Date : 02/07/2008

La préservation numérique à la Bibliothèque nationale de France : présentation technique et organisationnelle

Emmanuelle Bermes
Bibliothécaire du numérique, chef du service Prospective et services documentaires, département de l'Information bibliographique et numérique

Isabelle Dussert Carbone
Directrice du département de la conservation

Thomas Ledoux
Chef de projet, département des Systèmes d'information

Christian Lupovici
Directeur du département de l'Information bibliographique et numérique

Meeting: **84. Preservation and Conservation, (PAC), Information Technology, IFLA-CDNL Alliance for Bibliographic Standards (ICABS) and Law Libraries**

Simultaneous Interpretation: Not available

WORLD LIBRARY AND INFORMATION CONGRESS: 74TH IFLA GENERAL CONFERENCE AND COUNCIL
10-14 August 2008, Québec, Canada
http://www.ifla.org/IV/ifla74/index.htm

Les bibliothèques nationales passent au numérique : l'accélération de la croissance des collections numériques, le changement de nature et d'échelle du dépôt légal, le développement du dépôt légal de l'Internet et du *records management* confrontent la bibliothèque à de nouveaux défis, dont la question de la préservation de ce nouveau support sur le long terme. Ce changement d'environnement constitue un défi technique et organisationnel : il impose de repenser les missions de la bibliothèque en termes d'infrastructure, de gestion de projet, de ressources humaines, d'activités courantes, etc. La Bibliothèque nationale de France (BnF) est en train de créer son propre magasin numérique, SPAR (Système de Préservation et d'Archivage Réparti), une infrastructure sécurisée qui sera capable de collecter, stocker, préserver et diffuser en grande quantité des documents numériques de nature diverse, numérisés ou nés numériques.

Implémentation technique

SPAR est un entrepôt de confiance, conforme au standard OAIS (Open Archival Information System / Système ouvert d'archivage d'information – ISO 14721). Il possède une structure modulaire et évolutive. Il fournit un système complet en miroir avec des fonctions de suivi et d'alerte ainsi qu'un plan de récupération d'urgence.

Le système SPAR interagit avec diverses applications de production liées à la génération d'objets numériques dans la bibliothèque, comme les processus de numérisation, la création d'objets numériques (comme le *records management*), ou la collecte d'objets nés numériques dans le contexte de l'archivage de l'Internet.
Du côté de l'accès, ces objets numériques sont fournis par SPAR à des applications de diffusion, qui prennent en charge la remise des objets numériques aux utilisateurs finaux. Ces applications de diffusion incluent entre autres Gallica[1], la bibliothèque numérique de la BnF.

