Les stratégies de gestion et de conservation préventive des documents électroniques

« When all data are recorded as 0s and 1s, there is, essentially, no object that exists outside of the act of retrieval. The demand for access creates the "object", that is, the act of retrieval precipitates the temporary reassembling of 0s and 1s into a meaningful sequence that can be decoded by software and hardware ».

Abby Smith, *Preservation in the Future Tense*

La conservation préventive des documents a pour objectif de prendre les mesures nécessaires pour éviter leur vieillissement et leur altération du fait de leur utilisation ou simplement du fait du vieillissement naturel.

Catherine Lupovici

Bibliothèque nationale de France
catherine.lupovici@bnf.fr

En cas de dommage constaté, des actions de conservation curative peuvent également être menées pour remettre en état un document qui a généralement subi un accident.

Ces actions de préservation vont aller de l'application de normes sur la qualité de l'environnement jusqu'à des consignes d'utilisation qui prescriront, dans certains cas, la communication d'une reproduction à la place de l'original, en passant par des reconditionnements comme la reliure. La qualité de l'environnement est mesurée selon des critères de température, d'hygrométrie, de poussière, de luminosité et même de champ magnétique pour les documents dont le contenu est enregistré selon des procédés magnétiques.

La bonne connaissance des conditions de vieillissement des supports et des encres pour les imprimés par exemple nous permet de conduire des opérations de prévention, de restauration et même de donner des recommandations pour les producteurs des composants et les façonniers qui réalisent les documents imprimés.

De la même façon, les enregistrements sonores, les photographies, les films ont leurs propres normes facilitant la conservation des supports et des recommandations sur la composition chimique de toutes les composantes des supports et des couches sensibles.

Le document numérique apporte bien évidemment une nouvelle génération de problèmes, qui sont à la fois une extension des choses connues à ce nouveau type de document et des problèmes totalement nouveaux qui vont avoir un impact sur les traitements de conservation à leur appli-

quer. Le document numérique en est à sa genèse et il subit encore des évolutions importantes : le recul manque donc, qui donnerait une vision suffisamment générique pour définir des traitements pouvant s'appliquer à tous les types. L'évolution la plus récente avec la diffusion grand public du *e-book* met particulièrement en lumière un de ses aspects distinctifs des autres documents : il est immatériel et peut être transféré sur des supports divers y compris être momentanément stocké sur un support de communication tout en gardant son identité et son intégrité intellectuelle.

La problématique de conservation de l'information numérique

Une ressource électronique est un ensemble complexe de contenus d'informations et de paquetage de ces informations dans un format de document accompagné d'un programme informatique ou application offrant des fonctionnalités de manipulation telles que la recherche, l'affichage, la navigation. Le tout est stocké sur un support. La bonne conservation ou la restauration des documents nécessite de connaître toutes les composantes techniques pour continuer à garantir l'accès au contenu malgré l'obsolescence de la technologie utilisée lors de la publication, ou de la création de la ressource électronique par la bibliothèque. Elle nécessite également, comme pour les documents classiques, d'assurer la conservation des supports sur lesquels l'information électronique est stockée de manière à pouvoir extraire le flux intégral des données enregistrées.

L'approche de la conservation de l'information numérique a évolué, s'attachant d'abord à la préservation du support, puis proposant la migration à la fois du support et du contenu. Enfin, une autre approche est actuellement en phase exploratoire, qui consiste à combiner le rafraîchissement du support et l'émulation de l'environnement technique informatique d'origine du document. Tous les programmes de recherche en cours pour développer des modèles d'archivage à long terme des ressources électroniques explorent les deux possibilités techniques les plus récentes. L'émulation s'appuyant davantage sur une durée de vie plus longue du support physique ou sur un rafraîchissement conduit en parallèle.

La bonne conservation ou la restauration des documents nécessite de connaître toutes les composantes techniques pour continuer à garantir l'accès au contenu malgré l'obsolescence de la technologie utilisée lors de la publication ou de la création de la ressource électronique

Bien entendu d'autres techniques sont proposées telle que l'impression systématique des fichiers textuels et la conservation du papier seul, ou la conservation de microreproductions, les fichiers numériques étant réservés à la communication de l'information. Cette approche, qui peut se justifier pour certains documents textuels ou images, ne peut plus s'appliquer aux documents multimédias qui associent également des objets ayant un déroulement temporel tels que le son ou la vidéo et des objets en trois dimensions. De plus, les travaux en cours dans les archives visant à évaluer la durée de vie des impressions bureautiques obtenues sur des imprimantes laser ne sont pas très encourageants.

Le vieillissement du support

La conservation de l'information numérique a été examinée dès la décennie 80 par les organismes responsables de la préservation des informations produites par les institutions sous forme de traitements de textes tels que les archives. L'attention s'est d'abord portée sur la fragilité du support qui était alors essentiellement magnétique. Les archivistes ont mis en œuvre, au début des années 90, un rafraîchissement du support de l'information en recopiant périodiquement les données d'un support sur un autre. Cette technique reste efficace tant que l'information est encodée dans un format indépendant de la plate-forme matérielle et logicielle qui a servi à la produire et à l'utiliser, et tant que le logiciel qui sert à interpréter le format d'encodage est maintenu ou qu'il est remplacé par une nouvelle version qui assure la compatibilité ascendante avec au moins la version précédente.

Les supports de publication et d'archivage des documents électroniques ont d'abord été des supports magnétiques. Actuellement, ce sont essentiellement des disques compacts, en attendant de devenir des DVD (Digital Versatile Disk). Ces derniers permettent d'engranger une plus grande quantité de données sur un disque de même taille, ce qui ne peut qu'entraîner une diminution de la durée de vie, en raison de la seule augmentation de densité des données. À chaque apparition d'un nouveau support de publication et d'archivage, des études techniques contradictoires aux annonces des constructeurs sont conduites, afin de pouvoir prévoir leur durée de vie et donc de gérer correctement la préservation des informations qu'ils véhiculent.

