Module 7C - Section 2 : Établir un diagnostic

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

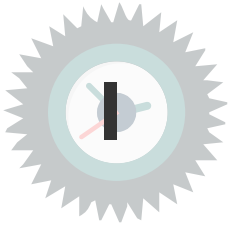
Pour bien préserver les documents d'archives sur support numérique, encore faut-il établir un diagnostic précis de la situation actuelle des ensembles de documents.

Comme pour les documents d'archives sur support physique ou analogique, cette action de diagnostic comporte deux aspects :

- connaître l'environnement de préservation dans son état actuel, pour identifier ses points forts et ses points faibles ;
- connaître la nature des fonds conservés, leur état de préservation et les risques associés pour leur accessibilité et leur intelligibilité sur le long terme.

Depuis quelques années, des outils tant méthodologiques que techniques ont été conçus, réalisés et mis à disposition des institutions et des archivistes pour réaliser ces opérations et, par conséquent, identifier quelles actions sont prioritaires à mettre en œuvre en matière de préservation numérique.

1. Connaître l'environnement de préservation et identifier ses points forts et ses points faibles



Introduction

La première étape consiste à connaître de manière globale l'environnement de préservation mis en œuvre par le service d'archives pour évaluer ses points forts et ses points faibles.

Ce diagnostic doit être global et porter sur :

- le contexte d'intervention dans lequel le service d'archives agit ;
- les procédures existantes pour la *collecte* [p.21], la prise en charge, la *description* [p.21], la *conservation* [p.21] et l'exploitation des documents d'archives sur support numérique ;
- l'existence ou non d'une stratégie de préservation ;
- les moyens financiers et humains à disposition ;
- les moyens techniques (matériel, logiciel) à disposition pour mettre en œuvre l'ensemble des procédures associées à la gestion des documents d'archives sur support numérique.
- la criticité
- la volumétrie (la volumétrie concerne à la fois le nombre de fichiers et/ou leur poids, ce qui a des conséquences sur les capacités de stockage, sur les modalités de transfert des données et sur la rapidité d'exécution des programmes).

Pour réaliser ce diagnostic, l'archiviste dispose désormais de deux outils méthodologiques faciles à prendre en main, proposés pour l'un par la *« National Digital Stewardship Alliance »*, pour l'autre par la *« Digital Preservation Coalition. »*

1.1. Les Niveaux de préservation numérique (NDSA Levels of Preservation) de la National Digital Stewardship Alliance (NDSA)

Ce premier outil méthodologique a été élaboré en 2013 puis révisé en 2018 par la *« National Digital Stewardship Alliance »* (NDSA), une association américaine présentée dans la section 6 de ce module.

Les *« Niveaux de préservation numérique (NDSA Levels of Preservation) »* constituent un guide permettant à tout service d'archives – quels que soient sa taille et son niveau de ressources – d'évaluer sa maturité en matière de préservation numérique.

1.1.1. Le tableau des Niveaux de préservation numérique

Cet outil se présente sous la forme d'un tableau comprenant 4 colonnes et 5 lignes :

- les colonnes définissent des niveaux de maturité, en fonction des réponses apportées aux questions posées dans les différentes lignes :
  - niveau 1 : protéger ses données ;
  - niveau 2 : connaître ses données ;
  - niveau 3 : surveiller ses données ;
  - niveau 4 : réparer ses données ;
- les lignes correspondent à des domaines fonctionnels qu'il convient d'analyser pour évaluer le niveau de maturité de son organisation :
  - le stockage ;
  - l'*intégrité* [p.22] des fichiers ;
  - la sécurité des informations ;
  - les *métadonnées* [p.22] ;
  - les formats de fichiers.

Remarque

L'évaluation peut être globale et porter sur l'ensemble de l'environnement, ou partielle et ne porter que sur tout ou partie des documents d'archives sur support numérique pris en charge.

