La reproduction de sauvegarde des documents patrimoniaux

Partie 2 – La préservation numérique

Stéphane REECHT
Experte préservation numérique
Bibliothèque nationale de France
Département de la conservation

BnF | Bibliothèque nationale de France

Assurer la pérennité des données produites par numérisation d’objets physiques ou par collecte de documents nés numériques suppose d’aller au-delà de la question du stockage. Il faut avoir en permanence à l’esprit qu’une suite de bits, c’est-à-dire de 0 et de 1, parfaitement conservés d’un point de vue physique mais sans un minimum d’informations documentant le contenu intellectuel, la structure, l’historique et les caractéristiques techniques de ces données, perd très rapidement son intelligibilité et sa valeur, et peut être considérée comme perdue à une échéance qui se compte sur une échelle d’années voire de mois.

Introduction

Au préalable de tout travail visant à la pérennisation de données numériques, il convient de se poser les questions suivantes :

- A-t-on bien la mission de les conserver ? N’y a-t-il pas une autre institution qui est en charge de la conservation de ces mêmes données ?
- Faut-il conserver toutes les données ? Dans le cadre d’un projet de numérisation par exemple, faut-il s’intéresser aussi à la conservation à long terme des images de consultation en JPEG ou en PDF, ou bien se concentre-t-on sur les masters au format TIFF ou JPEG2000 ?
- Quel niveau de responsabilité mon établissement est-il prêt à prendre sur leur conservation ? Est-il voué à assurer seul la pérennité de ces données, ou bien l’effort qu’il fait aujourd’hui sera-t-il dans un futur proche pris en charge par un autre établissement aux moyens supérieurs ?
- Quels risques prend-on à ne rien faire ?

1. Que conserver ?

Un objet numérique ne peut prétendre à être pérennisé dans toutes ses dimensions (physique, technique et intellectuelle) sans un certain nombre d'informations, qui doivent dans la mesure du possible être conservées au plus proche de l'objet, et selon les mêmes principes de stockage (cf. partie 2). Il s'agit ainsi de s'assurer de la conservation de métadonnées permettant de gérer les documents sur le long terme.

Ces informations sont de plusieurs types :

- **Information d'intégrité** : voir 3.1.1.
- **Information d'identification** : Il s'agit d'attribuer à chaque document numérique un identifiant unique et si possible durable. Celui-ci peut être repris dans le nom du ou des fichiers constituant le document numérique, mais ne doit pas reposer sur leur nommage. En d'autres termes, il ne faut pas faire d'un élément aussi facilement modifiable qu'un nom de fichier le seul endroit où retrouver l'identifiant du document numérique. Cet identifiant doit être référencé par ailleurs (SIGB, système de gestion de bibliothèque numérique, système de gestion d'archives, fichier tableur…). Certains systèmes d'identifiants sont pensés pour assurer un accès sur le long terme, comme ARK[1], DOI[2], PURL[3] ou Handle[4].
- **Lien à l'original** dans le cadre d'une numérisation, afin de savoir de quel objet le document numérique est la reproduction. Il peut s'agir du numéro unique dans le système où est référencé l'objet physique (SIGB, système de gestion d'archives…). Idéalement, cette information est conservée en tant que métadonnée avec le document lui-même.
- **Description** : une description sommaire du contenu intellectuel du document numérique (titre, auteur, date, type de document…) sera utile pour avoir une idée de ce que l'on manipule, sans faire appel à une base extérieure. Il y a mille manières d'exprimer ces informations, mais, à des fins d'interopérabilité, le schéma Dublin Core est à recommander.
- **Information de structure** : Il s'agit de permettre aux humains comme aux machines de savoir comment lire les fichiers afin d'appréhender correctement le document. Par exemple, l'association entre un fichier image issu de la numérisation d'une page et le fichier texte issu de la reconnaissance optique de caractères de la même page. Si cette information doit être donnée au cas par cas, le schéma METS fournit un bon moyen d'exprimer de telles associations. Si cette information peut être exprimée de manière générale pour un ensemble de documents, elle peut être recueillie dans un fichier de documentation unique, à conserver précieusement lui aussi ; il faut alors s'assurer que chaque document numérique porte la référence à cette documentation.
- **Informations techniques** (application de production, réglage de l'appareil de numérisation, caractéristiques propres au format…). Si un profil d'application a été défini avant la production, celui-ci doit être conservé, dans ses versions successives éventuelles. En outre, dans le cadre de la numérisation, il est possible de faire inclure des métadonnées techniques dans le fichier image, en utilisant les tags TIFF pour le TIFF, JFIF pour le JPEG, et XMP pour le JPEG2000.
- **Contexte** : Toute information sur le contexte de production ou de collecte (par qui, pourquoi, dans quel cadre juridique…) aidera dans le futur à la compréhension de la collection et à sa gestion. Les cahiers des charges, procédures, etc., documentant les choix de production rentrent dans cette catégorie.
- **Historique** : Il s'agit d'identifier les actions qu'ont subies les documents numériques avant leur prise en charge au titre de la conservation, ainsi que les acteurs de ces actions : production, transformation, réception… Cela permettra d'assurer la traçabilité des opérations subies et aidera à garantir l'authenticité des données conservées. Dans le cadre d'une chaîne de traitement automatisée, le format de métadonnées PREMIS[5] est particulièrement adapté à l'expression de ces informations ; sinon une documentation *ad hoc* doit être tenue, faisant état non du cadre de l'opération (point précédent) mais de ce qui s'est réellement passé (dates et acteurs de création ou de collecte, de transformation, de réception… pour chaque document, tenues au minimum dans un fichier tabulé). Une partie de ces informations peut aussi être incluse dans les fichiers eux-mêmes (cf. ci-dessus point sur les informations techniques).