Le CD est aujourd'hui très répandu comme CD pressé pour les enregistrements sonores et les applications multimédias, mais aussi très utilisé en tant que CD enregistrable comme support d'archivage des numérisations de documents patrimoniaux effectuées par les archives, les bibliothèques et les musées. Les constructeurs annoncent une longévité de 75 à 200 ans selon la nature des composants du disque, et en particulier de la couche sensible pour les CD enregistrables, et une fois le disque gravé. Les tests de vieillissement accélérés montrent que les CD enregistrables gravés sont plus fragiles que les CD pressés. Il n'y a malheureusement pas de norme établie pour mesurer la durée de vie probable d'un CD. Seuls des travaux de recherche effectués dans certaines bibliothèques patrimoniales comme la Library of Congress ou la Bibliothèque nationale de France permettent d'envisager la mise en œuvre de tests systématiques permettant de prédire la durée de vie de chaque CD, de manière à assurer son rafraîchissement avant que le disque ne soit plus exploitable. Les études de vieillissement « naturel » des premiers disques conservés dans de bonnes conditions montrent un vieillissement des vernis protecteurs de la couche métallique qui peut entraîner son oxydation. Aussi les durées de vies moyennes annoncées par des laboratoires de tests tels que le National Media Lab pour des CD enregistrables sont plutôt de l'ordre de 5 ans, et pour le CD pressé les chiffres communément avancés vont de 10 à 25 ans.

Le problème consiste donc à gérer le rafraîchissement des supports en disposant d'information de gestion de l'espérance de vie de tous les supports de tous les documents. Ceci, de façon à lancer les opérations de report éventuellement couplées à d'autres actions concernant l'obsolescence des technologies informatiques associées à chaque ressource électronique, dont on a décidé la conservation sur une période donnée, qui peut être le long terme se poursuivant indéfiniment. Il est cependant d'ores et déjà clair que la quantité de supports de données numériques à entretenir va augmenter de façon très importante et qu'une durée de vie des supports qui se rapprocherait de celle des microreproductions noir et blanc, qui est de l'ordre d'une centaine d'années, serait nécessaire pour rendre la conservation des informations numériques plus supportable économiquement.

L'obsolescence des technologies informatiques

Dans le domaine de l'information numérique, il est vite apparu que non seulement le support avait une durée de vie très limitée, même dans de bonnes conditions de conservation, mais que, de plus, les périphériques, les programmes informatiques, les méthodes de traitement de l'information disparaissaient au profit de nouvelles techniques selon un cycle de validité d'une durée de 2 à 5 ans. Avec une telle rapidité de l'évolution des techniques de stockage, de traitement et de recherche de l'information, un investissement minimum devait être consacré à la durée de vie du support au profit d'une étude plus poussée sur les moyens de pallier l'obsolescence des technologies informatiques.

Les évolutions technologiques que nous avons connues dans la dernière décennie montrent à l'évidence que les données numériques sont produites de manière très diversement dépendante des matériels et logiciels et que les compatibilités ascendantes des versions logicielles ne sont pas toujours garanties de façon satisfaisante.

Deux techniques sont ainsi proposées en réponse à l'obsolescence des technologies : la migration des données et l'émulation des environnements informatiques.

La migration est le transfert périodique d'une ressource électronique d'un environnement matériel/logiciel à un autre, ou d'une génération de technologie informatique à une autre. La migration doit préserver l'intégrité de l'objet numérique, tout en permettant à l'utilisateur de continuer à l'utiliser, c'est-à-dire de faire des recherches d'information et d'afficher les données de manière identique. Cette migration inclut implicitement un rafraîchissement du support, mais elle ne s'appuie pas sur une copie exacte des objets d'information qui doivent restituer les contenus informationnels dans le nouvel environnement technique.

La migration doit préserver l'intégrité de l'objet numérique, tout en permettant à l'utilisateur de continuer à l'utiliser, c'est-à-dire de faire des recherches d'information et d'afficher les données de manière identique

La migration opère en modifiant le format des objets numériques tout en ambitionnant de préserver leur contenu intellectuel. Elle est effectuée à chaque changement d'environnement informatique. Elle a donc des implications légales sur le droit à modifier des objets numériques, qui sont de plus en plus des objets complexes composés d'éléments gérés par des régimes de droits différents. Au fur et à mesure que la quantité d'information numé-

rique à traiter s'accroît, il semble nécessaire d'organiser les opérations de migration au niveau national et international sur une infrastructure coopérative forte.

L'émulation consiste à émuler un environnement informatique différent de la plate-forme sur laquelle se fait l'émulation. Par exemple, on peut imaginer d'émuler un MacIntosh d'une certaine version sur un PC avec Windows NT, de manière à pouvoir utiliser un cédérom initialement publié pour Mac Intosh seulement. La technique de l'émulation appliquée à la préservation à long terme des ressources électroniques signifie que l'on cherche à définir une méthode qui va permettre d'émuler des systèmes informatiques obsolètes sur les systèmes encore inconnus du futur. L'objectif étant d'arriver à utiliser les logiciels obsolètes associés au document électronique.

Une telle approche est encore en phase de recherche. Elle semble cependant actuellement la plus prometteuse, dès lors que l'on évoque les coûts qui seront engendrés par la migration permanente d'une quantité toujours grandissante de données numériques, chaque fois que les formats, les logiciels et les matériels changeront. Au-delà de la période actuelle où nous manquons de recul et où nous prenons des mesures de préservation ponctuelles en réaction aux évolutions des technologies, il apparaît indispensable aux tenants de l'émulation de commencer à explorer une solution viable sur le long terme. La mise en œuvre de l'émulation recouvre :

- le développement de techniques généralisables pour la spécification d'émulateurs qui tourneront sur des ordinateurs futurs et qui permettront d'enregistrer tous les attributs nécessaires à la recréation du comportement des documents actuels et futurs ;
- le développement de techniques pour enregistrer les métadonnées nécessaires pour chercher, accéder aux documents numériques et les recréer ;
- le développement de techniques pour encapsuler les documents, leurs métadonnées, les logiciels et les spécifications des émulateurs de façon à prévenir leur altération.

Les caractéristiques de l'objet d'information numérique

Les objets d'information que nous avons à traiter, stocker et communiquer sont de plus en plus des objets multimédias avec des liens hypertexte vers d'autres objets. L'objet d'information numérique étant techniquement complexe, il est nécessaire de définir ce qui est à préserver. Pour cela il est important de comprendre ce qui donne une identité distincte et ce qui constitue l'intégrité intellectuelle d'un objet d'information numérique.