NDSA

NIVEAUX DE PRESERVATION NUMERIQUE

Domaine fonctionnel	Niveaux			
	Niveau 1 (connaître vos contenus)	Niveau 2 (protéger vos contenus)	Niveau 3 (surveiller vos contenus)	Niveau 4 (pérenniser vos contenus)
Stockage	Posséder deux copies complètes dans des lieux distincts. Documenter tous les supports de stockage où les contenus sont stockés. Utiliser des supports de stockage stables.	Posséder trois copies complètes avec au moins une copie à un emplacement géographique distinct. Documenter le stockage et les supports de stockage en indiquant les ressources et dépendances nécessaires à leur fonctionnement.	Posséder au moins une copie à un emplacement géographique présentant un type de menace différent de ceux des autres emplacements. Posséder au moins une copie sur un support de stockage différent. Surveiller l'obsolescence du stockage et des supports.	Posséder au moins trois copies dans des emplacements géographiques présentant des types de menaces différents. Augmenter la variété des supports de stockage pour éviter les points de défaillance uniques. Avoir un plan et mener des actions pour remédier à l'obsolescence des supports de stockage, des logiciels et du matériel informatique.
Intégrité	Vérifier l'information d'intégrité si celle-ci a été fournie avec les contenus. Générer une information d'intégrité si aucune information n'est disponible. Contrôler la présence de virus. Le cas échéant, mettre les contenus en quarantaine.	Vérifier l'information d'intégrité lors de la migration ou de la copie des contenus. Utiliser des bloqueurs d'écriture lors des travaux sur les supports originaux. Sauvegarder l'information d'intégrité et stocker la copie dans un emplacement distinct de celui des contenus.	Vérifier l'information d'intégrité à intervalles réguliers. Documenter les processus et les résultats des vérifications de l'information d'intégrité. Mener des audits d'intégrité à la demande.	Vérifier l'information d'intégrité à la suite d'événements ou d'activités spécifiques. Remplacer ou réparer les contenus corrompus le cas échéant.
Contrôle	Déterminer les agents humains et logiciels autorisés à lire, écrire, mettre à jour et supprimer les contenus.	Documenter les droits de lecture, d'écriture, de mise à jour et de suppression des agents humains et logiciels.	Identifier les agents humains et logiciels qui mènent des actions sur les contenus et journaliser ces actions.	Examiner périodiquement les journaux des opérations et des accès.
Métadonnées	Créer un inventaire des contenus. Y documenter les emplacements utilisés pour le stockage. Sauvegarder cet inventaire et en conserver au moins une copie à part des contenus eux-mêmes.	Stocker suffisamment de métadonnées pour connaître les contenus (possibilité de combiner les métadonnées administratives, techniques, descriptives, de préservation et structurelles).	Déterminer quel standard de métadonnées appliquer. Trouver et combler les lacunes dans les métadonnées pour se conformer à ces standards.	Archiver les actions de préservation associées au contenu et les occurrences de ces actions. Choisir et implémenter des standards de métadonnées.
Contenu	Documenter les formats de fichiers et toutes les autres propriétés essentielles (*significant properties*) des contenus, y compris les modalités et la date d'acquisition de cette documentation.	Vérifier les formats de fichiers et les autres propriétés essentielles (*significant properties*) des contenus. Développer des relations avec les créateurs de contenus pour encourager des choix de formats de fichiers durables.	Surveiller l'obsolescence et les évolutions des technologies dont dépendent les contenus.	Mener des opérations de migration, de normalisation, d'émulation, etc. pour s'assurer que les contenus restent acccessibles.

Niveaux de préservation numérique v2 — 26 septembre 2019

DOI 10.17605/OSF.IO/QGZ98

Fig.1 : Niveaux de préservation numérique (source : https://bnf.hal.science/hal-02551807v1)

1.1.2. La feuille de calcul

Une feuille de calcul (disponible et téléchargeable à l'adresse suivante : https://bnf.hal.science/hal-02551807v1 ; DOI : 10.17605/OSF.IO/QGZ98[1]) permet de procéder à l'évaluation en entrant dans chaque case une valeur entre 0 et 2 :

- la valeur 2 permet de dire que le service a atteint l'objectif proposé ;
- la valeur 1, que le service travaille sur le sujet ;
- la valeur 0, que le service n'a rien fait en la matière.

Un code couleur (vert/jaune/rouge) permet d'identifier automatiquement les domaines les plus critiques et d'entamer le travail de réflexion nécessaire à l'amélioration de la situation : fixation d'objectifs et de priorités ; demande et allocation de ressources ; communication auprès des différents interlocuteurs du service (autorités hiérarchiques ou de tutelle, financeurs, collaborateurs internes, services producteurs, usagers).

1.1.3.Limites...

Conseil

L'opération d'évaluation peut être relancée régulièrement, ce qui permet de mesurer l'état d'avancement des actions mises en œuvre et la progression du service en maturité.

Attention

L'outil proposé est pratique, facile à prendre en main. Il présente cependant l'inconvénient de se concentrer sur les procédures existantes et les moyens techniques, et néglige l'analyse du contexte d'intervention du service d'archives, ainsi que les aspects stratégiques et les moyens disponibles.

1.2. Grille d'évaluation rapide (Rapid Assessment Model) de Digital Preservation Coalition

Le second outil est constitué par la « *Grille d'évaluation rapide* » conçue par « *Digital Preservation Coalition* », structure à but non lucratif britannique qui sera également présentée dans la section 6 de ce module.

La grille d'évaluation a globalement la même finalité que le premier outil. Il s'agit d'aider les organisations – y compris les services d'archives – à évaluer rapidement leur degré de maturité en matière de préservation numérique, ainsi que leur capacité à mettre en œuvre des actions en la matière.

Remarque

Son périmètre est plus large que celui des « *Niveaux de préservation* » de la NDSA, car elle couvre également les questions stratégiques et juridiques et les moyens disponibles.

1. https://dx.doi.org/10.17605/OSF.IO/QGZ98

Description de la grille d'évaluation conçue par Digital Preservation Coalition

La grille d'évaluation (disponible à l'adresse suivante : https://www.dpconline.org/docs/digital-preservation/ram/translations-4/2441-dpc-ram-2-0-fr/file) se décompose en deux niveaux regroupant 11 sections :

- le niveau stratégique :
- - section A : viabilité de l'organisation (mode de gouvernance, structure organisationnelle, dotation en personnels et en ressources) ;
  - section B : politique et stratégie ;
  - section C : bases légales (gestion des droits et responsabilités, éthique et déontologie) ;
  - section D : ressources informatiques ;
  - section E : amélioration continue (définition des objectifs, suivi de ceux-ci) ;
  - section F : communauté (engagement de l'organisation dans la communauté professionnelle) ;
- le niveau opérationnel :
- - section G : collecte, transfert et prise en charge ;
  - section H : préservation binaire (stockage, gestion de l'intégrité) ;
  - section I : préservation des informations et de leur accessibilité ;
  - section J : gestion des métadonnées ;
  - section K : recherche et accès.

