Géraldine Barron et Pauline Le Goff-Janton (dir.)

Intégrer des ressources numériques dans les collections

Presses de l'enssib

Gestion de la conservation des collections numéiques

Thierry Claerr et Jean-François Moufflet

DOI : 10.4000/books.pressesenssib.11738
Éditeur : Presses de l'enssib
Lieu d'édition : Villeurbanne
Année d'édition : 2014
Date de mise en ligne : 4 mai 2020
Collection : La Boîte à outils
ISBN électronique : 9782375460573

Gestion de la conservation des collections numéiques

Thierry Claerr et Jean-François Moufflet

1 L'objectif de cette contribution est de mettre en lumière quelques solutions et méthodes adaptées aux collections numériques acquises ou produites par les bibliothèques, pouvant servir de point de départ et de comparaison à toute réflexion sur la gestion de la conservation de ce patrimoine informationnel et documentaire fragile. Il vise à répondre à une préoccupation grandissante de la part de ces institutions de garantir un *continuum* des collections.

2 Avec le développement de la numérisation et du livre numérique natif (ebooks, revues en ligne), les institutions qui possèdent un patrimoine documentaire important sont également concernées par l'archivage de leurs collections numérisées.

3 La problématique est double :

- la production documentaire informatique ne cesse d'augmenter et prend diverses formes ;
- l'information concernée (produite ou acquise) est très volatile et peut se détériorer rapidement si certains moyens ne sont pas mis en œuvre.

4 Il convient de garder à l'esprit que la conservation du numérique se fonde sur la notion de cycle de vie des données. Elle peut s'envisager sur plusieurs paliers :

- la bonne gestion documentaire au quotidien : en préalable à la conservation du numérique, il faut mettre en place une politique de gestion qualitative des données quotidiennement produites ou reçues par la bibliothèque. Cela débouche concrètement sur une identification claire des contenus à préserver en priorité et des procédures, en s'assurant qu'ils sont traités dans des conditions qui en faciliteront la consultation et la conservation ultérieures.
- la conservation à court terme : une fois produits ou reçus, les contenus numériques sont conservés par la bibliothèque qui en fait une utilisation immédiate. Une partie de ces contenus peuvent être à court terme éliminés (documents ne devant pas être conservés au terme de la licence d'utilisation ou éliminables rapidement une fois leur utilité administrative échue) ; une autre partie est appelée d'ores-et-déjà à être pérennisée sur le long terme (collections numériques ou numérisées). Pendant cette phase, la bibliothèque peut mettre en place, par elle-même, quelques mesures pour garantir la bonne conservation

de ces contenus dans l'immédiat. Ce chapitre insistera particulièrement sur cet aspect en donnant quelques conseils pratiques en la matière.

- la conservation à long terme, ou conservation permanente : ce n'est pas la même chose de conserver des contenus sur une plage de cinq ans et d'en préserver d'autres sans limite de durée. Dans ce cas-là, on est clairement sur le secteur de la pérennisation et il faut se reposer sur des procédures et des outils d'une grande complexité et qui ne sont réellement maîtrisés que par un petit nombre d'acteurs. Ainsi pour les contenus à préserver définitivement, il faudra nécessairement se poser la question de faire appel à un partenaire spécialisé à qui l'on confiera une copie de nos données : cette contribution présentera quelques solutions tournées vers la pérennisation.

5 Avant de se lancer dans une politique de préservation du numérique, il convient de commencer par quelques questions simples :

