Module 7C-Section 1 : Les documents d'archives sur support numérique

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

Les documents d'archives sur support numérique présentent quelques spécificités techniques qu'il convient de comprendre et qui les distinguent des documents d'archives sur support physique ou sur support analogique.

Dans le premier cas, celui des documents d'archives sur support physique « traditionnel » (papiers de tous genres, parchemin, papyrus, cire (tablette ou sceau), argile, tissu, maquettes et plans-reliefs, tirages photographiques sur papier ou sur verre, etc.), 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 (photographies nécessitant un agrandissement comme une diapositive, microfilm). Certes, cela ne veut pas dire que la personne qui consulte ces documents pourra les comprendre et celle-ci devra mettre en œuvre pour ce faire des connaissances en matière de paléographie, de diplomatique, de linguistique ou d'analyse de l'image. Néanmoins, il n'est pas besoin d'appareils spécifiques pour rendre lisible l'information concernée.

Le deuxième cas est celui des documents d'archives sur support analogique, à savoir des enregistrements sonores et audiovisuels : 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.

Or, dans le cas **des documents d'archives sur support numérique**, la chose est encore plus compliquée :

- en aucune manière il n'est possible à un être humain de prendre connaissance directement de l'information stockée sur un support numérique. **Il y aura toujours besoin d'un appareil qui permettra de lire l'information stockée sur le support** ;
- cependant, un appareil de lecture ne suffit pas, car l'information est encodée sur le support numérique. L'appareil de lecture doit donc être muni d'un ensemble de **programmes informatiques** (système d'exploitation, logiciels divers et variés) **qui permettent d'accéder à l'information stockée sur le support numérique, de l'interpréter et de la restituer sous une forme lisible** pour un être humain ;
- même restituée sous une forme lisible par un être humain, **l'information peut avoir fait l'objet d'un codage particulier pour représenter l'information**, et qui ne peut devenir intelligible qu'au moyen d'une documentation précise. Sur ce point, le module 7B/section 1 pourra être consulté.

Et tous ces matériels, logiciels et supports de stockage évoluent dans le temps, souvent assez rapidement.

Par ailleurs, les documents d'archives sur support numérique sont plus facilement :

- **reproductibles** : il est plus facile de réaliser une copie d'archives sur support numérique que d'archives sur support physique ou analogique. Un même document d'archives peut être stocké sur de nombreux supports de stockage différents ;
- **modifiables** : falsifier un document d'archives existant sous format analogique est certes faisable, mais c'est sans commune mesure avec la capacité à modifier un document d'archives sur support numérique.

Ils nécessitent donc un traitement spécifique.

Le présent chapitre a pour objectif de :

- permettre aux archivistes de mieux comprendre comment les documents d'archives numériques sont créés, stockés et transmis ;
- comprendre ce qui rend les documents d'archives sur support numérique si spécifiques ;
- appréhender les risques associés à la préservation dans le temps de ces documents d'archives sur support numérique.

Il complète et précise des notions abordées dans le module 7B.

1. Comment les documents d'archives sur support numérique sont-ils créés, stockés et transmis ?

1.1. Comment les documents d'archives sur support numérique sont-ils créés ?

Introduction

Le mode de création des documents d'archives est spécifique sur support numérique, en raison de l'existence native d'un *encodage technique* [p.36] et d'un *codage métier* [p.36]. La présente section développe ces deux points. On consultera avec profit au préalable le module 5 Gestion et traitement des archives courantes et intermédiaires et le module 7B Gestion des documents numériques au stade courant.

1.1.1. L'encodage technique

Au sens le plus strict du terme, l'information numérique est codée en une série de chiffres binaires (des 0 et des 1), ce que l'on appelle des bits, stockés sur un support. À titre d'exemple, le caractère « C » est généralement encodé de la manière suivante en langage binaire : « 0100 0011 ».

La manière dont la série de bits est encodée dépend du type d'informations que l'on veut enregistrer. Il existe des manières différentes d'encoder du texte, des images, du son, de la vidéo, des informations géolocalisées.

La manière d'encoder chaque type d'information fait l'objet d'un ensemble de règles et de conventions, plus ou moins complexes. C'est ce que l'on appelle le *format de fichiers* [p.37]. Ce format est interprété par le matériel au moyen de *logiciels* [p.37] qui sont capables de les traduire pour les utilisateurs (figure 1).

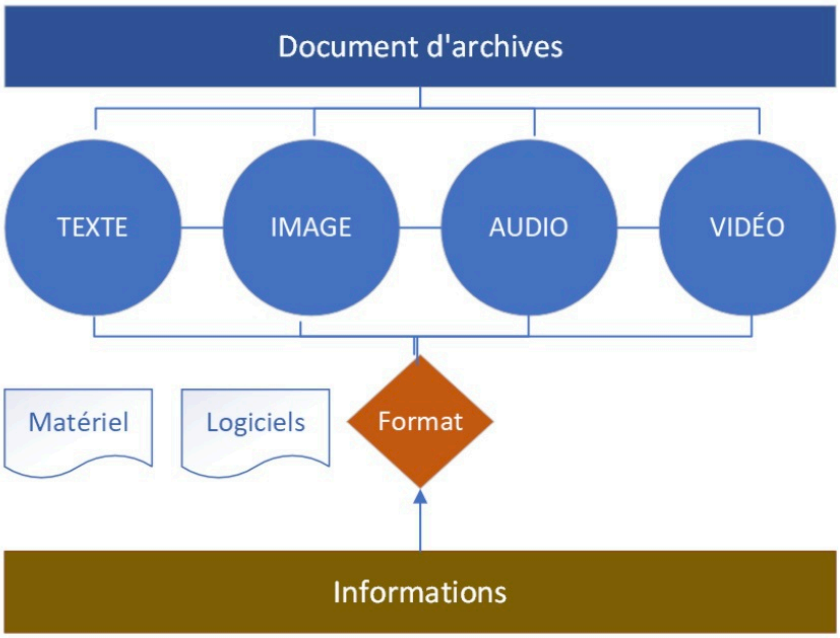

Fig. 1 : Relation entre informations numériques et document d'archives (crédits : B. Grailles/ PIAF)

Attention

Les règles et conventions d'encodage technique de l'information numérique sont plus ou moins clairement connues et sont parfois couvertes par le secret industriel et commercial :

- on parle de **format de fichiers ouvert** quand les règles et conventions (les spécifications du format) sont formalisées dans un document et portées à la connaissance des producteurs et des utilisateurs ;
- on parle de **format de fichiers propriétaire** quand les règles et conventions sont secrètes.

Certaines de ces règles et conventions ont fait l'objet d'une normalisation internationale, soit par un organisme de normalisation de type étatique (comme l'International Standard Organization -- ISO), soit par une organisation à l'origine plus informelle (comme le World Wide Web Consortium qui rassemble des acteurs soucieux de garantir la compatibilité des technologies utilisées sur le web). On parle alors de **format normalisé**.

Ces règles et conventions peuvent être élaborées par des particuliers, des organismes à but non lucratif ou des organismes à but lucratif (des entreprises). Dans ce cas, on parle de **format propriétaire**, qui peut être ouvert (comme le format PDF conçu par la société Adobe) ou fermé (comme le format DOC de Microsoft).

La compression

Complément

Dans certains cas, les règles et conventions prévoient que les informations ne sont pas enregistrées de manière « brute », mais subissent des traitements qui permettent de réduire la taille de l'information enregistrée et stockée. C'est ce que l'on appelle une opération de compression.

Cette compression peut être réalisée :

- **sans perte** : dans ce cas, une fois décompressée, l'information sera strictement identique à l'information d'origine. Des algorithmes de compression sans perte sont utilisés par exemple pour les fichiers qui permettent d'exécuter des programmes (les fichiers exécutables) ou qui encodent du texte. Ex. format Free Lossless Audio Codec (FLAC) ;
- **avec perte** : dans ce cas, une fois décompressée, l'information est plus ou moins identique à l'information d'origine et la qualité est plus ou moins bonne. L'utilisation d'algorithmes de compression avec perte est fréquente pour les informations de type image, son et vidéo, qui sont très volumineuses. La compression d'une image peut avoir tendance à effacer les détails de cette image (on parle souvent dans ce cas d'image pixelisée). Ex. format MPEG-1/2 Audio Layer 3 (mp3).

Encapsulage des différents types d'informations

Dans certains cas, les règles et conventions permettent d'encapsuler différents types d'informations et définissent la façon dont celles-ci s'organisent. On parle alors de **format de fichiers *conteneur*** p.36 (ex. formats ZIP mais aussi les formats de messagerie comme MBOX).

Ces conteneurs permettent souvent de faciliter les exports et les imports d'informations entre logiciels et évitent la manipulation de nombreux objets.

Exemple

Un fichier conteneur vidéo rassemble un ou plusieurs flux d'images, un ou plusieurs flux sonores, des sous-titres, des éléments de chapitrage ainsi que la description des différents flux.

Exemple

Les logiciels de messagerie permettent d'exporter le contenu de celle-ci sous la forme d'un fichier unique, qui comprend à la fois les messages envoyés et reçus, leurs pièces jointes, leurs indexations dans le logiciel, le carnet d'adresses et l'agenda si le logiciel offre cette fonctionnalité (ex. formats PST pour Microsoft Outlook).

