NORMES ET LIGNES DIRECTRICES TECHNIQUES ET ORGANISATIONNELLES POUR LES INITIATIVES DE NUMÉRISATION DES PATRIMOINES CULTURELS SOUTENUES PAR LA COMMUNAUTÉ FRANÇAISE

Table des matières

- Introduction
- Le processus de numérisation
- Les formats de fichiers
- Les formats de stockage
- Les métadonnées
- Architecture des contenus
- Architecture des plates-formes
- Ressources documentaires

INTRODUCTION

Pourquoi des normes et lignes directrices ?

Le gouvernement de la Communauté française a adopté, le 19 octobre 2007, le Plan de préservation et d'exploitation des patrimoines – le Plan Pep's[1]. Ce plan concerne tous les champs de compétence de la Communauté française et les patrimoines qui y sont liés. Le Plan Pep's a un double objectif : d'une part, la préservation des patrimoines culturels en veillant à la sauvegarde et à la pérennité des collections et, d'autre part, l'exploitation des patrimoines culturels en assurant un accès interopérable, notamment via un portail fédératif.

La numérisation des patrimoines culturels de la Communauté française nécessite des investissements techniques, humains et financiers importants. La valorisation de ces investissements va de pair avec l'adoption d'une approche réfléchie de la création, de la gestion, de l'édition, de la diffusion et de la conservation des ressources numériques ainsi constituées.

L'accessibilité (à toute la population de la Communauté française[2] mais aussi aux équipes successives de concepteurs et de développeurs), la visibilité (par exemple par une présence mieux assurée dans les résultats des moteurs de recherche), l'interopérabilité (des fichiers numérisés entre eux et dans le temps)[3], la viabilité et la préservation à long terme (notamment par l'organisation de migration technique périodique vers de nouvelles versions de navigateurs ou vers d'autres terminaux) de ce patrimoine pluriel doivent être assurées.

Pour y parvenir, *« le plus simple et le plus efficace est d'adopter un ensemble de règles et de normes techniques minimales qui peuvent être comprises, aujourd'hui et demain, par tous les fournisseurs de contenus culturels et par tous les développeurs de sites web et qui en faciliteront la maintenance et la migration vers de nouveaux développements technologiques* »[4].

L'adoption de standards techniques communs est de pratique courante.

---

[1] http://www.culture.be/index.php?m=document_view&fi_id=530 .

[2] Une attention particulière doit être réservée à concevoir et maintenir l'accessibilité aux personnes déficientes sensorielles.

[3] Selon le NISO-National Information Standards Organisation, l'interopérabilité peut être définie comme la « *capacité d'échanger des données entre systèmes multiples disposant de différentes caractéristiques en terme de matériels, logiciels, structures de données et interfaces, et avec le minimum de perte d'information et de fonctionnalités* », Understanding Metadata, 2004 (http://www.niso.org/standards/resources/UnderstandingMetadata.pdf)

[4] Culture canadienne en ligne, « *Exigences et recommandations techniques* », version 4.0, 26 octobre 2007 (http://www.pch.gc.ca/pgm/pcce-ccop/publctn/tech-fra.cfm).

De tels standards sont reconnus par des organismes internationaux tels que l'ISO-Organisation internationale de normalisation[5], le W3C-World Wide Web Consortium[6] ou le CEN-Comité européen de normalisation au niveau européen[7].

« *Un critère important pour définir les standards est leur « ouverture ».* (…) *trois aspects* (sont) *d'un intérêt majeur pour les utilisateurs de standards :*

- *L'accès ouvert (au standard et aux documents produits au cours de son développement) ;*
- *L'utilisation libre (implémenter le standard induit peu ou pas de coûts liés aux droits de propriété intellectuelle, au travers de licences d'utilisation par exemple) ; et*
- *Le support orienté vers les besoins des utilisateurs plutôt que vers l'intérêt des fournisseurs de standards.*

(…) *Les spécifications des formats, interfaces et protocoles utilisés par les fournisseurs de ressources sont librement disponibles. De multiples développeurs peuvent donc développer des outils et services similaires en évitant la dépendance vis-à-vis d'un seul outil ou d'une seule plate-forme* »[8].

A la suite de ces recommandations européennes, la préférence est accordée, en Communauté française, à des formats ouverts reposant sur des normes et standards dont les spécifications sont publiques, même si, dans certains cas, le recours à des standards « propriétaires » d'utilisation fréquente et dans leur version normalisée publique peut se justifier.

Si la technologie est un élément essentiel de la numérisation des patrimoines culturels, elle n'est qu'un outil au service d'objectifs culturels soutenus par des structures d'organisation des archives patrimoniales qui doivent être pensées elles aussi en terme d'accessibilité, de viabilité et d'interopérabilité.

---

[5] www.iso.org.

[6] http://www.w3.org. Il a en charge la normalisation de l'ensemble des protocoles d'internet : standards de base (http, HTML, XHTML, DOM, XML, XSL, …), standards autour de l'interopérabilité et des services web (SOAP, WSDL, Web,…), standards concernant l'accessibilité (WAI), standards liés à la sémantique et à la description de ressources (XML Schema, RDF, langages d'ontologies OWL). Les documents et recommandations du W3C sont disponibles en français sous http://www.la-grange.net/w3c/fr-trans1 .

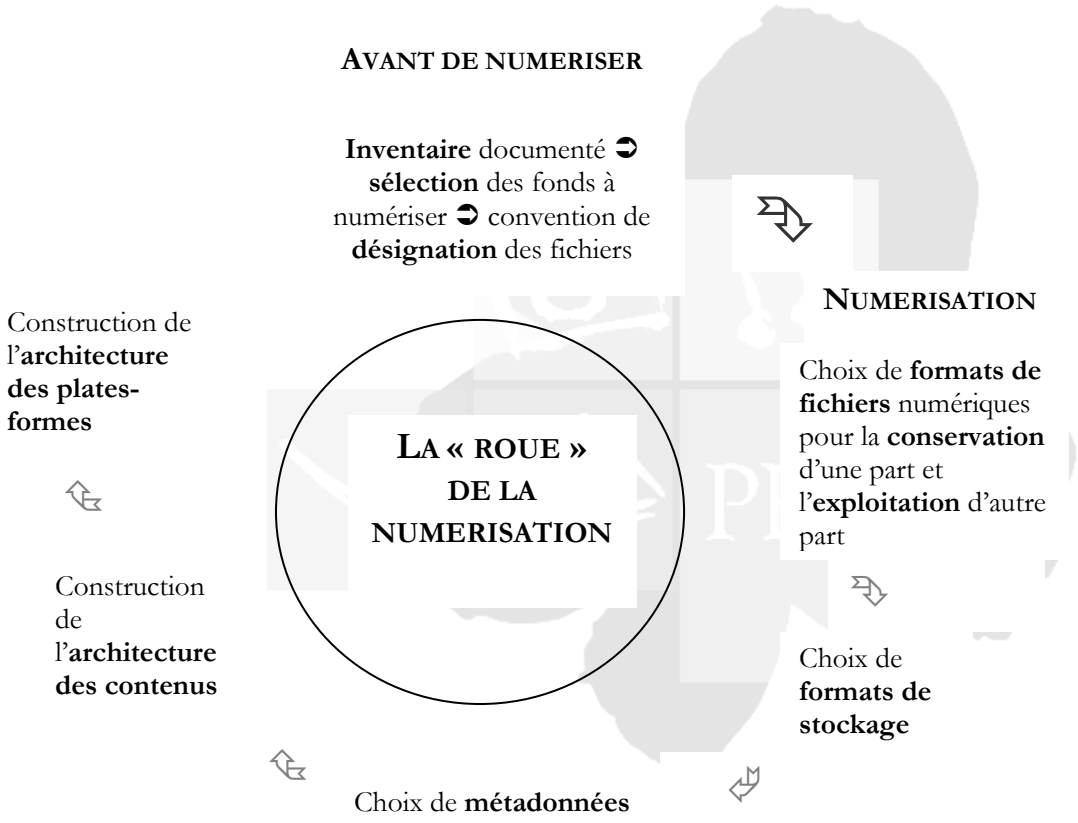
[7] http://www.cen.eu/cenorm/index.htm.

[8] UKOLN, Université de Bath, en collaboration avec MLA : le Conseil pour les musées, bibliothèques et archives, « *Recommandations techniques pour les programmes de création de contenus culturels numériques* », document rédigé dans le cadre du projet Minerva, version révisée le 7 mai 2004, p.10. La version française a été réalisée pour le compte de la Mission de la recherche et de la technologie du Ministère de la Culture et de la Communication (partenaire français du projet Minerva) par Muriel Foulonneau (Relais Culture Europe), Sarah Faraud (Relais Culture Europe) et Alexandra Bonnamy (traductrice). La version 2.0 de ce document a été diffusée (http://www.minervaeurope.org/interoperability/technicalguidelines.htm) en septembre 2008 en langue anglaise.

La « roue » de la numérisation

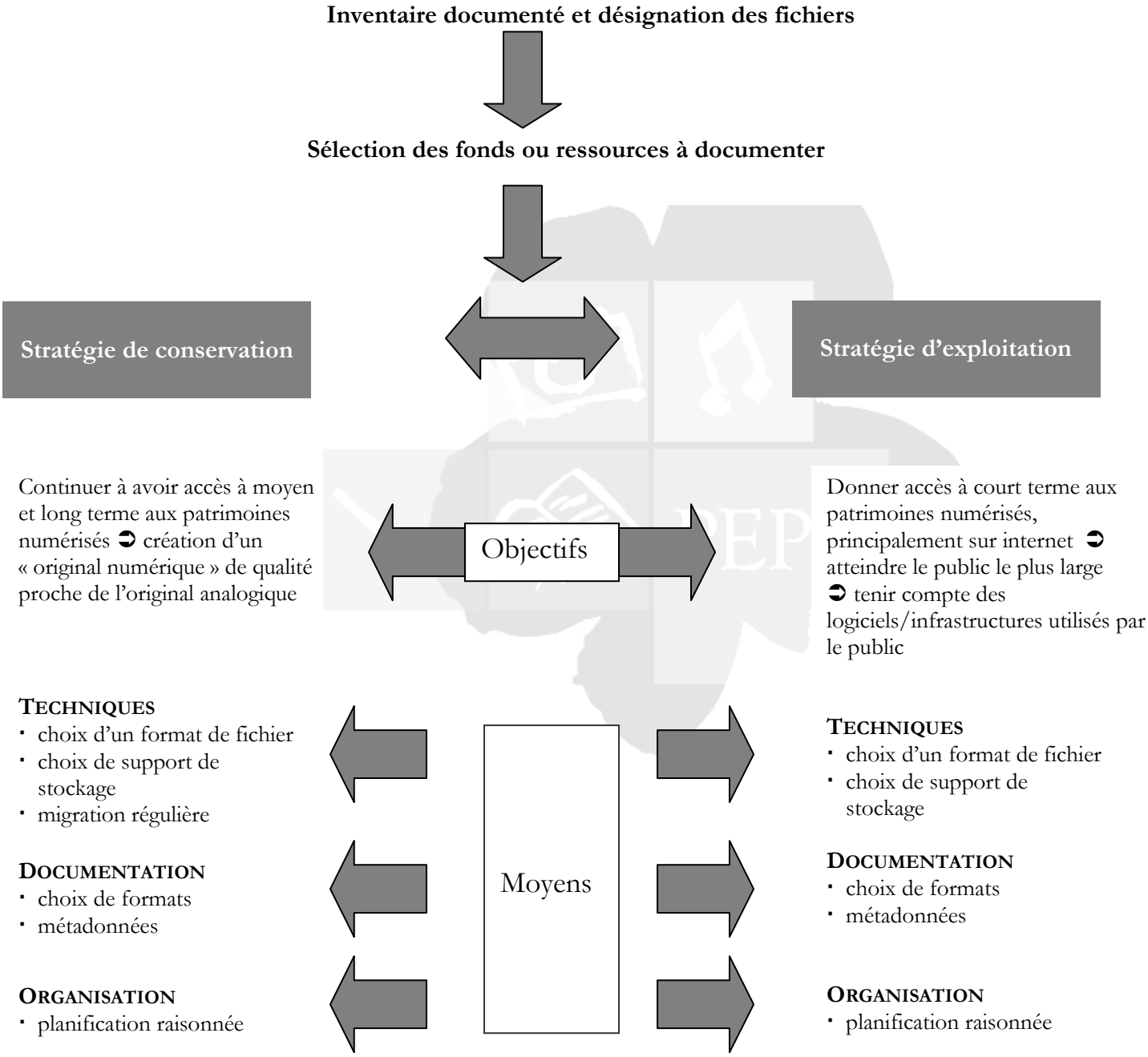
Numériser les fonds ou collections des patrimoines culturels de la Communauté française dans le double objectif d'en assurer la pérennité et l'interopérabilité nécessite une certaine organisation du travail.

Les différentes étapes de ce travail de numérisation sont présentées sous forme d'une « roue ». Pour chacune de ces étapes, des prescriptions que doivent respecter les initiatives de numérisation soutenues par la Communauté française sont explicitées et justifiées.



Les deux stratégies de la numérisation des patrimoines culturels

Deux stratégies sont parties intégrantes de tout processus de numérisation soutenu par la Communauté française de Belgique.



Les niveaux d'exigence

Dans la suite du document, différents niveaux d'exigence sont présentés en raison de leur importance respective pour atteindre les objectifs d'interopérabilité et de pérennité. Ils sont traduits par les verbes « doit, devrait et peut » :

- « doit » : il s'agit d'une prescription que tous les projets de numérisation doivent rencontrer ;
- « devrait » : il s'agit d'une recommandation de bonne pratique mais il convient d'en mesurer tous les effets pour chacun des projets de numérisation ou de rester attentifs aux évolutions futures ;
- « peut » : il s'agit de standards qui peuvent être utilisés, ce niveau de préconisation sera peu développé dans la suite du document car son objectif n'est pas de faire un inventaire exhaustif des normes et standards existants[9].

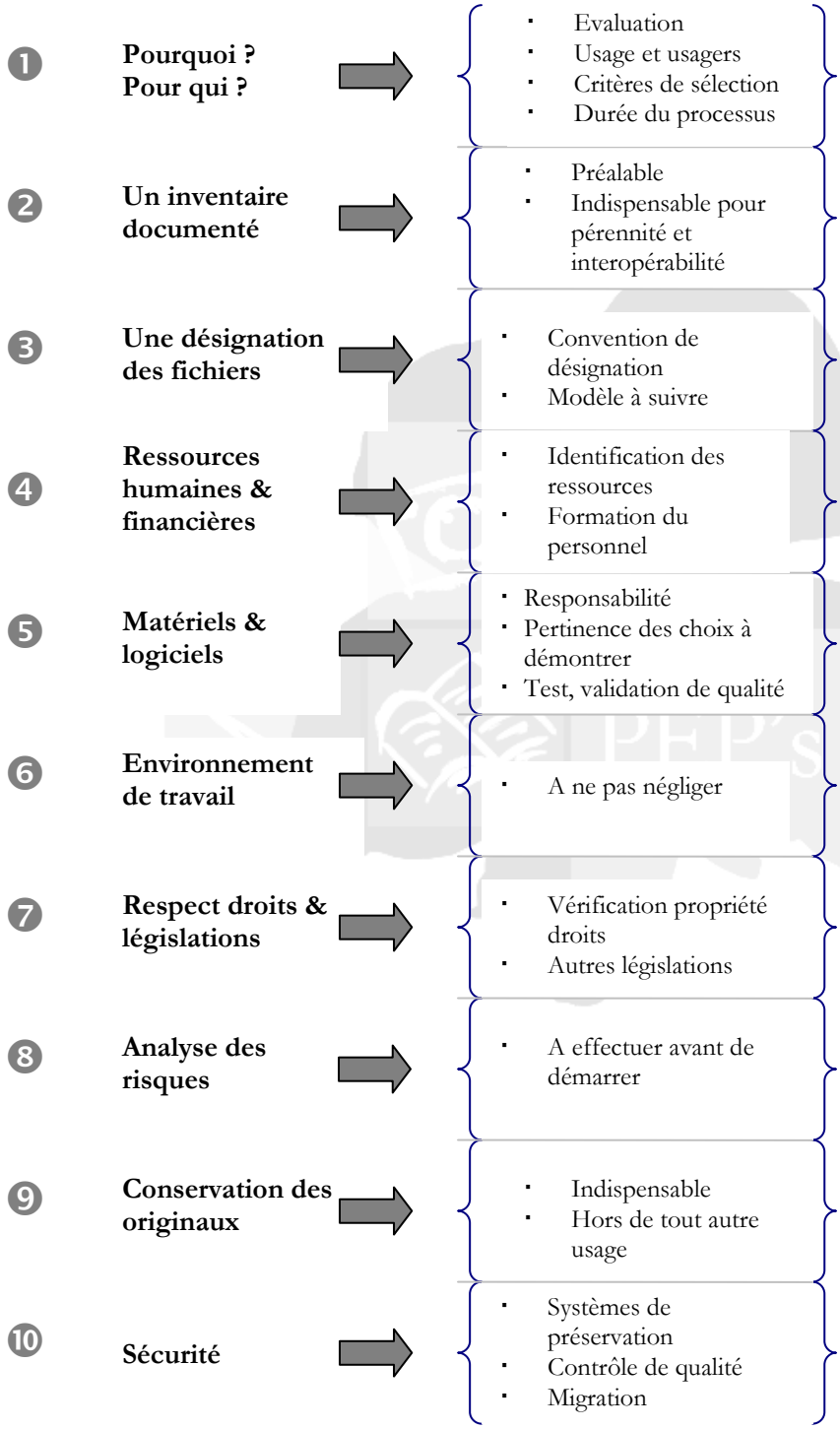
*
**

Il existe aujourd'hui un nombre impressionnant de guides de bonne conduite ou de lignes directrices en matière de numérisation des patrimoines. La plupart sont en langue anglaise. Dans la mesure de leur accessibilité sur le web, les sources en langue française sont privilégiées dans ce document. Nous nous en sommes largement inspirés, traduisant à notre réalité des procédures ou des concepts adoptés et éprouvés dans nombre d'institutions.

Ces recommandations seront régulièrement mises à jour.

[9] Ministère du Budget, des Comptes publics et de la Fonction publique de la République française, Direction générale de la modernisation de l'Etat, « *Référentiel général d'interopérabilité* », version 3, 22 juin 2007 (http://www.synergies-publiques.fr/article.php?id_article=746).