À chaque section, les organisations doivent évaluer leur niveau de maturité, sur une échelle de 0 à 4, en fonction de critères qui sont définis dans le manuel accompagnant la grille :

- niveau 0 : conscience faible ;
- niveau 1 : conscience ;
- niveau 2 : gestion minimale ;
- niveau 3 : gestion standard ;
- niveau 4 : gestion optimisée.

Un tableau récapitulatif permet à l'organisation procédant à l'évaluation de :

- reporter le niveau choisi pour chaque section (le niveau actuel), en documentant le choix ;
- identifier l'objectif à atteindre en matière de niveau (le niveau cible) ;
- définir les actions à mettre en œuvre pour atteindre ce niveau cible.

Conclusion

Comme on peut le constater, ces deux outils sont extrêmement simples d'utilisation et très ergonomiques.

Pour en savoir plus, consulter la page suivante, où ces deux méthodes d'évaluation ont été traduites en français : https://www.association-aristote.fr/sous-groupe-traduction/

2. Connaître la nature des documents d'archives conservés et les caractériser

Introduction

Parallèlement à l'évaluation du niveau de maturité, il est recommandé d'apprendre à mieux connaître les caractéristiques techniques des documents d'archives sur support numérique que le service d'archives conserve, indépendamment de leurs caractéristiques archivistiques (service producteur, processus métier de production, etc.).

Cette connaissance repose sur plusieurs actions :

- la connaissance de l'état des supports de stockage existants ;
- la connaissance des formats de fichiers conservés et leur caractérisation ;
- le rassemblement des informations de représentation et des métadonnées disponibles.

2.1. Connaissance et état des supports de stockage

Rappel

Comme nous l'avons vu dans la section précédente, les supports de stockage numériques sont en perpétuelle évolution, comme le matériel permettant de les lire.

Le recensement et le contrôle des supports de stockage existants, à défaut de politique de stockage mature, constituent donc des prérequis indispensables.

Ces actions permettent de :

- identifier les différents types de stockage existants (cassettes et autres supports magnétiques, disquettes, CD, DVD, disques durs, clés USB, serveurs, offre cloud) dans le service et leur nombre ;
- repérer les problèmes posés pour accéder aux fichiers stockés : absence de matériel ou de pilote de lecture, absence de formatage des supports ;
- dénombrer les fichiers présents sur chaque type de stockage existant ;
- vérifier l'existence ou non de sauvegardes des fichiers enregistrés sur ces différents supports, ainsi que la présence de doublons, en mettant en œuvre un calcul d'*empreintes* p.21 et en comparant celles-ci ;
- s'assurer qu'il n'y a pas de fichier chiffré pour lequel on ne dispose pas de la clé de déchiffrement.

Conseil

Si aucune action n'a été entreprise auparavant, l'idéal consiste à copier le contenu des différents supports sur un support unique (de type serveur ou *cloud computing*, en fonction du caractère critique de la confidentialité des informations) et à répliquer celui-ci, en veillant à ce que la copie générée soit régulièrement actualisée, stockée dans un lieu géographique différent et protégée contre d'éventuelles attaques (pour en savoir plus : section 3, chapitre 4 de ce module).

2.2. Connaissance et caractérisation des formats de fichiers conservés

Introduction

Connaître les *formats de fichiers* p.21 des documents d'archives sur support numérique conservés dans un service d'archives et caractériser ces documents permet de mieux identifier les problèmes de préservation numérique à affronter.

Cette étude repose sur deux opérations particulières :

- l'identification des formats de fichiers ;
- pour certaines catégories de formats de fichiers, la *validation du format des fichiers.* p.23

2.2.1. Identification des formats de fichiers

2.2.1.1. Les trois techniques d'identification du format d'un fichier numérique

L'identification est une opération technique qui permet de définir précisément quel est le format d'un fichier, quel qu'il soit. Elle permet de catégoriser les documents d'archives conservés et d'identifier les grandes catégories de formats de fichiers (traitement de texte, tableurs, bases de données, images fixes, son, images animées, sites web, etc.) que le service d'archives a pris en charge, qu'il doit préserver et auxquelles il doit donner accès sur le long terme.

Trois techniques d'identification du format d'un fichier numérique existent :

- **l'*extension du fichier*** p.21 (ex. .doc) : c'est le moyen employé par la plupart des systèmes d'exploitation pour identifier le format d'un fichier et définir quel logiciel sera proposé à l'utilisateur pour ouvrir celui-ci et visualiser son contenu. Malheureusement, aucune règle ou norme particulière ne précise la méthode de formatage des extensions et chaque système d'exploitation possède ses propres règles. Par ailleurs, les fichiers les plus anciens (antérieurs au développement des systèmes d'exploitation les plus courants) ne sont pas toujours dotés d'extension et certains systèmes d'exploitation n'utilisent les extensions que depuis peu de temps. Se baser sur l'extension d'un fichier pour déterminer son format est donc problématique, d'autant qu'il est très facile de modifier cette information ;
- **les *métadonnées techniques*** p.22 que le fichier embarque, non visibles directement par l'utilisateur, notamment son type MIME (Multipurpose Internet Mail Extensions) : apparu en 1991, le *type MIME* p.22 consiste en un système normalisé d'identifiants enregistrés par l'Internet Assigned Numbers Authority (IANA) – même si certaines organisations ont créé leur propre type MIME, sans l'enregistrer auprès de l'IANA. L'IANA maintient donc une liste presque exhaustive des types MIME disponible à l'adresse suivante : https://www.iana.org/assignments/media-types/media-types.xhtml. Le type MIME est destiné à faciliter la visualisation dans un navigateur web, mais est également utilisé par certains systèmes d'exploitation. Le type MIME est composé d'un type et d'un sous-type, séparés par un slash (ex. audio/mpeg, text/csv, image/jpeg). Se baser uniquement sur le type MIME est tout aussi risqué que se baser sur l'extension, car un même type MIME peut être partagé par plusieurs versions d'un même format de fichiers voire entre plusieurs formats de fichiers. Les erreurs d'identification sont donc nombreuses ;

