Module 7C - Section 5 : Appréhender les traitements spécifiques aux différentes catégories de formats de fichiers

Édouard Vasseur @AIAF - PIAF

VF 02/12/2024

Table des matières

Objectifs

Description du module :

La préservation des documents d'archives sur support numérique – ce que les Québécois nomment documents technologiques – constitue désormais un enjeu quotidien des archivistes. L'archiviste dispose désormais d'un important panorama de normes, de standards, d'outils et de retours d'expérience pour lui permettre d'appréhender les documents d'archives sur support numérique et envisager leur préservation dans le temps.

Le but du module est de :

- aider à évaluer la situation en matière de préservation des documents d'archives sur support numérique ;
- permettre de concevoir et mettre en œuvre une politique de préservation des documents d'archives sur support numérique.

L'apprenant doit être en mesure de :

- appréhender les spécificités en matière de préservation des documents d'archives sur support numérique ;
- dresser un état des lieux d'ensembles de documents d'archives sur support numérique ;
- définir et mettre en œuvre une politique de préservation des documents d'archives sur support numérique ;
- concevoir, mettre en œuvre et administrer un environnement permettant de gérer dans le temps les documents d'archives sur support numérique, quels que soient les moyens disponibles ;
- appréhender les différentes catégories de formats de fichiers numériques ;
- savoir comment aller plus loin dans la réflexion.

Positionnement :

Ce module s'inscrit naturellement dans la chaîne archivistique. S'il se concentre sur les questions de planification de la préservation, de mise en œuvre de la préservation et de stockage des documents d'archives sur support numérique, il fournit également des éléments à prendre en compte lors de la mise en place de politiques et procédures de gouvernance de l'information et de gestion de l'archivage/gestion des documents d'activité/gestion des documents institutionnels/records management, de collecte de documents d'archives définitifs et d'accès à ceux-ci.

Il ne s'intéresse en revanche pas à la numérisation de documents d'archives sur support physique ou d'enregistrements sonores et audiovisuels sur support analogique, sauf dans le cas où l'opération de numérisation vise à substituer la version du document sur support numérique à celle sur support physique ou analogique.

Point sur le vocabulaire employé :

- Le terme “préservation” est entendu comme recouvrant « les fonctions de conservation préventive et matérielle » [Direction des Archives de France, Dictionnaire de terminologie archivistique, 2002] ;
- Sont distingués :
    - **les documents d’archives sur support physique**, où l’information est directement accessible à l’œil humain ou ne nécessite, pour le devenir, que l’emploi d’un appareil optique (projecteur) permettant de faciliter son agrandissement
    - **les documents d’archives sur support analogique**, où l’information, pour être intelligible, a absolument besoin de la médiation d’un appareil pour permettre à l’utilisateur de prendre connaissance de l’information (projecteur, lecteur, etc.) ;
    - **les documents d’archives sur support numérique**, qu’ils aient été directement produits avec des outils numériques ou soient le produit de la numérisation de documents d’archives sur support physique ou analogique. L’information, pour être intelligible, a absolument besoin de la médiation d’un environnement matériel et logiciel pour permettre à l’utilisateur de prendre connaissance de l’information ;
- “Document d’archives” est l’expression utilisée pour identifier toute information sur un support qui a besoin d’être prise en charge et conservée, soit pour sa valeur de preuve, soit pour sa valeur informationnelle, soit pour sa valeur patrimoniale ou de recherche. En fonction du contexte, l’expression pourra concerner des documents, des records ou des archives au sens anglo-saxon des termes ;
- “Service d'archives” est l’expression utilisée pour désigner toute structure ou organisme souhaitant mettre en place une politique de préservation de documents d’archives sur support numérique. Ce service d’archives peut être
    - interne à une organisation productrice et en charge de la gouvernance de l’information et de la gestion de l’archivage/gestion des documents d’activité/records management ou de la gestion d’archives intermédiaires ;
    - externe à une organisation productrice, soit qu’il s’agisse d’un prestataire de tiers archivage, soit d’un service d’archives définitif.

Les notions abordées dans ce module peuvent être complétées par :

- le module 9 - Section 2 : Numériser les documents qui présente les techniques de base de transfert de support vers le numérique
- le module 5 Gestion et traitement des archives courantes et intermédiaires

Il est vivement conseillé de prendre connaissance du module 7B Gestion des documents numériques au stade courant avant d’entamer la lecture du module présentement proposé. Certaines notions de base, activées ici, sont exposées plus longuement dans ce premier module.

Le glossaire du PIAF doit être consulté pour les définitions des termes spécifiques.

Introduction

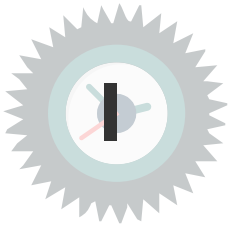
Dans l'univers physique et analogique, l'archiviste a pris l'habitude d'appréhender les différents types de supports et d'adapter sa stratégie de préservation – conservation préventive et restauration – aux contraintes de chacun d'entre eux. Parchemin, papier chiffon, papier chimique, papier calque, volume relié, fichier papier, plaque de verre, film en nitrate de cellulose, objet en trois dimensions ou bande magnétique ne se conservent pas tous de la même manière et nécessitent des conditions de préservation spécifiques.

De même, pour les documents d'archives sur support numérique, chaque catégorie de *format de fichiers* nécessite un traitement particulier, en raison de ses caractéristiques propres et de ses conditions particulières de production (logiciels utilisés, encodage de l'information dans les fichiers).

L'objectif du présent chapitre est de présenter les différentes catégories de formats de fichiers utilisés par les producteurs d'archives sur support numérique et d'examiner, pour chacune d'elles, l'état de l'art relatif à leurs conditions de préservation, ainsi que les différentes stratégies existantes pour assurer un accès à moyen et long terme aux documents d'archives que ces formats représentent.

Il s'ouvre par une discussion autour de la notion de format de fichiers pérenne.

1. Existe-t-il des formats de fichiers pérennes ?



Introduction

Les formats de fichiers sont conçus pour répondre aux besoins de production et de représentation d'archives portés par différentes communautés d'utilisateurs. Optimiser la production, le stockage et l'échange entre logiciels à un moment donné constitue les principaux objectifs des concepteurs de ces formats. La préservation à moyen et à long terme des documents d'archives représentés au moyen de ces fichiers ne constitue en revanche pas des objectifs essentiels.

Les services d'archives et les autres institutions patrimoniales se sont penchés depuis une vingtaine d'années sur la meilleure stratégie à adopter en matière de préservation des différents types de formats de fichiers.

Certaines institutions ont réfléchi à des critères permettant de décider si un format de fichiers est acceptable ou non pour la préservation à moyen ou long terme des documents d'archives sur support numérique.

1.1. Les critères à prendre en compte dans le choix des formats de fichiers

Les principaux critères pris en compte par ces institutions services d'archives et autres institutions patrimonailes) sont :

- **le caractère communautaire ou propriétaire du format de fichiers.** Les formats communautaires sont particulièrement prisés, mais leur adoption doit faire l'objet d'un examen approfondi. Comme les *formats propriétaires* sont dépendants de l'entreprise qui les a conçus et qui les maintient, les formats communautaires sont dépendants des communautés de développeurs qui les ont conçus et les maintiennent ; cependant, n'importe qui peut s'emparer de leur code, ce qui n'est pas nécessairement le cas pour les formats propriétaires. Par ailleurs, pour plusieurs catégories de formats de fichiers (ex. données géographiques et géospatiales, images tridimensionnelles), seuls des formats propriétaires sont disponibles ;
- **l'existence d'une documentation relative au format de fichiers.** L'existence de spécifications publiées et disponibles est un facteur souvent pris en compte dans la décision ou non de retenir un format de fichiers comme acceptable pour la préservation numérique. Les spécifications peuvent prendre ou non la forme d'un standard ou d'une norme internationale ;
- **la diffusion du format de fichiers.** La diffusion plus ou moins importante d'un format de fichiers peut constituer un critère décisif pour considérer un format de fichiers comme acceptable. En effet, un format de fichiers largement adopté, même s'il s'agit d'un format propriétaire, pourra être considéré comme acceptable dans la mesure où de nombreux acteurs sont intéressés par le maintien de l'accessibilité à son contenu. En revanche, un format de fichiers ouvert dont la diffusion est restreinte aux seules institutions patrimoniales aura tendance à reposer uniquement sur leurs ressources en termes de maintenance ;

- **les modalités de compression du format de fichiers.** Les formats de fichiers recourant à la compression avec perte et sans réversibilité possible ne sont souvent pas considérés comme acceptables pour la préservation à moyen et long terme. En revanche, ils sont souvent utilisés pour les versions de diffusion des documents d'archives sur support numérique ;
- **la prise en charge des métadonnées par le format de fichiers.** La capacité du format de fichiers à embarquer tout ou partie des métadonnées associées au document d'archives ou au fichier lui-même peut constituer un atout pour considérer qu'un format de fichiers est acceptable pour la préservation numérique.
- **la capacité du format de fichiers à préserver les propriétés porteuses de sens (*significant properties*) des documents d'archives.** Ce point est particulièrement important en cas de migration de format de fichiers, car, si le format cible ne préserve pas les propriétés porteuses de sens présentes dans le format de fichiers source, le document d'archives ne sera plus considéré comme étant digne de confiance.

1.2. Le classement des formats de préservation à long terme

Plusieurs services d'archives, sur la base de ces différents critères, ont adopté une politique en matière de formats de fichiers qui identifie :

- **les formats de fichiers considérés comme recommandés**. Il s'agit souvent des formats de fichiers documentés et pour lesquels l'institution dispose de connaissances et de compétences en interne et garantit donc à la fois la conservation binaire et l'accès à long terme à son contenu ;
- **les formats de fichiers considérés comme acceptables.** Il s'agit souvent des formats de fichiers largement répandus, souvent propriétaires, pour lesquels l'institution dispose de connaissances minimales, mais pas nécessairement de compétences internes, et dont l'institution garantit la conservation binaire et l'engagement de rechercher des solutions pour garantir l'accès à long terme à son contenu ;
- **les formats de fichiers considérés comme tolérés.** Il s'agit souvent des formats de fichiers moins largement répandus et pour lesquels l'institution ne dispose pas de connaissances et de compétences. Dans ce cas, seule la conservation binaire est assurée par l'institution, sans garantie que l'accès à long terme à son contenu soit possible. Le service d'archives assume ce risque au regard de l'intérêt de ces fichiers et de l'absence de versions dans des formats recommandés ou acceptables.

1.3. Choix des formats : une politique restrictive ou une politique souple ?

La définition d'une politique restrictive en matière de formats de fichiers acceptés :

- nécessite l'emploi de ressources pour contrôler et valider les formats de fichiers à l'entrée et pour procéder à l'uniformisation des formats de fichiers, par le biais d'opérations de migration ;
- implique de vérifier si l'uniformisation des formats de fichiers ne risque pas de rendre les documents d'archives non dignes de confiance ;
- est plus facile à mettre en œuvre dans le cadre d'une numérisation de documents physiques ou analogiques ;

La définition d'une politique souple en matière de formats de fichiers acceptés :

- prend davantage en compte les contraintes des services producteurs en matière de création des documents d'archives ;
- peut se révéler complexe en raison de la diversité des formats de fichiers proposés en entrée par les services producteurs.

