67th IFLA Council and General Conference
August 16-25, 2001

Code Number: **163-168-F**
Division Number: **VI**
Professional Group: Preservation and Conservation
Joint Meeting with: Information Technology
Meeting Number: 168
Simultaneous Interpretation: -

Les besoins et les données techniques de préservation

Catherine Lupovici
Head, Digital Library Department
Network and Services Direction
Bibliothèque nationale de France
Paris, France
E-mail: catherine.lupovici@bnf.fr

Gérer et préserver des collections numériques afin d'en garantir l'accès à long terme aux chercheurs est, comme c'est toujours le cas avec l'arrivée d'une nouvelle technologie, à la fois une continuité des objectifs et de l'organisation générale des services et une rupture des techniques utilisées et donc des compétences requises pour les personnels. L'acquisition de ressources numériques natives qui n'ont plus d'équivalent analogique ainsi que la numérisation de collections classiques existantes qui devient elle-même une composante des politiques de conservation, placent les bibliothèques face à un nouveau défi de préservation. Pour garantir la possibilité d'utiliser les ressources électroniques qu'elles conservent, elles vont devoir définir et collecter des métadonnées qui permettront aux gestionnaires de la préservation de prendre les mesures techniques nécessaires à la conservation des flux de bits qui constituent les objets numériques, de manière à permettre leur restitution et leur interprétation quelle que soit l'évolution technique future de l'informatique. Cette situation est identique pour toutes les institutions qui ont une mission de mémoire pour des communautés particulières d'utilisateur, qui conservent parfois des archives à des fins de preuve, et pour lesquelles l'information est déjà créée et conservée uniquement sous forme numérique depuis déjà plus de dix ans. Toutes ces communautés ont déjà commencé à construire des ensembles de métadonnées de préservation et la plupart s'appuient sur la norme OAIS (Open Archival Information System). La communauté des bibliothèques est sur la voie d'un consensus pour un ensemble de métadonnées de préservation.

1 Les nouveaux défis de la préservation du numérique

La préservation des ressources numériques, tout comme celle des ressources classiques, consiste en la préservation d'une médiation technique entre un objet et l'information qu'il véhicule. Dans le cadre de que nous avons à préserver couramment, la préservation du seul support physique, comme par exemple un livre, constitue l'essentiel du travail. Bien entendu nous savons que parfois nous avons besoin de conserver également des informations de contexte si nous voulons rendre le contenu de

l’objet intelligible. La pierre de Rosette est le symbole parfait de la différence entre la préservation de l’objet physique et la préservation de la capacité d’interpréter le codage de l’information inscrite sur le support. Nous avons également l’exemple plus récent des documents sonores analogiques pour lesquels nous avons besoin de la médiation technique d’un appareil qui transforme une vibration physique en ondes sonores et pour lesquels les caractéristiques techniques de la transformation ont changé avec la disparition de matériels remplacés par de nouvelles générations incompatibles. Dans ce dernier cas le défi de préservation était concentré uniquement sur la préservation du support et sur celle de l’obsolescence du matériel et les actions de préservation étaient de maintenir le matériel en ordre de marche le plus longtemps possible puis de transférer les enregistrements dans les nouveaux formats compatibles avec les matériels courants.

Les ressources numériques introduisent un niveau de complexité plus grand accompagné d’une dissociation entre le support et le contenu qui touche de plein fouet la législation sur le droit d’auteur et l’organisation des bibliothèques qui est fondée sur des objets physiques et sur les types de support, comme par exemple la législation du Dépôt Légal en France ou l’organisation de la conservation dans beaucoup de bibliothèques.
La conservation du support des ressources numériques, sur lequel les bibliothèques se sont concentrées depuis le milieu des années 90 en relation avec les publications électroniques hors ligne, est aujourd’hui davantage considérée comme une fonction classique de sauvegarde informatique courante. Nous avons en effet à rafraîchir régulièrement et préventivement tous ces supports selon un programme strictement planifié.
Mais l'obsolescence technologique qui affecte les ressources numériques est plus rapide que le vieillissement du support. Par exemple les versions du traitement de texte Word® changent en moyenne tous les trois ans et ne sont directement compatibles qu’avec la version immédiatement précédente. Le véritable défi auquel nous sommes confrontés est de comprendre et de savoir gérer la complexité de l’obsolescence technologique du contenu informationnel depuis le flux de bits jusqu’à son utilisation au travers de l’application avec laquelle nous dialoguons. Nous pouvons décider d’émuler un environnement technique complexe devenu obsolète ou d’effectuer une migration technique de la ressource, partiellement ou totalement selon ce qui peut être considéré comme le contenu à préserver. Dans la pratique, nous serons certainement conduits à utiliser les deux solutions et pour cela nous devons archiver des informations techniques appropriées qui constituent les **métadonnées de préservation**.