Comprendre les règles et conventions d'encodage technique des informations facilite la définition et la mise en œuvre d'opérations de préservation. Chaque type d'information disposant d'un encodage technique propre, le connaître permet de savoir quel procédé mettre en œuvre pour garantir la préservation à long terme de cette information.

1.1.2. La codification métier

Dans l'univers numérique, l'information peut également être codée selon des règles et conventions que nous pourrions qualifier de « métier ».

Deux cas de figures se présentent :

- L'utilisation d'une nomenclature reconnue nationalement ou internationalement (ex. pour la France, le Code officiel géographique ou la Nomenclature des activités françaises ou la Classification géographique type utilisée par Statistiques Canada) ;
- La création d'une codification permettant de simplifier les données enregistrées dans les fichiers, mais dont la saisie peut être transparente pour l'utilisateur (ex. Code de sécurité sociale en France).

Prenons l'exemple d'un registre.

Lorsqu'un individu ou une organisation demandent à être enregistrés, ils doivent fournir une série d'informations précises, comme leur nom, leur prénom (ou leur raison sociale si c'est une organisation), leur date de naissance (ou leur date de création si c'est une organisation), leur adresse, etc. Les informations fournies par l'individu ou l'organisation sont complétées par l'organisation qui tient le registre, avec d'autres informations comme la date d'enregistrement, par exemple.

Dans l'univers physique, ces registres prenaient généralement la forme de volumes reliés pré-imprimés, dont les doubles-pages étaient organisées en colonnes correspondant aux différentes informations fournies par le demandeur ou rajoutées par le gestionnaire du registre. Cette organisation en colonnes reflétait la structure du registre. Généralement, sur chaque double-page, l'en-tête de chaque colonne était répété et indiquait le type d'information qui était attendu dans la colonne. Chaque enregistrement dans le registre était matérialisé par une ligne sur la double-page. Les informations écrites dans chaque colonne, pour chaque enregistrement, étaient rédigées dans un langage majoritairement intelligible - même si on y trouve des abréviations – par la ou les personnes qui tenaient le registre.

Dans l'univers numérique, on retrouve une logique similaire :

- les registres sont structurés avec l'équivalent de « colonnes » (tout dépend de la complexité du registre) ;
- chaque ligne correspond à un enregistrement donné.

Exemple

La plateforme de données ouvertes du ministère de la Culture (France)[1]

1. https://data.culture.gouv.fr/explore/dataset/acces-anticipe-aux-archives-publiques/export/

Simplement, les informations saisies dans chaque « colonne » pour un enregistrement donné peuvent faire l'objet d'un codage complémentaire, défini au moment où celui-ci est passé d'un support physique à un support numérique. Pour rendre l'information intelligible, il faut donc décoder l'information.

Exemple

En France, l'exemple le plus simple pour illustrer cette question est celui du numéro d'immatriculation auprès de la Sécurité sociale dont dispose chaque individu, ce que l'on appelle le NIR (Numéro d'Identification au Répertoire).

Ce numéro est composé d'une série de 15 chiffres, structuré de la manière suivante (ce qui correspond aux colonnes du registre) :

- le sexe de la personne : 1 chiffre ;
- le millésime de l'année de naissance de la personne : 2 chiffres ;
- le mois de naissance de la personne : 2 chiffres ;
- le lieu de naissance de la personne : 2 chiffres pour la zone (numéro du département pour les personnes nées en France ou un code particulier pour les personnes nées à l'étranger) + 3 chiffres correspondant soit à la ville de naissance, soit au pays de naissance ;
- un numéro d'ordre dans le mois de naissance : 3 chiffres ;
- une clé de contrôle : 2 chiffres.

Le sexe et le lieu de naissance de la personne font l'objet d'un codage. Ainsi, le sexe de la personne peut être renseigné avec le chiffre 1 (si c'est un homme), 2 (si c'est une femme) ou 3 (si le sexe est indéterminé).

Les deux premiers chiffres du lieu de naissance correspondent à un code donné au département de naissance. Encore plusieurs départements peuvent-ils avoir le même code, à un moment différent. Avant 1962, le code 91 correspondait au département d'Alger. Depuis 1968, au département de l'Essonne. Le code 99 correspond à tous les pays étrangers.

Quant au code correspondant à la ville de naissance ou au pays de naissance, il est spécifique et n'est pas le même que celui utilisé par les services postaux.

Attention

Il y a donc bien un codage qu'il est important de documenter pour être en mesure de rendre l'information intelligible, quel que soit le format de fichiers utilisé pour encoder techniquement celle-ci.

D'autres cas de figure peuvent également être cités : le référentiel de localisation géographique pour une carte, les unités de mesure et l'échelle pour un plan, etc. Ces informations « métier » sont absolument nécessaires pour interpréter les informations. Encore faut-il les connaître … et les localiser.

1.1.3. Le document d'archives sur support numérique peut prendre la forme d'un agrégat

Dans l'environnement physique, les données constituant un document sont réunies sur un même support : une feuille, la page (ou double-page) d'un volume relié.

Dans l'univers numérique, ces données sont certes toutes sur un même support, numérique naturellement. Simplement, elles ne sont pas nécessairement toutes rassemblées au même endroit (ex. : dans le même fichier). On parle d'*agrégat* p.36.

C'est d'ailleurs ce que l'article 4 de la loi québécoise concernant le cadre juridique des technologies de l'information affirme : « Un document technologique, dont l'information est fragmentée et répartie sur un ou plusieurs supports situés en un ou plusieurs emplacements, doit être considéré comme formant un tout, lorsque des éléments logiques structurants permettent d'en relier les fragments, directement ou par référence, et que ces éléments assurent à la fois l'intégrité de chacun des fragments d'information et l'intégrité de la reconstitution du document antérieur à la fragmentation et à la répartition. » (RLRQ c C-1.1, art 4, en ligne, disponible sur https://canlii.ca/t/19b8#art4 [consulté le 18 novembre 2024]).

Simulation

Prenons l'exemple d'un enregistrement (ex. l'inventaire d'un objet d'art dans un musée). Dans l'environnement physique, en France, le registre d'inventaire prenait en France la forme d'un volume relié dont chaque double-page était composée de 17 colonnes, chacune correspondant à un type d'information (ex. numéro d'inventaire, titre de l'œuvre, provenance, date d'acquisition, matériaux, etc.).

Dans l'environnement numérique, le registre d'inventaire est géré sous la forme d'une application métier appuyé sur un système de gestion de bases de données relationnelles. Les données sont enregistrées dans des tables reliées entre elles. Il pourra y avoir une table pour les provenances, une table pour les matériaux, etc. Et la table centrale où va être enregistré le nouvel objet acquis par le musée va pointer vers des données présentes dans ces différentes tables. L'enregistrement correspondant à l'objet acquis est donc constitué de données réparties dans différentes tables. C'est ce que la norme ICA-Req - Principes et exigences fonctionnelles pour l'archivage dans un environnement électronique - appelle un agrégat de données (cf. schéma ci-dessous).

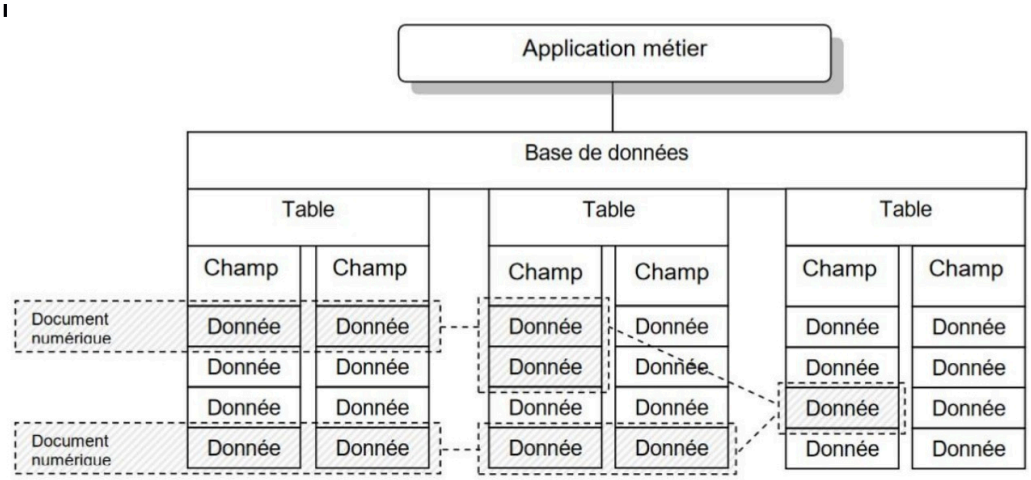

Fig. 2 : Représentation schématique d'un document d'archives sur support numérique comme agrégat de données (source : ICA, Principes et exigences fonctionnelles pour l'archivage dans un environnement électronique. Module 3. Recommandations et exigences fonctionnelles pour l'archivage des documents dans les applications métier, p. 15).

Dans bien des cas, dans l'environnement numérique, les documents d'archives prennent la forme d'agrégats logiques, même si toutes les données composant le document sont stockées sur le même support. L'objectif de l'archiviste, pour déterminer ce qui doit être intégré dans son système d'archivage, consiste à identifier ces agrégats (voir pour cela le module 5 et le module 6 du PIAF).