L'objet d'information numérique étant complexe, il est nécessaire de définir ce qui est à préserver, et de comprendre ce qui donne une identité distincte et ce qui constitue l'intégrité intellectuelle d'un objet d'information numérique

On distingue les caractéristiques suivantes à prendre en compte pour garantir l'intégrité d'un objet d'information comme un tout singulier : l'information contenue, la fixité de l'objet, la référence à l'objet, la provenance et le contexte de l'objet. Toutes ces informations qui conditionnent l'intégrité de l'objet sont créées et doivent être préservées en complément de l'objet lui-même tout au long du cycle de vie de l'objet. Ceci va de sa mise à disposition par son créateur jusqu'à son archivage dans un système ayant mission de mémoire électronique collective, en passant par sa publication chez un éditeur et sa distribution *via* des intermédiaires tels que les bibliothèques ou les centres de documentation.

L'information contenue

L'archivage à long terme des objets d'information numérique vise essentiellement à préserver l'essence intellectuelle de l'information qu'ils contiennent. Si, en première approche, tout le monde accepte cette idée comme une évidence, les choses sont plus complexes lorsque l'on s'attache à définir plus précisément ce qu'est le contenu intellectuel d'un objet d'information numérique. Le concept complexe de contenu peut s'analyser à plusieurs niveaux d'abstraction. L'archive numérique doit donc choisir le niveau d'abstraction auquel elle définit le contenu à préserver.

Au niveau le plus bas, toute information numérique est constituée d'une succession de bits ayant la valeur 0 ou 1 et chaque objet se distingue d'un autre par l'ensemble exact de la succession de bits qui le constitue. L'action de préservation à ce niveau technique de définition du contenu consiste à conserver cette exacte succession de bits. Il suffira donc de mettre en œuvre un outil de vérification d'équivalence de flux de bits tel qu'une « somme de contrôle » par exemple.

Cependant, une telle définition de contenu est assez limitée et ne rend pas compte du format et de la structure de l'information, pas plus que de

sa dépendance éventuelle de certaines plates-formes matérielles et logicielles qui évoluent particulièrement rapidement actuellement.

Si l'on considère *l'exemple du texte*, il est actuellement couvert par un standard international ASCII pour les caractères latins. UNICODE, le nouveau standard international, est en cours de déploiement et permet de prendre en compte un grand nombre d'autres systèmes d'écriture. Par ailleurs, IBM maintient son propre schéma de codage de caractère EBCDIC et Apple, où les ordinateurs personnels construits sur Intel ont leur propre méthode de mise en œuvre des caractères de l'ASCII étendu.

Il faut donc, dans une perspective d'archive, savoir effectuer la correspondance entre tous ces codes si l'on souhaite préserver l'intégrité de l'information. Cela sera particulièrement complexe si on manipule de l'information multilingue ou multi-écriture. Le texte lui-même pourra être encapsulé dans une structure logique et une présentation de mise en pages qui peuvent être codées selon des langages indépendants ou dépendants des environnements informatiques. Les langages TEX ou XML/SGML, qui sont indépendants du logiciel et du matériel qui sert à les produire et à les consulter, ne sont utilisés que par certains segments de marché comme les éditeurs scientifiques, techniques et médicaux et les chercheurs de ces domaines.

De plus, il arrive souvent que les documents qui sont produits selon des formats standards soient publiés et distribués dans des formats propriétaires liés à l'application de consultation qui les accompagne. HTML, qui est un sous-ensemble très simplifié de SGML, est utilisé comme base du codage du Web et garantit une certaine pérennité de ce qui est ainsi codé sur le Web. Cependant la grande masse du traitement de texte et de l'édition électronique est encore dominée par des codages de structure et de mise en forme propriétaires, dont la conversion dans des codages normalisés peut entraîner la perte de données servant au rendu correct du contenu telles que des éléments de mise en page.

Pour qu'un objet d'information numérique soit non seulement préservé, mais aussi accessible, il faut pouvoir y faire référence de manière univoque

L'exemple de l'image illustre d'autres aspects de définition de l'intégrité de l'information contenue dans l'objet numérique. La résolution et le rendu adéquat des couleurs qui poussent à une résolution et à un nombre de couleurs le plus élevé possible sont opposés aux impératifs économiques de limitation de la taille des objets pour leur archivage et leur transfert sur les réseaux. Ceci peut conduire à l'utilisation d'algorithmes de compression avec perte des informations comme dans le cas de l'utilisation de la norme JPEG avec perte. Là encore l'objet publié et diffusé peut être techniquement différent de l'objet créé.

Au niveau d'abstraction le plus élevé, on considère l'information contenue au-delà des limitations strictes du logiciel et du matériel employés initialement, et on essaye de définir la migration des données ou l'émulation de l'environnement comme devant permettre de retrouver le contenu intellectuel qui s'exprimera, en terme d'information brute, mais aussi, selon les options prises par l'archive, en termes de présentation, de recherche et de navigation.

Plus on considère un niveau d'abstraction élevé en terme de contenu de l'objet, plus on s'autorise à modifier des éléments techniques de codages des données à des fins de préservation et d'accès à long terme. De telles modifications peuvent être considérées comme des modifications de contenu intellectuel et ne peuvent se faire que par une autorisation du détenteur de la propriété intellectuelle *via* un transfert de responsabilité de conservation.

La fixité

L'action de préservation d'objets d'information numérique doit non seulement considérer le contenu, mais aussi la version stable et fixée de l'objet, qui ne doit pas être modifié sans accompagnement d'une numérotation ou identification rigoureuse des versions qui se succèdent dans le temps. On peut également marquer les versions canoniques des objets par un filigrane électronique. Les méthodes de marquage électronique des documents sont encore en plein développement et sont très peu normalisées. Il est donc difficile actuellement d'appliquer un codage informatique qui sera lui-même facile à préserver avec le contenu de l'information.

Dans le cas des objets d'information qui sont des bases de données en perpétuelle évolution et dont la nature même est le changement, la fixité se traduira plutôt en archivage des changements intervenus sur les notices individuelles de la base de données. Si un suivi de l'évolution des enregistrements ne peut être mis en œuvre, l'alternative est de fixer périodiquement des états de la base de données, en effectuant une prise de vue instantanée de la totalité de la base.