1.1 Les composantes de la préservation du contenu numérique

Lors de l’utilisation d’une ressource numérique, le contenu est traité par un ordinateur pour rendre le flux de bits intelligible. Ce traitement est composé d’une chaîne de sous traitements successifs qui créent une série de contraintes techniques que nous devons documenter à l’aide de métadonnées afin de pouvoir gérer la préservation. Tous ces sous traitements ne seront pas nécessairement obsolètes au même moment, mais il suffit que l’un d’eux ne puisse plus être réalisé pour que l’accès au document soit compromis. Nous pouvons analyser et représenter cette chaîne selon un modèle en couche, où chacune des couches représente un sous traitement qui rend un service à la couche immédiatement supérieure et sur lequel le sous traitement suivant de la chaîne opère. Depuis le bas jusqu’au niveau supérieur nous pouvons distinguer les sous traitements suivants ainsi que la catégorie de métadonnées que nous devons créer et conserver associées à la ressource :

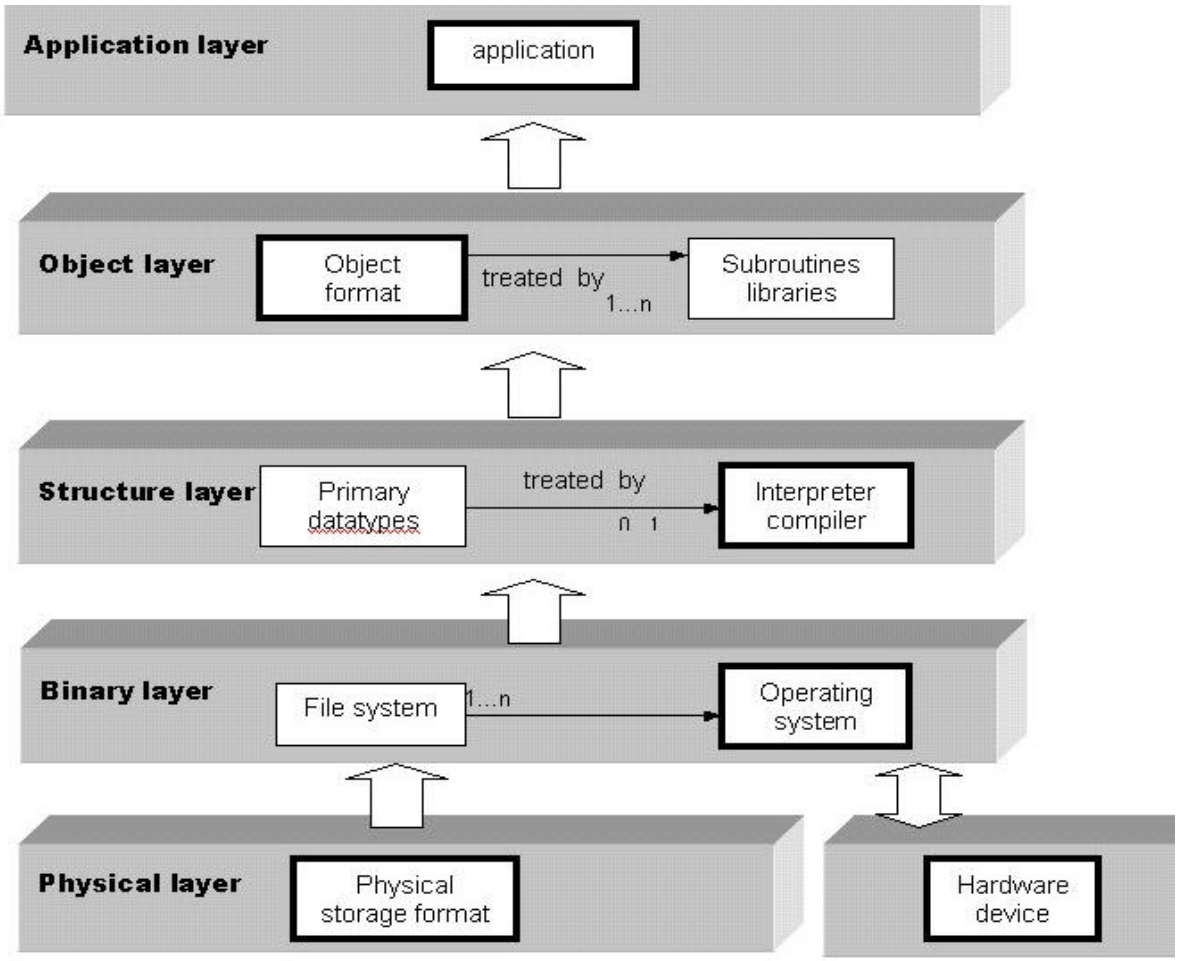

Fig.1. Le modèle d'information en couches (Extr. de NEDLIB Report 2. Fig.6, p. 8)

– **La couche physique**

La ressource numérique est stockée sur un support physique ou sur un support de communication, dans un format lié associé au support physique et qui est généralement normalisé (par exemple ISO 9660 pour un CD-ROM). Ce format sera modifié si l'on migre le document sur un autre type de support par exemple d'un CD-ROM à un DVD et l'on doit conserver toute l'information historique s'il faut pouvoir fournir le format original.

Nous devons également prendre en compte, au bas de la chaîne de traitements, l'information relative à des matériels périphériques dans le cas de programmes qui y sont attachés tels que ceux dépendants de périphériques additionnels (par exemple pour une application multimédia qui s'appuie sur des applications audio MIDI). Pour le matériel principal cette information est généralement redondante avec celle qui est associée au système d'exploitation.

– **La couche binaire**

