SKKB CSBC

Schweizerische Konferenz der Kantonsbibliotheken
Conférence suisse des bibliothèques cantonales
Conferenza svizzera delle biblioteche cantonali

Guide pour un archivage numérique à long terme

Leitfaden für Digitale Langzeitarchivierung

État 21.03.2024 (V9 FR)

Membres du groupe de travail
Sylvie Béguelin (Médiathèque Valais)
Romain Guedj (BCU Fribourg)
Théophile Naito (BCU Lausanne)
Alexis Rivier (Bibliothèque de Genève)
Brigitte Sacker (ZB Zürich)
Tobias Viegener (NB)
Philipp Wiemann (KB Vadiana St.Gallen)
Mirjam Zürcher (ZHB Luzern)

Table des matières

1 Introduction

L'archivage numérique à long terme du numérique est depuis plus de vingt ans un sujet important pour les institutions patrimoniales. La norme Open Archival Information System Reference Model (OAIS) est publiée en 2002 déjà. Ce modèle organisationnel pose les bases de tous les systèmes de pérennisation des données numériques et fait toujours autorité aujourd'hui.

Même si plusieurs installations ont été mises en service à la Confédération (Bibliothèque nationale, Archives fédérales) ou dans les cantons (Valais et Bâle par exemple), et bien qu'il existe des prestataires très actifs, la démarche de mise en place d'une telle infrastructure reste complexe pour une bibliothèque cantonale et il n'existe pas de solution « clé en main ».

Le contraste est frappant en comparaison avec la conservation des sources documentaires physiques, que les bibliothèques comme les archives réalisent et maîtrisent efficacement depuis des siècles. La simplicité inhérente aux documents physiques rend évidentes les moyens à mettre en œuvre : protéger les documents de la dégradation matérielle suffit à préserver l'accès au contenu pour les usages futurs.

Le *Message culture 2020-2024* de l'Office fédéral de la culture (OFC) rappelle la problématique du numérique : « Exploiter des informations numériques fait aujourd'hui partie de notre quotidien. Ce qui ne va pas de soi est la sauvegarde et l'exploitabilité à long terme de ces données. »[1] Il reconnaît que « garantir à long terme la conservation et l'exploitabilité des données numériques est une affaire complexe. » Et que la façon privilégiée d'aborder le sujet passe par la coopération : « Étant donné le montant des coûts, la tâche ne peut être maîtrisée qu'en coopérant. Sont sollicitées ici non seulement les bibliothèques cantonales et les archives cantonales, mais encore la BN et les Archives fédérales, qui sont les institutions mémorielles de la Confédération. »

La *Stratégie de la Bibliothèque nationale suisse BN 2020–2028*[2] exprime cet engagement à assurer la pérennité des ressources numériques et à coopérer avec les autres institutions fédérales et cantonales pour la conservation de la mémoire de la Suisse. Le groupe de travail DigiRep mandaté par la CSBC/SKKB s'inscrit directement dans cette perspective.

Le premier rapport du groupe de travail, *Lignes directrices pour une collection de contenus d'information numériques* (2020)[3] a dressé un tableau général des enjeux existants, tout particulièrement pour les bibliothèques cantonales.

Celles-ci doivent se poser une question fondamentale, à savoir leur positionnement face au numérique : mon institution a-t-elle la mission de conserver des données numériques et par conséquent doit-elle disposer d'un système d'archivage électronique (SAE) ?

Si la réponse à cette interrogation est affirmative, on doit se demander comment atteindre cet objectif. Le but du présent document est de passer en revue les points essentiels à considérer, afin de réaliser concrètement une infrastructure d'archivage pérenne.

---

1 https://www.fedlex.admin.ch/eli/fga/2020/725/fr

2 https://www.nb.admin.ch/dam/snl/fr/dokumente/nb_als_flag-einheit/studien_und_berichte/nb_strategie.pdf.download.pdf/nb_strategie.pdf

3 https://www.bibliosuisse.ch/Portals/0/InhalteFR/Sections/CSBC/Activit%C3%A9s/Lignes-directrices-collection-me%CC%81dias-nume%CC%81riques_FR_final_20201006.pdf?ver=Ose2EYd4CwTkOZLKjhE6ow%3d%3d

2 Questions générales

2.1 Stratégie

La mise en place d'un système d'archivage pérenne est une démarche d'innovation. Dans une situation de ressources financières et humaines dans le meilleur des cas stables, et étant donné que des services de base ne peuvent être abandonnés (notamment ceux qui concernent le traitement et l'accès aux documents physiques traditionnels), cet objectif doit être légitimé par une **orientation stratégique** forte de la bibliothèque.

Cette stratégie définit et justifie les arbitrages budgétaires en faveur d'un projet d'archivage pérenne. Elle doit également interroger les compétences nécessaires et vérifier si elles sont réunies en interne. Les recrutements, quand ils sont possibles, permettront de compléter les lacunes et de composer des équipes interdisciplinaires efficaces. Une stratégie bien pensée ne peut être complètement externalisée et exige donc l'acquisition d'un minimum de connaissances et de savoir-faire dans ce domaine. Il faut être également conscient qu'un SAE opérationnel engendre des coûts récurrents qui doivent être intégrés dans la planification budgétaire. Sachant que les processus décisionnels dans la fonction publique sont lents, ces évolutions devraient être pensées le plus en amont possible.

Un **projet pilote**, d'ampleur limitée, peut représenter une approche adaptée. Les résultats obtenus serviront alors de base à l'élaboration de la stratégie. Certaines opportunités peuvent rendre cette démarche intéressante : par exemple un producteur de données désireux d'expérimenter un workflow d'archivage numérique ou encore un fonds numérique important que la bibliothèque envisage d'acquérir et d'intégrer à ses collections.

2.2 Infrastructures

Les activités de la bibliothèque sont soutenues par un environnement technique. Les infrastructures existantes sont à prendre en compte et pourront servir de « briques », qui seront mises à profit pour une solution SAE. Il s'agit notamment :

- du ou des systèmes de gestion métier déjà utilisés: système intégré de gestion de bibliothèque (SIGB), système de gestion d'archives, système de gestion de musée, système de gestion de bibliothèques numériques;
- des infrastructures informatiques globales de l'administration de tutelle (université, canton, ville): serveurs de fichiers, système de GED (gestion électronique de document), Digital Asset Management (DAM), voire même un SAE déjà mis en place pour les besoins d'un autre service, tel celui des archives par exemple.

2.3 Périmètre

Le **périmètre** de données concernées est une question cruciale, car la réponse va déterminer la taille du système requis et sa croissance.

Le périmètre dépend de la mission de collecte de l'institution et de la base légale sur laquelle elle s'appuie. La situation est extrêmement diverse selon les cantons. Certains cantons de Suisse occidentale disposent d'un dépôt légal pour les publications (FR, GE, VD), mais qui ne couvre pas toujours de manière explicite les publications numériques. D'autres cantons ont des mandats

de collecte exprimés de façon variées dans des lois cantonales ou des règlements[4]. Par conséquent chaque institution doit soigneusement étudier le corpus législatif et réglementaire en vigueur et en déduire l'étendue des ressources numériques concernées[5].

Une fois le périmètre de ressources défini, son volume doit être quantifié :

- en fonction des **types de ressources concernés** (texte, texte numérisé, image fixe, son, image animée, etc.);
- en fonction du nombre d'objets à traiter (ou d'heures pour le son et l'image animée) pour chaque type de ressources;
- en fonction de l'accroissement attendu pour les prochaines années.

Cette quantification doit s'exprimer en mesure de **stockage informatique** (To). Cette unité permettra de chiffrer l'infrastructure de stockage pérenne ainsi que son évolution future. C'est également valable pour les prestations de SAE en sous-traitance comme celles des Archives fédérales pour les tiers qui sont exprimées en francs par To de données à pérenniser[6].

2.4 Partenaires

Pour la Bibliothèque nationale, la recherche de partenariats est une méthode privilégiée afin d'atteindre les objectifs de projets d'innovation et elle est inscrite dans son plan stratégique mentionné plus haut : « La mémoire de la Suisse repose sur une étroite coopération entre bibliothèques, archives, musées et établissements de recherche. La Bibliothèque nationale s'engage pour une coordination nationale et internationale efficace entre les différents acteurs concernés. »

Par le passé une démarche similaire a permis de mettre en place le réseau Memoriav. Au lieu de créer une nouvelle institution pour sauver le patrimoine audiovisuel, Memoriav a privilégié la mise en réseau des compétences présentes sur le territoire.

Pour les bibliothèques cantonales, les possibilités de partenariat peuvent se présenter selon deux dimensions :

- Une dimension locale ou verticale : la tutelle (canton, université, ville) dispose de compétences clés (équipes informatiques) ou d'institutions dont les besoins en archivage numérique sont proches (services d'archives, musées).
- Une dimension nationale ou plus horizontale : les bibliothèques cantonales et la BN ont des besoins similaires dans leurs périmètres respectifs et mettent en commun leurs compétences.

L'approche par la dimension nationale a permis aux bibliothèques cantonales de participer à des initiatives très importantes et structurantes du paysage documentaire, comme SLSP, e-rara, e-manuscripta, e-periodica, e-newspaperarchives.

Les exemples de mise en place de SAE ont plutôt été réalisés sur la base de partenariats locaux. Par exemple le projet de SAE pour le Service de la culture valaisan bénéficie aux Archives de l'Etat du Valais, à la Médiathèque Valais et aux Musées cantonaux. La Confédération a mis en place un SAE pour la Bibliothèque nationale et pour les Archives fédérales. L'appel à des sociétés

---

4 *Ligne directrices pour une collection de contenus d'information numériques* (CSBC/SKKB 2020), chap. 5.

5 Pour plus de détails, voir le chapitre « 2. Sélection », du rapport déjà cité.

6 *Archivage numérique aux AFS: Prestations pour les clients extérieurs à l'administration fédérale*, 01.09.2021. https://www.bar.admin.ch/bar/fr/home/archivage/documents-numeriques/archivage-numerique-pour-tiers.html

spécialisées dans la gestion d'archives numériques actives sur tout le territoire sont des interlocutrices précieuses permettant de garantir une bonne compréhension des besoins métiers et des exigences des services informatiques.

2.5 Conservation et/ou accès

Le modèle de référence OAIS conceptualise l'acquisition de contenus numériques (*Ingest*) et leur gestion dans l'archive, mais aussi leur accessibilité pour les usagers et usagères (*Access*).

Le système d'accès a un intérêt citoyen et politique évident, mais il ajoute une couche de complexité au projet. En effet l'accès par le public est limité par d'éventuelles restrictions que le système doit pouvoir gérer : droit d'auteur, protection des données, délais de consultation.

De plus l'accès doit être articulé avec les systèmes de gestion existants. Dans certains cas le système de gestion propose une interface publique. La Bibliothèque nationale par ex. intègre à son portail HelveticAll (Ex Libris Alma) différents niveaux d'accès pour les documents numériques archivés dans le cadre du programme e-Helvetica : accès ouvert, accès réservé sur des postes dédiés. Dans d'autres cas, il peut être plus simple de mettre en place une plateforme de diffusion hors du système de gestion existant. Par exemple, la Bibliothèque cantonale et universitaire de Fribourg utilise Ex Libris Alma comme système mais diffuse via AtoM.

3 Réalisation d'un système d'archivage électronique

3.1 Le modèle de référence OAIS

La clarification des conditions organisationnelles, financières et techniques, ainsi que l'évaluation des besoins réels sont des préalables à la réalisation d'un système d'archivage électronique (SAE). La plupart des systèmes actuels sont fondés sur le modèle de référence Open Archival Information System (OAIS)[7]. Ce modèle décrit des unités fonctionnelles qui, en combinant des moyens humains avec des systèmes informatiques, servent à préserver les informations à long terme, c'est-à-dire dans un avenir indéterminé[8]. On distingue les unités fonctionnelles suivantes :

- **Entrée des données** (*Ingest*) : Reçoit les paquets d'information du producteur.
- **Administration** (*Administration*) : Pilote l'ensemble des unités fonctionnelles.
- **Stockage des données** (*Archival Storage*) : Assure le stockage et la récupération des données archivées.
- **Gestion des données** (*Data Management*) : Gère les métadonnées en lien avec les catalogues et les inventaires, mais également les accès, les contrôles de sécurité, les algorithmes de traitement des données, les statistiques entre autres.
- **Planification de la pérennisation** (*Preservation Planning*) : Assure que les données préservées dans l'OAIS restent accessibles, compréhensibles et utilisables à long terme.
- **Accès** (*Access*) : Rend les données pérennisées dans le système visibles pour l'usager final, ainsi que les fonctions associées (recherche, fourniture des informations).