AVANT DE NUMÉRISER



« *La numérisation est la conversion de fonds analogiques ou physiques en fonds numériques en vue d'une utilisation par des logiciels. Les décisions prises au moment de la numérisation ont un impact sur la facilité avec laquelle les ressources créées pourront être gérées et accessibles, mais également sur leur viabilité* »[10].

Une planification est préalable à tout projet de numérisation.

Elle doit prendre en compte l'ensemble des aspects suivants.

❶ Tout projet de numérisation doit répondre à des **objectifs** à énoncer précisément et explicitement[11]. Ces objectifs – qui doivent être réalistes au regard des ressources - déterminent les critères de sélection des fonds/collections ou parties de ceux-ci (ressources) à numériser. Numériser l'ensemble des collections est rarement faisable, des choix sont dès lors à réaliser. Ces objectifs influencent également l'ensemble des choix en cours de processus de numérisation.

Doivent ainsi être présentés de façon détaillée :

- l'apport que la numérisation fournira aux institutions,
- les critères d'évaluation de l'intérêt et de la qualité du ou des fonds/collections : unicité ou la rareté, valeur culturelle, pédagogique, esthétique, documentaire, historique ou symbolique, ou encore valeur d'ensemble ou de « représentativité »,
- le respect des critères de qualité dans le cas d'une diffusion en ligne : être identifiable facilement, présenter des contenus pertinents, assurer la maintenance et les mises à jour des contenus, être accessible à tous les utilisateurs, être adapté aux besoins des utilisateurs, être réactif, être multilingue, s'efforcer d'être interopérable, être respectueux des droits, assurer la pérennité du site et des contenus en adoptant des stratégies et des standards adaptés[12],
- l'usage et les usagers des fonds/collections numérisés, les coûts inhérents au processus,
- le temps nécessaire ainsi que l'adéquation des fonds/collections à une présentation en ligne.

Les moyens techniques à mettre en œuvre dépendent de l'ensemble de ces facteurs.

❷ Un **inventaire documenté** des fonds/collections et de leur support doit toujours précéder tout projet de numérisation.

---

[10] « Recommandations techniques pour les programmes de création de contenus culturels numériques », *op.cit.*, p.17.

[11] La version 1.2 du Guide technique établi dans le cadre du groupe de travail Minerva par UKOLN précise que les objectifs devraient être « *SMART : Specific : expressed singularly ; Measurable : ideally in quantitative termes ; Acceptable : to stakeholders ; Realistic : in terms of achievement ; Time-bound : a timeframe is stade* », *op.cit.*, p.6.

[12] Minerva, « *Principes de qualité des sites internet culturels : guide pratique* », groupe de travail n°5, 2005, (http://www.minervaeurope.org/publications/qualitycommentary/qualitycommentary_fr.pdf).

Cet inventaire doit permettre de détecter les documents et les supports en péril, les documents/objets à caractère unique, … et d'opérer des choix et de fixer des priorités. La numérisation est en effet intrinsèquement liée à la création et l'existence de bases de données.

La valeur d'un fonds numérisé dépend certes de critères culturels, esthétiques et techniques mais aussi documentaires et pédagogiques.

La documentation de l'inventaire doit dépasser la simple liste des pièces y figurant, pour comprendre des informations d'identification et de description plus complètes.

A titre d'exemple, un tableau de ce type peut être dressé pour faciliter vos décisions en matière de numérisation de vos collections.

Format	Volume	Âge	Stockage	Caractéristique	Conditions	Actions	Délai
16mm négatifs films B&N		1950 à 1970	Archive, non circulant	Matériel original unique	Bonnes	Maintien dans de bonnes conditions, vérification périodique	5 ans
16mm Ektachrome		1968 à 1982	5 ans dans l'institution puis archivage	Informations, réutilisation fréquente	Certaines couleurs s'estompent	Copies d'accès en digibeta et DVD	2 ans si budget
16m cassette audio (mag sound track)		1950 à 1980	Archive	Originaux	Syndrome du vinaigre	Numérisation destruction des originaux	2 ans
16m mag sound track		1950 à 1980	Archive	Copies, sans valeur de conservation	Syndrome du vinaigre	Destruction après vérification originaux numérisés	2 ans

Source : PrestoSpace, http://prestospace-sam.ssl.co.uk/

❸ Chaque image numérique génère un fichier auquel il faut attribuer un nom spécifique. Le nommage des fichiers est une opération essentielle pour son exploitation future. Les fichiers ne doivent pas être nommés au coup par coup, un plan de nommage doit être élaboré en envisageant, pour la structure, tous les cas de figure pour l'ensemble des collections.

Une **convention de désignation des fichiers** aide à gérer le travail de numérisation et diminue le risque de perte d'informations. Elle accroît aussi leur interopérabilité, une meilleure découverte par les moteurs de recherche et la visibilité pour les utilisateurs (fournit aux utilisateurs des renseignements sur l'origine des objets, permet leur réutilisation et évite qu'ils ne soient « écrasés » lorsqu'ils sont utilisés à des fins d'agrégation, par exemple dans des collections d'images miniatures).

Chaque institution possède un système d'identification dans son catalogage. Les fichiers représentant les documents/objets de ces catalogues et ceux représentant les métadonnées et les structures doivent recevoir des identificateurs (noms de fichiers ; noms de dossiers) suivant un système cohérent et bien documenté. Dans des systèmes de gestion active, la solution usuelle est de mettre en place une base de données ou d'en acquérir une généraliste

(DAM/MAM[13]) qui se charge de dénommer ou renommer les fichiers de manière unique mais sans codage signifiant. Lorsque l'objectif est de mettre en place un système de gestion pérenne, il est souhaitable de ne pas devoir dépendre d'une base de données et de l'algorithme de dénomination d'un fournisseur (que l'on voudra peut-être remplacer un jour ou l'autre ou qui pourrait faire faillite). De plus, il est souhaitable de laisser dans les noms de fichiers et de dossiers le reflet des identificateurs des catalogues, lesquels sont souvent collés avec code-barres sur les pièces ou sur les boîtes. C'est pour avoir négligé cette remarque que la NASA ne retrouve plus les bandes originales des premiers pas de l'homme sur la lune. Il est aussi important que les noms de fichiers et de dossiers soient des noms uniques (des URN) et que les identifiants de localisation (des URL) soient construits comme la séquence de l'URL d'un dossier (ou volume) suivi de l'URN du fichier (dossier) concerné. Ceci est particulièrement important pour les archives parce qu'elles seront copiées en plusieurs exemplaires et sur divers supports pour être stockées en plusieurs endroits

Le modèle à suivre en Communauté française est le suivant :

- Un code en 3 lettres en majuscule déterminant le secteur d'activités : quatre identifiants ont ainsi été choisis : ARC pour les archives ; AVC pour les institutions audiovisuelles et cinématographiques ; ELB pour les éditions et bibliothèques ; MAR pour les musées et les arts plastiques ;
- Un « tiret » comme premier séparateur ;
- Un code en trois ou quatre lettres en majuscule identifiant l'institution (voir en annexe) ;
- Un « tiret » comme deuxième séparateur ;
- Une identification ou numéro d'immatriculation des fichiers concernés (numérisations ; fichiers de métadonnées ; empaqueteurs ; …). Pour les organismes qui gèrent plusieurs fonds/collections, ce champ documentaire doit commencer par le code du fonds/collection concerné suivi d'un code basé sur l'identification du catalogue existant (code-barres, …) ;
- Un « tiret » comme troisième séparateur ;
- Un code de variante qui permet de distinguer plusieurs fichiers ou dossiers qui doivent avoir le même début. Par exemple : diverses versions différant par la qualité (divers débits binaires par exemple) ;
- Quand il s'agit de fichiers, un troisième séparateur doit être placé : un « point » ;
- Quand il s'agit de fichiers, l'extension usuelle doit être ajoutée : Exemples : mp3 ; jpg ; wav ; xml.

En résumé, le codage d'un fichier est construit comme suit :
« code de secteur »-«code de l'institution »-« code de fonds »-«code d'item »-«variante ». «extension »

Exemple de nom de fichiers:
Pour un document des Archives et musée de la littérature : MAR-AML-4785325914-V5.odt

[13] Digital Asset Management (http://en.wikipedia.org/wiki/Digital_asset_management).

Les zones alphanumériques sont encodées en ISO-Latin-1 ; elles ne doivent comporter aucun signe diacritique.

Les directives conduisent donc à :

- des noms des fichiers et les structures de répertoire qui sont descriptifs et intuitifs, à la fois pour les créateurs de site et pour les utilisateurs ;
- assurer que les objets peuvent être identifiés au moyen d'une URL persistante, à des fins de citation, d'établissement de référence croisée et d'accès intégré. Les objets extraits à partir d'une base de données ne doivent pas comporter d'identificateurs attribués dynamiquement (par exemple de clés de session) intégrés à l'URL (ce qui va à l'encontre de la persistance).

❹ Les **ressources humaines et financières**[14] doivent être identifiées pour atteindre les objectifs du plan de numérisation et la formation du personnel affecté à cette tâche.

Décision doit être prise d'assurer la numérisation dans l'institution même ou faire appel à un sous-traitant. Le niveau adéquat de ressources internes à mobiliser (personnel, locaux, équipements, logistique, …) n'est évidemment pas le même dans un cas et dans l'autre.

❺ Des **solutions logicielles et matérielles** doivent être déterminées. Cette phase du processus de numérisation ressort de la responsabilité des institutions qui définiront les solutions techniques de capture numérique qui conviennent aux ressources à numériser en fonction de leurs caractéristiques particulières et eu égard aux usages recherchés.

Les institutions doivent en effet s'assurer – et démontrer – que les choix qu'ils opèrent en terme de matériels tels que scanners, appareils photos numériques, supports de copie, …, d'infrastructure informatique à laquelle ce matériel sera connecté, de logiciels de capture de l'image et son traitement, de logiciels pour les métadonnées et de « contrôle qualité », soient adéquats et appropriés pour atteindre les objectifs et les usages du projet de numérisation et cela à des coûts acceptables et proportionnés.

Pour ce faire, il convient de bien prendre en compte les caractéristiques des documents/objets à numériser (état de conservation, format, taille, couleurs, …) et de ne pas se fier, sans vérification, aux seules paroles des vendeurs de ces solutions logicielles et matérielles[15].

---

[14] Voir notamment le document de Minerva, « *Handbook on Cost Reduction in Digitisation* », septembre 2006 (http://www.minervaeurope.org/publications/CostReductioninDigitisation_v1_0610.pdf).

[15] Quelques références utiles :

- le site du Patrimoine canadien propose une évaluation de ressources en matériel et logiciels : http://www.chin.gc.ca/Francais/Contenu_Numerique/Materiel_Logiciel/index.html ,
- pour les photographies, voir notamment les recommandations du programme SEPIA : http://www.knaw.nl/ecpa/photo;

❻ **L'environnement de travail** - qualité de la lumière, taux d'humidité, vibrations, rangement, maniement et déplacement des originaux, mesure de risques de l'exposition d'originaux à la numérisation, … - doit être pris en considération, et le cas échéant des solutions trouvées, avant le démarrage du processus de numérisation.

❼ La vérification de la propriété des **droits** des documents/objets (droit de copie et droit de propriété intellectuelle) doit être effectuée avant – ou à tout le moins pendant – le processus de numérisation, et cela en fonction des besoins de l'institution concernée et des usages des sources numérisées.

❽ La planification de la numérisation doit comprendre une **analyse des risques**[16], de ses probabilités de survenance et des moyens d'y répondre.

Prévoir un « plan bis » peut éviter bien des désagréments.

❾ Les **originaux** doivent toujours être conservés en dehors de tout autre usage. Il est préférable de conserver les originaux de vos collections dans l'état où ils sont, en veillant toutefois à ne pas aggraver leur situation.

La numérisation est toujours liée à des pertes d'information. Elle ne remplace en aucun cas la conservation des originaux qui témoignent de bien d'autres choses que des informations que la numérisation peut aider à protéger (notamment en limitant la fréquence des utilisations des originaux) et à diffuser (simplification de l'accès rapide à l'information quand il s'agit de la reproduire et de la diffuser).

Par ailleurs, le fait que les originaux existent n'exonère pas de la conservation des images numériques, en raison notamment de leur propre dégradation. Une version doit être déposée en dehors de l'institution concernée et de stocker les données sur un format indépendant de l'application qui l'aura généré, donc de dissocier conservation et exploitation.

❿ La **sécurité des documents numérisés** doit être assurée régulièrement par des systèmes de préservation de l'information, un contrôle de qualité des fichiers et une migration des données.

- pour les images, voir le site du service du JISC britannique, Technical Advisory Services for Images, qui est hébergé à l'Université de Bristol (Institute for Learning and Research Technology) : http://www.tasi.ac.uk/ ;
- pour le son, voir les travaux du Comité technique de l'IASA, et notamment le document « *Sauvegarde du patrimoine sonore : Ethique, principes et stratégies de conservation* », IASA-TC 03, décembre 2005 : http://www.iasa-web.org/downloads/publications/TC03_French.pdf.

16 Voir notamment le document du service du JISC TASI, « *Risk Assessment* », mai 2006 (http://www.tasi.ac.uk/advice/managing/risk.html).

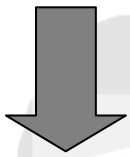
*
**

Vous n'êtes pas seul. N'hésitez pas à interroger la Délégation générale à la préservation et à l'exploitation des patrimoines de la Communauté française (02/413.26.45) pour toute question de numérisation de vos patrimoines culturels.

Notre site www.numeriques.be vous donne régulièrement des renseignements ou ressources complémentaires.

LES FORMATS DE FICHIERS

Objectifs :
Pérennité, qualité et fonctionnalité



Choix d'un format ouvert, non propriétaire
ou, à défaut,
Choix d'une version normalisée publique
d'un format propriétaire

Maintien du choix
Sauf migration requise en cas de format propriétaire

C'est quoi un format de fichier ?

Les fichiers numériques répondent à des principes prédéfinis de structure et d'organisation. Ils se distinguent par leur format[17] que l'on identifie grâce à l'extension du nom de fichiers (par exemple : .mp3) et qui sont en général explicités dans un document – plus ou moins important – appelé « définition de format ».

Le format de fichier **doit** être connu pour pouvoir correctement rendre intelligibles les informations le concernant. Sans cela, le document/objet numérisé ne serait qu'une succession de bits, de 0 et de 1, indifférenciés. Ce n'est toutefois pas suffisant, il convient aussi de connaître la syntaxe et les règles sémantiques que le format en question utilise pour transformer une information sous une forme numérique.

Les formats de fichiers se sont multipliés ces dernières années. Ils changent très rapidement, par exemple en raison de modifications induites par une utilisation particulière (par exemple : le format TIFF/EP pour les photographies digitales et le format TIFF/IT pour l'imprimerie professionnelle). Les versions se succèdent. Des registres de formats de fichiers ont été constitués pour aider à s'y retrouver[18].

Choix d'un format de fichier

Dans le choix d'un format de numérisation, des facteurs de pérennité, de qualité et de fonctionnalité **doivent** entrer en ligne de compte.

Comme tout logiciel peut devenir obsolescent, tout format de fichier traité par un logiciel est également vulnérable. Un format de fichier est d'autant plus pérenne qu'il cumule « *une utilisation répandue, des antécédents de rétrocompatibilité, une bonne prise en charge des métadonnées (dans un format ouvert tel que XML), une bonne gamme de fonctions mais une complexité raisonnable, la disponibilité d'interfaces avec des formes utilisables, des dispositifs intégrés de vérification d'erreurs, un cycle de mise à niveau raisonnable* »[19]. Disposer d'une documentation complète sur les spécifications techniques et les outils de vérification de l'intégrité du format choisi est au moins aussi important que de s'assurer que le format choisi est reconnu par un institut de normalisation.

Les facteurs de qualité et de fonctionnalité ont trait à la capacité du format de représenter les caractéristiques significatives d'un contenu donné pour les utilisateurs actuels et futurs.

---

[17] La Bibliothèque du Congrès définit les formats comme des "*packages of information than can be stored as data files or sent via network as data streams (aka bitstreams, byte streams)* (…) *A format is a fixed, byte-serialized encoding of an information model*».

[18] National Science Foundation/DELOS, *Invest to Save: Report and Recommendations of the NSF-DELOS Working Group on Digital Archiving and Preservation*, 2003. D'autres initiatives existent, comme celle des Archives nationales du Royaume-Uni (http://nationalarchives.gov.uk/pronom/) ou de la bibliothèque de l'Université d'Harvard (http://hul.harvard.edu/gdfr/ ou http://www.gdfr.info/).

[19] L'Université de Cornell et la société des archivistes, « *Guide de la conservation de collections numériques* » dont une version française est disponible sous http://www.library.cornell.edu/iris/tutorial/dpm-french/index.html .

Ces facteurs varient selon les genres particuliers ou les formes d'expression des contenus. Par exemple, les caractéristiques significatives pour le son sont différentes de celles qui sont pertinentes pour les images fixes de même que tous les formats numériques pour les images ne sont pas appropriés pour tous les types d'images fixes[20].

Certains formats sont non exclusifs et « ouverts » (ex : JPEG, MPEG-2, SVG, XML)[21]. Leurs définitions sont produites par des organismes internationaux de normalisation (ISO/CEI, CEN, W3C). Ils sont dès lors les plus sûrs du point de vue de leur disponibilité à long terme. D'autres formats sont « propriétaires » en ce sens que leurs spécifications ne sont pas nécessairement publiées et qu'ils font l'objet de licence commerciale ou d'un accord d'utilisation (ex : DOC, WMA ou WMV, PDF-8). Il existe des versions normalisées publiques de certains formats propriétaires (ex : PDF/A[22]). Ces versions sont préférables aux versions propriétaires car leur pérennité est assurée par la publicité, voire la normalisation (PDF/A est une norme ISO), de leurs spécifications techniques.

Des formats ouverts **doivent** être utilisés.

Si un logiciel « exclusif » ou « propriétaire » est nécessaire pour afficher, écouter ou utiliser certains contenus, sa version normalisée publique **doit** être choisie. Ce logiciel **doit** pouvoir être obtenu gratuitement par les utilisateurs, être téléchargeable avant que le contenu soit présenté et être indépendant de la plateforme. Dans l'hypothèse où l'utilisation de formats propriétaires est inévitable, une migration ultérieure **devrait** être organisée vers des formats ouverts.

