Date submitted: 10/08/2010

Une approche orientée données pour la préservation du numérique : le projet SPAR

Emmanuelle BERMES
Louise FAUDUET
et
Sébastien PEYRARD
Bibliothèque nationale de France
Paris, France

Traduction :
Emmanuelle BERMES
(Bibliothèque nationale de France, Paris, France)

Meeting: **157. ICADS with Information Technology**

WORLD LIBRARY AND INFORMATION CONGRESS: 76TH IFLA GENERAL CONFERENCE AND ASSEMBLY
10-15 August 2010, Gothenburg, Sweden
http://www.ifla.org/en/ifla76

Résumé :

En 2010, SPAR, le système de préservation du numérique de la Bibliothèque nationale de France, est entré en production. SPAR a été conçu pour recevoir et préserver plus d'1.5 pétaoctets de données issues de collections numériques variées, comprenant aussi bien le produit des programmes de numérisation, que les documents audiovisuels ou nés numériques, et les archives du Web. Le système a vocation à s'intégrer dans le flux des activités quotidiennes de la bibliothèque, en permettant aux bibliothécaires de participer activement à la gestion du cycle de vie de leurs collections numériques. La BnF bénéficie aujourd'hui d'une expérience de plusieurs années de travail sur ce projet, et peut partager les enseignements que lui a apportés le développement de ce système de préservation à grande échelle.

La manière dont le système utilise les métadonnées constitue une des grandes forces de SPAR. L'approche orientée données confère une grande souplesse au système, car elle permet de changer le comportement de ce dernier en se contentant de changer les données et non les logiciels ou les processus. Ce système est extensible : il est possible d'y verser de nouveaux types de collections numériques en se cantonnant à améliorer le cœur du système si de nouvelles exigences voient le jour.

Enfin, en créant un système qui fait du modèle de données une « lingua franca » entre les informaticiens, les administrateurs et les bibliothécaires, SPAR a favorisé la mise en commun des compétences réparties dans la bibliothèque. La création d'un nouveau profil de poste d'« expert de préservation » au cours du projet, au croisement du système et des données, s'inscrit dans cette tendance. Ces experts sont responsables du modèle de données

et de la négociation d'accords sur la qualité de service avec les bibliothécaires en charge des collections numériques.

---

1. Le contexte : la BnF et le projet SPAR

En 2006, la Bibliothèque nationale de France (BnF) a lancé un travail de fond visant à bâtir un système de préservation[1] de ses collections numériques. Quatre ans plus tard, ce système, nommé SPAR (Système de Préservation et d'Archivage Réparti), est enfin sorti de sa phase de conception, et les premiers objets numériques. Ce projet a été la source de nombreux enseignements, bien au-delà d'une optique de stricte préservation du numérique.

L'un des principaux défis, lors de la conception d'un système tel que SPAR, est de prendre en compte le risque d'obsolescence du système lui-même. Lors de la création d'un logiciel supposé durer plusieurs années, il est très difficile de garantir que les évolutions de l'environnement technologique, d'une part, et des objets numériques, d'autre part, ne seront pas de taille à nécessiter une refonte complète du système. Par conséquent, le principal défi est de garantir l'aptitude à gérer la collection numérique, quelles que soient les évolutions qui affecteront le système d'information, et ses composants logiciels en particulier.

Afin de nous protéger d'un tel risque, nous avons conçu SPAR dès le début comme un système modulaire : chacune de ses fonctions devait pouvoir être améliorée à son propre rythme, afin de faciliter l'intégration de nouvelles technologies. Ainsi le système a été divisé en modules fondés sur les entités du modèle fonctionnel OAIS : Versement, Gestion de données, Stockage, Accès, Administration et Planification de la préservation, ce dernier devant être développé à une date ultérieure. Ces modules forment le « cœur » de SPAR.

Des modules supplémentaires, sans équivalent direct dans le modèle fonctionnel OAIS, ont été conçus, comme le module de Gestion des droits, qui n'est pas encore implémenté, ou un module Pré-versement pour chaque grand ensemble d'objets homogènes. L'étape de Pré-versement a pour rôle d'harmoniser les différents documents numériques pour en faire un SIP (Submission Information Package — Paquet d'informations à verser) conforme à SPAR et apte à être traité par le reste du système de manière générique.

