laquelle sera porté un texte lisible, décrivant le contenu de l'élément encapsulé ainsi que la manière de l'utiliser. Une autre interne sur laquelle figureront les caractéristiques du logiciel, du système d'exploitation et du matériel à reconstituer en vue de la lecture de l'objet lui-même. L'encapsulation, qui permet de rendre autonome le contenu et l'information qui lui est attachée, semble être une méthode relativement viable pour la conservation à long terme en particulier pour des fichiers texte. Elle est encore incertaine pour les autres documents qui souffrent de la dispersion technologique et du trop grand nombre de nouveaux logiciels, de systèmes de compression ou de formats mis sur le marché chaque année.

4- Les outils techniques nécessaires à la préservation

Préserver sur le long terme, nécessite la mise en œuvre d'un certain nombre d'outils techniques**.** L'intégrité des documents électroniques conservés dépend, pour partie, de l'association cohérente de ces différents outils. Il est donc important de connaître leurs caractéristiques principales.

4.1-Les formats

Tout processus de conservation électronique, repose sur l'emploi d'un format de codage. Le codage exprime la manière dont l'information est structurée au sein du fichier de façon à pouvoir être conservée, transmise et échangée. Dans le contexte des documents électroniques, tout format de codage est inclus dans une chaîne d'autres codages et fait encore partie d'une chaîne d'éléments qui le rendent intelligible (par exemple, un fichier Word est lié au logiciel Word, lui-même mis en œuvre sur un certain modèle d'ordinateur et par un certain système d'exploitation. Par conséquent la notion de lisibilité d'un document électronique ne peut être comprise en dehors de l'interaction de l'ensemble de ces encodages avec le logiciel et le matériel informatique conçus pour les interpréter.

Les formats d'encodage se distingue entre eux par le fait qu'il soit ouvert ou fermés.. Des formats sont dits :

Ouverts quand les spécifications sont publiques
Fermés quand les spécifications sont tenues secrètes par le propriétaire,
Propriétaires lorsqu'ils sont définis par un organisme propriétaire et leur utilisation soumise à des droits,
Standards lorsqu'ils sont définis et adoptés par un organisme de normalisation, et que leur utilisation est libre et publié.

Il convient donc de mesurer à quel point les choix techniques opérés a priori peuvent retentir sur la restitution du document a posteriori. A ce moment là il est indispensable de choisir, dès l'origine, des formats considérés comme pérennes et d'effectuer, en temps voulu, les conversions nécessaires pour maintenir la lisibilité des données.

En effet, à travers toutes les expériences menées jusqu'ici en matière de stockage et de préservation des données électronique de part le monde, XML s'avère être le format privilégié pour plusieurs raisons. C'est un format universel et normalisé, créé pour structurer, stocker, sauvegarder et échanger les données. Il garantie à ses utilisateurs, l'indépendance de leurs documents de toute technologie propriétaire.
XML joue donc plusieurs rôles. D'abord pour l'échange des données, l'un des principaux challenges pour Internet aujourd'hui et dans le futur, par rapport aux obstacles rencontrés liés aux problèmes de l'hétérogénéité des formats.

Du côté de la pérennisation des données, l'avantage de ce langage réside dans ses caractéristiques à savoir :

- structurer un document,
- Séparer la présentation du contenu,
- styler un document,
- Le respect de la langue de l'auteur,
- La possibilité de restituer un même document dans plusieurs formats.

Tenant compte du faite qu'il soit un format universel et normalisé, indépendant de toute technologie propriétaire et gratuit.

4.2-Les supports

Le choix d'un format pérenne permet une meilleure préservation et assure une longévité aux données numériques. Mais il est toujours utile de faire des sauvegardes de sécurité sur des supports de stockage adéquats.

Le choix du support de stockage à utiliser est très difficile car la plupart des supports existant ont des avantages et des inconvénients. Ce qui nous amène à définir des critères pour le choix. Parmi ces critères nous citons :

- La capacité de stockage
- Le temps d'accès aux données
- Le coût
- La pérennité du contenu et la sécurité d'accès

Les supports actuellement utilisés pour l'archivage sont les suivants :

Technologie/support	Capacité	Longévité	Utilisation
WORM 12	30 Go	50 à 100 ans	Archivage permanent à très long terme
MO(*) 5.25	9 Go	30 ans	Application data-intensive
CD et DVD 120mm	3 Go	30 ans	Archivage économique
Bandes magnétiques	200 Go	10 ans	Très fortes capacités

(*) Disque magnéto-optiques

Caractéristiques des technologies d'archivage[11]

