CollEx-Persée

Sélectionner les objets numériques à préserver

Note du sous-groupe « ligne directrices » du GT préservation numérique de Collex-Persée

Présentation générale

Le GT « Préservation numérique » de l'infrastructure de recherche CollEx-Persée travaille sur la préservation à long terme des ressources produites et collectées par les établissements documentaires afin d'enrichir leur collection d'excellence, à destination des chercheurs.

Ce groupe a décidé de créer, en son sein, un sous-groupe "lignes directrices" afin de produire des éléments d'instruction et de décision sur la préservation des ressources numériques. Ces lignes directrices sont plutôt destinées aux personnes et aux entités amenées à créer ou recueillir des données numériques, notamment dans le cadre de projets. Elles doivent leur permettre de définir ce qu'il faut conserver – à court, moyen et long terme – et leur fournir des pistes et des critères pour choisir la solution d'archivage.

L'objectif de ce sous-groupe n'est pas de produire de nouvelles normes ou référentiels, mais de fournir des documents à des fins pédagogiques et d'accompagnement, en s'appuyant sur des documents de référence comme le modèle OAIS, la norme NF Z42-013 sur l'archivage électronique, les référentiels du CINES ou de la section Aurore du SIAF...

Ce sous-groupe est constitué de représentants du SIAF, de la BNU, de la BnF, de la bibliothèque de Sciences-Po et du MNHN.

Nous avons d'abord produit un logigramme qui indique, en fonction de la ressource, et d'un certain nombre de critères, quels objectifs et quelles solutions de préservation doivent être privilégiées.

Périmètre de travail et éléments de définition

Les lignes directrices peuvent concerner toutes les ressources, sous forme numérique, destinées à rejoindre la collection d'excellence d'un établissement documentaire.

Par conséquence, ceci couvre aussi bien des ressources numérisées (à partir d'un original sur support) que des ressources reçues ou produites directement sous forme numérique. Cette distinction fondamentale en croise une autre : les ressources, quelle que soit leur forme, peuvent avoir été publiées ou non. Enfin, les types documentaires peuvent être variés : texte, image, son, documents audiovisuel, logiciels, etc.

On trouvera ainsi comme ressources, de façon non exhaustive :

- les ressources, imprimées ou non, qui ont fait l'objet d'une numérisation rétrospective ;
- les archives orales ;
- des documents audiovisuels acquis ou produits par l'établissement ou des chercheurs ;
- les archives institutionnelles d'un établissement, numérisées ou recueillies sous forme directement numérique ;
- les archives scientifiques d'un chercheur ou d'un groupe de chercheur, recueillies sous forme directement numérique ;
- …

On trouvera en illustration, en annexe, une typologie des documents qui ont vocation à être conservés à la BNU.

Deux types de documents ne sont cependant pas traités par ce sous-groupe, car ils font l'objet d'instructions ou de programmes spécifiques.

- Les revues électroniques éditées par de grands éditeurs scientifiques relèvent du périmètre de travail du projet Istex (https://www.istex.fr/).
- Les données de la recherche entrent dans le périmètre d'étude du COSO (Comité pour la Science Ouverte), notamment de son collège « données ». Celles-ci comportent des enjeux spécifiques (en matière de description, de diffusion, de réutilisation, etc.), bénéficient souvent d'outils de gestion dédiés (souvent qualifiés d'« entrepôts de données), et par ailleurs ne sont pas nécessairement sous la responsabilité des établissements documentaires. Cependant, la limite peut être floue entre les données de la recherche et les ressources qui entrent dans le périmètre de Collex-Persée : par exemple, les documents bureautiques produits par des chercheurs, ou des enregistrements réalisés dans le cadre d'enquêtes orales, sont des archives scientifiques (donc dans le périmètre de Collex-Persée) mais elles sont aussi des données de la recherche. À ce titre, un lien devra être établi avec le COSO sur ces questions.

Glossaire des termes employés

Ces lignes directrices traitent des enjeux et des solutions de conservation numérique à court, moyen et long terme. Dans ce document, les termes suivants seront employés ; leur définition sont tirées ou adaptées de documents normatifs ou de référence.

Archivage : « Ensemble des actions, outils et méthodes mis en œuvre pour identifier, gérer et conserver des documents et des données dans le but de les rendre accessibles, exploitables et intelligibles pendant toute la durée nécessaire à la satisfaction des obligations légales, pour les besoins des utilisateurs et/ou à des fins patrimoniales. L'archivage implique d'identifier précisément les responsabilités des différents acteurs. » [*Les Archives électroniques*[1]].