Toutefois, cette approche modulaire ne suffisait pas à assurer que ces modules pourraient réellement évoluer indépendamment les uns des autres. Il leur fallait également reposer sur un modèle de données pérenne afin de permettre la gestion des données, même si le logiciel est amené à changer. Dans une optique de préservation du numérique, autoriser le système à traiter les données de manière opaque n'est pas acceptable. L'objectif principal du modèle OAIS est de rendre les processus ayant un impact sur les données aussi transparents que possible. C'est ce que nous appelons l'approche orientée données [6].

Cette approche repose en très grande partie sur la création et la maintenance des métadonnées. Le système est entièrement auto-décrit : processus, agents (y compris logiciels) et formats utilisés dans le système sont documentés et versés dans SPAR sous

---

[1] Par système, les auteurs entendent la partie logicielle de la structure d'ensemble visant à la préservation du numérique. La partie matérielle, ou infrastructure, a été mise en place par la BnF à partir de 2005 et n'est pas décrite dans cet article. Voir [1] pour plus d'informations.

forme de Paquets d'information destinés eux-mêmes à être préservés. Le comportement du système est défini par des contrats qui sont négociés entre les bibliothécaires et les administrateurs. Le module de Gestion de données contient toute l'information nécessaire pour assurer le suivi du système et planifier des opérations de préservation.

A cet égard aussi, le problème ne réside pas uniquement dans la création des métadonnées, mais aussi dans leur utilisation. Le type de métadonnées à générer [4], et le format de métadonnées le plus approprié pour les stocker et les échanger [5], sont des questions qui ont beaucoup mobilisé la communauté de la préservation du numérique et sont désormais classiques. En revanche, peu a été dit ou écrit sur la manière dont ces données doivent être utilisées concrètement pour des opérations de préservation, ou pour la gestion de l'Archive au jour le jour. L'approche orientée données de SPAR a donc suivi une règle fondamentale : si des métadonnées sont créées, elles doivent être utilisées. C'est ce que nous allons développer dans cette communication, en montrant comment la gestion de collections est implémentée dans le système, tout en détaillant les avantages d'une approche orientée données d'un point de vue organisationnel.

2. Préserver des objets, gérer des collections

2.1. La notion de « filière »

La notion de « filière » fait partie des principaux concepts élaborés au début du projet SPAR. Une filière est une collection d'objets partageant les mêmes exigences en termes de préservation. Les filières identifiées sont les suivantes :

- numérisation de conservation
- contenus audiovisuels
- dépôt légal automatique (archivage du Web)
- acquisitions de contenus numériques
- tiers archivage
- production administrative.

Les critères permettant d'identifier une filière sont basés sur l'homogénéité des contenus, mais aussi sur des politiques de haut niveau définissant les contraintes et exigences associées à chaque filière. Par exemple, la nécessité d'accepter tous types de formats constitue une contrainte de la filière de dépôt légal, car nos obligations légales supposent que l'on collecte tous les contenus produits, quelles que soient leur forme ou leur destination. Cette filière a également pour contrainte de préserver pour toujours ce patrimoine ; par conséquent, cette filière n'autorise aucune suppression des données originales versées. À l'inverse, pour les documents administratifs, les contraintes légales imposent la suppression à l'issue d'une période donnée.

Ces exemples montrent que les critères définissant les filières sont fondés sur des considérations politiques et parfois légales, mais que ces mêmes considérations ont des implications techniques très directes et importantes sur les fonctions que le système doit être en mesure d'offrir. Ainsi, la décision de regrouper des objets dans des filières afin de gérer leur politique de préservation constitue un aspect essentiel de SPAR. Les filières sont définies pour un ensemble particulier de contenus numériques homogènes demandant les mêmes services de la part du système.

Ces services sont réalisés par le biais d'un ensemble d'exigences formalisées, qui

gouvernent la relation entre les responsables de ces ensembles particuliers d'objets et les administrateurs du système de préservation. Ces acteurs doivent formaliser la nature exacte de leurs engagements l'un envers l'autre dans une politique, afin d'assurer le transfert de responsabilité du Producteur à l'Archive. Ce processus garantit une bonne connaissance des risques liés aux objets numériques concernés, engage le Producteur à soumettre et verser des contenus appropriés, et certifie que l'Archive effectue toutes les opérations nécessaires pour assurer le service de préservation demandé.