- quels contenus numériques sont à préserver en priorité ? Pour y répondre, cela suppose de recenser l'ensemble des ressources numériques produites ou reçues par la bibliothèque, d'en connaître les droits reconnus à l'institution sur ces ressources, d'évaluer la nécessité de les conserver ou non, et si oui, la durée de cette conservation (par exemple, un court terme pour des documents relevant davantage du fonctionnement administratif et, à l'inverse, un très long terme pour les ressources numériques patrimoniales) ;
- a-t-on pour mission de conserver ces contenus ? Est-il nécessaire de conserver par soimême des contenus numériques qui sont déjà préservés ou qui, réglementairement, doivent être préservés par d'autres ? Par exemple, les revues électroniques en ligne, auxquelles l'institution est abonnée et auxquelles les usagers accèdent en ligne, ne sont sans doute pas à conserver, étant donné que cela est déjà assuré par l'éditeur de la revue et, parfois, par les bibliothèques nationales des pays d'origine des éditeurs – c'est ainsi que la Bibliothèque royale des Pays-Bas conserve les collections d'Elsevier. Les documents d'archives électroniques de l'institution ne sont pas nécessairement à faire entrer dans le périmètre de la conservation permanente, étant donné que cela peut être assuré par le service public d'archives dont elle dépend ;
- peut-on tout faire par soi-même ? La conservation permanente du numérique fait appel à des compétences extrêmement poussées et repose sur des outils complexes et coûteux. La plupart des institutions n'ont pas les moyens nécessaires ni les missions pour mettre en place et gérer un véritable système d'archivage électronique (SAE) tourné vers la pérennisation. Dans ce contexte, la tendance est donc plutôt de mutualiser les efforts entre plusieurs partenaires, qui partagent les frais et les outils de conservation, ou de passer une convention avec un partenaire spécialisé. Certaines institutions (Persée, bibliothèque Cujas…) se sont tournées à juste titre vers les offres de tiersarchivage, comme celle proposée par le Centre informatique national de l'enseignement supérieur (CINES) et la BnF.

Panorama des ressources numériques en bibliothèque

Livres numériques et e-publications (ressources numériques sur abonnement, dont bases de données)

6 Ces données sont complexes et il convient de prendre en compte le plus possible en amont leur statut et la technicité de ces collections numériques. Elles doivent faire l'objet d'un traitement spécifique qu'elles soient conservées ou détruites. Les

documents à archiver sont en grande partie issus du monde éditorial et conservés au format EPUB ou PDF. La conservation de ces collections peut s'avérer complexe :

- pour faire face au risque de piratage, il est fréquent que les éditeurs équipent ces fichiers de mesures techniques de protection (MTP), qui interdisent la copie numérique ou en limitent le nombre ;
- de nombreuses bibliothèques n'acquièrent pas physiquement les fichiers, mais achètent simplement un accès (temporaire ou pérenne) à des collections hébergées sur les sites des éditeurs.

7 Une institution souhaitant garantir elle-même la conservation des e-publications qu'elle acquiert doit donc s'assurer qu'elle est en mesure d'obtenir de l'éditeur les fichiers, et qu'ils soient dépourvus de MTP. C'est le sens de la démarche qu'a entrepris la BnF pour son dépôt légal des livres numériques[1].

Ressources issues de la numérisation

8 La numérisation permet de créer un support de substitution numérique pour la consultation de ressources documentaires très demandées ou dont le support d'origine peut s'avérer trop fragile pour être communiqué directement aux lecteurs. Parfois même, la numérisation est un transfert de support : c'est particulièrement le cas pour les ressources informationnelles dont le support d'origine est voué à une dégradation irrémédiable, comme les supports audiovisuels et sonores analogiques par exemple. Dans ce cas, l'effort de conservation de l'information portera donc sur le substitut numérique en tant que tel.

9 De manière plus générale, compte tenu du coût et des efforts que demandent les opérations de numérisation, il importe de conserver correctement les données qui en sont issues : cela revient à pérenniser cet investissement. Rien ne serait plus dommageable pour les collections que de recommencer une opération de numérisation dans l'hypothèse où les fichiers issus de la numérisation auraient été mal gérés et conservés, sous prétexte qu'il ne s'agit « que » de substituts numériques.

Production documentaire numérique

10 Comme toute autre administration, une bibliothèque produit dans le cadre de ses activités des documents qui en sont le reflet. Une partie de ces documents ont une utilité administrative ou juridique. La préservation numérique d'une partie d'entre eux peut se poser : par exemple, la conservation du fichier de récolement conçu lors d'une opération de numérisation est précieuse pour garantir la bonne gestion et la bonne conservation des ressources issues de la numérisation.

11 La sélection, l'évaluation et la conservation des documents d'activité d'une bibliothèque relève davantage d'une démarche archivistique et il convient de définir clairement à qui revient ce mandat.

Les risques posés par le support numérique

12 L'accessibilité de l'information numérique est fragile car elle repose sur des couches technologiques condamnées par essence à l'obsolescence.

13 Les principales menaces qui reposent autour des données sont les suivantes :

- l'inaccessibilité physique ou obsolescence du couple matériel/support ;
- l'inaccessibilité logique ou obsolescence du couple format de données/logiciel ;
- l'inaccessibilité intellectuelle ou l'absence de métadonnées.