- ***la signature du fichier*** [p.22] (le magic number) : la signature d'un fichier est constituée d'un ensemble de caractères propre à chaque format. À l'origine, les 2 octets stockés au début de chaque fichier suffisaient à identifier le format de fichiers. Aujourd'hui, il est nécessaire de définir une chaîne d'octets plus complexe afin d'identifier de manière certaine le format de fichiers. Cet ensemble de caractères n'est généralement pas visible par l'utilisateur et nécessite l'utilisation d'outils spécifiques – un éditeur hexadécimal – pour être identifié. Même si tous les formats de fichiers ne disposent pas de signatures – notamment les fichiers HTML ou XML –, les signatures offrent de meilleures garanties pour identifier finement le format d'un fichier. Cette méthode est largement utilisée par les outils développés pour la préservation numérique.

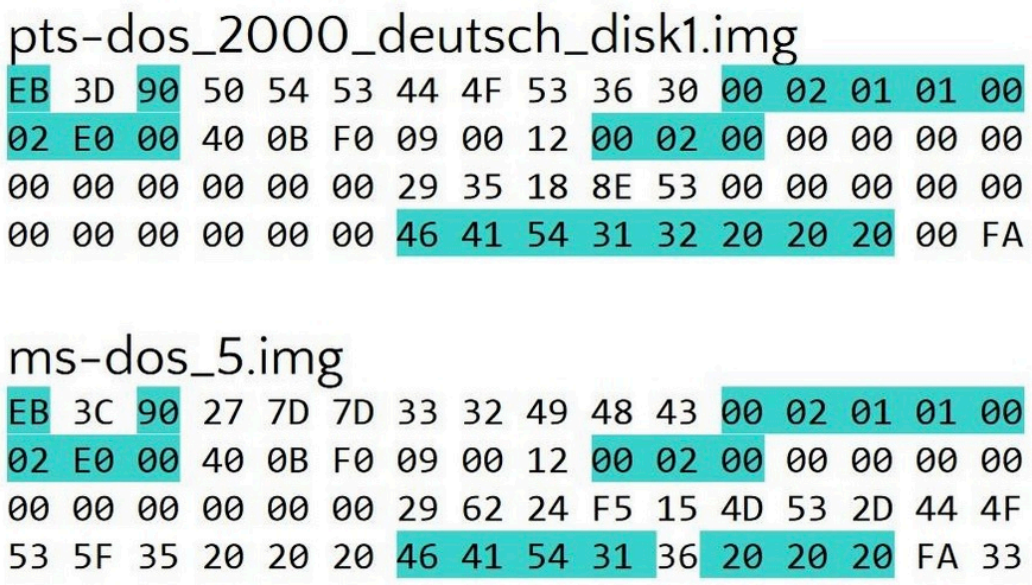


Fig.2 : Signature d'un fichier FAT12, identifiée au moyen d'un éditeur hexadécimal (source : David Clipsham, Nick Krabbenhoeft, Shira Peltzman, Justin Simpson, Carl Wilson, *« PRONOM in practice »*, IPRES 2018, page 62, https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwiultK6_ZqAAxWsTaQEHYnuBrAQFnoECBcQAQ&url=https%3A%2F%2Fosf.io%2Fy28h3%2Fdownload&usg=AOvVaw3AoEr6yMPdeIKiHCwi_7ce&opi=89978449[2])

Remarque

Extensions, types MIME et signatures des formats sont enregistrés, avec la documentation trouvée sur les différents formats, par les spécialistes de la présentation numérique dans des bases de données – des registres – décrivant les différents formats, recensant les moyens de les identifier et les outils permettant de les traiter et de les rendre accessibles.

2.2.1.2. Le registre PRONOM

Si plusieurs registres ont été créés – dont certains ont été abandonnés – l'un d'entre eux peut aujourd'hui être considéré comme la base de connaissance de référence : le *registre PRONOM* [p.22] (https://www.nationalarchives.gov.uk/PRONOM/) maintenu par The National Archives (Royaume-Uni).

Conçu en 2002 et mis en ligne à partir de 2004, le registre PRONOM attribue à chaque format (et à chaque version de ceux-ci) un identifiant unique et enregistre les informations suivantes :

Un résumé des informations essentielles disponibles :

- nom du format ;
- version du format ;

2. https://osf.io/y28h3/download

- identifiants (identifiant unique attribué par PRONOM – PRONOM Persistent Unique Identifier (PUID), type MIME) ;
- catégorie de format de fichiers (ex. image fixe) ;
- description ;
- formats de fichiers associés ;
- date de mise à disposition ;
- dates de création et de mise à jour de la notice ;

La documentation disponible, notamment si le format a fait l'objet d'une normalisation ;

Les moyens d'identification disponibles :

- extensions ;
- nombres magiques (magic number), correspondant à la signature d'un fichier (cf. supra) ;

Les méthodes de compression utilisées, si elles existent ;

Les méthodes d'encodage des caractères, si nécessaire ;

Les droits de propriété intellectuelle associés ;

Les fichiers de référence ;

Les propriétés du format.