1.2. Comment les documents d'archives sur support numérique sont-ils stockés ?

2.1. Introduction

Le mode de stockage des documents d'archives sur support numérique a considérablement évolué depuis la création de l'informatique, allant vers une miniaturisation et une extension de la capacité de stockage croissantes (on estime que les capacités de stockage augmentent de 25 % par an).

La question du stockage doit être envisagée sous deux angles différents :

- celle des supports de stockage ;

- celle des méthodes de stockage.

Il convient également d'évoquer le cloud computing, *informatique en nuage ou infonuagique* [p.37], ainsi que la protection de l'accès à des espaces de stockage via le chiffrement.

1.2.1. Les supports de stockage

1.2.1.1. La première génération

Depuis le développement de l'informatique, plusieurs générations de supports de stockage se sont succédé.

La première génération restait fortement dépendante des méthodes employées antérieurement et reposait principalement sur l'utilisation de cartes perforées. Celles-ci sont constituées de morceaux de papier rigide où l'information est stockée sous forme de perforations, selon une logique définie spécifiquement.

Les premières cartes perforées ont fait leur apparition au 18e siècle, en particulier dans le textile et la fabrique d'instruments de musique. Les premières machines à cartes perforées pour le stockage d'informations alphanumériques ont vu le jour à la fin du 19e siècle (recensements de population) et se sont multipliées dans la première moitié du 20e siècle.

Le modèle le plus courant des cartes perforées est la carte à 80 colonnes, feuille de bristol rectangulaire dont un coin est tronqué et où les caractères alphanumériques étaient traduits par des perforations rectangulaires disposées en colonnes parallèles à la largeur et sur 12 lignes parallèles à la longueur.

Les cartes perforées ont progressivement disparu à partir des années 1970 pour le stockage des informations. Elles ont cependant été utilisées pour la saisie des programmes logiciels jusqu'au milieu des années 1980.

Fig. 3 : carte perforée à 80 colonnes (source : https://commons.wikimedia.org/wiki/File:Punched_card.jpg; crédits : Mutatis mutandis, CC BY-SA 3.0 <http://creativecommons.org/licenses/by-sa/3.0/>, via Wikimedia Commons)

1.2.1.2 La deuxième génération

La deuxième génération de supports de stockage repose sur plusieurs technologies magnétiques successives :

Les bandes magnétiques

Les bandes magnétiques : rubans de film plastique enroulés sur une bobine dont une des faces est recouverte d'une couche de matériau magnétique. Elles permettent l'enregistrement et la lecture d'informations analogiques ou numériques à l'aide de différents appareils (magnétophones, magnétoscopes, enregistreur-lecteur de bandes magnétiques). Les bandes magnétiques sont utilisées comme mémoire de masse dès les années 1950. Malgré le développement des disques magnétiques et optiques, les bandes magnétiques (ex. bandes LTO – Linear Tape-Open) restent encore aujourd'hui un support de stockage privilégié en raison de leur grande capacité, de leur bon rapport qualité/prix, de leur caractère amovible, de leur facilité de transport et de leur solidité et fiabilité physiques dans le temps. Elles sont regroupées en baies de stockage ;

Fig. 4 : bande magnétique (source : https://commons.wikimedia.org/wiki/File:Magtape1.jpg ; crédits :Daniel P. B. Smith., CC BY-SA 3.0 <http://creativecommons.org/licenses/by-sa/3.0/>, via Wikimedia Commons)

Les cassettes

Parallélépipède rectangle composé d'une bande magnétique enroulée autour de deux bobines. L'entraînement par un moteur permet de faire se déplacer la bande qui est alors lisible par une tête de lecture. La cassette audio a été utilisée comme support de stockage sur les premiers ordinateurs personnels, via un enregistrement analogique (transformation des signaux numériques en signaux sonores). Le volume de données enregistré était cependant limité. Ce support a été vite abandonné au profit de la disquette dans les années 1980 ;

Fig. 5 : Cassette audio (source : https://commons.wikimedia.org/wiki/File:Audio_cassette_tapes.jpg ; Seth Ilys, Public domain, via Wikimedia Commons)

Les disques durs

Système de mémoire de masse à disque magnétique tournant, utilisé dans divers appareils (ordinateurs, baladeurs numériques, caméscopes, lecteurs/enregistreurs de DVD – Digital Versatile Disc, consoles de jeux vidéo). Inventés dans les années 1950, les disques durs ont vu leur capacité augmenter progressivement (multiplication par 100 000 de celle-ci entre les années 1950 et les années 2000) et leur utilisation s'est progressivement généralisée en informatique, jusqu'à l'apparition des disques SSD (Solid State Drive). Leur capacité se mesure aujourd'hui en téraoctets (To, 1 To = 1000 Go). La rapidité de leur lecture est dépendante de la vitesse de rotation du disque et du temps de positionnement des têtes de lecture ;

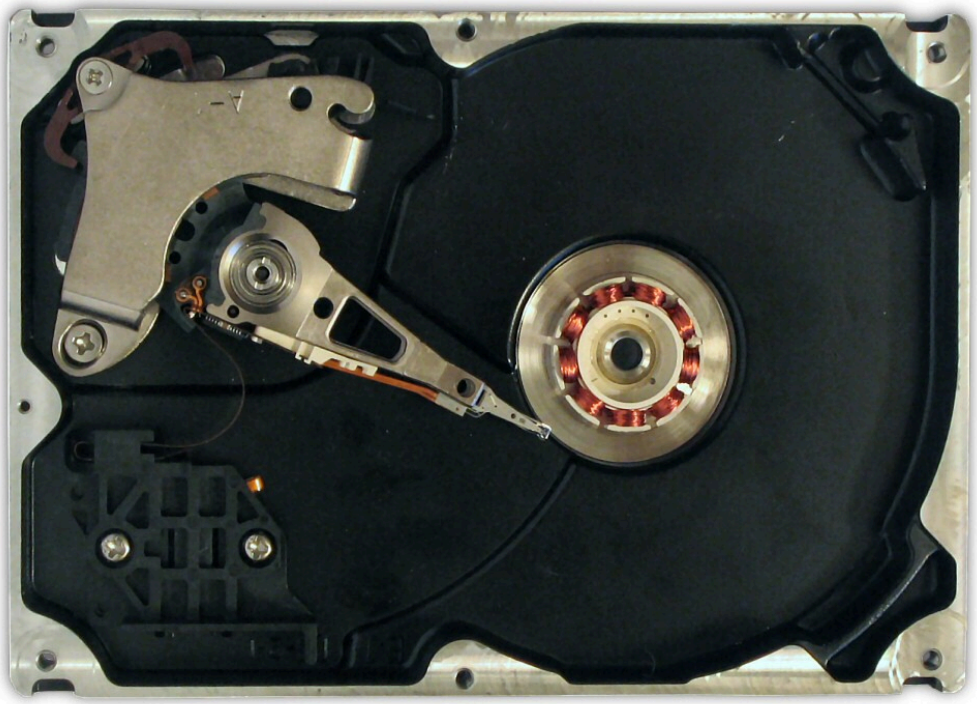

Fig.6 : disque dur (source : https://commons.wikimedia.org/wiki/File:Hard_disk_dismantled.jpg; crédits :No machine-readable author provided. Ed g2s assumed (based on copyright claims)., CC BY-SA 3.0 <http://creativecommons.org/licenses/by-sa/3.0/>, via Wikimedia Commons)

Les disquettes

Support de stockage souple et amovible, composé d'un fin disque de plastique souple renforcé en son centre sur lequel est apposé un substrat magnétique, et enveloppé d'une coque de protection en matière plastique. Deux formats ont été principalement utilisés : 5 pouces 1/4 et 3 pouces 1/2. La commercialisation des disquettes, commencée dans les années 1960, s'est arrêtée en 2010.

Fig. 7 : disquettes (source : https://commons.wikimedia.org/wiki/File:Diskettes.jpg, crédits :Veronidae, CC BY-SA 3.0 <https://creativecommons.org/licenses/by-sa/3.0>, via Wikimedia Commons).

1.2.1.3. La troisième génération

La troisième génération de supports repose cette fois-ci sur des technologies optiques :

Le disque compact ou Compact Disc (CD)

Disque optique utilisé pour stocker des données sous forme numérique, commercialisé à partir du début des années 1980 et largement diffusé à partir du début des années 1990. Portée par l'industrie musicale, sa diffusion s'est ralentie dans les années 2000 avec l'apparition de lecteurs portatifs à mémoire flash intégrée et avec la création de plateformes web proposant des albums entiers en téléchargement ou en écoute instantanée ;

Fig. 8 : disque compact (source : https://commons.wikimedia.org/wiki/File:CD_autolev_crop.jpg, crédits :Ubern00b, CC BY-SA 3.0 <http://creativecommons.org/licenses/by-sa/3.0/>, via Wikimedia)

Le Digital Versatile Disc (DVD)

Disque optique utilisé pour stocker des données audiovisuelles sous forme numérique, créé en 1995. Lui ont succédé les disques Blu-Ray et High Definition Digital Versatile Disc (HD-DVD).

Fig. 9 : DVD (source : https://commons.wikimedia.org/wiki/File:Sony_DVD%2BRW.jpg, crédits : Thigalepranav, Public domain, via Wikimedia Commons)

Si cette génération de support a largement été portée par l'industrie musicale et audiovisuelle, elle a également été utilisée pour stocker tous types d'archives. Certains disques étaient réinscriptibles (ex. DVD-RAM) – c'est-à-dire que l'on pouvait modifier les données enregistrées –, d'autres non (ex. CD-R).