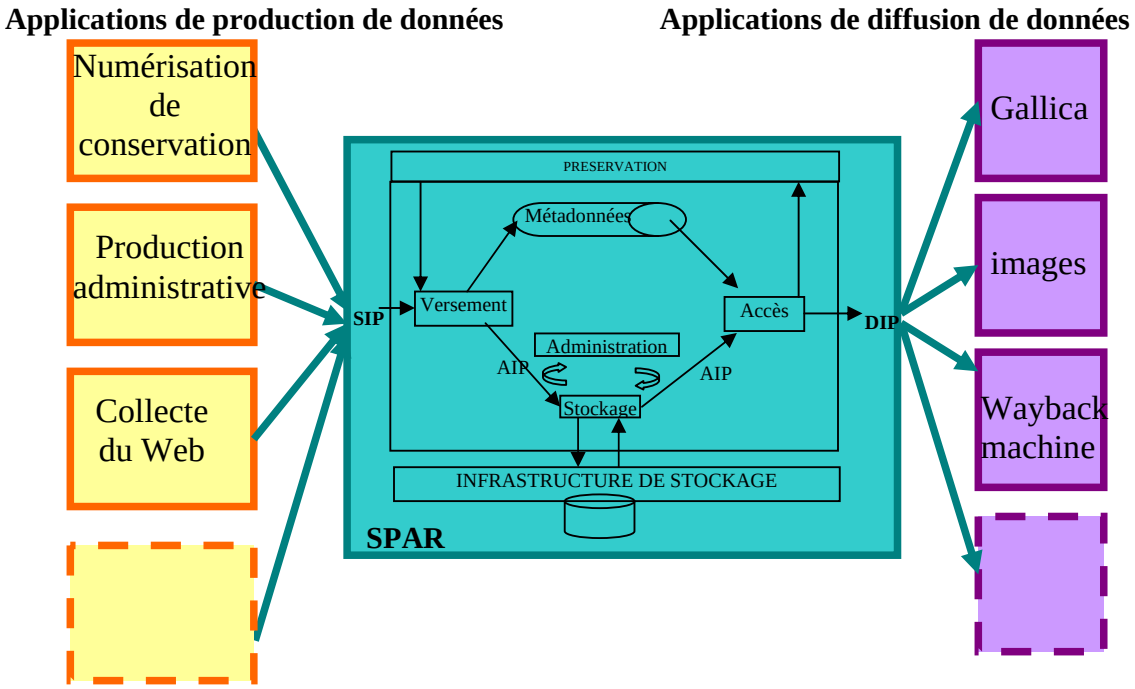

Fig. 1 – Présentation du système SPAR – schéma fonctionnel

L'implémentation du système SPAR se divise en deux sous-projets :

- Le sous-projet « SPAR Infrastructure » vise à acquérir et implémenter l'infrastructure technique requise par le système.
- Le sous-projet « SPAR Réalisation » vise à bâtir le système de manière à stocker, préserver sur le long terme et communiquer les documents numériques de la bibliothèque à partir de l'infrastructure disponible.

Architecture logicielle

Ainsi, SPAR est un système de préservation de contenus numériques sur le long terme. Il possède une structure **modulaire** et évolutive, qui se compose des modules génériques suivants :

- Module de versement
- Module de stockage
- Module de gestion des données
- Module de gestion des droits

[1] http://gallica.bnf.fr.

- Module d'accès
- Module d'administration
- Module de préservation
- Module technique (« Service d'Abstraction de Stockage »)

D'autres modules plus spécifiques s'y ajoutent :

- Pré-versement (construction des paquets de versement)
- Livraison des paquets de diffusion.

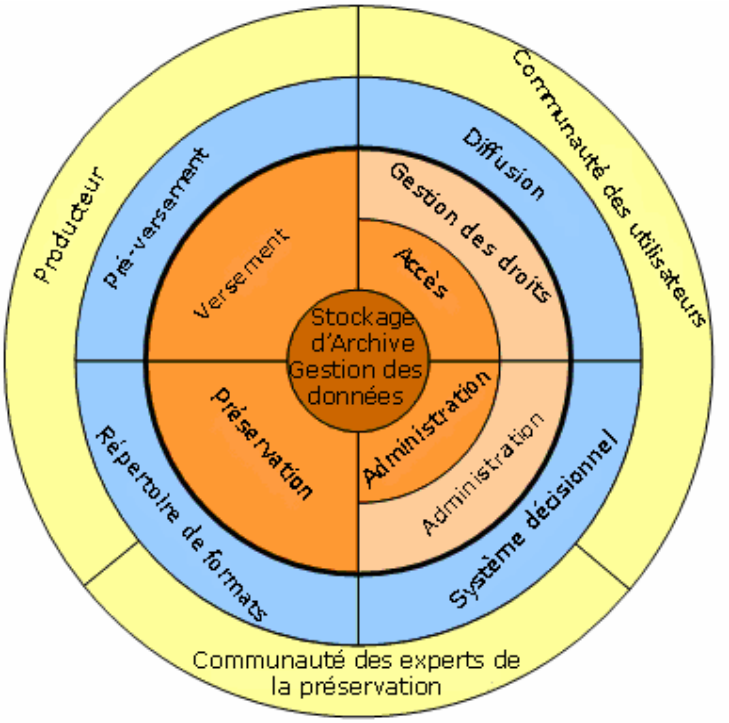

Fig. 2 – Présentation conceptuelle du système SPAR.