D'autres différenciations sont opérées entre les types de formats disponibles. Sont ainsi distingués les formats « source » qu'il est possible de retraiter (ex : Word, RTF, ODT, SGML, XML) des formats « de présentation des données » que l'on ne peut modifier (ex : PDF, Postcript, Bitmap, TIFF, JPEG).

Une fois le choix de format arrêté, on **doit** maintenir ce choix tout au long du processus de numérisation des documents concernés. Un changement de format (par exemple du TIFF à JPEG) en cours d'une opération de numérisation posera des problèmes de traçabilité, de nommage, voire des difficultés en cas de retraitements ultérieurs. Il convient de réduire autant que possible le nombre de formats utilisés afin de minimiser les risques et ne pas obérer les possibilités de migration future.

---

[20] Library of Congress, "*Formats, Evaluation Factors and Relationships*, in *Sustainability* …", *op.cit*.

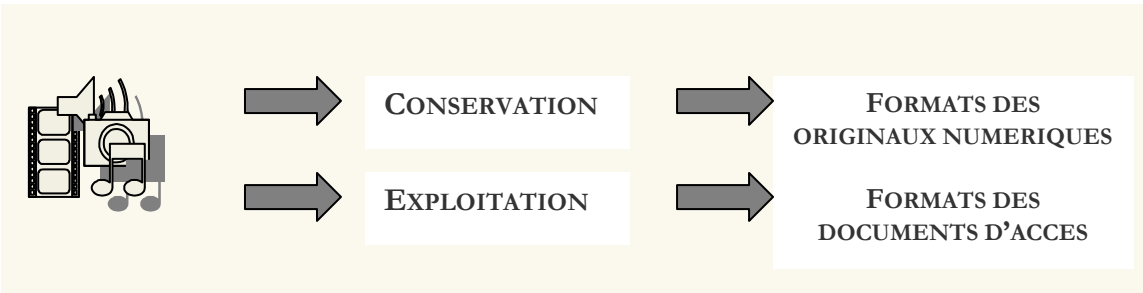
[21] Voir Open Source Observatory, IDABC (http://www.ec.europa.eu/idabc/en/chapter/452, devenu http://osor.eu; http://www.oasis-open.org/); OSS Watch Briefing Document, JISC OSS Watch (http://www.oss-watch.ac.uk/resources/fulllist.xml); Free & Open Source Software Portal, Unesco (http://www.unesco.org/cgi-bin/webworld/portal_freesoftware/cgi/page.cgi?d=1).

[22] Le format PDF est devenu une norme ISO en juillet 2008 (ISO 32000-1) ce qui signifie que la responsabilité de la publication des spécifications de la version actuelle (1.7), des mises à jour et du développement des versions ultérieures est du ressort de l'ISO (http://www.iso.org/iso/pressrelease.htm?refid=Ref1141).

Des ressources d'évaluation des possibilités de migration commencent à faire leur apparition[23].

Enfin, il convient de garder raison ; nous connaissons tous des exemples de formats considérés comme les meilleurs mais qui ne l'ont pas emporté sur le marché pour des raisons diverses.

Les deux stratégies de numérisation



Pour chaque type de contenus analogiques (texte, son, image fixe, image animée, 3D, archives web), des formats de fichiers pérennes sont recommandés. Les fonds/collections numérisés dans ces formats constituent les enregistrements de référence (« masters »), qui sont la représentation la plus fidèle possible ou la plus économiquement acceptable, de l'original. Ils doivent être conservés comme tels. Ce sont les originaux numériques.

Des versions d'utilisation et de consultation doivent parallèlement être créées, sous des formats de fichiers d'accès (« access »), souvent de moindre qualité, notamment à destination du web.

Les uns et les autres sont conformes aux Recommandations de Minerva, du JISC et de la Bibliothèque du Congrès.

[23] Voir la base de données PRONOM des archives nationales britanniques (http://www.nationalarchives.gov.uk/pronom/), la méthode de mesure du potentiel de durée de conservation de formats numériques (http://www.dlib.org/dlib/november04/stanescu/11stanescu.html); ainsi que le rapport de l'Université de Cornell sur les risques encourus par les formats de fichiers lors de migration (http://www.clir.org/pubs/abstract/pub93abst.html).

VOS DOCUMENTS CONSISTENT EN	FORMATS DE FICHIERS NUMERIQUES
	▪ Conservation : ODF ▪ Exploitation : PDF/A ▪ Balisage : XML
	▪ Images matricielles : Conservation : TIFF, PNG Exploitation : JPEG ▪ Images vectorielles : Conservation : SVG Exploitation : PDF/A, JPEG
	▪ Conservation : MPEG-2,-4, MJPEG2000 ▪ Exploitation : MPEG-2,-4
	▪ Conservation : WAVE ▪ Exploitation : OGG, MP3, WMA

NUMÉRISATION DES TEXTES

Conservation : ODF
Exploitation : PDF/A
Balisage : XML

Pas de format unique

Il n'existe pas à ce jour un format unique qui serait le meilleur quelque soit les documents textes concernés. Le choix doit tenir compte des caractéristiques des documents textes en question que les créateurs-éditeurs ou les utilisateurs considèrent comme les plus importantes : lisibilité, qualité de la présentation, intégrité (du texte, des diagrammes, des illustrations, des graphiques et des formules mathématiques ou autres), intégrité de la pagination et de la présentation, compréhension du contexte dans lequel le texte a été créé, navigation aisée (via une table des matières, par section ou via des liens explicites), recherche plein texte dans le document, capacité de citer page/section/chapitre et d'y être directement redirigé, capacité d'imprimer le texte en entier ou des passages sélectionnés, capacité de copier-coller des citations,…

Certains formats numériques rendent exactement le texte dans sa présentation originelle (ex : PDF), d'autres ont d'abord pour objectif de représenter la structure logique du document, la présentation du texte (caractères, layout,…) passant par l'utilisation de feuilles de style (ex : XML). Les documents textes ont été créés dans des buts très divers et leur utilisation au moment de leur création peut être supplantée par un autre type d'utilisation à l'avenir. Certains formats numériques sont d'utilisation aisée aujourd'hui mais offrent des fonctionnalités limitées pour l'avenir. Tous ces paramètres sont à prendre en compte.

Codage de caractères

Il est indispensable, pour l'échange d'informations sur l'internet par exemple, de préciser le codage de caractères utilisé : « *Un codage de caractères est un algorithme permettant de représenter des caractères sous une forme numérique en définissant une équivalence entre des séquences de code de caractères (les entiers correspondant à des caractères dans un répertoire) et des séquences de valeur de 8 bits (octets). Pour pouvoir interpréter les bits qui composent un objet numérique, une application doit connaître le mode de codage des caractères utilisé* »[24].

Le codage de caractère ASCII est utilisé pour les textes « pleine page » sans caractères accentués et l'ISO 8859-1 (latin-1) pour les textes avec caractères accentués. Le jeu de

[24] Recommandations techniques …, *op.cit.*, p. 20.

caractères ASCII est insuffisant pour un système informationnel tel que l'internet. HTLM utilise un jeu de caractères plus important, codés sur plusieurs octets, défini par la norme ISO 10646 et par la recommandation du consortium Unicode[25].

Conservation : formats des originaux numériques

Les originaux numériques doivent être réalisés dans un format conforme à la norme ODF-Open Document Format[26]. Très proche de la suite OpenOffice.org, de nombreux logiciels utilisent cette norme.

Les fichiers auront comme extensions selon le type de fichier : .odt pour du texte formaté ; .ods pour un tableur ; .odp pour la présentation ; .odg pour un dessin ; .odf pour une formule ; .odc pour un diagramme ; .odm pour le document principal, … Il existe aussi des modèles dans OpenDocument qui contiennent des informations sur le formatage de documents ; les extensions sont alors, par exemple, .ott pour le texte formaté ; .ots pour un tableur ; .otp pour la présentation.

Exploitation : formats des documents d'accès

Pour les documents d'exploitation, un format « propriétaire » peut être choisi : par exemple, le format PDF d'Adobe. Dans ce cas, il convient de choisir la version publique de ce format, à savoir PDF/A[27], qui est une norme ISO.

Tout contenu en mode texte peut aussi être produit dans des formats tels que RTF ou ASCII ou encore sous forme de fichier texte délimité. Ces fichiers ne sont acceptables que si les fichiers sont destinés à être téléchargés, stockés ou manipulés par les utilisateurs à l'extérieur de l'environnement du navigateur et s'ils ne sont pas destinés à remplacer un contenu créé en (X)HTML.

Balisage

Pour faciliter sa viabilité à long terme, tout contenu en mode texte doit être créé et géré dans un format structuré adapté à sa présentation sous la forme de fichiers HTML[28] (ou

---

[25] http://www.unicode.org.

[26] Le format ODF (OASIS Open Document Format for Office Application) est un format non propriétaire et une norme ISO depuis 2006 (ISO 26300). Il est soutenu par la Commission européenne et rendu obligatoire par le gouvernement fédéral belge dans son administration à partir de septembre 2008 (http://www.oasis-open.org/; http://www.odfalliance.org/). Il est compatible avec les normes de métadonnées du Dublin Core.

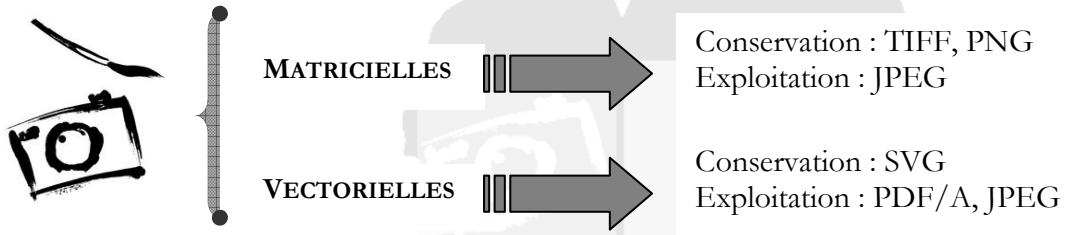
[27] http://www.pdfa.org/doku.php , http://www.digitalpreservation.gov/formats/fdd/fdd000125.shtml.

[28] HTML-HyperText Markup Language est un format ouvert, créé et utilisé pour décrire des pages Web, et en particulier pour construire les interfaces d'applications Web (http://www.w3.org/TR/html4/; http://www.la-grange.net/w3c/html4.01/cover.html). Le format XHTML est une reformulation de HTML. Il a pour caractéristique que le contenu est séparé de la forme. Il respecte la syntaxe XML (http://www.w3.org/TR/xhtml1 ; http://www.la-grange.net/w3c/xhtml1/).

XHTML) ou XML[29], ces deux langages « de balisage » sont recommandés par le W3C pour la description de la structure sémantique des contenus (identifier un en-tête, un paragraphe, une liste,… et les liens entre ces éléments). Il est préférable d'utiliser la dernière version disponible et de vérifier que le code soit être exempt d'erreurs et validé en fonction de la définition du Consortium W3C s'appliquant au langage choisi.

Dans la plupart des cas, la meilleure solution est de stocker le contenu sous la forme de textes en XML (voir aussi au chapitre « Métadonnées »).

NUMÉRISATION DES IMAGES FIXES



Les images matricielles et vectorielles

Deux types d'images sont à distinguer : les images matricielles et les images vectorielles.

Les images matricielles sont des images en mode point (en anglais « bipmap ou raster »). Le modèle « matriciel » repose sur la décomposition de l'image en une matrice de cellules, par un simple quadrillage. Ces cellules, habituellement carrées, sont les plus petits éléments de l'image, caractérisés par une couleur unique et indivisible, appelés « pixel » (PIcture ELement). Plus la taille des pixels est faible, plus le modèle sera proche de l'image représentée. La résolution de l'image correspond au nombre de pixels par unité de longueur ou de largeur, elle est exprimée en points ou en pixels.

Les images vectorielles (en anglais « vector images ») contiennent des représentations mathématiques d'objets, comme des points, les lignes et des cercles, des courbes, ... Chaque objet possède des paramètres de couleur, de forme, de position, de taille. Le modèle « vectoriel » décompose l'image en éléments simples comme des segments de droite, des cercles et, plus généralement, des formes géométriques prédéfinies. Une image vectorielle permet une résolution d'image quasiment infinie. A résolution égale, elle est souvent moins volumineuse qu'une image matricielle.

[29] XML-Extensive Markup Language est un format ouvert. De nombreux langages respectent la syntaxe XML (XHTML, SVG, XSLT, …). Objectif : faciliter l'échange automatisé de contenus entre systèmes d'informations hétérogènes (http://www.w3.org/TR/xhtml1; http://w3.org/TR/REC-xml).

Critères de choix de formats

Avant tout choix de formats, il convient d'être attentif, pour les images matricielles, à des paramètres de qualité, en choisissant une résolution spatiale élevée (la qualité de la résolution affectera la capacité de « zooming » et le potentiel d'agrandissement) et une résolution de couleur. La Bibliothèque du Congrès inclut dans ces paramètres des caractéristiques relatives à la compression pour les couleurs et les échelles de gris[30] et à des fonctionnalités particulières requises par les « dépositaires » ou attendues par les utilisateurs finaux.

Pour les images vectorielles, l'attention doit porter d'abord sur le caractère ouvert ou standard du format et sa large utilisation.

Formats pour images matricielles

Conservation : formats des originaux numériques

Pour les originaux numériques, le choix porte préférentiellement sur le format TIFF[31]. L'échantillonnage devrait être fait avec 600dpi et 24 bits par pixel pour les documents de taille A4. Pour les autres tailles de document, il convient d'adapter cette résolution. Un critère de référence est de prendre comme cible 10 millions de pixels par « item ».

Le format PNG[32] peut également être choisi, notamment pour les images fixes non photographiques telles que des dessins, des icônes, des logos ou des schémas ; il sera de toute façon préféré au format GIF[33]. Des formats liés au format PNG sont disponibles

[30] "*Uncompressed bitmapped images preferred over use of lossy compression, lossless compression scheme that are not fully disclosed or are subject to use-based licence fees; lower compression ratios are preferred over higher for same compression scheme and resolution; Discrete Wavelet Transform (DWT) compression (e.g. JPEG 2000) preferred over Discrete Cosine Transform (DCT) compression (e.g. JPEG)*".

[31] TIFF-Tagged Image File Format, développé par Aldus qui a été racheté par Adobe, est un format propriétaire (http://partners.adobe.com/public/developer/tiff/index.html). Le format TIFF/EP (TIFF for Electronic Photography) est toutefois une norme ISO (ISO 12234-2 :2001). Ce format est fréquemment proposé comme format par défaut dans des logiciels de numérisation. Il gère toutes les profondeurs de couleurs et intègre des informations de correction Gamma. Il comporte de nombreuses variantes utilisées dans des applications très diverses, des scanners industriels aux appareils photo numériques en passant par des imprimantes. Utilise des modes de compression avec ou sans pertes.

[32] PNG-Portable Network Graphics est un format reconnu par le consortium W3C et est une norme ISO/CEI (15948 :2004 ; IETF RFC 2083) (http://www.w3.org/TR/PNG/). Ce format d'images matriciel sans perte a été développé pour fournir une alternative au format propriétaire GIF. Il est basé sur l'algorithme de compression LZW (Lempel-Ziv-Welch, propriété d'Unisys). Il peut aussi remplacer beaucoup d'usages habituels du TIFF. Il supporte tous les styles d'images bitmap. Il peut gérer une couche alpha de transparence. Il peut intégrer le codage de correction Gamma et des métadonnées. Il est adapté à l'enregistrement d'images pour l'internet. Il comporte aussi des fonctions telles que la signature électronique.

[33] GIF-Graphical Interchange Format est développé par ComputerServe Interactive Services Inc. (Skale Industries) (http://www.w3.org/Graphics/GIF/spec-gif87.txt). C'est un des formats d'images numériques les plus courants sur l'internet. Il utilise l'algorithme de compression LZW (Lempel-Ziv-Welch), propriété également d'Unisys. Il ne code pas plus de 256 couleurs par pixel, au-delà les images subissent une perte de qualité.

pour des applications particulières comme l'animation simple et/ou de courte durée d'images (MNG-Multiple-Image Network Graphics, APNG-Animated Portable Network Graphics conçu pour faire des animations pour le Web). Le format DNG[34] est un format ouvert d'enregistrement des signaux générés par les capteurs d'appareils numériques ; basé sur un format TIFF/EP, il a pour but de standardiser les nombreux et incompatibles formats propriétaires RAW, de plus en plus utilisés en photographie numérique.

Des formats « spécialisés » à des types de documents sont également disponibles : par exemple, le format CGM[35] pour les objets graphiques à deux dimensions (bitmap, texte ou vectoriel), ou encore le format X3D[36] pour la description d'objets et d'univers virtuels en 3D (visite virtuelle de musée, art graphique, imagerie médicale ; conception d'architecture, simulateur pour la formation, applications géospatiales, …).

Le format doit être sans perte d'information (« lossless »). De préférence, ne pas utiliser de compression.

Exploitation : formats des documents d'accès

Pour les documents d'accès, le choix porte préférentiellement sur le format JPEG à différents niveaux de compression et de résolution suivant les besoins[37]. Pour proposer un accès adapté au contexte d'utilisation, plusieurs tailles différentes devraient être disponibles.

Formats pour images vectorielles

Conservation : formats des originaux numériques

[34] DNG-Digital Negative (http://www.adobe.com/fr/products/dng/).

[35] CGM-Computer Graphics Metafile, norme ISO/CEI 8632-1 :1999, est un format adapté à de nombreux secteurs industriels comme l'aéronautique, l'automobile, défense, télécoms, … CGM est utilisé dans ODA (Office Document Architecture ISO 8613) (www.cgmopen.org/; www.oasis-open.org/news/oasis-news-2007-01-30.php).

[36] X3D-Extensible 3D a été créé par le consortium W3C dans le but de succéder à VRML 2.0 (Virtual Reality Modeling Language) ; ce format est une norme ISO 19775-1 :2004 et ISO/CEI 19776-1 :2006 (http://www.web3d.org/about/overview/). Un guide de bonne pratique est disponible sous http://www.vads.ahds.ac.uk/services/advice/guides.html.