Archivage électronique : « Fonction essentielle d'un organisme en charge de collecter, classer, conserver et communiquer des archives. L'objectif de cette fonction est d'apporter la confiance nécessaire pour que les activités entre les organismes et les individus puissent s'appuyer sur des documents et données fiables et durables » [NF Z42-013[2]].

Archives électroniques : « Enregistrements électroniques, quelles que soient leurs origines et quelles que soient les raisons de leur conservation, [qui nécessitent d'être conservée de manière sécurisée lorsque ces informations relèvent d'une obligation réglementaire ou sont identifiées comme étant un actif important pour la mémoire, la sécurité ou l'activité opérationnelle d'une organisation ou d'un individu]. Les archives électroniques peuvent être issues :

- de systèmes de production d'objets nativement numériques ;
- de systèmes de numérisation d'objets analogiques (papier, microformes, supports analogiques audio ou vidéo…) ;
- d'autres systèmes de conservation d'objets numériques. [Adapté de NF Z42-013].

Conservation (numérique) : voir « Préservation ».

Durée d'utilité administrative (DUA) : « Durée légale ou pratique pendant laquelle un document est susceptible d'être utilisé par le service producteur ou son successeur, au terme de laquelle est appliquée la décision concernant son traitement final. Le document ne peut être détruit pendant cette période qui constitue sa durée minimale de conservation. » [*Abrégé d'archivistique*][3].

Entrepôt de données [de recherche] (*Research Data Repository* ou *Data Repository*) : base de données destinée à accueillir, conserver, rendre visibles et accessibles des données de recherche. Son rôle est de permettre le dépôt ou la collecte de données, leur description, leur accès, et leur partage en vue de leur réutilisation [*CoopIST*[4]].

---

[1] Béchard (Lorène), Fuentes Hashimoto (Lourdes), Vasseur (Édouard), *Les archives électroniques*, Paris : Association des archivistes français, 2020.

[2] Norme NF Z42-013. - Archivage électronique**. -** Spécifications relatives à la conception et à l'exploitation de systèmes informatiques en vue d'assurer la conservation et l'intégrité des documents stockés dans ces systèmes. Nous utilisons ici non la version publiée mais la version en cours de validation, qui sera publiée prochainement sur le site de l'AFNOR.

[3] Gueit-Montchal (Lydiane), dir., *Abrégé d'archivistique*, 4e édition, Paris, AAF, 2020.

[4] Dedieu (L.), Barale (M.), *Déposer des données dans un entrepôt, en 6 points*, Montpellier, CIRAD, 2020, https://doi.org/10.18167/coopist/0070.

Format de document : Convention de structure d'un objet numérique et de représentation de l'information contenue [*NF Z42-013*]. Certains formats sont « réputés pérennes »[5].

Migration de format : ré-encodage des informations numériques dans un nouveau format, notamment parce qu'il est réputé plus pérenne.

Sauvegarde : « Opération technique destinée à assurer, par une copie de sécurité, la continuité de l'exploitation d'un système informatique en cas d'incident. Par extension, résultat de cette opération. » [*Les Archives électroniques*].

Stockage : « Action d'entreposer des contenus numériques sur un support, quel qu'il soit (disque dur, support amovible, serveur physique, etc.), servant de base à leur traitement ultérieur). » [*Les Archives électroniques*].

Préservation : « Opérations préventives ou curatives de régénérescence de la lisibilité d'un document. » [*Les Archives électroniques*]. Le terme de préservation est souvent utilisé comme synonyme de « conservation » pour les objets numériques.

La sélection des objets à préserver

La première tâche de ce groupe est de répondre à la question quoi sélectionner. L'une des difficultés majeures en matière de conservation numérique, est la volumétrie souvent considérable des données. Une volumétrie importante signifie un coût important, et donc une menace sur la soutenabilité des projets de conservation. Il est donc indispensable de définir la durée pendant laquelle chaque élément d'une collection doit être conservé ; et d'identifier, au sein de l'ensemble des ressources produites lors d'un programme ou un projet spécifique, celles qui doivent faire l'objet d'un archivage sur le long terme à proprement parler.