Les fiches correspondant aux formats de fichiers sont :

- consultables sur le site internet de The National Archives (UK) ;
- récupérables dans leur intégralité uniquement en utilisant un script.

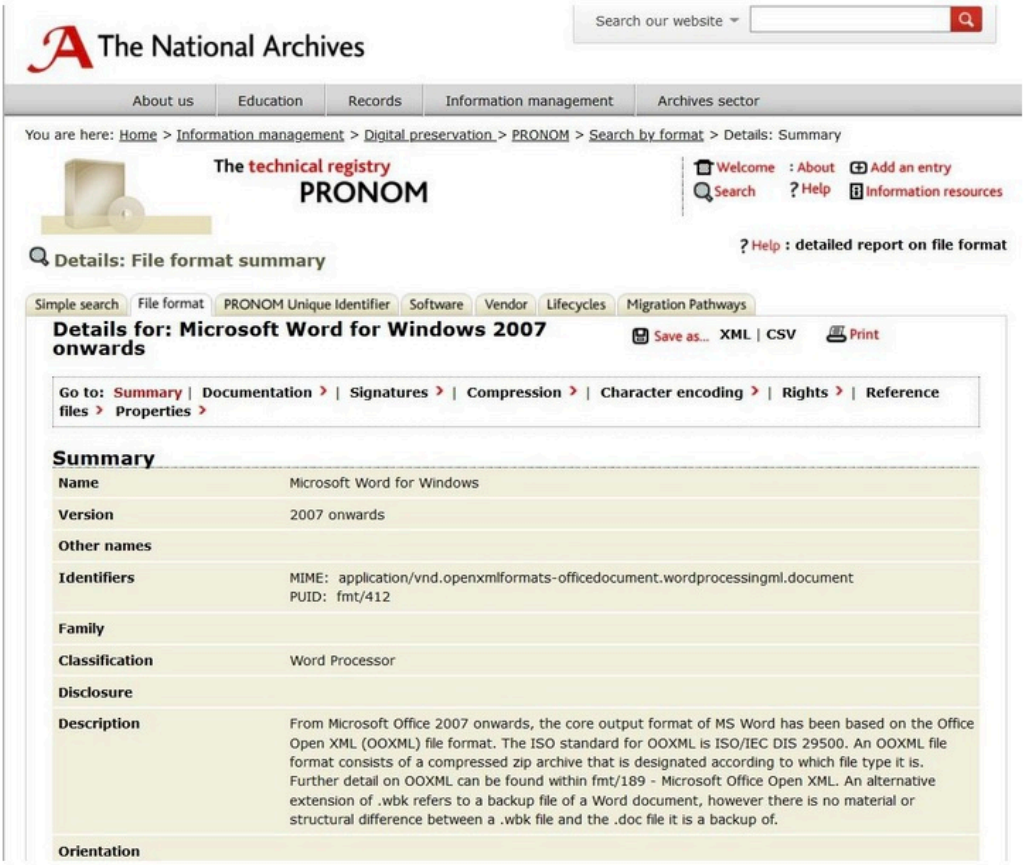


Fig.3 : Fiche PRONOM correspondant au format de fichier Microsoft Word for Windows 2007 onwards (https://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=1160[3])

Remarque

Si The National Archives administre le registre et l'enrichit régulièrement, elle accueille également les contributions de tous les spécialistes de la préservation numérique, ce qui fait du registre un outil collaboratif international.

Des recensements ont été initiés dans d'autres institutions comme la Bibliothèque du Congrès de Washington.

2.2.1.3. Les logiciels d'identification ayant PRONOM comme référentiel

Le registre PRONOM est le principal registre utilisé pour initier le travail de caractérisation des fonds d'archives sur support numérique conservés par les services d'archives. Plusieurs logiciels d'identification de formats l'utilisent comme référentiel :

- le logiciel Digital Record Object Identification (DROID – https://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/), conçu, développé et maintenu par The National Archives (Royaume-Uni), est le principal d'entre eux. Il présente l'avantage d'être téléchargeable et de disposer d'une interface graphique facile à manipuler et permettant des exports de résultats sous forme de tableur ;
- le logiciel Siegfried (https://www.itforarchivists.com/siegfried/), qui est conçu, développé et maintenu par un particulier, ne dispose en revanche pas d'interface graphique, mais peut être utilisé en ligne de commande ;
- le logiciel Format Identification for Digital Objects (FIDO – https://github.com/openpreserve/fido), qui ne dispose pas non plus d'interface graphique ;
- le logiciel File Information Tool Set (FITS – https://projects.iq.harvard.edu/fits/home), conçu, développé et maintenu par l'Université Harvard, qui ne dispose pas non plus d'interface graphique.