[37] JPEG-Joint Photographic Expert Group, groupe d'experts ISO/CEI qui a donné son nom à la norme de compression (haut taux avec perte) d'images numériques puis au format de données et au format de fichier le plus utilisé pour contenir ces données (http://www.w3.org/Graphics/JPEG/; http://www.jpeg.org/public/jfif.pdf; http://www.itu.int/rec/T-REC-T.800-200208-I/fr). C'est un standard pour l'internet et pour les appareils photos numériques. JPEG/SPIFF-JPEG Still Picture File Interchange File Format, ou JFIF, est le format de fichier adapté à l'image compressée en JPEG (non spécifié dans la norme ISO). La norme JPEG2000 est une norme publiée en 2004 et fait l'objet d'une Recommandation de l'UIT(T.800). Elle permet de traiter les images en bitonal, en niveaux de gris ou en couleurs et intègre le mode pyramidal. On peut choisir un mode de compression paramétrable en fonction du résultat souhaité, avec ou sans perte. A propos du format JPEG2000, voir notamment l'étude récente (février 2008) de Robert Buckley publiée par Digital Preservation Coalition (http://www.dpconline.org/docs/reports/dpctw08-01.pdf) et les lignes directrices de Minerva, 31 janvier 2008 (http://www.minervaeurope.org/structure/wg/3D_IT.html).

Pour les originaux numériques, les images vectorielles devraient être créées et stockées de préférence en utilisant le format ouvert SVG[38].

Le format propriétaire Macromedia Flash Payer SWF[39], développé par Adobe, peut également être approprié. Toutefois, une migration vers des formats ouverts devra être organisée dès que possible.

Exploitation : formats des documents d'accès

Pour les documents d'accès, le format est de préférence le PDF/A. Dans certains cas, cependant, le format JPEG s'impose. Pour la publication sur internet, le format PNG peut être utilisé.

NUMERISATION DES IMAGES EN MOUVEMENT

Conservation : MPEG-2,-4, MJPEG2000
Exploitation : MPEG-2,-4

La numérisation de la vidéo ajoute la dimension « temps » à la numérisation des images fixes. En principe, le processus de numérisation est similaire à celui relatif aux images matricielles avec la vitesse de succession des images en plus. Le nombre d'images par seconde est appelé en anglais « frame rate ». La qualité d'une vidéo numérique dépend de trois facteurs : la résolution, la profondeur de l'intensité des couleurs et la fréquence de trame (nombre de fois par seconde où l'image est redessinée). Comme les vidéos numérisées produisent de grande quantité de données, résultante de la succession d'images (en général 25-30 par seconde), le mode de compression est très important.

Critères de choix d'un format

La difficulté réside dans le très grand nombre de formats analogiques utilisés et dans l'absence de standardisation. Souvent, les originaux sont encodés dans des formats propriétaires. Les critères de choix vont dépendre des besoins de l'institution qui numérise ses collections et du type d'images. La Bibliothèque du Congrès a établi des tableaux qui déterminent l'importance respective des paramètres selon le type d'images à traiter et propose les formats qu'elle estime « recommandables » ou « acceptables ».

---

[38] SVG-Scalable Vector Graphics, développé par W3C, est basé sur XML (http://www.w3.org/TR/SVG/). Il soutient le schéma Xlink, a la possibilité d'être agrandi ou basculé, interface le langage SMIL et est très utilisé en cartographie et sur téléphone portable.

[39] http://download.macromedia.com/pub/flash/licensing/flash_fileformat_specification.pdf.

Description	Clarity & fidelity (picture & sound resolution)	Sound field (beyond stereo)	Special functionality	Effect of technical protection	Formats préférables
Moving image productions for theatral distribution or specialized copies for archiving by creators or distributors. May have surround sound.	Very important. Frame integrity likely to be important, extended dyna mic range may be important	If surround sound, retain with minimal change	Downsample take excerpts,etc., without artifacting	Must not affect clarity	*DPX_2[40] together with suitable format for sound information *DCDM_1_0[41] *MXF[42]
Video productions fully realized prior to dissemination via terres trial, cable or satellite broadcast (e.g. made-for-cable programs, TV dramas, documentaries). May be high definition, may have surround sound	Very important retain HD if present. Frame integrity may be important	If surround sound, retain with minimal change	Downsample take excerpts,etc., without artifacting	Must not affect clarity or normal rendering	*MXF *MJP2_FF_LL[43] *uncompressed or loss less compressed in others wrappers (AVI, QuickTime[44], WMV) *MPEG-2,-4
Video productions fully realized prior to dissemination via videotape, DVD disk or internet (e.g. promotional programs, independent productions, oral histories). May be high definition, may have surround sound	Very important or important depending on item, retain HD if present.	If surround, may be normalized to stereo	Downsample take excerpts,etc., without artifacting	Must not affect clarity or normal rendering	*MPEG-2 *MXF file containing MXF_GC_MPEG-2 with HD and AAC surround *MPEG-4_AVC *MPEG-4_V
Video programs for terrestrial, cable or satellite broadcast assembled at transmission time (e.g. scheluded newscasts, news specials, studio talk shows). May be HD	Less important, retain HD if present			Must not affect normal rendering	*MPEG-2 *MPEG-4_AVC *MPEG-4_V
Cybercasts streamed over the internet, other than program material covered in M3 or M4	Less important			Not important	*MPEG-2 *MPEG-4_AVC *MPEG-4_V
Video incidental to web harvesting (e.g. short animations that illustrate a web page)	Not important				Any
Encoding for dynamically generating animations and/or interactive programs, e.g. animated shorts for web delivery or for playback on personal computers and the animated output of CAD-CAM systems, but not deemed appropriate for "save as video"	Retain precision of original	N/A	Retain functionality of original via performance & composition software	Must not affect functionalit y for end users	*FLA *SVG

Source : Library of Congress, « Curator's View for Moving Image Content », *op.cit.*

[40] DPX-Digital Picture Exchange est un format de fichier couramment utilisé en cinéma numérique ; c'est un standard reconnu par l'ANSI-Academy of Motion Pictures Arts and Sciences et le SMPTE-Society of Motion Picture and Television Engineers (http://www.oscars.org/ et http://www.smpte.org/home).

[41] DCDM-Digital Cinema Initiative Distribution Master (http://www.dcdm.com.au/) est la matrice à partir de laquelle sont réalisées les copies numériques des longs-métrages, baptisées DCP par les studios hollywoodiens (équivalent de la copie 35mm).

[42] MXF-Material Exchange Format est un conteneur ou « wrapper » audio et vidéo développé par la Society of Motion Picture and Television Engineers-SMPTE.

[43] MJP2-Motion JPEG 2000 File with Lossless Encoding.

[44] QuickTime est un format de fichier développé par Apple mais disponible aussi sous Windows (http://www.apple.com/quicktime/whyqt/).

Conservation : formats des originaux numériques

Minerva **recommande** de préférer une forme non comprimée (format RAW AVI), sans traitement ultérieur, sans aucun codec (un codec est un algorithme permettant de réduire significativement les flux de données en compressant/décompressant les données vidéo), à une taille de trame de 720x576 pixels, une fréquence de trame de 25 images par seconde, avec 24 bits couleur. La vidéo **devrait** être créée à la plus haute résolution possible, à une intensité de couleur et une fréquence de trame abordables et pratiques compte tenu des utilisations prévues. Le codage de couleurs PAL **devrait** être utilisé. La vidéo **peut** être créée et stockée en utilisant le format MPEG approprié [45].

Si l'on veut augmenter la vitesse de consultation des fichiers ou minimiser l'espace de stockage nécessaire pour les documents numérisés, il est nécessaire d'avoir recours à des systèmes de compression. Plusieurs systèmes existent, avec perte ou sans perte d'informations[46]. Une fois le mode de compression choisi, il faut éviter des cycles successifs de compression et de décompression.

La Fédération internationale des archives de télévision-FIAT[47] annonce qu'elle mettra prochainement des lignes directrices techniques à disposition sur son site.

En attendant une standardisation, pour les documents à très haute valeur culturelle, le choix du format M-JPEG2K sans perte (« lossless ») et à la résolution initiale est envisageable mais **doit** être strictement réservé à ces cas.

Sont **recommandés** pour la conservation de :

- séquences vidéo basse définition : MPEG-2 ; pour les flux audio en MPEG-2, choisir AAC[48] ;

[45] MPEG désigne le Moving Picture Experts Group, groupe de travail du comité technique mixte de l'ISO et de la CEI pour les technologies de l'information qui est chargé du développement de normes internationales pour la compression, la décompression, le traitement et le codage de la vidéo et de l'audio (http://www.iso.org/; http://www.mpeg.org/ ; http://www.itu.int/rec/T-REC-H.264.2.fr). La norme MPEG-1 est une norme de compression pour la vidéo numérique. La norme MPEG-2 définit les aspects compression de l'image et du son et le transport à travers les réseaux pour la télévision numérique. Ce format est utilisé pour les DVD, CVD et SVCD avec différentes résolutions d'images, pour la diffusion de la télévision numérique. La norme MPEG-4 est une norme de compression pour les images animées, englobe les nouvelles applications multimédias comme le téléchargement et le streaming sur le réseau internet, le multimédia sur mobile, la radio numérique, les jeux vidéo, la télévision et les supports haute définition. Applicable aux bas débits. MPEG a aussi développé les normes MPEG-7 (norme de description pour la recherche de contenu multimédia), MPEG-21 (norme proposant une architecture pour l'interopérabilité et l'utilisation de contenus multimédias) et MPEG-A (applications multimédias). Références : MPEG-1 ISO 11172 :1993 ; MPEG-2 ISO/IEC 13818-1 :2000 ; MPEG-4 ISO 14496 :2004 et UIT-T Rec H.264 2005.

[46] http://fr.wikipedia.org/wiki/Motion-JPEG.

[47] http://www.fiatifta.org/cont/index.aspx.

[48] AAC-Advanced Audio Coding est un algorithme de compression audio avec perte de données (http://www.iis.fraunhofer.de/EN/bf/amm/index.jsp). C'est un standard reconnu par le Moving Pictures Expert Group.

- séquences vidéo en haute définition, services audiovisuels, vidéo-conférences et visiophonie : MPEG-4 (avec un débit de référence de 15 Mbps pour l'équivalent SD et 45 Mbps pour l'équivalent HD) ;

tous formats compressés dans des conteneurs (« wrappers ») comme AVI[49], QuickTime ou WMV[50].

Exploitation : formats des documents d'accès

Il convient d'être attentif au contexte technique (largeur de la bande passante) de l'utilisateur.

La vidéo destinée à être téléchargée sur internet **devrait** être publiée dans le format MPEG approprié (voir ci-dessus) ou les formats propriétaires WMF, AVI ou QuickTime.

Aucune recommandation pour la diffusion en mode continu (« streaming ») de flux audio ou vidéo ; Real Media et Windows Media sont considérés comme des formats possibles[51].

Les fichiers vidéo qui sont chargés dans un lecteur incorporé au navigateur **ne doivent pas** démarrer automatiquement et le lecteur **doit** être doté de commandes de démarrage et d'arrêt de la vidéo. Tous les liens à des grands fichiers vidéo, c'est-à-dire les fichiers de plus de 50 Ko, **doivent** être accompagnés d'une étiquette indiquant la durée et la taille du fichier. Si un codec est utilisé pour produire des fichiers vidéo, il doit pouvoir être obtenu gratuitement afin d'être installé par l'utilisateur. Un lien au codec (logiciel qui code et décode ou qui comprime et décomprime) **doit** être fourni aux utilisateurs qui ont besoin de la télécharger et de l'installer. Si des fichiers vidéo sur internet sont préparés pour être transmis dans un environnement à large bande, une version à bande étroite doit aussi être fournie aux utilisateurs. Tous les fichiers vidéo **devraient** être accompagnés d'un audio-script ou d'un sommaire en mode texte (X)HTML permettant à l'utilisateur, en particulier l'utilisateur déficient sensoriel, d'en comprendre le contenu sans avoir à les visionner[52].

---

[49] AVI-Audio Video Interleave est un standard propriétaire (Microsoft). Dans un fichier AVI, chaque composante audio ou vidéo peut être compressée par n'importe quel codec. Le format AVI permet de réunir en un seul fichier une piste vidéo et jusqu'à 99 pistes audio, ce qui permet de bénéficier, par exemple, de plusieurs langues pour un même film.

[50] WMV-Windows Media Video est un format propriétaire (Microsoft). Ce codec, fréquent sur internet tant en streaming qu'en téléchargement, est directement en concurrence avec MPEG-4 (http://www.microsoft.com/windows/windowsmedia/fr/content_provider/film/hdvideo.aspx).

[51] "Référentiel d'interopérabilité…", *op.cit.*

[52] Culture canadienne en ligne, *op.cit.*

NUMÉRISATION DU SON

Conservation : WAVE
Exploitation : OGG, MP3, WMA

Critères de choix d'un format

Les critères déterminants pour la numérisation du son (musique et parole) tiennent à la nécessaire fidélité de sa reproduction. La numérisation du son se réalise par la technique d'échantillonnage. En général, la qualité du son numérisé est déterminée par deux facteurs : le taux d'échantillonnage et le nombre de bits par échantillon. Par exemple, un CD normal comprend 44.100 échantillons par seconde et 16 bits par canal (gauche et droite) et par échantillon. Pour la fidélité de la reproduction du son, il est préférable de choisir un haut taux d'échantillonnage, un format non compressé et un « higher data rate » (par exemple, 128 kilobits par seconde), sans traitement ultérieur tel que la réduction de bruit.

Conservation : formats des originaux numériques

Les enregistrements audio devraient être créés et stockés sous un format non comprimé tel que WAV (ou WAVE[53]) ou AIFF[54], mais peuvent aussi être créés et stockés sous des formats comprimés tels que MP3[55], WMA[56], Ogg-Speex[57], Ogg-Vorbis, sans perte de données pour les séquences sonores de haute qualité (par exemple sur CD audio en haute

[53] Standard propriétaire développé par Microsoft et IBM (http://www.partners.adobe.com ; http://fr.wikipedia.org/wiki/WAVEform_audio_format.) Conteneur basé sur le format RIFF-Resource Interchange File Format, sans compression. La technique d'échantillonnage utilisée est la modulation par impulsion et codage (en anglais PCM-Pulse Code Modulation).

[54] AIFF-Audio Interchange File Format est un format de fichier audionumérique développé par Apple. Les données sont codées en PCM big-endian sans compression. Il existe un format compressé : AIFFC.

[55] Norme ISO-ITU (ISO/IEC 11172 :1993 ; 13818 ; 14496), MP3 est la spécification sonore du format MPEG-1 qui est destinée à être appliquée à un support numérique assurant un débit de transfert total continu d'environ 1,5Mbit/s (CD, DAT, disques durs magnétiques). Compression avec pertes et à débit constant. Format courant sur internet (http://www.chiariglione.org/mpeg/; http://www.iso.org).

[56] WMA-Windows Media Audio est un format propriétaire développé par Microsoft. C'est le format concurrent de MP3. Il permet de stocker de grandes quantités de sons (musique, voix) avec ou sans pertes et de protéger les fichiers de sortie contre la copie illégale (technique dénommée DRM-Digital Right Management ou GDN-Gestion des droits numériques). Il existe sous plusieurs formats. Optimisé pour diffuser du son sur internet, en particulier en streaming.

[57] Ogg-Speex (http://www.speex.org/) et Ogg-Vorbis (http://www.vorbis.com/) sont des formats ouverts développé par la fondation Xiph.org (http://xiph.org/). Ogg Vorbis – fichier audio au format Ogg contenant des données audio compressées en Vorbis - est un codec plus performant que MP3.Ogg-Speex, exclusivement conçu pour compresser la voix, est le meilleur compromis entre qualité et bande passante pour cet usage.

fidélité et stéréophonie)[58]. Dans certains cas, ils peuvent avoir été compressés « lossless » en FLAC[59]. Si la source est en MP3, il est préférable de conserver le document sous ce format ; toutefois, s'il est nécessaire de le numériser, le choix de WAVE devrait être fait.

L'Association internationale des archives sonores et audiovisuelles-IASA recommande que les documents sonores soient stockés sous forme de fichiers aux formats WAV ou BWF[60]. « *Les convertisseurs A/N d'échantillonnage 192 kHz et de résolution d'amplitude 24 bits sont les standards actuels. En ce qui concerne le transfert des signaux analogiques, l'IASA recommande une résolution numérique minimale de taux d'échantillonnage à 48kHz et une longueur de mot de 24 bits. Pour les institutions en charge de documents patrimoniaux, une résolution de 96 kHz/24 bits s'est largement répandue. Les composantes sonores non souhaitées étant transférées dans de telles conditions, l'élimination des artefacts au moyen de traitements numériques du signal s'effectuera plus facilement à partir de copies ainsi réalisées. Les enregistrements de parole, par le caractère transitoire des consonnes qu'ils comportent, doivent être traités comme les enregistrements musicaux* »[61]. Un son stéréo de 24 bits à un taux d'échantillonnage de 48/96 KHz devrait être utilisé pour des copies de référence. Ce taux est recommandé par la Société d'ingénierie du son-AES et par l'IASA.

Exploitation : formats des documents d'accès

Comme pour la vidéo, il convient de tenir compte de l'environnement technique des utilisateurs (largeur de bande passante).

Les documents d'accès seront codés conformément à la norme OGG avec un débit binaire recommandé supérieur à 48 kbps. En parallèle ou alternativement, les fichiers peuvent être codés en MP4 (>48 kbps) ou en MP3 (>96 kbps).

Le son destiné à être téléchargé sur internet devrait être publié en MP3 sous forme compressée ou en WAVE sous forme non compressée.

Les fichiers audio qui sont chargés dans un lecteur incorporé au navigateur ne doivent pas démarrer automatiquement et le lecteur doit être doté de commandes de démarrage et d'arrêt de l'audio. Tous les liens à des grands fichiers audio (+ de 50Ko) doivent être accompagnés d'une étiquette indiquant la durée et la taille du fichier. Un lien au codec utilisé pour produire les fichiers audio doit être fourni aux utilisateurs qui ont besoin de le télécharger et de l'installer. De même, si des fichiers audio sur internet sont préparés pour transmission dans un environnement à large bande, une version à bande étroite doit être fournie aux utilisateurs[62]. Tous les fichiers audio devraient être accompagnés d'audio-scripts ou de sommaires en mode texte (X)HTML.

---

[58] "Référentiel d'interopérabilité…", *op.cit.*

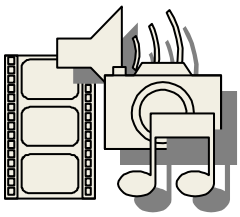
[59] FLAC-Free Lossless Audio Compression est un codec de compression sans perte. Il est disponible sur presque toutes les plateformes (http://www.flac.sourceforge.net).