2.2. La négociation entre le Producteur et l'Archive

De 2008 à 2010, nous avons assuré la conception de trois filières : la numérisation de conservation, le tiers stockage (sous-ensemble du tiers archivage) et les contenus audiovisuels. Nous nous sommes rapidement rendu compte que le niveau de la « filière », que nous voyions comme le niveau élémentaire pour gérer les objets, était tout à fait pertinent du point de vue des interlocuteurs de l'Archive (Producteurs et gestionnaires de collections) mais n'était pas adapté à des objectifs techniques. Nous avons dû définir des sous-ensembles d'objets numériques plus restreints qui soient homogènes non seulement du point de vue d'une politique, mais également d'un point de vue technique ; nous avions notamment besoin que les objets aient en commun les mêmes exigences sur les formats, les processus et le stockage, afin que le système puisse leur appliquer des règles globales. Cette distinction nous a amené à définir un niveau plus fin de gestion de collection, nommé « chaîne », niveau auquel tous les objets partagent à la fois les mêmes exigences en termes de politique et les mêmes caractéristiques techniques.

Pour chaque chaîne, le producteur des objets et l'administrateur de l'Archive négocient trois types de contrat : un pour le versement, un pour la préservation et un pour la diffusion. Ces contrats sont des procédures formalisées aidant le Producteur à exprimer ses besoins sous une forme quantifiable, transformée par la suite en règles formelles destinées à être utilisées par le système.

À ce niveau de description, il est possible d'arriver à un point où les politiques portant sur les collections trouvent, telles que définies par leurs responsables, une correspondance concrète dans les exigences du système, d'un point de vue technique. Il devient alors possible de créer des données directement exploitables par une machine, qui décrivent de manière formelle les caractéristiques d'un ensemble de documents, c'est-à-dire une chaîne, et les politiques qui s'appliquent à ceux-là. Ces données peuvent être utilisées par le système afin qu'il puisse savoir comment il doit se comporter et manipuler les objets numériques. Dans SPAR, ce sont les accords sur la qualité de service (AQS) qui jouent ce rôle.

2.3. L'accord sur la qualité de service

L'accord sur la qualité de service est un document formalisé décrivant de manière approfondie les processus, acteurs, contenus et stratégies associés à une chaîne. Il est accompagné d'un « cahier technique détaillé » décrivant précisément la structure des objets à préserver, en particulier, l'origine des métadonnées et le niveau de granularité auquel un paquet correspond.

Chaque document numérique est versé dans le système de préservation SPAR sous la forme d'un Paquet d'informations tel que défini dans le modèle OAIS, accompagné d'un manifeste METS stocké dans chaque paquet et qui constitue l'information d'empaquetage de ce dernier. L'implémentation systématique de METS pour les collections numériques versées dans le système n'est qu'une des facettes de notre approche orientée données ; nous avons également

besoin de décrire les processus du système et les choix effectués lors de sa conception, afin que tout ce qu'un expert de préservation ou un chargé de collections numériques a besoin de savoir sur SPAR soit documenté à l'intérieur même du système.

Dans cette perspective, les politiques (et donc les AQS) doivent être préservées dans le système aussi bien que les objets eux-mêmes, de manière à ce que le système soit entièrement auto-décrit. Dans SPAR, chaque paquet appartient à une filière, que l'on peut définir comme une famille de documents possédant les mêmes caractéristiques intellectuelles et légales ; chaque filière comporte une chaîne pour chaque ensemble ayant des caractéristiques techniques homogènes[2]. La description de chacune des chaînes et des filières est factorisée dans un paquet d'informations dédié.

Les paquets d'information de chaîne contiennent les AQS sous la forme de 3 fichiers exploitables par une machine :

- l'AQS de versement (formats autorisés, volume, niveaux de sécurité…), qui permet de valider les versements du Producteur, et formalise les responsabilités de l'Archive pour chaque catégorie de format ;
- l'AQS de préservation (temps de rétention, niveaux de garantie…), qui définit où les paquets d'information archivés (AIP) sont stockés et comment leur cycle de vie est géré ;
- l'AQS de diffusion (formats de diffusion, temps de diffusion, disponibilité…)[3]

Les AQS sont écrits en XML, et définissent quatre types d'exigences. Les exigences portant sur la chaîne incluent par exemple les dates de validité des AQS, les horaires d'ouverture et de fermeture du service, ou la durée maximale d'indisponibilité. Il existe aussi des exigences portant sur le paquet (taille minimum et maximum d'un paquet, types de formats autorisés et interdits pour cette chaîne, durée de détention d'un AIP, etc.), sur le stockage (nombre de copies, présence de cryptage, etc.), et sur les processus, qui déterminent de quelle manière les ressources du système peuvent être mobilisées par la chaîne (nombre minimum et maximum d'invocations d'un processus sur une période donnée, etc.)