La référence

Pour qu'un objet d'information numérique soit non seulement préservé, mais aussi accessible, il faut pouvoir y faire référence de manière

univoque. Les systèmes de citation et de description traditionnels tels que les bibliographies, les catalogues, les index, les dictionnaires de données, les instruments de recherche, offrent les éléments nécessaires à l'identification et à la recherche d'information.

Dans l'espace du numérique, il faut également que la référence soit un lien actif qui permette, par activation du lien, d'avoir directement accès à la ressource, ou au minimum à des explications sur les conditions d'accès. Il est indispensable que ce lien reste persistant quelle que soit la localisation de l'objet sur le réseau. Plusieurs méthodes sont mises en œuvre de manière complémentaire afin de faciliter l'accès aux objets d'information numériques :

- Certains des systèmes traditionnels de description se sont déjà étendus au nouvel environnement de navigation d'un objet d'information à un autre en ajoutant le champ 856 au format MARC de notice bibliographique, afin d'inclure le lien si possible normalisé à l'objet numérique décrit dans la référence bibliographique.
- Un moyen d'obtenir des références cohérentes à un objet électronique d'information est d'inclure les éléments de la référence à la source dans une partie distincte de l'objet numérique même. C'est ce principe qu'applique le format TEI (Text Encoding Initiative), qui contient les informations descriptives et d'identification dans l'en-tête du document. C'est aussi ce principe qui est appliqué dans les métadonnées créées avec la syntaxe « méta » du format HTML. Malheureusement peu de documents codés selon ces techniques contiennent des informations descriptives de qualité suffisante.
- Les URLs (Uniform Resource Locators) encodés en HTML dans les pages Web sont des codes de localisation de l'objet numérique sur un ordinateur du réseau Internet, exécutable par simple click sur le lien pour afficher directement l'objet à partir d'une référence. L'URL n'est pas persistant et évolue avec la localisation physique de l'objet sur le réseau.

> L'objet numérique est dépendant pour sa création et pour son utilisation d'un environnement informatique comportant une dimension matérielle et une dimension logicielle

Ces solutions sont une réponse associée à un format et à un contexte particulier. L'IETF (Internet Engineering Task Force), qui effectue la normalisation d'Internet, a créé un cadre générique pour l'identification, la numérotation et la localisation des objets d'information sur le Web : l'URI (Uniform Resource Identification). Ce cadre générique englobe l'URL (Uniform Resource Locator), qui est un lien vers la localisation physique d'un objet, et qui change avec le changement de cette localisation, l'URN (Uniform Resource Name) qui est un identifiant unique et persistant de la ressource et qui est enregistré dans un répertoire permettant de résoudre ce nom dans une ou plusieurs localisations physiques du même objet, et l'URC (Uniform Resource Characteristic), qui permet d'associer des informations sur l'identité et sur les conditions d'accès à un objet.

La provenance

C'est un des concepts centraux de l'archivage moderne : on enregistre l'origine de l'objet ainsi que tout l'historique de sa conservation et des environnements techniques qui ont accompagné sa vie. Dans le contexte du numérique, la provenance est connectée au problème de la fixité, et il devient important de tracer les versions et éditions multiples d'une publication. Il devient également important de garder la trace des migrations que l'objet a subies pour préserver son utilisation dans un changement de contexte technique. Enfin, si l'information a été produite au travers d'une instrumentation ou d'un équipement, il devient important de conserver l'information relative à cette instrumentation et aux conditions d'enregistrement des données. Pour les données issues de l'activité d'une organisation, il est nécessaire que l'organisation elle-même enregistre l'historique de provenance et de migration des données et de toutes les transformations qu'elles peuvent avoir connues, et que ces informations accompagnent l'objet lors d'un transfert de compétence d'archivage.

Le fait de conserver l'historique de la chaîne de conservation de la ressource depuis sa création sert la préservation de deux manières. Il permet :

- de s'assurer de l'authenticité de l'objet ;
- de connaître le contexte qui est un autre point important de la garantie de l'intégrité.

Le contexte

L'objet d'information numérique est dépendant de son contexte technique à plusieurs niveaux :

- il est dépendant d'un contexte technique qui conditionne son utilisation ;
- il est également lié à d'autres objets par des liens encodés dans son propre contenu ;
- enfin il est lié à son vecteur de communication.

L'objet numérique est complètement dépendant pour sa création et pour son utilisation d'un environnement informatique comportant une dimension matérielle et une dimension logicielle. Selon les cas il sera plus ou moins étroitement lié

à une configuration spécifique. Par exemple, un ouvrage sur cédérom fonctionnant uniquement sur Macintosh, processeur 68030, 6MB RAM, système 7.01 ou supérieur. À l'opposé, un objet du Web encodé en HTML aura une contrainte technique beaucoup plus générique, puisqu'il sera utilisable avec un navigateur disponible sur pratiquement tous les matériels. La contrainte peut être également logicielle dans le cas d'un objet qui ne pourra être utilisé qu'avec une version d'un traitement de texte très particulier. L'archivage de l'objet doit donc être accompagné d'informations suffisantes pour pouvoir traiter cet environnement technique non seulement de manière courante lorsque la technique est disponible, mais aussi à long terme lorsqu'elle sera dépassée et que l'environnement d'origine aura disparu.

La technique de liens qui se répand de plus en plus pose un autre problème pour la conservation à long terme. Le lien écrit en HTML dans un objet du Web par exemple permet de cliquer et d'ouvrir un autre objet situé sur un autre site. La préservation de l'intégrité de l'objet voudrait que l'on préserve également tous les objets liés à un objet préservé. Étant donné la généralisation actuelle de la pratique du lien, cela conduit certains à préconiser un archivage unique périodique de l'ensemble du Web qui assurerait ainsi la préservation de tous les objets liés les uns aux autres.

Les caractéristiques des supports de communication des objets numériques diffèrent selon que l'on est hors ligne ou en ligne. Le cédérom par exemple a ses propres formats d'organisation de fichiers. Cependant, le contexte de plus en plus fréquent de distribution sur le réseau introduit d'autres caractéristiques à bien cerner pour assurer l'intégrité des objets transmis.

L'organisation de l'archivage numérique

Le volume de l'information numérique créée uniquement sous cette forme augmente de façon très importante ainsi que les volumes d'information classique numérisée.