1.2.1.4. La dernière génération

La dernière génération de supports de stockage repose sur la technologie de la mémoire flash, ultra rapide et à programmation électrique :

Les clés USB (Universal Serial Bus)

Support de stockage amovible inventé dans les années 2000, contenant une mémoire flash mais pas ou très peu d'éléments mécaniques ;

Fig. 9 : clé USB (source : https://commons.wikimedia.org/wiki/File:SanDisk-Cruzer-USB-4GB-ThumbDrive.jpg, crédits :Evan-Amos, Public domain, via Wikimedia Commons)

Les cartes SD (Secure Digital)

Carte mémoire amovible de stockage de données créée en 2000, principalement pour les appareils de photographie, les caméscopes, les systèmes de navigation GPS (Global Positioning System), les consoles de jeux vidéo, les téléphones mobiles et les systèmes embarqués ;

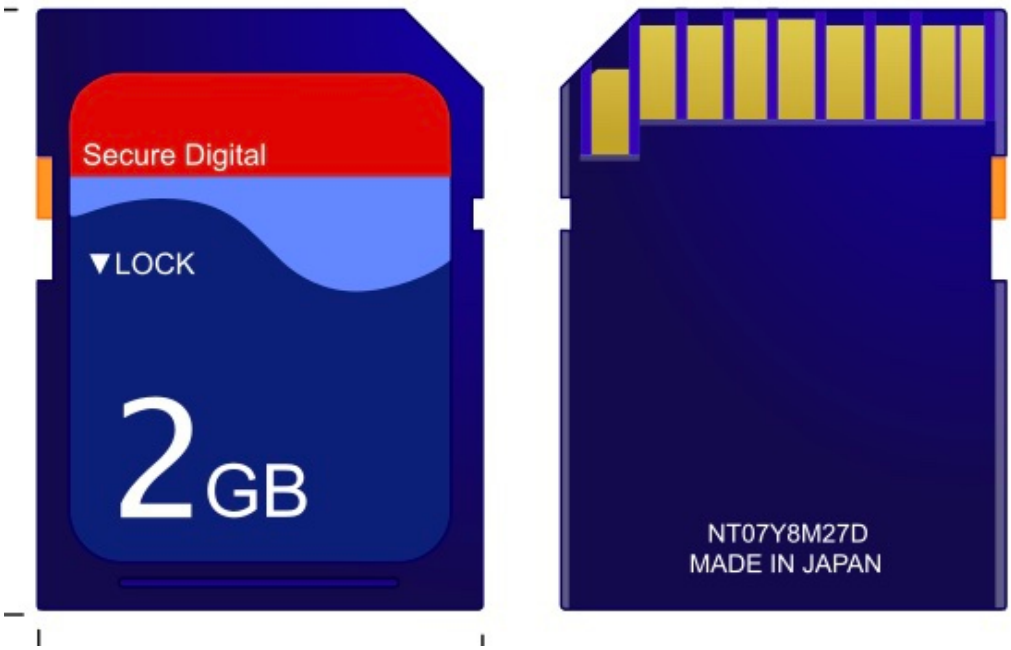

Fig. 10 : carte SD

Fig. 11 : carte SD (source : 毛抜き Derivative work: Tkgd2007, CC BY-SA 3.0 ; https://commons.wikimedia.org/wiki/File:SD_Cards.svg?lang=fr)

Les disques SSD (Solid State Drive)

Lecteur de stockage sans aucune pièce mécanique en mouvement, ce qui leur confère une résistance aux chocs supérieurs aux disques durs.

Fig. 12 : disque SSD (source : https://commons.wikimedia.org/wiki/File:SSD_120_GB_2.5_inch_with_mounting_frame_for_3.5-inch_h, crédits :smial (FAL or GFDL 1.2 <http://www.gnu.org/licenses/old-licenses/fdl-1.2.html>), via Wikime)

1.2.1.5. Durée de vie des supports de stockage

Tous les supports ont des durées de vie limitées dans le temps (figure 13).

Type de support	Durée de vie
CD non inscriptibles (gravés en usine)	50 à 100 ans
CD-R	5 à 100 ans suivant le type de colorant et la couche métallique utilisés
CD-RW	20 à 50 ans
DVD en lecture seule	10 à 20 ans
DVD-R	10 à 50 ans
DVD-RW	5 à 20 ans
Bande analogique	10 à 30 ans
Bande audionumérique	5 à 10 ans
Disquettes	5 à 15 ans
Disques durs	2 à 5 ans

Figure 13 : Durée de vie de différents supports de stockage (source : Christopher Dicks, [en ligne] disponible sur https://www.canada.ca/fr/institut-conservation/services/soin-objets/supports-electroniques/disques-durs-disquettes-faq.html

Les supports les plus récents permettent le stockage d'une grande quantité d'informations à un coût raisonnable. Leur durée de vie est cependant également limitée (4 à 8 ans) et ils doivent donc être remplacés très régulièrement.

1.2.2. Les méthodes de stockage

Plusieurs critères déterminent la méthode de stockage utilisée par les particuliers et les organisations :

- **La fréquence d'utilisation** des documents d'archives sur support numérique ;
- **La capacité totale de stockage nécessaire** pour les documents d'archives sur support numérique ;
- **La criticité** des documents d'archives sur support numérique ;
- **La vitesse d'accès** aux documents d'archives sur support numérique.

On distingue particulièrement :

- **le stockage chaud ou à chaud**, réalisé sur des supports de stockage immédiatement accessibles et rapides (ex. disques SSD) ;
- **le stockage froid ou à froid**, réalisé sur des supports de stockage moins rapidement accessibles et plus lents (ex. bandes magnétiques) et moins facilement accessibles aux pirates informatiques.

Pour des questions de sécurité, le stockage des documents d'archives sur support numérique implique de disposer de plusieurs copies de celles-ci, à des emplacements différents.

La réalisation de ces copies résulte de l'utilisation de deux méthodes souvent confondues :

- **la sauvegarde** : il s'agit, périodiquement, de **créer une copie** de tous les documents qui se trouvent **sur les environnements de production**, ou du moins de ceux qui ont fait l'objet d'une modification depuis la dernière sauvegarde. Cela permet, en cas de corruption de fichiers, de défaillance du système ou de panne, de restaurer les documents concernés. La sauvegarde peut se faire sur une grande diversité de supports de stockage. La réalisation de l'opération peut prendre plusieurs heures et est souvent initiée à des horaires de plus faible activité, afin de réduire son impact sur le fonctionnement des systèmes en production ;
- **la réplication** : il s'agit de **copier automatiquement les documents à différents emplacements**, ce qui garantit un accès à distance depuis le site de secours en cas de panne ou d'urgence. La réplication peut être faite en temps réel (synchrone) ou programmée (asynchrone). En cas de sinistre, le basculement vers le site de secours est quasi instantané.

1.2.3. Informatique en nuage ou infonuagique (cloud computing)

Depuis désormais une dizaine d'années, le cloud computing, informatique en nuage ou infonuagique, a connu un rapide développement. Ce dernier est souvent réduit à la question du stockage, ce qui constitue un défaut de compréhension.

Le cloud computing désigne en réalité l'utilisation de la mémoire et des capacités de calcul fournies par des ordinateurs et des serveurs informatiques répartis dans le monde entier et mis en réseau. Les applications et les informations ne sont plus stockées sur un ordinateur ou un serveur donné mais dans un nuage composé de serveurs distants interconnectés. Flexible, offrant une grande élasticité avec une forte capacité à s'adapter à une demande croissante, cette architecture facilite réplication des données, accès à distance et audit.

Remarque

Trois niveaux de service sont offerts par les architectures de cloud computing :

- Infrastructure as a service (Iaas) : il consiste à offrir un accès à un matériel informatique sur lequel le client peut installer système d'exploitation et applications informatiques. Le client est ainsi dispensé de l'achat et de la maintenance du matériel ;
- Platform as a service (Paas) : il consiste à fournir, en plus du matériel et du moyen de le faire fonctionner, le système d'exploitation. Le client garde le contrôle des autres applications qu'il installe sur la plateforme ;
- Software as a service (Saas) : il consiste à mettre également à disposition du client des applications qui sont accessibles au moyen d'un navigateur. Le fournisseur de service opère les mises à jour et garantit la disponibilité du service.

L'architecture de cloud computing peut être :

- publique : le service est hébergé à l'extérieur de l'organisation, mutualisé avec d'autres clients et accessible depuis Internet. Les documents d'archives sont donc conservées en dehors de l'organisation ;
- privée : le service est privatif, hébergé soit à l'intérieur, soit à l'extérieur de l'organisation. Les documents d'archives peuvent être conservés soit dans l'enceinte de l'organisation, soit en dehors de l'organisation ;
- hybride : le service résulte d'une combinaison des deux précédents.