Parmi les institutions ayant adopté une politique en matière de formats de fichiers, on peut citer...

Exemple

- les Archives fédérales suisses : https://www.bar.admin.ch/bar/fr/home/archivage/versement-de-documents/documents-numeriques.html ;
- Bibliothèque et Archives Canada : https://www.bac-lac.gc.ca/fra/services/gestion-ressources-documentaires-gouvernement/lignes-directrices/Documents/formats-fichier-rdvc.pdf ;
- la Bibliothèque nationale de France (BnF) : https://hal-bnf.archives-ouvertes.fr/hal-03374030/document ;

2. Les fichiers bureautiques

1. Introduction

Les formats de fichiers bureautiques sont sans doute ceux que l'archiviste a le plus l'habitude de manipuler dans sa vie quotidienne, et ceux avec lesquels il est le plus familier. Sont regroupés sous cette catégorie des fichiers correspondant à des traitements de texte, à des tableurs, à des présentations sous forme de diapositives et aux impressions numériques de ces fichiers qui recourent au format PDF.

Chaque catégorie est successivement étudiée.

2.1. Les fichiers correspondant à des traitements de texte

Introduction

Les fichiers correspondant à des traitements de texte se décomposent en deux catégories :

- les fichiers plein texte
- les fichiers de traitement de texte à proprement parler

2.1.1. les fichiers plein texte

Les fichiers plein texte, réalisés à partir d'éditeurs de texte comme BlocNotes ou Notepad, qui contiennent du texte sans aucun formatage ou objet intégré. Ce type de fichier est abondamment utilisé dans le monde informatique pour l'écriture du code des logiciels ou pour l'enregistrement des traces générées par les systèmes informatiques (logs).

Cette catégorie de fichiers ne pose aucun problème de préservation, car il ne contient que du texte dans un encodage standard, n'incorpore pas d'autre type de contenu et n'est jamais enregistré sous la forme d'un conteneur.

2.1.2. Les fichiers de traitement de texte à proprement parler

Les fichiers de traitement de texte à proprement parler, produits à partir de logiciels de traitement de texte (ex. Microsoft Word, OpenOffice Writer, etc.). Ces fichiers ne contiennent pas seulement du texte, mais peuvent aussi incorporer d'autres types de contenus (images, son, vidéo, carte, etc.). Les formats de fichiers les plus récents de cette catégorie sont en fait des formats conteneurs de type ZIP intégrant de multiples fichiers, notamment des fichiers XML compressés (il est possible de le constater en changeant l'extension des fichiers correspondants).

a) La préservation à long terme de ces fichiers présente plusieurs difficultés

- il peut ne plus exister de logiciel pour lire les fichiers concernés. Les versions récentes **des logiciels de traitement de texte ne permettent pas systématiquement de relire des versions obsolètes des formats de fichiers concernés.** Bien souvent, les résultats sont souvent meilleurs avec les logiciels ouverts qu'avec les logiciels propriétaires, les premiers prenant en charge les versions obsolètes des formats de fichiers. Par ailleurs, lire avec un logiciel de traitement de texte exploité sous Windows un fichier créé avec un logiciel de traitement de texte exploité sous Mac n'est pas évident ;

- **les métadonnées embarquées dans les fichiers eux-mêmes** (accessibles depuis les propriétés) **ne sont pas nécessairement fiables.** Les dates embarquées (notamment la date de dernière modification) peuvent par exemple correspondre à des événements techniques (enregistrement sur un espace de stockage) et non à des événements « métier » ;
- les fichiers peuvent contenir des **données dynamiques** (par exemple des dates) qui se mettent à jour automatiquement à chaque lecture du fichier ;
- des **erreurs d'enregistrement** ont pu intervenir, notamment en cas de téléchargement des fichiers depuis une source externe ;
- tous les logiciels ne prennent pas en compte de la même manière **l'encodage des caractères** ;
- les **liens** vers des ressources extérieures au fichier (adresse d'un site internet par exemple) peuvent être **cassés** suite à la disparition de la ressource ou à son changement de localisation ;
- des **dispositifs de sécurité** (saisie de mots de passe, chiffrement) peuvent empêcher l'accès au contenu des fichiers.

b) Stratégies et solutions à considérer/adopter

Plusieurs stratégies sont possibles pour conserver ces formats de fichiers :

- **maintenir le fichier dans son format d'origine et identifier un logiciel permettant de lire les versions actuelles et les versions antérieures du format** :
  - cette solution est la meilleure pour les fichiers plein texte. Pour ceux de traitement de texte, elle est acceptable si les fichiers sont aux formats ODT ou DOCX ;
  - cette solution est financièrement avantageuse, mais la lecture dans un logiciel d'une version plus récente que celui ayant servi à créer le fichier peut générer des changements d'apparence, de mise en page ou de pagination ;
- **convertir le fichier dans un format d'une version plus récente** ou l'imprimer au format PDF. Cette stratégie, valable pour la seconde catégorie, a le mérite de faciliter la lecture par les logiciels de lecture les plus récents, mais, comme toute opération de migration de formats, elle est consommatrice de ressources et peut porter atteinte aux propriétés essentielles du fichier d'origine ;
- **recourir à l'émulation**, notamment pour les formats de fichiers les plus anciens. Cette solution, comme cela a été expliqué dans la section 3 de ce module, nécessite cependant de mettre en œuvre une plateforme technique complexe et d'avoir le droit d'utiliser les logiciels anciens.

Conseil

Comme pour tous les autres formats de fichiers, la stratégie retenue doit dépendre de la finalité de la préservation et des propriétés et fonctionnalités des documents représentés par les fichiers qu'il convient de préserver.

2.1.3. Points d'attention

Il faut noter que :

- **le principal problème de la préservation des fichiers de traitement de texte est archivistique :** identifier les fichiers à archiver et récupérer les éléments de contexte permettant de les rendre intelligibles. Il s'agit d'une question d'évaluation, de sélection et de tri qui doit être étudiée lors de la définition de la politique d'archivage ;
- **toute opération de migration d'un fichier** dans un format d'une version plus récente ou d'impression au format PDF **a un coût**, nécessite des ressources et la mise en place d'un protocole de test et de validation des résultats obtenus. Par ailleurs, le processus de migration

peut générer des erreurs multiples qui pourront être identifiées lors d'une opération de validation du format obtenu, mais qui ne sont pas nécessairement faciles à interpréter et à traiter ;

- toute opération de migration implique de **s'interroger sur la pertinence de conserver le fichier d'origine**, que ce soit pour permettre des traitements futurs (ex. émulation), ou parce que c'est lui qui porte la valeur juridique ;
- **en cas de recours à une impression au format PDF**, il est recommandé d'utiliser le module d'impression intégré au logiciel de traitement de texte ayant permis de créer le fichier. Cette opération permet d'embarquer plus facilement les contenus, notamment les caractères utilisant des polices de caractères spécifiques, dans le fichier PDF ;
- les opérations d'identification et de validation de formats permettent de repérer d'éventuels problèmes dans les fichiers (malformations, blocage par un mot de passe, chiffrement) ;
- **ces fichiers peuvent contenir des formules de calcul mathématiques et des macros**, c'est-à-dire des morceaux de code informatique qui exécutent des tâches répétitives, et qui ne sont pas nécessairement pris en charge par des versions plus récentes des logiciels avec lesquels ces fichiers ont été créés, ou par des logiciels de même nature, mais reposant sur un code informatique différent. Par ailleurs, les logiciels d'antivirus peuvent considérer ces codes embarqués malveillants et déclarer infectés les fichiers concernés.

2.2. Les tableurs

2.2.1. La préservation à long terme des fichiers de tableurs

Les fichiers correspondant à des tableurs permettent de stocker des données (texte, nombre, date, pourcentage, formule de calcul, etc.) sous forme de tableaux, divisés en colonnes et en lignes. Les formats de fichiers les plus récents de cette catégorie sont en fait des formats conteneurs de type .zip intégrant de multiples fichiers (il est possible de le constater en changeant l'extension des fichiers correspondants).

a) La préservation à long terme des tableurs présente plusieurs difficultés

On rencontre avec ces fichiers les mêmes problématiques que celles posées par les fichiers de traitement de texte (obsolescence des logiciels de lecture, fiabilité des métadonnées embarquées, présence de données dynamiques, lien vers des ressources externes, présence de dispositifs de sécurité).

À cela s'ajoutent cependant des problèmes spécifiques :

- le formatage des cellules (choix d'un type comme texte, date, nombre) est fragile, et peut facilement être cassé, ce qui peut générer des difficultés d'interprétation des données ;
- plusieurs tableaux peuvent être intégrés dans un même fichier et être liés entre eux ou à des tableaux contenus dans d'autres fichiers de manière dynamique, la modification d'une valeur dans un tableau entraînant la modification des valeurs liées dans les autres tableaux (on parle alors de tableau croisé dynamique). Le simple déplacement d'un des fichiers peut occasionner la perte du lien avec les autres fichiers ;
- ces fichiers peuvent contenir des équations mathématiques et des visualisations statistiques sous forme de diagrammes ou de graphiques (entre autres) qui sont difficiles à récupérer dans certains contextes (notamment quand les fichiers ont été créés à partir de logiciels de tableurs disponibles en ligne) ;
- ces fichiers peuvent naturellement contenir des formules de calcul mathématiques et des macros.

b) Les stratégies et solutions à adopter

Les stratégies de préservation sont globalement identiques à celles mises en œuvre pour les fichiers de traitement de texte, cependant avec les particularités suivantes :

- toute migration vers un format de fichiers plus simple – type CSV (Comma Separated Value) – et toute impression au format PDF risque de faire perdre certains éléments des fichiers : formules mathématiques et logiques, graphiques, diagrammes, macros, liens avec d'autres tableaux ;
- l'impression au format PDF n'est pas adaptée dès lors qu'il s'agit de préserver des tableaux aux nombreuses colonnes et lignes et dont les données devront pouvoir être récupérées et ré-exploitées pour d'autres usages. Il en va de même pour les fichiers contenant plusieurs tableaux, notamment quand ils sont liés entre eux.

2.2.2. Points d'attention

Points d'attention, outre ceux déjà identifiés pour les fichiers de traitement de texte :

- la structure des tableaux (intitulé des colonnes et des lignes) doit être claire et documentée pour permettre l'interprétation des données, tout comme les formules existantes. Il peut s'avérer utile de rédiger une documentation extérieure au fichier ;
- il peut être utile de conserver les macros dans un fichier à part.

2.3. Les impressions PDF

2.3.1. Les caractéristiques du format PDF

Le format PDF (Portable Document Format) a été inventé en 1991 par John Warnock, cofondateur de la société Adobe, pour permettre de créer des documents, d'en échanger des versions par voie électronique indépendamment de l'environnement matériel et logiciel et de visualiser leur contenu sur n'importe quel environnement matériel et logiciel.