À cette fin, nous avons proposé le modèle du logigramme : à la suite d'une série de questions, il sera possible de déterminer le devenir de la ressource étudiée. Les questions de ce logigramme doivent être appliquées à tous les types de ressources ou d'objets numériques produits (les types de données pouvant s'entendre en termes de catégorie documentaire, de format, de volume, etc.). Par ailleurs, au sein d'un même type de données, il est envisageable, là encore pour des raisons d'économie, de procéder à des échantillonnages en ne gardant qu'une sous-partie de l'ensemble.

Le logigramme proposé est en réalité conçu comme une série de trois logigrammes.

[5] Pour être réputé pérenne, « *a minima* le format doit être spécifié (c'est-à-dire qu'il doit être documenté) et l'on doit pouvoir accéder librement à ses spécifications. De plus, il est préférable que le format soit utilisé par une large communauté, voire normalisé, ce qui lui assure une certaine durabilité. Cette sélection n'a pour seul objectif que d'essayer de maîtriser le format dans lequel sont représentées les données afin de faire face à l'obsolescence technologique, inévitable dans le domaine numérique. Cette réflexion doit être envisagée en tenant compte de la durée pendant laquelle les données doivent être conservées » [*Les Archives électroniques*].

Le premier logigramme cherche à identifier le statut de la ressource : archive publique ou non. En effet, la qualification d'archive publique entraîne des règles spécifiques de traitement et de conservation.

Le second logigramme, en conséquence, est applicable aux archives publiques. Les règles de traitement sont fixées par la loi ; on peut donc s'appuyer sur des référentiels comme celui de la section Aurore pour déterminer si la ressource doit être supprimée à l'issue d'une durée définie ou être conservée indéfiniment. Le logigramme détaille plus longuement, en revanche, la question technique du format. Il s'agit de déterminer si le format de la donnée peut tenir dans le temps. Si le format offre des garanties de pérennité, une conservation sur le long terme est envisageable. Sinon, on tentera de convertir la ressource dans un format plus maîtrisé.

- Si cette conversion est réalisée sans perte d'information, on peut envisager, pour des raisons économiques, de supprimer le format initial – ce choix dépend aussi des volumétries en question et des ressources de l'établissement.
- Si cette conversion est réalisée avec perte, il est recommandé de conserver en parallèle le format initial. Il est en possible que le format original devienne maîtrisé à l'avenir.
- Si aucune conversion n'est possible, la seule solution est de garder le format original tel quel, en espérant qu'il devienne maîtrisé à l'avenir.

Le troisième logigramme s'applique aux ressources qui ne relèvent pas du statut d'archives publiques.

- Il est possible que la conservation de la ressource en question, même s'il ne s'agit pas d'une archive publique, soit requise au titre d'une obligation légale (par exemple, le dépôt légal) ou contractuelle (il peut ainsi arriver qu'un établissement se soit engagé à conserver un objet sur le long terme, auprès de son donateur ou de son dépositaire). C'est l'objet de la **première question du logigramme**. Dans ce cas, il faut d'emblée considérer qu'elles doivent faire l'objet d'un archivage définitif.
- À l'inverse, la conservation à long terme d'une ressource peut être interdite. Les législations protectrices des données personnelles, comme le RGPD, doivent ainsi être prises en compte. C'est l'objet de la **deuxième question**. Dans ce cas, il faut supprimer les données à l'issue de leur **période d'utilité courante** ; ou les transformer (les anonymiser, par exemple).

- Au-delà des strictes obligations et interdictions légales, la décision de conservation doit être prises sur des critères scientifiques. Dans ce domaine, deux textes peuvent éclairer la décision : la charte de la conservation dans les bibliothèques (2011), élaborée par les ministères de la Culture et de l'ESR[6], et le « décret n° 2020-195 du 4 mars 2020, portant diverses dispositions relatives aux bibliothèques »[7].
    - Le premier document stipule : « Est dit patrimonial un document, un objet ou un fonds auquel est attachée une décision de conservation sans limitation de durée ».

[6] https://www.culture.gouv.fr/Sites-thematiques/Livre-et-lecture/Patrimoine-des-bibliotheques/Gerer-le-patrimoine-en-bibliotheque/La-charte-de-la-conservation-dans-les-bibliotheques.

[7] https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000041686954&dateTexte=&categorieLien=id

Réciproquement, la conservation à long terme doit être réservée aux documents patrimoniaux.