La spécificité de ces deux derniers modules est directement liée à la notion de **chaîne**. Une chaîne se caractérise par les relations entre la production ou l'utilisation des objets numériques (besoins et exigences) et le système d'archivage (contrat de service).
Le contenu devant être versé dans le futur système est classé en 7 chaînes différentes :

- numérisation de conservation : contenus numérisé par la bibliothèque à partir de supports analogiques dans un but de conservation préventive
- numérisation de reproduction : contenus numérisés à la demande (commande de clients) par la bibliothèque à partir de supports analogiques
- dépôt légal automatique : archives de l'Internet collectées par le biais de processus automatiques (moissonneurs)
- dépôt légal négocié : contenus versés dans le cadre du dépôt légal, collectés selon des modalités à négocier avec les producteurs
- *records management* : contenus numériques produits par la bibliothèque dans le cadre de ses activités
- tiers-archivage : contenus archivés pour le compte d'un tiers
- acquisitions : contenus acquis par achat ou collectés par don, legs, etc.

Infrastructure

Le module technique « Service d'Abstraction de Stockage » garantit l'indépendance vis-à-vis de l'infrastructure physique par le biais de **capsules de stockage**. Ces capsules se définissent par un ensemble de caractéristiques et de services qui permet de communiquer avec un composant de stockage particulier ou, plus couramment, avec un ensemble de ces éléments (par exemple, en associant un ensemble de bandes avec une partition dans une baie de disques). Ces différentes capsules sont configurées par les administrateurs à partir des éléments d'infrastructure à leur disposition, pour se conformer au mieux aux exigences définies par les politiques.

L'infrastructure actuelle est constituée des composants suivants :

- l'infrastructure est répartie sur 2 sites géographiquement distincts,
- chaque site possède sa propre librairie de bandes magnétiques utilisées pour héberger l'AIP,
- les échanges (dépôt ou livraison) se font par le biais de baies de disques nommées SSS (pour les entrées) et SSC (pour la diffusion) ;
- les serveurs de gestion (SUG) et de traitement (SUT) hébergent la partie logicielle,
- les outils de supervision des supports offrent des fonctionnalités de suivi et d'alerte.

Réalisation

Le développement de ce système est planifié dans un contrat quadriennal. La première année vise à réaliser le noyau commun du système ainsi que la chaîne de numérisation de conservation. Les autres chaînes seront développées de manière itérative.

Le système d'ensemble est construit à partir de composants réutilisables assemblés dans un *framework* en Open Source : FedoraCommons. Le système dans son ensemble offre les avantages suivants : une version de développement stable, une garantie de fiabilité, et une longue durée de vie car il est associé à une large communauté.

Modèle d'information

Cependant, l'infrastructure technique ne suffit pas à remplir l'objectif de la préservation sur le long terme : le projet SPAR connaît un engagement fort en faveur d'une gestion de données en continu, et une viabilité organisationnelle propre à éviter que le système ne tombe en obsolescence ou ne souffre d'un manque de constance dans le suivi au jour le jour.
Du point de vue de la gestion de données, toute la conception est fondée sur des standards faisant autorité tels que METS (Metadata Encoding & Transmission Standard / Standard d'encodage et de transmission de métadonnées) et PREMIS (Preservation Metadata Maintenance Activity – Activité de maintenance des métadonnées de préservation).

Standards de métadonnées

Le standard METS a été retenu comme format d'empaquetage. Il relie ensemble les Objets-données (fichiers numériques ou flux de bits pris en charge par le système Fedora) et leurs métadonnées. Ces dernières se composent au minimum d'un sous-ensemble de métadonnées descriptives encodées en Dublin Core, importées depuis le catalogue bibliographique par le biais d'un entrepôt OAI-PMH, mais les métadonnées sont en grande partie des métadonnées

administratives et techniques générées par le système lui-même. Les métadonnées de provenance, qui se chargent de la chaîne d'audit des changements qui se produisent à l'intérieur du système, sont enregistrées sous la forme d'une série d'événements encodés conformément au standard PREMIS. Les métadonnées techniques sont extraites des fichiers numériques pendant le processus de versement, et encodées dans la section des métadonnées techniques du fichier METS, en utilisant des schémas d'extension appropriés tels que MIX pour les images fixes et TextMD pour les documents en mode texte. Le manifeste METS ainsi obtenu fournit une vue d'ensemble de l'objet numérique à préserver, et est stocké avec les fichiers numériques correspondants dans le système de préservation. Toutefois, pour assurer l'accessibilité de ces métadonnées et la possibilité de les interroger de manière souple, un système de gestion des métadonnées sera développé. La gestion de données sera basée sur un *mapping* depuis METS vers RDF, et les données en RDF qui en résulteront seront entreposées dans un entrepôt RDF. Cette technologie permettra un accès optimal à l'ensemble de l'information stockée dans le système et permettra un suivi rigoureux des objets numériques dans la perspective de stratégies de préservation telles que la migration et l'émulation.