Attention

Si le recours à une architecture de cloud computing est de plus en plus fréquente, y compris pour des services courants (bureautique, messagerie, stockage de fichiers, visioconférence), il n'est pas sans présenter plusieurs risques :

- respect de la confidentialité des documents d'archives et de la souveraineté numérique. La réglementation européenne impose par exemple que certaines catégories de données soient stockées sur des infrastructures installées dans des pays de l'Union européenne, et interdit leur stockage à l'extérieur de celle-ci ;
- capacité à récupérer les documents au terme de la relation contractuelle, dans une forme qui soit ré-exploitable (structuration, métadonnées, etc.) ;
- confiance dans la capacité du prestataire à garantir l'intégrité des documents.

1.2.4. Le chiffrement

Afin de protéger la confidentialité et l'intégrité des documents d'archives enregistrés sur un espace de stockage informatique ou transmises par quelque technologie que ce soit, il est possible d'utiliser un procédé de *chiffrement*. Cela consiste à appliquer un algorithme à ces dernières afin de rendre leur signification intelligible à toute personne ne disposant pas des moyens techniques de déchiffrer, au moyen d'une clé.

Deux systèmes de chiffrement existent :

- le chiffrement symétrique, quand la même clé est utilisée pour chiffrer et déchiffrer ;
- le chiffrement asymétrique, quand des clés différentes sont utilisées : une clé publique pour chiffrer et une clé privée pour déchiffrer. La clé privée est impossible à déduire de la clé publique, ce qui sécurise le système.

Attention

Si le chiffrement constitue un moyen pratique de sécuriser la transmission et le stockage des documents, il peut également constituer une menace pour ceux-ci à long terme. À défaut d'une gestion dans le temps des clés de chiffrement et de déchiffrement, le contenu des documents peut devenir totalement inintelligible.

Prendre en charge un document d'archives sans sa clé de déchiffrement ne sert donc à rien, car il ne sera jamais compréhensible.

1.3. Comment les documents d'archives sur support numérique sont-ils transférés ?

Introduction

Le transfert de fichiers est une opération informatique qui consiste à acheminer des documents d'archives d'un environnement à un autre (d'un ordinateur à un autre par exemple). Elle permet de rendre le fichier transféré disponible sur la machine distante, sans nécessairement recourir à un support physique (CD, DVD, clé USB, etc.).

Le transfert de fichiers se décompose en deux éléments distincts :

- un mode physique de communication :
- un protocole de communication.

1.3.1. Le mode physique de communication

Des documents d'archives sur support numérique peuvent être acheminés d'un environnement informatique à un autre en utilisant plusieurs modes physiques de communication :

- le transfert par supports physiques (cartes, bandes, cassettes, disquettes, clés, disque dur externe). Dans ce cas, les documents sont écrits sur le support concerné à partir de l'environnement de départ et écrits sur l'environnement d'arrivée à partir du support d'acheminement. Le recours à cette méthode suppose que les deux environnements, celui de départ et celui d'arrivée, disposent du moyen de se connecter au support et d'interagir avec lui (existence d'un pilote) ;
- le transfert filaire. Dans ce cas, un câble connecte physiquement les deux environnements ;
- le transfert par ondes électromagnétiques. **Une onde électromagnétique peut se déplacer dans un milieu de propagation comme le vide ou l'air,** avec une vitesse avoisinant celle de la lumière. Les ondes électromagnétiques peuvent transporter de l'énergie, mais aussi de l'information. Elles sont donc très utilisées dans le domaine de la communication. Les ondes électromagnétiques se caractérisent par leur fréquence – le nombre d'oscillations par seconde, exprimé en Hertz –, et par leur longueur – la distance qui sépare deux oscillations de l'onde, inversement proportionnelle à la fréquence. Aujourd'hui, le transfert d'archives sur support numérique se fait principalement en utilisant cette méthode, que l'on utilise un réseau mobile téléphonique (du 2G au 5G), le Wi-Fi ou le Bluetooth.

1.3.2. Les protocoles de communication

L'acheminement d'archives sur support numérique d'un environnement informatique à un autre peut nécessiter le recours à un protocole de communication, c'est-à-dire un ensemble de règles qui définissent la manière dont les informations sont transmises entre les environnements.

Il en existe de plusieurs types :

- le plus ancien est le protocole de communication entre un ordinateur et un périphérique (clé USB par exemple), qui nécessite la mise en œuvre d'un pilote ;
- un transfert via un réseau informatique se fait à l'aide d'un protocole réseau. Il en existe plusieurs :
    - le protocole de réseau standard *File Transfer Protocol* (FTP), établi en 1985, qui permet à un appareil (ordinateur, serveur) de se connecter à un autre appareil faisant office de serveur de fichiers et de transférer des fichiers entre les deux appareils, au moyen d'une identification ;
    - le protocole de contrôle de transmission *Transmission Control Protocol/Internet Protocol* (TCP/IP)
    - le protocole *Hypertext Transfer Protocol* (HTTP) qui définit le format des messages et des échanges qui sont réalisés sur le World Wide Web
    - le protocole *Secure Copy* (SCP)
    - le protocole *Secure File Transfer Protocol* (SFTP)

Ces protocoles peuvent être accélérés et sécurisés via l'utilisation d'une couche de compression et de chiffrement. On parle alors de protocole sécurisé, comme le SFTP.

1.3.3. Les problématiques à prendre en compte

En matière de transfert d'archives sur support numérique, deux questions essentielles sont à prendre en compte :

- la sécurité de la transmission, pour éviter toute interception non souhaitée du contenu. Celle-ci peut impliquer le recours à une technique dite de chiffrement ;
- le débit de la transmission au regard de la volumétrie des documents à transférer et du risque d'interruption de l'opération induit par des dysfonctionnements électriques ou des déconnexions provoquées par des épuisements de délais.

2. Les spécificités des documents d'archives sur support numérique

II

1. Introduction

Dès lors qu'ont été expliquées les conditions de création, de stockage et de transmission des documents d'archives sur support numérique, il convient de présenter leur modélisation et leurs caractéristiques et de s'interroger sur ce qui les rend dignes de confiance.

2.1. La modélisation des documents d'archives sur support numérique

Description du modèle OAIS

Le modèle de référence OAIS propose une modélisation des documents d'archives (l'objet d'information ou *Information Object* au sens du modèle), physique, analogique comme numérique, qui est conçue comme une combinaison d'objets et d'informations sur les données (donc des métadonnées) qui permettent de rendre celles-ci intelligibles par un individu.

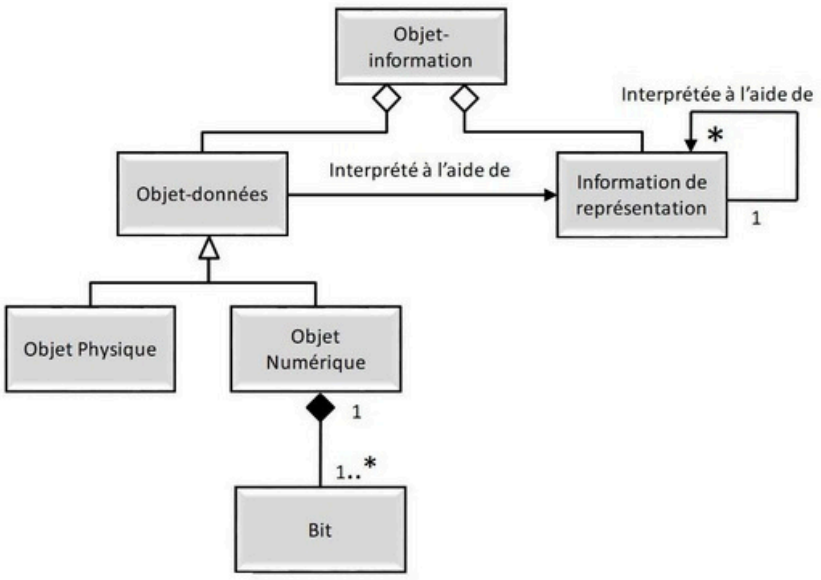

Fig. 14 : modélisation des objets d'information au sens de la norme OAIS (source: Modèle de référence pour un Système ouvert d'archivage d'information (OAIS), livre magenta, pratique recommandée par le comité consultatif pour les systèmes de données spatiales, 2e édition, 2017, , http s://public.ccsds.org/Pubs/650x0m2%28F%29.pdf[2], p. 4-24)

NB : l'astérisque correspond au caractère répétable d'une entité ou d'une relation. Par exemple, une information de représentation peut avoir elle-même besoin d'une information de représentation.

2. https://public.ccsds.org/Pubs/650x0m2(F).pdf

Qu'est-ce qu'un objet d'information dans le modèle OAIS ?

Le modèle de référence OAIS explique qu'un objet d'information est constitué des composantes suivantes :

- Des données, écrites sur un support, physique, analogique ou numérique, qui constituent l'objet de données (*data object*), donc le cœur de l'information à préserver. Pour les données numériques, il s'agit du train binaire correspondant à l'encodage technique ;
- Les informations dites de représentation (Representation Information), qui permettent de rendre les données intelligibles et les transforment en objets d'information. Ces informations peuvent prendre plusieurs formes :
    - Des informations de structure (*Structure Representation Information*) : organisation des fichiers composant un site web ou un enregistrement audiovisuel encodé sous forme d'un format de fichiers conteneur ; spécifications des formats de fichiers, permettant de comprendre comment les données sont encodées techniquement ; description de la structure d'un tableur ou des tables d'une base de données ;
    - Des informations sémantiques (*Semantic Representation Information*) : il peut s'agir des outils permettant de comprendre la langue utilisée pour coder l'information (dictionnaire, grammaire), mais aussi de toute la documentation présentant le codage métier des informations dans un registre (ex. dictionnaire des codes utilisés dans une base de données) ;
    - Toute autre information utile (*Other Representation Information*).