L'ensemble de ces exigences sont incorporées au module Gestion de données lorsqu'un paquet de référence de chaîne est ingéré. Dès lors, les autres modules de SPAR peuvent interroger cette information afin d'exécuter les tâches nécessitant de vérifier certains de ces paramètres.

Afin de voir comment ces données sont utilisées dans le fonctionnement quotidien de SPAR, et quel rôle joue Gestion de données vis-à-vis de ces données, nous pouvons prendre l'exemple du cas d'utilisation « verser un SIP » :

- À chaque fois que le module Versement reçoit la notification d'un nouveau SIP, ce dernier est audité, et son manifeste METS est validé en utilisant l'information du paquet de chaîne qui a été versé dans le module Gestion de données : quels utilisateurs sont autorisés à soumettre des paquets pour cette chaîne, ou quel est le profil METS pour les SIPs de cette chaîne.

---

[2] Par exemple, la chaîne B de la filière audiovisuelle comprend le produit de la numérisation de documents audio et vidéo analogiques acquis dans le cadre du dépôt légal, ce qui correspond à des formats de production bien documentés et faciles à gérer ; en regard, la chaîne A de la même filière concerne le dépôt légal de contenus nés numériques (à l'exclusion des documents collectés sur le Web), qui ont pour contrainte d'être ingérés « tels quels », dont immanquablement avec des formats inconnus ou mal utilisés.

[3] L'AQS de diffusion n'est pas encore complètement implémenté dans SPAR, car notre module Accès se contente pour le moment de communiquer l'AIP « tel quel ».

- Les caractéristiques du SIP sont vérifiées à partir de l'AQS de la chaîne concernée, afin de vérifier que les exigences de versement, telles que la taille maximum ou le nombre d'objets autorisés dans le paquet, sont respectées.
- Chaque fichier est identifié, caractérisé et validé. Le résultat est comparé avec la liste des formats acceptés dans cette chaîne, exprimée dans les AQS. Le comportement que le système doit adopter si les exigences ne sont pas respectées (rejet du paquet ou simple alerte aux administrateurs) est également précisé dans les AQS.

Ce cas d'utilisation montre qu'au sein de SPAR, le concept d'AQS n'est pas une notion abstraite ou un pur outil organisationnel. Il se matérialise dans un document formalisé de manière à être exploitable par une machine, et le système utilise concrètement ces données afin de déterminer certaines de ses opérations les plus cruciales. Ce mécanisme fait partie intégrante de ce que nous nommons l'approche orientée données, parce que les métadonnées n'y jouent pas un rôle uniquement informatif mais sont également utilisées comme paramètres par le système.

La conséquence principale en est que la négociation entre le Producteur et l'Archive, telle que formalisée dans les AQS, fournit aux gestionnaires et aux chargés de collections une véritable garantie que le système va se comporter réellement tel que décrit. Pour les administrateurs, cela facilite le pilotage du flux de travail de la chaîne et aide à déterminer parmi les évolutions du système lesquelles reposent sur de nouveaux AQS et lesquelles nécessitent des modifications du logiciel. Cet aspect accroît la confiance dans le système du point de vue des bibliothécaires comme des administrateurs.

Quant à la pérennité du système, elle s'en trouve également améliorée, puisque des changements dans le flux de travail du Producteur (par exemple l'ajout d'un nouveau format de production, ou l'augmentation de la taille moyenne d'un paquet) se traduisent par un changement dans les AQS, pas dans le code source du système. Les évolutions logicielles sont donc limitées, et la nécessité de flexibilité dans le temps repose sur les données.

3. Données et organisation : les avantages de l'approche orientée données

Les AQS ne sont qu'un exemple de données utilisées par le système pour piloter ses fonctions les plus critiques. Viennent s'ajouter aux AQS des ensembles de données de nature différente, qui sont versées à des fins de gestion du système. Dans leur ensemble, elles constituent l'information de référence de SPAR.

3.1. Une filière de référence

Dans la mesure où SPAR est un système conforme à l'OAIS, toute information de préservation doit être versée et stockée sous la forme d'un Paquet d'informations. À cet égard, le système utilise des paquets de référence de trois types distincts : contexte, formats et agents.