14 Les métadonnées sont des informations complémentaires sur les données et il en existe plusieurs types : *descriptives*, pour définir intellectuellement un contenu (titre, auteur, éditeur, date de publication d'un ouvrage par exemple) ; *administratives*, pour les gérer (droits d'accès, durées de conservation) ; *techniques*, pour indiquer au système les dépendances nécessaires à leur exploitation (format, version de format, logiciel de création, date de création, etc.) ; et enfin *sémantiques*, lorsqu'il s'agit d'expliciter et d'apporter des informations supplémentaires ou contextuelles sur le contenu d'information primaire.

15 L'utilité des métadonnées est multiple : elles servent à retrouver les fichiers, à les manipuler, à les décoder informatiquement mais aussi intellectuellement. Une information insuffisamment qualifiée est une information perdue, avec le risque de ne pas comprendre d'où vient ce fichier, ni savoir qui l'a produit et dans quel but.

16 En résumé, la question de l'environnement de création des données peut devenir redoutable avec le temps, notamment dès lors que l'on essaie de les exploiter dans un autre environnement ou système d'information, ou auprès d'une autre communauté que celle qui les a produites et utilisées.

Conserver et pérenniser ses données

Polysémie et finalités de la conservation du numérique

17 Une politique de conservation des ressources numériques a pour but d'atténuer les risques précédemment évoqués en vue de garantir l'accès à l'information et sa qualité sur une durée plus ou moins longue. Il est primordial avant tout de définir la finalité de cette conservation. Pour simplifier, les acteurs et éditeurs se rattachent généralement à l'un ou l'autre de ces deux périmètres, de manière plus ou moins consciente, et qui recouvrent des besoins différents :

- conservation ou archivage à valeur probante : il s'agit de l'archivage de documents numériques ayant une forte valeur juridique et engageant leurs producteurs (documents signés électroniquement, dossiers de marchés publics électroniques...). Cette conservation met l'accent sur l'*intégrité* (non altération) des données confiées à un SAE et la *traçabilité* des actions effectuées sur les données au sein du système. *L'objectif premier est de respecter les exigences juridiques qui portent sur des écrits numériques pouvant constituer une preuve* ;
- conservation à long terme ou archivage patrimonial : cette conservation touche des objets numériques dont on estime qu'ils revêtent un grand intérêt scientifique, historique, patrimonial et qui sont censés être préservés longtemps si ce n'est indéfiniment dans le temps. Les solutions qui ont été mises au point accordent ainsi une très grande importance au choix des formats de données et à la qualité des métadonnées, deux composantes essentielles pour garantir la lisibilité et l'accessibilité des contenus dans le temps. On parle plutôt de *préservation* ou de *pérennisation*. *L'objectif premier est de donner accès à des contenus scientifiques et patrimoniaux à une communauté élargie d'utilisateurs.*

Focus

Il ne faut pas confondre conservation et sauvegarde de collections numériques courantes : la sauvegarde sécurisée ne fournit qu'une copie de sécurité d'un ensemble d'informations numériques. Le but n'est pas de documenter les informations à préserver mais d'avoir un espace « miroir » permettant seulement de se prémunir des pertes liées à des accidents au sein de l'infrastructure (pannes de matériel, dégâts...). Une sauvegarde ne peut donc concerner que des documents à conserver sur du court terme et n'est qu'un des aspects de la conservation.

Une gestion qualitative de ses données

18 On fait souvent le lien entre politique d'archivage numérique et *records management* puisque cette discipline anglo-saxonne définit une gestion documentaire qualitative de l'information, qui repose notamment sur le repérage et l'évaluation des contenus cruciaux d'une institution, l'organisation intellectuelle de ces ressources, la qualification de ces contenus (c'est-à-dire, dans un contexte numérique, l'enrichissement homogène des métadonnées), les droits d'accès aux informations, leurs durées de conservation. Dans le monde des bibliothèques, cette politique d'archivage numérique ainsi définie constitue dès lors une branche de la politique documentaire de l'établissement.

Évaluer la production documentaire

19 La bibliothèque devra commencer par une enquête sur les contenus numériques qu'elle produit ou reçoit et éventuellement les documents d'activité. Il s'agira d'estimer, pour chacun d'eux, lesquels constituent des informations importantes et pour lesquelles l'institution a un mandat de conservation. Un archiviste peut apporter son expertise, en définissant notamment des durées de conservation pour les documents ayant un statut d'archives publiques.