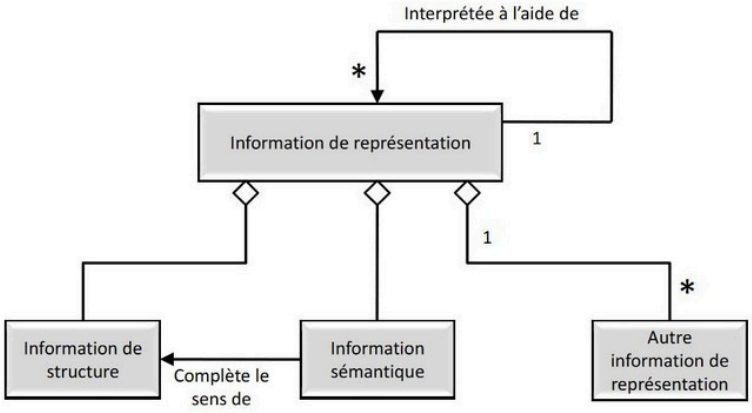

Fig. 15 : modélisation des informations de représentation au sens de la norme OAIS (source: Modèle de référence pour un Système ouvert d'archivage d'information (OAIS), livre magenta, pratique recommandée par le comité consultatif pour les systèmes de données spatiales, 2e édition, 2017, https://public.ccsds.org/Pubs/650x0m2%28F%29.pdf[3], p. 4-27)

2.1.3 Une modélisation de l'information applicable aux documents d'archives sur support physique

Cette modélisation de l'information s'applique très bien aux documents d'archives sur support physique, par exemple à un instrument de ratification d'un traité international :

- information de structure : relation entre l'acte de ratification d'un traité et le sceau qui valide l'acte ;
- information sémantique : langue utilisée pour rédiger l'acte (ici le français).

3. https://public.ccsds.org/Pubs/650x0m2(F).pdf

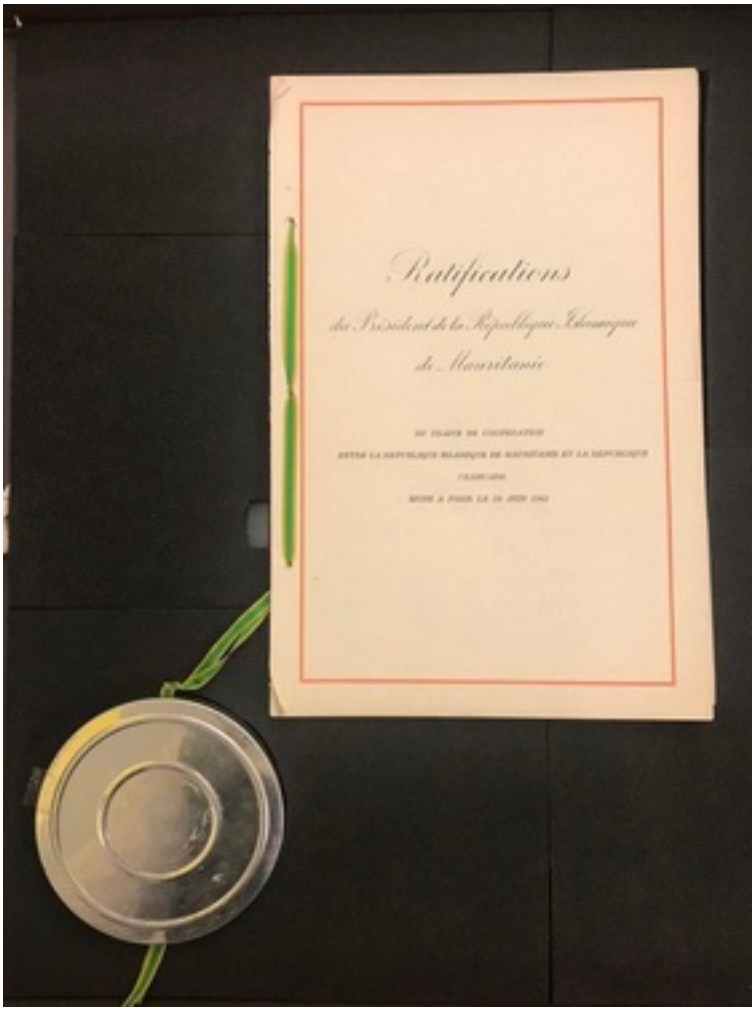

Fig. 16 : Instrument de ratification par la Mauritanie du traité de coopération avec la France signé le 19 juin 1961 (source : ministère des Affaires étrangères - France)

2.2. Les caractéristiques des documents d'archives sur support numérique

Introduction

Les objets d'information au sens du modèle de référence OAIS présentent un certain nombre de caractéristiques qu'il est important d'analyser et de comprendre en vue d'assurer leur préservation numérique.

2.2.1. Propriétés porteuses de sens (Significant properties)

Une des caractéristiques essentielles des objets d'information est constituée par ce qui a été désigné par l'expression de propriétés porteuses de sens (*significant properties*).

Historique de la notion

Ce concept a été défini à la fin des années 1990 dans le cadre de plusieurs projets de recherche et a fait l'objet d'une formalisation en 2008 dans le cadre du rapport-cadre rédigé par les membres du projet InSPECT (*Investigating the Significant Properties of Electronic Content Over Time* – accessible à l'adresse suivante : https://significantproperties.kdl.kcl.ac.uk/inspect-finalreport.pdf).

Définition

La définition donnée par ce rapport (p. 3) des propriétés porteuses de sens est la suivante [NB : la traduction est de l'auteur] : « Les caractéristiques des objets numériques qui doivent être préservées au fil du temps afin d'assurer l'accessibilité, l'utilisation et la signification continues de ceux-ci, ainsi que leur capacité à être acceptés comme preuve de ce qu'ils sont censés enregistrer ».

Utiliser des propriétés porteuses de sens

Les *propriétés porteuses de sens* [p.37] sont donc les caractéristiques les plus importantes d'un objet d'information à préserver dans le temps, au-delà des changements technologiques subis par les objets de données. Elles doivent faire l'objet d'un recensement avant toute opération de préservation et leur maintien après l'opération doit faire l'objet de vérifications soigneuses.

Méthode

Le rapport-cadre du groupe InSPECT fournit une méthodologie formalisée et documentée permettant d'étudier, pour une catégorie d'objets d'information, les propriétés porteuses de sens. Cette méthodologie repose successivement sur :

- une analyse des objets eux-mêmes : sélection du type d'objet à analyser (ex. enregistrement sonore) ; recensement des propriétés au moyen des spécifications techniques ou normes disponibles ; identification de la finalité des différentes propriétés – contenu (ex. nombre de mots), contexte (ex. auteur), rendu (ex. taille de caractère), structure (ex. existence de pièces jointes), comportement (ex. existence d'animations) ; détermination des différents usages des objets et catégorisation des comportements attendus (ex. recréation de l'apparence visuelle) ; dépendance entre comportements attendus et propriétés ;
- une analyse des besoins des utilisateurs de ces objets : identification des différentes catégories d'utilisateurs (producteur, chercheur, archiviste, etc.) ; identification des objets types manipulés par chaque catégorie d'utilisateur ; détermination des différents usages des objets et catégorisation des comportements actuels ; confrontation entre comportements attendus déterminés lors de la première phase et comportements actuels pour déterminer la liste des propriétés indispensables ; identification des propriétés non indispensables, dont la perte est acceptable lors des opérations de préservation.

Retour d'expérience : les Archives fédérales américaines

Des analyses ont d'ores et déjà été effectuées pour les images fixes matricielles, les messages électroniques, les enregistrements sonores, les textes structurés et les tableurs.

Les Archives fédérales américaines (*National Archives and Records Administration*) ont de leur côté procédé, pour chaque catégorie de format de fichiers, à ce recensement des propriétés porteuses de sens (https://github.com/usnationalarchives/digital-preservation).

Exemple

À titre d'exemple, pour les images fixes (https://github.com/usnationalarchives/digital-preservation/blob/master/Still_Image_Formats/NARA_PreservationActionPlan_DigitalStillImage_20220714.pdf), il s'agit des propriétés suivantes :

- taille de l'image ;
- espace colorimétrique ;
- profondeur de bits ;
- orientation ;

- présentation ;
- compression ;
- résolution ;
- restitution visuelle ;
- identifiant assigné lors de la capture ;
- date de capture ;
- type de caméra ;
- titre ;
- copyright ;
- auteur
- etc.

2.2.2. Empreinte

Il est possible d'associer à chaque fichier une *empreinte* [p.36], c'est-à-dire une chaîne de caractères générée au moyen d'une analyse cryptographique mettant en œuvre un algorithme dit de hachage (ex. MD5, SHA-1, SHA-253, SHA-512). Deux fichiers strictement identiques auront la même empreinte, dans la mesure où l'analyse cryptographique a été effectuée avec le même algorithme. Par ailleurs, plus l'algorithme est puissant, plus l'empreinte générée pour le fichier est complexe et plus il est donc difficile de modifier délibérément le contenu du fichier sans que cela soit détecté.

