La cellule Formats

Axes de travail et premiers résultats

#PINFormats

Webinaires PIN - 10 novembre 2020

Composition de la cellule et programme de travail 2019-2020

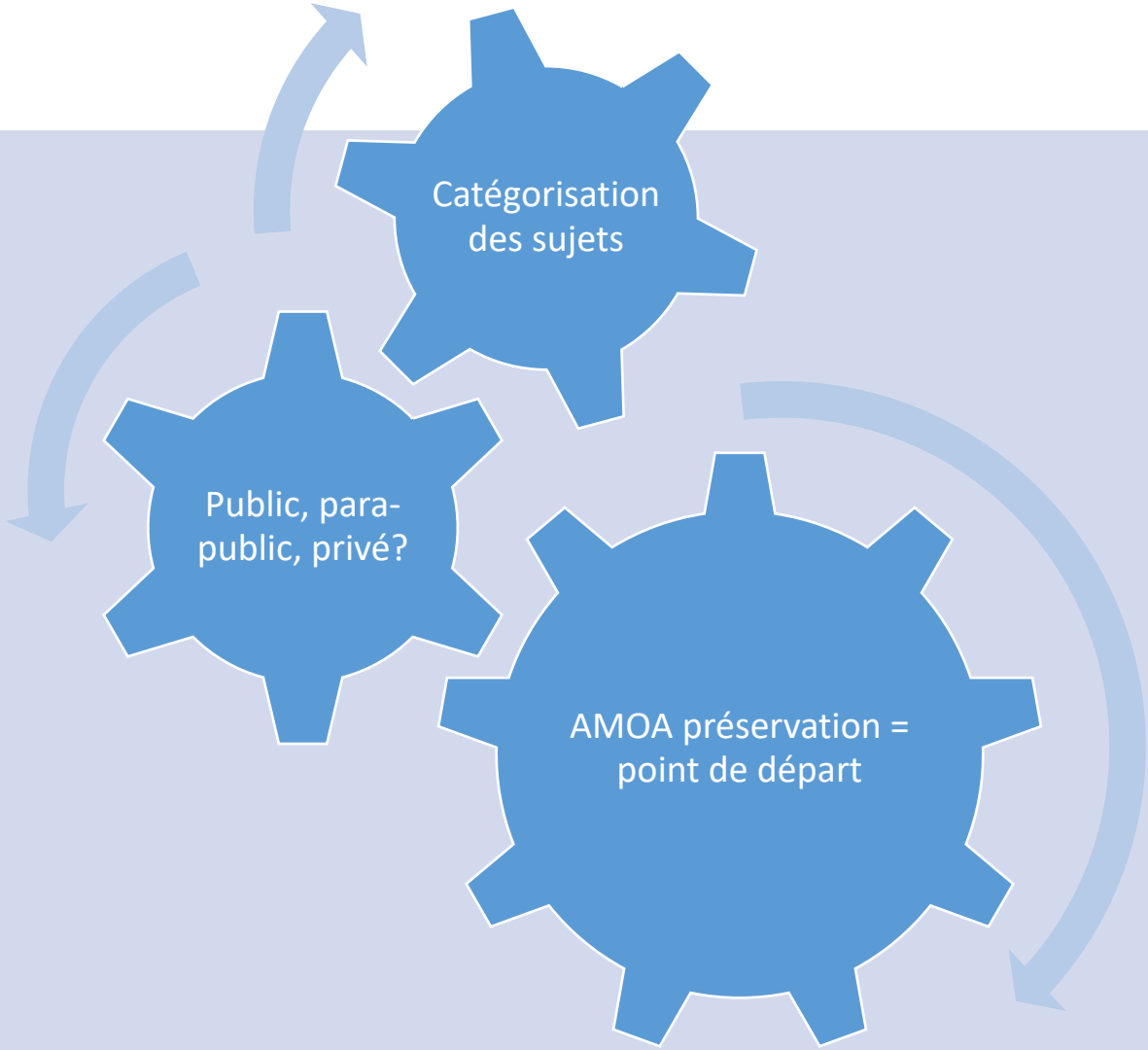
Sous-groupe « Identification des expertises » Présentation Dominique Naud (Siaf)	• Constitution d'un annuaire des expertises en France et à l'étranger
Sous-groupe « Connaissance des formats et critères d'obsolescence et de pérennisation» Présentation Emeline Levasseur (AN)	• Recensement des formats existants, notamment les formats spécifiques • Définition de critères d'obsolescence et de pérennité pour les formats
Sous-groupe « Outils et Corpus » Présentation Thomas Ledoux (BNF)	• Recensement des corpus de fichiers librement réutilisables • Utilisation de Wikidata pour recenser les outils
Sous-groupe « Traduction » Présentation Edouard Vasseur (ENC)	• Traduction de la grille d'évaluation de la NDSA • Traduction du Rapid Assessment Model du DPC • Traduction du Handbook sur la préservation numérique élaboré par le DPC