[60] http://en.wikipedia.org/wiki/Broadcast_Wave_Format.

[61] "*Guidelines on the production and preservation of digital audio objects*", "*The Safeguarding of the audio heritage : ethics, principles and preservation strategy*", IASA Technical Committee, http://www.iasa-web.org.

[62] Culture canadienne en ligne, *op.cit.*, p. 21.

LES SUPPORTS DE STOCKAGE



COPIES

- Faire plusieurs copies de sauvegarde des fichiers de données, des logiciels et des systèmes d'exploitation (si possible de manière séparée)
- Faire plusieurs copies de sauvegarde sous des formats différents et/ou des marques différentes de supports de stockage
- Conserver au moins une copie ailleurs que dans l'institution

ENVIRONNEMENT

- Adopter de bonnes pratiques et conditions d'entreposage
- Mettre en place des routines d'entretien et de nettoyage des supports

TECHNOLOGIES

- Faire des mises à niveau successives
- Effectuer au bon moment des transitions ou les acquisitions essentielles
- Vérification : faire régulièrement des tests de lisibilité et d'intégrité des données
- Assurer le suivi des corrections d'erreur (« rafraîchissement ») et remplacer le support avant que les erreurs ne deviennent impossibles à corriger (« migration »)

COUTS

- Budgéter le coût des supports de stockage et de leur maintenance

LES ENJEUX DE LA CONSERVATION À LONG TERME

La conservation d'objets numériques regroupe une vaste gamme d'activités qui visent à allonger la vie utile des fichiers informatiques et à les protéger contre les défaillances des supports, la perte physique et l'obsolescence. Il convient en effet d'assurer une possibilité de restitution et d'intelligibilité des contenus, ce qui signifie à la fois conserver le contenu, mais aussi sa forme, son style, son apparence et les fonctions sous-jacentes. Tout support de stockage est, par définition, dépendant d'une combinaison de hardware et de software. L'accessibilité à l'information ainsi stockée est très vulnérable eu égard à l'environnement technologique en rapide évolution.

L'obsolescence rapide du matériel informatique est une constante depuis son origine. Les changements technologiques – qui se sont accélérés – ont accru la vitesse des processeurs, la densité des puces de mémoire, la capacité des dispositifs de stockage, la vitesse de traitement vidéo et le débit de traitement de données. D'autres facteurs interviennent également dans la détérioration des supports : conditions d'entreposage inadéquates, sur-utilisation, erreurs humaines (manipulations inappropriées), désastres naturels (incendie, inondation), …

Éviter l'obsolescence ou contrer ses conséquences implique d'assurer un suivi des développements technologiques (mises à niveau successives, effectuer au bon moment des transitions ou les acquisitions essentielles) et de budgéter leurs coûts (qui ne doivent pas être excessifs par rapport à la valeur des actifs à protéger). Il faut aussi à intervalles réguliers évaluer les exigences de sécurité à la lumière de nouvelles menaces, des nouveaux règlements, des changements technologiques et de l'évolution des besoins d'archivage. Un équilibre est à trouver entre sécurité et facilité d'accès.

La durée de vie des supports de stockage fait partie des arguments de vente des producteurs de supports et des critères de choix des consommateurs. Aucune étude indépendante des producteurs ne permet cependant à ce jour d'avoir une idée précise de la pérennité des supports matériels tandis que les durées de vie annoncées sont des moyennes statistiques qui ne tiennent aucunement compte d'autres facteurs, humains et techniques. Au lieu de se fier à leurs déclarations, il vaut mieux adopter une attitude prudente et proactive.

Le « Didacticiel sur la conservation d'objets numériques » de la Bibliothèque Cornell fait les recommandations suivantes :

- *« adopter de bonnes pratiques d'entreposage*
- *acheter des supports de qualité supérieure*

- *prendre en note le nom du fabricant et les numéros de lot des supports afin de permettre un suivi des performances et de la qualité*
- *ne pas acheter en trop grande quantité*
- *ne pas oublier que, dans certains cas (en particulier les disques optiques et magnéto-optiques), un support vierge a une durée de vie moindre qu'un support contenant des données*
- *acheter des supports compatibles avec la vitesse et la capacité des lecteurs dans lesquels ils seront utilisés*

Tous les supports de stockage doivent faire périodiquement l'objet de tests d'intégrité des données. Prévoir au minimum les procédures suivantes :

- *confirmer la fidélité de tout support immédiatement après l'enregistrement de données*
- *par la suite et à intervalles réguliers, lire en entier des supports pris au hasard (selon le fabricant et le code du lot) ainsi que des fichiers de plusieurs supports*
- *déterminer les lots, les fabricants ou les conditions d'entreposage susceptibles de poser problèmes et faire dans ces cas des tests plus poussés*
- *tester les supports vierges (cela peut être long et coûteux)*
- *faire un suivi des corrections d'erreur et remplacer le support avant que les erreurs ne deviennent impossibles à corriger »* [63].

Par prudence, il convient de faire plusieurs copies de sauvegarde de l'original numérique. **Doivent** être sauvegardés les fichiers de données mais aussi les logiciels d'application et les systèmes d'exploitation. Une copie au moins **doit** être entreposée dans un lieu différent de celui de l'institution ou du fonds concerné. Si le stockage des données est effectué dans un site externe par une entreprise spécialisée, il faut que l'entreposage soit accompagné d'un service de gestion de données. Dans le cas de copies multiples et pour réduire la dépendance à l'égard d'une technologie, il peut être intéressant de choisir des supports différents pour chacune des copies, basés sur des technologies différentes (par exemple, magnétique et optique). Si le même type de support est utilisé pour les copies, des marques différentes **devraient** être choisies afin de minimiser les risques de perte de données.

Un point à considérer dans le choix d'une stratégie de sauvegarde est la possibilité de perte totale (équipement et données) de l'installation principale dans un sinistre. Un plan « bis » **doit** exister. Il ne doit pas compter sur les capacités – variables – des entreprises spécialisées dans la récupération des données.

Des référentiels de conservation d'objets numériques ont été développés et des groupes de travail ont procédé à leur homologation afin de permettre aux institutions de procéder à une évaluation interne[64].

[63] http://www.library.cornell.edu/iris/tutorial/dpm-french/oldmedia/threats.html .
[64] http://www.oclc.org/programs/ourwork/past/repositorycert.htm .

LES DIFFÉRENTS SUPPORTS DE STOCKAGE

Les ressources générées au cours du projet de numérisation sont, en général, stockées sur les disques durs d'un ou plusieurs serveurs de fichiers, ainsi que sur des supports de stockage portables (bande magnétique et supports optiques tels que CD-R et DVD).

Les supports de stockage numérique ont des spécificités logicielles et matérielles différentes. Leurs caractéristiques de stockage et de gestion diffèrent également. De nouveaux supports sont testés régulièrement. Il en va ainsi, par exemple, des systèmes de stockage holographique.

Les supports numériques de stockage couramment utilisés se répartissent en trois catégories :

- les disques magnétiques fixes, magnétiques amovibles, magnéto-optiques (à lecture unique, à lecture-écriture), optiques (à lecture seule, à écriture unique, inscriptibles, à lecture-écriture : CD, DVD, Blue-Ray,…)
- les bandes magnétiques : AIT/SAIT (Advanced Intelligent Tape et Super Advanced Intelligent Tape), LTO (Linear Tape Open), SDLT, …
- les semi-conducteurs ou mémoires nomades : cartes mémoire CompactFlash, modules de mémoire MemoryStick, modules de mémoire SmartMedia (mémoire d'appareil photographique numérique) ; clés ou modules de mémoire USB ; lecteurs Flash.

Pour le stockage des photographies, Memoriav considère que les supports magnétiques les plus appropriés sont les bandes magnétiques en cassette et les DVD.

	Avantages	Inconvénients
Bandes magnétiques	-grande capacité de stockage -stabilité et sécurité satisfaisantes -coût très avantageux	-dommages mécaniques possibles -pas d'accès direct, seule une lecture séquentielle est possible -recherche de fichier lente -entretien nécessaire toutes les quelques années pour assurer la préservation des données -lecteurs/enregistreurs relativement chers automatisation très coûteuse
DVD	-faible capacité de stockage -accès direct aux données -bon taux de transfert -faible coût des graveurs -automatisation d'un coût raisonnable	-trop grande multitude de stockage -dommages possibles (absence de boîtier de protection) -lenteur de la gravure

Source : http://fr.memoriav.ch/dokument/Empfehlungen/recommandations_photo_fr.pdf

Memoriav a également établi un état des lieux, arrêté en 2006, des formats vidéo[65] en indiquant notamment le degré d'obsolescence et de risques pour la bande, la méthode d'enregistrement et le domaine d'utilisation.

Pour le son, le Comité technique de l'IASA précise qu' « *aucun système d'enregistrement numérique dédié spécifiquement au son n'a montré une stabilité industrielle probante, ils demeurent à part parmi les documents d'archives. A l'exception du CD audio, du DVD audio et du MiniDisc, tous les formats numériques spécifiques audio sont devenus obsolètes après une courte période de commercialisation, laissant de nombreux supports en bon état mais dépourvus d'appareillage permettant d'accéder aux sons. On a assisté, ces dernières années, à une évolution marquée de formats audio tels que R-DAT et CD-R (audio) vers des formats de stockage de données, c'est-à-dire des formats de fichiers dans un environnement informatique. Bien qu'en principe les formats de fichiers, les systèmes d'exploitation et les supports de stockage informatiques soient aussi menacés d'obsolescence, de tels environnements professionnels rendent la gestion plus aisée que les formats audionumériques grand public* »[66].

Pour l'IASA, l'utilisation de disques CD et DVD enregistrables doit être considérée comme étant potentiellement dangereux pour la survie des enregistrements (cf IASA-TC 04, 6.6). En raison de l'absence de problèmes vis-à-vis des normes et compatibilités, les CD-R et DVD-R peuvent être reconnus comme des supports fiables, mais seulement si des tests sont effectués. Comme ceci prend du temps et implique des investissements significatifs, l'IASA ne les recommande pas.

Le consortium PrestoSpace – soutenu par le 6ème programme européen (FP6-2002-IST-1) - a produit un guide très complet qui propose des solutions techniques et des systèmes intégrés pour la préservation des collections audiovisuelles[67].

LES CRITÈRES DE CHOIX DES SUPPORTS DE STOCKAGE

De nombreux facteurs interviennent dans le choix de supports numériques de stockage à long terme.

Les Archives nationales britanniques[68] proposent une méthode de pondération de 6 critères pour 6 supports de stockage :

---

[65] http://fr.memoriav.ch/dokument/Empfehlungen/empfehlungen_video_fr.pdf

[66] Comité technique de l'IASA, « *Sauvegarde du patrimoine sonore : Ethique, principes et stratégies de conservation*", IASA-TC 03, décembre 2005.

[67] http://prestospace.org/ et http://prestospace-sam.ssl.co.uk/ .

[68] The National Archives, « *Digital Preservation Guidance Note 2 : Selecting Storage Media for Long-Term preservation* », août 2008 (http://www.nationalarchives.gov.uk/documents/selecting-storage-media.pdf). Ce site présente aussi les conditions idoines de conservation selon les supports.

- **longévité** : avoir un horizon de 10 ans ; viser une longévité plus grande n'est pas nécessaire au vu de l'obsolescence des techniques ;
- **capacité** : ajuster la capacité de stockage au volume des données et à la dimension physique des équipements de stockage ;
- **viabilité** : les médias et les « drives » devraient supporter des modes de détection d'erreurs convenables. Prévoir de tester l'intégrité du média après écriture est un plus. Prévoir des mécanismes de protection pour éviter « d'écraser » accidentellement des données et maintenir l'intégrité de celles-ci ;
- **obsolescence** : préférer des technologies matures, largement disponibles et utilisées sur le marché aux technologies « à la pointe ». Choisir des standards ouverts pour les médias et les « drives » est préférable ;
- **coût** : deux éléments sont à prendre en considération, à savoir le coût du média lui-même et le coût total. L'élément pertinent de comparaison pour le média est le coût par Gigabyte. Le coût total doit inclure les coûts d'achat et de maintenance des matériels et logiciels et celui des équipements de stockage. Si le coût du stockage en lui-même diminue, celui du management du stockage augmente fortement ;
- **vulnérabilité** : le média doit être peu vulnérable aux dommages physiques et supporter des environnements divers sans perte de données.

Chaque support reçoit pour chaque critère une note de 1 (ne répond pas au critère) à 3 (répond tout à fait au critère). Un support de stockage doit avoir une note totale d'au moins 10 pour être considéré comme une bonne solution de stockage.

Supports de stockage numérique à long-terme

Support	CD-R	DVD-R	Hard Disk	Flash Memory Stick and Card	Linear Tape Open (LTO)
Longévité	3	3	2	1	3
Capacité	1	3	3	2	3
Viabilité	2	2	2	1	3
Obsolescence	1	2	2	2	2
Coût	3	3	1	3	3
Vulnérabilité	1	1	3	1	3
Total	11	14	13	10	17

Source : The National Archives, *Digital Preservation Guidance Note 2 : Selecting storage media for long-term preservation*, p.6.

Le consortium PrestoSpace a mis au point des instruments pratiques, comme un outil d'analyse de stockage, un module d'estimation des coûts ou encore un « Preservation calculator »[69].

[69] http://prestospace-sam.ssl.co.uk/hosted/d12.2/calc4.php; http://prestospace-sam.ssl.co.uk/hosted/d13.2/newcalc.php; http://prestospace-sam.ssl.co.uk/hosted/d14.2/newcalc.php.

NORMES DE DOCUMENTATION - MÉTADONNÉES

Métadonnées,
maillon essentiel pour l'interopérabilité des fonds/collections

Choix de la standardisation XML pour l'encodage des métadonnées

Adopter le schéma « Dublin Core »
utiliser un identifiant unique et pérenne (URL réputée persistante)

Adopter les métadonnées sectorielles de standards internationaux

Adopter des ressources terminologiques et des ontologies partagées
Compatibilité avec les portails européens Michael et Europeana

LES MÉTADONNÉES

Que sont les métadonnées ?

Les métadonnées sont « des données sur les données », c'est un ensemble structuré d'informations décrivant une « « ressource quelconque (document, objet, fonds ou collection). Les ressources décrites par les métadonnées ne sont pas nécessairement sous forme digitale : un catalogue de bibliothèque ou de musée contient aussi des métadonnées décrivant les ressources que sont les livres de la bibliothèque ou les objets du musée.

Les métadonnées peuvent être utilisées à des fins multiples. Parmi les opérations qu'il est possible de réaliser grâce aux métadonnées, figurent la description et la recherche de ressources, la gestion de collections de ressources (y compris la gestion des droits), la diffusion et l'utilisation des ressources, l'échange d'informations entre bases de données, la migration des données vers de nouveaux systèmes et la préservation à long terme des ressources. L'utilisation de métadonnées permet aussi de favoriser la cohérence des données sur les ressources, d'assurer que l'information importante est bien enregistrée ou d'accroître la sécurité des collections.

Les métadonnées sont particulièrement importantes pour les ressources visuelles : les utilisateurs dépendent en effet des informations ajoutées aux images et vidéos pour effectuer des recherches pertinentes. Elles sont aussi un élément essentiel de l'architecture web.

Les métadonnées peuvent être contenues, encapsulées, dans le document ou la ressource même (elles correspondent à des marqueurs que l'on introduit dans les fichiers ou dans les langages de programmation) ou externes à la ressource mais fournies en même temps. Pour les ressources non digitales, les métadonnées sont externes aux ressources. Dans la plupart des systèmes informatisés, les métadonnées sont stockées dans une base de données spécifique. Si la ressource est elle-même sous forme digitale (une image JPEG par exemple) et qu'elle est utilisée en dehors de la base de données qui la référence, les métadonnées associées sont perdues sauf si elles sont exportées séparément et à nouveau associées à la ressource.

Les métadonnées et l'interopérabilité

L'ensemble des données structurées décrivant les ressources physiques ou numériques - que sont les métadonnées - sont un maillon essentiel pour l'interopérabilité des collections et de sa gestion. Il est aussi important d'assurer l'interopérabilité des métadonnées elles-mêmes.

L'interopérabilité se réalise à trois niveaux techniques complémentaires[70] :

- « *une description des ressources avec des sémantiques communes issues de différents jeux de métadonnées standardisés ;*
- *un contexte générique d'implémentation de ces descriptions dans des langages structurés, interprétables par les machines ;*
- *des protocoles informatiques d'échange de ces données normalisées* ».

Le tableau suivant repositionne en ces termes différents standards bien connus.

	Standards traditionnels	Standards récents
Jeux de métadonnées	MARC	Dublin Core MARC-XML, MODS EAD LOM…
Cadre générique d'implémentation	ISO 2709 ISAD(G)	XML RDF Espaces de nom URL
Protocoles	WAIS FTP Z39.50	HTTP OAI-PMH SRU/SRW

Source : Catherine Morel-Pair, *op.cit.*, p. 4.

«*Les métadonnées sont toujours implémentées dans un langage structuré, (X)HTML, XML, RDF, et échangées par des protocoles. Selon l'usage attendu, les éléments sont présents soit dans la ressource elle-même (documents HTML, métadonnées natives des images …), soit dans un fichier associé ; un « enregistrement » (« record ») correspond à la description d'une ressource dans un format particulier.*

Un quatrième niveau d'interopérabilité, plus organisationnel que technique, implique des pratiques communes dans l'utilisation des éléments et dans les valeurs de ces éléments (codes, vocabulaires, données d'autorités standardisés, ontologies). Dans ce but, de plus en plus de producteurs de jeux de métadonnées éditent des documents, registres de métadonnées décrivant chaque élément (caractéristiques, liens avec d'autres éléments, usage, valeurs attendues, parfois « mapping » avec d'autre jeux, …) et des « guidelines » ou recommandations de bonnes pratiques. Pour les valeurs, il convient dans tous les cas d'utiliser au maximum des outils reconnus et accessibles en ligne, codes ISO ou W3C, vocabulaires, thesaurus, classifications, ontologies et listes d'autorité existantes ou à venir »[71].

Les métadonnées et la standardisation

L'interopérabilité implique aussi un certain niveau de standardisation et de normalisation.

[70] Catherine Morel-Pair, « *Panorama : des métadonnées pour les ressources électroniques* », INIST-CNRS, 2005 (http://hal.archives-ouvertes.fr/docs/00/04/04/73/PDF/Metas_panorama_CMO.pdf).
[71] Idem, p.4.