Désormais normalisé à l'ISO (norme ISO 32300, 2008), le format PDF préserve la mise en page (police de caractères, images, objets graphiques) définie par l'auteur d'un document d'archives.

a) Le format PDF est disponible dans plusieurs versions, aux fonctionnalités différenciées :

- **les versions courantes** offrent des options personnalisées comme la compression des images et des textes ;
- **des versions interactives** permettent de définir des zones de texte modifiables, d'ajouter des notes et des corrections, offrent des menus déroulants, permettent de rajouter des liens vers des sites webs, etc. ;
- **une version PDF/A** a été conçue pour imprimer des documents destinés à être conservés sur le long terme. Cette version interdit certaines fonctionnalités inadaptées à cet usage, comme le fait de lier le document à une police de caractères externe – dans un fichier au format PDF/A, la police de caractères est embarquée dans le fichier –, le chiffrement, l'intégration d'annotations, la présence de fichiers Javascript, etc. Plusieurs sous-versions existent : PDF/A-1 (2005), PDF/A-2 (2011), PDF/A-3 (2012), PDF/A-4 (2020)

Attention

Les fichiers PDF se présentent sous la forme d'un conteneur qui peut comprendre du texte, des images fixes, des fichiers multimédias, des signatures électroniques, des pièces jointes qui ne sont pas directement visibles mais que les logiciels de visualisation restituent – certains permettant même de les voir dans un volet particulier.

b) Les fichiers PDF peuvent être créés par :

- **des logiciels intégrés à des scanners ou des imprimantes multicopieurs**, à l'occasion de la numérisation de documents d'archives sur support physique ;
- **des suites bureautiques**, des logiciels de messagerie ou des logiciels de conception assistée par ordinateur (CAO), pour imprimer des documents en cours de conception ;
- **des logiciels spécifiques de conversion**, disponibles soit sur l'ordinateur de l'utilisateur, soit en ligne, en mode service (comme IlovePDF).

2.3.2. La préservation à long terme des fichiers PDF présente plusieurs difficultés

Très répandus, les fichiers PDF présentent pourtant de nombreuses difficultés de préservation :

- la norme PDF étant relativement complexe, elle est souvent interprétée différemment par les éditeurs de logiciels. Les logiciels de visualisation des fichiers PDF sont ainsi particulièrement permissifs aux non-conformités par rapport à la norme ;
- la création de fichiers PDF par numérisation de documents d'archives sur support physique n'est pas nécessairement associée à un processus de contrôle de la qualité visuelle comme technique des fichiers produits. C'est tout particulièrement le cas des numérisations « à la volée », à partir de scanners et d'imprimantes/copieurs multifonctions ;
- la création et la transmission de fichiers PDF peuvent conduire à la compression avec perte des informations, voire au chiffrement de celles-ci ;
- selon la version du format PDF choisie, certains éléments du document et de sa mise en forme (police de caractère, images, ressources multimédias) ne seront pas intégrés dans le fichier au moment de sa création ;
- a contrario, du fait que le format PDF – hors version PDF/A – permet d'intégrer tout type de contenus, les fichiers PDF peuvent inclure des contenus susceptibles de mettre en danger un système d'archivage : fonctions JavaScript, programme malveillant, etc. ;
- les fichiers PDF sont en réalité facilement modifiables ;

Comme tout format conteneur, un fichier PDF est à la fois un tout et une somme de parties ayant chacune ses propres contraintes en matière de préservation.

Le format PDF dans sa version PDF/A a longtemps été recommandé pour la préservation de tous les documents d'archives sur support numérique.

Or cette solution n'est pas viable, pour plusieurs raisons :

- la migration au format PDF prend du temps et est consommatrice de ressources (humaines, financières, techniques). Elle n'est donc pas toujours rentable ;
- la migration au format PDF n'est pas adaptée à certains types de formats de fichiers : présentations intégrant des animations (l'impression au format PDF/A supprime celles-ci), tableurs (l'impression ne prend pas en compte la multiplicité des onglets). Ce problème est exacerbé lorsque l'utilisateur souhaite réutiliser les données imprimées (ex. coordonnées mathématiques utilisés dans les formats de fichiers de dessin, données enregistrées dans une base de données) ;
- comme les autres versions du format PDF, la version PDF/A est normalisée d'une manière sujette à interprétations. Il est donc quasiment impossible de disposer de fichiers PDF/A strictement conformes à la norme elle-même.

2.3.3. Points d'attention

Le principal problème de la préservation des fichiers au format PDF est archivistique : identifier les fichiers à archiver et récupérer les éléments de contexte permettant de les rendre intelligibles. Il s'agit d'une question d'évaluation, de sélection et de tri qui doit être étudiée lors de la définition de la politique d'archivage ;

Les opérations d'identification et de validation de formats permettent d'identifier d'éventuels problèmes dans les fichiers (malformations, blocage par un mot de passe, chiffrement).

Des outils de validation de formats existent (ex. veraPDF), mais les résultats fournis par ceux-ci ne sont pas toujours cohérents et dépendent de la capacité des personnes qui les ont spécifiés à interpréter les normes décrivant le format des fichiers correspondants.

En cas d'intégration de ressources externes dans le PDF (tableurs, contenu multimédia), d'interroger la possibilité de conserver à part les objets concernés.

3. Les messages électroniques

3.1. Les caractéristiques du message électronique

Les premiers messages électroniques ont été envoyés dans les années 1970, dans les universités américaines.

C'est cependant dans les années 1990 que leur usage se développe, notamment grâce à l'ouverture du réseau internet et du World Wide Web, mettant cet outil à disposition d'une population sans cesse croissante.

En 2023, ce sont des centaines de milliards de messages qui sont échangés chaque année.

Techniquement, un message électronique tel que spécifié par les standards Internet Message Format (IMF) (Requests for comments – RFC 5322, 2008) et Multipurpose Internet Mail Extensions (MIME), est un document textuel structuré comprenant :

- **un en-tête** comprenant des informations visibles par l'utilisateur au moyen d'interfaces graphiques mises à disposition par des éditeurs de logiciels – nom de l'expéditeur, noms des destinataires, date d'expédition, objet du message, identifiant du message auquel le message répond (permettant de créer des fils de discussion) –, ainsi que des informations de routage, d'authentification et d'horodatage qui ne sont visibles qu'en analysant le message au moyen d'un éditeur de texte. Le destinataire du message peut ajouter des métadonnées supplémentaires à réception – statut, importance, nécessité d'un suivi, mots-clés sous forme de tags ;
- **le corps du message** lui-même, comprenant du texte non structuré mais pouvant également intégrer d'autres types de fichiers (image, son, vidéo) ;
- **des pièces jointes**, parfois limitées en termes de nombre, de taille et de format de fichiers.

Un utilisateur peut organiser ses messages dans le logiciel qui est mis à sa disposition pour les consulter, en créant une arborescence de répertoires plus ou moins complexe.

Fondamental

Le format de fichiers utilisé pour encoder ces messages est l'EMail File (EML). Ce fichier peut embarquer les pièces jointes ou se contenter de les référencer. Si le format EML est fondé sur les normes Requests for comments (RFC), il n'est pas officiellement spécifié.

3.2. La préservation à long terme des messages électroniques présente plusieurs difficultés

Malgré un caractère relativement rudimentaire, les messages électroniques présentent plusieurs difficultés de conservation.

3.2.1. Les principales ne sont pas d'ordre technique mais d'ordre archivistique...

- Les messages électroniques véhiculent indifféremment contenus professionnels et contenus personnels, y compris de nombreuses données à caractère personnel, dans le corps du message comme dans les pièces jointes. Les messages électroniques sont donc couverts par le secret des correspondances et ne peuvent être pris en charge et gérés sans encadrement juridique ;
- L'identification des messages à prendre en charge, à quel moment, pour quelle finalité, est capitale ;
- L'organisation retenue pour les messageries diffère grandement en fonction des utilisateurs .

3.2.2. ... mais il existe également des contraintes techniques

- les messages sont tout d'abord stockés par les logiciels de messagerie et exportés de leur environnement d'origine. Si des exports unitaires au format EML sont possibles, la plupart des exports possibles se font sous forme de lots. Deux grands types de conteneurs existent :
    - les conteneurs au format PST générés par le logiciel Outlook de la société Microsoft, qui embarquent messages, mais aussi agendas, tâches et carnets d'adresses, pour tout ou partie d'une messagerie. La structuration de ces conteneurs est définie par Microsoft et évolue dans le temps, rendant toute rétrocompatibilité difficile – les versions actuelles d'Outlook prennent difficilement en charge des exports effectués depuis plus de 5 ans ;
    - les conteneurs au format MBOX, chaque fichier MBOX contenant un dossier de messages avec des pièces jointes intégrées sous forme de texte codé. Là encore, plusieurs variantes de ce format existent qui ne sont pas entièrement compatibles entre elles ;
- les messages peuvent comprendre des liens vers des ressources qui leur sont extérieures (adresse d'un site internet par exemple) et qui peuvent être cassés suite à la disparition de la ressource ou à son changement de localisation ;
- les pièces jointes correspondent à des fichiers enregistrés dans tous les types de formats possibles. Leur préservation à long terme peut impliquer de mettre en œuvre des stratégies spécifiques ;
- la destruction d'un message au sein d'un fil de conversation peut provoquer une perte d'intelligibilité de la chaîne ;
- la taille des messageries peut constituer un frein lors des opérations d'export et d'import.

3.3. Stratégies et solutions de préservation à adopter

Depuis une dizaine d'années, de nombreux outils logiciels ont été développés pour assurer le traitement des messages électroniques, que ce soit par des institutions de préservation ou des entreprises commerciales.

Plusieurs stratégies sont par ailleurs possibles pour conserver ces formats de fichiers :

- **extraire les messages au format EML à partir des conteneurs exportés depuis les logiciels de gestion eux-mêmes**. De nombreux outils, payants ou ouverts (logiciels ou librairies) sont désormais disponibles pour réaliser cette opération ;
- **conserver les messages exportés sous forme de conteneurs MBOX**, considéré comme un format de fichiers privilégié ou accepté par certaines institutions en raison de son caractère très diffusé ;
- **convertir au format XML** la structure de la messagerie et des messages eux-mêmes (ex. utilisation du logiciel DarcMail).

3.4. Points d'attention

- toute opération de préservation retenue doit **prendre en compte les caractéristiques à préserver et les fonctionnalités attendues pour l'accès et la réutilisation des contenus** eux-mêmes. La réalisation de tests est vivement recommandée ;
- **il est préférable de collecter des messageries entières**, afin de préserver les fils de discussion et l'organisation retenue par la ou les personnes titulaires de la messagerie ;
- dans la mesure où les infrastructures de messagerie imposent des quotas de stockage sur les serveurs mutualisés, les utilisateurs peuvent être encouragés à télécharger les messages sur un ou plusieurs postes de travail (ordinateur professionnel ou personnel). **Il est donc essentiel de s'assurer de la complétude de l'export réalisé** ;
- **l'impression au format PDF ou PDF/A des messages fait perdre un nombre important de métadonnées.** Des travaux de recherche sont cependant en cours à ce sujet ;
- **la conversion des messages au format MSG n'est généralement pas recommandée** par les institutions de préservation, même si le format de fichiers est largement répandu et ses spécifications maintenues par la société Microsoft ;
- **il est essentiel de documenter les pratiques d'utilisation des messageries,** dans la mesure où cela est possible ;
- en cas d'extraction des messages EML de leur conteneur d'origine, ou de conversion du conteneur .pst au format MBOX, il peut s'avérer utile de **s'interroger sur la pertinence de conserver le fichier d'origine**, que ce soit pour permettre des traitements futurs (ex. émulation), ou parce que c'est lui qui porte la valeur juridique ;
- **une réflexion sur l'accès aux messages archivés est capitale** et doit être entreprise rapidement, afin de rendre possible les recherches et l'accès aux différentes catégories d'utilisateurs intéressés (services producteurs, juges, journalistes, chercheurs, grand public). Des travaux de recherche sont actuellement en cours pour tester le recours aux technologies d'intelligence artificielle (traitement automatisé du langage naturel, extraction d'entités nommées) afin d'améliorer cet accès. Leur utilisation nécessite cependant la mise à disposition de connaissances et compétences spécifiques.