- L'information de **contexte** porte sur des ensembles d'objets : en font partie les paquets de filières et de chaînes, ces derniers contenant les AQS.
- Nous donnons également de l'information de représentation sur tout **format** pour lequel nous avons défini une stratégie de préservation et sur le suivi duquel nous sommes engagés. Cela peut être des normes comme TIFF 6.0, ou des profils BnF qui sont la restriction de ces formats, par exemple du TIFF non compressé en 24 bits avec une résolution de 300 dpi.

- Enfin, SPAR recueille des informations de référence sur les **agents** réalisant des opérations de préservation, qui peuvent être des humains (administrateur, expert de préservation), des logiciels (outils d'identification, de caractérisation et de validation) et des processus de SPAR (comme les processus de versement et de mise à jour d'un paquet). À l'avenir, nous avons l'intention d'utiliser ces paquets de référence pour décrire les environnements logiciels dans une perspective d'émulation.

Le regroupement d'informations communes à plusieurs objets numériques n'est que l'un des rôles des paquets de référence. Ils en améliorent également la maintenance : mettre à jour cette information essentielle n'oblige pas à mettre à jour tout paquet d'informations qui s'y réfère.

Comme nous l'avons expliqué plus haut pour les AQS, ces Paquets d'informations nous permettent de définir des paramètres du système par le biais de fichiers exploitables par une machine. Par exemple, le système peut vérifier que les images sont conformes à un profil spécifique de TIFF utilisé à la BnF (TIFF 6.0, 24 bits, résolution de 300 dpi, *watermarking* BnF, etc.) à chaque fois qu'est versé un paquet contenant des fichiers dont le format est identifié comme du TIFF. De cette manière, les données définissent et configurent les processus et non l'inverse. Cet aspect améliore le contrôle des processus du système par les utilisateurs non informaticiens. Enfin et surtout, les chargés de collections numériques et les experts de préservation ont la possibilité de récupérer un fichier d'exemple de format ou le code source d'un outil, assortis d'une description à destination d'un humain dans la documentation de tout Paquet d'informations portant sur un format ou un outil. Tout aspect des fonctionnalités du système ayant un impact sur les bibliothécaires est documenté dans SPAR.

En réalité, la dernière filière née dans SPAR, que nous n'avions pas prévue lors de l'étape de conception, a été dédiée à ces types d'objets que nous plaçons sous le terme d'information de référence. Toutes ces informations sur le contexte, les formats et les agents sont organisées non seulement sous la forme de Paquets d'information comprenant un manifeste METS, mais encore sur le modèle de données de SPAR, sous la forme d'une filière (la filière de référence) avec ses chaînes (contexte, format, agent) avec, pour chacune de ces chaînes, ses AQS propres. Nous avons finalement constitué une nouvelle collection : une collection d'informations de référence.

3.2. Un graphe global

Comme nous l'avons démontré plus haut, l'information de référence dans SPAR a été conçue dans l'idée de faciliter la manipulation de l'information par le système, en agrégeant toute information portant sur un ensemble d'objets similaires, ce qui permet ainsi d'éviter la redondance. En outre, cette information n'est pas seulement présente dans le système à des seules fins de documentation, mais est aussi utilisée par le système afin de traiter les objets versés.

Ce besoin ne pouvait être couvert qu'en utilisant un modèle de données prédisposé à lier les données de manière souple, et offrant des mécanismes normalisés d'encodage et d'interrogation, ainsi qu'une indépendance complète des choix d'implémentation. C'est pour ces raisons que nous avons choisi la norme RDF comme cadre principal de manipulation des données dans SPAR.

L'entrepôt de métadonnées, initialement encodé en XML, est ainsi transformé en RDF lors de son versement dans le module de Gestion de données. Le choix de RDF a été fait à l'issue d'une analyse de risques fondée sur les fonctionnalités souhaitées pour les principaux

entrepôts de métadonnées dans SPAR [3]. Le RDF (Resource Description Framework - Cadre de Description de Ressources) fournit un modèle de données extrêmement générique et versatile : toute information y est exprimée sous la forme de triplets, selon la syntaxe sujet/prédicat/objet. Il est sorti en tête de l'analyse en raison de son langage de requêtes très souple, SPARQL, de ses bonnes aptitudes en conversion des métadonnées XML existantes, et de son caractère réversible si l'on envisageait de changer de format de données à l'avenir.