- C'est dans le second texte que l'on trouve une décision plus précise d'un document patrimonial en bibliothèque. Le décret n°2020-195 modifie ainsi l'article R. 311-1 du code du patrimoine : « Sont des documents patrimoniaux, au sens du présent livre, les biens conservés par les bibliothèques relevant d'une personne publique, qui présentent un intérêt public du point de vue de l'histoire, de l'art, de l'archéologie, de la science ou de la technique, notamment les exemplaires identifiés de chacun des documents dont le dépôt est prescrit aux fins de constitution d'une mémoire nationale par l'article L. 131-2 du présent code et les documents anciens, rares ou précieux ».
- Le premier critère justifiant d'une conservation à long terme est donc « l'intérêt public du point de vue de l'histoire, de l'art, de l'archéologie, de la science ou de la technique » de l'objet à conserver. Ce critère, naturellement subjectif, doit être apprécié par le personnel scientifique de l'établissement ; et il doit d'abord apprécié au regard des missions de l'établissement qui conserve cet objet. C'est l'objet de la **troisième question**.
- Le décret fournit également des critères plus précis. Tout d'abord, le décret indique que les documents relevant du dépôt légal sont de nature patrimoniale et doivent être conservés. Notons que ce cas-là était déjà couvert par la **première question du logigramme**.
- Le décret évoque ensuite le caractère « précieux » d'un objet. Celui-ci peut s'apprécier de deux manières. Soit il est précieux au regard de sa valeur (scientifique ou artistique) intrinsèque ; dans ce cas il rejoint le critère des ressources « présentent un intérêt public du point de vue de l'histoire, de l'art (…) », critère couvert par la **troisième question**. Soit il est précieux au regard de sa valeur vénale d'acquisition, que l'objet ait été acheté auprès d'un tiers, ou qu'il ait été produit par l'institution. Les numérisations de documents sur support sont par exemple « précieuses » dès lors que le coût du processus de numérisation lui-même est important. C'est pourquoi la **quatrième** question demande si l'objet peut être re-généré à un coût limité (par exemple, une version de diffusion n'a pas à être conservée si une version master existe déjà). Dans le cas contraire, on peut considérer l'objet comme précieux.
- La question de la rareté se pose de façon spécifique pour les documents numériques, par essence réplicables à l'infini (sauf présence de mesures technique de protection ou impossibilité technique de les lire). Cependant, en pratique, une ressource peut être conservée par un nombre limité, voire par une seule institution. L'importance de conserver une ressource est inversement proportionnelle au nombre d'établissements qui la conservent. Dans le logigramme, la question de la rareté ou de l'unicité s'exprime au fil d'une série de questions (**questions 5 à 8 du logigramme**) : il s'agit de déterminer si le document est unique (au sein de l'établissement ou globalement), et dans le cas contraire, de déterminer si la version possédée est la meilleure d'un point de vue technique (seule la meilleure version ayant vocation à être conservée). À noter : l'unicité ne doit pas se penser à un instant T, mais sur le long terme : si des établissements possèdent une ressource numérique mais ne se sont pas engagées à la conserver, on ne peut pas considérer que celle-ci soit correctement archivée.
- Enfin, le critère d'ancienneté est aussi plus difficilement applicable aux collections numériques, qui par définition sont récentes. Cependant, des ressources anciennes

à l'échelle de l'ère numérique (avant les années 2000, par exemple) pourraient entrer dans cette catégorie. Cela peut représenter une exception au critère d'unicité ou de rareté : des versions précédentes d'un objet numérique existant dans une qualité supérieure pourraient être conservées au titre de la documentation de l'histoire d'une technique. Ce cas étant assez particulier, il ne fait pas l'objet d'une question dans le logigramme.

- La question du format de conservation est également à prendre en compte (voir les questions sur ce sujet dans le logigramme n°2).

- Enfin, même en ayant éliminé tous les documents qui ne répondent pas aux critères exposés ci-dessus, il peut arriver que l'établissement n'ait pas les moyens financiers d'assurer la bonne conservation des objets qu'il possède, car leur volumétrie est trop importante. Dans ce cas, une étape supplémentaire de tri peut être nécessaire. Plusieurs critères de priorisation et de tri peuvent être appliqués :
    - Du point de vue de la **rareté** : on peut décider de ne pas assurer la conservation des objets qui sont simplement « rares » (i.e. conservés par un nombre restreint d'établissements) pour se concentrer sur ceux qui sont uniques (i.e. conservés par son propre établissement seulement).
    - Du point de vue du **format** : On peut décider d'éliminer les documents au format original dès lors qu'on peut le convertir vers un nouveau format, sans perte, ou avec un niveau de pertes limité. On peut aussi choisir de ne pas conserver les objets dans des formats qui ne sont pas maîtrisés, considérant qu'on ne pourra pas les conserver.
    - Du point de vue de l'« **intérêt public** [de l'objet] du point de vue de l'histoire, de l'art, de l'archéologie, de la science ou de la technique » : on peut décider de ne pas conserver l'ensemble des objets d'un lot, mais de réaliser un échantillonnage.