Pour ce qui concerne la description des ressources, l'établissement de standards de métadonnées a d'abord fait l'objet d'initiatives aux Etats-Unis : le standard « Dublin Core » a ainsi été défini dès 1995 par le DCMI (voir ci-dessous). En Europe, c'est essentiellement l'Ukoln (http://www.ukoln.ac.uk) qui a développé des projets soutenus par l'Union européenne.

Toutefois, il n'existe pas de standard universel et unique pour rencontrer toutes les fonctions[72], ou même chacune d'entre elles, qui peuvent être exercées grâce aux métadonnées. En pratique, les schémas de métadonnées ont souvent de multiples fonctions. Des applications sont développées par de très nombreuses institutions pour répondre à des besoins particuliers, ce qui ne dispense pas de devoir créer, dans certains cas, un profil d'application propre.

Implémenter des métadonnées doit s'inscrire dans la démarche de gestion du projet de numérisation et correspondre à ses objectifs. Il faut éviter de se laisser convaincre par les effets de mode[73].

---

[72] Les **métadonnées descriptives** permettent de décrire, d'identifier et de localiser la ressource. Le schéma le plus simple et le plus courant est l'ensemble des éléments du « Dublin Core ». Il existe des normes plus sophistiquées telles que MODS-Medadata Object Description Schema, utilisé notamment par la Bibliothèque du Congrès et particulièrement destiné au secteur des bibliothèques (il est basé sur un sous-ensemble du standard MARC, http://www.loc.gov/standards/mods/). Les **métadonnées techniques** décrivent les caractéristiques techniques d'une ressource numérique (équipement de numérisation utilisé et ses paramètres comme les formats ou les types de compression). Il existe un standard très largement utilisé pour les images fixes et pour les fichiers texte. Pour les images fixes, il s'agit de MIX (Medatada for Images in XML), conçu par le NISO-National Information Standards Organisation (organisation agréée par l'ANSI-American National Standards Institute qui développe des standards spécifiquement à destination des bibliothèques, des services d'information et du secteur de l'édition (http://www.niso.org/publications/tr/; http://www.loc.gov/standards/mix/). Pour les fichiers texte, il s'agit de TEI Headers (TEI-Text Encoding Initiative, http://www.tei-c.org) ou un schéma plus simple comme le Schema for Technical Metadata for Text développé par l'Université de New-York (http://dlib.nyu.edu/METS/textmd.htm). Pour les fichiers audio et vidéo, aucun standard ne semble prévaloir à ce jour. La Bibliothèque du Congrès a produit les schémas AUDIOMD-Audio Technical Metadata Extension Schema (http://lcweb2.loc.gov/mets/Schemas/AMD.xsd) et VIDEOMD-Video Technical Metadata Extension Schema (http://lcweb2.loc.gov/mets/Schemas/VMD.xsd), tandis que le service public de télévision américain a produit le schéma usuel PBCORE (http://www.pbcore.org/). Les **métadonnées administratives** permettent de gérer un objet numérique et de fournir des informations sur sa création et sur toute contrainte liée à son utilisation. Elles peuvent inclure :

- les **métadonnées sur la source**, qui décrivent l'objet à partir duquel la ressource numérique a été produite. Elles ne sont pas nécessaires si l'objet d'origine est numérique. En général, les standards MODS ou DC sont utilisés, accompagnés si nécessaire de schémas plus spécialisés ;
- les **métadonnées de provenance numérique** décrivent l'historique des opérations effectuées sur un objet numérique depuis sa création ou sa saisie. Le schéma « events » de PREMIS remplit cette fonction ;
- les **métadonnées de gestion de droits** : décrivent les droits d'auteur, les restrictions d'utilisation et les accords de licence. La gestion des droits (déclaration de propriété et contrôle d'accès) peut être gérée par le schéma de droits METS, ou par des schémas plus complexes, soutenus par le secteur commercial tels que XrML-eXtensible Rights Markup Language développé à l'origine par Microsoft et Xerox et qui est inclus dans le standard MPEG-21 (http://www.xrml.org/) ou encore le modèle ouvert ODRL-Open DigitalRights Language (http://odrl.net).

[73] Le NISO a défini six principes pour qualifier les bonnes métadonnées:"*Principle 1: Good metadata conforms to community standards in a way that is appropriate to the materials in the collection, users of the collection, and current and*

Les métadonnées dont question ici concernent le niveau de la ressource. Il existe des métadonnées dites structurelles qui énumèrent les éléments d'information à enregistrer pour documenter correctement une collection[74] et qui fournissent les informations permettant de relier les différents composants d'une ressource ou objet complexe.

LE LANGAGE DE BALISAGE XML

Le JISC considère que la standardisation XML des métadonnées descriptives, techniques et administratives des contenus numériques est un élément fondamental de l'interopérabilité des contenus et des applications associées[75].

A l'origine, XML, eXtensible Markup Language, était un langage permettant de « marquer » des textes électroniques par le biais de balises ayant une signification sémantique ; il devient de plus en plus un mécanisme à part entière d'encodage des métadonnées.

L'intérêt du XML est multiple :

- il s'agit d'une norme ouverte enregistrée à l'ISO, complètement indépendante de toute application informatique ;
- un document XML est « lisible » par un être humain sans qu'il nécessite une interface particulière ;
- la simplicité de sa syntaxe et sa capacité à encoder des structures hiérarchiques complexes doivent être mise en évidence. XML fournit ainsi un moyen élégant de représenter des ensembles complexes de métadonnées par le biais de *schémas XML* qui peuvent être échangés et interprétés par différents systèmes informatiques ;
- XML permet la définition d'« espaces de nommage » (« *namespace* ») : un espace de nommage est une collection de noms, identifiées par une référence d'URI (Uniform Ressource Identifier), qui sont utilisés dans les documents XML comme types d'éléments ou noms d'attributs. Un espace de nommage peut dès lors donner la définition formelle des éléments d'un ensemble de métadonnées, de manière à

---

potential future uses of the collection. Metadata Principle 2: Good metadata supports interoperability. Metadata Principle 3: Good metadata uses authority control and content standards to describe objects and collocate related objects. Principle 4: Good metadata includes a clear statement of the conditions and terms of use for the digital object. Principle 5: Good metadata supports the long-term curation and preservation of objects in collections. Principle 6: Good metadata records are objects themselves and therefore should have the qualities of good objects, including authority, authenticity, archivability, persistence, and unique identification" (http://framework.niso.org/node/5).

[74] Minerva a défini un schéma de description de niveau « collection » (pour les standards : http://www.minervaeurope.org/intranet/reports/D3_2.pdf, pour les recommandations http://www.minervaeurope.org/intranet/reports/D3_1.pdf ainsi que le nouveau profil de fonction de Dublin Core : http://www.dublincore.org/groups/collections).

[75] Voir Richard Gartner, "*Metadata for digital libraries : state of the art and future directions*", JISC Technology & Standards Watch, avril 2008 (http://www.jisc.ac.uk/techwatch).

éviter toute ambigüité dans son interprétation. On peut ainsi constituer ses propres profils d'applications en combinant des éléments provenant de différents schémas de métadonnées, ou encore spécialiser des structures générales en y accrochant des schémas spécifiques (METS utilise ce mécanisme).

Si les métadonnées sont déjà encodées dans d'autres formats, l'importation et l'exportation des métadonnées (et des schémas correspondants) en XML **doit** être envisagée en tenant compte du coût d'une conversion vers des bases de données XML-natives.

LE DUBLIN CORE METADATA INITIATIVE

La multiplication des besoins ainsi que la diversité des structures et nomenclatures de métadonnées ont conduit à la recherche d'un standard minimal.

Le NCSA[76] et l'OCLC [77], rassemblant des professionnels provenant de diverses disciplines, ont défini en 1995 à Dublin (Ohio, USA) un ensemble de métadonnées communes à différents secteurs d'activités, applicables à presque tous les formats de fichiers, appelé le « Dublin Core »[78]. Ce référentiel commun, qui fait l'objet d'un large consensus et d'une large utilisation, a été proposé pour faciliter la recherche de ressources peu complexes. Il ne prétend pas répondre à toutes les fonctions et besoins. « *Il est insuffisant pour la description des collections, la gestion administrative et technique, limité pour la description d'objets physiques et peu contraignant en matière d'architecture* »[79]. C'est pourquoi il doit souvent être complété par des champs additionnels ou des schémas complémentaires, développés au niveau sectoriel ou selon les fonctions à représenter (voir ci-dessous).

Le schéma « Dublin Core », avec un seul niveau d'arborescence, comprend 15 éléments, répartis autour de 3 domaines qui permettent d'identifier et de décrire les ressources documentaires :

- contenu : titre, description, sujet, source, couverture, type, relation,
- propriété intellectuelle : créateur, contributeur, éditeur, droits (droits d'auteur,…)
- version : date, format, identifiant, langue.

Un modèle, appelé Qualified Dublin Core[80], a été mis au point pour préciser la signification des éléments simples de la norme Dublin Core à l'aide de qualificatifs ou de schémas de codage.

---

[76] NCSA-National Center for Super Computing Applications (http://www.ncsa.uiuc.edu/).
[77] OCLC-Online Computer Library Center (http://www.oclc.org/fr/fr/default.htm).
[78] Cette norme a été approuvée comme norme ANSI (Z39.85-2001) et comme norme ISO (15836 :2003) (http://dublincore.org et en français http://www.bibl.ulaval.ca/dublincore/usageguide-20000716fr.htm).
[79] Catherine Morel-Pair, *op.cit.*, p.10.
[80] http://dublincore.org/documents/2000/07/11/dcmes-qualifiers/.

Différentes syntaxes permettent d'implémenter les éléments du Dublin Core : XML (utilisé en particulier dans l'enregistrement OAI-PMH et dans des projets d'échange et de mutualisation de données), HTML et XHTML (non utilisé par les moteurs actuels) et RDF.

Dans un contexte d'interopérabilité des ressources, la notion d'URI, identifiant unique et pérenne définissant chacune d'elles, est fondamentale. Ainsi, l'OCLC préconise d'utiliser une URL réputée persistante[81] pour faire référence à un élément du Dublin Core.

Les 15 éléments du Dublin Core

Elément	Nom	Identifiant	Définition	Commentaires
Titre	Titre	Title	Le nom donné à la ressource	le nom par lequel la ressource est officiellement connue
Créateur	Créateur	creator	L'entité principalement res ponsable de la création du contenu de la ressource	une personne, une organisation ou un service.
Sujet	sujet et mots-clefs	subject	Le sujet du contenu de la ressource	décrit par un ensemble de mots-clefs ou de phrases ou un code de classification qui précise le sujet de la ressource. L'utilisation de vocabulaires contrôlés et de schémas formels de classification est encouragée.
Description	description	description	Une description du contenu de la ressource	peut contenir, mais ce n'est pas limitatif, un résumé, une table des matières, une référence à une représentation graphique du contenu ou un texte libre sur le contenu
Editeur	éditeur	publisher	L'entité responsable de la diffusion de la ressource, dans sa forme actuelle	une personne, une organisation, ou un service. Ex : le nom d'une maison d'édition
Contributeur	contributeur	contributor	Une entité qui a contribué à la création du contenu de la ressource -	une personne, une organisation ou un service.
Date	date	date	Une date associée avec un événement dans le cycle de vie de la ressource	Ex : une date associée à la création ou à la publication d'une ressource. Il est fortement recommandé d'encoder la valeur de la date en utilisant le format défini par l'ISO 8601 sous la forme AAAA-MM-JJ
Type	type de la ressource	Type	La nature ou le genre du contenu de la ressource	inclut des termes décrivant des catégories, fonctions ou genres généraux pour le contenu ou des niveaux d'agrégation. Il est recommandé de choisir la valeur du type dans une liste de vocabulaire contrôlé (par exemple, la liste provisoire de Types du Dublin Core). Pour décrire la matérialisation physique ou digitale de la ressource, il faut utiliser l'élément Format
Format	format	format	La matérialisation physique ou digitale de la ressource	peut inclure le media ou les dimensions de la ressource. Peut être utilisé pour préciser le logiciel, le matériel ou tout autre équipement nécessaire pour afficher ou faire fonctionner la ressource. Il est recommandé de choisir la valeur du format dans une liste de vocabulaire contrôlé (ex. : la liste des types de media définis sur Internet).

[81] Une URL-Uniform Resource Locator ou « adresse web » est une chaîne de caractères, codée en ASCII, utilisée pour identifier les pages et les sites web. Une PURL-Persistent Uniform Resource Locator est une URL qui a la caractéristique de ne pas pointer directement vers une adresse web mais vers un service intermédiaire qui vérifie l'adresse et la renvoie vers l'utilisateur mise à jour si nécessaire (http://purl.org).

Identifiant	identifiant de la ressource	identifier	Une référence non ambiguë à la ressource dans un contexte donné	Il est recommandé d'identifier la ressource par une chaîne de caractère ou un nombre conforme à un système formel d'identification (URI-Uniform Resource Identifier, DOI-Digital Object Identifier et ISBN-International Standard Book Number)
Source	source	source	Une référence à une ressource à partir de laquelle la ressource actuelle a été dérivée	La ressource actuelle peut avoir été dérivée d'une autre ressource, en totalité ou en partie. Il est recommandé de référencer cette source par une chaîne de caractère ou un nombre conforme à un système formel d'identification
Langue	langue	language	La langue du contenu intellectuel de la ressource	Il est recommandé d'utiliser la RFC 1766 qui comprend un code de langage à deux caractères (venant du standard ISO 639), éventuellement suivi d'un code à deux lettres pour le pays (venant du standard ISO 3166 ou en français). Ex : 'fr' pour le français
Relation	relation	relation	Une référence à une autre ressource qui a un rapport avec cette ressource	Il est recommandé de référencer cette ressource par une chaîne de caractères ou un numéro conforme à un système formel d'identification.
Couverture	couverture	coverage	La portée ou la couverture spatio-temporelle de la ressource	inclut une position géographique (le nom d'un lieu ou ses coordonnées), une période de temps (un nom de période, une date ou un intervalle de temps) ou une juridiction (telle que le nom d'une entité administrative). Il est recommandé de choisir la valeur de Couverture dans un vocabulaire contrôlé (par exemple, un thésaurus de noms géographiques, comme [TGN]) et, quand cela est approprié, des noms de lieux ou de périodes plutôt que des identifiants numériques tels que des coordonnées ou des intervalles de date.
Droits	gestion des droits	rights	Information sur les droits sur et au sujet de la ressource	contiendra un état du droit à gérer une ressource, ou la référence à un service fournissant cette information.

LES MÉTADONNÉES SECTORIELLES

Les ressources sont en général partagées par différentes institutions. Des collectivités par « métier » collaborent depuis longtemps pour établir des modes de description et de documentation de leurs ressources qui soient compatibles entre elles.

Certains standards ont ainsi été développés par secteur d'activités : par exemple, SPECTRUM[82] et CDWA[83] pour les musées, ISAD(G)[84], ISAAR(CPF)[85], EAD[86] pour les

[82] SPECTRUM-Standard Procedures for Collections Recording Used in Museums est une norme ouverte, développée par MDA-Museum Documentation Association qui conseille et assiste les musées au Royaume-Uni pour ce qui est de la documentation de leurs collections et qui est reconnue à l'échelle internationale pour les échanges de données muséologiques (http://www.mda.org.uk/spectrum.htm).

[83] CDWA-Categories for the Description of Works of Art, dont l'éditeur est l'AITF-Art Information Task Force avec l'appui financier du J. Paul Getty Trust, décrivent le contenu d'une base de données dans le domaine des arts en proposant un cadre pour la description et l'accès à l'information sur des objets et des images (http://www.getty.edu/research/conducting_research/standards/cdwa/index.html).

centres d'archives, MARC[87] et ses dérivés pour la description des ouvrages dans les bibliothèques ou ISBD(S)[88] pour la description des publications en série, CIMI[89] pour la description des ressources muséographiques, RKMS[90] pour la description des ressources audio, LOM[91] pour la description des ressources liées à l'éducation, …

Il existe des normes qui portent sur la description ou la documentation générale des collections (par exemple, pour les musées la norme RSLP pour Research Support Libraries Program, fondée sur la norme Dublin Core), d'autres sur la description d'objets précis dans une collection (par exemple, pour les musées la norme CDWA, ou pour les objets archéologiques l'International Core Data Standard for Archaeological Objects du CIDOC[92]).

Avant d'opérer un choix parmi cette multitude de standards, il convient de s'assurer que les schémas de métadonnées choisis sont entièrement documentés.

RESSOURCES TERMINOLOGIQUES ET ONTOLOGIES

La transmission des informations contenues dans les enregistrements de métadonnées dépend aussi d'une compréhension partagée des termes, concepts et modèles utilisés.

Or, il n'existe pas de terminologie standard, d'application intersectorielle et multilingue.

---

[84] ISAD(G)-International Standard for Archival Description (General) dont l'éditeur est l'ICA, le Conseil international des archives (http://www.ica.org/fr/node/30001).

[85] ISAAR(CPF)-International Standard Archival Authority Record for Corporate Bodies, Persons and Families dont l'éditeur est l'ICA (http://www.ica.org/fr/node/30231).

[86] EAD-Encoded Archival Description dont la maintenance est assurée dans le Network Development and MARC Standards Office de la Bibliothèque du Congrès en collaboration avec la Society of American Archivists (http://www.archivists.org/saagroups/ead/; http://www.loc.gov/ead/; http://www.archivesdefrance.culture.gouv.fr/gerer/classement/normes-outils/ead/).

[87] MARC-Machine-Readable Cataloging, norme ISO 2709 développée par la Bibliothèque du Congrès, se présente sous forme d'une grille (succession de champs et de sous-champs de données). Certains champs sont définis de manière précise par l'IFLA. Existe plusieurs variantes : InterMarc utilisé par la Bibliothèque nationale de France ; MARC21 reconnu par l'IFLA comme format d'échange et dont la maintenance est assurée par la Bibliothèque du Congrès ; UNIMARC, initialement créé par l'IFLA comme format interface, devenu le format d'échanges de données en France (http://www.loc.gov/marc/ http://www.collectionscanada.gc.ca/marc/index-f.html; http://www.bnf.fr/pages/infopro/normes/no-acuni.htm).

[88] ISBD-International Standard Bibliographic description, norme développée par l'IFLA-International federation of Library Association (http://www.ifla.org/VI/3/nd1/isbdlist.htm).