Toutes ces informations sont à conserver pour chaque objet numérique à préserver, avec l'objet lui-même, dans un ou plusieurs fichiers de métadonnées qui constituent en quelque sorte la carte d'identité du document. L'ensemble forme **un paquet d'informations[6]**. À l'exception des informations contenues dans des documents autonomes (procédures, cahiers des charges, contrats d'acquisition…), dont il convient de ne conserver que la référence au niveau du paquet à préserver, et ce a fortiori s'ils sont valables pour plusieurs paquets.

Les schémas idoines pour exprimer ces métadonnées sont METS et PREMIS[7], basés sur XML. Si l'on ne dispose pas des moyens de générer ou traiter du XML, il est recommandé au moins d'exprimer ces informations de manière structurée et dans un format favorisant l'interopérabilité (TXT, CSV, TSV, JSON…). Dans tous les cas, on veillera à définir un encodage des caractères en UTF-8. Il convient également de documenter les choix faits pour la structuration des fichiers de métadonnées et de conserver cette documentation, afin de préserver leur intelligibilité.

[1] https://www.bnf.fr/fr/lidentifiant-ark-archival-resource-key
[2] https://www.doi.org/
[3] http://www.oclc.org/research/themes/data-science/purl.html
[4] https://www.handle.net/

[5] https://www.loc.gov/standards/premis/ ; http://www.bnf.fr/fr/professionnels/formats_catalogage/a.f_premis.html
[6] Pour une définition complète du concept de « paquet d'informations », voir la norme OAIS (https://public.ccsds.org/pubs/650x0m2.pdf).
[7] https://www.bnf.fr/fr/premis-preservation-metadata-implementation-strategies

2. Où conserver ?

Le stockage en interne

Si l'établissement prend en charge lui-même le stockage des documents numériques, le premier choix à faire concerne le support. Dans une logique de maîtrise des coûts, on préfèrera :

- Le **disque dur externe** pour des volumes restreints.
- Le stockage sur **disques en réseau** (baie de stockage), pour des volumes plus importants avec des besoins d'accès rapide au master numérique (par exemple dans le cas où l'on ne dispose pas d'un espace de stockage dédié à la diffusion, cf. partie 2 de la fiche consacrée à la numérisation). Le coût peut être important, mais une mutualisation est bien sûr possible, à condition de pouvoir effectuer des contrôles d'intégrité réguliers (cf. infra).
- Le stockage sur **bande magnétique de format LTO**, pour des volumes importants sans besoin d'accès rapide au master. L'essentiel du coût correspondant à celui de l'appareil de lecture-écriture, il faut prêter une attention particulière aux conditions de garantie et de maintenance de celui-ci.