2.2.1.4. Les limites du registre de référence PRONOM

Certains points d'attention méritent cependant d'être signalés :

- le registre PRONOM – comme les autres registres d'ailleurs – est loin d'être exhaustif. Seuls les principaux formats sont recensés. Il est donc possible que les logiciels utilisant PRONOM ne soient pas en mesure d'identifier le format de tous les fichiers correspondant aux documents d'archives sur support numérique que votre service d'archives conserve ;
- le format des fichiers anciens ne disposant pas d'extension ne peut être identifié, sauf existence d'une signature connue d'un registre comme PRONOM ;
- en cas d'incertitude sur le format exact du fichier, ces logiciels peuvent proposer plusieurs identifications, ou proposer une identification par défaut qui peut s'avérer fausse ;
- les logiciels qui utilisent le registre PRONOM comme référentiel ne paramètrent pas son utilisation de la même manière. Aussi les résultats obtenus peuvent-ils être différents en fonction du logiciel retenu. Pour un même fichier, en cas de doute, DROID peut remonter plusieurs PUID. Siegfried le fait aussi. Mais il peut être décidé par les logiciels de forcer le choix d'un PUID ;

3. Fiche PRONOM - https://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=1160

- enfin, ce n'est pas parce que le format d'un fichier n'a pas pu être identifié que celui-ci ne pourra pas être interprété par un logiciel, rendant son contenu accessible aux utilisateurs.

i) Conclusion

L'opération d'identification de format, et notamment les résultats générés par les logiciels et que l'on peut récupérer sous forme de tableur, permettent aux services d'archives d'alimenter un recensement des fichiers qu'ils conservent – avec de nombreuses informations les décrivant, dont leur format. Ce recensement permet ensuite de générer des vues graphiques macroscopiques des documents conservés, ce qui permet de repérer les catégories de formats de fichiers dominantes (ex. traitement de texte, tableurs, images fixes, son, images animées, sites web, messages électroniques, etc.) et de réfléchir aux actions prioritaires à mettre en œuvre en matière de préservation.

2.2.2. Validation des formats de fichiers

2.2.2.1. Description de l'opération de validation

La validation est une opération technique qui consiste à vérifier la conformité du format d'un fichier avec les spécifications de celui-ci, d'un point de vue syntaxique comme sémantique. Si le format du fichier est déclaré valide, tout logiciel susceptible d'exécuter le fichier pour en proposer l'accès sera capable de le faire.

La validation de formats repose sur un examen technique intégral du fichier et une comparaison de sa structure et de sa syntaxe avec les spécifications du format, telles qu'elles ont été interprétées au moyen de tests préalables conçus à la lecture des spécifications. En cas de non-conformité, les outils logiciels développés pour réaliser ces opérations peuvent proposer des actions correctives permettant de repérer l'erreur identifiée ou simplement se contenter de signaler la non-conformité.

Attention

Pour être réussie, l'opération de validation de formats suppose donc que :

- le format dispose de spécifications écrites et accessibles ;
- les spécifications du format ont fait l'objet d'une interprétation intégrale ne laissant pas de place au doute. Si les spécifications de certains fichiers sont très simples, d'autres sont très complexes (ex. PDF).

2.2.2.2. Les logiciels de validation

Plusieurs logiciels ont été conçus, développés et maintenus par des concepteurs de formats, des éditeurs de logiciels ou des experts en préservation numérique pour réaliser des opérations de validation de formats :

- certains sont génériques, c'est-à-dire qu'ils permettent de procéder à la validation de plusieurs formats de fichiers :
    - JSTOR Harvard Object Validation Environment (JHOVE) : conçu et développé en 2003 par JSTOR et la bibliothèque de l'Université Harvard, l'outil est désormais maintenu par l'Open Preservation Foundation. Basé sur une logique modulaire, il permet de valider plusieurs types de formats de fichiers : HTML, XML, GIF, JPEG, JPEG2000, TIFF, PDF, AIFF, WAVE. Il est utilisable soit via une interface graphique, soit en ligne de commande et génère des résultats sous forme textuelle ou sous forme de fichier XML. Les résultats obtenus sont cependant plus ou moins satisfaisants en fonction des formats ;

Fig. 4 : Sélection d'un module dans JHOVE

- certains sont spécifiques à un format de fichiers donné :
  - veraPDF pour le format PDF/A ;
  - ImageMagick ou Jpylyzer pour les fichiers correspondant à des images fixes ;
  - plusieurs outils existent pour les fichiers aux formats XML ou JSON, dont certains sont accessibles sur internet.

2.2.2.3. Les écueils à éviter...

Certains points d'attention méritent cependant d'être signalés :

- l'opération de validation de formats est une opération plus complexe que l'identification des formats ;
- seuls quelques formats de fichiers bénéficient d'outils de validation. Nombreux sont les formats de fichiers qui n'en bénéficient pas ;
- toutes les spécifications d'un format n'ont pas nécessairement été interprétées et testées. Certaines non-conformités peuvent donc ne pas être repérées par les outils de validation de formats ;
- certains outils se contentent de signaler la non-conformité du fichier sans décrire précisément celle-ci ;
- les non-conformités signalées par les outils ne peuvent parfois être interprétées que par des spécialistes du format lui-même ;
- le fait qu'un fichier ne soit pas conforme aux spécifications de son format ne signifie pas qu'il ne pourra pas être exécuté par un logiciel et que son contenu ne pourra pas être accessible. Les logiciels de visualisation sont en effet très permissifs aux non-conformités. À titre d'exemple, un fichier au format PDF considéré comme non valide par un outil comme veraPDF pourra tout à fait être ouvert avec un logiciel de visualisation de documents PDF, et son contenu sera parfaitement intelligible ;
- les fichiers produits dans le cadre d'opérations de numérisation faisant l'objet de contrôles qualité poussés sont souvent de meilleure qualité technique que les fichiers produits de manière nativement numérique par les services producteurs avec les logiciels du marché ;
- toute correction automatique lancée suite à la détection d'une non-conformité porte atteinte à l'intégrité binaire du fichier pris en charge par le service d'archives. Elle ne doit être lancée qu'en étant sûr qu'elle ne va pas porter atteinte à l'intégrité intellectuelle du contenu du fichier. En cas de réalisation d'une telle opération, il peut être prudent de conserver le fichier d'origine ;

- l'opération de validation de formats reposant sur une analyse exhaustive des fichiers, elle est consommatrice de puissance de calcul. Elle peut donc nécessiter un matériel puissant ou une segmentation de l'opération en petits lots.