[89] CIMI-Computer Interchange of Museum Information (http://www.cni.org/pub/CIMI/framework.html)

[90] RKMS-Recordkeeping Metadata Schema (http://www.sims.monash.edu.au/research/rcrg/research/spirt/)

[91] LOM-IEEE Learning Object Metadata (http://ltsc.ieee.org/wg12/).

[92] Comité international pour la documentation du Conseil international des musées (http://cidoc.mediahost.org/home(fr)(E1).xml).

Dès lors, **doivent** être utilisées des terminologies communes[93] ou des terminologies différentes pour lesquelles les relations entre les termes sont clairement définies. De même, il est d'une grande importance d'indiquer sans ambiguïté dans les enregistrements de métadonnées la terminologie choisie.

Les enregistrements des métadonnées **devraient** se faire en utilisant les terminologies compatibles avec le schéma de description de collection du portail multilingue européen Michael[94].

Les ontologies[95] jouent un rôle important dans les développements récents en matière de métadonnées descriptives « métiers ». Une ontologie est une définition formelle des termes et concepts d'un domaine d'activités, ainsi que de leurs relations. Une ontologie n'a pas pour but premier de fournir un modèle de métadonnées, mais les concepts et relations qu'elle définit peuvent servir à la définition d'un tel modèle.

Une ontologie est donc en quelque sorte le « corpus conceptuel » permettant de raisonner dans un domaine d'activités, un choix quant à la manière de décrire un domaine. C'est aussi sa description formelle. Les termes, concepts et relations qu'elle définit permettent de construire des jeux de métadonnées. En ce sens, elle peut être considérée comme une « colle sémantique » facilitant la correspondance entre différents jeux de métadonnées propres à un domaine d'activités, dans la mesure où l'interopérabilité entre deux jeux de métadonnées sera facilitée s'ils reposent sur la même ontologie sous-jacente[96].

Pour le domaine du patrimoine culturel, l'ontologie de référence est le CRM-Conceptual Reference Model[97]. Pour la description des ressources bibliographiques, les FRBR-Fonctional Requirements for Bibliographic Records peuvent être considérés comme

---

[93] Voir par exemple le thesaurus de l'Unesco (http://www2.ulcc.ac.uk/unesco /).

[94] Minerva, « *Inventories, discovery of digitised content & multilingual issues : Feasibility survey of the common platform*" (http://www.minervaeurope.org/intranet/reports/D3_2.pdf); et le portail multingue Michael (http://www.michael-culture.eu/documents/MICHAELDataModel_FR.pdf).

[95] Les ontologies peuvent être classées selon leur niveau de généralités. Sont distinguées ainsi :

- les ontologies de haut niveau (« Top-level Ontologies ») qui décrivent les concepts très généraux comme l'espace, le temps, la matière, les objets, les événements, les actions, etc., qui sont indépendants d'un problème ou d'un domaine d'application particulier ;
- les ontologies de domaine (« Domain Ontologies ») et les ontologies de tâches (« Task Ontologies ») qui décrivent, respectivement, le vocabulaire lié à un domaine générique (comme les musées ou la littérature) ou une tâche ou une activité générique (comme la vente), en spécialisant les concepts présentés dans les ontologies de hauts niveaux. Elles donnent une représentation formelle des concepts du domaine étudié ainsi que des différentes relations qui lient ces derniers ;
- les ontologies d'application (« Application ontologies ») qui décrivent des concepts dépendant à la fois d'un domaine et d'une tâche particulière.

Source : Nicola Guarino, "*Formal Ontology and Information Systems* », 1998 (http://www.loa-cnr.it/Papers/FOIS98.pdf).

[96] Voir l'article de Patrick Leboeuf, « *Le modèle CRM pour la documentation muséographique : s'attacher au sens pour ne pas être piégé par la forme* », http://cidoc.ics.forth.gr/docs/adbs_crm.pdf . Sa description des avantages du CRM est généralisable pour toutes les ontologies.

[97] http://cidoc.ics.forth.gr/docs/cidoc_crm_version_4.3_Nov08.pdf. Elle est gérée par le comité international pour la documentation des musées, le CIDOC-ICOM. C'est une norme ISO (ISO 21127 :2006).

l'ontologie de référence[98]. Le modèle FRBR sert également de référence pour le développement en cours de la seconde partie du standard CEN pour l'identification des œuvres cinématographiques (CWS-Cinematographic Works Standard)[99]. Un travail d'harmonisation de ces deux modèles a pour objectif d'exprimer le modèle FRBR avec les concepts, outils, mécanismes et conventions de notation fournis par le modèle CRM.

Les ontologies sont appelées à jouer un rôle majeur dans le développement du web sémantique, notamment par le développement d'outils de recherche basés sur des langages formels proches du langage naturel.

Les ontologies sont développées soit en RDF-Ressource Description Framework[100] (notamment avec OWL-Ontologies Writing Language[101] qui est le standard le plus utilisé), soit dans des langages et concepts proches de RDF (Topics Maps[102] ou SKOS-Simple Knowledge Organisation System for the Web[103]). Le langage RDF permet de construire le web sémantique.

[98] http://www.ifla.org/VII/s13/frbr/. Certains auteurs considèrent qu'il ne s'agit pas à proprement parler d'une ontologie dans la mesure où ce modèle n'est pas conçu selon un formalisme orienté objet. L'extension vers FRBRoo (pour object oriented) (http://cidoc.ics.forth.gr/frbr_inro.html) devrait remédier à ce problème.
[99] http://ec.europa.eu/avpolicy/docs/reg/cinema/june08/cen_en.pdf.
[100] Élaboré en 1997 à l'initiative du W3C par un ensemble de professionnels venant d'horizons différents, il repose sur un schéma XML (http://www.w3.org/TR/rdf-primer/). C'est une technique générale de description de ressources à l'aide de métadonnées, que l'on peut comparer aux mots-clés d'une fiche de catalogage.
[101] http://www.w3.org/2004/OWL/ .
[102] http://www.isotopicmaps.org/; http://www.topicmaps.org/xtm/1.0/.
[103] http://www.w3.org/2004/02/skos/ .

Stratégie de conservation

Adoption de l'OAIS comme modèle de référence :

- **terminologie unique et fiable**
- **inventaire des questions à se poser dans la mise en place du système de préservation**
- **description des composantes du système de préservation**

GESTION DE LA CONSERVATION

Un programme de conservation d'objets numériques existe dans un contexte organisationnel. Une archive numérique doit être organisée de telle manière qu'elle puisse conserver l'information sur une longue période. Or, cette exigence peut être contradictoire avec le cycle de vie - de plus en plus court - des produits informatiques (supports, formats, logiciels,…). Il est dès lors nécessaire d'établir un modèle d'organisation des contenus qui permette d'assurer la préservation à long terme des objets numériques.

Cette organisation de la représentation des contenus comprend des aspects de structuration et des aspects technologiques à appliquer de concert. Elle dépend de la nature et de l'ampleur des actifs numériques dont l'organisation en question doit prendre en charge. Une telle organisation implique des ressources et une responsabilité administrative qui doivent en assurer la viabilité et la sécurité.

La référence internationale en la matière est le Modèle de référence pour un système ouvert d'archivage d'information-OAIS.

LE MODÈLE OAIS

Le Modèle de référence pour un système ouvert d'archivage d'information-OAIS[104] a été produit par un groupe international de chercheurs et praticiens réunis par le Comité consultatif pour les systèmes de données spatiales de la NASA-National Aeronautics and Space Administration. OAIS est une norme ISO (ISO 14721 :2003).

Le modèle de référence de l'OAIS propose un modèle conceptuel de ce qui doit être géré pour obtenir la persistance. En effet, il ne suffit pas de conserver la ressource numérisée à archiver mais il faut également conserver toute l'information permettant de représenter la ressource sous une forme directement compréhensible par l'être humain, ce qui nécessite le recours à une médiation (processus qui permet d'interpréter le contenu en information de la ressource).

[104] Comité consultatif pour les systèmes de données spatiales, « *Recommandation pour les normes sur les systèmes de données spatiales* », « *Modèle de référence pour un Système ouvert d'archivage d'information (OAIS)* », janvier 2002 et mars 2005 pour la version en langue française (http://public.ccsds.org/publications/archive/650x0b1.pdf, http://public.ccsds.org/publications/archive/650x0b1(F).pdf).

Concepts

Le modèle de référence de l'OAIS comprend des définitions et décrit les relations entre les participants et les composantes d'un système d'archivage.

« *Ce que fait OAIS :*

- *il donne une terminologie fiable et unique pour manipuler tous les concepts liés à la préservation des données numériques*
- *il fait le tour de toutes les questions à se poser au moment de mettre en place un système de préservation*
- *il décrit les composantes d'un tel système au niveau de l'organisation interne et externe*

Ce qu'il ne fait pas :

- *il ne donne pas de formats, schémas, règles ou techniques pour préserver les documents numériques*
- *il ne décrit pas les applications informatiques et techniques à mettre en œuvre, ni logicielles, ni matérielles*
- *il ne donne pas de méthodologie concrète de réalisation d'un tel système* »[105].

L'organisation du modèle de référence de l'OAIS va du producteur (l'organisme qui fournit les objets à numériser) à l'utilisateur (les personnes ou les organismes qui ont accès aux objets numérisés), le « management » assurant la fonction de décideur chargé de l'organisation de l'archive. Le modèle OAIS cartographie le système d'archivage lui-même en 6 grands domaines fonctionnels articulés et interagissant entre eux (en grisé dans le schéma ci-dessous).

« *Le modèle OAIS repose sur l'idée que l'information constitue des paquets, et que ces paquets ne sont pas les mêmes suivant que l'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[106]) préparés par les producteurs à destination de l'archive*
- *les paquets d'archivage (AIP[107]) transformés par l'archive à partir du SIP sous une forme plus facile à conserver dans le temps*
- *les paquets de diffusion (DIP[108]) 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 »[109].

Le modèle de référence de l'OAIS prévoit également l'interopérabilité des archives et traite des niveaux techniques d'interaction entre archives OAIS.

---

[105] http://figoblog.org/document1089.php . Cette source présente le modèle de manière claire et simple.
[106] Submission Information Package.
[107] Archival Information Package.
[108] Dissemination Information Pakage.
[109] Figoblog, *op.cit.*

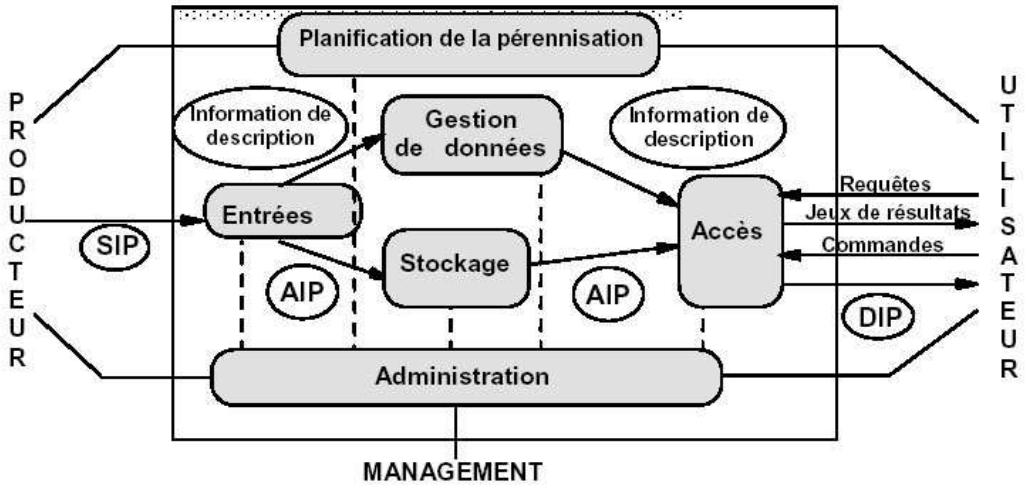


Schéma 4-1 : Entités fonctionnelles OAIS

Source : Comité consultatif…, *op.cit.*, p.4-1.

Migrations

Quelque soit le niveau de qualité avec lequel une institution maintient ses collections numérisées, il sera certainement nécessaire d'en migrer un grand nombre vers des supports différents et/ou vers des plates-formes logicielles et matérielles pour en préserver l'accès.

Le modèle OAIS identifie quatre types de migrations numériques, classés par ordre croissant de risque de perte d'information. Le choix dépend du type de problème rencontré :

- rafraîchissement de support : copie de bit à bit de l'information, dans laquelle un support généralement ancien est remplacé par un support identique généralement neuf (exemple : dans le cas d'un support qui se dégrade mais dont l'enregistrement qui est dessus ne pose pas de problème);
- duplication : recopie des objets archivés vers un nouveau type de support, sans changement de l'organisation logique du stockage ;
- ré-empaquetage : recopie des objets archivés vers un nouveau type de support, nécessitant une nouvelle organisation logique du stockage ;
- transformation ou émulation : il s'agit d'une réelle modification du contenu de l'information, portant notamment sur sa forme (on s'efforce de conserver ou reproduire les conditions d'accès au document : c'est le mode utilisé de préférence pour les documents dans des formats propriétaires). La transformation peut ou non être réversible[110].

[110] http://www.cines.fr/spip.php?article131&var_recherche=recherche .

Implémentation

Le modèle de référence de l'OAIS n'impose aucune modalité de mise en œuvre.
Différents schémas conformes au modèle de l'OAIS ont été élaborés, notamment par des fédérations professionnelles. Ces schémas concernent essentiellement les formats d'échanges entre les systèmes d'acquisition et de systèmes basés sur la norme OAIS. Il s'agit de suggestions de formats pour représenter les paquets de versement SIP-Submission Information Package. Deux exemples : METS et ZIP.

L'encapsulation via METS

METS-Metadata Encoding and Transmission Standard[111] est un schéma XML développé à l'initiative de la Digital Library Federation et maintenu par la Bibliothèque du Congrès. Il s'agit d'un standard non propriétaire, ouvert et extensible, qui autorise la création et la description intégrale d'objets numériques textuels ou graphiques et permet les échanges entre institutions patrimoniales.

La représentation du SIP via METS encapsule les données et normalise la représentation des métadonnées, de leurs liens avec les essences – qui peuvent y être encapsulées (binary) ou externalisées (links) – et les formats ainsi les structures liant plusieurs items, ce qui autorise le transfert d'objets complexes. Il permet la création de « documents METS » en XML contenant la description de la structure hiérarchique d'objets numériques constituant une ressource numérique, répertoriant les noms et la localisation des fichiers correspondant à ces objets (et assurant des liens entre eux), contenant toutes les métadonnées (de structure, administratives et descriptives) associées et associant des exécutables (tourne page,…). On dispose alors d'un « objet METS » échangeable qui comprend la ressource numérique et le « document METS ».

L'encapsulation via ZIP

Les études faites par diverses organisations internationales montrent que, dans la plupart des institutions culturelles, la documentation, la représentation et la présentation se fait sur base d'une structuration autour de chaque item d'une collection. La méthode de description de ceux-ci peut être choisie par l'institution en fonction de ses habitudes, normes et recommandations de la profession, … La définition du SIP se fait alors en concertation avec l'organisation en charge du système de dépôt pérenne.

Dans cette approche, le SIP est construit en séparant l'encapsulation de l'encapsulé. De telles encapsulations peuvent être réalisées par un processus d'emballage de dossiers et fichiers dont le plus connu est le ZIP (.zip est un format d'encapsulation ouvert, ayant même des variantes sécurisées [PK-ZIP])[112]. Les programmes d'encapsulation et de dés-encapsulation sont spécifiques aux plateformes mais le format encapsulé est lui

[111] http://www.loc.gov/standards/mets/mets-schemadocs.html .
[112] http://www.commentcamarche.net/telecharger/telecharger-145-winzip.

indépendant des plateformes. ZIP permet, par exemple, une copie fiable du contenu de dossiers d'Apple vers Windows (et vice-versa).

En ce qui concerne le contenu, l'informatique transpose le contenu de la « *FICHE* » d'inventaire en un petit fichier XML beaucoup plus simple qu'un fichier XML conforme à la norme METS. Ce petit fichier reprenant les métadonnées associées à cet item pourrait aisément, si nécessaire ou souhaité un jour, être recodé pour être incorporé dans une structure METS. Il suffit alors à l'institution de veiller à décrire dans un document distinct les formats, conventions, codages et normes utilisées : cela correspond aux données dites de « Préservation » dans l'OAIS.

Stratégie d'exploitation

Choix d'une architecture distribuée

Adoption par la Communauté française du protocole OAI-PMH
Prescriptions

- **identification des ressources : un identifiant et une localisation unique**
- **métadonnées sur les ressources : au minimum au format Dublin Core non qualifié & stockage dans une base de données (ex : SQL)**
- **référentiels : utiliser les référentiels nécessaires à l'interprétation des relations et des équivalences entre les enregistrements de données, organiser leur structure hiérarchique et leurs références externes (« alias »)**
- **encodage en XML (attention aux problèmes de syntaxe)**
- **définition des droits de propriété et des utilisations autorisées**
- **rencontre à optimaliser entre les enregistrements des ressources et les manières dont les utilisateurs souhaitent y accéder**

GESTION DE L'INTEROPÉRABILITÉ

La Communauté française va mettre en œuvre un portail d'interrogation et d'accès commun aux contenus numérisés des institutions culturelles.

Interroger simultanément ou offrir un accès commun à des bases de données hétérogènes (contenant tout type de contenu numérique) tout en laissant les ressources numériques dans les institutions (plus de 150) qui les gèrent et en gardent la responsabilité en Wallonie et à Bruxelles demande une certaine organisation.

La Communauté française fait choix d'utiliser, à cette fin, le protocole OAI-PMH qui est une solution adoptée par de plus en plus d'institutions et de portails, dont les portails européens Michael[113] et Europeana[114] ainsi que les sites de la bibliothèque numérique Gallica de la BNF[115] ou encore celui de la Bibliothèque du Congrès[116].

PROTOCOLE OAI-PMH

Le protocole OAI-PMH - Open Archives Initiative Protocol for Metadata Harvesting (ou protocole OAI) est un outil important, simple et standard pour la mise en commun de ressources documentaires et pour atteindre une plus grande interopérabilité. L'OAI-PMH a été mis au point par l'Open Archive Initiative[117]. Le protocole a été publié en 2002[118].