Gérer la granularité, les versions et les stratégies de préservation

En complément de ce modèle de métadonnées, un modèle générique d'information pour la gestion de la granularité et des versions a été conçu. Ce modèle inclut quatre niveaux de granularité (set – group – object – file) qui recouvrent toutes les différentes configurations possibles d'objets complexes, le niveau « set » étant récursif. Dans le manifeste METS, cette structure est définie dans la section « carte de structure » (structural map) :

- **set** : ce niveau est utilisé pour une collection d'objets numériques, le titre d'un périodique ou un document multimédia
- **group** : ce niveau correspond à l'objet numérique : par exemple, une monographie, un numéro de périodique, un film
- **object** : une partie de l'objet numérique, comme une image, une page, une piste de CD audio
- **file** : l'Objet-données (fichier numérique ou flux de bits).

Par exemple, un numéro d'un journal numérisé est considéré comme un « group ». Le « set » correspond au titre du journal. Chaque page du numéro constitue un « objet », lui-même composé de deux « files » : un pour l'image de la page et un pour la transcription OCR. Les niveaux « set » et « group » sont tous deux considérés comme des paquets indépendants et faiblement liés, accompagnés de leur propre manifeste METS. Les différents types de métadonnées s'appliquent aux niveaux appropriés.

Trouver un équilibre entre la nécessité de mettre à jour les objets numériques et celle de la sécurité et de la fiabilité de l'Archive constituait un défi. Aussi était-il nécessaire de permettre des mises à jour régulières de métadonnées (y compris les métadonnées descriptives) et une amélioration progressive des Objets-données (par exemple l'ajout d'une nouvelle version du fichier OCR, lorsque la précision de l'OCR est améliorée grâce à l'évolution de l'état de l'art). Le système de gestion de versions prend cet aspect en compte et fournit une gestion souple du cycle de vie fondé sur des versions et des éditions. Quand les Objets-données sont touchés par la mise à jour (par ex. en remplaçant ou en supprimant un fichier numérique), une nouvelle version est créée et la version antérieure est conservée. À l'inverse, quand seules les métadonnées sont modifiées, ou que des Objets-données sont ajoutés, mais qu'aucun des deux n'est modifié ou supprimé, une nouvelle édition est créée, remplaçant l'état précédent de l'objet numérique.