4. Les bases de données

4.1. Les caractéristiques d'une base de donnée

Définition

Une *base de données* est un outil logiciel ou un ensemble d'outils logiciels permettant de collecter et de structurer des données de manière à permettre d'y accéder au moyen de requêtes plus ou moins complexes.

La structure des bases de données peut aller de simples tables ressemblant à des tableurs à des ensembles plus complexes de schémas, requêtes, vues, tables et autres éléments fonctionnant ensemble et permettant l'ajout, la suppression, la modification, le stockage et l'interprétation de données par des utilisateurs.

Plusieurs types de bases de données existent :

- les bases de données relationnelles, interrogeables par exemple au moyen du langage de requête structurée (*Structured Query Language*, SQL). Plusieurs systèmes de gestion de bases de données relationnelles existent, comme Access, Maria DB, MySQL, Oracle, PostgreSQL, etc. ;
- les bases de données non tabulaires ;
- les bases de données orientées texte.

Les bases de données sont au cœur de nombreux systèmes informatiques plus importants permettant de créer des documents d'archives, de les traiter, de les classer, de les rechercher et de les consulter.

4.2. La préservation à long terme des bases de données présente plusieurs difficultés

Fondamental

Avant toute chose, il convient de signaler que la préservation de la base de données n'est pas nécessairement une fin en soi.

Tout va dépendre :

- du contenu que l'on souhaite archiver,
- de la finalité de cet archivage,
- des besoins d'accès ultérieurs aux contenus archivés.

Avant d'entamer toute opération d'archivage d'une base de données, il convient de s'interroger sur ces points absolument essentiels.

4.2.1. Les difficultés de préservation peuvent être d'ordre archivistique...

- les contenus manipulés dans des bases de données et à archiver peuvent prendre la forme d'agrégats plus ou moins complexes, avec des données réparties dans plusieurs tables d'une même base de données voire entre plusieurs bases de données ;
- les données saisies dans les bases de données peuvent faire l'objet de mise à jour avec suppression des données antérieures ;
- le degré de qualité des données saisies peut être très inégal, notamment quand leur création est décentralisée et mise entre les mains de nombreux utilisateurs ;
- la connexion des utilisateurs, la mise en œuvre des processus de travail (création, modification, lecture et suppression des données) font systématiquement l'objet d'enregistrements sous forme de journaux ou de logs. Leur intégration au contenu à archiver doit être interrogée, notamment s'il est essentiel de garantir l'authenticité des données ;
- préserver les méthodes de travail et l'apparence des interfaces utilisées par les usagers de la base peut mériter d'être examiné

4.2.2. ... ou d'ordre technique

- **l'architecture** des bases de données peut être **très complexe** (nombre de tables par exemple) ;
- **l'architecture des bases de données et la manière dont les données sont codées à l'intérieur de la base ne sont pas toujours bien documentées**. Même si une documentation existe, elle peut avoir été établie au moment de la création de la base et ne pas refléter les évolutions fonctionnelles et techniques de celle-ci. Or une bonne documentation est capitale pour garantir un accès sur le long terme aux contenus eux-mêmes ;
- **les bases de données peuvent comprendre des liens vers des ressources qui leur sont extérieures** (adresse d'un site web par exemple) et qui peuvent être cassés suite à la disparition de la ressource ou à son changement de localisation ;
- d'autres documents d'archives sous format numérique se présentant sous la forme de **fichiers indépendants, stockés à part de la base de données**, mais intimement liés aux documents présents dans la base (ex. pièces justificatives). Leur export combiné à celui des données enregistrées en base peut se révéler problématique.

4.3. Stratégies et solutions à adopter pour préserver une base de données

Plusieurs stratégies sont possibles pour préserver les bases de données, en fonction du type de contenu à archiver .

4.3.1. Si l'intégralité de la base doit être archivée,

Plusieurs possibilités techniques sont à envisager :

- **exporter l'intégralité de la base sous forme d'un dump, un export des données d'une table et/ou les requêtes SQL correspondantes dans un fichier texte**, ce qui permettra éventuellement son import dans une autre base. Cependant cette stratégie n'est possible que pour des données à conserver sur des durées très courtes (moins de 2 ans) ;
- **exporter les données contenues dans la base table par table sous forme de tableurs à plat, au format CSV**. Cette méthode préserve les données mais ne permet en aucun cas de préserver les formulaires, les interfaces, les requêtes ;
- lorsque les données font l'objet d'une mise à jour sans historisation, il peut s'avérer utile de **réaliser des exports réguliers de l'ensemble de la base**, sous forme d'états annuels ;

- **exporter toutes les tables de la base dans un format de fichiers pivot**. Dans cette logique a été conçu et développé le format *Software Independent Archiving of Relational Databases* (SIARD) par les Archives fédérales suisses qui permet d'encoder les données exportées de bases Microsoft Access, DB, MySQL, MariaDB, Oracle, PostgreSQL et SQL Server et de régénérer des exports aux formats SQL Server, MySQL, Oracle et PostgreSQL. Ce format ouvert, basé sur le format XML et les normes et standards SQL2008, Unicode et Zip64, a fait l'objet d'extensions dans des versions ultérieures (SIARD 1.0 et 2.0). Des outils – SIARD Suite et Database Preservation Toolkit (DBPTK – https://database-preservation.com/) – ont été développés pour générer les fichiers SIARD et consulter les données conservées dans ceux-ci, ainsi que la documentation de ces bases. C'est la suite SIARD qui a été privilégiée en France pour l'archivage de la matrice cadastrale numérique ;
- **recourir à l'émulation**.

4.3.2. Si l'intégralité de la base ne doit pas être archivée

Dans ce cas, notamment quand il s'agit d'exporter des documents d'archives sous forme d'agrégats, avec ou sans autres fichiers joints, les exports constituent une combinaison de ces fichiers joints et d'exports des documents voire des registres les décrivant sous une forme à étudier. Le choix de la méthode d'export dépend alors de la finalité de l'opération d'archivage et des besoins d'accès aux contenus archivés (accès en totalité ? Accès par extrait ?).

4.4. Les méthodes d'exports sous forme de tableurs à plat ou sous forme de fichier SIARD

Les méthodes d'exports sous forme de tableurs à plat ou sous forme de fichier SIARD présentent chacune des avantages et des inconvénients.

4.4.1. Avantages de la conservation sous forme de tableurs à plat

- toutes les communautés d'utilisateurs sont familières de ces formats de fichiers ;
- le format de fichiers n'est pas dépendant d'outils spécialisés ;
- les fichiers ont une taille maîtrisable ;
- le format de fichiers peut être utilisé pour collecter des données exportées de bases de données en production comme de bases de données décommissionnées, sans qu'il y ait besoin d'une connexion technique à la base en production ;
- une sélection des tables, des données et des fichiers joints est possible ;
- le format de fichiers permet de fournir des options d'accès multiples (en totalité, par extrait).

4.4.2. Avantages de la conservation sous forme de fichiers SIARD

- le format de fichiers est ouvert ;
- des outils ouverts (open source) sont disponibles et leur code source est disponible pour concevoir et mettre en production de nouveaux outils ;
- données et métadonnées sont stockées dans des fichiers séparés, mais rassemblés dans un unique conteneur ;
- l'extraction des données et des métadonnées depuis la base de données source est automatique ;
- la base de données archivée est immédiatement accessible et interrogeable, sans avoir besoin de reconstruire la base.

Attention

Donner accès à des contenus exportés d'une base de données, sous forme de fichiers à plat ou de fichiers SIARD, n'est pas évident pour les utilisateurs. Il peut s'avérer intéressant de disposer de la base archivée sous plusieurs formes, en fonction des usages attendus (version de conservation, version de diffusion facilement exploitable).

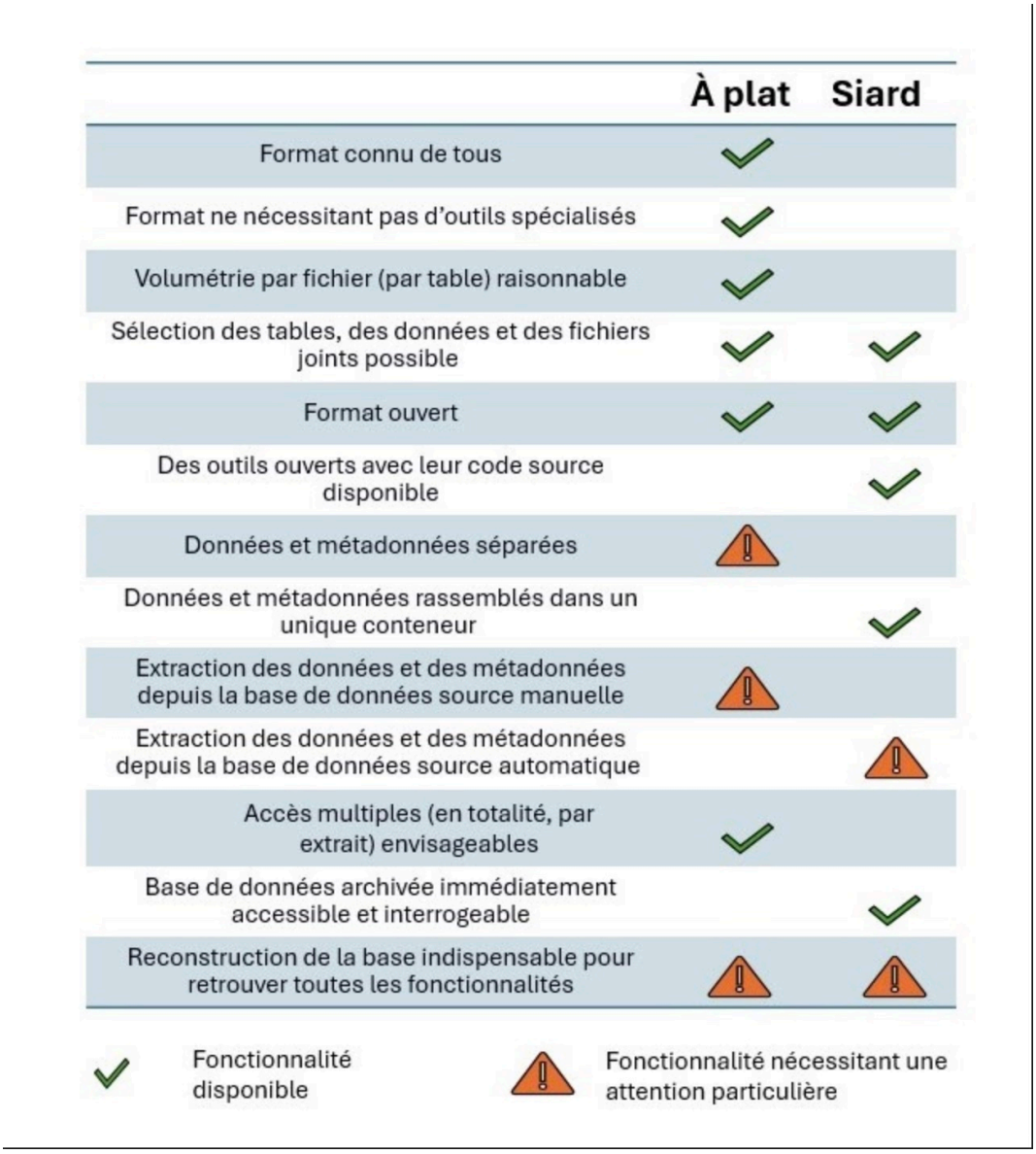


Fig.1 : Comparaison des avantages et inconvénients des deux solutions d'archivage de base de données (crédits : B.Grailles/PIAF)

4.5. Point d'attention

Fondamental

Quelle que soit la méthode d'archivage retenue, il faut s'assurer de la complétude des exports, que ce soit en nombre de tables, en nombre de données ou en nombre de fichiers joints. Le contrôle qualité des exports est essentiel. Ce contrôle doit s'étendre à la structure de l'export et à la conformité de celui-ci avec la documentation recueillie ou constituée.