---

7 https://public.ccsds.org/Pubs/650x0m2%28F%29.pdf

8 «Das Referenzmodell OAIS – Open Arch Information Model», in *nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung*, éd. par H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, K. Huth, versio 2.3, 2010. Rédigé dans le cadre du projet *nestor – Kompetenznetzwek Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen für Deutschland* (*nestor – Network of Expertise in Long-Term Storage of Digital Resources*). http://nestor.sub.uni-goettingen.de/handbuch/artikel/nestor_handbuch_artikel_183.pdf

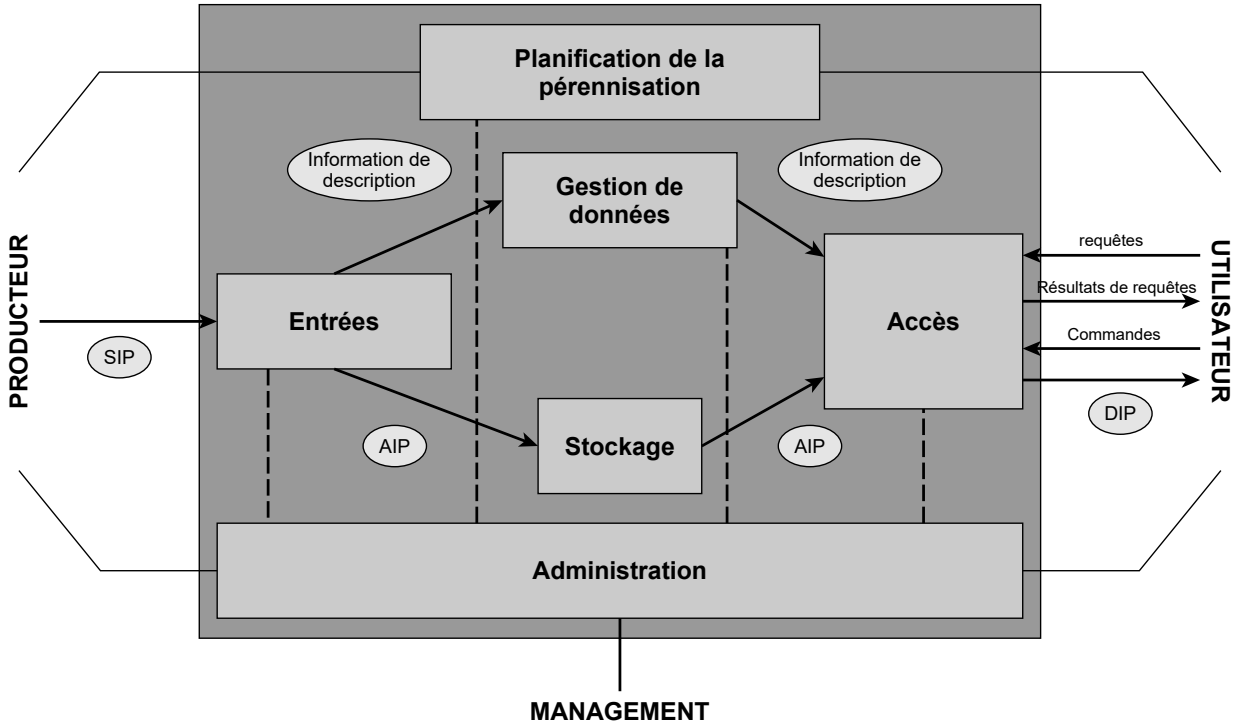

Figure 1: Interaction des unités fonctionnelles (selon « Modèle de référence pour un Système ouvert d'archivage d'information (OAIS) »)

La notion de paquet d'informations est un autre concept essentiel d'OAIS. Un paquet d'informations comprend les contenus à archiver, les métadonnées nécessaires à leur préservation, ainsi que des informations à communiquer aux utilisateurs. Dans le modèle OAIS, on distingue les paquets d'informations à verser (SIP, *Submission Information Packages*), qui sont fournis par les producteurs de ces informations, les paquets d'informations archivés (AIP, *Archival Information Packages*), qui sont stockés dans le système, et les paquets d'informations diffusés (DIP, *Dissemination Information Packages*), qui sont livrés aux utilisateurs finaux. Les relations entre les paquets peuvent être diverses. Il est ainsi possible que les SIP et les AIP aient la même structure et ne se distinguent que de façon marginale au niveau des formats de contenus. A l'inverse, SIP et AIP peuvent également être de nature structurellement différente. Dans tous les cas l'AIP contiendra des informations supplémentaires nécessaires à la préservation des objets numériques par rapport aux SIP. Il peut en être de même pour les relations entre AIP et DIP, bien que la différence est plus probable, car les exigences imposées par l'archivage divergent de celles pour l'affichage et la livraison.

Le fait que les paquets soient analogues ou différents en termes de structure, de contenu ou de formats dépend d'une part du système utilisé et, d'autre part, de la façon dont les objets numériques sont élaborés par l'institution. Par exemple, dans un projet de numérisation, les images TIFF pourraient être transférées en continu du scanner vers le SAE sous forme de SIP individuels. Elles y sont toutefois stockées dans un AIP unique, comprenant les centaines de pages d'un livre numérisé au format JPEG2000. Lors de l'accès par l'utilisateur, seules les pages demandées sont livrées sous la forme d'un DIP au format PDF multipages.

3.2 Conception d'un SAE

3.2.1 Développement

Le modèle OAIS est aujourd'hui une norme ISO reconnue[9]. En définissant clairement les rôles, les responsabilités et les processus, il procure une base commode qui aide à la formulation des exigences d'un SAE, sans pour autant imposer un système technique. Cela laisse une latitude pour le développement de solutions réelles plus ou moins complexes et qui peuvent tenir compte des besoins des différentes unités organisationnelles. Par exemple, il n'est pas nécessaire que chaque processus soit entièrement automatisé. Dans certaines circonstances, la création manuelle de SIP peut s'avérer plus appropriée, tandis que dans d'autres, une interface technique qui se charge de la création des SIP et de leur prise en charge par le SAE de façon entièrement automatique peut être préférable. En même temps, cette flexibilité constitue une difficulté pour l'acquisition d'un tel système : comme les solutions logicielles pour l'archivage numérique sont souvent paramétrées en fonction de cas d'applications concrets, des clarifications préalables complètes sont souvent nécessaires pour définir l'architecture logicielle précise. Les conditions cadres et organisationnelles discutées plus haut constituent une base pour l'établissement d'un cahier des charges et la recherche d'un fournisseur approprié. Grâce au modèle de référence OAIS et à sa terminologie, il est toutefois possible pour le fournisseur et le client de se mettre d'accord sur les fonctionnalités sans avoir besoin de discuter de la mise en œuvre technique dans tous ses détails.

D'autres données sont importantes à rassembler avant le développement du système. Au niveau des producteurs, il faut savoir quelles sont les quantités attendues de livraisons, d'objets livrés, et quels sont les formats de données, les systèmes source ainsi que leurs interfaces. Au niveau des utilisateurs des données, on déterminera s'il s'agit de personnes, d'institutions ou encore de systèmes. Par exemple, un SAE peut n'être accessible qu'à quelques personnes autorisées ou au contraire les utilisateurs peuvent accéder directement aux contenus sur Internet via une interface de consultation interne ou externe. Dans le premier cas, on parle également de *dark archive*. Dans le second cas, la question se pose de savoir si l'interface de présentation (ou le catalogue de la bibliothèque) génèrent et livrent les DIP au moment où un utilisateur les demande ou si les DIP de tous les contenus pertinents sont constitués préalablement à cette demande, dans les formats les plus courants. Cette décision doit être prise en fonction de chaque cas d'usage et en mettant en balance les coûts et les avantages.

3.2.2 Implémentation

Les systèmes d'archivage numérique à long terme sont souvent basés sur des systèmes de type DAM tels que *Fedora Commons*, complétés de façon modulaire par d'autres logiciels, afin de pouvoir couvrir l'ensemble des exigences fonctionnelles (Figure 2, p. 9). Il existe sur le marché aussi bien des solutions open source que des solutions propriétaires de fournisseurs commerciaux. Si l'organisation dispose de ressources informatiques personnelles et techniques importantes, la mise en œuvre d'une solution open source ***on premise*** (sur site), soit sur sa propre infrastructure de serveurs (applications, bases de données, serveurs de fichiers) peut être une solution appropriée en raison de la bonne adaptation aux besoins et des coûts de licence faibles ou inexistants. A l'inverse les organisations plus petites envisageront plutôt d'envisager une **solution basée sur le cloud**, voire une prestation de ***preservation as a service***.

[9] https://www.iso.org/standard/57284.html

Dans le cas d'une solution basée sur le cloud, l'exploitation de l'infrastructure du serveur relève de la responsabilité du prestataire de services. La capacité du client d'influencer le système sont souvent limitées, car les prestataires doivent valider les modifications, les développements à apporter, à l'exception de simples options de configuration. En outre, il convient de vérifier si, dans le cas d'une solution cloud, c'est seulement le système qui opère dans le cloud ou si l'ensemble des données archivées se trouve également dans le cloud.

Il existe des offres qui proposent un système d'archivage, mais pour lequel le stockage redondant des données est en partie sous la responsabilité du client, ce qui peut représenter un poste de coûts supplémentaires. Dans le cas d'une offre de *preservation as a service* en revanche, la responsabilité sur les données archivées, mais aussi sur les mesures de conservation, incombent entièrement au prestataire de services. Pour les deux types de services, les coûts sont généralement calculés en fonction des besoins de stockage des contenus à archiver. Il convient d'y ajouter les coûts de maintenance et de licence.

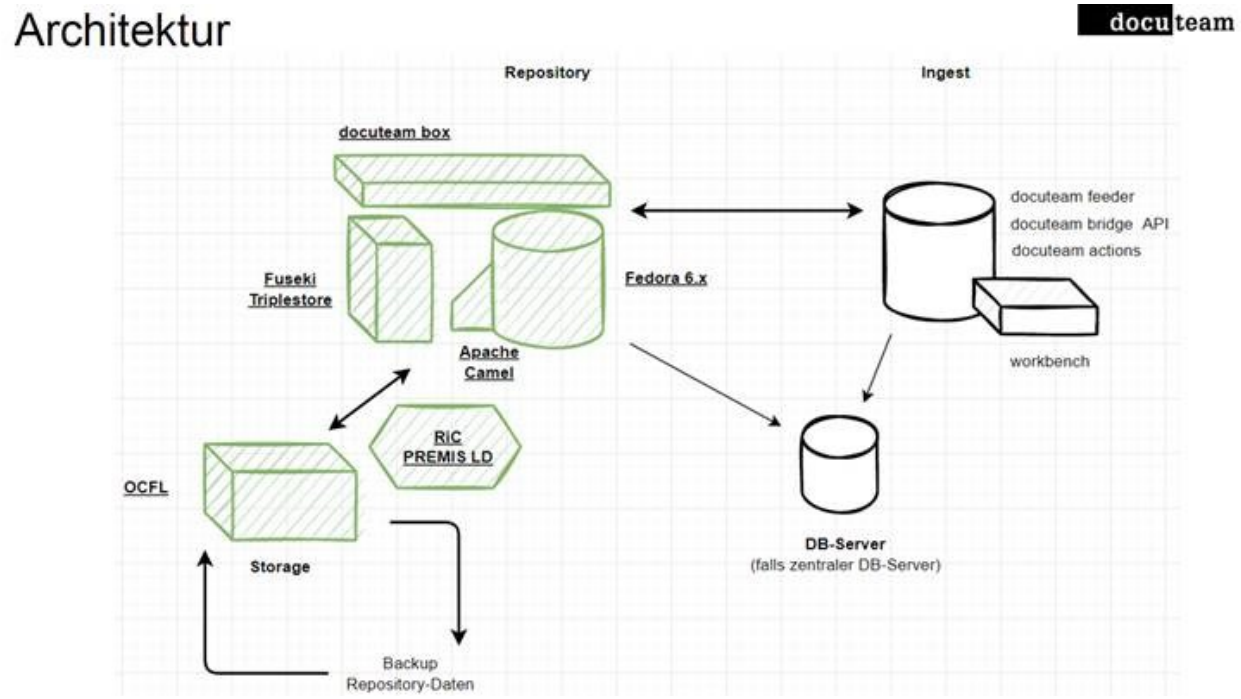

Figure 2: Schéma de l'architecture modulaire d'un système d'archivage numérique à long terme : exemple du système de l'entreprise docuteam SA (crédit : docuteam SA)

En mettant en œuvre une solution, il faut toujours prévoir un plan de sortie en cas de difficultés. Si, par exemple, l'implémentation en interne d'une solution open source ne peut pas être menée à bien en raison de changements de personnel, l'achat d'une solution propriétaire doit pouvoir être envisagé. Au cas où les augmentations de coût ne seraient plus supportables dans le cadre d'un modèle de *preservation as a service*, il convient d'envisager dès le départ une stratégie de retrait documentée qui permette d'exporter les données et métadonnées dans des formats standards et reconnus, afin de se prémunir d'un *cloud lock-in* ou *vendor lock-in*.