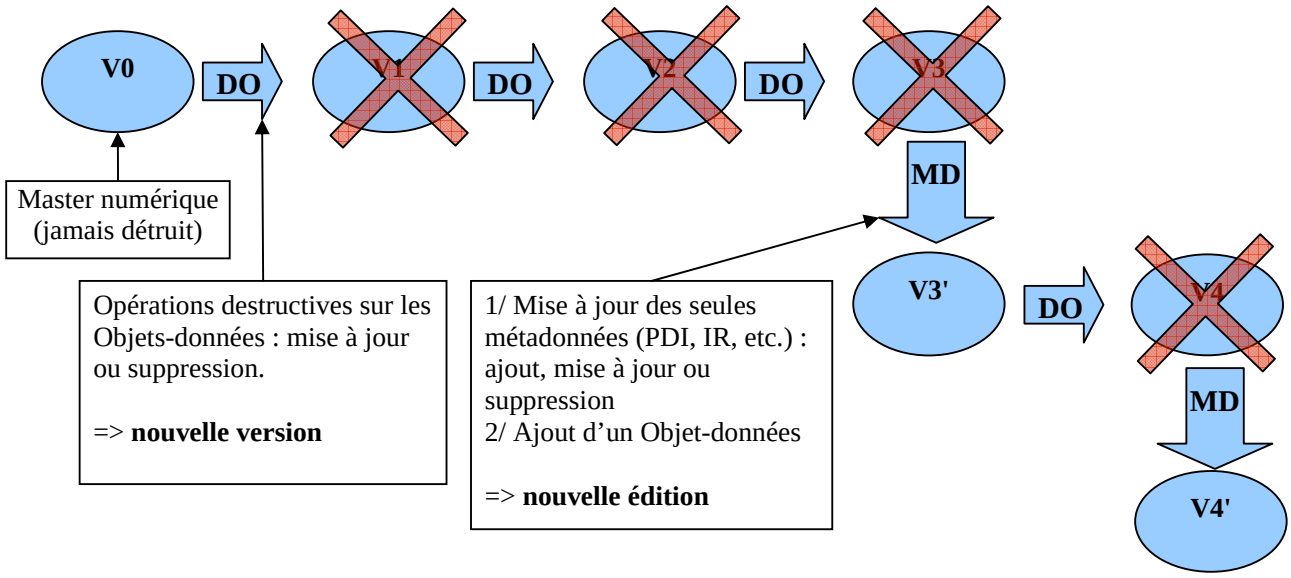

Fig. 3 – Gestion du cycle de vie pour les versions et les éditions de paquets

Grâce à ce modèle, nous escomptons que les futures stratégies de préservation pourront être gérées de manière souple. Les opérations qui présentent un risque de perte de données seront protégées par le système de gestion de versions ; toutefois, le système ne sera pas paralysé par ce choix de sécurité.

La gestion de collection numérique

Étant donné que le système SPAR a vocation à atteindre un haut niveau de modularité et de flexibilité, un fort engagement en faveur d'une bonne gestion des collections numériques est nécessaire. Cet objectif sera atteint par le biais de procédures normalisées qui régissent la relation entre le producteur et l'Archive. Cette négociation se fonde sur le standard PAIMAS et sur des accords sur la qualité de service. Pour chaque chaîne, le producteur et l'Archive doivent négocier trois types de politiques : une pour le versement, une pour la préservation et une pour la diffusion. Ces politiques, ou accords sur la qualité de service, sont des procédures contractualisées qui obligeront le producteur à exprimer ses besoins en termes quantifiables, qui ont ensuite à être convertis en règles formelles exploitables par le système. Les politiques sont destinées à être archivées dans le système tout autant que les objets eux-mêmes, de manière à ce que le système soit complètement auto-descriptif.

L'accord sur la qualité de service est un document contractualisé qui décrit de manière exhaustive les processus, les acteurs, les contenus et les stratégies associées à une chaîne. Le cahier technique détaillé décrit précisément la provenance des métadonnées et la structure de la granularité des paquets. L'accord sur la qualité de service comprend trois politiques pour chaque chaîne :

- la politique de versement (formats acceptés, volumétrie, niveaux de sécurité, …),

- la politique de préservation (temps de rétention, niveaux d'assurance, …),
- la politique d'accès (formats de diffusion, temps de mise à disposition, disponibilité, …).

La politique de versement permet de valider les versements effectués par le producteur et de prendre ses responsabilités en fonction de la catégorie de format :

Code	Catégorie de format	Description
00	Stocké	Format dont les caractéristiques techniques sont inconnues (non identifiées) et pour lequel la seule préservation du flux de bits est assurée.
01	Identifié	Format dont les caractéristiques techniques sont connues (identifiées dans un répertoire de formats), mais pour lequel ni la migration ni l'émulation ne sont envisagées. Un format identifié devient maîtrisé ou connu si un tel plan est mis en œuvre.
10	Connu	Format identifié pour lequel la BnF détient au moins un outil de référence, connaît ses utilisations, suit les évolutions, et pour lequel la BnF a défini un plan soit pour le transformer en un format maîtrisé, soit pour l'émuler.
11	Maîtrisé	Format connu pour lequel la BnF détient la documentation publiée et au moins un outil de référence, dont elle suit les évolutions et pour lequel elle a défini des contraintes spécifiques avec les producteurs.

La politique de préservation définit l'endroit où les paquets d'information archivés (AIP) sont stockés et la gestion de leur cycle de vie. Un format maîtrisé est requis pour mettre en place une stratégie de migration, tandis qu'un format connu est requis pour une émulation.

La politique d'accès définit en particulier les formats de diffusion ainsi que les contraintes d'accès qui déterminent si l'objet diffusé doit être pré-calculé ou généré à la volée.