Complément

La réalisation d'une opération de validation de formats donne donc au service d'archives une indication quant à la qualité technique des fichiers correspondant aux documents d'archives sur support numérique qu'il a pris en charge. Elle facilite l'identification des fichiers pour lesquels une surveillance particulière devra être mise en œuvre, pour garantir la préservation et l'accessibilité à long terme.

Conclusion

Les opérations d'identification et de validation des formats de fichiers fournissent aux services d'archives le moyen d'alimenter un registre des fichiers pris en charge, ce qui donnera par la suite la possibilité de réaliser des opérations de *récolement* p.22. Elles permettent également de catégoriser les fonds, de définir les actions prioritaires à entreprendre en matière de préservation. La démarche est comparable à celle entreprise pour les documents d'archives sur support physique et analogique permettant d'établir une planification des actions à prévoir, en fonction de la proportion de photographies, plans, documents audiovisuels, etc.

2.3. Rassemblement des informations de représentation et extraction des métadonnées

Introduction

Rassembler les informations de représentation disponibles et extraire les métadonnées présentes dans les fichiers eux-mêmes constitue enfin un bon moyen de connaître les documents d'archives pris en charge.

2.3.1. Rassembler les informations de représentation disponibles

Les informations de représentation nécessaires pour interpréter des documents d'archives sur support numérique peuvent être nombreuses.

Dans la plupart des cas, ces informations sont les mêmes que celles qui sont nécessaires pour la conservation des documents d'archives sur support physique ou analogique :

- *métadonnées descriptives* p.22 : service producteur, service versant, intitulé du document, date du document, etc. ;
- *métadonnées de gestion* p.22: durée d'utilité administrative, délai de communicabilité, niveau de protection du document au titre de la réglementation du secret, existence de droits d'auteur associés et titulaires de ceux-ci, etc.

Cependant, comme on l'a vu dans la section 1 de ce module., les documents d'archives sur support numérique sont sujets à des *codages métier* p.21. Faute d'information sur ceux-ci, les documents ne peuvent pas être interprétés.

Fondamental

Il importe donc de vérifier si toutes les informations de représentation au sens du modèle de référence OAIS sont présentes et exhaustives, notamment :

- la description du processus métier ;
- les informations relatives à la structure du document d'archives ;
- les informations relatives à la sémantique utilisée dans le document d'archives.

À défaut de ces informations ou dans le cas où il est impossible de les (re)constituer, c'est la conservation même du document d'archives qui devra être interrogée.

2.3.2. Extraction de métadonnées présentes dans les fichiers

Les fichiers contiennent nativement des métadonnées ou des propriétés qui peuvent s'avérer utiles pour la préservation numérique et la planification des opérations de préservation (taille du fichier, par exemple) ou en cas d'absence ou de manque d'informations de représentation.

Pour certains formats de fichiers, existent même des standards de métadonnées (par exemple pour les images fixes et les images animées, cf. section 5 de ce module).

Ces métadonnées embarquées dans le fichier peuvent faire l'objet d'une extraction automatique, afin d'enrichir les outils de recherche du service d'archives et de vérifier, lorsque des actions de préservation sont mises en œuvre (notamment les migrations de format), que celles-ci sont restées intègres et exactes.

Méthode

Pour que l'extraction soit possible et efficace, deux prérequis doivent cependant être remplis :

- le format de fichier doit disposer de spécifications écrites et disponibles, permettant de savoir comment et sous quelle forme ces métadonnées sont encapsulées dans les fichiers ;
- les spécifications du format ne doivent pas être sujettes à interprétation, ce qui peut malheureusement être le cas.

L'extraction de métadonnées internes est réalisée au moyen de logiciels conçus et édités par différents organismes (concepteurs de formats de fichiers, éditeurs de logiciels, experts de la préservation numérique). Ces outils repèrent et extraient les métadonnées encapsulées dans le fichier et les mettent en forme selon une grammaire et une syntaxe propre.

Ces outils peuvent être :

- génériques, permettant de procéder à la validation de plusieurs catégories de formats de fichiers : ex. FITS, Tika ;
- propres à une catégorie de formats de fichiers donnée : ex. Jpylizer pour le format JPEG 2000.