3.2.3 Questions préalables

Plusieurs questions se posent à l'institution, à la lumière des différentes méthodes de mise en œuvre ci-dessus. Il est essentiel d'avoir les réponses avant d'entreprendre la réalisation du SAE :

- **On premise ou dans le cloud**
  Ai-je les ressources nécessaires pour exploiter une solution de SAE sur mes propres ser-

veurs et avec mon personnel ?
Dans le cas d'une préférence pour un SAE dans le cloud : quel degré d'influence je souhaite conserver sur la solution (p. ex. au niveau des mesures de préservation) ? Le lieu de stockage des données et le cadre légal existant sont-ils compatibles avec le besoin de protection de mes données ?

- **Conséquences en termes de coûts**
  Quel est le volume des données que je souhaite archiver ? Combien de transferts de données sont à prévoir annuellement et est-ce que les sources de données sont variées ? Est-il facile d'adapter le processus de transfert à chaque nouvelle source de données ? Quelle est le volume d'accroissement par an ?
- **Intégration du système**
  Avec quels autres systèmes faut-il prévoir des interfaces pour le transfert ou l'accès aux données ? Quelles métadonnées doivent être échangés avec ces systèmes et dans quels formats ? Quel doit être le niveau de performance attendu ?

4 Recommandations pour les acquisitions

4.1 Provenance

Afin d'enrichir leurs collections, les bibliothèques acquièrent des documents numériques auprès de différents organismes ou particuliers :

- production interne;
- équipe de numérisation interne;
- entreprises de numérisation externes;
- éditeurs, librairies, distributeurs, etc.;
- personnalités diverses (écrivains, compositeurs, photographes, etc.);
- institutions diverses.

Comme la production interne est souvent quantitativement peu importante, voire inexistante, car ce n'est en général pas une de leurs missions principales, les bibliothèques ont un contrôle important surtout sur les documents numériques acquis dans le cadre de la numérisation.

Les données acquises auprès de personnes et d'institutions qui ne sont pas soumises aux règles de production des bibliothèques, sont en revanche susceptibles de se présenter sous des formes variées, tant en termes de formats de fichiers, de formats de métadonnées lorsqu'elles existent, que de supports de données et de moyens de livraison. Dans un tel contexte, la souplesse et la capacité d'adaptation sont des qualités importantes, de même que la disponibilité de ressources humaines, financière et de l'infrastructure technique requise. Il peut également être nécessaire de fixer des règles afin d'éviter d'avoir à traiter des documents difficilement gérables en raison de leur quantité ou de leurs caractéristiques techniques.

4.2 Représentation des données

4.2.1 Niveaux de données

Sur un support, les données numériques peuvent être considérées à plusieurs niveaux. Ceci est illustré par la figure ci-dessous :

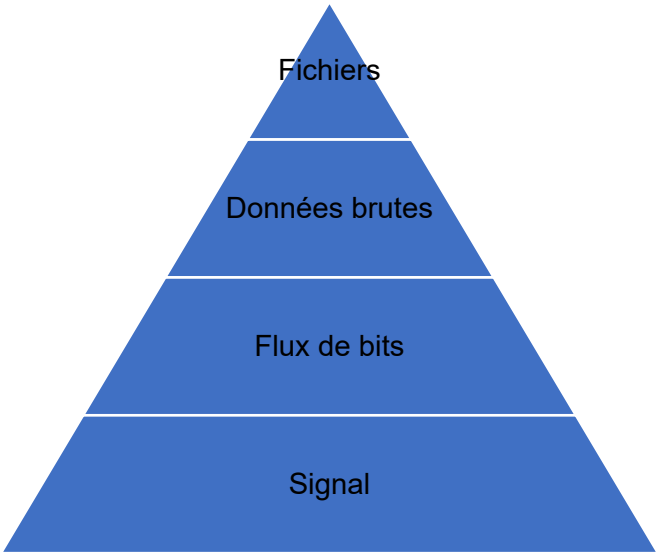

Figure 3: Niveaux de données numériques sur un support

L'information est d'abord codée comme une suite de valeurs binaires enregistrées selon les spécifications du support (par exemple des creux et des bosses sur un CD). Les données lues par le dispositif de lecture adapté correspondent au **signal**, qui est transformé en un **flux de bits** par le dispositif de lecture. Le flux de bits est constitué de **données brutes** et de données de contrôles (sommes de contrôle ou *checksums* par ex.), alors que les données brutes sont elle-même constitutives des fichiers ou de métadonnées techniques (date de modification par ex.).

Plus le niveau est bas, plus la quantité de données est grande.

Il est nécessaire de prêter attention à cette structure pyramidale pour traiter et archiver les données ainsi que les métadonnées de manière pertinente. Cela concerne tout particulièrement les données livrées sur des supports originaux.

4.2.2 Format de fichiers

Pour les fichiers produits en interne ou sur mandat de la bibliothèque, il est nécessaire d'en préciser les caractéristiques attendues. Celles-ci doivent être définies et formalisées en prenant en compte les impératifs de l'archivage numérique.

Il n'y a pas de standards universels dans le domaine. Toutefois beaucoup d'institutions disposent de standards qu'il est possible de prendre comme point de départ pour créer son propre modèle.

Dans d'autres cas, il est souvent impossible d'exiger des fichiers respectant des formats et des caractéristiques spécifiques. Les actions possibles suivantes sont en revanche possibles :

- **Informer**
  Informer le fournisseur de données à propos des standards et des exigences en vigueur localement.

- **Sensibiliser**
  Mettre en place des actions de sensibilisation auprès des producteurs de documents. Cela peut permettre de faire connaître des bonnes pratiques à même d'assurer un meilleur archivage des données produites. A titre d'exemple, nous pouvons noter les recommandations que Memoriav met en place et qui sont pour partie destinés à des publics sans expertise particulière[10].
- **Créer des fichiers dérivés**
  Lorsque des fichiers n'entrent pas dans la politique de conservation mise en place ou ne correspondent pas aux possibilités techniques d'un système d'archivage, alors il est nécessaire de déterminer si la création de fichiers dérivés plus aptes à l'archivage permet de résoudre la situation.
- **Refuser des fichiers**
  Pour éviter d'avoir à accepter et traiter des fichiers qui nécessiteraient un travail disproportionné, il peut être utile de mettre en place une politique d'acquisition qui fixe également des limites aux types de documents et de fichiers acceptés.

Lorsque le système d'archivage ou les standards de la bibliothèque prévoient un format spécifique pour les SIP, il est certainement utile de communiquer cette information au fournisseur de données. Dans le cas le plus favorable, le fournisseur de données pourra transmettre des SIP valides.

4.2.3 Format de métadonnées

La réception de documents numériques doit autant que possible être accompagnée des métadonnées correspondantes. Cela permet de créer aisément les notices descriptives de ces documents. En particulier dans le cas des documents d'archives, cela permet aussi d'obtenir des informations qu'il ne serait pas possible d'acquérir autrement. Par exemple, un photographe peut attribuer un titre et une description à ses photographies dans son outil de gestion. Ces informations peuvent se perdre si elles ne sont pas automatiquement intégrées dans les photographies ou transmises dans un fichier séparé.

Tout comme les fichiers, il n'est souvent pas possible de recevoir les métadonnées dans un format qui permet un import direct dans le catalogue de la bibliothèque ou dans le système d'archivage. Il est alors nécessaire de créer une mise en correspondance du modèle de données originale avec le modèle de données de la bibliothèque (mapping).

A noter que les métadonnées peuvent être externes aux fichiers numériques acquis mais que de nombreuses métadonnées existent aussi dans le fichier lui-même : date de création, date de dernière modification, propriétaire du fichier, etc. Il est nécessaire d'exploiter ces deux types de métadonnées afin de rassembler, sauvegarder et utiliser les informations les plus complètes.

4.3 Transfert des données

4.3.1 Transfert en ligne

Le transfert de documents numériques peut se faire en ligne selon différentes modalités.

Quels que soient les systèmes de transfert de données mis en place par l'institution, il faudra au sein du projet prévoir en amont une infrastructure afin d'accueillir ces fichiers pendant toute la

[10] https://memoriav.ch/fr/recommandations

phase de traitement pour l'archivage. Si une équipe informatique est disponible au sein de l'institution, il est fortement recommandé de leur demander une estimation des espaces de stockage temporaires requis. Cette évaluation concerne la taille de stockage annuelle des fichiers à accueillir et le temps nécessaire pour le traitement d'archivage, afin d'avoir suffisamment d'espace libre en permanence.

4.3.2 Système de stockage et de partage de fichiers

Un espace commun entre le fournisseur de données et la bibliothèque peut être mis en place pour qu'un transfert puisse avoir lieu de l'un à l'autre.

Cet espace peut être mis en place par le fournisseur de données ou la bibliothèque, via par exemple un serveur SFTP. Il peut aussi être fourni par des acteurs tiers, dont des services plus ou moins connus : Dropbox, Google drive, kDrive, OneDrive, Proton drive, Switch drive, etc. Le service choisi doit évidemment respecter les exigences de la bibliothèque et du fournisseur de données, notamment en matière de protection des données.

4.3.3 Système d'envoi de fichiers

Des systèmes spécialisés dans l'envoi de fichiers avec des fonctionnalités spécifiques peuvent également être utilisés : SwissTransfer, Switch Filesender, WeTransfer, par ex.

Comme pour les systèmes de stockage, il est également nécessaire de vérifier que le service choisi respecte les exigences de la bibliothèque et du fournisseur de données.

4.3.4 Support de données

Un transfert de données peut également avoir lieu à l'aide d'un support physique livré à la bibliothèque par le fournisseur de données. Il peut s'agir d'une simple clé USB ou d'un disque dur externe, voire aussi de supports devenus aujourd'hui obsolètes.

4.3.5 Automatisation

Lorsqu'une organisation (un éditeur par exemple) fournit régulièrement des documents à une bibliothèque, il est intéressant de vérifier s'il est possible d'automatiser la transmission des données.

Dans le cas le plus favorable, la bibliothèque peut recevoir des données et les intégrer dans un SAE sans aucune intervention manuelle. Par exemple, ceci est possible techniquement si, d'une part, le fournisseur de données met à disposition une API permettant d'obtenir ces données et, d'autre part, le système d'archivage dispose d'une API permettant de sauvegarder les données dans le système.

Automatiser un tel processus peut être réalisé en interne si la bibliothèque dispose de développeurs ou de développeuses informatique au sein de son personnel. Si tel n'est pas le cas, des offres peuvent être sollicitées auprès d'une ou plusieurs sociétés informatiques afin d'en étudier la faisabilité.

4.4 Lecture des données

La livraison des données et des métadonnées, ainsi qu'un contrôle de leur lecture, est une étape importante à laquelle les ressources nécessaires pour garantir leur qualité doivent être consacrées.

Avant toute livraison de données, il est important de demander au déposant de réaliser un pré-inventaire. Cette étape permet de préparer le travail de réception des pièces comme :

- vérifier à l'issue du versement que la totalité a bien été reçue;
- repérer en amont des fichiers qui ne semblent pas pertinents à conserver; vous pourrez communiquer en amont du versement et demander au déposant par exemple s'il ne dispose pas de formats de fichiers alternatifs pour les mêmes objets;
- proposer d'aller récupérer les objets in situ au cas où les quantité de données seraient trop importantes.

En procédant de cette manière, on assure que les données pourront être accueillies sans entraves.

De nombreuses difficultés peuvent apparaître pour lire et copier les données à partir d'un support.

Un disque dur externe n'est pas forcément compatible avec tous les ordinateurs, en raison du système d'exploitation (Mac, Windows ou Linux) et du système de fichiers utilisé (HFS+, NTFS, etc.). Il peut aussi y avoir une incompatibilité matérielle au niveau du connecteur (USB-A, USB-C, Thunderbolt, etc.).

Lorsque le support de données est obsolète, copier les données peut demander un effort supplémentaire. R. François et R. Rochat[11] proposent un processus en plusieurs étapes, soutenus par des outils logiciels et matériels. Il est également possible de mandater une entreprise spécialisée pour ce travail.

Ce processus suit les étapes suivantes:

1. **Localiser et inventorier les supports de données**.
   La conformité des supports reçus est vérifiée avec le pré-inventaire.
2. **Identifier le type de support et préparer la lecture.**
   Par exemple, le matériel de lecture nécessaire pour un support spécifique peut manquer et doit être acheté. De plus, la procédure de lecture peut nécessiter des tests préalables si elle n'est pas encore en place.
3. **Générer une image du support.**
   Il s'agit de créer une image au niveau le plus bas (voir Figure 3 en p. 11) en fonction des besoins et des possibilités techniques. L'image d'un support est un fichier contenant l'intégralité des données présentes sur ce support. L'article de R. François et R. Rochat détaille le matériel et les logiciels utiles pour cette étape.
4. **Extraire les fichiers et les métadonnées de l'image du support.**
   Il est nécessaire de commencer par extraire les fichiers et métadonnées pertinentes avant de travailler avec ces données. Le logiciel Aaru[12] est un exemple d'outil pour cette étape lorsqu'il prend en charge le système de fichier du support de données considéré.