RDF est nativement apte à gérer des données riches en liens, comme c'était le cas dans SPAR du fait de la factorisation de l'information de référence dans des paquets dédiés. L'utilisation d'XML est appropriée pour le stockage et l'échange de fichiers, et nous estimons particulièrement commode d'avoir un fichier METS rassemblant dans chaque paquet l'information qui lui correspond. Toutefois, même s'il existe des liens internes aux fichiers METS, et d'un fichier METS vers d'autres paquets, ces liens manquent de souplesse et sont extrêmement lourds à traiter par une machine.

Lorsque les données sont converties en RDF, chaque élément d'information présent dans un manifeste METS devient un triplet autonome, qui peut être compris et manipulé sans avoir à analyser l'information en XML qui le contextualise. En définitive, nous obtenons un réseau d'informations où chaque objet ou partie d'objet est identifié de manière unique afin que nous puissions faire des déclarations à leur sujet. Les déclarations provenant du manifeste METS de tel objet sont intégrées en continu avec l'information provenant de paquets de référence, de sorte que l'ensemble de l'information dans SPAR peut être manipulée de manière globale. Le fait que l'information soit stockée à l'origine dans des paquets distincts n'entraîne pas une baisse de performance pour le traitement des requêtes ou de modèles complexes pour les formuler. Le module Gestion de données contient le graphe global des informations nécessaire à la gestion de nos collections numériques sur la longue durée.

Les avantages de RDF énumérés plus haut s'avèrent particulièrement précieux concernant les questions de récupération de données. Les données sont maîtrisées, donc l'accès est maîtrisé : les mêmes concepts ont toujours le même nom. Les requêtes sont précises car elles sont faites depuis des points d'accès contrôlés et avec des données structurées. De plus, à l'inverse des technologies de bases de données relationnelles, il n'est pas obligatoire de connaître à l'avance le nom des catégories de données pour formuler une requête : leur nom peut être déduit de la manière dont les données sont structurées, par le biais de requêtes successives.

Voici quelques exemples de requêtes que nous pouvons formuler sur les contenus issus des collections de livres et lots d'images numérisés :

- Quels paquets comprennent des pages signalées comme contenant une table des matières, mais aucun fichier de table des matières en XML permettant une navigation dynamique dans le document ? La réponse à cette question aide à planifier la création rétrospective de tables des matières structurées.
- Combien de paquets ont été ingérés dans SPAR le mois dernier, combien de fichiers comprennent-ils, quels sont les formats de ces fichiers, et le taux de qualité de leur OCR ? Cette question classique montre que les données aident également les administrateurs à assurer le suivi du système.
- Dans notre chaîne de numérisation, quels paquets contiennent des fichiers de table des matières en HTML non valide ? L'invalidité d'un fichier HTML n'entrave pas forcément l'accès au document, mais le rend assurément plus difficile à préserver ; une telle requête aide les experts de préservation à planifier une régénération de fichiers HTML non valides.

Nous voyons de nombreux avantages à utiliser RDF pour gérer les données dans notre

entrepôt numérique sécurisé, mais nous devons aussi admettre que des problèmes subsistent, en rapport avec l'adoption d'une technologie relativement nouvelle. Premièrement, en regard d'autres technologies, peu de fournisseurs de logiciels existent pour les entrepôts RDF, et beaucoup d'adaptation et d'optimisation ont été nécessaires lors de l'implémentation. Actuellement, les performances de notre entrepôt RDF sont également inférieures à celles de bases de données relationnelles classiques. Même s'il ne s'agit pas d'une question de premier plan dans une optique de préservation, des temps de réponse rapides apportent un confort non négligeable aux chargés de collections numériques. En outre, des tests menés en 2008 ont montré que notre implémentation d'un entrepôt RDF atteindrait ses limites lorsque le volume de données approchera de 2 milliards de triplets — notons toutefois que les performances des technologies RDF s'améliorent régulièrement. Compte tenu du fait que la première chaîne d'objets à entrer dans SPAR correspond déjà à une quantité estimée à 1 milliard de triplets, nous savons que l'extensibilité sera un enjeu dans les années à venir.

3.3. Une fonction émergente: l'expert de préservation

Le besoin de formations *ad hoc*, du côté des informaticiens comme des bibliothécaires, a été un autre grand enjeu.