L'existence d'une documentation précise des processus de travail mis en œuvre, de la structure de la base, de la structure de chaque table et de la manière dont les informations sont codées est essentielle. La récupération de la documentation existante, externe (spécifications, manuel utilisateur, supports de formation, document d'architecture, autorisations juridiques) ou interne à la base, est essentielle pour garantir un accès sur le long terme aux contenus archivés.

Lorsqu'il s'agit d'exporter des documents d'archives sous forme d'agrégats, avec des fichiers joints – donc des dossiers – il est essentiel d'envisager l'export, à côté de ceux-ci, du registre présent dans la base de données les décrivant. Quand l'architecture de la base de données est particulièrement complexe, les données correspondant au traitement subi par chaque affaire peuvent faire l'objet d'un export sous la forme d'un fichier généré spécifiquement au moment de l'export – par exemple aux formats XML ou PDF – et joint aux fichiers correspondant au dossier.

5. Les fichiers structurés/balisés (XML, JSON)

L'**Extensible Markup Language (XML)** et le **JavaScript Object Notation (JSON)** sont deux formats de données textuels qui permettent de représenter et de transmettre des données de manière structurée.

Tous deux ont pour finalité première de **faciliter les échanges automatisés** de contenus complexes entre systèmes d'informations hétérogènes, dans une logique d'interopérabilité.

Tous deux sont normalisés au moyen de *RFC* et ont une syntaxe générique mais extensible qui leur permet de structurer une grande variété de contenus tout en les rendant facilement malléables et adaptables à des besoins divers et en restant définis et validables par des schémas qui permettent de contrôler leur structuration.

Ces formats de fichiers ne présentent d'autre difficulté de préservation que l'identification et la conservation associée des schémas permettant de contrôler la syntaxe et la sémantique de chacune de leurs déclinaisons. De nombreux outils de validation génériques sont par ailleurs disponibles sur internet.

6. Les sites web (internet, intranet, blogs, etc.)

6.1. Caractéristiques des sites internet

Les années 1990 ont vu l'apparition et la multiplication des sites accessibles à distance, qui ont connu, depuis cette date, plusieurs générations (web 1.0, web 2.0, etc.).

Leurs usages sont aujourd'hui multiples :

- **diffusion d'informations auprès du public**, en remplacement des anciens outils des services de communication et de marketing : plaquettes institutionnelles, dossiers et communiqués de presse, journaux d'information, affichage des obligations légales, etc. ;
- **diffusion de publicités génériques** ou cibles en fonction du profilage du consommateur ;
- **point d'accès à des services génériques ou personnalisés** : correspondance, réalisation de transactions (consultation de compte client, accès à des documents, acquisitions et règlements de prestations, etc.), consultation de catalogues (vente en ligne, bibliothèques, etc.) ;
- **échanges** entre personnes et réactions à des contenus publiés.

Définition

Ces sites sont accessibles depuis un réseau public (on parle alors de **site internet**) ou privé (on parle alors de **site intranet**). Ils peuvent être institutionnels ou privés (sites de blogs).

La production de ces sites est basée sur l'existence de contenus variés (textes, images, vidéos, etc.) reliés entre eux par des hyperliens. Leur création et leur mise à jour sont la plupart du temps réalisées au moyen de logiciels dits de ***Content Management Systems*** **(CMS)** qui permettent aux utilisateurs de créer le site, de le structurer et de choisir sa mise en page, de créer et de mettre à jour les contenus, de décider si ces contenus doivent être diffusés ou au contraire retirés de la diffusion.

6.2. La préservation des sites web présente plusieurs difficultés

Les sites web présentent plusieurs difficultés en matière de préservation :

- leur contenu est fréquemment mis à jour, parfois plusieurs fois lors d'une même journée (ex. site de presse en ligne) ;
- certains contenus ne sont accessibles qu'après avoir souscrit un abonnement ;
- certains contenus sont dépendants de l'utilisation de technologies (ex. flash) dont la maintenance peut s'arrêter et qui rendent ces contenus inaccessibles dans le temps ;
- un site peut être interrelié avec plusieurs autres sites ;
- certains contenus intégrés sur ces sites ne sont pas facilement accessibles par les outils disponibles pour réaliser des opérations de collecte : fichiers multimédias diffusés en streaming, contenus générés par des bases de données, structure hiérarchique générée dynamiquement ;
- les contenus sont souvent protégés au titre de la propriété intellectuelle ;

- des logiciels malveillants peuvent être cachés.

6.3. Stratégies et solutions à adopter

Depuis la fin des années 1990, des stratégies de collecte ont été mises en place pour assurer une conservation sur le long terme **des sites diffusés sur un réseau public.**

Remarque

Dans plusieurs pays, celle-ci est encadrée par les dispositifs de dépôt légal (ex. France). Un consortium international, l'International Internet Preservation Consortium (IIPC), a également été créé en 2003 à l'initiative de la Bibliothèque nationale de France pour identifier et diffuser les meilleures pratiques, favoriser une large couverture internationale de la collecte et encourager l'adoption de législations permettant la conservation des sites.

Techniquement, la collecte du contenu des sites peut être réalisée de plusieurs manières :

6.3.1. La réalisation d'exports depuis les CMS

Cette méthode présente cependant l'inconvénient de ne pas prendre en compte la mise en page des contenus et rend plus complexe la navigation dans les contenus collectés.

6.3.2. Le téléchargement unitaire de chaque page

et leur enregistrement dans une arborescence de fichiers. Cette méthode est longue et présente les mêmes problèmes que la précédente.

6.3.3. La collecte automatique des pages par moissonnage

La collecte automatique des pages par *moissonnage*, en utilisant des robots, opérée au moyen d'outils :

- soit développés par les institutions chargées de la conservation. Plusieurs outils ont été proposés : PANDORA *Digital Archiving System* (PANDAS) par la Bibliothèque nationale d'Australie (outil propriétaire) ; *Web Curator Tool* (WCT) par la Bibliothèque nationale de Nouvelle-Zélande, la British Library et une société privée (outil ouvert) ; *NetarchiveSuite*, développé par Det Kongelige Bibliotek et *Statsbiblioteket* du Danemark et devenu libre avec de larges contributions de la Bibliothèque nationale de France et de l'Österreichische Nationalbibliothek ;
- soit mis à disposition par des fournisseurs de service bien établis comme *Internet Archive* ou *Internet Memory Foundation*. Internet Archive a notamment développé, avec plusieurs bibliothèques nationales, le logiciel ouvert Heritrix. Ces fournisseurs proposent deux approches :
  - la réalisation par eux-mêmes de la collecte ;
  - la mise à disposition de services et d'outils (cf. Archive-It qui offre un service de collecte et d'hébergement de sites web – https://archive-it.org/)
- soit développés par des particuliers comme l'aspirateur de sites libre HTRack.

Stratégie possible pour préserver les sites web

- Empaquetage dans un conteneur au format WARC (Web Archive), développé et maintenu par l'International Internet Preservation Consortium et normalisé par l'ISO (ISO 28500:2009)

6.4. Points d'attention

- La collecte régulière d’un site par moissonnage peut se révéler volumineuse et redondante, une même page pouvant être collectée sans qu’aucun changement ne soit intervenu sur celle-ci. La définition d’une périodicité de collecte adaptée aux mises à jour des contenus est essentielle ;
- Le recours à un prestataire nécessite la mise en place d’un dispositif d’assurance qualité ;
- Les fournisseurs de service de moissonnage offrent des services de collecte et d’hébergement, mais pas de services de préservation ;
- L’accès aux sites collectés doit être conforme aux réglementations en matière de protection de la propriété intellectuelle.

7. Les réseaux sociaux

7.1. Caractéristiques des réseaux sociaux

Les années 2000 ont vu l'apparition de plateformes offrant **la possibilité à des utilisateurs de constituer et d'animer des réseaux ou des communautés dynamiques permettant de diffuser et d'échanger des informations.** Elles sont connues sous l'appellation de **réseaux sociaux** dont les plus utilisés sont Facebook, X (ex-Twitter), Instagram, Snapshat, LinkedIn, Viadeo ou TikTok.

Utilisables de manière gratuite ou payante, proposant une large gamme de fonctionnalités, ces plateformes fondent leur modèle économique sur l'exploitation des données diffusées et échangées par leurs utilisateurs, ce qui leur permet d'identifier leur profil et de faciliter la diffusion de publicités ciblées. Le volume des données collectées par ces plateformes intéresse par ailleurs de plus en plus de chercheurs désireux de comprendre et d'analyser ces communautés et leurs réactions aux événements.

7.2. La préservation à long terme de ces plateformes de réseaux sociaux présente plusieurs difficultés

La préservation de ces plateformes de réseaux sociaux présente plusieurs difficultés :

- les contenus publiés et échangés sur ces plateformes sont soit diffusés publiquement, soit accessibles uniquement à des utilisateurs abonnés ;
- les contenus publiés et échangés sur ces plateformes sont **volumineux** et **liés entre eux** par un système de liens externes ;
- les contenus publiés et échangés sur ces plateformes sont **modifiables** (ex. changement par le titulaire du compte de l'apparence de celui-ci) et peuvent faire l'objet d'interactions avec d'autres utilisateurs sur une longue durée (ex. ajout de commentaires, réactions sous forme de « like ») ;
- la collecte et la réutilisation des contenus diffusés sur ces plateformes sont souvent restreintes par les conditions générales d'utilisation imposées par leurs éditeurs – ainsi, le *Developer Agreement and Policy* de la plateforme Twitter restreint la quantité de données récupérables et interdit leur diffusion sur internet ;
- les contenus mis à disposition sont souvent sélectionnés au moyen d'un algorithme utilisé par ces plateformes, et dont les spécifications ne sont pas nécessairement connues.