L'accord de qualité de service prend aussi en compte une stratégie de gestion des risques afin d'assurer le transfert de responsabilité du producteur à l'Archive. Le processus d'ensemble est une garantie d'avoir une bonne connaissance des risques liés aux objets numériques, et des engagements appropriés du côté du producteur, pour verser les contenus appropriés, et du côté de l'Archive, pour lancer les actions nécessaires afin d'assurer le service de préservation.

Problématique organisationnelle

D'un point de vue organisationnel, les activités spécifiques à la bibliothèque sont en cours de développement : la gestion des métadonnées, l'analyse du document, la gestion des collections, les définitions de politiques, la gestion de droits, etc.

La numérisation de masse et les objectifs de préservation

Ces nouvelles activités spécifiques qu'implique le développement de SPAR ne relèvent pas seulement du domaine des technologies de l'information, mais aussi de la gestion de collections. Les facilités d'accès fournies par la numérisation et la conversion OCR améliorent considérablement le confort et l'expérience utilisateur par rapport à l'utilisation du papier ou des microformes. Pour cette raison, la demande de programmes de numérisation

s'accroît, et provient de tous les départements chargés de gestion de collections. La numérisation est aujourd'hui également envisagée dans une perspective de conservation, le document numérisé étant considéré comme un nouveau type de substitut, capable d'empêcher l'original de subir des dégradations liées à son utilisation par le public. Dernier point et non des moindres, la numérisation est entreprise dans un but intellectuel, dans le cadre de portails thématiques sur le Web et de collections virtuelles. Ces différentes motivations pour la numérisation peuvent prêter à confusion lors de la mise en place de projets de numérisation de masse : la politique documentaire (d'un point de vue scientifique et intellectuel), comprenant des présentations thématiques spécifiques à certaines collections, est parfois opposée à la politique de numérisation de conservation. Le travail sur l'organisation de SPAR a accru la confiance dans la préservation sous forme numérique, au point que la bibliothèque est en train d'abandonner les microformes en faveur de substituts numériques. Cette orientation doit être planifiée financièrement. Dans l'univers analogique, le coût des étagères et des opérations traditionnelles de conservation des collections est planifié et calculé ; dans l'univers numérique, la bibliothèque doit également prévoir et planifier les coûts, qui comprennent :

- le stockage et la préservation d'une masse gigantesque de contenus numériques,
- le développement et la maintenance logiciels,
- les mises à jour et la maintenance de l'infrastructure matérielle.

Ces prévisions sont comparables à celles d'une construction de bibliothèque, et présentent des besoins tout à fait similaires :

Univers analogique	Univers numérique
bâtiment	salle des machines
rayonnages	stockage
catalogue en ligne	moteur de recherche
système de gestion	administration
restauration	système de préservation
numéro de téléphone	identifiant
réservation et prêt	systèmes de diffusion

Il est n'est pas encore sûr aujourd'hui que la bibliothèque ait bien évalué les coûts réels d'un programme aussi ambitieux. De plus, au-delà des aspects techniques, il est nécessaire de prendre en compte un ensemble d'activités spécifiques de préservation comme facteur de réussite d'un programme de préservation numérique sur le long terme.

Les activités de préservation

La préservation numérique est un véritable enjeu de conservation ; par conséquent, des méthodologies familières telles que les stratégies d'évaluation des risques, peuvent s'appliquer au numérique. Les stratégies de préservation sont définies dans des programmes comprenant des mesures préventives et curatives, afin de se maintenir à un bon niveau technologique et assurer le rafraîchissement, la migration, la copie, l'émulation. Passer en revue la collection numérique est le seul moyen de vérifier que le contenu versé dans le système SPAR n'a pas été altéré, et cette activité de suivi est également bien connue des experts de conservation. Les métadonnées de préservation équivalent aux informations de restauration collectées lors d'activités de conservation de contenus analogiques, et elles garantissent la possibilité de retracer l'histoire du contenu numérique. Les plans de recouvrement en cas de désastre doivent inclure les systèmes de préservation numérique. Les activités particulières de négociation entre le producteur et l'Archive sont bien connues des services de restauration, et doivent être étendues à la discussion sur les contenus numériques.