Choisir des formats de données pertinents

20 La migration de format reste une solution : des contenus produits dans un format propriétaire peuvent être transformés dans un format standardisé, avec toutefois le risque d'amoindrir les fonctionnalités voire de perdre une partie des informations.

21 Pour les contenus numériques à préserver en priorité, il faut se reposer sur des formats sur lesquels on a une certaine maîtrise : formats largement utilisés, non propriétaires, documentés et interopérables (non liés à un logiciel unique et/ou fermé ou à un système d'exploitation particulier).

22 En prenant en compte ces critères, la communauté des bibliothèques s'appuie sur une liste restreinte de formats :

- pour les ebooks : le format EPUB apparaît comme le standard émergent. La bibliothèque nationale des Pays-Bas l'a d'ailleurs étudié pour évaluer son caractère pérenne[2]. Le risque posé par le format peut être la possible présence de DRM limitant son utilisation. Le format PDF peut constituer aussi une bonne alternative ;

- pour les fichiers images : TIFF, JFIF/JPEG et PNG (bien que plus marginalement utilisé) pour les fichiers haute qualité. JPEG 2000 il présente un ratio entre la qualité de compression et le volume de données beaucoup plus intéressant mais c'est un format complexe ;
- pour l'océrisation : le format XML-ALTO, bien qu'encore assez peu utilisé, constitue un gage certain de pérennité en tant que format documenté et standard ;
- pour les documents bureautiques : le format PDF/A (versions PDF/A-2 ou PDF/A-3) ;
- pour les ressources informationnelles produites par la bibliothèque : parmi les données cruciales d'une bibliothèque figure bien entendu le catalogue des collections. Il est important de disposer d'un module d'export des données catalographiques : celles-ci doivent pouvoir être récupérées dans un format standardisé (comme UNIMARC).

Gérer correctement ses fichiers et les documenter

- disposer d'un plan de nommage rigoureux des fichiers : les choix des règles de nommage doivent être documentés en partant de la norme ISO 9660, niveau 2, et cette documentation préservée ;
- renseigner les métadonnées lors de la création des objets ;
- structurer l'arborescence de stockage ;
- conserver la documentation afférente aux contenus.

23 Toutes ces tâches peuvent être encadrées par un logiciel de gestion électronique de documents (GED) qui présente le grand intérêt d'assister l'utilisateur dans la création des contenus. Une GED peut être représentée comme une couche de gestion unifiée de contenus mixtes et initialement produits avec des outils différents.

Surveiller les supports de stockage et les faire migrer[3]

Choix des supports de stockage en fonction des usages

24 Une politique de conservation doit davantage isoler les contenus numériques importants de la production courante. Il faut disposer d'une copie des données qu'il faut le moins possible manipuler.

25 Les préconisations minimales suivantes sont à appliquer :

- le stockage sur serveurs doit absolument être redondant ;
- il faut utiliser différents supports de stockage.

La conservation des supports amovibles

26 Une bibliothèque amenée à conserver ses données sur support amovible doit faire attention :

- à la qualité du support de stockage[4] ;
- à l'environnement de conservation du support de stockage ;
- aux procédures de manipulation et de maintenance des supports.

27 La qualité des supports d'enregistrement est évidemment essentielle pour pouvoir un jour transmettre ou confier des données à un partenaire de pérennisation.

Pérenniser les données sur un long temps : solutions techniques et modalités existantes

28 La pérennisation des données est une activité complexe et coûteuse. Pour cette raison, seul un faible nombre d'acteurs en maîtrise réellement les compétences indispensables.

Les grands standards du système de pérennisation

29 Les acteurs de la pérennisation s'appuient sur le modèle *Open Archival Information System* (OAIS), norme issue du monde de l'aérospatiale[5].

30 L'OAIS définit les grandes fonctions d'un système d'archivage ouvert, censé effectuer des traitements sur les données que les producteurs lui confient : contrôle des données et métadonnées soumises au système, qualification et enrichissement des métadonnées, enregistrement des données sur une architecture de stockage, gestion des opérations sur les données à partir des métadonnées, recherche et communication des contenus demandés, planification de la pérennisation (procédures de migration des supports de stockage et des formats, enrichissement des métadonnées au fil du temps) et administration du système avec une gestion fine des droits d'accès. De même l'OAIS définit les grandes catégories de métadonnées nécessaires pour préserver dans le temps les contenus. S'appuyant sur une liste limitée de formats[6], et dont la documentation est normalisée à l'ISO, les acteurs de la pérennisation accordent une importance fondamentale aux métadonnées. Le format de métadonnées Preservation Metadata Implementation Strategies (PREMIS) fournit ainsi une liste très complète des informations nécessaires à la préservation des contenus numériques. Toutefois, dans le système de pérennisation, d'autres formats de métadonnées, plus documentaires, comme UNIMARC ou Dublin Core, peuvent côtoyer les métadonnées PREMIS, le format METS permettant d'encapsuler des métadonnées afférentes à un même objet numérique mais relevant de différents schémas.