7.3. Stratégies et solutions à adopter

Plusieurs stratégies de collecte et de préservation de ces contenus sont possibles :

- le recours aux techniques utilisées pour la collecte et la préservation des sites web et l'utilisation des formats de préservation correspondants (cf. supra) ;
- la collecte des contenus (données et métadonnées) en utilisant directement les interfaces de programmation d'application (API) mis à disposition par ces plateformes et leur récupération sous la forme de fichiers structurés de type JSON ou XML ;

- l’achat de contenus à des revendeurs tiers, titulaires de contrats particuliers avec les éditeurs de ces plateformes, ce qui leur offre des fonctionnalités supplémentaires. La plateforme Twitter dispose par exemple d’un revendeur officiel, Gnip, qui constitue une de ses filiales ;
- la négociation d’un accord avec ces plateformes, ce qui n’est possible que pour des institutions ayant une certaine assise (Archives ou Bibliothèques nationales) ou dans le cadre d’une recherche scientifique.

7.4. Points d'attention

- la quantité de contenus (données et métadonnées) collectées via les API varie en fonction des plateformes. Certaines proposent même différents types d’API avec différentes fonctionnalités ;
- les fonctionnalités de collecte offertes aux titulaires des comptes eux-mêmes sont souvent plus fournies que celles autorisées à des tiers. Il s’agit sans doute de la méthode la plus adaptée pour un archiviste désireux de récupérer les billets publiés par son organisation sur les réseaux sociaux ;
- la collecte des contenus par API permet de récupérer davantage de contenus qu’une collecte par moissonnage. Ainsi, la plateforme Twitter permet via ses API de récupérer l’identifiant des utilisateurs, la géolocalisation de celui-ci, les opérations réalisées sur chaque contenu (partage, « like »). Les contenus sont considérés comme étant plus authentiques et plus complets ;
- la collecte des contenus diffusés par un utilisateur donné peut être restreinte par les paramètres de confidentialité adoptés par celui-ci.

8. Les images fixes

8.1. Caractéristiques des images fixes

Les images fixes, héritières de la photographie et produites soit par des outils de numérisation soit par des appareils de prise de vues numériques au moyen de capteurs photographiques, se présentent sous deux formes :

8.1.1. les images dites matricielles

- la représentation prend la forme d'une grille de pixels qui constituent une codification de la couleur sous forme de bits ;
- le nombre de pixels contenus dans une image constitue sa résolution, exprimée en points par pouce (ppi ou dpi en anglais). Plus la résolution est faible, plus les formes sont floues ;
- chaque pixel peut représenter de 1 à 48 couleurs, ce que l'on nomme la profondeur de bits. Pour information, l'oeil humain perçoit des couleurs représentés par une profondeur de 24 bits ;
- les couleurs sont exprimées sous forme de nombres, ce que l'on nomme l'espace colorimétrique. Le plus courant est le RVB (rouge, vert, bleu).

8.1.2. les images dites vectorielles

Sur les images dites vectorielles, la représentation prend la forme de points reliés entre eux par des lignes et des courbes, constituant ainsi des formes géométriques diverses, Celles-ci sont enregistrées sous forme de formules mathématiques qui peuvent être particulièrement complexes.

8.2. La préservation à long terme des images fixes présente plusieurs difficultés

8.2.1. Le nombre de formats de fichiers pour représenter des images fixes est très important

Le nombre de formats de fichiers développé par l'industrie pour représenter des images fixes est très important, même si certains formats de fichiers sont relativement répandus (ex. JPEG). Certains d'entre eux sont propriétaires.

8.2.2. Certains formats de fichiers ne sont pas pris en charge par les logiciels de visualisation

Certains formats de fichiers ne sont pas pris en charge par les logiciels de visualisation d'images disponibles par défaut sur les appareils grand public (ex. fichiers au format RAW ou JPEG 2000).

8.2.3. Certains formats de fichiers utilisent des algorithmes de compression

Certains formats de fichiers correspondant à des images matricielles utilisent des algorithmes de compression pour réduire la taille des fichiers. Cette compression peut être faite avec perte (JPEG) ou sans perte (JPEG 2000). Une compression avec perte réduit la taille du fichier et la qualité de l'image. Une compression sans perte ne réduit que la taille du fichier.

8.2.4. Les propriétés des images sont à prendre en compte

Les propriétés des images (résolution, profondeur de bits, espace colorimétrique, compression) sont importantes à prendre en compte lors de toute opération de préservation. Leur conservation constitue un impératif.

8.2.5. La taille des images fixes (en octets)

Le nombre et la taille en octets des images fixes numériques sont beaucoup plus importants que celle des images fixes argentiques. Une sélection s'avère souvent indispensable.

Complément

Il convient de signaler qu'existent pour les images fixes plusieurs normes de métadonnées :

- **norme *Exchangeable image file format* (EXIF)**, dans le cas d'images générées au moyen de scanners et d'appareils photographiques. Ces métadonnées ne peuvent être modifiées une fois la capture effectuée ;
- **norme *International Press Telecommunications Council* (IPTC)**. Elle offre la possibilité d'enrichir la description des images fixes avec de nombreuses informations descriptives comme le nom du photographe, des mots-clés, des informations sur les droits d'auteur, des informations sur la localisation de la prise de vue ;
- **norme *Extreme Memory Profile* (XMP)** développée par la société Adobe pour enregistrer différents types de métadonnées sous forme de balises.

8.3. Stratégies et solutions à adopter

Plusieurs stratégies de préservation sont possibles pour les images fixes, même s'il n'existe pas de format parfait :

8.3.1. Le format Tagged Image File Format (TIFF)

Longtemps, le format ***Tagged Image File Format* (TIFF)** a été privilégié pour la conservation des images fixes, mais il présente l'inconvénient de ne permettre aucune compression, même sans perte.

Par conséquent, la taille des fichiers enregistrés dans ce format est souvent importante. Depuis quelques années, certaines institutions ont fait le choix de privilégier le format *Joint Photographic Experts Group 2000* (JPEG 2000) qui prend en charge la compression avec ou sans perte. Il s'agit cependant d'un format complexe, qui ne permet pas facilement la migration vers d'autres formats.

8.3.2. Les formats JPEG et Graphic Interchange Format (GIF)

Ces deux formats sont généralement considérés comme des formats acceptables pour la préservation sur le long terme.

Attention

Il faut faire attention, **pour le format JPEG**, au fait qu'il peut avoir fait l'objet de déclinaisons conformément aux choix opérés par certains industriels producteurs de matériel. Par ailleurs, lorsque des modifications sont enregistrées, des altérations peuvent se produire.

8.4. Points d'attention

- des outils de validation de formats existent (ex. JHOVE, Jpylyzer et DPF Manager), mais **les résultats fournis par ceux-ci ne sont pas toujours cohérents** et dépendent de la capacité des personnes qui les ont spécifiés à interpréter les normes décrivant le format des fichiers correspondants ;
- les formats sans compression ou avec compression sans perte sont à utiliser en priorité pour une conservation à long terme ;
- le choix d'un format de préservation d'une image fixe doit concilier maintien des caractéristiques de celle-ci, taille du stockage et besoins des utilisateurs ;
- en cas d'opération de migration, il est indispensable de mettre en place un contrôle qualité garantissant que les caractéristiques des images (profondeur de bits, espace colorimétrique, métadonnées embarquées) ne sont pas altérées par celle-ci.

9. Les données géographiques et géospatiales

9.1. Caractéristiques des données géographiques et géospatiales

Les formats de fichiers permettant de représenter des données géographiques et géospatiales se sont largement répandus depuis les années 1990, à la faveur de la diffusion des systèmes d'information géographiques (SIG) dans les structures chargées d'urbanisme, d'architecture, d'aménagement du territoire, de planification urbaine, d'archéologie, de transport, etc.

Techniquement, **les catégories de formats de fichiers sont les mêmes que pour les images fixes** (images matricielles et images vectorielles). Cependant, les formats de fichiers correspondant à des données géographiques ou géospatiales présentent un certain nombre de spécificités :

- le référentiel géospatial constitue un élément clé de compréhension des données enregistrées dans les fichiers (ex. système de projection utilisé) ;
- les données sont organisées en couches correspondant chacune à une catégorie d'objets géolocalisés (rues, restaurants, rivières, arbres, bouches d'incendies) ;
- les normes internationales de métadonnées spécifiques à cette catégorie de données se sont multipliées (ex. ISO 19115 Informations géographiques - Métadonnées).

9.2. La préservation à long terme des données géographiques et géospatiales présente plusieurs difficultés

Les formats de fichiers permettant de représenter des données géographiques et géospatiales présentent les défis suivants en matière de préservation :

- **la mise à jour des données est fréquente** – sans historisation systématique –, pour prendre en compte les évolutions de la structure de la terre et surtout de l'occupation humaine de celle-ci. La définition de la fréquence de collecte des données est donc essentielle ;
- **la documentation du système de projection géographique utilisé est essentielle** à la préservation à long terme des données ;
- **plusieurs formats de fichiers sont propriétaires**, voire dépendants de l'outil avec lequel ils ont été créés (ex. format ArcGIS Pro project file) ;
- la visualisation par les utilisateurs des données géographiques associe souvent différents composants (tuiles matricielles, couches de données vectorielles, interface graphique) liés entre eux pour constituer une carte. La gestion des données nécessaires pour recréer ces vues est essentielle, mais peut s'avérer complexe. L'utilisation d'un SIG pour réaliser cette restitution peut s'avérer indispensable ;
- **aucun outil de validation des formats de fichiers correspondant à des données géographiques n'existe**.

9.3. Stratégies et solutions à adopter

En matière de stratégie de préservation, il est actuellement recommandé d’utiliser :

- soit des formats de fichiers correspondant à des normes ouvertes (ex. GeoJSON, GeoTIFF),
- soit des formats propriétaires largement répandus dans les outils commerciaux, permettant un accès facile via les outils disponibles sur le marché (ex. ESRI Shapefile ou fichier de forme qui permet de stocker des données vectorielles).

9.4. Points d'attention

- Dans bien des cas, il convient de conserver le format originel ;
- métadonnées et documentation sont essentielles, notamment pour garantir une restitution des données avec les outils disponibles sur le marché.

10. Les formats de dessins conçus par ordinateur

10.1. Caractéristiques des formats de dessins conçus par ordinateur

Les formats de fichiers de conception assistée par ordinateurs (CAO) permettent de concevoir les spécifications techniques de produits et de modèles au moyen de dessins volumineux et complexes.

Ils sont utilisés dans de nombreux métiers : architecture, ingénierie, génie civil, mécanique, construction automobile, etc.

Ils se distinguent de l'imagerie tridimensionnelle par le maintien d'une représentation en deux dimensions.

10.2. La préservation à long terme des formats de dessins conçus par ordinateur présente plusieurs difficultés

Les formats de fichiers de CAO présentent un certain nombre de défis en matière de préservation :

- la **dépendance vis-à-vis des logiciels de conception coûteux est forte**. L'interopérabilité est rendue difficile par la multiplication des formats concurrents et l'obsolescence rapide des logiciels et des formats ;
- **les systèmes sont rarement compatibles entre eux**, car dépendants de noyaux de modélisation mathématique complexes et différents entre systèmes. La migration d'un format vers un autre est donc rendue difficile ou du moins risquée en raison de pertes de données ;
- **les liens vers des données externes sont nombreux** ;
- le **nombre de fichiers** produits pour chaque projet est conséquent et la **taille des fichiers** est importante.