Du côté des informaticiens, les technologies du Web Sémantique étaient inédites à la BnF, et exigeaient une formation, d'abord pour l'équipe de préservation du numérique, puis pour leurs collaborateurs. Le suivi quotidien du module Gestion de données est également plus difficile à mener, car les compétences ou retours d'expérience de pairs sont encore rares.

Du côté des bibliothécaires, la question de la formation est encore plus importante, puisque SPAR n'est pas seulement destiné à l'usage des experts de préservation, mais aussi à celui des producteurs d'Objets numériques et des chargés de collections numériques. Il leur faut assimiler le modèle de données de SPAR afin de pouvoir formuler leurs besoins en informations. Dans l'idéal, toute personne ayant maille à partir avec des collections numériques devrait être en mesure de récupérer l'information dont elle a besoin depuis le module de Gestion de données, ce qui implique d'apprendre comment l'interroger en SPARQL.

En outre, l'absence de bonnes pratiques en modélisation RDF pour la préservation du numérique nous a forcés à bâtir « au fil de l'eau » le modèle de données de SPAR et les ontologies exprimant les métadonnées de préservation en RDF, en nous servant notre bon sens et de notre expérience professionnelle en matière de modélisation de données.

Cette phase d'« apprentissage sur le tas » a vu l'émergence d'une nouvelle fonction au sein du personnel de la BnF, celle d'expert de préservation. Les experts de préservation ont une formation de bibliothécaire mais ont acquis les connaissances techniques nécessaires à la compréhension des principales fonctions du système de préservation, et en particulier de son modèle de données. Ils jouent un rôle d'intermédiaire entre les informaticiens et les responsables de collections numériques ; ils participent à la négociation des filières et la définition des chaînes ; ils analysent les données et définissent de nouveaux modèles ; enfin et surtout, ils gèrent une partie importante de la filière de référence, devenant par là les chargés d'une nouvelle collection, celle qui a été créée pour les besoins du système.

Une fois que les experts de préservation ont acquis les compétences permettant de créer, modéliser et interroger les données, la conception orientée données du système leur a conféré un plus grand contrôle sur le fonctionnement de ce système et sur la manière dont il manipule les collections. Ainsi, dans une perspective à long terme, RDF avait de réels atouts

en termes d'organisation en ce qu'il permettait de séparer les questions techniques / informatiques de celles concernant les données / les bibliothécaires. Malgré la complexité de RDF et SPARQL dans un premier temps, ils apportent aux bibliothécaires une meilleure maîtrise de leurs données, ce qui, dans une approche orientée données, dénote aussi une meilleure maîtrise des processus du système.

Au bout du compte, nous espérons que le modèle de données de SPAR, et son utilisation des technologies RDF, permettra au personnel de la BnF en lien avec la préservation et la gestion de collections numériques de parler un langage commun, qui s'adaptera à différentes missions et différentes temporalités. Chaque individu en interaction avec l'Archive devra se référer au même modèle de données, en utilisant le même langage de requêtes, qu'il soit en train de planifier des opérations de préservation sur le long terme telles que des migrations, qu'il doive prendre des décisions à court terme, par exemple demander une nouvelle océrisation de certains documents, ou qu'il ait besoin des statistiques les plus à jour. En fin de compte, tous ces utilisateurs devront définir ensemble les évolutions nécessaires du modèle de données.

4. Conclusion

Si nous examinons SPAR aujourd'hui, nous voyons un système où les collections sont gérées de manière équilibrée par les administrateurs informaticiens et par les bibliothécaires, en utilisant comme une « lingua franca » le graphe global d'informations constitué par l'ensemble des données versées dans le système, soit à la fois les métadonnées sur les objets et les données de référence sur les collections. Ces données nous permettent d'assurer un suivi efficace de ces collections, tout en apportant aux responsables de collections une visibilité et une transparence qu'ils n'avaient que très peu auparavant.

Mais qu'en est-il de la préservation ? Aujourd'hui, les opérations de préservation dans SPAR se limitent à améliorer la qualité des données patrimoniales avant leur versement dans le système. Le module de planification de la préservation, dont le développement est prévu pour 2011, nous permettra de planifier, tester et mettre en œuvre ce type de stratégie pour des collections numériques de masse.