Toutefois le disque optique non-réinscriptible WORM (Write once-read many) est selon la norme AF Z 42-013 un procédé d'archivage privilégié. Non-réinscriptible, il empêche ainsi toute modification des données enregistrées et a une durée de vie estimée de 10 a 30 ans

4.3-L'identification et le référencement

Dans le cas d'une collection de document numérique et particulièrement ceux en ligne, le référencement constitue aussi une condition indispensable pour la conservation. Par conséquent, et suite à une opération de migration, le fait de connaître les formats de données utilisés dans le document, le système d'exploitation, le langage dans lequel le document a été écrit sont des informations fondamentales. Elles permettent de repérer les fichiers qui peuvent être mis en danger par une obsolescence logicielle ou matérielle et donc de procéder à une nouvelle migration. Ce type d'information ne peut être obtenu qu'à travers une opération d'identification et de référencement

[11] CD Rom techniques de l'ingénieur

Nous allons donc aborder ces deux aspects de la description des documents numériques conservés

4.3.1-L'identification unique

Plusieurs solutions sont envisageables pour attribuer aux documents numérique dont les sites web un identifiant unique :

- L'URL (Uniform Resource Location) est un identifiant unique pour chaque site Web sans autant être persistant dans la mesure où un même site peut changer d'URL c'est-à-dire de localisation

- L'URN (Uniform Resource Name). Comme l'URL constitue un nom d'identification unique pour chaque site Web et persistant. L'enregistrement de cet URN s'effectue auprès de l'IETF (Internet Engeneering Task Force).

- Le PURL (Persistent Uniform Resource Locator) a été conçu par l'OCLC. Il s'agit d'un identifiant de localisation comme dans le cas de l'URL, mais qui, contrairement à l'URL, serait persistant. Il se présente sous la forme d'un alias public. Il est créé par un administrateur de site Web et est enregistré comme « propriétaire » de PURL. Il maintient une mise en correspondance du PURL avec une URL.

- Le DOI (Digital Object Identifier) a été élaboré par « l'Association of American Publishers » et la « Corporation for National Research Initiatives ». Il s'agit d'un numéro d'identification unique et persistant ressemblant à l'ISBN. Il doit permettre d'identifier les objets numériques quels qu'ils soient. Ce système a été créé de façon à améliorer la défense du copyright pour les ressources en ligne et à faciliter le commerce en ligne de ressources électroniques.

4.4- Les métadonnées de pérennisation

On définit les « métadonnées » comme étant l'ensemble des informations techniques et descriptives ajoutées au document pour mieux le qualifier. Ces informations décrivent le contexte, la structure et le contenu des documents, ainsi que sa gestion dans le temps. Les métadonnées sont indispensables pour retrouver et accéder aux informations tout au long du cycle de vie des documents.

Le point de départ de tout travail autour des métadonnées est très certainement l'ensemble de métadonnées de référence du Dublin Core. Ces 15 éléments fondamentaux destinés à décrire toute ressource, au sens large, disponible sur Internet.

Mais à ce premier niveau généraliste de métadonnées descriptives, il est nécessaire d'ajouter des métadonnées plus techniques, spécialisées dans l'activité de pérennisation proprement dite. A savoir :

- Les métadonnées structurelles ou de conservation Il s'agit de rassembler un ensemble d'informations sur la structure informatique des documents : l'arborescence des fichiers, les formats des fichiers, le langage, les code de caractères….

- Les métadonnées de gestion comprennent les informations sur l'histoire du document dans l'institution de conservation. Il s'agit par exemple de savoir à quel moment la ressource a été enregistrée, le nombre de modifications (migrations) qu'elle a subies, les formats successifs dans lesquels les fichiers ont été enregistrés… Ces données renferment également des données sur la gestion de droits de consultation de la ressource : qui a le droit de la consulter ? Sous quelles conditions ? Quels sont les droits de reproduction des usagers?…

L'utilisation des métadonnées est l'un des éléments permettant de documenter et de suivre le processus de conservation. C'est notamment le cas des migrations de documents signés par un procédé de signature cryptographique à clef publique où il sera nécessaire d'inscrire dans les métadonnées associées au document les informations issues des procédures de vérification de signature Cette place fondamentale des métadonnées est reconnue au niveau international. L'utilisation des métadonnées permet également de participer à l'intégrité de la conservation. Les métadonnées participent, par les informations qu'elles contiennent, à la vérification de certains critères permettant de s'assurer de l'intégrité du document (spécialement la traçabilité).