[11] Robin François et Rebecca Rochat, "Digital Preservation Pipeline for Data Storage Media at the Cinémathèque Suisse. Imaging and extracting data and metadata from Special Collections media", in: *18th International Conference on Digital Preservation*, 2022.

[12] https://www.aaru.app

5. **Convertir, trier et documenter.**
6. **Archiver.**

4.5 Acquisition des données

4.5.1 Principes et précautions à prendre

De manière similaire à l'acquisition d'objets analogiques, l'acquisition des objets numériques suit plusieurs étapes. Deux situations :

- les objets numériques présents dans un fonds d'archives;
- les objets numériques issus de la numérisation ou d'une GED.

Objets numériques issus d'un fonds d'archive

Un fonds d'archives est un ensemble de documents constitués par un ou plusieurs producteurs. La nature de cet ensemble dépend du contexte de production.

Par exemple, tout au long de son existence, un bureau d'architecture produira des maquettes, des plans, des photographies, des films, des factures, de la correspondance. Ces documents auront été produits à différentes périodes et pour des destinataires variés. Par conséquent la diversité des documents numériques est importante. Si le début de l'activité du bureau est antérieur aux années 2000, il est très probable que les plans ou les modélisation 3D aient été réalisées sur des logiciels qui n'existent plus aujourd'hui.

De plus l'organisation du travail, les flux de documents à l'interne, et leur stockage varient d'un bureau à un autre. Il est assez rare dans les années 1990 que les bureaux d'architecture soient équipés de DAM ou de GED. Ainsi chaque collaborateur organise le stockage de ses fichiers de manière différente, même si plusieurs personnes collaborent sur un même projet. Par conséquent, un même document peut se trouver à de multiples emplacements, être nommés de manières différentes, ou encore être représentés par des fichiers de formats différents. Ainsi un plan peut être représenté par un fichier Maya (Autodesk), puis exporté dans un fichier JPEG, PNG ou encore PDF qui aurait été envoyé à un collaborateur externe.

Par conséquent, la diversité des fichiers peut être importante, le nombre de doublons non négligeable, le nommage des fichiers hétérogènes, avec très peu de métadonnées.

Objets issus de la numérisation ou d'une GED

Dans le contexte de la numérisation, le commanditaire décrit de façon précise le format des fichiers, la résolution des images, le taux de compression éventuel, les métadonnées et d'autres métadonnées spécialisées incorporées (espaces de couleurs par ex.). Ces règles normalisent les types de fichiers qui seront acquis et conservés dans l'archive.

Par conséquent, dès la réception des objets, un contrôle de conformité aux standards exigés peut être effectué. Les fichiers qui ne répondent pas au cahier des charges de numérisation pourront être rejetés.

Au sein d'une GED, des processus permettent également d'assurer des normes définies :

- chaque fichier appartient à une liste restreinte de formats;
- les doublons sont évités;
- le nommage des fichiers est contrôlé;

- les métadonnées nécessaires (date de production, versioning, producteur, etc.) sont présentes.

Au-delà des spécificités expliquées ci-dessus, le processus d'acquisition suit les étapes suivantes :

- établissement d'un pré-inventaire;
- élaboration d'une convention de don, de dépôt afin de clarifier les droits d'auteur, d'exploitation, et de diffusion des objets numériques.
- versement;
- constat d'état des supports et des fichiers numériques, par sondage;
- sélection des objets à conserver et à éliminer;
- ingest;
- archivage.

4.5.2 Pré-inventaire

Avant de procéder à l'acquisition d'un fonds, il convient d'établir un pré-inventaire en collaboration avec le donateur ou le producteur. Cela permettra d'obtenir une première évaluation des données sous différents aspects :

- *Diversité et du nombre des supports de données*. Il convient de s'assurer que l'accès aux différents supports de données sera possible : disquette ZIP, syquest, CD-R, etc.
- *Matériels (hardware) encore en état de marche ou potentiellement réparables* pour lire certains supports de données,
- *Diversité des formats de fichiers*. Cette évaluation permet d'anticiper les éventuels besoins en logiciels pour lire ces fichiers.
- *Logiciels spécialisés éventuellement encore présents dans le fonds d'archives*. Il est en effet précieux de pouvoir conserver les logiciels contemporains de l'époque de production. Dans de tels cas, il est important de conserver toute trace qui documente ces logiciels (factures, correspondances). Par ailleurs, il est nécessaire de conserver les clefs et mot de passe relatifs aux licences. Ces logiciels peuvent être parfois compliqués à trouver, leur usage permet un accès "original", mais peuvent aussi exiger une émulation du système d'exploitation correspondant.
- *Localisation des supports de données*. En effet il n'est pas rare que des fonds d'archives soient disséminés sur plusieurs lieux de production ou d'entreposage.
- *Fréquence des téléversements* pour chaque type de donnée dans le cas d'une GED, étant donné qu'il s'agit d'un flux continu. Cette information permet d'anticiper les ressources nécessaires pour traiter ce flux.
- *Sélection des objets à conserver ou à éliminer*.
- *Stratégie pour le transfert des documents vers l'institution*. Elles permettent de répondre aux questions suivantes :
  - Comment est faite la copie des données ? est-elle faite dans les locaux du producteur par un archiviste numérique ? Par ex. une réponse possible est la réalisation d'une image disque des données. Une autre serait d'utiliser rsync pour réaliser une copie des données, puis les sauvegarder sous forme de bags selon la spécification BagIt.
  - Demande au producteur de faire des copies sur disque dur ?
  - Transfert des supports originaux : CD-R, Digital S, DVD-R, disque dur, NAS, etc. ?
  - Envoi des fichiers via un serveur SFTP ?
  - Envoi des fichiers via une API ?

- Comment l'intégrité des données est assurée lors du transfert ?

La précision et la granularité avec laquelle cette pré-évaluation est réalisée dépend de l'ampleur et de la taille du fonds d'archive ou de la complexité de la GED. Plus le fonds est important plus la granularité est grossière. On réalise généralement quelques sondages pour en extrapoler une description générale.

4.5.3 Versement

Les versements d'objets numériques peuvent commencer, une fois le pré-inventaire effectué. Certains objets sont priorisés, d'autres peuvent être rejetés, s'ils sont déjà présents dans la collection. Les versements peuvent être effectués in situ chez le producteur, ou être transférés vers l'institution patrimoniale.

Dans le cas d'une GED, les versements sont effectués de façon plus ou moins continue, le plus souvent en ligne.

A noter que le transfert des objets numériques vers l'institution n'est pas toujours la meilleure stratégie. Par exemple, le producteur peut copier les objets numériques vers un support de données temporaire (disque dur) pour en assurer le transfert, mais ce faisant des informations essentielles peuvent être perdues ou altérées telles que :

- dates de création des fichiers;
- problèmes de jeu de caractères pour le nommage des fichiers et dossiers;
- structure hiérarchique originale des dossiers depuis laquelle les fichiers sont copiés;
- dépendances éventuelles vers d'autres fichiers localisés sur d'autres emplacement qui ne font pas l'objet du dépôt en cours;
- intégrité des fichiers;
- etc.

Suivant le contexte de production, il convient d'établir un protocole avec les ressources matérielles et humaines pour la récolte in situ de ces objets.

4.5.4 Constat d'état

Le constat d'état d'un fonds d'archive numérique est réalisé en deux temps. L'objet matériel est évalué dans un premier temps, puis l'objet numérique dans un second temps.

Objet matériel

Le constat d'état intervient dès le premier versement effectué. Chaque support de données doit être inventorié, identifié (CD-R, DVD-RAM, disque dur externe, etc.) et faire l'objet d'un constat d'état sommaire (moisissure, bris, rayure, empoussièrement).

Objet numérique

Le constat d'état de l'objet numérique précède l'ingest, c'est à dire le moment où les SIP sont préparés. On peut le définir comme une série d'opération de diagnostic, d'analyse et de vérification de l'état des fichiers numériques. Pour éviter toute perte ou transformation des données lors du constat d'état, il est nécessaire d'effectuer une copie de travail. Pour plus de sécurité, une autre copie peut être réalisée et servir de copie de sauvegarde.

Étant donné les grandes quantités de fichiers pouvant constituer un fonds d'archives numérique, il n'est pas toujours possible de vérifier la lisibilité de tous les fichiers. Le constat d'état concerne alors notamment les opérations suivantes : extraction des métadonnées, identification des formats de fichiers, validation de ces formats de fichiers, vérification de la lisibilité du fichier.

- *Extraction des métadonnées*
  Certaines métadonnées spécialisées peuvent être vérifiées par extraction. Il est par exemple possible d'extraire les métadonnées des images PNG et JPG dans un fichier csv au moyen de l'utilitaire exiftool :

  ```
  exiftool -ext jpg -ext png -r -csv "/répertoire/source/desImages" > listeMetadonneeI-
  mage.csv
  ```

  Une extraction réussie ne garantit pas que l'image est elle aussi accessible, mais fournit déjà une première indication sur l'état du fichier.

- *Identification des formats de fichier*
  L'extension d'un fichier est un premier indice sur sa nature, mais ne constitue pas une identification, car l'extension peut facilement être modifiée : par exemple un fichier .tiff est renommé en .txt. Des fichiers peuvent aussi ne pas présenter d'extension.
  Des outils open source peuvent être utilisés pour identifier les fichiers :
  - Fido (Format Identification for Digital Objects) est un programme en python développé par la fondation *Open Preservation*[13].
  - DROID est un outil développé par les National Archives du Royaume-Uni. Il a l'avantage de disposer d'une interface graphique (GUI), mais s'utilise aussi en ligne de commande. Comme Fido, Il utilise la base de données des National Archives comme référentiel (Pronom).
  - FITS est un programme Java pour l'identification des fichiers qui est notamment très utilisé côté serveur via son API. C'est un outil rapide, et il a aussi l'avantage d'agréger plusieurs outils d'identification sous une seule application java comme DROID, JHOVE, Exiftool ou encore Tika (liste non exhaustive). En revanche, il a l'inconvénient d'utiliser beaucoup de mémoire vive.
  - Siegfried est aussi un outil basé sur la base de référence PRONOM. Son installation est décrite sur le site *IT for archivists*[14].

- *Validation des formats de fichiers*
  Elle doit être distingué de l'identification des formats de fichiers. Elle consiste à vérifier que la structure du fichier corresponde à un standard ou une norme. Par exemple, un fichier peut être un PDFA/1b lisible, exploitable mais peut contenir des métadonnées ou des éléments qui ne respectent pas la norme PDFA/1b. Un fichier peut donc être identifié comme étant un PDFA/1b sans respecter toutes les spécifications de la norme PDFA/1b. Les outils de validation sont souvent spécialisés pour un type de format de fichier. D'une part la validation est un processus plus ou moins aisé suivant la complexité du format de fichier, d'autre part les logiciels et le référentiel ne sont pas toujours à jour. Par conséquent, il est fréquent que les résultats donnés par différents logiciels de validation différent ne soient pas identiques. Il convient d'interpréter les résultats au cas par cas. Les logiciels open sources les plus communs sont :
  - Fichiers PDF: VeraPDF, JHOVE. Il existe d'autres solutions propriétaires comme PDFtools;

[13] https://openpreservation.org. Voir le guide d'installation ici https://github.com/openpreserve/fido#installation.
[14] https://www.itforarchivists.com/siegfried/

- fichiers TIFF: DPFManager, libtiff, JHOVE;
- fichiers JPEG: Bad Peggy, JHOVE;
- Fichiers JPEG2000: jpylyzer.

- *Liens et dépendances*
  Certains fichiers comportent des liens pointant vers d'autres fichiers. Ainsi il n'est pas possible d'afficher ou de lire correctement un fichier si les liens vers d'autres fichiers sont rompus ou si ces fichiers secondaires sont absents de l'archive. Le fichier principal peut ne pas être lu ou seulement en partie : des cadres vides apparaissent dans le cas d'image, des parties restent muettes dans un montage audiovisuel, par ex. Il convient alors de faire un constat d'état de ces liens afin de rechercher à travers l'archive l'existence de ces dépendances.

4.5.5 Submission information package (SIP)

Lorsque la bibliothèque dispose d'un système d'archivage, il est utile de créer un SIP à la réception des données, si elles ne sont pas déjà livrées sous cette forme.

Ce SIP contient les sommes de contrôle pertinentes permettant le suivi régulier de l'intégrité des données.

En l'absence d'un système d'archivage et de tout standard pour les SIP, il est recommandé d'utiliser le format BagIt[15], établi en collaboration par la Library of Congress et la California Digital Library. Divers outils existent pour créer et gérer des « bags » dans le respect du format BagIt, tel que Bagger mis à disposition par la Library of Congress[16].

4.5.6 Besoins matériels (local de tri)

Avant de pouvoir préparer les SIP qui seront traités par le système d'archivage, il convient d'aménager des espaces de stockage temporaire (« local de tri » ou « local des dons ») :

- pour les supports physiques: armoire, étagère;
- pour les fichiers à extraire: serveur(s) de stockage

Les supports physiques et les fichiers peuvent rester entreposés pour un temps plus ou moins long. En effet, en fonction des ressources de l'institution, une mise en attente sera nécessaire. Une capacité de ces espaces de l'ordre d'une à deux années est une bonne recommandation.

