



HAL
open science

La conservation à long terme des documents numérisés : conseils pratiques

Judith Hannoun

► **To cite this version:**

Judith Hannoun. La conservation à long terme des documents numérisés : conseils pratiques. École thématique. France. 2015. cel-01813928

HAL Id: cel-01813928

<https://hal.science/cel-01813928>

Submitted on 12 Jun 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

La conservation à long terme des fichiers numériques

CONSEILS PRATIQUES

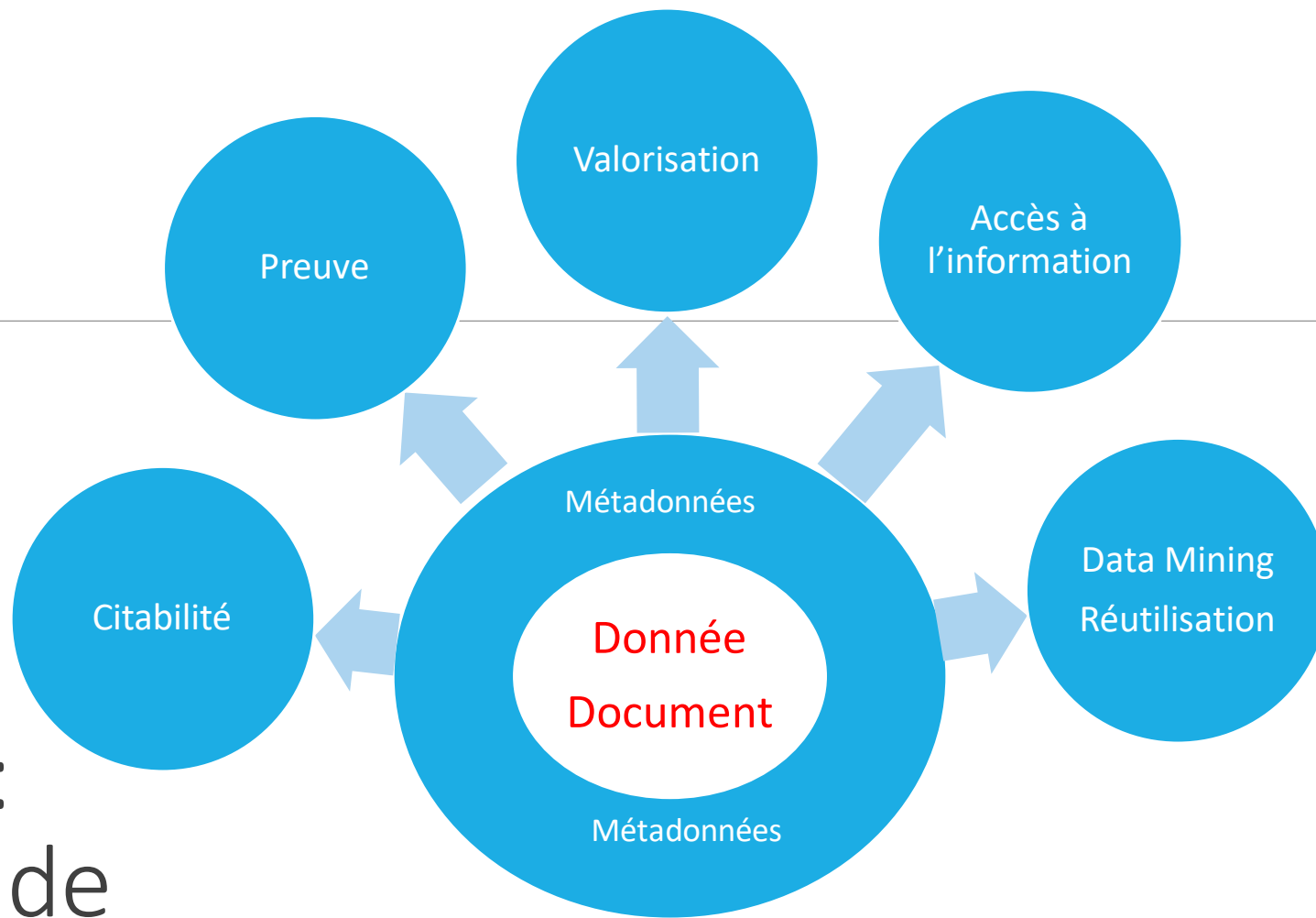
25 JUIN 2015 – JUDITH HANNOUN

Conservation

Ensemble des actions, outils et méthodes mis en œuvre pour :