2.3.3. Limites de l'extraction

Points d'attention :

- tous les formats de fichiers ne disposent pas d'outils d'extraction de métadonnées ;
- les outils génériques sont peu nombreux. Il est donc souvent nécessaire d'utiliser plusieurs outils pour couvrir le maximum de formats de fichiers ;
- l'extraction de métadonnées est une opération technique qui prend du temps (jusqu'à 10 heures pour extraire les métadonnées de 100 000 fichiers avec un outil comme PreScan, par exemple) ;

- les outils d'extraction de métadonnées peuvent renvoyer des messages d'erreur difficilement compréhensibles, car très techniques ;
- le nombre de métadonnées extraites automatiquement peut être très important et celles-ci peuvent n'avoir aucun intérêt pour la préservation numérique (ex. numéro de série d'un appareil de prise de vue photographique). Il est souvent nécessaire d'opérer une sélection des métadonnées extraites ;
- la manière dont les métadonnées extraites sont formatées est souvent propre à chaque outil. Leur interprétation peut s'avérer délicate ;
- la valeur et la qualité des métadonnées extraites dépendent de la manière dont celles-ci ont été générées ou produites dans le fichier :
    - à titre d'exemple, dans un fichier bureautique, la métadonnée Author peut correspondre non seulement à l'auteur du document au sens diplomatique (la personne qui valide le document) mais aussi à l'auteur du modèle de document (celui qui a conçu le formulaire, par exemple). Il peut être renseigné nommément (nom, prénom) ou sous la forme de l'identifiant enregistré dans un annuaire (ce qui ne garantit pas une identification précise de la personne concernée) ;
    - les métadonnées d'ordre temporel (date de création du fichier, par exemple) nécessitent, pour être interprétées et interprétables, de savoir quel référentiel de temps est utilisé par le système de production (temps universel, temps observé sur le lieu de création du fichier) ;
- il convient de savoir comment exploiter ces métadonnées extraites. La meilleure solution consiste à les réinjecter dans les outils de recherche du service d'archives.

Conclusion

Cette section a permis d'apprendre à établir un diagnostic sur les fichiers correspondant aux documents d'archives pris en charge. C'est sur la base de ce diagnostic que l'archiviste peut désormais définir ses modalités d'intervention.

Glossaire

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)

Collecte

Action de rechercher et de recueillir des documents, des objets, des informations.

Conservation matérielle

Ensemble de techniques, méthodes et procédés destinés à assurer l'intégrité physique des documents.

Description

Ensemble des opérations d'identification d'une unité archivistique, de sa description matérielle au contexte de sa production en passant par l'analyse du contenu et l'indexation. L'expression désigne à la fois le processus de représentation et son résultat.

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).

Extension d'un fichier

Une extension de nom de fichier (ou simplement extension de fichier, voire extension) est un suffixe donné au nom d'un fichier pour identifier son format. Ainsi, on dira qu'un fichier nommé truc.doc a l'extension doc ou .doc.

https://www.techno-science.net/definition/7661.html

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).

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.

Métadonnées

Données décrivant le « contexte, le contenu et la structure ainsi que leur gestion dans le temps » (ISO15489, 2016). On les utilise notamment pour définir les « spécifications techniques, l'organisation intellectuelle, les conditions d'utilisation, la préservation, l'échange entre systèmes et l'administration des données » (Turner, s.d. cité dans EBSI, 2018).

Métadonnées de gestion

Les métadonnées de gestion ou métadonnées administratives comprennent les métadonnées d'identification (cote par exemple), les métadonnées d'intégrité (signature électronique, horodatage, empreinte par exemple), les métadonnées décrivant les droits (gérant l'accès, la communicabilité, la réutilisation).

Métadonnées descriptives

Les métadonnées descriptives sont les métadonnées qui servent à organiser la connaissance. Ce sont les métadonnées qui identifient, classifient, hiérarchisent l'information contenue dans l'objet numérique.

Métadonnées techniques

Les métadonnées techniques sont les métadonnées qui servent à identifier, caractériser, définir l'environnement technique des objets numériques.

Récolement

Opération consistant à dresser la liste topographique des articles conservés dans un service d'archives ou un fonds. Désigne aussi l'opération destinée à vérifier l'intégralité des fonds et collections d'un service d'archives périodiquement ou lors du changement de responsable d'un service d'archives.

RegistrePRONOM

Le registre PRONOM (https://www.nationalarchives.gov.uk/PRONOM/) maintenu par The National Archives (Royaume-Uni) est considéré comme la base de connaissance de référence. Il attribue à chaque format (et à chaque version de ceux-ci) un identifiant unique et enregistre les informations suivantes : un résumé des informations essentielles disponibles ; la documentation disponible ; les moyens d'identification disponibles ; les méthodes de compression utilisées, si elles existent ; les méthodes d'encodage des caractères, si nécessaire ; les droits de propriété intellectuelle associés ; les fichiers de référence ; les propriétés du format. (PIAF/Vasseur/Module 7C)

signature (d'un fichier)

La signature d'un fichier est constituée d'un ensemble de caractères propre à chaque format. À l'origine, les 2 octets stockés au début de chaque fichier suffisaient à identifier le format de fichiers. Aujourd'hui, il est nécessaire de définir une chaîne d'octets plus complexe afin d'identifier de manière certaine le format de fichiers (PIAF/Vasseur/Module 7C).

Type MIME

Le type MIME (Multipurpose Internet Mail Extensions) consiste en un système normalisé d'identifiants enregistrés par l'Internet Assigned Numbers Authority (IANA) – même si certaines organisations ont créé leur propre type MIME, sans l'enregistrer auprès de l'IANA (PIAF/Vasseur/Module 7C).

Validation de format de fichier

La validation de format cherche à vérifier si un fichier numérique est conforme aux spécifications connues du format, d'un point de vue syntaxique - le fichier est bien formé - comme sémantique - le fichier est valide.