Ainsi les caractéristiques de l'espace dévolu au stockage temporaire des objets numériques sont déterminées comme suit :

- Évaluation de la masse de données à archiver annuellement;
- Calcul de l'espace de stockage temporaire avant archivage;
- Établir des procédures de droit d'accès à cet espace pour les collaborateurs (différencier les droits d'écriture des droits de lecture);
- Penser à la localisation et à la propriété des serveurs de stockage par rapport au niveau de protection (RGPD) et de sensibilité des données. Par exemple, toute entreprise américaine

[15] http://www.digitalpreservation.gov/documents/bagitspec.pdf
[16] https://github.com/LibraryOfCongress/bagger

via le Cloud Act se doit d'accorder l'accès aux données si un juge américain le demande, même si ce serveur est localisé en Suisse[17].
- Mettre en place une procédure de backup de cet espace.

Il est possible de s'inspirer du *Minimum preservation tool* proposé par la British Library pour établir un local de tri[18].

4.5.7 Besoins en ressources humaines

Les ressources humaines devant être disponibles pour chaque étape de l'acquisition peuvent être identifiées en fonction des rôles suivants :

- Conservateur pour l'évaluation de la pertinence du fonds ou des objets;
- Spécialiste en droit d'auteur et sur les données à caractère privé;
- Restaurateur ou archiviste numérique pour l'évaluation des supports et des fichiers;
- Spécialiste IT pour la mise à disposition d'espaces de stockage temporaire, la mise en place des sauvegardes, la gestion des droits d'accès. Si ces ressources manquent en interne, des espaces de stockage peuvent être loués auprès de prestataires externes. Il faut veiller cependant à s'assurer des conditions d'accès, comme le coût de la bande passante, les sauvegardes, l'emplacement des données (sur le territoire suisse ou non);
- Spécialiste IT pour le développement d'applications ou de connexions pour le transfert – manuel ou automatique – des objets numériques.

Dans le plan de traitement des acquisitions, les rôles sont attribués à des personnes identifiées en interne ou en externe.

5 Plan de préservation (Preservation-Planning)

5.1 Définition

La planification de la préservation d'objet numérique n'est pas une activité standardisée et son périmètre est plus ou moins défini suivant les institutions. Son but est de formaliser toutes les étapes de traitement dans un document de référence. Celui-ci constitue un guide à l'attention de toutes les personnes qui participent à la préservation des objets numériques.

A minima, cette planification devrait comprendre les étapes suivantes :

- la politique d'acquisition;
- l'ingest;
- l'archivage à long terme;
- la diffusion;
- la migration et la mise à jour de l'archive;
- l'évaluation des procédures;
- l'attribution des tâches principales;
- l'élaboration des plans d'urgence et des risques.

Ce travail est chronophage et relativement complexe, car les questions posées sont transversales à toute l'institution (service informatique, registrariat, spécialiste des droits d'auteurs, conserva-

[17] P. Fischer et S. Pittet, *US Cloud Act - un aperçu*, 08.11.2021. https://swissprivacy.law/101
[18] https://www.dpconline.org/blog/minimum-preservation-tool-mpt

teur et conservatrices, etc.). La ou les personnes responsables de sa rédaction devront faire appel à la collaboration de divers départements ou secteurs de l'institution.

5.2 Politique d'acquisition

Afin de bien comprendre la diversité et l'ampleur des matériaux numériques que l'institution doit préserver, une bonne option est de débuter le plan de préservation par l'analyse des politiques d'acquisition.

Cette analyse permettra de calibrer le volume annuel de nouvelles acquisitions, au travers des différents flux d'entrée dans les collections. Il peut s'agir d'entrées en vertu du dépôt légal, par l'acquisition de fonds d'archives, l'acquisition de pièces isolées, ou le versement régulier de pièces basé sur des conventions avec les donateurs ou d'autre obligations légales comme les services d'état.

Une fois les flux définis, il est intéressant d'évaluer les types d'objets numériques que chacun peut apporter. Cette étude doit être tant qualitative que quantitative (fichiers numériques, supports de donnés).

Dans le cas par ex. d'une institution responsable du dépôt légal, il convient de vérifier les types de documents que celui-ci concerne. S'il s'agit uniquement de documents textuels et photographiques, cette information permet d'établir d'emblée une liste de formats de fichiers susceptibles d'entrer par ce flux : PDFA/1b, ou PDFA/2U, ODT, SVG, TIFF, PNG, JPEG (liste non exhaustive). Par contre si les œuvres audiovisuelles sont également concernées par le dépôt légal, la liste de formats de fichiers s'élargit vers des formats comme : MP4, WAV, MKV, MOV (liste non exhaustive).

Cette vérification doit être également réalisée du point de vue des supports. En cas d'acquisition de données inscrites sur support optique, il convient de s'équiper d'une station de travail capable de lire les CD-R, DVD-R, Blu-ray, etc. Si l'acquisition de fonds photographiques des années 1990 constitue une activité centrale de la bibliothèque, alors l'achat de lecteur ZIP ou JAZZ (Iomega) pour accéder à ces supports d'information serait à envisager.

Une évaluation quantitative est préconisée pour estimer le volume de données à traiter annuellement, détaillée idéalement pour chaque type de document, comme :

- nombre d'heures de fichiers audio ou vidéo;
- nombre de pages de documents texte;
- nombre de photographies;
- etc.

Une analyse des métadonnées descriptives minimales nécessaires à cette étape doit être également faite pour renseigner au plus tôt dans le processus d'acquisition les métadonnées utiles pour le suivi des objets dans le processus d'acquisition.

L'analyse de la politique d'acquisition est conduite en collaboration avec les responsables des services suivants :

- collections;
- dépôt légal (si existant);
- services versants des documents.
- métadonnées (si existant)

- droits d'auteurs (pour l'établissement de conventions).

5.3 Processus d'acquisition des objets numériques

Il convient de définir dans le plan de préservation le processus de récolte des objets numériques, lorsqu'il n'est pas piloté par une GED ou un processus automatique de versement. Deux scénarios sont possibles (voir 4.5.3 Versement) :

- le producteur transmet les objets numériques;
- la bibliothèque récupère les objets chez le producteur.

Chaque scénario a des avantages et des inconvénients. Les contextes peuvent être variés et différents facteurs rentrent en ligne de compte, notamment la compétence du producteur, le type de support sur lesquels les objets numériques sont inscrits, les systèmes de fichier de ces supports.

5.3.1 Contextes techniques chez le producteur

Le producteur peut avoir stocké les objets numériques sur des supports variés :

- NAS;
- disque optique;
- disque externe (mécanique ou électronique);
- support amovible magnétique sur disque (disquette, ZIP, Syquest, etc.);
- support amovible magnéto-optique (miniDisc);
- disque dur interne (ordinateur);
- clefs USB;
- carte électronique (SD, smartMedia, microSD, etc.);
- serveur cloud ou relié à une application web.

5.3.2 État de conservation

L'état de conservation des supports d'information peut fortement varier. Ceux-ci peuvent exiger un traitement particulier pour permettre d'accéder aux informations qu'ils contiennent.

5.3.3 Obsolescence

Des supports d'un certain âge peuvent nécessiter des lecteurs spécifiques qui ne sont plus présents chez le producteur, ni à la bibliothèque. Par ex. les disquettes ZIP ou Jazz de la marque Iomega requièrent des lecteurs particuliers que les producteurs ne possèdent peut-être plus. Enfin, dans l'éventualité où ce dernier existe, encore faut-il disposer du bon matériel pour assurer leur connexion avec le matériel informatique contemporain.

5.3.4 Système de fichier et accès à l'objet numérique

Le support de l'information ne décrit pas comment l'information y est inscrite. Par ex. un CD-R peut stocker des documents numériques de différentes manières, selon le système de fichier utilisé : CDDA, ISO9660 (plusieurs variantes existent), CD-i (interactive), SACD.

Afin de préserver l'intégralité de l'information, il convient dans la mesure du possible de récolter non seulement les documents numériques mais le système de fichiers dans son ensemble. Cette opération peut être plus ou moins complexe, mais elle permet d'assurer de conserver toutes les métadonnées. Elle permet également de re-présenter l'objet numérique dans son apparence originale. Par ex. dans le cas d'un DVD, une récolte de l'objet numérique film ne permet pas de conserver les menus du DVD, alors qu'une récolte du système de fichiers dans son ensemble permet de présenter les menus, qui sont des fichiers différents de ceux du film.

5.3.5 Préparation des données et des supports

A l'exception de la GED ou éventuellement de la numérisation pour lesquelles le respect de spécifications déterminée peut être exigé, il est rare que les fichiers soient susceptibles d'être envoyés directement dans le système informatique qui réalise l'ingest. Des travaux préparatoires sont nécessaires, parfois nombreux, complexes et chronophages, qu'il convient de prendre en compte dans le plan de préservation.

5.3.6 Pré-ingest

Toutes les opérations de préparation des paquets ainsi que la normalisation doivent être documentées, idéalement au moyen des standards existants comme le METS ou PREMIS. Les outils choisis doivent être capables de récupérer les métadonnées des objets numériques et de les écrire dans un fichier METS.

Il convient d'établir un plan de gouvernance des métadonnées qui définit quelles métadonnées descriptives et spécialisées sont nécessaires pour l'archivage.

5.4 Ingest

5.4.1 Définition

L'ingest est un ensemble d'opérations dont la finalité est de documenter et de transformer les objets numériques en un paquet de données archivable (AIP) ainsi qu'en un paquet de données de diffusion (DIP). L'ingest est une des étapes centrales de l'archivage.

Les notions suivantes lui sont associées :

- **Producteur**
  Le producteur est l'agent qui a réuni les objets numériques qui sont transmis pour l'archivage. Il n'est pas nécessairement l'auteur de l'œuvre ou un participant à l'élaboration de l'œuvre.
  Dans le contexte archivistique, ce terme ne doit pas être confondu avec le producteur d'une œuvre audiovisuelle, dans laquelle par exemple l'étalonneur et le monteur sont deux producteurs de l'objet film numérique (l'étalonneur produit la version finale du film à la suite du monteur).
- **Conversion**
  La conversion est la transformation du format original d'un fichier en un autre format, dans le meilleur des cas sans que le contenu en soit modifié. Cependant la conversion entraîne toujours une modification de l'information du fichier.
  Par exemple une conversion est réalisée lorsqu'on exporte un fichier Excel vers un nouveau fichier csv.

La conversion peut représenter une perte d'information ou non, par exemple dans le cas d'images numériques.

- **Migration**
  La migration est une opération qui consiste à transformer un fichier dans un même format de fichier, mais pour lequel les spécificités sont différentes. Généralement, la migration est effectuée d'un format de fichier ancien vers une version plus récente. Exemple: conversion d'un format TIFF 5.0.

5.4.2 Opérations de l'ingest

Les principales opérations réalisées dans le cadre de l'ingest sont les suivantes (liste non exhaustive) :

- Identifier les fichiers et la structure de l'organisation de l'archive ou du versement original.
- Identifier les formats de fichiers.
- Calculer la somme de contrôle (checksum) des fichiers. Si un bag a été utilisé comme paquet SIP, cette somme de contrôle sera comparée avec celle du SIP, afin de s'assurer que l'intégrité des données est conservée depuis la phase d'empaquetage SIP à celle de l'ingest.
- Sélectionner les fichiers, supprimer des fichiers inutiles comme les « .db » ou certains fichiers cachés des systèmes de fichiers.
- Récolter les métadonnées spécialisées des fichiers.
- Convertir les fichiers sources en fichiers de préservation et/ou de diffusion.
- Valider les fichiers convertis. Une validation des fichiers sources peut être réalisée en option.
- Enregistrer toutes les opérations précédentes dans un fichier qui accompagnera l'AIP. Ce fichier doit être lisible dans un format non propriétaire et si possible sous forme standardisé afin d'être validé à la fin du processus ingest. Le standard METS est conseillé.
- Réaliser l'ingest. De nombreux outils propriétaire ou open source existent à cet effet.
- Stocker l'AIP et le DIP dans le SAE. Le stockage à court ou long terme est une opération réalisée au terme de la chaîne ingest, mais il peut être complété ou modifié en dehors de l'ingest. La question du stockage à long terme est évoquée à la section suivante.

Chaque opération présentée ci-dessus peut être réalisée par des logiciels qui automatisent ces tâches. Chaque opération est paramétrable selon les standards et recommandations que l'institution souhaite suivre.

Il existe différents outils propriétaires ou open source pour la réalisation de l'ingest. Les Archives nationales britanniques en ont établi une liste non exhaustive[19].

5.5 Stockage pérenne

Une fois l'AIP produit à l'issue de l'ingest, il convient de le stocker sur un support pérenne. Différentes solutions techniques existent. Leur analyse est généralement réalisée par le service IT de la bibliothèque ou de l'administration dont elle dépend, en collaboration étroite avec les responsables du plan de préservation.