10.3. Stratégies et solutions à adopter

Plusieurs stratégies de préservation ont été expérimentées par le *Massachusetts Institute of Technology* (MIT) ou les services d'archives spécialisés dans l'architecture :

- s'il n'existe pas de format unique et parfait, plusieurs formats de fichiers sont jugés acceptables pour la préservation : ***Industry Foundation Classes*** **(IFC)** pour les données de construction ; ***STandard for Exchange of Product model data*** **(STEP)** – même si ce format est peu pris en charge par les logiciels métier ; ***AutoCAD Drawing*** ou ***Microstation Drawing***, malgré leur caractère propriétaire, en raison de leur large prise en charge par les logiciels de CAO ; ***Initial Graphics Exchange Specification*** **(IGES)**, format ouvert utilisé pour l'échange de modèles géométriques et pris en charge par un grand nombre de systèmes de CAO – avec cependant une mise en œuvre de la norme différente ;
- **l'émulation** constitue une alternative sérieuse à la réalisation de migrations de format.

10.4. Points d'attention

- La dépendance aux références externes doit être résolue au moment de la collecte des fichiers ;
- lors des opérations de migration de format, il faut absolument vérifier que la géométrie des modèles de données a été correctement préservée (notamment les données paramétriques). La conservation des fichiers d'origine constitue une solution de secours souvent indispensable ;
- le nombre de couches de données peut être important. Il convient d'envisager toutes les contraintes associées à leur collecte et à leur migration ;
- des normes de métadonnées existent : Buildm pour les objets architecturaux, Ifcm pour les fichiers *Industry Foundation Classes* (IFC), *Construction-Operations Building information exchange* (COBie) pour l'installation et la réception de chantier des bâtiments.

11. L’imagerie tridimensionnelle (3D)

11.1. Caractéristiques de l''’imagerie tridimensionnelle (3D)

L’imagerie tridimensionnelle (3D) connaît un essor certain depuis les années 1990 dans le domaine du cinéma et du jeu vidéo, mais aussi dans ceux de l’industrie, de l’armée, de la médecine, de l’architecture et du patrimoine (musées, monuments historiques, archéologie, par exemple).

Utilisée pour représenter personnages et objets dans un espace tridimensionnel, l’imagerie 3D présente la particularité de recourir à diverses techniques :

- les nuages de points, permettant de représenter un objet ou un espace sous forme de millions de points placés selon un repère de coordonnées mais non reliés entre eux ;
- les surfaces maillées (mesh), elles aussi composées de points, mais reliés sous forme de triangles ;
- l’orthorectification, permettant de supprimer des informations de perspective et de terrain d’une image ;
- les squelettes, qui permettent de placer des axes de rotation sur un maillage et ainsi de donner les mouvements désirés à la figure créée en 3D, comme le fait l’articulation des os et des tendons pour un squelette humain ou animal ;
- le skinning, qui permet de relier les surfaces maillées au squelette et de lui donner ainsi l’équivalent d’une peau.
- les voxels, équivalents des pixels, qui permettent de stocker une information physique (couleur, densité, intensité) correspondant à un point d’un volume sur un maillage régulier ;

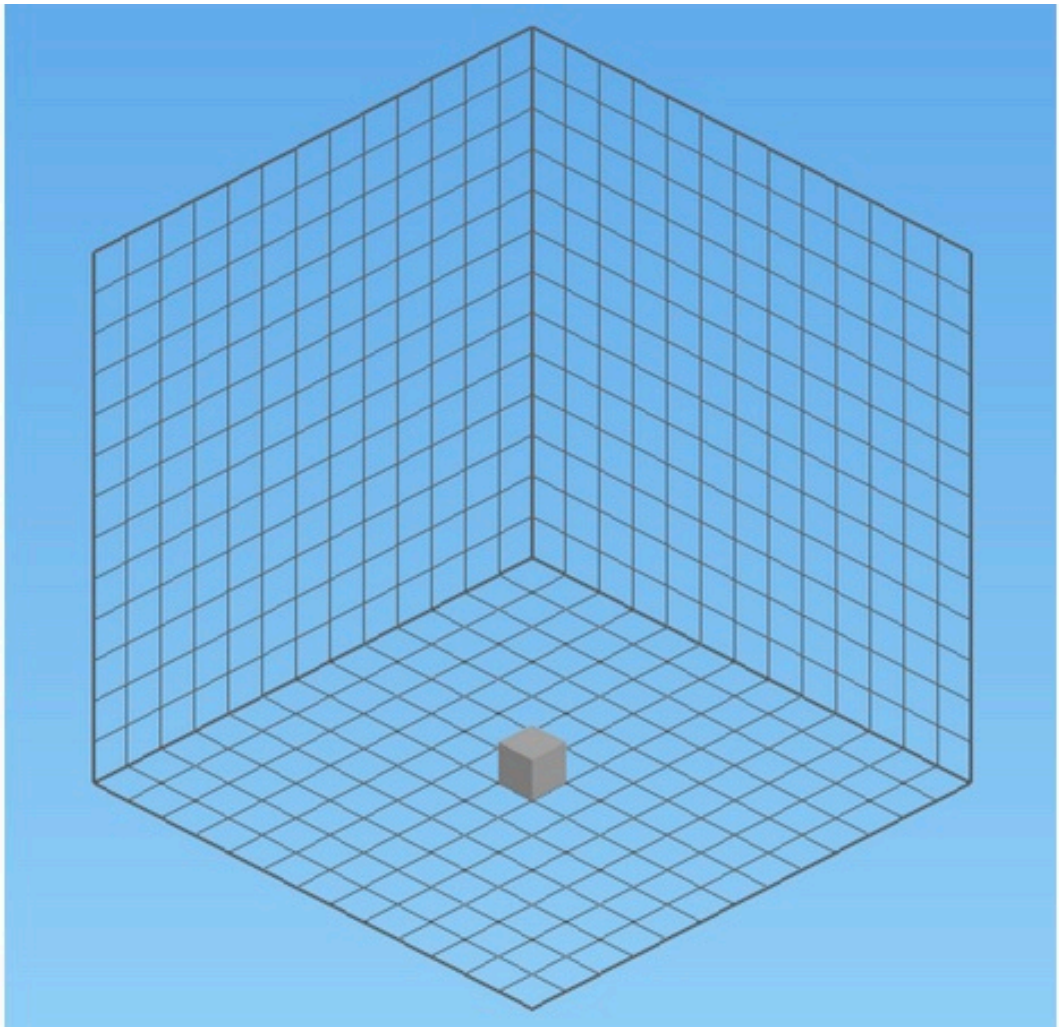


Fig.2 : illustration d'un voxel (source : https://www.megavoxels.com/learn/what-is-a-voxel/)

11.2. La préservation à long terme de l'imagerie tridimensionnelle présente plusieurs difficultés

Les formats de fichiers correspondant à l'imagerie tridimensionnelle présentent un certain nombre de défis en matière de préservation :

- les données sont souvent complexes, réparties entre plusieurs fichiers sources ;
- les formats de fichiers sont principalement propriétaires, et dépendants des plateformes matérielles et logicielles, ce qui limite l'interopérabilité entre environnements techniques ;
- les données sont souvent dépendantes d'autres données ou des systèmes nécessaires pour y accéder (système d'information géographique, moteurs de jeu, équipements).

11.3. Stratégies et solutions à adopter

À ce stade, il n'existe pas de consensus sur les formats de préservation à retenir pour l'imagerie tridimensionnelle. Si certains formats sont considérés comme acceptables par des institutions comme la Bibliothèque du Congrès en raison de leur caractère ouvert, l'émulation constitue également une solution à étudier de près.

11.4. Points d'attention

- La *Community Standards for 3D Data Preservation* (CS3DP – https://cs3dp.org/) offre un certain nombre de ressources en ligne sur la préservation des fichiers d’imagerie tridimensionnelle (travaux des groupes de travail sur les thèmes de l’accès, des bonnes pratiques, des métadonnées, de la gestion des droits, etc.) ;
- des normes de métadonnées existent (ex. Buildm pour l’imagerie architecturale) ;
- un travail en amont avec les producteurs d’imagerie tridimensionnelle s’avère souvent indispensable pour s’assurer que les formats choisis pour créer les fichiers sont gérables et que métadonnées et documentation existent et sont suffisantes ;
- une documentation du processus de production (environnement matériel et logiciel de production, convertisseurs utilisés, etc.) peut s’avérer utile ;
- dans la mesure où il n’existe pas de consensus sur les formats de préservation, il est important de conserver les formats de fichiers d’origine ;
- il n’existe aucun outil de validation des formats de fichiers d’imagerie tridimensionnelle et seuls quelques formats sont identifiables au moyen de PRONOM.

12. Les enregistrements sonores et audiovisuels

12.1. Caractéristiques des enregistrements sonores et audiovisuels

Les enregistrements sonores et audiovisuels sur support sont omniprésents aujourd'hui, qu'ils aient nativement été produits sous cette forme ou soient le produit de la numérisation de documents créés sous forme analogique.

12.1.1. Les enregistrements sonores et audiovisuels sur support numérique sont structurés de deux manières

- **sous la forme de fichiers (file-based)** : il existe des centaines de formats de fichiers possibles, ouverts ou propriétaires, dépendants ou non d'environnements matériels et logiciels spécifiques ;
- **sous la forme de flux (stream-based)** : les enregistrements sont de plus en plus souvent créés à la volée et diffusés directement sur des plateformes de diffusion comme YouTube et Instagram. Ces enregistrements n'existent que sous la forme d'un flux de données et n'aboutissent pas à la création d'un fichier au sens traditionnel du terme. Même si une version sous forme de fichier existe, elle n'est pas nécessairement facilement disponible.

12.2.2. Caractéristiques des formats de fichiers correspondant à des enregistrements sonores et audiovisuels

Les formats de fichiers correspondant à des enregistrements sonores et audiovisuels présentent un certain nombre de caractéristiques qui nécessitent d'être explicitées :

- **il s'agit de formats conteneur** (ex. le format WMV de Windows Media Player) qui non seulement embarquent des composants différents dont chacun doit être pris en compte (flux vidéo, flux audio, flux audios supplémentaires pour les doublages avec une piste par langue ou des sous-titres), mais aussi permettent d'identifier les données, de comprendre les types de flux de données présents et les informations à leur sujet, de stocker des données temporelles ou des métadonnées d'identification. Les conteneurs déterminent l'extension appropriée pour le fichier ;
- **les flux de données audio et vidéo sont encodés et décodés au moyen d'un codec**, qui peut utiliser une compression avec ou sans perte pour mettre en œuvre une transmission ou un stockage ;
- **ils sont constitués de trames**, correspondant à une série d'images pour la vidéo – chaque trame correspond à une image affichée pendant une durée déterminée – et aux échantillons audio pris pendant l'intervalle de la trame vidéo ;