Une fois les données extraites du support, elles sont réorganisées en blocs étiquetés de manière indépendante du format du niveau physique. Le système d'exploitation (nom et version) qui gère le système de fichiers fournit ce service et dans la plupart des cas le système d'exploitation correspond implicitement à un système de fichier mais pas nécessairement. Par exemple il suffit de dire Windows NT 4.0 pour savoir associer toutes les composantes qu'il faut recenser à ce niveau là dans la perspective de la préservation.

– **La couche structure**

Les données sont ensuite assemblées selon une structure de primitives de données qui est interprétable par des langages de programmation de haut niveau. Si la préservation concerne un logiciel non compilé ou non interprété, il sera faudra conserver l'information relative à l'interpréteur ou au compilateur nécessaire pour le réutiliser.

– **La couche objet**

Les traitements correspondant à cette couche organisent les données en objets signifiants pour la couche application et au travers d'elle pour l'utilisateur. Le format objet peut être soit un format ouvert, soit un format propriétaire. Les objets étant rendus intelligibles par l'application, seules les informations relatives au format des objets sont nécessaires dans la perspective de préservation. Les formats objet sont par exemple une image codée en JPEG, une page codée en HTML, une vidéo codée en MPEG.

– **La couche application**

Les programmes de la couche application manipulent les objets transmis par la couche précédente pour les présenter à l'utilisateur. Le nom et la version de l'application peuvent être redondants avec le format objet, et le nom et la version de l'application peuvent avoir une correspondance exacte et unique avec le format objet (par exemple le format PDF et l'application Acrobat reader sont des informations redondantes). Dans d'autres cas plusieurs applications peuvent correspondre à un même format objet (par exemple le format JPEG pour lequel plusieurs applications de visualisation peuvent être utilisées). Enfin il existe des cas où seule l'application est connue et où le format propriétaire qu'elle utilise reste caché (c'est souvent le cas dans les premières publications électroniques sur CD-ROM)

1.2 Les différents types techniques de ressources électroniques

Cette analyse nous conduit à distinguer deux types de ressources électroniques qui ne demanderont pas les mêmes types d'actions de préservation et qui ne seront pas archivées avec les mêmes métadonnées de préservation :