[19] https://cdn.nationalarchives.gov.uk/documents/archives/digital-preservation-repository-systems-for-archives.xlsx. Fedora ne permet pas l'ingest mais davantage un repository.

Le stockage pérenne peut être interne à l'institution ou externalisé. Dans tous les cas, il convient de s'assurer que les prestataires externes répondent à la réglementation en matière de protection des données[20], ainsi qu'aux standards de sécurité, de localisation des serveurs et de backup.

5.5.1 Typologie de stockage

On peut distinguer deux grands types de stockage :

- offline: principalement sur bande LTO;
- online: disque dur.

Les stockages online sont coûteux, énergivores, sensibles aux attaques informatiques, et leur durée de vie est limitée à 5 ans. En revanche ils permettent des accès rapides pour les opérations telles que calcul d'intégrité, copie, suppression.

Les stockages offline sont lents et peu coûteux. Leur consommation électrique est très réduite, ils sont protégés des attaques. La durée de vie du support s'élève jusqu'à 30 ans, en revanche celle de la technologie est plus réduite : 6 à 15 ans.

Il est important de noter que les DIP ne sont pas systématiquement enregistrés sur des stockages pérennes, mais plutôt sur des espaces d'accès rapide dont l'intégrité et la pérennité ne sont pas les qualités premières exigées. En effet, comme ils sont diffusés auprès du public, ils sont davantage tributaires de l'évolution technique du web. La bibliothèque sera amenée plus fréquemment à les changer pour des versions plus en adéquation avec l'écosystème technique du web. Cette adaptation prend la forme d'une migration.

5.5.2 Migration

Une migration de données est une transformation de l'information ou de son organisation sous une forme plus récente. Dans le cas d'un fichier numérique, cela correspond à la conversion d'un fichier dans un format plus récent. Par ex. la transformation d'un fichier Word docx daté de 2014 vers un nouveau fichier docx (version 19.0).

Afin des respecter le modèle OAIS, il convient de conserver les traces suivantes :

- de l'ancien fichier;
- des logiciels et des versions utilisées pour procéder à la conversion;
- des rapports fournis par les outils de validation.

Pour les raisons mentionnées plus haut, les DIP sont susceptibles d'être régulièrement migrés, tandis que la fréquence de migration est plus faible pour les AIP.

Le plan de préservation prévoit l'organisation d'une veille technique, qui permet de décider à quel moment et selon quels standards une migration doit être effectuée.

[20] Ordonnance sur les certifications en matière de protection des données (OCPD). https://www.fedlex.admin.ch/eli/cc/2007/701/fr

5.6 Stratégie de stockage

Le plan de préservation définit précisément comment le stockage des données dans le système est mis en œuvre, afin d'en assurer la conservation sur le long terme.

5.6.1 Préservation du flux binaire

Un flux binaire est une séquence de bits (unité d'information ayant la valeur de 0 ou 1). A différents stades du processus, l'intégrité des données est vérifiée : sur les fichiers (objets numériques ou métadonnées) au moment de l'acquisition, dans le serveur d'archives, sur les backups. Le processus de préservation du flux binaire, calcule la somme de contrôle et compare la valeur avec la somme de contrôle présente dans les métadonnées.

5.6.2 Copies redondantes

Les AIP, DIP ainsi que les étapes intermédiaires de travail sont copiées sur une autre ressource, afin d'améliorer la fiabilité. Le processus de copie peut être synchrone ou asynchrone. Dans le second cas la copie n'est pas immédiatement identique à la source.

5.6.3 Sauvegarde (backup)

La sauvegarde est la copie à un instant donné des objets numériques dans l'archive à des fins de sécurité. On définit la fréquence à laquelle on réitère cette copie et si elle est complète ou incrémentielle (seules les différences par rapport à la copie précédente sont enregistrées, afin de minimiser l'espace de stockage). On précise également si elle se fait au niveau des fichiers, des volumes de disques, ou des machines virtuelles.

La sauvegarde doit être présente sur différents supports à des endroits différents. Selon le cadre institutionnel elle est conservée pendant un temps déterminé.

Mentionnons également le snapshot (instantané), qui est la sauvegarde du système à un moment donné. Plus fréquent que les backups, les snapshots ne sont conservés que pour une durée limitée, sur le même système ou sur un autre.

Le backup permet de rétablir le système à un état antérieur : par exemple la veille au soir. Cela signifie qu'au pire le travail d'une journée risque ainsi d'être perdu.

La combinaison d'un backup et d'un snapshot permet en revanche de reconstituer le système à un moment très précis : heure, minute.

5.6.4 Interdépendance de la sauvegarde et de la préservation du flux binaire

Les cycles de sauvegarde et de préservation du flux binaire doivent être organisés dans le temps.

Si, par exemple, il n'y a pas eu de préservation du flux binaire, alors les erreurs vont être copiées dans la sauvegarde. A la fin de la durée de vie de la sauvegarde (par exemple trois mois), il ne sera plus possible de restaurer les données correctes. Pour cette raison le cycle de préservation du flux binaire doit intervenir à un rythme plus fréquent que trois mois, en comptant le temps pour traiter les erreurs éventuelles.

Dans le cas contraire et en cas d'erreurs, il faut demander une copie de la sauvegarde sur un disque.

5.6.5 Stockage à froid (cold storage)

L'accès à ce type de stockage est très lent. Son but est uniquement d'être présent pour d'éventuelles lectures. Il n'est pas connecté à une application, comme par exemple une salle de lecture virtuelle. Ce mode de stockage ne permet pas de travailler sur les données qu'il contient : pour cela il faut demander une recopie sur un support plus performant. Selon le modèle de coûts du centre informatique, le prix du stockage à froid est faible, mais celui des transactions demandées peut être élevé.

Les bandes LTO sur lesquelles rien n'est effacé est un mode typique de stockage à froid.

Dans le cadre de l'archivage numérique, la question de savoir ce que l'on conserve sur un stockage à froid se pose : les AIP uniquement ou également les DIP ? Le contrôle de qualité est également une préoccupation : comment le vérifier et à quelle fréquence ?

5.7 Veille technique

La veille technique est une évaluation pro-active des évolutions techniques qui apparaissent au cours du temps. Elle est cruciale dans l'objectif d'assurer la continuité de la préservation des archives numériques et de mitiger le risque éventuel de perte d'accès. Elle concerne les matériels, les logiciels, les formats de fichiers, les systèmes de fichiers (*file systems*). Elle doit s'intéresser à tous les éléments techniques qui font partie de l'écosystème techniques qui assure le maintien de l'archive numérique.

Il convient notamment de s'assurer que le matériel soit toujours « supporté » et réparable facilement.

Par exemple dans le cas du stockage à froid, le standard LTO8 évolue actuellement vers le standard LTO9. Les fabricants de lecteurs ou de robots supportent encore la version LTO8. Cependant dans quelques années, les contrats de maintenance deviendront plus onéreux ou les fabricants ne proposeront plus de maintenance des versions LTO8. La veille technique permet de repérer ces évolutions afin de planifier le changement de standard avant de se trouver dans la situation critique où des lecteurs LTO8 ne fonctionneraient plus et ne seraient plus réparables.

La veille technologique concerne aussi les formats de fichiers. Par exemple le codec H265 pour la vidéo est de plus en plus supporté par les principaux navigateurs web du marché. Étant donné son taux de compression important pour un même niveau de qualité, il peut être intéressant de surveiller son évolution afin d'envisager de produire des DIP avec ce codec. Une analyse des différences entre le codec H264 et le codec H265 constitue un exercice typique de veille technologique. La bande passante est presque diminuée par deux.

5.8 Métadonnées

Dans une archive numérique à long terme, les métadonnées bibliographiques ou descriptives ont pour fonction d'identifier clairement un objet numérique enregistré dans l'AIP. Ceci est important à deux points de vue. D'une part, ces métadonnées sont nécessaires à l'identification de l'objet numérique pour sa gestion et son utilisation dans l'archive. D'autre part, une bonne pratique est de fournir à l'AIP toutes les métadonnées descriptives qui servent à l'identification de l'objet nu-

mérique auquel elles sont associées, et pas seulement celles dont l'archive a besoin pour la gestion. Ainsi on augmente les chances d'identifier l'objet numérique de manière univoque même après un export depuis l'archive (ou au moyen de sauvegardes en cas de catastrophe).

5.8.1 Choix d'un standard

Le choix d'un standard de métadonnées dépend du type d'utilisation des archives numériques de l'institution patrimoniale. En général celle-ci emploie à côté du SAE un système de gestion des collections au moyen duquel elle gère les objets sur supports physiques, comme un SIGB) Elle s'aligne pour cela sur les pratiques et normes de description habituelles dans son domaine, lesquelles définissent aussi bien la structure des métadonnées que l'ensemble des règles de saisie (RDA, ISAD(G), SPECTRUM, …). Étant donné l'existence d'un tel système, il est judicieux de l'utiliser également pour la saisie des objets numériques, afin que l'ensemble de la collection soit décrite dans un seul système. Par conséquent les métadonnées des objets numériques sont déjà disponibles dans le format du système de gestion des collections et elles peuvent être réemployées dans le SAE.

Si cette approche est avantageuse car elle permet d'organiser de façon aussi simple que possible les liens entre ces systèmes, elle n'est pas sans poser des difficultés. Par exemple le catalogage des documents d'archives dans un SIGB restreint beaucoup les possibilités de description. Dans certains cas, il faudrait pouvoir saisir des informations descriptives qui ne seraient pas visibles pour les utilisateurs ou utilisatrices, mais qui sont importantes pour la conservation à long terme, comme certaines informations sur l'acquisition ou des descriptions de contenu de données sensibles. De nombreux SIGB ne permettent pas de limiter la visibilité des descriptions.

Par ailleurs, plusieurs systèmes fournisseurs de données (repositories) peuvent être déjà en production dans le contexte de l'institution. Dans un tel cas, il est recommandé de relier ces différents systèmes, de charger leurs métadonnées respectives dans le SAE par le biais d'API et d'établir une correspondance (mapping) entre elles.

Le choix d'un standard de métadonnées dépend aussi de l'utilisation des métadonnées bibliographiques dans le SAE. Ainsi, il est judicieux d'utiliser une norme dans laquelle les métadonnées nécessaires à l'identification des objets numériques soient bien supportées par le SAE. Dans le cadre de l'ingest, les informations qui seront nécessaires à l'administration des archives numériques doivent être transférées à partir des métadonnées employées pour la gestion des collections. Si l'institution patrimoniale impose des exigences dans ce domaine, cela peut avoir des conséquences négatives, comme celle de restreindre le choix d'un SAE, dans un marché où l'offre de prestataires est limitée. Cette situation peut pourtant être contournée, car la plupart des standards de métadonnées modernes sont basés sur XML, et il est généralement facile de développer des mappings lors du développement et de les indexer pour la recherche. Les formats conteneurs comme METS sont en outre ouverts à différentes normes de métadonnées descriptives qui peuvent y être intégrées.

S'il s'agit d'enregistrer les métadonnées bibliographiques comme composants de l'AIP, elles doivent être saisies dans le format utilisé et connu par l'institution patrimoniale.

Quels sont les formats normalisés de métadonnées qui garantissent la lisibilité par machine des structures de métadonnées? Les institutions patrimoniales sont familières de différents formats normalisés spécifiques à un domaine comme MARC21, MODS, EAD, RiC, LIDO. Il n'existe pas de norme interdomaine qui permette de structurer les métadonnées de différents types d'institutions mémorielles sans perte d'informations.

C’est le cas du standard Dublin Core. Les métadonnées spécifiques d’un domaine peuvent certes être converties en Dublin Core, mais la conversion inverse n’est pas possible sans pertes d’information. Il n’est donc pas recommandé d’utiliser Dublin Core pour l’enregistrement des métadonnées bibliographiques des AIP. Il vaut mieux joindre les métadonnées bibliographiques à l’AIP dans le format utilisé par l’institution.

Il faut également veiller à garder accessibles toutes les informations nécessaires pour l’exploitation du format par machine ou pour en comprendre la signification. Sans cela il n’est pas possible de garantir que les métadonnées jointes à l’AIP puissent encore être lues et comprises sur le long terme. Si le système de gestion (SIGB, système d’archivage, système de musée) offre plusieurs possibilités d’exportation pour les métadonnées bibliographiques, il convient de choisir de préférence un format normalisé, répandu au niveau international et bien documenté. Grâce à une documentation existante et mise à jour, les métadonnées exprimées dans ce format ont une probabilité plus grande de pouvoir être encore lues et comprises sur long terme.

Si l’institution patrimoniale préfère définir de façon autonome sa propre structure de métadonnées et de champs, sans référence à un standard international répandu, elle doit impérativement tenir à disposition toute la documentation. Elle assurera les mises à jour du format et de sa documentation. Dans le cas d’AIP autonomes, la documentation du format devrait obligatoirement être déposée dans chaque AIP, et pas seulement quelque part au sein de l’institution.