- Stocker
- Sécuriser
- Pérenniser
- Restituer
- Tracer
- Transférer

L'enjeu :
ingénierie de
la
connaissance



Accès sécurisé

Ethique et droit

Niveau scientifique

Processus de numérisation

Stockage et archivage à long terme

Etapes d'un projet de numérisation

- Préparation des documents
- Rédaction de conventions avec inventaire le plus précis
- rédaction d'un cahier des charges
- **Numérisation**
- **Contrôle qualité**
- **Conversion et traitement des fichiers**
- **Sauvegarde sur les différents supports**
- **Mise en ligne sur la plateforme**
- **Renseignement des métadonnées**
- Traitement textuel : OCRisation, transcription, annotation,
- Choix des accès, définition des rôles de chacun,...

Ne pas confondre

Archivage / Sauvegarde

Contenu figé, trié et classé	Contenu évolutif
Objectif : consultation et conservation	Objectif : Restauration en prévention d'une erreur ou d'un dysfonctionnement technique
Moyen et long terme	Court et moyen terme

Défis

- Financier : coût élevé de la numérisation et la préservation ultérieure, évaluer la pertinence d'une externalisation
 - Technique : définir les exigences d'archivage et de conservation ; aspect lié à la diversité et au degré de développement de la technologie impliquée
 - Organisationnel : identifier les acteurs, mettre en place des procédures,...
 - Juridique : connaître les différents droits, établir des conventions, évaluer les enjeux éthiques...
- => savoir écrire un projet, établir les postes de dépenses, évaluer les coûts de la numérisation, proposer des solutions techniques, évaluer les délais, tenir compte des facteurs d'échec et de réussite.**

Coût réels de la numérisation et de l'archivage des données

- Coût en personnel ou prestations de numérisation : 20% du coût d'un projet
- Coût en matériel de reproduction et de conservation
- Coût en hébergement, développement et maintenance. Les coûts globaux de conservation de l'information par To et par an varient entre 2000 et 3000 euros en fonction du support de stockage.

Principes des archives numériques et de leur conservation

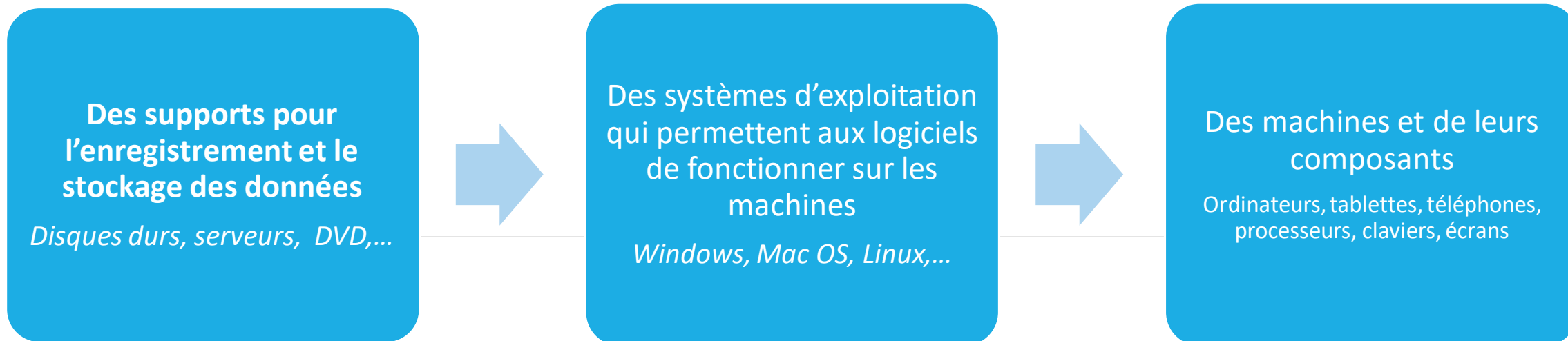
Il faut garantir :