– Les ressources électroniques qui sont dépendantes de systèmes propriétaires spécifiques, pour lesquelles nous devons collecter des métadonnées sur l'Application (au niveau couche application) et sur le Système d'exploitation (au niveau couche binaire). Elles correspondent généralement à des publications hors-ligne sur CD-ROM pour lesquelles une application spécifique manipule des formats inconnus. Dans ce cas le seul accès au contenu est l'application propriétaire (par exemple cdu.exe pour le CD-ROM de l' *Encyclopaedia Universalis*)

– Des ressources électroniques construites sur des formats qui sont indépendants de systèmes spécifiques. Ici il nous faut collecter des métadonnées sur le Format objet (au niveau couche objet) et si nécessaire sur l'Application, au moins aussi longtemps que la bibliothèque est obligée par son contrat d'utilisation à se servir de l'application d'origine pour accéder aux données de la ressource. Ce type de ressource correspond généralement aux ressources du Web. Evidemment les documents numériques créés par la bibliothèque dans un format connu et documenté doivent être archivés indépendamment des applications courantes de bibliothèque numérique comme des contenus neutres.

Nous constatons donc que les ressources les plus difficiles à conserver sont les applications dépendantes de systèmes spécifiques, ce qui est le cas principalement des publications hors ligne avec des accès protégés et contrôlés par le bais d'applications propriétaires. Actuellement le Web est plus ouvert mais nous devons suivre l'évolution technique du Web commercial et du Web profond qui peuvent devenir aussi propriétaires que certaines publications hors ligne. La communauté des bibliothèques doit également sensibiliser les auteurs et les éditeurs pour qu'ils déposent des

applications plus ouvertes s'appuyant sur des formats standards documentés de manière à permettre la préservation à long terme de leurs publications.

2 Les métadonnées de préservation

Plusieurs ensembles de métadonnées ont été définis en relation avec les ressources numériques orientées vers des fonctions précises. Les catégories de métadonnées couramment associées aux collections numériques sont :

- Les métadonnées descriptives qui permettent la découverte et l'identification. Elles ont reçu une attention particulière de la part de la communauté des bibliothèques car elles peuvent être considérées comme proches du catalogage
- Les métadonnées administratives qui permettent la gestion d'une ressource à l'intérieur d'une collection de ressources numériques
- Les métadonnées de structure qui attachent les composants d'une ressource numérique complexe

Cependant ces métadonnées sont destinées à des systèmes de gestion de collections de bibliothèques numériques afin que les utilisateurs puissent rechercher et feuilleter à l'intérieur des collections. Elles ne sont pas faites pour la préservation qui correspond davantage à des fonctions de système d'archivage.

Des premières tentatives pour définir des métadonnées de préservation et d'archivage ont été faites à l'occasion de plusieurs projets de bibliothèques. Par exemple RLG (The Research Libraries Group) a publié en mai 1998 une recommandation d'un ensemble de 16 métadonnées essentielles à créer en même temps que les reproductions numériques réalisées dans les projets de numérisation. Ce travail initial a conduit à des travaux préparatoires pour une norme NISO sur les Métadonnées techniques pour les images fixes numériques. Les différentes approches reflètent des besoins spécifiques dans des domaines particuliers et il est nécessaire de les intégrer dans un cadre général qui soit axé sur les fonctions d'archivage et qui couvre toute la gamme des besoins de préservation que l'on peut rencontrer dans les bibliothèques.

Le modèle de référence OAIS (Open Archive Information System), développé sous les auspices du Consultative Committee for Space Data Systems (CCSDS) de la NASA, avant projet de norme ISO DIS/14721, fournit un tel cadre. C'est un modèle de référence de haut niveau qui peut être utilisé pour le cadre de métadonnées d'archivage dont les métadonnées de préservation. Plusieurs projets de bibliothèque s'appuient déjà sur le modèle OAIS pour développer des ensembles de métadonnées de préservation.