Pour les contenus nés numériques tels que les archives de l'Internet et d'autres types de supports numériques (CDs, DVDs, jeux vidéo…), le défi de préservation est encore plus grand. D'un point de vue technique, ce type de contenu peut poser des problèmes spécifiques tels qu'un environnement matériel et logiciel spécifique, des protections avec DRM (digital rights management / gestion des droits numériques), des formats de fichiers rares ou inconnus…, qui rendent les conditions de préservation sur le long terme plus difficiles à créer. En outre, ces contenus, en particulier les archives de l'Internet, sont soumis à un changement d'échelle : les masses de documents collectés requièrent de lourdes capacités de stockage et impliquent une approche de sélection et d'échantillonnage peu coutumière des bibliothécaires dans des contextes de dépôt légal et de conservation. Cette approche semblerait plus voisine des pratiques des archivistes et suppose que l'on s'efforce d'informer les futures communautés d'utilisateurs sur l'état des collections et les contraintes techniques liées au processus de collecte.
Toutes ces activités propres à la préservation (planification de la préservation, suivi des collections, création et interrogation de métadonnées, plans de recouvrement, négociation entre le producteur et l'Archive) soulignent le besoin de délimiter les responsabilités quant à l'évolution et la gestion courante de SPAR.
L'expertise requise pour évaluer des collections particulières afin de faire des choix, définir des priorités et des actions de planification de la préservation, constituent des compétences traditionnelles du bibliothécaire qui doivent être adaptées à l'environnement numérique. Cela implique que le personnel reçoive une formation spécifique lui permettant de développer des compétences en préservation numérique.

La coopération au royaume de la préservation numérique

SPAR est un projet ambitieux que la BnF n'a pas l'intention de réaliser seule. La coopération à l'échelle mondiale est considérée comme un aspect essentiel du travail, ce qui inclut une participation active à des communautés open source (Fedora et autres), ou la mise en place de groupes d'intérêts comme PASIG (Preservation and Archiving Special Interest Group / Groupe de travail spécifique sur la préservation et l'archivage, en partenariat avec Sun Microsystems).
La coopération par le biais de groupes spécifiques comme PIN (Pérennisation de l'Information numérique, groupe de travail français d'expertise et de diffusion) ou IIPC (International Internet Preservation Consortium / Consortium international pour la préservation de l'Internet) est essentielle dans le cas de la préservation d'objets numériques, afin de faire de l'évangélisation ou de définir les bonnes pratiques, et pour élaborer des solutions techniques et organisationnelles dont nous avons tous besoin face au défi de la préservation numérique.

Bibliographie

Bermès Emmanuelle, "Risk Management: methodological principles" in *International preservation news*, n°41, july 2007.

Calderan, Lisette (dir) et al. *Pérenniser le document numérique. Séminaire INRIA, 2-6 octobre 2006, Amboise.* ADBS, 2006.

Fedora commons. http://www.fedora-commons.org/

Kaczmarek Joanne et al. “Using the Audit Checklist for the Certification of a Trusted Digital Repository as a Framework for Evaluating Repository Software Applications. A Progress Report.” *D-Lib Magazine*, Volume 12 Number 12, December 2006.

Kolding Nielsen, Erland. “Digitisation of library material in Europe: problems, obstacles and perspectives anno 2007.” *Library quaterly*, Vol 18, 2008, n°1

Lavoie Brian, Gartner, Richard. *Preservation metadata. Technology watch report*. OCLC/DPC, september 2005.

METS. *Metadata encoding and transmission standard*. http://www.loc.gov/standards/mets/

MIX. *NISO metadata for images in XML format.* http://www.loc.gov/standards/mix/

PREMIS. *Preservation metadata.* http://www.loc.gov/standards/premis/

Producer-Archive methodology abstract standard. CCSDS, Blue book, may 2004. http://public.ccsds.org/publications/archive/651x0b1.pdf

Reference model for an open archival information system (OAIS). CCSDS, Blue book, January 2002. http://public.ccsds.org/publications/archive/650x0b1.pdf

Rieger Oya Y., *Preservation in the Age of Large-Scale Digitization. A White Paper*. CLIR, February, 2008.

TextMD. Technical metadata for text. http://www.loc.gov/standards/textMD/

Verheul Ingebord, *Networking for digital preservation Current practice in 15 national libraries*. KB/IFLA/Saur, 2006

Verheusen, Astrid. “Mass digitisation by libraries : issues concerning organisation, quality and efficiency.” *Library quaterly*, Vol 18, 2008, n°1