Rôles et responsabilités

L'appel d'offre en cours à la Bibliothèque Royale des Pays-Bas pour son système informatique d'archive des documents numériques estime le volume de ses documents numériques à préserver à 12 téra octets* en 2000, 75 téra octets en 2002 et 340 téra octets en 2007. Cet archivage devra couvrir les besoins de fonction de Dépôt (le dépôt étant volontaire aux Pays-Bas), de serveur de documents en ligne pour les besoins de traitement et de consultation, d'archivage du Web, de numérisation des collections de la bibliothèque.

> Les bibliothèques nationales examinent tous les problèmes techniques et juridiques relatifs à la collecte et à la conservation des publications numériques sur réseau

La mise en place de l'organisation des responsabilités au niveau international, en particulier pour la préservation de l'information publiée sur réseau, est en cours de construction au niveau international entre les différentes institutions qui créent et celles qui ont une mission ou une sensibilité forte à la préservation de l'information pour en garantir l'accès sur le long terme. Beaucoup d'acteurs de la chaîne de création et de distribution de l'information électronique s'intéressent à la fonction d'archive dans sa dimension de gestion des droits d'accès à cette information et donc de commerce électronique qui peut y être associé. On distingue ainsi les grandes tendances d'organisation suivantes :

- *Les créateurs/fournisseurs/propriétaires des droits* ont la responsabilité initiale d'archivage des objets d'information numérique et assurent ainsi leur préservation tant qu'ils ne transfèrent pas leur responsabilité. Ils peuvent transférer cette responsabilité par des accords avec des systèmes d'archive qui prennent en charge tout ou partie de la responsabilité de l'archivage. Les bibliothèques et les archives peuvent ainsi se positionner en sous-contractant des créateurs et fournisseurs pour maintenir les archives numériques pendant la durée de vie active des objets d'information et au-delà sur le long terme. Les bibliothèques, les archives et les musées qui numérisent leurs collections sont responsables de la sauvegarde des reproductions numériques qui ont représenté un certain investissement et qui doivent toujours doubler la copie de communication.
- *Les bibliothèques nationales* examinent tous les problèmes techniques et juridiques relatifs à la collecte et à la conservation des publications numériques sur réseau. Elles projettent sur ce nouveau type d'objet leurs missions traditionnelles de collecte et de préservation de l'héritage culturel et intellectuel de leur pays pour en assurer l'accès actuel et futur.
- *Des communautés de créateurs* dont les chercheurs qui sont à l'origine de l'Internet, souhaitent organiser l'archivage sur le long terme des

* Un téra octet = un million de millions d'octets (10^{12}).

publications du réseau, indépendamment de l'organisation de l'accès à ces données. Ils militent pour un archivage au niveau international de l'Internet par une organisation coopérative indépendante, s'appuyant sur une organisation technique d'archive certifiée par rapport à un modèle en cours de définition.

De grands courants se dessinent pour l'archivage et la disponibilité à long terme des objets numériques ayant pour support le réseau, qui devront trouver leur cohérence par rapport à l'organisation déjà en place pour les objets numériques sur support. L'aspect réseau de l'objet laisse ouvertes toutes les possibilités techniques de mise en œuvre de manière répartie ou centralisée. Les aspects économiques pèseront cependant de manière décisive sur la construction des modèles viables sur la durée. Les contraintes techniques et économiques font parallèlement renaître de vieux débats sur l'assise géographique de la mémoire culturelle et scientifique de chaque pays.

Les contraintes techniques et économiques font renaître de vieux débats sur l'assise géographique de la mémoire culturelle et scientifique de chaque pays

Le Danemark, la Finlande et la Norvège ont déjà modifié leur législation sur le dépôt légal, afin de couvrir les ressources électroniques publiées sur le réseau, et beaucoup d'autres pays sont en train de préparer une révision de leur législation, très largement à l'initiative et avec une importante contribution de leur bibliothèque nationale. L'approche est fonction des caractéristiques nationales concernant la situation de l'industrie de l'édition électronique, les accords qui existent déjà pour les autres supports et la politique de collecte de la bibliothèque nationale. En Allemagne et aux Pays-Bas par exemple, il existe une industrie forte d'édition électronique internationale assurée par les éditeurs traditionnels du domaine, ce qui influence les réflexions en cours vers le dépôt volontaire des documents de ces éditeurs. La Bibliothèque Royale des Pays-Bas va même jusqu'à proposer qu'elle puisse être l'archive officielle des éditeurs tels qu'Elsevier ou Kluwer. À l'opposé, en Australie, en Norvège, au Danemark ou en Suède, l'activité d'édition électronique sur réseau est faite par de nouveaux éditeurs complètement en dehors de l'activité d'édition traditionnelle. Ceci affecte les modalités de collecte et de transfert des ressources électroniques dans l'archive numérique nationale : la Suède et la Finlande pratiquent la collecte systématique des ressources en ligne du domaine national du Web à l'aide de robots. L'Australie sélectionne les ressources du Web correspondant à la politique documentaire définie et publiée, et moissonne (*harvest*) ces ressources sur le réseau. La Norvège demande le dépôt de toutes les publications par les éditeurs qui assurent la mise en forme. Enfin, au Danemark, les éditeurs notifient la bibliothèque nationale de la publication sur le réseau, et la bibliothèque va chercher le document à la source dans la base de l'éditeur.