2.1 Le modèle d'information de l'OAIS

Au niveau abstrait le modèle d'information de l'OAIS[1] manipule des *Objets d'information*. Un Objet d'Information est l'information dérivée d'un *Objet de Données* (physique ou numérique) traduite en information qui a du sens grâce à la Base de Connaissances de l'utilisateur et surtout du point de vue de la préservation grâce à l'*Information de Représentation*.
Quatre classes d'Objets d'Information constituent les *Paquetages d'Information d'Archive* que le système reçoit (SIP), stocke (AIP) et fournit (DIP) à la demande d'un utilisateur. Les types d'Objet d'Information sont :

- L'*Information Contenue* qui est l'information majeure qui doit être préservée. La définition de l'Information Contenue n'est pas évidente. Par exemple pour un périodique en ligne, on peut décider qu'il ne s'agit que du texte de chaque article et de ses illustrations, ou bien qu'il s'agit non

[1] La traduction française de la terminologie OAIS utilisée dans cette présentation est celle réalisée pour l'avant projet de norme par le Centre National d'Etudes Spatiales (France) en 1999. Dans la mesure où elle n'est pas encore validée au niveau international, la terminologie anglaise qui est quant à elle tout à fait stable est conservée dans le schéma comme référence des concepts qui sont utilisés.

seulement du contenu mais aussi de la présentation des articles par exemple dans des fichiers PDF. Mais pour le même périodique, on peut décider que le contenu est également toute l'application qui permet de faire de la recherche dans la base de données du périodique ou de feuilleter toute la collection.

- L'*Information complémentaire à la Pérennisation* (PDI), qui contient l'information nécessaire à la gestion de la préservation de l'Information Contenue
- L'*Information d'Empaquetage*, qui relie ou regroupe l'Objet Numérique et les métadonnées relatives à son paquetage en une entité identifiable sur un support spécifique
- L'*Information de Description*, qui facilite l'accès à l'Information Contenue au travers des outils de recherche de l'archive.

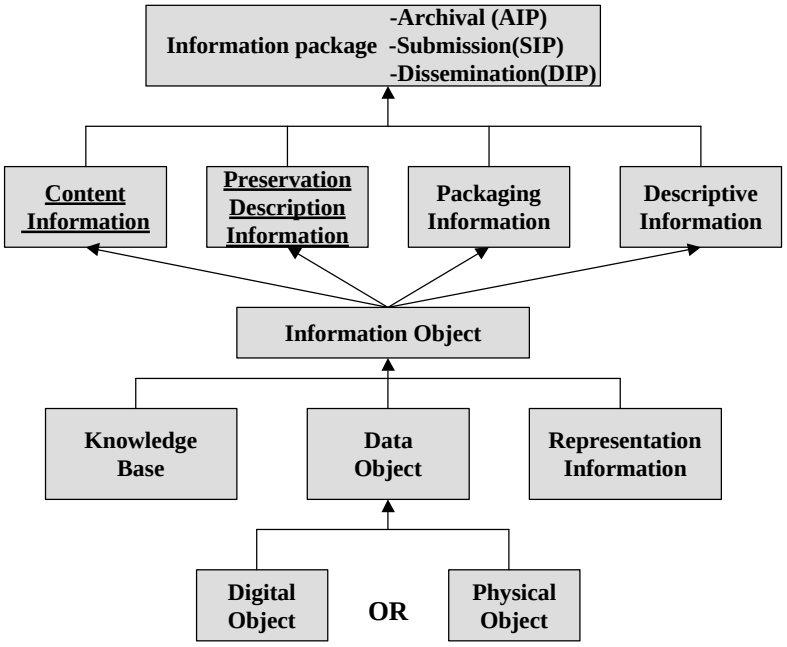

Fig.2. Le modèle d'information OAIS (Extr. de OCLC:RLG White paper. FiG2, p. 12)

L'Information Contenue et l'Information complémentaire à la Pérennisation sont les deux classes d'Objets d'Information essentielles pour la préservation à long terme, et elles fournissent le cadre pour la création des métadonnées de préservation.

L'Information Contenue d'un objet numérique doit être rendue en utilisant l'*Information de Représentation* et nous devons donc créer des métadonnées de cette catégorie, par exemple des métadonnées sur le nom et la version du Système d'exploitation.