Identification des expertises

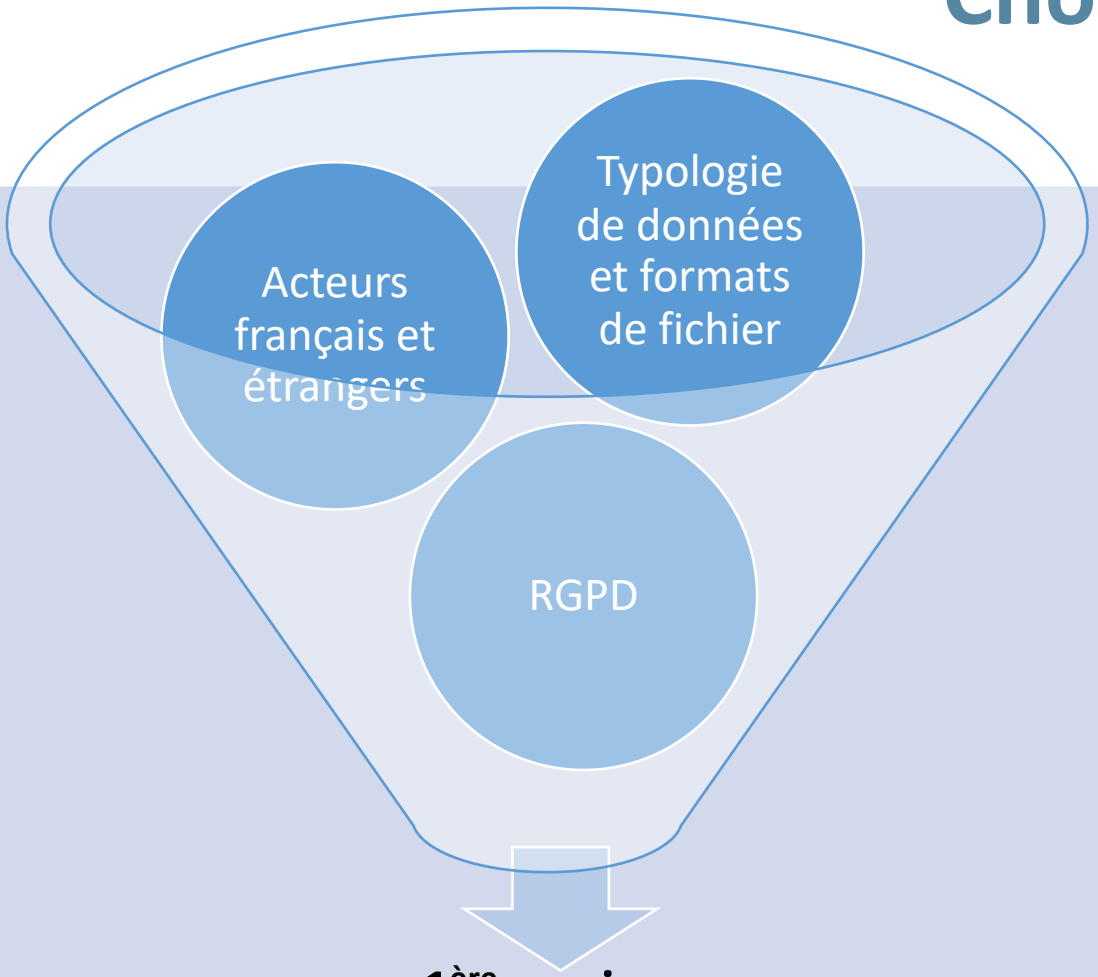
Composé du Ministère des Armées (MINARM), du Commissariat à l'énergie atomique (CEA) et du Service interministériel des Archives de France (SIAF)