On évitera l'utilisation du CD-ROM. Pour des collections déjà présentes sur CR-ROM, voir pour leur entretien la fiche *La Conservation des documents audiovisuels*. Prévoir une trajectoire de migration vers un des supports ci-dessus, ou vers le *cloud* (cf. infra).

Quel que soit le medium de stockage choisi, il est indispensable de répliquer les données numériques en deux ou, mieux, trois exemplaires, et de conserver au moins un exemplaire à distance des autres, dans un bâtiment distinct si possible assez éloigné pour ne pas être exposé aux mêmes risques (inondation, incendie, explosion…). Tous les exemplaires doivent absolument être conservés dans un local sécurisé contre l'incendie et l'intrusion humaine, et ayant une température et une hygrométrie modérées et stables (cf. fiche *La Conservation des documents audiovisuels*); en outre, une attention particulière doit être portée à la sécurisation de l'alimentation électrique dans le cas de stockage sur disque en ligne, car une coupure électrique inopinée peut causer des pertes de données irréparables.

Il est souhaitable que les copies ne soient pas toutes conservées sur des supports de même modèle, pour éviter la dépendance à un constructeur ; par exemple une copie peut être conservée sur disque dur (sur un site) et une autre sur bande magnétique (sur un autre site). Mais il ne faut pas pour autant trop diversifier les supports, sinon cela augmente le coût de maintenance.

Stockage en externe et tiers-archivage

Si l'établissement souhaite externaliser le stockage, il convient de s'assurer que le prestataire offre des garanties suffisantes en termes de localisation (territoire national pour la France), d'accès aux données, et de récupération de celles-ci en cas de cessation d'activité. La liste des prestataires agréés par le Service interministériel des Archives de France pour la conservation d'archives courantes et intermédiaires sur support numérique offre un large choix[8]. Cet agrément est d'ailleurs obligatoire dans le cas de documents qui ont le statut d'archives publiques.

Dans tous les cas, il faut s'assurer que le prestataire de stockage procède bien à une **vérification d'intégrité** régulière par calcul de l'empreinte numérique (cf. infra). Les principaux services de stockage dans le *cloud* (Oracle, Amazon, OVH…) ne communiquant pas sur ce sujet, il est difficile de savoir s'ils y procèdent réellement.

[8] https://francearchives.fr/fr/article/26287437

3. Comment conserver ?

Pérenniser des données numériques suppose un certain nombre d'actions, à effectuer de manière régulière ou non.

Les contrôles

S'assurer de l'intégrité des fichiers

Dès réception des données, il s'agit de vérifier que les enregistrements informatiques n'ont pas subi de dégradation lors du transfert (passage d'un médium de stockage à un autre, par opération de copie manuelle ou par transmission FTP). On demandera ainsi que **l'empreinte numérique** de chaque fichier soit fournie en même temps dans un fichier texte par le producteur de données, afin de la recalculer à réception, puis lors de chaque transfert d'un support de stockage à un autre, et même en l'absence de transfert de support une fois par an au moins. L'algorithme le plus commode à utiliser pour produire et contrôler les empreintes numériques est le MD5, mais pour des données sensibles on préfèrera un algorithme plus robuste comme SHA-1 ou l'un de la famille SHA-2[9].

Si l'empreinte n'est pas fournie avant réception, il convient de la générer avant toute autre opération, afin de garantir que les données reçues ne sont pas modifiées indûment. La modification du nom de fichier et de son extension est la seule modification qui ne provoque pas de changement d'empreinte.

S'assurer de la conformité des fichiers.