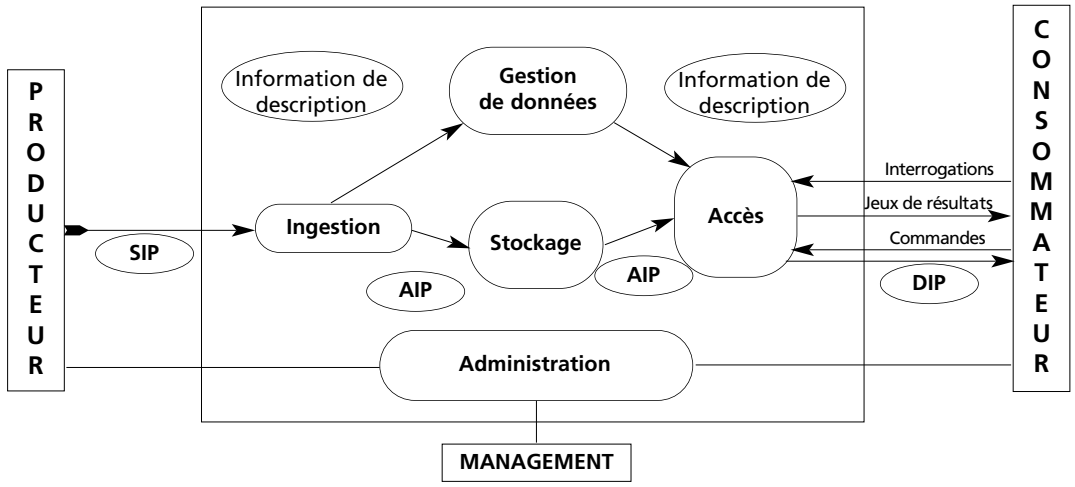
Une différence apparaît également entre les publications gratuites et les publications dont l'accès est payant. En Norvège par exemple, la bibliothèque nationale reçoit les publications payantes et les communique sur site avec les mêmes restrictions que le dépôt légal traditionnel et a, par ailleurs, le droit de collecter directement sur le réseau les publications gratuites et d'en offrir un accès gratuit sur le réseau. Les nouvelles législations tentent donc de définir de manière plus large la notion de publication, d'acte de publication et de lieu de publication. Les nouvelles législations autorisent également les bibliothèques nationales à effectuer des reformatages des documents pour leur préservation. Enfin, les réflexions en cours examinent l'inscription dans la loi d'autoriser la sélection des documents électroniques en réseau généralement sur des bases proches des politiques de développement de collection des autres formats de publications.

En France, la loi du Dépôt légal de 1992 et ses décrets d'application de 1993 règlent l'obligation du dépôt pour les publications électroniques sur support. Le texte ne s'applique pas aux publications en ligne qui ne sont pas actuellement enregistrables par la Bibliothèque nationale de France au titre du dépôt légal. Une extension de la loi est nécessaire pour couvrir ce dernier type de publications.

Aux États-Unis des discussions ont lieu pour savoir s'il faut une extension du *Copyright Act* de 1976 qui définit les droits et obligations des bibliothèques et des archives pour assurer la préservation des documents imprimés de manière obligatoire si l'institution qui assure naturellement cette fonction ne remplit plus ses obligations. D'autres personnes avancent la nécessité de créer un nombre restreint de l'ordre de trois centres d'archives indépendants pour effectuer la sauvegarde des publications sur Internet, sachant que d'autres institutions pourraient se partager la responsabilité d'offrir l'accès aux publications stockées dans les centres d'archivage.

Dans tous les cas, le concept d'archive assurant la préservation à long terme des objets d'information numérique est présent à l'esprit. Il s'agit d'un système qui permet l'enregistre-

Entités fonctionnelles de l'archive OAIS



ment des objets, leur gestion documentaire et technique et qui sait les communiquer à la demande. Une norme générique définissant les fonctions et les métadonnées de gestion des objets d'information est en cours d'élaboration : c'est le Modèle de référence d'un Système d'Archive Ouvert (OAIS).

Le modèle de référence d'un Système d'Archivage Ouvert

Le Comité consultatif pour les systèmes de données spatiales (CCSDS) a élaboré une recommandation pour la normalisation d'un modèle de référence d'un système d'archivage ouvert (OAIS) qui a été introduit comme projet de norme internationale ISO dans sa version finale de juin 1999.

Le modèle ainsi défini peut s'appliquer à n'importe quelle archive et plus spécialement aux organisations ayant la responsabilité de rendre l'information disponible à long terme. C'est un modèle générique pour tous les types de documents, qu'ils soient numériques ou traditionnels, et sur lequel on peut construire des normes d'application dans des domaines particuliers. Il n'indique pas une conception ni une implémentation qui reste libre de regrouper ou de décomposer les fonctions présentées dans le modèle. La NASA (National Space Agency) aux États-Unis, mais aussi le CNES (Centre national d'études spatiales) ont participé activement à l'élaboration de ce modèle de référence. La NARA (National Archives and Records Administration), ainsi que le National Institute of Health, s'intéressent également à ce modèle. Enfin, les grands projets d'archivage des publications électroniques des bibliothèques ont tous décidé d'adopter ce modèle générique pour définir les fonctions d'archivage des bibliothèques numériques et pour réaliser leurs systèmes.

Une archive OAIS interagit avec son environnement en :
- recevant des informations à pérenniser des producteurs de données ;
- gérant l'archive conformément aux règles de gestion définies par la réglementation générale (droits, tarifs, relations avec les producteurs et les consommateurs de données) ;
- fournissant à la demande les données au consommateur de données, qui peut être une personne ou un système client.

L'archive OAIS doit avoir un niveau suffisant de contrôle concédé par les producteurs d'information pour pouvoir assurer la pérennisation de l'information contenue en ayant autorité pour assurer les opérations de migration ou d'émulation. Les questions à résoudre sont les problèmes liés au droit d'auteur et à l'autorisation de modifier la représentation des contenus pour s'affranchir de la plate-forme technique initiale d'utilisation. L'OAIS doit mettre en place des procédures de soumission avec les producteurs et des procédures de diffusion qui permettent de réaliser la préservation et l'accès.

L'archive OAIS est composée des entités fonctionnelles figurées dans le schéma ci-dessus :

L'archive OAIS reçoit des *paquetages d'information de soumission (SIP)* qu'elle transforme en *paquetages d'information d'archive (AIP)* et restitue à la demande en *paquetages d'information de diffusion (DIP)*.

– *L'entité d'ingestion* réceptionne les SIP, contrôle leur qualité et génère les AIP conformément aux normes de l'archive et les transmet à l'entité de stockage. Elle extrait également les informations de description qui sont transmises à l'entité de gestion de données.

– *L'entité de stockage* réceptionne les AIP, gère le stockage, renouvelle les supports, fournit les moyens de sauvegarde en cas de catastrophe et transmet les AIP à l'entité d'accès pour satisfaire les commandes.

– *L'entité de gestion de données* administre les fonctions de la base de données de l'archive pour les données de description bibliographique et technique et les données administratives de l'archive.

– *L'entité d'accès* communique avec les consommateurs de données en recevant les demandes, en filtrant l'accès aux données qui ont une protection particulière, et en transmettant les réponses aux consommateurs.