- Constitution d'un annuaire des expertises en France et à l'étranger

Evolution 2019-2020



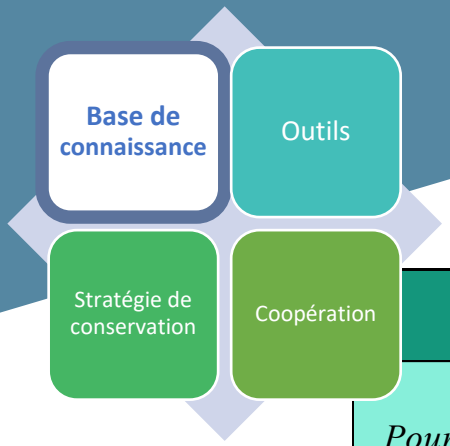
Choix de publication



1ère version

Acteurs français du secteur public

Sujet d'expertise = type de contenu



1ère version publiable

GLOSSAIRE

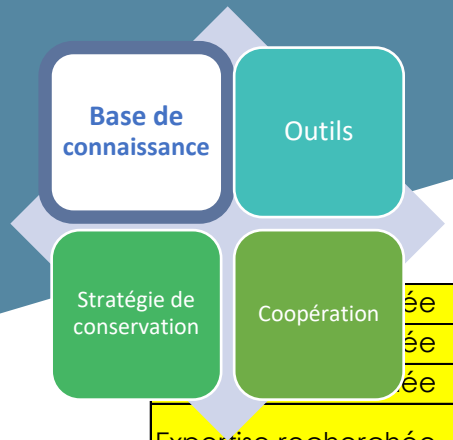
Pour faciliter votre circulation dans cet annuaire et vous permettre d'affiner votre recherche, ***les sujets d'expertises ont été regroupés selon la classification COPTR****, dont voici ci-dessous leur équivalent en français.*

Content type	Type de contenu
Audio	Audio
Binary Data	Données binaires
Container	Conteneur
Database	Base de données
Disk Image	Image disque
Document	Document
EBook	Livres électroniques
Email	Courriel
Geospatial	Géospatial
Image	Image
Project Management Data	Gestion de projet
Research Data	Données de la recherche
Software	Logiciel
Spreadsheet	Feuille de calcul
Video	Vidéo
Web	Web
Not content type specific	Hors classement

Pour toute contribution, et lorsque cela est possible, nous vous invitons à vous servir des listes déroulantes pour améliorer l'uniformité de nos données.

1ère version publiable

Acteurs (personnes physiques ou morales)	[Pays]	Lieu d'exercice actuel (depuis le)	Content type	Précision du sujet	Contact	Références/sources	Public/privé
Centre Informatique National de l'Enseignement Supérieur	***France***	Centre Informatique National de l'Enseignement Supérieur	Image	TIFF/GeoTIFF, PNG, JPEG, JPEG2000, 3D (.ply, .dae)	svp@cines.fr	https://www.cines.fr/archivage/ https://alfresco.cines.fr/alfresco/faces/jsp/browse/browse.jsp	Personne morale secteur public
Institut national de l'audiovisuel	***France***	Institut national de l'audiovisuel	Video	Tout type de format	expertises@ina.fr	https://www.ina.fr/	Personne morale



1ère version publiable

Expertise recherchée		Expertise recherchée	Document	PDF/A
Expertise recherchée		Expertise recherchée	Software	Design (.agdb .anf .axd)
Expertise recherchée		Expertise recherchée	Web	.html .htm .mht
Expertise recherchée		Expertise recherchée	Software	programmation (.cpp), codes sources (.f)
Expertise recherchée		Expertise recherchée	Software	développement (.inc .list), implémentation (.m)
Expertise recherchée		Expertise recherchée	Document	feuilles de style (.css)
Expertise recherchée		Expertise recherchée	Document	mise en page (.fm)
Expertise recherchée		Expertise recherchée	Image	2D/3D (.igs)
Expertise recherchée		Expertise recherchée	Software	Unix (.unix) Archive Unix (.tar)
Expertise recherchée		Expertise recherchée	Document	LaTeX (.tex)
Expertise recherchée		Expertise recherchée	Research Data	HDF
Expertise recherchée		Expertise recherchée	Geospatial	FITS
Expertise recherchée		Expertise recherchée	Image	CAO : STEP (.step .stp .p21)
Expertise recherchée		Expertise recherchée	Image	CAO : .cat
Expertise recherchée		Expertise recherchée	Image	
Expertise recherchée		Expertise recherchée	specific	modeling (BIM) : formats ICF
Expertise recherchée		Expertise recherchée	Not content type specific	Imagerie médicale : format DICOM (.dcm)
Expertise recherchée		Expertise recherchée	Research Data	données brutes scientifiques (applications LIMS)

NOUS AVONS BESOIN DE VOUS !