Ainsi, le principal enseignement donné par SPAR, c'est que la préservation ne revient pas à préserver des objets mais à gérer des collections. Qui souhaite gérer une collection doit avoir le contrôle des données. Nous avons besoin d'assurer ce contrôle en utilisant les moyens à notre disposition :

- des procédures organisationnelles, telles que les accords sur la qualité de service, qui permettent des accords entre les parties sur les opérations sur lesquelles s'engager ;
- des métadonnées, sous la forme initialement prévue avec l'utilisation de METS, mais également sous la forme plus concentrée de l'information de référence, qui peut, et doit, être factorisée et exploitée ;
- des normes, car elles sont une garantie de pérennité et de fiabilité des données ;
- une infrastructure technique, mais qui ne soit pas une boîte noire : c'est le système qui dépend des données, non les données qui sont modélisées pour s'adapter aux processus ;
- enfin, des ressources humaines, formées et mobilisées en connaissance de cause.

Les points mentionnés ci-dessus font tous partie d'une approche globale de gestion des risques car, dans un environnement où il est difficile de prévoir les évolutions, toute prise de décision est une tentative de maintenir autant que possible le niveau de risque au plus bas.

Une collection se définit en grande partie par la manière dont on gère son contenu, et en

particulier par le fait que les responsables de collection sont clairement identifiés. Il est vain de déclarer des politiques de préservation sans y associer une vision organisationnelle de la façon dont ces politiques vont être mises en œuvre sur des contenus. En réalité, à l'intérieur d'une filière donnée, les gestionnaires ne peuvent prendre de décision de traitement sans se fonder sur une connaissance du contenu, de sa communauté d'utilisateurs, de ses spécificités (rareté, fragilité, etc.), selon un modèle tout à fait similaire aux collections traditionnelles. Une décision de préservation n'est jamais uniquement technique.

SPAR n'est pas un coffre-fort sécurisé où les collections numériques seront entreposées sans qu'il soit besoin d'agir sur elles. Notre système constitue une architecture fiable pour garantir la préservation du train de bits, et constitue une aide à la prise de décision pour les opérations de préservation : le logiciel ne prend pas en charge par lui-même ces opérations, mais facilite leur définition et leur mise en œuvre. Nous avons bâti un système de gestion de collections, un outil pour assurer le suivi une collection numérique de façon fiable et durable, à commencer par les activités quotidiennes telles que la gestion et l'amélioration de la qualité des collections. La collection numérique sera-t-elle en mesure de durer plusieurs décennies ? Plusieurs siècles ? Aujourd'hui plus que jamais, cette question paraît moins une question technique qu'un véritable défi pour les bibliothèques.

BIBLIOGRAPHIE

[1] Bermès, E. et al. "Digital preservation at the National Library of France: a technical and organizational overview", *World Library And Information Congress: 74th IFLA General Conference And Council*, 2008. En ligne : http://archive.ifla.org/IV/ifla74/papers/084-Bermes_Carbone_Ledoux_Lupovici-en.pdf [Dernière consultation 2010-05-04].

[2] Bermès, E. and Fauduet, L. "The Human Face of Digital Preservation: Organizational and Staff Challenges, and Initiatives at the Bibliothèque nationale de France", *Proceedings of iPRES 2009: the Sixth International Conference on Preservation of Digital Objects*, 2009. En ligne : http://escholarship.org/uc/item/6bt4v3zs [Dernière consultation 2010-04-20].

[3] Bermès, E. and Poupeau, G. "Semantic Web technologies for digital preservation: the SPAR project", *Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference (ISWC2008)*, 2008. En ligne : http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-401/iswc2008pd_submission_14.pdf [Dernière consultation 2010-05-04].

[4] Farquhar, A. "Implementing Metadata that Guides Digital Preservation Services." *Proceedings of iPRES 2009: the Sixth International Conference on Preservation of Digital Objects*, 2009. En ligne : http://www.escholarship.org/uc/item/12p437bd [Dernière consultation 2010-04-20].

[5] Guenther R., Wolfe R., "Integrating Metadata Standards to Support Long-Term Preservation of Digital Assets: Developing Best Practices for Expressing Preservation Metadata in a Container Format." *Proceedings of iPRES 2009: the Sixth International Conference on Preservation of Digital Objects*, 2009. En ligne : http://www.escholarship.org/uc/item/0s38n5w4 [Dernière consultation 2010-04-20].

[6] Mazocchi, S., "Data First vs. Structure First", *Stefano's Linotype*, July 28th, 2005. En ligne : http://www.betaversion.org/~stefano/linotype/news/93/ [Dernière consultation 2010-06-01].