– *L'entité administration* gère le fonctionnement global de l'archive.

Cette description se limite volontairement au niveau très général, mais le modèle est très détaillé. De plus, des normes complémentaires sont envisagées, dont des normes pour la livraison de sources numériques pour l'archive, des normes pour la soumission de métadonnées numériques à l'archive, des normes pour l'identification de données numériques dans l'archive, des normes pour la migration de l'information d'un support ou format vers un autre support ou format, des normes pour l'accréditation des archives. Toutes ces normes complémentaires pourront être développées dans des domaines spécifiques, mais cependant de façon coordonnée.

Le modèle couvre toutes les fonctions d'une archive depuis la collecte des données jusqu'à la fourniture des informations aux utilisateurs, qui peuvent faire des recherches dans la base bibliographique de l'archive et commander des ressources. Dans le contexte d'une bibliothèque ayant déjà un système qui permet l'acquisition, l'enregistrement des documents dans les collections, la description bibliographique et la recherche dans le catalogue, il est évident que certaines des fonctions prévues généralement pour une archive sont déjà fournies par le système de gestion de la bibliothèque. La fonction de l'archive devra donc compléter le système de gestion de la bibliothèque sur les aspects stockage des documents numériques et gestion de la préservation.

Un certain nombre de projets pilotes d'archives de ressources électroniques dans des bibliothèques sont en cours, tous basés sur le modèle OAIS

Projets et réalisations en bibliothèques

Un certain nombre de projets pilotes d'archives de ressources électroniques dans des bibliothèques sont en cours, tous basés sur le modèle OAIS. Ils ont pour objectif d'expérimenter la mise en place de systèmes d'archives, afin d'évaluer la charge de travail et de tester diverses méthodes de préservation à long terme. Tous les projets s'intéressent aux métadonnées de gestion technique des objets d'information archivés dans le but de préparer une normalisation et de bien cerner leurs modalités de création et d'utilisation. Trois de ces projets sont particulièrement marquants et échangent de façon continue leurs résultats afin de progresser plus rapidement en évitant les divergences et le travail en double. Ils sont également en contact permanent avec les initiatives nord-américaines et en particulier le programme PRESERV. Selon les cas, les projets s'intéressent à l'archivage de documents numériques d'origine et/ou de documents numériques créés par la bibliothèque dans le cadre de programmes de numérisation.

CEDARS

Le projet CEDARS (CURL Examplars in Digital ARchiveS) est un projet de trois ans du Consortium of University Research Libraries (CURL), subventionné dans le cadre du programme britannique eLib (Electronic Libraries Programme), et destiné à conduire une expérimentation d'archivage distribué pour la préservation à long terme des documents numériques. Le projet associe des bibliothèques nationales, des bibliothèques universitaires et des bibliothèques de recherche d'Angleterre, d'Écosse, du Pays de Galles et d'Irlande du Nord. Il est piloté par les universités d'Oxford, Cambridge et Leeds, qui construisent trois archives numériques test en s'appuyant sur le modèle OAIS. CEDARS explore l'aspect Management dans les relations avec les éditeurs sur le droit d'archiver et converge avec d'autres projets sur le fait qu'il semblerait utile d'avoir un accord sur l'archivage immédiat, séparé de l'accord sur l'accès si l'on veut effectivement préserver les publications numériques sur le plus long terme que les intérêts commerciaux des éditeurs. Par ailleurs, CEDARS considère les aspects de préservation en relation avec l'accès par l'expérimentation de la technique d'émulation et la création de métadonnées d'accès en relation avec cette technique.

NEDLIB

NEDLIB (Networked European Deposit Library) est un projet de recherche subventionné par le

Programme Télématique de la Commission européenne qui se déroule de 1998 à la fin de l'année 2000. Il regroupe la Bibliothèque Royale et les Archives des Pays-Bas, la Bibliothèque nationale de France, la Bibliothèque nationale de Norvège, la Bibliothèque universitaire d'Helsinki, la Bibliothèque nationale allemande, la Bibliothèque nationale du Portugal, la Bibliothèque nationale de Suisse, la Bibliothèque nationale de Florence et deux sociétés informatiques. Les éditeurs Elsevier Science BV, Kluwer Academic et Springer Verlag contribuent également au projet en fournissant des documents pour les tests.

NEDLIB a défini un modèle des spécifications fonctionnelles pour un Système de dépôt des publications électroniques qui s'appuie sur le modèle OAIS, en développant particulièrement la fonction technique de préservation à long terme des documents publiés, que ce soit sur support ou sur le réseau. Un système de démonstration est en cours d'élaboration utilisant les deux techniques de migration et d'émulation. Le système de démonstration sera interfacé avec les modules fonctionnels d'acquisition, de catalogage et d'OPAC des systèmes de gestion classiques des bibliothèques où il sera implémenté. Tout comme CEDARS, NEDLIB fait en parallèle un travail plus théorique sur les métadonnées techniques dont une archive a besoin pour piloter les fonctions de préservation des documents archivés dans le Système de dépôt des publications électroniques.

PANDORA

Le projet PANDORA (Preserving and Accessing Networked Documentary Resources of Australia), commencé en juin 1996, a pour objectif de réaliser l'archive numérique d'une sélection de publications australiennes sur Internet telles que des périodiques électroniques, des sites d'institutions, des publications gouvernementales et des publications éphémères. Le projet a déjà produit des *Recommandations* pour la sélection des documents sur Internet, un modèle économique de l'archive, un modèle des données nécessaires à la gestion de l'archive pour une préservation à long terme. La Bibliothèque nationale d'Australie, considérant que la sauvegarde des publications nationales sur Internet ne peut être réalisée par une seule institution, a également pris des contacts avec les bibliothèques des États d'Australie et avec les Archives nationales du film et du son pour mettre en place un modèle national pour l'archivage partagé des publications sur le réseau. Dans le cadre de PANDORA, la Bibliothèque nationale a développé les métadonnées qu'elle utilise pour l'archive expérimentale, mais elle a aussi effectué une réflexion plus générale sur les métadonnées techniques d'une archive numérique pratiquant à la fois la migration et l'émulation. La bibliothèque a également mis en place l'attribution d'un identifiant persistant pour les publications du réseau archivées au niveau national. Sur la base de toute cette activité, la Bibliothèque nationale d'Australie a lancé un appel d'offres pour un système intégré de gestion des collections numériques fondé sur le modèle OAIS. Malheureusement, aucune offre intégrée n'existe actuellement et la bibliothèque a finalement opté pour une approche modulaire : un système de collecte des ressources sur le réseau, qui sera développé en interne sur la base de l'expérience PANDORA, et qui devrait être terminé en octobre 2000, un système de gestion des objets d'information numérique, qui doit permettre de gérer les données techniques de l'archivage, et qui devrait être disponible en juin 2001, un système d'archivage des objets d'information numériques, qui devrait être terminé en juin 2001. Elle a déjà acquis un système d'indexation et de recherche documentaire pour les métadonnées descriptives des publications numériques qui représente une partie du module d'accès OAIS.