À l'issue de ce processus, on identifie les objets qui doivent faire l'objet d'un archivage. Cependant, à l'exception des documents dont la conservation est une obligation légale ou contractuelle, le caractère patrimonial (et donc la décision d'archivage pérenne) des documents est sujet est réexamen. C'est ce qu'indique la charte de la conservation dans les bibliothèques, dans la note de son article 5 : « Le statut patrimonial conféré à un document, un objet ou un fonds peut lui être retiré. Cette procédure doit reposer sur une réflexion scientifique, s'inscrire dans une démarche professionnelle et collective et se conformer au droit de la domanialité des personnes publiques ».

Cette réévaluation peut avoir lieu à l'issue d'une période d'archivage intermédiaire (cela peut par exemple correspondre à la fin de contrat avec un prestataire d'archivage externe). Dans ce cas, il faut réexaminer l'ensemble des types d'objets à conserver, au regard des critères précédemment évoluer. En effet, la décision peut être différente : par exemple, un ou plusieurs autres établissements peu(ven)t avoir pris la responsabilité de conservation de l'objet, rendant inutile la conservation par son propre établissement. Autre possibilité, l'établissement peut avoir renumérisé un objet sur support, rendant la version précédente obsolète… À l'inverse, la conservation peut être devenue une obligation légale formelle de l'établissement, et donc faire l'objet d'une décision de conservation définitive.

À l'issue de la période d'archivage intermédiaire, trois décisions sont alors possibles : l'objet peut être supprimé ; il peut faire l'objet d'une nouvelle période d'archivage intermédiaire à l'issue de laquelle il sera évalué à nouveau ; il peut faire l'objet d'une décision de conservation définitive.

Annexe : exemple de typologie des documents à préserver

Nature	***Exemples***	***Précision***
Archives administratives de l'établissement	Documents administratifs, archives de l'activité de l'établissement	Archives historiques (jusqu'à 2nde GM). Versement en 2016 aux AD des archives depuis la 2nde GM
Documents patrimoniaux numérisés (domaine public)	Manuscrits médiévaux, livres anciens, cartes historiques, etc.	Images
		Métadonnées bibliographiques
		Enrichissement scientifique / documentaire (OCR corrigé, structuration XML-TEI etc)
Documents patrimoniaux numérisés (couverts par le droit d'auteur)	Correspondances ; Affiches européennes du 20e siècle	Images
		Métadonnées bibliographiques
		Enrichissement scientifique / documentaire (transcription, etc.)
Documents du dépôt légal, nativement numériques	Aujourd'hui, le DL imprimeur arrive sous forme physique. Envisageable de demander les fichiers numériques à la place	Images
		Métadonnées bibliographiques
Ressources numériques éditées, acquises	Revues électroniques, E-books, publiés	Accès "pérenne" à la plateforme de l'éditeur ou du diffuseur
		Livraison des données sur support (faire un inventaire à partir des licences signées)

Dépôt légal du web	Sélection par l'établissement, gestion technique par la BnF	Sites internet français
		Métadonnées et indexation des pages web à des fins de TDM
Sources historiques nativement numériques	Archives d'écrivains et compositeurs de la région ; "user-generated content" : photographies, textes, produits par les citoyens ; archives photographiques ;	Images
		Métadonnées bibliographiques
Archives scientifiques nativement numériques	Archives de chercheurs des domaines couverts par l'établissement ; tri et sélection par l'établissement.	Responsabilité est normalement au service des archives de l'université de rattachement du chercheur. Néanmoins, possibilité de dons.
Documents audiovisuels numériques	Films, enregistrements audio-visuels	Responsabilité de l'INA / CNC
Editions de l'établissement	Revue de l'établissement, catalogues…	Diffusion en ligne de la Revue ; possible diffusion en ligne des autres produits éditoriaux de l'établissement.