Utiliser des outils de **contrôle de conformité** des fichiers vis-à-vis des spécifications du format, et **d'extraction de métadonnées techniques**. Ces dernières permettent d'effectuer des contrôles supplémentaires en fonction d'un cahier des charges particulier ; par exemple, dans le cadre de la numérisation, on pourra ainsi s'assurer que les fichiers livrés respectent la résolution, la profondeur de codage, la compression et la colorimétrie demandées, et que sont bien présents la définition et tel ou tel tag de métadonnée (cf. *Numérisation et conservation*, partie 3). Les logiciels spécialisés dans ces actions peuvent la plupart du temps être installés à la fois en mode serveur (pour des traitements de masse) ou en mode local. Des services de vérification en ligne existent, comme celui du CINES (https://facile.cines.fr/), mais il faut avoir à l'esprit que leur utilisation suppose l'envoi des fichiers sur le réseau internet et leur stockage sur un serveur distant. De plus, les informations retournées sont souvent minimales.

En fonction des moyens humains et techniques dont l'on dispose et du volume de données à traiter, on fera une vérification systématique ou par échantillonnage. Dans le second cas, on s'assurera que l'échantillon soit représentatif.

Dans l'idéal, la trace de la validation du fichier est conservée au titre de l'historique du paquet (préciser a minima la date, l'outil utilisé et sa version, et le résultat donné par l'outil).

[9] https://fr.wikipedia.org/wiki/SHA-2

S'assurer de la complétude et de la cohérence des métadonnées

Selon le degré d'automatisation auquel on peut parvenir, on procèdera soit :

- par contrôle visuel, de manière exhaustive ou sur un échantillon représentatif ;

- par contrôle automatique. Si les métadonnées sont exprimées en XML, la validation au regard du schéma (schéma XML, DTD ou RelaxNG) est fortement recommandée. Dans le cas où l'institution souhaite ajouter des contraintes spécifiques à celles déjà définies par le schéma de métadonnées, un langage tel que Schematron est particulièrement adapté.

Les migrations

La migration de support

Il s'agit du passage des copies d'un support de stockage à un autre, pour des raisons d'obsolescence (du support lui-même ou de son appareil de lecture).

Au préalable de toute migration de support, qu'elle concerne le passage à une technologie différente ou non, il convient d'effectuer un calcul d'empreinte (cf. 3.1). Une fois la migration effectuée, les empreintes doivent être recalculées et confrontées aux empreintes de départ.

La migration de format

Il s'agit de convertir les fichiers constituant les documents numériques dans un autre format de données, soit parce que le format d'origine est obsolète ou en voie de le devenir, soit parce que l'on souhaite adopter un format présentant des avantages supplémentaires (par exemple, le passage du TIFF au JPEG2000 peut être motivé par des gains de place de stockage, alors que le TIFF n'est pas encore frappé d'obsolescence). Une migration de format est une opération lourde, potentiellement complexe, et qui comporte de nombreux risques. Elle a d'autant plus de chances de réussir que les formats d'origine et de destination sont maîtrisés, que l'on dispose d'outils pour les manipuler, et que les fichiers ont bien été contrôlés lors de la prise en charge initiale.

Il convient avant de se lancer dans de telles opérations de définir une méthodologie de travail et de test, en définissant les propriétés significatives que l'on doit retrouver à l'identique dans le document de départ et dans le résultat de la migration. Pour des documents issus de la numérisation, la définition et la résolution doivent être identiques, mais c'est surtout à la colorimétrie que l'on prêtera attention, car les risques de dérives sont nombreux. De manière générale, on prendra bien garde à effectuer des tests suffisants sur les données migrées avant de supprimer toute donnée originale.

Bibliographie

BANAT-BERGER Françoise, DUPLOUY Laurent, HUC Claude. *L'archivage numérique à long terme, les débuts de la maturité* ? Paris : La documentation française, 2009. 284 p.

PEYRARD Sébastien. « Préserver ses collections numériques ». Dans CLAERR Thierry et WESTEEL Isabelle (dir). *Manuel de constitution de bibliothèques numériques*. Paris : Éditions du Cercle de la Librairie, 2013. 407 p. Pages 307-382.