PRESERV

PRESERV est le programme de préservation de RLG (Research Libraries Group). La préservation a toujours été une des préoccupations centrales de RLG depuis 1974. L'activité a tout d'abord été consacrée à la préservation par le microfilmage des collections qui s'est concrétisée par des programmes de microfilmage et par la publication de manuels de base au début des années 90 sur les techniques de production de microfilms de haute qualité et de conservation des *masters* de microfilms.

Cette phase achevée, RLG s'est consacré aux techniques numériques de reproduction pour définir des modèles d'archivage des documents numériques qu'ils soient numériques d'origine ou le fruit d'une opération de numérisation. Les jalons de cette démarche sont marqués par un rapport très important publié en mai 1996, et qui relate les travaux d'une Task Force commandée par la Commission on Preservation and Access (CPA) et RLG. Ce document pose les bases de toutes les réflexions en cours sur l'archivage à long terme des documents numériques. Un groupe de travail spécialisé sur l'archivage numérique a ensuite été créé en mars 1997 pour la mise en place de différentes actions. Une étude a été faite en 1998 auprès des bibliothèques membres nord-américaines, mais aussi en Europe et en Australie sur leur perception et leurs besoins en termes de préservation des ressources numériques.

- 66 % des bibliothèques n'avaient aucune politique écrite et 18 % avaient des éléments de politique écrite portant soit sur la collecte, soit sur le stockage, soit sur la stratégie de migration des données, soit sur le rafraîchissement des supports ;
- l'importance des préoccupations était, par ordre décroissant : l'obsolescence des technologies, le manque de moyens, le manque de planification, l'état physique des collections.

Une autre Task Force sur les politiques et les pratiques pour la préservation à long terme des collections numériques a été lancée en partenariat avec la Digital Library Federation (DLF) en 1999. Elle explore les pratiques d'archivage pour trois catégories de documents : les archives numériques institutionnelles, les publications numériques et les documents numérisés localement. Enfin, en mars 2000, les deux grandes associations OCLC et RLG ont signé un accord de coopération pour explorer l'archivage des documents numériques. La première étape de cette coopération concerne la rédaction de deux documents de travail :

- les attributs des archives numériques pour la recherche : définition de l'archive de collections hétérogènes de recherche, les objectifs, les éléments, les mesures de performance, la faisabilité d'une certification ;
- les métadonnées de préservation pour le long terme : métadonnées descriptives et métadonnées techniques ;

Les travaux actuels de RLG s'appuient sur le modèle OAIS et les différentes composantes de l'archive qu'il définit.

Conclusion

Les stratégies de préservation préventives des documents numériques sont guidées principalement, au contraire des publications analogiques qui s'appuient sur les mêmes supports, par la préservation et l'accès au code informatique des contenus et non plus par la seule préservation du support. Le support reste bien évidemment indispensable, mais il suit la stratégie technique de préservation du contenu numérique.

Ces nouvelles stratégies ainsi que les volumes concernés fondent les fonctions de préservation sur des techniques d'archivage électronique

> La préservation à long terme des publications numériques a des implications sur la négociation des droits pour le traitement de sauvegarde

et impulsent une convergence forte des pratiques de sélection et de description du monde des archives et du monde des bibliothèques, qui se concrétise dans le modèle de référence OAIS et dans toutes les normes d'application qui en découleront. La sensibilisation à ce modèle, au-delà de l'investissement initial du CNES, commence tout juste en France à partir de la participation à des projets de recherche tels que le projet européen NEDLIB. Elle devrait conduire à la préparation d'archives numériques de qualité qui doivent impérativement se mettre en place dans tous les établissements qui ont des collections numériques qui doivent être conservées.

Pour les institutions qui créent des informations numériques à caractère patrimonial, il est très important de bien comprendre la complexité de la préservation des contenus dans sa relation avec les formats de données et leur indépendance par rapport aux matériels et aux logiciels qui ont servi à les produire et qui servent à en assurer la communication immédiate.

Enfin, la préservation à long terme des publications numériques a des implications sur la négociation des droits pour le traitement de sauvegarde et sur l'évolution de la législation sur le dépôt légal.

Avril 2000

BIBLIOGRAPHIE

The Cedars Project (CURL exemplars in digital archives). http://www.leeds.ac.uk/cedars (visité le 3 avril 2000)

NEDLIB (Networked European Deposit Library). http://www.konbib.nl/nedlib (visité le 3 avril 2000).

PADI (Preserving Access to Digital Information). Site Web spécialisé créé et maintenu par la National Library of Australia. http://www.nla.gov.au/padi (visité le 3 avril 2000)

PANDORA Project (Preserving and Accessing Networked Documentary Resources of Australia). http://pandora.nla.gov.au/pandora (visité le 3 avril 2000)

PRESERV – The RLG preservation programme. http://www.rlg.org/preserv (visité le 3 avril 2000) Preserving digital information : report of the task force on archiving of digital information, commissioned by The Commission on preservation and access and The Research libraries group, Inc. 1996 http://www.ifla.org/documents/libraries/net/tfadi-fr.pdf (visité le 3 avril 2000)

Reference Model for an **Open Archive Information System** (OAIS) : red book. – Consultative Committee for Space Data Systems, May 1999. Pagination multiple. http://ssdoo.gsfc.nasa.gov/nost/isoas (visité le 3 avril 2000)

Lupovici, Catherine. – « Identification des ressources sur Internet et métadonnées : diversité des standards ». – *Documentaliste-sciences de l'information*, 1999, vol. 36, n° 6, p. 321-325