- Existence, pérennité et accessibilité des documents
- Authenticité
- Fiabilité et intégrité
- Exploitation possible des documents
- Traçabilité et respect des droits d'accès
- Interopérabilité
- Citabilité

Caractéristiques techniques

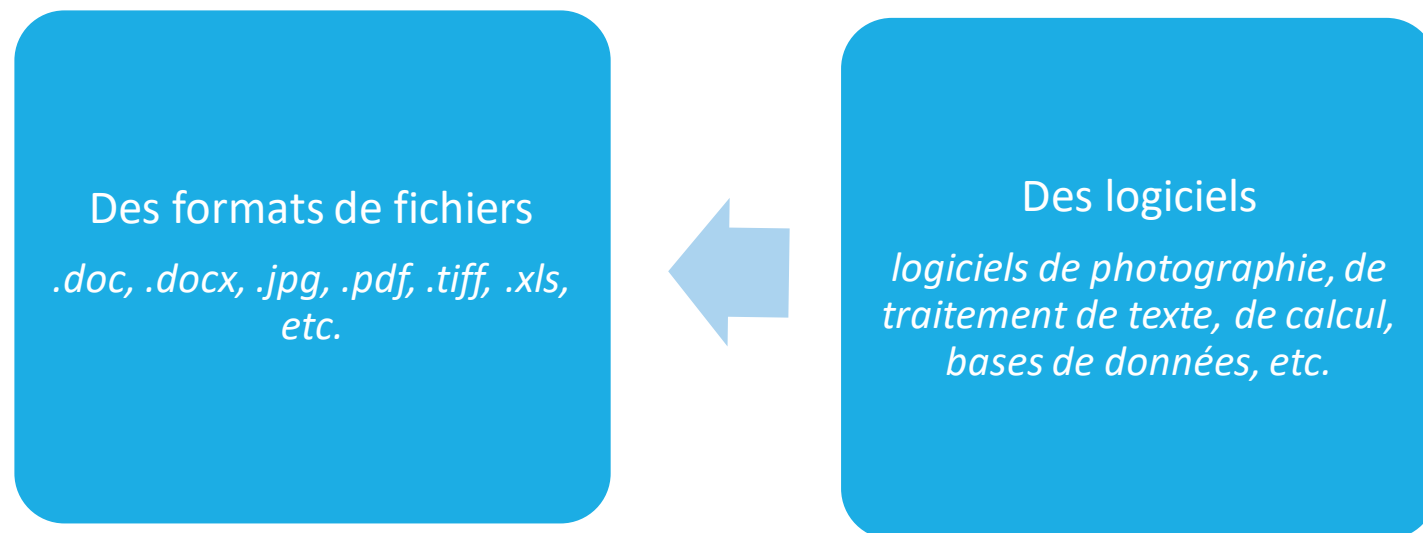
Les 4 principaux éléments qui permettent de caractériser un fichier numérique sont :

- **son format d'enregistrement**, reconnaissable à l'extension du fichier (.jpeg, .tiff, png, raw, ...)
- **Sa taille en pixels** : Attention à ne pas perdre l'information de la taille en cm du document
- **sa profondeur** (8 à 64 bits) : la profondeur définit la quantité d'informations chromatiques disponibles pour chaque pixel dans une image.
- **sa définition ou résolution** (72 à 4800 dpi en moyenne) : plus le document est petit, plus la résolution doit être élevée : pour du texte au format A4, en général de 300 dpi à 600 dpi ; pour une diapo, de 2400 dpi à 4800 dpi ; pour une photographie, 2400 dpi.
- **son profil couleur**