A l’heure actuelle, les formats conteneurs d’AIP sont en général représentés en XML et il est par conséquent recommandé d’utiliser des formats de métadonnées représentés en XML (par exemple EAD, LIDO ou MODS), ou pour lesquels une version XML existe (par exemple MARCXML). Cela permet d’intégrer les métadonnées bibliographiques directement dans la structure de données de l’AIP, ce que permet le standard XML au moyen d’une combinaison de schémas XML.

La pertinence d’une telle intégration dans le contexte de l’archive numérique doit être discutée et vérifiée avec le fournisseur du système. Lorsque les métadonnées descriptives sont intégrées dans le conteneur XML, elles peuvent être indexées et utilisées pour la recherche au sein de l’archive. Si l’intégration des métadonnées bibliographiques dans le conteneur XML de l’archive n’est pas possible, elles doivent être ajoutées à l’AIP en tant que fichier indépendant. Dans ce cas l’indexation des métadonnées n’est souvent pas possible ou seulement au prix d’un développement supplémentaire. Il faut en être conscient et, le cas échéant, trouver avec le fournisseur une solution qui permette de répondre à l’objectif de pouvoir rechercher les métadonnées descriptives.

5.8.2 Métadonnées techniques ou spécialisées

Les métadonnées descriptives renseignent sur l’objet dans sa dimension intellectuelle et concernent des éléments généraux comme le titre, l’auteur, la durée, la dimension, la date, etc.

Les métadonnées techniques donnent des informations qui dépendent davantage de la spécificité de l’objet, comme le montrent les métadonnées techniques d’une image en mouvement :

- la résolution;
- l’algorithme de compression audio et vidéo (codec);
- l’espace de couleur;
- le niveau échantillonnage des couleurs;
- le nombre d’image par seconde;

- etc.

Ces métadonnées sont généralement inscrites au sein des fichiers numériques. Leur organisation et leur description sont standardisées, afin d'assurer leur lecture par des logiciels variés.

Pour chaque catégorie d'objet, des standards pour la gestion des métadonnées techniques ont été définis. Il en existe notamment pour l'image fixe, l'audiovisuel, les partitions de musique, la 3D:

- métadonnées techniques pour l'image fixe : EXIF, IPTC, XMP;
- métadonnées techniques pour l'audiovisuel : EBUcore, MXF, MPEG7, Quicktime, etc.

5.8.3 Métadonnées pour la préservation

Les informations relatives aux traitements effectués pendant l'ingest et l'archivage constituent des métadonnées utilisées pour la préservation. Dans une acception plus large, on peut ranger dans cette catégorie toutes les métadonnées nécessaires pour assurer la préservation de l'archive. La frontière entre les métadonnées techniques et les métadonnées pour la préservation n'est pas étanche. Leur différence ne réside pas dans leur nature, mais dans leur usage.

Par exemple, une métadonnée technique sur la version du codec employé pour l'encodage d'un fichier vidéo est une information qui pourra intéresser un chercheur étudiant le fond d'un vidéaste. Mais il est également pertinent de conserver cette information pour assurer la préservation des objets, car elle est importante pour prendre des décisions de conversion ou de migration de la vidéo. L'usage de cette métadonnée technique dans le contexte de la préservation diffère donc de son usage dans le contexte de la production (EBUcore). Pour cette raison, d'autres standards ont été créés pour répondre au mieux aux besoins de la préservation. Trois standards sont communément utilisés : METS, PREMIS, SEDA (France).

6 Accessibilité et utilisation

Les objets numériques archivés numériquement peuvent être rendus accessibles de différentes manières. Dans ce qui suit, il n'est pas fait de distinction entre un accès par téléchargement ou par streaming, particulièrement utile pour les fichiers vidéos et audios.

Premièrement, le SAE peut proposer une interface propre permettant d'accéder aux fichiers de l'archive. Il s'agit d'un module tel qu'un lecteur ou player, spécialement conçus à cet effet.

Une seconde possibilité consiste à utiliser un portail propre, relié à l'archive par une interface. Les objets archivés sont alors consultés et téléchargés par les utilisateurs via ce portail. Le transfert des fichiers des archives vers le portail peut se faire de manière automatisée ou par un contrôle humain, selon les exigences du service d'archives.

Une troisième possibilité consiste à rendre accessibles les objets archivés via des portails externes. Cela suppose toutefois qu'il existe une autorisation d'utilisation et que les types de fichiers soient compatibles avec les spécifications du portail. Il existe de nombreuses plates-formes, souvent spécialisées par médias ou par thématiques : manuscrits, affiches, audiovisuels, etc. Elles uniformisent les descriptions et facilitent la recherche scientifique dans leurs domaines.

Il subsiste dans tous les cas un « reste » de fichiers qui ne peuvent être rendus accessibles et utilisés selon l'une ou l'autre de ces méthodes. La difficulté peut provenir du format de fichiers. Ainsi, les fichiers dans des formats de données spécifiques à certaines professions ou à certains

domaines peuvent être difficilement accessibles ou ouverts en ligne. Citons par ex. les partitions musicales créées numériquement, les fichiers créés à l'aide d'un outil de publication assistée par ordinateur comme InDesign, ou encore les plans d'architecture qui nécessitent des outils de consultation ne pouvant pas être intégrés dans l'interface de présentation des archives numériques à long terme. Une solution consiste à migrer les fichiers dans un format courant tel que le PDF, soit lors de leur consultation dans le SAE, soit le cas échéant lors du téléchargement.

Les fichiers soumis à des restrictions d'accès légales, telles que des délais d'embargo, ne peuvent pas non plus être réutilisés facilement. Il en va de même pour les fichiers qui, pour des raisons de droits d'auteur, ne peuvent pas être consultés librement sur Internet, mais dont l'accès doit être strictement contrôlé et limité. Il est alors judicieux d'envisager la mise à disposition dans un espace en ligne sécurisé. Une alternative est la consultation sur place via un poste de travail dédié dans les murs de la bibliothèque. Il existe toute une série de possibilités pour mettre en œuvre techniquement des restrictions d'accès sur un poste, par exemple en bloquant l'ouverture d'un lecteur, ou en ne transmettant pas le fichier au portail de consultation.

Les fichiers dont l'accès et la réutilisation sont restreints posent des exigences particulières à un système d'accès tel qu'un portail. Dans le domaine des archives, le concept de « salles de lecture virtuelles » s'est imposé. L'ensemble du fonds archivé y est répertorié, y compris les objets soumis à un délai d'embargo. Si un fichier est demandé à être utilisé, il est migré dans un format dans lequel il peut être consulté et téléchargé.

Enfin il convient de mentionner le concept d'*Emulation as a service infrastructure (Eaasi)*[21]. Une émulation de l'(ancien) environnement système est créée dans le navigateur, au sein duquel le fichier est affiché. Pour l'instant, Eaasi est encore au stade d'un projet en tant que tel et n'est pas intégré dans un SAE. On peut imaginer qu'Eaasi puisse être installé comme un module d'une salle de lecture virtuelle. Les métadonnées jointes à l'AIP permettent de renseigner sur le type d'émulation devant être fourni. Dans la salle de lecture virtuelle, il est alors indiqué à l'utilisateur qu'une émulation existe. En activant sur le lien, le navigateur ouvre alors l'émulateur correspondant, avec les données originales correspondantes de l'AIP en question.

Grâce au développement et à l'adoption d'Eaasi dans le cadre de l'archivage à long terme, la migration des formats de fichiers pourrait devenir inutile, ce qui faciliterait énormément l'accessibilité purement technique des fichiers au fil du temps. Les dispositions légales relatives à l'accès et à la réutilisation des archives numériques doivent être respectées dans tous les cas. En outre, Eaasi ne rend pas obsolète la migration des fichiers. Celle-ci reste nécessaire dans le cadre du plan de sauvegarde pour garantir la bonne conservation des données.

7 Synthèse et conclusion

En juin 2023, la Confédération a rendu public le *Message culture* pour la prochaine période 2025-2028[22]. Il manifeste la volonté de renforcer la mission de la Bibliothèque nationale dans le domaine numérique. Il propose une révision de la loi sur la bibliothèque nationale, afin de clarifier le périmètre de son action, soit l'information concernant la Suisse, qu'elle soit avec ou sans support physique. Une mesure emblématique est l'introduction d'un dépôt légal numérique.

Le rôle de coordination de la bibliothèque nationale avec les bibliothèques cantonales et les autres institutions mémorielles (archives, musées) dans le domaine de l'archivage numérique est

[21] https://www.softwarepreservationnetwork.org/emulation-as-a-service-infrastructure

[22] https://www.bak.admin.ch/bak/fr/home/themes/le-message-culture.html

réaffirmé. Le contexte est donc particulièrement favorable et encourageant pour la concrétisation ou l'approfondissement des projets de SAE des institutions mémorielles.

Les contenus numériques sont largement prédominants. Non seulement parce que la majorité d'entre eux sont produits d'emblée sous cette forme (« nés numériques »), quel que soit la forme (texte, son, image fixe et image animée), mais aussi parce que le numérique est devenu un support privilégié pour la conservation de certains médias analogiques, notamment audiovisuels, fragiles et peu durables (il s'agit alors de documents « numérisés », générés au terme d'une conversion numérique par le moyen d'un scanner).

Ensemble, tous ces contenus sont donc des témoignages historiques essentiels pour les 20e et 21e siècles. Les bibliothèques cantonales prennent conscience de leur responsabilité sur ce patrimoine.

D'une certaine manière, la situation actuelle pour l'archivage numérique ressemble à celle qui prévalait il y a 30 ou 40 ans dans l'informatisation des fonctions bibliothéconomiques : quelques institutions seulement, notamment universitaires, avaient mis en place un système informatique. Peu à peu les SIGB se sont imposés aux bibliothèques de tous types et de toutes tailles.

Ce guide montre que les institutions ne sont pas démunies pour la concrétisation de leur archivage numérique : les outils, les méthodes et les standards existent. La familiarité avec l'archivage numérique augmente grâce à la formation des professionnel-le-s et les expériences acquises. De plus il existe un tissu d'entreprises, dont certaines spécialisées dans le domaine des bibliothèques et des archives, qui peuvent accompagner les institutions, de même que les hautes écoles ou des organismes tels que KOST/CECO ou Memoriav.

Il est cependant un facteur exogène qu'il n'est plus possible de négliger et qui marque l'ensemble de la société : la crise énergétique et la réflexion nécessaire sur la sobriété de fonctionnement. De même qu'il existe des nouveaux bâtiments de conservation pour les documents physiques capables de réguler les conditions climatiques avec un minimum d'énergie, les SAE devront également être économes en ressources. Cela implique des décisions à différentes étapes du chemin documentaire :

- Le périmètre des contenus électroniques à archiver, imposant une sélection ou un tri.
- Les formats de fichiers qui représentent ces contenus, en ciblant en priorité les contenus audiovisuels, gros consommateurs d'espace de stockage.
- L'équilibre entre les méthodes d'archivage on line ou off line et le choix d'un nombre de copies raisonnables.

Les rédacteur-trice-s de ce guide, issu-e-s de bibliothèques patrimoniales dans tout le pays, sont également convaincu-e-s que les institutions mémorielles ont tout à gagner d'une collaboration dans le domaine de l'archivage numérique. Par l'échange de pratiques, mais aussi la mise en réseau de leurs moyens : on peut parfaitement imaginer par exemple qu'une copie d'archivage sur bande des données patrimoniales de Saint-Gall soit conservée en Valais et réciproquement.

Gageons que des solutions et des modes d'organisation collectifs seront mis en place au cours de ces prochaines années pour assurer l'archivage du patrimoine numérique de tous les cantons, de façon rationnelle et efficiente.

8 Annexe

8.1 Questions clés

8.1.1 Chapitre 2 : Questions générales

Resources (human and materials) inside, financial funds

- Do you have IT resources in your institution from Human Resources and hardware perspectives? (Have you already planned or evaluated how many IT resources could be available for the digital long-term archive?)
- Do you already collaborate with some a network of institutions or mutualize resources for other project (storage facilities, IT resources, conservation lab, exhibition production ...)? (may be placed somewhere else)
- Do you have the funds - and subsequently the personnel - not only to ingest digital resources but to preserve them?
- Is your institution already capable of long-term preservation of digital objects in terms of personnel? If not - are you capable of allocating the personnel?

Scheduling

- Do you have already a deadline or have you scheduled a plan to setup a long-term archival system?

evaluation of the archivable material

- Have you evaluated approximatively the volume of data your institution would preserve? Have you evaluated the increase of your archive per year?
- What kind of data your institution would preserve? Born digital material from your institution; born digital material created outside of the institution; digital material from digitization.
- Do you already have a collection and preservation strategy for your digital data? (What you want to collect and how you want it to preserve)
- Do you have a collection management system for your analogue collection?

collaboration

- Is the data you want to preserve already part of another’s institutions collection strategy?

legal framework of your institution

- Does your legal framework allow the long-term preservation of digital objects?
- What is the scope of your long-term preservation, do you want simply to preserve your collection or additionally open it to the public? (I think we can recommend people with limited resources to focus on preservation project first; then when ingest-archiving process is done, implement the access part)

diversity of platform: management part and the access part

- If you open your collection to the public, how do want it to be presented - separately as a digital collection or together with your main collection of analogue materials?

management of the collection (system and human resources)

- How should the preservation of digital objects be integrated in your institution's workflows? Is a separate team doing it or do you want it to be integrated with your current workflows and teams?
- continuing education or new team?