Le modèle OAIS identifie quatre types d'Information complémentaire à la Pérennisation pour lesquelles des métadonnées doivent également être définies :

- L'*Information de Référence* qui énumère et décrit les identifiants de l'Information Contenue
- L'*Information de Provenance* qui documente l'historique de l'Information Contenue
- L'*Information contextuelle* qui documente les relations entre l'Information Contenue et son environnement

– L'*Information de Fixité* qui documente les mécanismes d'authentification utilisés pour vérifier que l'Information Contenue n'a pas été altérée de manière non documentée

2.2 Etat de l'art du développement des métadonnées de préservation

OCLC et RLG ont annoncé, en mars 2000, leur engagement pour collaborer à l'identification et au soutien de bonnes pratiques destinées à favoriser l'accès à long terme aux objets numériques. L'un des domaines de cette collaboration est l'utilisation de métadonnées pour faciliter les traitements de préservation et ils ont mis sur pied un Groupe de travail sur les métadonnées de préservation. La première tâche du groupe a été de publier un White paper présentant et comparant quatre ensembles de métadonnées afin de préparer leur rapprochement pour la construction d'un consensus international. Trois de ces ensembles de métadonnées peuvent être mis en correspondance avec le modèle OAIS. Il s'agit de :

– L'ensemble des métadonnées publiées en 1999 par la National Library of Australia, qui ont été développées dans le cadre du projet PANDORA (Preserving and Accessing Networked Documentary Resources of Australia). Ces métadonnées ont été créées à la fois pour les objets numériques natifs et pour les reproductions numériques de substitution.

– Le projet CEDARS (CURL Exemplar in Digital Archives) a élaboré un ensemble de métadonnées, publié début 2000. Le projet CEDARS est conduit par les universités de Leeds, Cambridge et Oxford au Royaume Uni. Les métadonnées couvrent l'ensemble des informations administratives, techniques et juridiques pour l'ensemble des fonctions d'archivage, dont la préservation.

– Le projet NEDLIB (Networked European Deposit Library) a publié son ensemble de métadonnées fin 2000. Le projet NEDLIB a été piloté par la Bibliothèque Royale des Pays-Bas et a rassemblé les Archives nationales des Pays Bas ainsi que les bibliothèques nationales d'Allemagne, Finlande, France, Italie à Florence, Norvège, Portugal, Suisse. Des éditeurs (Elsevier, Kluwers et Springer Verlag) ont également été associés. Les métadonnées ont été vues comme le minimum requis pour arriver à gérer la préservation d'une grande quantité de ressources lors de l'archivage des publications en ligne dont le Web dans la perspective d'une fonction nationale de dépôt.

Le quatrième ensemble de métadonnées examinées par le Groupe de travail OCLC/RLG ne suit pas le modèle de référence OAIS. Il a été développé par le Digital Repository Service de Harvad University. Il démontre comment on peut utiliser des structures XML pour encapsuler les métadonnées de préservation dans les objets lors de leur soumission à l'archivage.

Ce document est donc une base pour construire un consensus pour un standard de métadonnées de préservation qui corresponde à l'ensemble des besoins des bibliothèques.

3 Conclusion

Nous constatons que des progrès très importants ont été réalisés durant les cinq dernières années sur la préservation de l'ensemble des ressources électroniques que les bibliothèques sont susceptibles d'avoir dans leurs collections. Le modèle de référence OAIS offre une excellente base sur laquelle les bibliothèques peuvent construire leur propre standard de métadonnées sans s'isoler des autres communautés. Le travail déjà effectué peut permettre d'espérer un standard de bibliothèque dans un temps relativement court.
Cependant les questions fondamentales sur la nature du travail à effectuer pour créer ces métadonnées ne sont pas résolues. Pour certains projets, elles peuvent être créées pendant le catalogage descriptif par des catalogueurs spécialisés puisque des données techniques sont déjà enregistrées dans les notices descriptives en format MARC. Pour d'autres c'est un sujet très technique plus en relation avec une qualification en informatique et il serait très important de pouvoir les générer automatiquement plutôt

que de les saisir manuellement. Il est donc nécessaire de continuer à expérimenter l'implémentation de telles métadonnées et d'évaluer leur efficacité dans les traitements de préservation sur une durée de plusieurs années. La route est donc encore longue pour pouvoir garantir l'accès à long terme aux ressources numériques.

Bibliographie

Reference model for an open archival information system (OAIS) Red book issue 1 / Consultative committee on space data systems. 1999. 140 p. http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-1.pdf (visité le 20/07/2001)

Preservation metadata for digital objects : a review of the state of the art / A white paper by the OCLC/RLG Working group on preservation metadata. January 31, 2001. 50 p. http://www.oclc.org/digitalpreservation/presmeta_wp.pdf (visité le 20/07/2001)