Lire et comprendre les données dépend :

Attention à l'obsolescence



Les supports d'enregistrement et de stockage

- Attention à l'obsolescence (10 ans environ). Il existe plusieurs types : magnétiques (bandes vhs, disques durs), optiques (CD-Rom, DVD, Blu-Ray) et flash (SSD, USB)
- Attention aux problèmes de compatibilité de matériel
- Migration régulière et systématique des fichiers : la durée de vie moyenne d'un DVD est de 5 ans, un disque dur externe tombe; un serveur tombe en panne,...
- Penser à tout dater : date de gravure, d'enregistrement, et à planifier les contrôles et les migrations..., coller des étiquettes sur les tranches
- A chaque étape, effectuer un contrôle qualité soigneux (numérisation, conversion, mise en ligne)
- Choisir des conditions optimales de conservation des supports (température, humidité, poussière, verrous,...) et choisir plusieurs sites.

Choix du support : critères

1. Le type de support (optique, magnétique, flash ou serveur)
2. La capacité de stockage
2. Le temps d'accès aux données
4. La durée de vie du support
5. Le coût à l'unité de mémoire
6. La matériel compatible ou nécessaire
7. Le nombre d'accès simultanés

Tableau des choix de supports de conservation (usage courant, à moyen et à long terme)

TYPE DE SUPPORTS	TYPE D'USAGE/CONSERVATION					
	Usage courant (0 à 5 ans)		Moyen terme (de 5 à 10 ans)		Long terme (+ de 10 ans)	
	Faible volume de données	Volume important de données	Faible volume de données	Volume important de données	Faible volume de données	Volume important de données
Disquette	OBSOLÈTE					
Disque Dur	X	X	X	X	X*	X*
Bande magnétique		X		X		X
Cd	X		X			
Dvd	X		X			
Blu-Ray		X		X		
Clé USB	X					

Format des fichiers

- Choisir si possible pour la conservation, des formats ouverts (opposés à propriétaires) reposant sur des normes et standards, dont les spécifications sont publiques, et qui sont largement diffusés. L'indication du format de fichier doit préciser si possible le nom du format et sa version.
- Attention au vieillissement des formats et procéder aux conversions tant que vos fichiers sont lisibles
- Organisez plusieurs copies de vos fichiers et conservées dans des endroits différents

TYPE DE DOCUMENT		UTILISATION	
		Exploitation	Conservation à long terme
Documents bureautiques	Texte	.docx, .odt, .texte, .pdf (PDF)	.pdf (PDF/A)
	Tableau / feuille de calcul	.xis, .odf	.pdf (PDF/A) (si statique)
			.xml, .csv (si dynamique)
	Présentation	.ppt, .odp	.xml, .svg
Base de données	.mdb, .accdb, .fmj	.xml, .csv	
Multimédia	Images	.jpg (JPEG), .jpg (JPEG2000), PNG	.tiff
	Son	.mp3, .wav	.wav, .pcm
	Vidéo	.avi	.aaf, .mxf
Plus spécifique	Carte / plan	.dwg, .svg	.dxf, .svg

JPEG VS TIFF

JPEG

- C'est le format le plus répandu. C'est un format compressé.
- Utilisé pour la diffusion et le web
- Le jpeg est un format exclusivement 8 bits et **est dégradé à chaque nouvelle sauvegarde**
- Pour la prise de vue avec un appareil photo numérique, privilégier le format raw

TIFF

- Format non compressé, le plus répandu pour numérisation avec un scanner
- Utilisé dans le traitement numérique de l'image et pour la sauvegarde des fichiers hautes résolution
- Il peut stocker les données sur 8, 16 ou 32 bits et ne se dégrade pas
- Poids des fichiers très important

Le nommage des fichiers

Il faut établir des règles de nommage strictes qui prennent en compte les contraintes liées aux documents, celles de la plateforme de diffusion et celles liées à la sauvegarde