8.1.2 Chapitre 3 : Réalisation d'un système d'archivage électronique

Core question: How to build a digital long-term archive from scratch?

Includes aspects such as the conception of a digital long-term archive, the development of cross-system workflows and project planning, an overview of vendors, architectures and standards, important principles, cost issues and maintenance...

- Which organizational questions should be answered, if you plan to design a long-term archive?
  - What is our goal, our mission statement? What do we want to archive, for what and for whom? Who (department, team) is responsible for this?
  - Priorities: Are priorities set, with which datatype you want to begin? Do you have a strategy begin with the most urgent, the biggest heap, the most complex use case?
  - Are there external factors that influence the roadmap, the available budget or the choice of systems?
  - Which legal and contractual regulations are relevant for our institution?
  - What form of access and usage are we aiming for?
  - Who are our internal and external stakeholders? What are their needs?
  - What is the internal perception of the topic? Do we need to promote awareness?
  - What could our exit strategy look like? Is there an institution that could take over services and data in an emergency?
- Which questions regarding resources and cooperation should be answered, if you plan to design a long-term archive?
  - What are the skills / is the job description of the person (to be) in charge of digital long-term preservation within a specific organisation?
  - Do we have the financial and human resources to accompany the development and operation of a long-term archive?
  - Who should work with the software? Librarians, archivists or IT staff? It could be necessary to do some tasks manually, e.g., build structure of a SIP.
  - Do we need or can we use an existing service instead of building our own long-term archive? How do we deal with such issues in principle (outsourcing versus self-determination)?
  - With which partners will we necessarily or usefully have to work with (cantonal or university IT, providers of storage infrastructures, providers of library and archive systems ...)?
  - Is your future digital long-term digital archive to be integrated in an existing IT architecture?
- Which technical-conceptual questions should be answered, if you plan to design a long-term archive?
  - Who are our data providers (persons, institutions, systems)? Which file formats and data sources have to be integrated (e.g., data carriers in archives; digital publications like e-books, journals, databases, websites; digitized cultural heritage)? If you have a DAM, do you think it is relevant for the LTA?

    - What else do you want to integrate in your AIP apart from the digital object and its administrative metadata? Bibliographic metadata? If so, do you want this data to be up-to-date? If it has to be up-to-date in which system, do you plan to keep it up-to-date?
    - What volume of data - initially and in estimated annual growth - will be ingested into the long-term archive?
    - Who are our data users (persons, institutions, systems)?
    - Which level of security is required for your stored data? Are the objects that are going to make up your collection legally sensitive? If so in which sense and what are the consequences? (Localisation of the storage)
- Which financial questions should be answered, if you plan to design a long-term archive?
    - What are the initial costs (points) for setting up and developing a long-term archive?
    - What are the annual costs (points) of running a long-term archive?
    - What hidden costs should be taken into account (e.g., internal project support, knowledge development, costs in cooperation with partner ITs ...)?
- What are the characteristics that a minimal (maybe temporary, until a complete solution is available) digital preservation system should have (see also: https://www.scimeeting.cn/m/video/play/2?vid=172068)?

8.1.3 Chapitre 4 : Recommandations pour les acquisitions

Core question: What are our guidelines, ideas and wishes for acquisition and preservation of digital objects?

- Do you have already set channel for producer to deposit digital material? Do these producers have their own wishes and ideas that are of relevance for preservation and accessibility?
- Which entities (persons, departments, systems) should deliver data in the future? Can these be connected automatically via interfaces or do delivery tools have to be included in the planning?
- Which wishes do we already have for accessibility during acquisition?

Subchapter 4.1: Born digital material and acquisition through APIs

- Are there existing standards for data delivery (APIs, data formats, file formats, SIP)?
- How can you ensure the integrity of the data?
- Which possibilities have data providers which we are cooperating with to deliver the file formats and metadata we are demanding?

Subchapter 4.2: Digital historical material or material from analogue carriers

- Do we have the necessary hardware for reading data carriers or do we need them? Or do we need partner organizations which can fulfil this task for us?
- How can you ensure the integrity of the data if it isn’t ingested immediately in the long-term archive?
- how to handle original carriers or media before and after ingest?

8.1.4 Chapitre 5 : Plan de préservation (Preservation-Planning)

Core question: How do we take care of the objects in a long-term archive?

Includes aspects such as setting up ingest workflows (I would say that ingest may be describe in chapter 2 - better not, because the ingest is a detailed process with many aspects. Chapter 2 is more focused on the pre-ingest and the fact, how data comes from a producer to our infrastructure), migrating objects of different natures (text, image, AV ...), tool recommendations, storage strategies, recommendations for formats suitable for archiving or where to find them ...

Subchapter 5.1: Ingest workflow creation and optimization

- Which work steps in an ingest workflow have to be carried out and to what extent can they be automated? Which steps are optional or recommendable? (may be placed in Chapter 2)
- To which extent can I or do I want to influence my workflow?
- Potential for optimization (workflows, hardware, parallelization, ...)
- Risks of optimization

Quality control and data curation

Subchapter 5.2: Strategy for storage

- How do you ensure bit preservation?
- What is the strategy for redundant copies? Which partners are involved?
- What is the backup strategy for the LTA (AIP-only or with surrounding (meta-)data -> file structure)?

Subchapter 5.3: File formats and recommendations

- What recommendations or sources of recommendations are there regarding file formats suitable for archiving?

Subchapter 5.4: Metadata

- What strategy do we pursue with regard to metadata: should they be archived as completely and up-to-date as possible in the AIPs or do they primarily serve to correctly identify an AIP and link it to the single source of truth in a library system, for example?
- Which type of identifier do you need for your long-term archive? Should it be linked to the persistent identifier in other systems or of your DIP?

8.1.5 Chapitre 6: Accessibilité et utilisation

- Which access levels do we need to support?
  - public on internet, public in the library, authenticated users, special authorization for researchers, inaccessible
- Which legal regulations are relevant for us?
  - Copyright law, protection of personality, archive law …
- What access platforms already exist that we can use for accessibility? For what does no platform exist yet?
- Is it necessary to create consultation copies (example: derive jpg from tiff)?
- Which type of persistent identifier is possible for you and which do you prefer? Do you need identifiers for a whole DIP (Dissemination information package, Zugangskopie) or also for parts of it (pages, parts of a file, annotations)? Is the PID of the DIP linked to the internal identifier of the AIP (Archival information package)?

- Which level of security is needed for the access of your data? Are the objects that are going to make up your collection legally sensitive? If so in which sense and what are the consequences (localization of the storage)?

8.2 Glossaire

AARU	Aaru Data Preservation Suite Outil permettant de préserver les contenus de différents supports de données https://www.aaru.app
AFS	Archives fédérales suisses https://www.bar.admin.ch
AIP	Archive information package Paquet d'information archivé dans un SAE conforme au modèle de référence OAIS
API	Application programming interface Une interface de programmation d'application est une façade par laquelle un logiciel offre des services à d'autres logiciels
AtoM	Access to Memory Système de gestion d'archives online diffusé notamment par le Conseil international des Archives (ICA) https://www.accesstomemory.org
BagIt	Spécification décrivant une manière de transmettre ou sauvegarder du contenu numérique. Est utilisée pour créer des SIP ou des AIP.
BCU/F	Bibliothèque cantonale et universitaire, Fribourg https://www.fr.ch
BN	Bibliothèque nationale suisse https://www.nb.admin.ch
CDDA	Compact Disc Digital Audio
DAM	Digital Asset Management Système permettant de stocker, organiser et partager les ressources numériques d'une organisation de manière centralisée.
DIP	Dissemination Information Package Paquet d'informations diffusé, dans un SAE conforme au modèle de référence OAIS

DROID	Digital Record and Object Identification Outil développé par les Archives nationales du Royaume-Uni permettant d'identifier automatiquement les formats de fichiers.
EAASI	Emulation as a service infrastructure
EAD	Encoded Archival Description Standard XML pour l'encodage d'inventaires d'archives, développé par l'Université de Californie à Berkeley et maintenu par la Société des archivistes américains en partenariat avec la Bibliothèque du Congrès.
EBUCore	Ensemble de métadonnées descriptives et techniques pour les médias audiovisuels, développé par l'European Broadcasting Union (EBU)
EXIF	Exchangeable Image File Format Spécifications pour l'enregistrement de métadonnées descriptives et techniques dans les formats d'encodage d'images comme JPEG ou TIFF.
FITS	Flexible Image Transport System Format de fichier spécialisé pour les images scientifiques
GED	Gestion électronique de document
HFS+	Hierarchical File System (extended) Système de fichiers conçu par Apple pour les systèmes d'exploitation Mac OS
Image disque	Une image disque est un fichier archive proposant la copie conforme d’un disque optique ou magnétique
IPTC	International Press Telecommunications Council Jeu de métadonnées descriptives et techniques pour les images
ISAD(G)	International Standard Archival Description (General)
JHOVE	JSTOR/Harvard Object Validation Environment Utilitaire d'identification d'objets numériques
KOST/CECO	Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen. Centre de coordination pour l’archivage à long terme de documents électroniques http://kost-ceco.ch

LIDO	Lightweight Information Describing Objects Format XML normalisé pour la description d'objets de musées, développé par le Conseil international des musées (ICOM)
LTO	Linear Tape-Open Technique de stockage de données sur bande magnétique selon un format ouvert
MARC21	Machine-readable cataloging Format normalisé de description bibliographique. Développé à partir des années 1960, il est mis en œuvre par les bibliothèques du monde entier.
MARCXML	Schéma permettant de représenter le format MARC21 en XML
METS	Metadata Encoding and Transmission Standard Standard visant à réunir dans un même fichier XML toutes les métadonnées nécessaires à la description d'un document
MKV	MatrosKa Video Format conteneur ouvert pour données audiovisuels (vidéos, images, audios, sous-titres)
MODS	Metadata Object Description Schema Standard XML de description bibliographique, souvent associé à METS
MOV	Format d'encodage vidéo développé par Apple
MP4	Moving Picture (Experts Group), 4e norme Format normalisé d'encodage vidéo
MPEG7	Moving Picture (Experts Group), 7e norme Norme descriptive pour la recherche de contenus multimédias.
MXF	Material eXchange Format Format conteneur pour données vidéo et audio
NAS	Network-attached storage Serveur de fichiers relié à un réseau destiné au stockage centralisé de données
NTFS	New Technology File System Système de fichiers développé par Microsoft pour les systèmes d'exploitation Windows

OAIS	Open archival information system Reference Model Modèle de référence pour un Système ouvert d'archivage d'information. Norme ISO 14721
OCPD	Ordonnance sur les certifications en matière de protection des données https://www.fedlex.admin.ch/eli/cc/2007/701/fr
ODT	OpenDocument Text Format ouvert de documents issus de traitements de texte
OFC	Office fédéral de la culture https://www.bak.admin.ch
PREMIS	PREservation Metadata: Implementation Strategies Standard pour les métadonnées nécessaire pour la préservation des documents numériques
RDA	Resource Description and Access Règles de catalogage appliqué par les bibliothèques dans le monde entier
RGPD	Règlement général sur la protection des données Texte de référence mis en œuvre par l'Union européenne pour la protection des données des personnes
RiC	Records in Contexts Norme de description de documents d'archives publiée par le Conseil international des Archives (ICA)
rsync	Remote synchronization Logiciel libre de synchronisation/copie de fichiers
SACD	Super Audio CD Support audio numérique sur disque optique
SEDA	Standard d'échange de données pour l'archivage (France)
SAE	Système d'archivage électronique Son objectif est la conservation et mise à disposition de façon durable de l'information numérique. L'information peut être acquise par numérisation d'une source physique ou déjà sous forme numérique (numérique natif)
SFTP	Secure File Transfer Protocol – SSH File Transfer Protocol Protocole de transfert sécurisé de fichiers entre systèmes par Internet

SIGB	Système intégré de gestion de bibliothèque
SIP	Submission Information Package Paquet d'informations à verser dans un SAE, conforme au modèle de référence OAIS
SLSP	Swiss Library Service Platform Réseau de bibliothèques scientifiques et universitaires suisses
SPECTRUM	Standard de description pour les objets muséaux
WAV	Waveform Audio File Format Format de fichier audio développé par IBM et Microsoft
XMP	eXtensible Metadata Platform Format de métadonnées en XML enregistré sur les fichiers images développé par Adobe
ZIP	Support de données sur disquettes amovibles mis sur le marché en 1994 par Iomega