5- La normalisation dans le domaine de la préservation des documents électronique

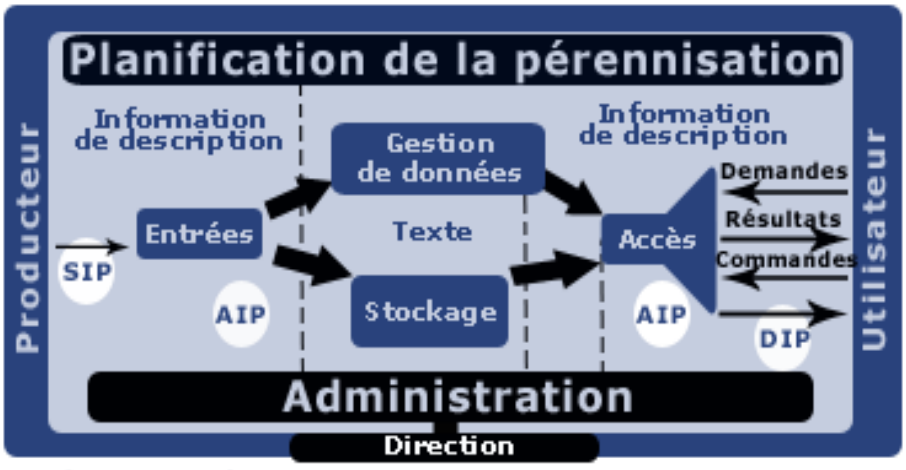
Actuellement, de nombreux processus de conservation électronique des documents font référence à une norme, internationale ou locale. Ces normes n'ont pas de caractère obligatoire mais peuvent être utilisées comme base technique.
Au niveau international, et dans le domaine de la préservation des documents électroniques, on peut retenir trois normes : la norme ISO 14 721 ou norme OAIS

(Open Archival Information System), la norme ISO 15 489-1 sur le records management et la norme ISO 19 005-1 sur l'utilisation du format PDF pour l'archivage (PDF/A-1).

L'OAIS constitue une référence décrivant dans ces grandes lignes les fonctions, les responsabilités et l'organisation d'un système qui voudrait préserver de l'information, en particulier des données numériques, sur le long terme. Il est l'aboutissement de réflexions menées à l'apparition de problèmes inédits : les risques de perte de données liés à l'obsolescence des technologies (codage ou supports).

L'OAIS est devenu en 2002 une norme (ISO 14721 : 2002) adoptée par tous les programmes de préservation et d'accès à long terme des bibliothèques. Ce modèle peut s'appliquer à n'importe quelle archive. C'est un modèle générique pour tous les types de documents qu'ils soient numériques ou traditionnels.

En gros, le modèle est conçu comme une boîte (voir figure ci-dessous) dans laquelle on manipule des paquets. Cette boîte a un ou plusieurs rôles ou missions, et elle interagit avec les producteurs des données en amont, les administrateurs du système, et la communauté d'utilisateurs en aval.



Modèle de système ouvert d'archivage d'information (OAIS)

Le modèle OAIS repose sur l'idée que l'information constitue des paquets, et que ces paquets ne sont pas les mêmes suivant qu'on est en train de produire l'information, d'essayer de la conserver, ou de la communiquer à un utilisateur. On a donc trois sortes de paquets :

- les paquets de versement (SIP) préparés par les producteurs à destination de l'archive

- les paquets d'archivage (AIP) transformés par l'archive à partir du SIP dans une forme plus facile à conserver dans le temps
- les paquets de diffusion (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 va trouver 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. Trois catégories de Métadonnées ont été définies

Métadonnées descriptives :

- Description bibliographique du document ; auteur, titre,…

Métadonnées administratives :

- Gestion des objets composant du document y compris les informations techniques qui permettront la préservation à long terme.
- Gestion des droits d'accès relatifs au document

Métadonnées de structure :

- Carte de la structure logique permettant d'assembler les différents composants logiques du document.
- Lien avec les objets numériques composant le document (le fichier).

Ainsi donc la mise en œuvre d'une archive OAIS requière que soit définis une institution responsable de la préservation à long terme des documents qu'on lui confie en vue de les communiquer à une communauté définie d'utilisateurs.

Pour la mise en œuvre d'un système d'archivage dans le cadre de l'OAIS, il faut tout d'abords définir quel sera le stockage, quels seront les formats acceptés dans les paquets, les formats de Métadonnées utilisés, les relations avec les producteurs, les services d'accès et de recherche. Ces paramètres sont à implémenter dans des logiciels adéquats. Toutefois, ces pré requis ne suffisent pas. Le modèle OAIS ne donne pas de clefs pour mettre tout ceci en œuvre et c'est sans doute la principale difficulté : comment passer du niveau conceptuel théorique à l'application sur le terrain en l'absence de logiciels de type « MyOAIS » clef en main.