Quelques précisions :

- La numérotation : faire précéder d'un nombre suffisant de zéros les chiffres pur garantir un tri cohérent
- Indiquer un minimum d'informations dans le nom des fichiers (cela facilite leur manipulation) et choisir des noms de fichiers qui font sens.
- Attentions aux caractères spéciaux, espaces ou accents
- Séparer les informations par un « _ underscore »
- Paramétrer le matériel pour une numérotation séquentielle et automatique
- utiliser des logiciels pour renommer les fichiers

something_0001

something_0002

something_0013

Traçabilité, métadonnées

- Ne jamais perdre le lien entre le document d'origine et le fichier numérique
- Récupérer le maximum d'informations sur le document d'origine et les reporter dans les métadonnées.
- Utiliser les outils de la gestion de projet pour garder des traces de toutes les opérations effectuées pour un fonds d'archives : journal de travail, diagramme de Gantt, datations multiples, tableau des tâches et des responsables,....
- Ne pas perdre les informations de taille du document
- Attribuer un identifiant unique et une URI

Documentation

Documenter une donnée revient à noter comment elle a été conçue et saisie, ce qu'elle signifie, à préciser son contenu et sa structure. La documentation est également décisive pour assurer une longue conservation des données.

Une bonne documentation comprend des informations sur:

- le contexte de la collecte de données (historique du projet, ses objectifs, ses hypothèses)
- les méthodes de collecte des données
- la structure des données et les liens entre elles
- la validation des données, les procédures de test, de preuves, etc.
- les éventuels changements dans le temps
- informations sur l'accès et les conditions d'usage ou de confidentialité

Type de métadonnées

- Métadonnées techniques : elles renseignent sur le fichier numérique et le processus de numérisation, soit enregistrées à l'intérieur même du fichier informatique auxquelles elles s'appliquent, soit stockées à l'extérieur du fichier, en lien avec lui.
- Métadonnées descriptives ou documentaires : elles renvoient à la problématique de recherche ou documentaire, à l'utilisation des données
- Métadonnées génériques : elles proposent au minimum les 15 champs Dublin core.

Il est bon de publier le schéma de métadonnées

Métadonnées adaptées au type d'objet

- Pour chaque projet, bien définir la grille de renseignement des métadonnées
- Chaque type de document demande des métadonnées différentes. Bien les définir.
- Conserver le lien entre le document d'origine, les différents fichiers (conservés, diffusés, sauvegardés), les métadonnées, la notice du catalogue d'archives
- Penser à se référer à des listes d'autorité (auteurs, localisations géographiques,...)
- Penser l'aspect multilingue de la plateforme et des métadonnées
- Evolution vers le web sémantique : extraction de mots clés et d'ontologies dans les titres, annotations, transcriptions; OCRisation

Pour aller plus loin :

- BANAT-BERGER F., DUPLOUY L. et HUC C., L'archivage numérique à long terme. Les débuts de la maturité ?, La documentation française, 2009 ;
- HOURCADE J.-C., LALOE F., SPITZ E., Longévité de l'information numérique, les données que nous voulons garder vont-elles s'effacer ?, EDP Sciences, Les Ulis, 2010 ;
- HUC Claude (dir.), Préserver son patrimoine numérique, classer et archiver ses e-mails, photos, vidéos et documents administratifs, Eyrolles, Paris, 2011 ;
- OFFENSTEIN S., La conservation numérique du dépôt légal à la Cinémathèque québécoise, méthodologie de conduite de projet pour un plan de sauvegarde et de numérisation des fonds, Editions universitaires européennes, Sarrebruck, 2010.
- www.piaf-archives.org: Portail international archivistique francophone qui propose un cours en ligne sur la gestion et la conservation des documents sur tous supports.
- Archives générales du Royaume et de Archives de l'Etat dans les Provinces. 2013. « Conserver des documents numériques ». Consulté le 24 juin 2015, http://www.arch.be/docs/brochures/conserver_documents_numeriques.pdf