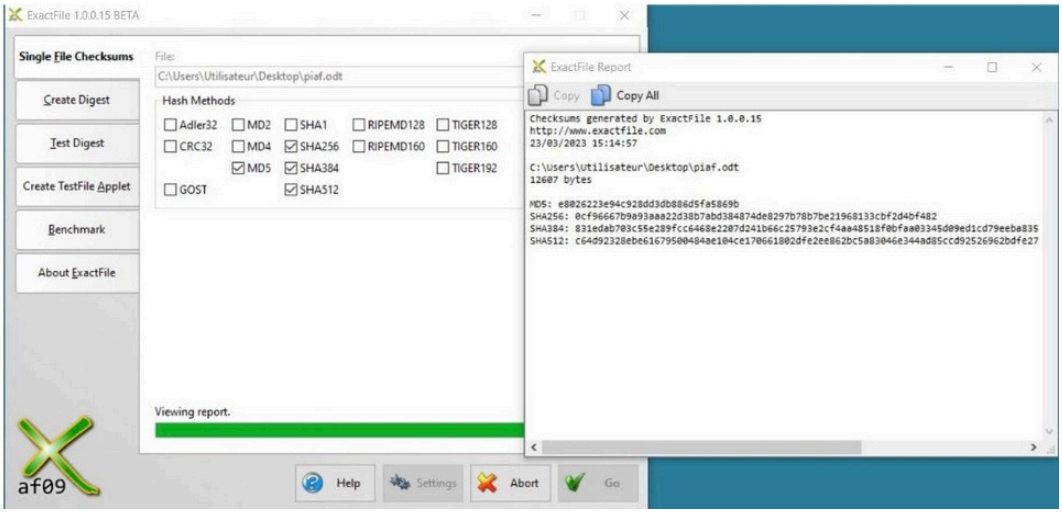

Fig. 17 : Résultats du calcul d'empreintes à l'aide de différents algorithmes proposés par le logiciel ExactFile

Fondamental

Toute modification d'un seul des éléments du fichier – parfois même la simple ouverture de celui-ci (ex. cas d'un fichier bureautique comportant une date se mettant à jour automatiquement à la simple ouverture du fichier) – se répercute sur son empreinte, ce qui permet d'identifier facilement toute perte d'intégrité subie par celui-ci. Si un changement d'empreinte permet d'identifier qu'un fichier a été modifié, il ne permet cependant pas de détecter à quel endroit cette modification a eu lieu (d'autres méthodes pourront être employées, comme une comparaison au moyen d'un outil appelé éditeur hexadécimal, voir section 2 de ce module).

Les empreintes sont utilisées dans deux contextes :

- lors de la transmission d'un fichier, pour garantir que le fichier reçu par le destinataire est bien le même que celui envoyé par l'expéditeur, par comparaison de l'empreinte du fichier avant envoi et celle calculée ;
- lors du stockage du fichier, pour garantir que le fichier conservé correspond bien au fichier qui a été écrit sur le support de stockage et qu'en cas de duplication, les deux copies d'un fichier sont strictement identiques.

2.3. Caractéristiques des documents d'archives dignes de confiance

Il est un fait que les documents d'archives sur support numérique sont très faciles à copier de manière exacte sur différents supports mais que, en même temps, l'évolution très rapide des technologies sous-jacentes (stockage, formats de fichier) invite à réinterroger les concepts d'originalité, d'authenticité et d'intégrité sur lesquels était fondé le caractère probant des documents d'archives sur support physique ou analogique.

Durant plusieurs années de travail, le groupe international InterPARES piloté par l'Université de Colombie-Britannique (Canada) s'est efforcé de définir ce qu'était un document d'archives sur support numérique digne de confiance.

Cette qualité s'évalue, selon lui, au regard de trois critères :

- sa *fiabilité* p.36 (authenticité diplomatique) : le document d'archives est reconnu fiable s'il traduit fidèlement les faits auxquels il se rapporte. On peut donc lui accorder foi en tant qu'énoncé des faits. Cette qualité peut se déduire de deux choses : la complétude de la forme du document et le degré de contrôle exercé sur la procédure au cours de laquelle il a été créé. La fiabilité est du ressort de la personne ou de l'organisation qui crée le document ;
- son exactitude (authenticité historique) : le document d'archives est reconnu exact s'il contient des données correctes, précises et justes. L'exactitude est présumée lorsque le document est produit et utilisé au cours du processus métier qui aboutit à sa création. L'exactitude du document doit cependant faire l'objet d'une vérification systématique en cas de changement de support et de format de fichiers (donc lors d'opérations de préservation numérique). Cette vérification relève de la responsabilité de la personne ou de l'organisation qui gère le document, même si seule la personne ou l'organisation qui ont produit le document sont responsables de son exactitude initiale (un faux document transmis à un service d'archives reste un faux document ensuite) ;
- son *authenticité* p.36 (authenticité juridique) : le document d'archives est reconnu authentique s'il est bien ce qu'il prétend être et qu'il n'a été ni corrompu ni altéré. L'authenticité d'un document peut être maintenue et vérifiée en préservant son identité (ensemble des attributs d'un document qui le caractérisent comme unique, et qui le distinguent des autres documents) et son *intégrité* p.37 (absence d'altération du message que le document est censé porter).

Rappel

L'original d'un document d'archives sur support numérique n'est en réalité que la première version (celle du moment de sa création) et la plus parfaite (celle qui est apte à produire les effets que son auteur attend de lui) de celui-ci . Par la suite, il n'est possible d'utiliser une autre version de celui-ci, que sous réserve que celle-ci ait été produite dans des conditions qui garantissent le caractère digne de confiance du document (lors d'une extraction de l'environnement d'origine avec migration de format de fichiers, par exemple).

Attention

L'archiviste doit être en mesure de garantir le maintien de l'exactitude et de l'authenticité (identité et intégrité) à partir du moment où il prend en charge le document d'archives, quelles que soient les opérations de préservation qu’il met en œuvre pour leur conserver un accès continu.

3. Quels sont les risques associés à la préservation des documents d'archives sur support numérique ?

Introduction

Les caractéristiques des documents d'archives sur support numérique font que leur préservation à long terme est confrontée à plusieurs risques qu'il convient soigneusement d'évaluer.

Ces risques concernent à la fois le support sur lequel ces documents d'archives sont stockés, les objets de données et les informations de représentation au sens du modèle de référence OAIS mais aussi et surtout la manière dont les documents d'archives sont gérés et utilisés par l'ensemble des acteurs au sens de la norme OAIS (service producteur, service informatique, service d'archives).

L'évaluation et la maîtrise des risques sont donc essentielles pour garantir que les documents d'archives sur support numérique restent et demeurent dignes de confiance et utilisables tout au long de leur durée de conservation.

3.1 Risques sur les supports

Le premier risque auquel sont confrontés les documents d'archives sur support numérique est la conséquence de la fragilité des supports sur lesquels ces documents sont stockés.

Comme nous l'avons vu, les supports informatiques :

- peuvent être perdus ou détruits par suite d'une malveillance humaine (destruction volontaire ou simple oubli) ou d'un événement externe (incendie, tremblement de terre, inondation, panne électrique, etc.), comme les supports physiques ou analogiques ;
- sont sujets à une obsolescence technologique rapide, dans la mesure où ils peuvent cesser d'être fabriqués et où les appareils de dernière génération ne sont pas nécessairement équipés pour lire des supports anciens (les ordinateurs aujourd'hui ne sont plus équipés de lecteurs de disquettes ou de CD, par exemple). On peut estimer à 5 ans maximum la durée pendant laquelle un support reste utilisable ;
- présentent des fragilités techniques intrinsèques : cassures (ex. pour les disques en verre), rayures, démagnétisation suite à une exposition à un champ magnétique très fort, dégradations chimiques, etc. .

Fondamental

Une bonne politique et un plan de gestion des supports sont donc essentiels à la préservation à long terme des documents d'archives sur support numérique.

3.2. Risques sur les objets de données (Data Object)

Le deuxième risque auquel sont confrontés les documents d'archives sur support numérique est la conséquence de la fragilité des objets de données (Data Object) au sens du modèle de référence OAIS.

Comme nous l'avons vu, il est très facile de modifier le contenu d'un objet de données – parfois simplement en ouvrant le fichier correspondant avec un nouveau logiciel – et la modification d'un simple octet dans un fichier modifie l'intégrité binaire de celui-ci.

Une bonne politique d'audit de l'intégrité des fichiers est donc essentielle à la préservation à long terme des documents d'archives sur support numérique.

3.3. Risques sur les informations de représentation (Representation Information)

Le troisième risque auquel sont confrontés les documents d'archives sur support numérique porte sur les informations de représentation (*Representation Information*) au sens du modèle de référence OAIS. Ces informations sont en effet indispensables pour transformer les objets de données en objets d'informations et pour rendre les documents d'archives sur support numérique intelligibles et accessibles.

Ce risque peut se décomposer en plusieurs éléments :

- le risque d'obsolescence de l'encodage technique utilisé (obsolescence des formats) ;
- le risque d'indisponibilité ou de mauvaise qualité des autres informations de représentation.