12.1.3. Pour ce qui est des flux vidéos…

- chaque image est composée d'une **matrice de pixels** plus ou moins fine (la résolution) ;
- la **couleur** est une composante essentielle du flux vidéo. Différents espaces colorimétriques sont utilisés dans les vidéos, en fonction de leur source et des migrations qui ont eu lieu. Les espaces colorimétriques les plus courants pour les documents audiovisuels sont le Rouge-Vert-Bleu (RVB), le YUV et le YCbCr ;
- la **profondeur de bits** fait référence à la quantité d'information stockée pour les images qui apparaissent à l'écran. La profondeur généralement recommandée pour la vidéo est de 8 bits, ce qui signifie qu'il y a 256 couleurs possibles pour un pixel particulier ;
- les fichiers vidéos sont également caractérisés par leur **fréquence d'images** qui déterminent la vitesse à laquelle les choses se déroulent. Pour les enregistrements sur support numérique, les fréquences sont variées ;
- les **rapports d'aspect** déterminent la largeur et la hauteur d'une image et la façon dont elle est affichée. Les plus connus sont le 4:3, utilisé dans la télévision traditionnelle à définition standard, le 16:9, utilisé par la télévision haute définition, le 21:9 utilisé dans le cinéma moderne et le 19:10 utilisé dans les fils IMAX. Avec le développement des réseaux sociaux, les formats carré (1:1) et portrait (9:16) sont de plus en plus populaires ;
- **l'entrelacement** permet d'optimiser la perception du mouvement dans un matériel vidéo avec perte. Il peut être repéré quand des lignes irrégulières apparaissent aux endroits où il y a du mouvement. L'entrelacement a été fréquent quand les signaux vidéo devaient être envoyés plus rapidement que ne le permettait le transfert de chaque image complète. Cette pratique n'est plus utilisée dans la vidéo contemporaine où l'optimisation de la bande passante est réalisée autrement ;
- Les **timecodes** attribuent un numéro à chaque image, selon le format heures, minutes, secondes et images (HH:MM:SS:FF). Les timecodes peuvent être intégrés dans les images elles-mêmes, et ils apparaissent donc à l'écran pour chaque image. Ils peuvent sinon être soit stockés dans le fichier, soit être inscrits dans une piste séparée. Les formats les plus importants de timecodes sont les suivants :
    - *Burnt-In Time Code* (BITC) : les données stockées dans la trame de l'image et ne peuvent pas être supprimées.
    - *Linear Timecode* (LTC) : les données sont sur une piste audio séparée.
    - *Vertical Interval Time Code* (VITC) : les données sont stockées dans l'intervalle de suppression verticale d'une piste vidéo. Cela signifie que les données sont stockées dans le flux vidéo, sur une seule ligne de balayage non visible.

12.1.4. Pour ce qui est des flux audios

- **les échantillons correspondent à des valeurs à un moment précis dans le temps**, sachant que l'audio est souvent décrit par son taux d'échantillonnage, généralement exprimé en échantillons (en Hertz – Hz ou en cycles par seconde) par seconde ;
- **les flux audio peuvent englober plusieurs canaux** : les fichiers audio mono comprennent un ou 2 canaux avec le même contenu ; les fichiers stéréo contiennent 2 canaux distincts ; le son surround tente de créer une expérience auditive.

12.2. La préservation à long terme des enregistrements sonores et audiovisuels présente plusieurs difficultés

Les formats de fichiers correspondant aux enregistrements sonores et audiovisuels présentent un certain nombre de défis en matière de préservation :

- **les fichiers correspondants sont souvent volumineux** et gourmands en espace de stockage. Ils sont difficiles à transférer et à prendre en charge dans un système d'archivage. Le temps nécessaire pour procéder aux vérifications et au contrôle de qualité, ainsi qu'aux opérations de caractérisation (identification et validation de format, extraction de métadonnées) est important ;
- **la connaissance des méthodes d'encodage des flux et des caractéristiques techniques** de ceux-ci est essentielle pour bien gérer la préservation ;
- il est essentiel de capturer **l'exhaustivité des flux et données** présents dans un conteneur ;
- les enregistrements sont la plupart du temps soumis à des **droits de propriété intellectuelle complexes**, voire à une gestion des droits numériques (DRM) ou protégés par d'autres moyens de lutte contre le piratage qui rendront la capture inopérante ;
- **la capture des enregistrements structurés sous forme de flux est plus complexe que celle des enregistrements structurés sous forme de fichiers**. Des outils de moissonnage de sites web peuvent être utilisés comme Webrecorder ou Wayback d'Internet Archive. Une autre option consiste à télécharger le site web à l'aide d'un outil en ligne de commande (wget), ce qui permet de récupérer les fichiers intégrés ;
- la manipulation des enregistrements sonores et audiovisuels peut nécessiter un **démultiplexage** et un **remultiplexage** des flux contenus dans les formats de fichiers conteneurs, pour leur faire subir des transformations séparées, opérations qui peuvent ne pas être anodines en portant atteinte aux flux eux-mêmes.

12.3. Stratégies et solutions à adopter

La conservation des enregistrements sonores et audiovisuels étant un domaine d'ores et déjà bien établi, un certain consensus se dégage sur les formats de fichiers à privilégier pour assurer une bonne conservation de ceux-ci, même si aucun format ne peut être considéré comme parfait :

- **pour les enregistrements sonores**, il est recommandé de choisir un format de fichiers non compressé comme WAV et FLAC, avec un échantillonnage à 96 kHz avec un profondeur de 24 bits ;
- **pour les enregistrements audiovisuels**, il n'existe pas de norme bien établie et acceptée, ce qui peut rendre difficile la prise de décision. Certains formats conteneur sont privilégiés comme le Matroska et le codec vidéo FFV1 ou comme le *Material Exchange Format* (MXF).

12.4. Points d'attention

12.4.1. Plusieurs normes de métadonnées sont disponibles pour décrire les enregistrements sonores et audiovisuels

- AES57, norme de l'Audio Engineering Society, pour les enregistrements audio ;
- EBUCore, fondée sur le Dublin Core, pour les enregistrements audiovisuels ;
- EN 15907, norme européenne pour la description des œuvres cinématographiques ;
- ID3, métadonnées pour des fichiers audio ;

- PBCore, également fondée sur le Dublin Core, développée pour les enregistrements radiophoniques ;
- VRA Code, élaborée par la Bibliothèque du Congrès, pour les œuvres d'art visuel.

12.4.2. Des outils de validation existent pour quelques formats de fichiers

- MediaConch pour les fichiers encodés en Matroska (MKV), LPCM ou FFV1 ;
- En matière d'extraction de métadonnées, MediaInfo, FFmpeg, ou Exiftool peuvent traiter des enregistrements sonores et audiovisuels.

12.4.3. Une communauté importante existe autour de la préservation des images animées et du son

Cette communauté peut être d'un grand recours en cas de question.

Exemple

Parmi d'autres, on peut citer l'Association of Moving Image Archivists, la Fédération internationale des archives du film, l'Association internationale des archives sonores et audiovisuelles, l'Association for Recorded Sound Collections, l'Audio Engineering Society, l'International Association of Sound and Audiovisual Archives. Par ailleurs, de nombreuses institutions nationales (comme la BnF et l'INA en France ou BAC et BanQ au Canada) disposent d'une expérience solide en matière de préservation des enregistrements sonores et audiovisuels.

13. Les logiciels

13.1. Caractéristiques des logiciels

Définition

Les logiciels sont les outils permettant de créer, modifier, consulter et supprimer les documents d'archives sur support numérique. Ils constituent en eux-mêmes une trace des activités humaines, un élément essentiel de l'histoire des techniques.

D'un point de vue technique, les logiciels sont constitués d'un code source, traduit par un compilateur et exécuté par un ordinateur. Le code source est rédigé dans un langage qui a pu varier dans le temps.

Fondamental

La question de leur conservation est d'autant plus cruciale qu'ils peuvent constituer un outil de préservation, que ce soit pour consulter des documents créés dans un format donné ou pour réaliser des opérations de préservation, notamment des opérations de migration de format.

13.2. La préservation à long terme des logiciels présente plusieurs difficultés

13.2.1. Le code source des logiciels

Il peut être écrit dans un langage qui n'est plus utilisé dans l'écriture de nouveaux logiciels. Comprendre le fonctionnement d'un logiciel peut donc amener à apprendre ou ré-apprendre un langage informatique, ce qui suppose des compétences spécifiques et peut s'avérer critique pour identifier la présence d'éléments malveillants.

13.2.2. Tout logiciel évolue régulièrement

Tout logiciel évolue régulièrement, au gré des versions majeures ou mineures corrigeant des dysfonctionnements (les bugs) ou offrant de nouvelles fonctionnalités en réponse aux attentes de leurs utilisateurs.

Par ailleurs, chaque version majeure ou mineure peut disposer de sous-versions adaptées à un environnement matériel et logiciel (ex. version pour ordinateurs PC ou pour ordinateurs MacIntosh).

Méthode

Il convient donc de définir quelle version et quelle sous-version préserver, en fonction de l'objectif de l'opération de préservation .

13.2.3. Les logiciels sont souvent des objets composites

Ils sont constitués de briques et composants qui évoluent dans le temps indépendamment les uns des autres. Ils font par ailleurs l'objet de paramétrages et de configurations génériques et spécifiques à chacun de leurs environnements d'implémentation. Tout ceci rend leur préservation extrêmement complexe et rend d'autant plus nécessaire leur documentation

13.2.4. Pour garantir un bon usage des logiciels conservés

Pour garantir un bon usage des logiciels conservés, il convient de disposer d'une documentation suffisante relative tant à son installation, qu'à son exploitation et à son fonctionnement.

Il est donc important de collecter, en même temps que le code source, les procédures d'installation, les manuels utilisateurs, les supports de communication diffusés par les éditeurs, voire les échanges sur les forums de discussion qui ont amené à la correction des dysfonctionnements et au développement des nouvelles fonctionnalités.

Cette documentation, souvent dispersée et évoluant au fur et à mesure de l'évolution du logiciel lui-même, est parfois difficile à repérer, quand elle existe.

13.2.5. Les logiciels sont soumis à des droits de propriété intellectuelle

L'utilisation de ceux d'entre eux qui sont propriétaires est soumise à l'octroi d'une licence par l'éditeur du logiciel, ce qui n'est pas sans poser de problème pour les logiciels qui ne sont plus édités ou dont l'éditeur a disparu.

13.3. Stratégies et solutions à adopter

La préservation des logiciels bénéficie cependant de l'existence de nombreuses équipes et communautés intéressées à leur développement et à leur maintenance.

Exemple

Les musées de l'informatique comme le *National Museum of Computing* anglais ou le *Computer History Museum* américain, disposent d'une certaine expérience en la matière.

Exemple

Plus récemment s'est constituée une organisation à but non lucratif, *Software Heritage* (https://www.softwareheritage.org/?lang=fr), résultant d'un partenariat entre l'Unesco et l'Institut national de recherche en sciences et technologies du numérique (Inria), qui a entrepris de collecter, de préserver et de partager le code source de tous les logiciels accessibles au public.

La collecte à laquelle elle procède combine des méthodes manuelles et automatiques, en s'appuyant sur les ressources disponibles dans les dépôts publics comme GitHub, BitBucket, Debian, Google Code et GNU.

Les logiciels collectés sont déposés au Centre européen pour la recherche nucléaire (CERN).

Exemple

Existe également le *Software Preservation Network* (https://www.softwarepreservationnetwork.org/) qui fournit conseils et outils pour la préservation des logiciels.

Conclusion

Les catégories de formats de fichiers sont nombreuses et les formats en évolution constante, ce qui a nécessairement des conséquences sur la préservation numérique. Il est donc indispensable de continuer à s'informer régulièrement pour adapter toujours son action à l'évolution de l'état de l'art.