Conclusion

La pérennisation du document électronique est un problème complexe et en perpétuelle évolution. Plusieurs aspects sont à prendre en compte : le support, le matériel, les programmes et les données, ainsi que la formation des acteurs de l'information numérique. Il n'existe pas encore une solution unique et universelle pour rendre pérenne le document numérique (sur support physique ou en ligne), Et ce en dépit des multiples tentatives émanant de grandes bibliothèques nationales à travers le monde. En effet, chacun d'entre elles adopte une solution propre à ses besoins locaux en matière de préservation. D'où la diversité des approches et des solutions.

Devant cet état de fait, on s'aperçoit que les efforts à faire en priorité ne sont pas seulement d'ordre technique. Ils sont plutôt :

- d'ordre culturel : un renforcement des collaborations entre informaticiens, archivistes et bibliothécaires est indispensable,

- d'ordre organisationnel : de nouveaux processus métier et de nouvelles méthodes sont à définir, à mettre en œuvre et à maîtriser.

En effet, la préservation des documents électroniques est une activité continue. Elle exige l'engagement et la participation non seulement des institutions qui s'occupent du patrimoine mais également des pouvoirs publics, des producteurs et utilisateurs de l'information, des fabricants de logiciels et des organisations et associations professionnelles internationales. Les solutions supposent une coopération à une vaste échelle et la mise en place d'une infrastructure durable.

Références :

1. BULLOCK, Alison. La conservation de l'information numérique : ses divers aspects et la situation actuelle. [En ligne] http://www.nlc-bnc.ca/9/1/p1-259-f.html

2. Consultative Committee for Space and Data System (CCSDS). Reference model for an Open Archival Information System (OAIS): blue book. In CCSDS. [en ligne] http://www.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf

3. DHERENT, Catherine. L'archivage à long terme des documents électroniques en France. In 7° conférence Microlib 2000. Lisbonne, Juin 2000. [En ligne] http://www.archivesdefrance.culture.gouv.fr/fr/notices/archi2.html

4. DHERENT, Catherine. Les archives électroniques : Manuel pratique. Site Web de la Direction des Archives de France.. [En ligne]. http://www.archivesdefrance.culture.gouv.fr/fr/archivistique.index.h

5. JACQUESSON, Alain et RIVIER, Alexis. Bibliothèques et documents numériques : Concepts, composantes techniques et enjeux. Paris : Editions du Cercle de la Librairie. 1999, 377p. Coll. Bibliothèques.

6. LUPOVICI, Catherine. Les stratégies de gestion et de conservation des documents électroniques. Bulletin des Bibliothèques de France. 2000, T.45,n°4, p.43-54. [En ligne] http://bbf.enssib.fr/bbf/html/2000_45_4/2000-4-p43-lupovici.xml.asp

7. LUPOVICI, Catherine. Les besoins et les données techniques de préservation. In 67th. IFLA council and general conference. August 16th-25th. 2001. [En ligne] http://www.ifla.org/IV/ifla67/papers/163-168f.pdf

8. LUPOVICI, Catherine. Les principes techniques et organisationnels de la préservation des documents numériques. Actes du 31° congrès de l'ADBU à l'Université de Provence, le 14/09/01. In Site de l'ADBU. [En ligne] http://www-sv.cict.fr/adbu/actes_et_je/je2001/cathLUPO_140901.html

9. LUPOVICI, Christian. La chaîne de traitement des documents numériques. Bulletin des Bibliothèques de France. 2002, t.47, n°1, p.86-91. [En ligne] http://bbf.enssib.fr/bbf/html/2002_47_1/2002-1-p86-lupovici.xml.asp

10. LUPOVICI, Catherine et MASANES, Julien. Metadata for long-term preservation. In NEDLIB. [En ligne] http://www.kb.nl/coop/nedlib/results/D4.2/D4.2.htm

11. Marcoux, Y.- Les formats des documents électroniques en archivistique : La solution au problème des archives électroniques passe-t-elle obligatoirement par les formats normalisés de documents structurés ?.- in : Archives, vol.26, n° (1-2).- pp. 85-100

12. MARTIN, Julia et COLEMAN, David. The archive as an ecosystem. In Michigan University. [En ligne] http://www.press.umich.edu.jep/07-03/martin.html

13. OCLC (Online Computer Library Center). Web characterization. In OCLC's Website. http://wcp.oclc.org