3.3.1. Le risque d'obsolescence des formats

Les formats de fichiers qui encodent les documents d'archives sur support numérique sont sujets à plusieurs types de défaillances :

- ils peuvent être dépendants d'un matériel et d'un système d'exploitation donnés ;
- ils évoluent régulièrement à mesure que concepteurs et utilisateurs adoptent de nouvelles fonctionnalités. Les logiciels actuellement utilisés pour représenter des fichiers dans un format donné peuvent ne pas supporter les versions les plus anciennes de ce format, ce qui peut rendre les objets de données inaccessibles ;
- ils peuvent ne plus être maintenus du tout, quelle que soit leur nature (ouvert/fermé, propriétaire, standardisé), en raison de la défaillance de la communauté ou de l'entreprise qui les a créés et maintenus ;
- les logiciels qui permettent de représenter des objets de données enregistrées peuvent être sujets à des dysfonctionnements (bugs, logiciels malveillants).

Attention

Les risques portant sur les formats de fichiers sont donc importants, même s'il ne faut pas les exagérer outre mesure. De nombreux formats de fichiers anciens restent pris en charge par les actuelles versions des logiciels. Le moment exact où l'obsolescence d'un format de fichiers se révèle est donc difficile à déterminer. L'obsolescence est davantage relative - liée à un environnement matériel et logiciel donné – qu'absolue.

3.3.2 Les risques d'indisponibilité ou de mauvaise qualité des autres informations de représentation

Mais le risque peut également porter sur l'insuffisance ou l'absence des autres informations de représentation, hors formats des fichiers :

- la documentation du codage métier (structuration intellectuelle, codage des informations, choix des unités de mesure utilisées, etc.), indépendamment de l'encodage technique, est indispensable à la compréhension des documents d'archives sur support numérique. À défaut de création de cette documentation lors de l'élaboration des documents ou du système informatique permettant leur création, l'intelligibilité des documents sur le moyen comme le long terme ne peut pas être garantie. La perte ou la corruption de cette documentation ont les mêmes conséquences ;
- les documents d'archives sur support numérique pouvant être constitués de fichiers dépendant les uns des autres. Toute perte d'information externe (ex. disparition d'un annuaire, absence de conservation d'une nomenclature ou d'une codification) aux fichiers concernés peut rendre les documents inexploitables ;
- à défaut de connaissance du processus métier ayant abouti à l'élaboration ou à la réception des documents d'archives sur support numérique, leur intelligibilité sera tout autant mise en péril.

Fondamental

L'accessibilité et l'intelligibilité à long terme des documents d'archives sur support numérique impliquent donc de prendre en compte et de gérer ces risques directement liés à leur exploitabilité technique et humaine. Une bonne politique de gestion des informations de représentation et la mise en œuvre des opérations humaines et techniques correspondantes sont donc essentielles à la préservation à long terme des documents d'archives sur support numérique.

3.4. Risques organisationnels

Indépendamment des risques techniques, les documents d'archives sur support numérique sont sujets à tout un ensemble de risques organisationnels qui peuvent mettre en péril leur préservation à long terme.

On peut citer, entre autres :

- le manque de connaissance et de compétences des différents acteurs de la chaîne de préservation (services producteurs, services informatiques, services d'archives), ce qui peut entraîner des prises de décisions inappropriées ;
- l'absence ou la défaillance des procédures mises en œuvre par les différents acteurs. Ces dysfonctionnements peuvent porter atteinte à la confiance portée aux documents eux-mêmes ;
- le manque de documentation et de traçabilité des actions mises en œuvre, ce qui peut rendre indétectables des atteintes portées à l'intégrité des objets de données ou remettre en question l'authenticité et l'exactitude des informations correspondantes.

Fondamental

Une bonne organisation ainsi qu'une bonne sensibilisation et formation de l'ensemble des acteurs sont donc indispensables à la préservation à moyen et à long terme des documents d'archives sur support numérique. Dans le cas d'une conservation à très court terme (moins d'un an), le risque est moindre.

3.5. Risques financiers

Enfin, une mauvaise préservation des documents d'archives sur support numérique peut induire des risques financiers non négligeables qui peuvent résulter de :

- sanctions financières prononcées à défaut de présentation d'un document d'archives digne de confiance, par exemple dans le cadre d'une investigation (discovery) ou d'une procédure contentieuse ;
- coût d'opérations de préservation (migration de format, émulation, etc.) non anticipées ;
- pertes financières occasionnées par une opération de réingénierie documentaire et technique de documents d'archives dont la préservation n'a pas été prise en compte de manière suffisamment précoce ;
- etc.

4. Conclusion

On retrouve avec les documents d'archives sur support numérique un certain nombre de caractères communs avec les documents d'archives sonores et audiovisuelles sur support analogique. Dans les deux cas, l'obsolescence technique des supports et des appareils de lecture rend nécessaire une préservation dynamique des documents d'archives concernés qui, à défaut, risquent de devenir inexploitables et inaccessibles sur le moyen comme le long terme. L'existence d'un encodage technique et d'un codage métier rend cependant les documents d'archives sur support numérique encore plus fragiles.

Les problèmes de préservation à long terme des documents d'archives sur support numérique doivent être pris en compte le plus tôt possible. C'est souvent dès la production et dès la mise en œuvre de la première opération d'archivage que se joue la préservation à long terme des documents d'archives sur support numérique.

Glossaire

Agrégat

Toute accumulation d'entités document d'activité à un niveau supérieur à l'objet document d'activité (d'après ISO 16175-2:2011)

Authenticité

un document authentique est un document dont on peut prouver

a) qu'il est bien ce qu'il prétend être,

b) qu'il a été effectivement produit ou reçu par la personne qui prétend l'avoir produit ou reçu, et

c) qu'il a été produit ou reçu au moment où il prétend l'avoir été.

(ISO 15489 « records management », 2016)

Chiffrement

Opération par laquelle est substitué, à un texte en clair, un texte inintelligible, inexploitable pour quiconque ne possède pas la clé permettant de le ramener à sa forme initiale (Office québécois de la langue française).

Codage métier

Liste contrôlée de toutes les valeurs acceptables en langage naturel et/ou en tant que chaîne de caractères encodée conçue pour le traitement par machine (d'après ISO 23081-1:2017)

Conteneur

Un conteneur (wrapper ou container en anglais) est est une enveloppe virtuelle utilisée pour stocker des fichiers, services, librairies etc. sous une forme organisée qui suit des règles d'accès spécifiques.

Empreinte

Empreinte (empreinte numérique ou condensat ou hash) : Résultat d'une fonction de hachage appliquée sur une chaîne de caractères de longueur quelconque visant à réduire celle-ci en une donnée de longueur fixe représentative de cette chaîne de caractères. L'empreinte est l'un des éléments permettant de vérifier l'intégrité d'un document, d'un flux, d'un lot, d'une transmission,... (comparaison d'empreintes).

Encodage technique

Voir codage technique

Règles ou conventions qui déterminent un format et permettent de représenter un texte ou un medium afin qu'il soit lisible.

Fiabilité

Pour qu'un document soit fiable, « son contenu doit pouvoir être considéré comme la représentation complète et exacte des activités ou des opérations qu'il décrit, de façon que sa consultation ultérieure permettre de comprendre comment l'opération qu'il décrit a été réalisée ». (ISO15489, 2016).

Format de données, ou format de fichier ou format de représentation de l'information :

le format de données peut être défini par l'ensemble des règles et algorithmes permettant d'organiser l'information dans un objet numérique.

Par exemple, le format de données permettra de :

* spécifier le codage des couleurs des pixels d'une image, définir un algorithme de compression des données et l'organisation de ces données dans un fichier (formats PNG, TIFF...),

* spécifier l'organisation et la structuration d'informations textuelles à partir de l'encodage élémentaire des caractères (formats SGML, XML) ;

en réalité, SGML et XML sont en premier lieu des langages comportant un ensemble de règles, une syntaxe, des mots clés permettant de constituer des documents structurés ; lorsqu'un document a été structuré par le langage XML, on connaît en pratique l'ensemble des règles d'organisation de l'information au sein de ce document ; à ce titre, XML (comme SGML) peut donc être considéré comme un format,

* définir comment les quatre informations élémentaires que sont la mantisse (nombre entier positif), l'exposant (nombre entier positif), le signe de l'exposant et le signe de la mantisse (caractères + et -) sont organisées pour représenter un nombre réel sous forme numérique (cf. standard ANSI/IEEE 754-1985).

Infonuagique

Le cloud ou l'informatique en nuage en français, est une technologie qui permet de mettre sur des serveurs localisés à distance des données de stockage ou des logiciels qui sont habituellement stockés sur l'ordinateur d'un usager, voire sur des serveurs installés en réseau local au sein d'une entreprise (PIAF, Glossaire).

Intégrité

L'intégrité d'un document renvoie au caractère complet et non altéré de son état, (ISO 15489 « records management »). Le document n'a subi aucune modification non tracée.

Logiciel

Un logiciel est un ensemble des programmes constituant une unité destinée à effectuer un traitement particulier sur un ordinateur.

Propriétés porteuses de sens (Significant properties)

Les propriétés porteuses de sens sont les caractéristiques des objets numériques qui doivent être préservées au fil du temps afin d'assurer l'accessibilité, l'utilisation et la signification continues de ceux-ci, ainsi que leur capacité à être acceptés comme preuve de ce qu'ils sont censés enregistrer (NARA, traduction Édouard Vasseur, module 7C)