Concepts

« *Ce protocole d'échange permet de créer, d'alimenter et de tenir à jour, par des procédures automatisées, des réservoirs d'enregistrements qui signalent, décrivent et rendent accessibles des documents, sans les dupliquer ni modifier leur localisation d'origine.* (…) *Enfin, le protocole OAI permet de faire communiquer entre elles des bases de données diverses et hétérogènes, et donc de réaliser des partenariats entre plusieurs établissements*

[113] Michael se concentre sur la représentation ouverte des institutions et des collections (http://www.michael-culture.eu.).
[114] Europeana se concentre sur la recherche et l'accès multilingue aux contenus qui auront été récoltés (« harvesting » en anglais) (http://www.europeana.eu/portal/).
[115] http://gallica.bnf.fr.
[116] "*OAI at the library*"(http://memory.loc.gov/ammem/oamh).
[117] http://www.openarchives.org .
[118] http://www.openarchives.org/OAI/openarchivesprotocol.html . Une version arrêtée au 7 décembre 2008 est disponible.

que rapprochent leurs collections (complémentarité des fonds) ou leurs publics (services culturels d'une même collectivité) »[119].

OAI-PMH est un protocole qui normalise le transfert de métadonnées (pas de modification des structures existantes, ce n'est qu'une « couche » ajoutée au-dessus de l'architecture de l'archive). Il est basé sur HTTP et XML (standards du web). Il est indépendant des logiciels et des plates-formes.

OAI-PMH se préoccupe uniquement des métadonnées et ne touche pas aux ressources numériques. Il permet la centralisation des métadonnées décrivant des ressources diverses, tout en laissant les ressources à leur place originelle. La question de savoir si oui ou non les ressources sont accessibles en ligne est en dehors du protocole.

L'OAI-PMH définit plusieurs rôles dans une architecture exploitant les métadonnées[120]. Le « fournisseur de données » donne accès à ses métadonnées (dans un ou plusieurs formats de description) à travers ce que l'on nomme un « entrepôt OAI » (base de métadonnées sur la machine du fournisseur de données) qui est un outil chargé de répondre aux requêtes formulées par un « fournisseur de services » (moissonneur), qui recueille ces métadonnées dans une représentation spécifique (le schéma de description des métadonnées), et offre aux utilisateurs des services de recherche sur ces métadonnées. La réponse donnée est au format XML et contient, selon la requête formulée, des informations sur l'entrepôt, une liste d'identifiants, de références (métadonnées) ou de « sets » (regroupement de notices correspondant à un thème donné). Le « fournisseur de données » doit fournir ses données au format Dublin Core à tout le moins (et peut proposer d'autres schémas de métadonnées en supplément, sous format XML)[121]. La qualité du contenu de l'entrepôt OAI repose sur la qualité et l'organisation des systèmes d'information source dans les institutions.

Dans le cas du portail fédératif de la Communauté française, un troisième opérateur va intervenir puisqu'un « agrégateur » va rassembler les métadonnées des différentes institutions culturelles (ressources hétérogènes et éventuellement redondantes) et les rendre accessibles dans un entrepôt OAI avec la construction d'un service dont la valeur ajoutée est l'accès « éditorialisé » aux collections. Dans ce cas, l'architecture sera un peu plus complexe puisqu'il s'agira de réaliser des recherches croisées dans plusieurs bases de données, ce que ne fait pas le protocole OAI.

---

[119] François Nawrocki, « *Le protocole OAI et ses usages en bibliothèque* », Ministère de la Culture et de la Communication, Direction du livre et de la lecture, Bureau des politiques documentaires, 15 février 2005, http://www.culture.gouv.fr/culture/dll/OAI-PMH.htm .

[120] Voir Muriel Foulonneau, « *Le protocole OAI-PMH : une opportunité pour le patrimoine numérique* », Relais Culture Europe, Mission de recherche et de la Technologie du Ministère de la Culture et de la Communication, janvier 2003 (http://www.culture.gouv.fr/culture/mrt/numerisation/fr/technique/documents/oai.pdf) et Muriel Foulonneau, « *Collaborer pour de nouveaux services en ligne* », Relais Culture Europe, janvier 2004 (http://www.culture.gouv.fr/culture/mrt/numerisation/fr/technique/documents/guide_oai.pdf).

[121] L'Ukoln propose des recommandations relatives aux équivalences entre modèles de métadonnées (http://www.ukoln.ac.uk/metadata/interoperability/).

Configuration

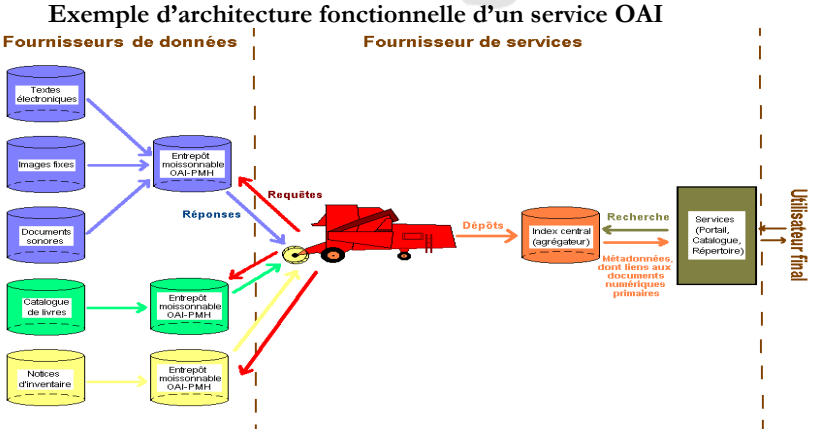
Le protocole OAI comprend six requêtes distinctes qu'un portail peut effectuer vers un entrepôt. Chaque notice dans un entrepôt possède un identifiant unique qui doit se conformer à la syntaxe des URI. Pour assurer la stabilité dans le référencement dans les moteurs de recherche, il est primordial d'assurer la permanence des identifiants. La combinaison des « Verbes » avec leurs différents «Paramètres » permet d'obtenir des ensembles précis et de récupérer les métadonnées contenues dans l'entrepôt.

Verbes	Rôle	Paramètres
GetRecord	Demande d'un enregistrement donné	Identifier MetadataPrefix
Identify	Demande d'informations à propos de l'entrepôt OAI	Aucun
ListIdentifiers	Demande de la liste des éléments contenus dans l'entrepôt	From : date de début Until : date de fin MetadataPrefix Set ResumptionToken
ListMetadataFormats	Demande de la liste des formats de métadonnées disponibles dans l'entrepôt.	Identifier
ListRecords	Demande des enregistrements des métadonnées contenues dans l'entrepôt	From : date de début Until : date de fin MetadataPrefix Set ResumptionToken
ListSets	Demande de la liste des ensembles OAI définis dans l'entrepôt.	ResumptionToken

Source : Wikipedia et Muriel Foulonneau

Un enregistrement OAI est composé de trois sections, « l'en-tête » (« header ») qui contient notamment l'identifiant OAI et la date de collecte, les métadonnées collectées et une section « à propos » (« about ») pour fournir des informations complémentaires sur l'enregistrement de métadonnées (provenance, droits associés).

Exemple d'architecture fonctionnelle d'un service OAI



Source : François Nawrocki, *op.cit.*

Implémentation

L'entrepôt OAI consiste en un ensemble de logiciels capables d'extraire des notices à partir d'une base de données (catalogue ou autres), de les mettre en forme (par exemple, MODS), de les attribuer à des lots distincts selon le cas, et à répondre aux requêtes d'un moissonneur.

La mise en œuvre d'un entrepôt peut se faire en utilisant un module interne au logiciel gérant la base de données ou en réalisant un logiciel externe (il existe des logiciels libres qui le permettent ; ils sont référencés sur le site Openarchives.org) qui obtient les notices de la base de données soit par l'entremise de requêtes envoyées à la base de données, soit à la suite d'exports des notices effectués manuellement ou de façon programmée.

L'Open Archives Initiative propose un guide de mise en œuvre du protocole (http://www.openarchives.org/OAI/2.0/guidelines.htm). De son côté, l'Open Archives Forum, financé dans le cadre du 6ème programme européen (IST-2001-320015), dont les partenaires sont l'Université de Bath-Ukoln (Royaume Uni), l'Istituto di Scienza e Tecnologie della Informazione-CNR (Italie) et Computer and Media Service (Computing Center) de l'Université de Humboldt (Allemagne), met à disposition un didacticiel, (voir sous http://www.oaforum.org/tutorial/index.php). Des exemples émaillent ces didacticiels.

Pour favoriser la coopération des institutions ayant une interface OAI-PMH, un référencement est encouragé par l'OAI. En Communauté française, le Musée royal de Mariemont est inscrit en tant que fournisseur de données.

RESSOURCES DOCUMENTAIRES

Les ressources documentaires en matière de numérisation sont multiples et se multiplient. Voici quelques références qui ont inspiré ce document.

En premier lieu, les travaux réalisés par les groupes de travail mis en place dans le cadre du programme **Minerva** soutenu par la **Commission européenne**, principalement :

- le « *Guide des bonnes pratiques* » version 1.3 édité par le groupe de travail n°6 du 3 mars 2004 qui présente en arrière plan les Principes de Lund et le projet Minerva et propose des directives pratiques en matière de planification du projet de numérisation, de sélection des sources pour la numérisation, de préparation à la numérisation, de maniement des originaux, du processus de numérisation (utilisation de scanners, d'appareils photos numériques, applications pour le reconnaissance optique de caractères), la préservation d'originaux numérisés, les métadonnées, la publication (traitement de l'image, questions liées à la 3D et à la réalité virtuelle, publication en ligne), à la propriété intellectuelle et au droit de copie et, enfin, à la direction de projets de numérisation (http://www.minervaeurope.org/publications/goodhand.htm) ;
- les « *Principes de qualité des sites internet culturels : guide pratique* » publié par le groupe de travail n°5 en 2005 qui définit les dix principes européens pour la qualité de ces sites : être identifiable facilement, présenter des contenus pertinents, assurer la maintenance et les mises à jour des contenus, être accessible à tous les utilisateurs, être adapté aux besoins des utilisateurs, être réactif, être multilingue, s'efforcer d'être interopérable, être respectueux des droits, assurer la pérennité du site et des contenus en adoptant des stratégies et des standards adaptés (http://www2.cfwb.be/qualite-bruxelles/pg001.asp) ;
- « *Recommandations techniques pour les programmes de création de contenus culturels numériques* », UKOLN, Université de Bath, en collaboration avec MLA : le Conseil pour les musées, bibliothèques et archives, document rédigé dans le cadre du projet Minerva, version révisée le 7 mai 2004. La version française a été réalisée pour le compte de la Mission de la recherche et de la technologie du Ministère de la culture et de la communication (partenaire français du projet Minerva) par Muriel Foulonneau (Relais Culture Europe), Sarah Faraud (Relais Culture Europe) et Alexandra Bonnamy (traductrice). Ce document renvoie à beaucoup d'autres références. La version 2.0 a été diffusée en septembre 2008 (http://www.minervaeurope.org/publications.htm/);
- « *Handbook on Cost reduction in Digitisation* », section 3.3, Minerva Plus, septembre 2006, (http://www.minervaeurope.org/publications/CostReductioninDigitisation_v1_0610.pdf).

Le programme **PrestoSpace** coordonné par l'INA-Institut national de l'audiovisuel français et soutenu par la **Commission européenne**, est une source importante en matière de préservation et d'accès aux archives audiovisuelles (http://prestospace.org; http://prestospace-sam.ssl.co.uk).

Trois plates-formes soutenues par la **Commission européenne :**

- DPE-DigitalPreservationEurope dont un des objectifs est la création d'une plateforme de collaboration entre des initiatives européennes issues de domaines différents (universités, musées, centres d'archives,…), publie des documents, dont des synthèses sur l'interopérabilité ou les métadonnées (http://www.digitalpreservationeurope.eu/). A developpé le site avec les plates-formes européennes
- CASPAR-Cultural, Artistic and Scientific Knowledge for Preservation, Access and Retrieval (http://www.casparpreserves.eu/);

- PLANETS-Preservation and Long-Term Access through Networked Services (http://www.planets-project.eu/).

Ces trois plates-formes ont développé ensemble le site www.wepreserve.eu.

Les **organismes internationaux de normalisation** :

- W3C-World Wide Web Consortium : il a en charge la normalisation de l'ensemble des protocoles d'internet : standards de base (http, HTML, XHTML, DOM, XML, XSL, …), standards autour de l'interopérabilité et des services web (SOAP, WSDL, Web,…), standards concernant l'accessibilité (WAI), standards liés à la sémantique et à la description de ressources (XML Schema, RDF, langages d'ontologies OWL) (http://www.w3.org). Les documents et recommandations du W3C sont disponibles en français sous http://www.la-grange.net/w3c/fr-trans1;
- ISO-Organisation internationale de normalisation (www.iso.org);
- CEN-Comité européen de normalisation (http://www.cen.eu/cenorm/index.htm).

L'**Unesco**, dans son programme « Mémoire du monde » a établi en 1998 des « *Principes directeurs pour la sauvegarde du Patrimoine documentaire* », version revisée en 2002 (http://unesdoc.unesco.org/images/0012/001256/125637f.pdf).

Les **fédérations (coordinations) professionnelles ou sectorielles** fournissent documentations et recommandations et dans certains cas proposent des programmes de formation ou participent à des programmes de recherche :

- ACE-Association des cinémathèques européennes (http://acefilm.de/index.php?id=9&L=2);
- AMIA-Association of Moving Image Archivists (http://www.amianet.org/);
- ARSC-Association for Recorded Sound collections (http://www.arsc-audio.org/);
- CCAAA- Coordinating Council of Audiovisual Archive Associations, forum constitué en 2000 par les associations reconnues par l'UNESCO (AMIA, IASA, ICA, FIAF, IFLA, FIAT/IFTA et SEAPAVAA) (http://www.ccaaa.org/);
- FIAF-Fédération internationale des archives du film : édite notamment le "Journal of Film Preservation" qui propose un forum de discussions sur les aspects théoriques et techniques de l'archivage des images en mouvement (http://www.fiafnet.org/fr/);
- FIAT-Fédération internationale des archives de télévision : des standards seront prochainement disponibles en ligne (http://www.fiatifta.org/cont/what_is_fiat.aspx);
- IASA-Association internationale d'archives sonores et audiovisuelles : a développé des guides techniques disponibles en ligne qui font référence pour les archives sonores (http://www.iasa-web.org/technical_guidelines.asp). Voir notamment le document du Comité technique de l'IASA, « *Sauvegarde du patrimoine sonore : Ethique, principes et stratégies de conservation*", IASA-TC 03, décembre 2005, version revue en 2008 (http://www.iasa-web.org/downloads/publications/TC03_French.pdf);
- ICA-Conseil international des archives (http://www.ica.org/fr);
- IFLA-Fédération internationale des associations de bibliothécaires et des bibliothèques (http://www.ifla.org/);
- ICOM-Conseil international des musées (http://icom.museum/mission_fr.html);
- PIAF-Portail international archivistique francophone : a conçu un didacticiel, voir notamment son chapitre relatif à la « Gestion et archivage des documents numériques ». (http://www.piaf-archives.org/)

En **France**, le Ministère de la Culture et de la Communication a ouvert un site réservé à la numérisation du patrimoine culturel qui comprend des recommandations techniques, des lexiques, une importante base documentaire (http://www.numerique.culture.fr). Le document « *Conservation à long terme des documents numérisés* » a été mis à jour en 2008 (http://www.culture.gouv.fr/culture/mrt/numerisation/fr/technique/documents/conservation.pdf)

Le Ministère du Budget, des Comptes publics et de la Fonction publique de la République française, Direction générale de la modernisation de l'État, a édité un « *Référentiel général d'interopérabilité* », version 3, 22 juin 2007 (http://www.synergies-publiques.fr/article.php?id_article=746).

Le groupe de travail et d'échanges français PIN-Pérennisation de l'information Numérique, créé en 2000, comprend des institutions patrimoniales, des organismes à caractère scientifique et technique, de grands groupes publics et privés, des experts, … (http://vds.cnes.fr/pin/index.html).

Au **Royaume-Uni**, deux coalitions ont été constituées : Digital Preservation Coalition (http://www.dpconline.org/graphics/index.html) et Digital Curation Center (http://www.dcc.ac.uk/).

Parmi les autres sources les plus utiles :

- les documents du JISC-Joint Information System Comitee (http://www.jisc.ac.uk/) et de ses services dont le TASI- Technical Advisory Services for Images, qui est hébergé à l'Université de Bristol (Institute for Learning and Research Technology) (http://www.tasi.ac.uk/);
- les documents de l'Ukoln, basé à l'Université de Bath et fondé par le Council for Museums, Librairies ans Archives-MLA, le JISC et d'autres institutions du Royaume-Uni (http://www.ukoln.ac.uk/).

En **Suisse**, l'association Memoriav présente sur son site « Préserver le patrimoine audiovisuel » des recommandations sur la sauvegarde du patrimoine audiovisuel (photo, film, vidéo, son) (http://www.memoriav.ch/).

Au **Canada**, le site de Patrimoine canadien contient une rubrique relative à la création et la gestion de contenu numérique (http://www.chin.gc.ca/Francais/Contenu_Numerique/index.html) ainsi qu'un document contenant les « Exigences et recommandations techniques pour les projets soutenus par Culture canadienne en ligne » (http://www.pch.gc.ca/pgm/pcce-ccop/publctn/tech-fra.cfm).

Aux **Etats-Unis**, deux sources parmi les plus essentielles :

- la Bibliothèque du Congrès propose en ligne une mine d'informations, notamment sous la rubrique "Sustainability of Digital Formats Planning for Library of Congress Collections" qui recommande des formats pour les images fixes et animées, le son, le texte, les images animées et les archives en ligne. C'est une des sources les plus précises et les plus complètes en la matière (http://www.digitalpreservation.gov/formats/intro/intro.shtml);
- l'Université de Cornell et la société des archivistes ont rédigé un didacticiel « *Guide de la conservation de collections numériques* » dont une version française est disponible sous http://www.library.cornell.edu/iris/tutorial/dpm-french/index.html..

Sont à relever aussi les documents du :

- NISO-National Information Standards Organisation, organisation agréée par l'ANSI-American National Standards Institute (http://www.niso.org);
- CLIR-Council on Libraries and Information Resources (http://www.clir.org/);
- l'Université de Stanford, « Conservation on line – Cool» (http://palimpsest.stanford.edu).

En **Australie**, la Bibliothèque nationale consacre un site web, PADI, à la préservation et l'accès à long terme de l'information numérisée, « The National Library of Australia's Preserving Access to Digital Information-PADI » (http://www.nla.gov.au/padi/index.html).