Le traitement des données dans le système de pérennisation

31 Lorsqu'un producteur confie à un prestataire ses données, le système va effectuer un ensemble de contrôles sur les objets numériques.

32 Identification du format des données entrantes :

- validation du formatage des données ;
- caractérisation des données (récupération et enrichissement des métadonnées) ;
- migration des données (préalable ou *a posteriori*).

Solutions techniques existantes sur le marché

33 Des plates-formes de dépôt, d'archivage et de diffusion de publications électroniques, libres et paramétrables (par exemple FEDORA, E-PRINT et DSPACE) permettent la constitution d'une base de métadonnées associées et la gestion-distribution à long terme des documents.

Encadré Principales institutions de pérennisation

Au niveau national, différents acteurs ont mis en œuvre des solutions afin de gérer leur patrimoine informationnel numérique, notamment les données documentaires :

- le CINES se positionne à la croisée des différents types d'archivage à la fois scientifique, patrimonial et administratif. La plate-forme d'archivage au CINES (PAC) propose depuis 2006 des services de tiers-archivage pérenne à destination de l'ensemble de la communauté enseignement supérieur et recherche* ;

- la BnF a mis en place un système de préservation et d'archivage réparti (SPAR) pour l'archivage stocké sur bandes magnétiques, de ses collections numérisées, des publications électroniques et du dépôt légal de l'internet. Dans une volonté de mutualisation des expertises et des coûts, la BnF a ouvert son système à d'autres partenaires ou institutions offrant ainsi un service de « tiers archiveur » du patrimoine numérique.

Outre la formation du bibliothécaire/documentaliste chargé de la politique documentaire et des collections devenues hybrides, la gestion de la conservation des ressources numériques en bibliothèque requiert un ensemble de compétences très diverses et techniques :

- approche juridique : maîtrise de la réglementation en vigueur dans le domaine de la propriété intellectuelle ;

- approche archivistique : expertise en gestion de l'information et du cycle de vie du document ;

- approche informatique : connaissance des risques liés à l'environnement numérique.

* < http://www.cines.fr > et *La Gazette du CINES*, dossier spécial « Archivage numérique pérenne », février 2013, notamment Olivier Rouchon, « Le service d'archivage à long terme du CINES et la plate-forme PAC », pp. 10-15.

** < http://www.bnf.fr/fr/professionnels/spar_systeme_preservation_numerique.html >.

NOTES

1. Voir l'intervention de Sophie Derrot et de Clément Oury sur le sujet lors de la conférence satellite IFLA, été 2014.

2. < http://www.openplanetsfoundation.org/system/files/epubForArchivalPreservation20072012ExternalDistribution.pdf >.

3. Pour un panorama complet des supports et des stratégies de stockage, voir le chapitre de Laurent Duplouy dans *L'archivage numérique à long terme. Les débuts de la maturité ?*, Paris, La Documentation française, 2009, pp. 81-102.

4. On pourra consulter des études et recommandations sur le choix des supports optiques sur le site des Archives de France < http://www.archivesdefrance.culture.gouv.fr/>.

5. Pour un développement complet sur l'OAIS, voir Jean-François Moufflet et Sébastien Peyrard, « Préserver ses collections numériques », in Thierry Claerr et Isabelle Westeel (dir), *Manuel de constitution de bibliothèques numériques*, Paris, Éditions du Cercle de la Librairie, 2013, pp. 307-382.

6. Voir, par exemple, les listes de formats retenus par la BnF ou le CINES.

AUTEURS

THIERRY CLAERR

Chef du bureau de la lecture publique, Service du livre et de la lecture, ministère de la Culture et de la Communication (Paris)

JEAN-FRANÇOIS MOUFFLET

Conservateur du patrimoine, adjoint au directeur des études du département des conservateurs, chargé de la formation initiale Institut national du patrimoine (Paris)