Et après ?

Acteurs (personnes physiques ou morales)	[Pays]	Lieu d'exercice actuel (depuis le)	Content type	Précision du sujet	Contact	Références/sources	Public/privé

Les acteurs des autres pays

Connaissance des formats et critères d'obsolescence et de pérennisation

Composé des Archives nationales, du ministère de l'Europe et des Affaires étrangères, du Ministère des Armées, de la Bibliothèque nationale de France et du département de la Moselle.

- Contextes et enjeux très différents parmi les participants.
- Plusieurs constats : difficultés de définition, difficultés à formuler des recommandations.

Critères de pérennisation

Questions

- Quels sont les critères de pérennisation adoptés par d'autres institutions ?
- Quelle synthèse peut-on en proposer ?

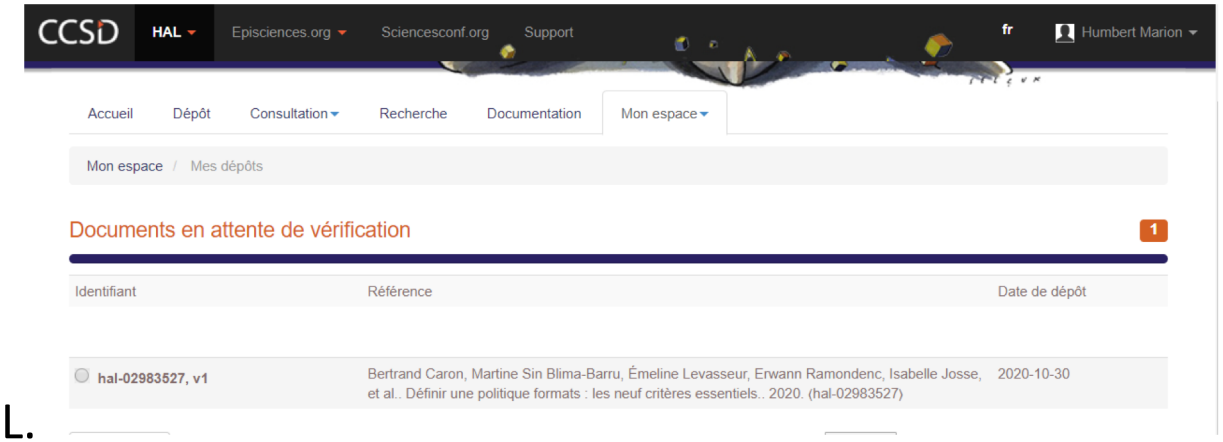
Réponse

Définir une politique formats : les neuf critères essentiels.

CELLULE NATIONALE DE VEILLE SUR LES FORMATS

Sous-groupe « connaissance des formats existants et émergents et définition de critères d'obsolescence et de pérennisation »

Avril 2020



Projet de publication sur HAL.

Connaissance des formats conservés

Questions

- Quels formats sont conservés au sein de nos établissements ?

Réponse

Vers un recensement des formats conservés : échanges, réflexions, documents de travail

CELLULE NATIONALE DE VEILLE SUR LES FORMATS

Sous-groupe « connaissance des formats existants et émergeants et définition de critères d'obsolescence et de pérennisation »

Février 2020

DOMAINES (prendre la catégorisation PRONOM)	Intitulé du format	Extension	Type MIME	PUID	Wikidata	Caractéristiques	Outil particulier pour les manipuler au sein de l'institution	outil pour générer un fichier (en fonction des contextes institutionnels) *facultatif*	Nombre d'objets par institution	Volume en % par institution	Contexte de production (éléments de compréhension propre à chaque institution) *facultatif*	Commentaire sur l'entrée *facultatif*
	OpenDocument Text (ODF)	.odt				Ouvert						
	Portable Document Format (Adobe Systems)	.pdf				Propriétaire ouvert						

Connaissance technique des formats

Questions

- Quelles sont les caractéristiques d'un format précis ?

Réponse

Travaux en cours : TIFF, Mp3, formats bureautiques, messageries.

Composition de la cellule et programme de travail 2019-2020

Sous-groupe « Outils et Corpus »

Composé de la BnF, du CINES, du TGIR HumaNum, de Mintika, du programme VITAM

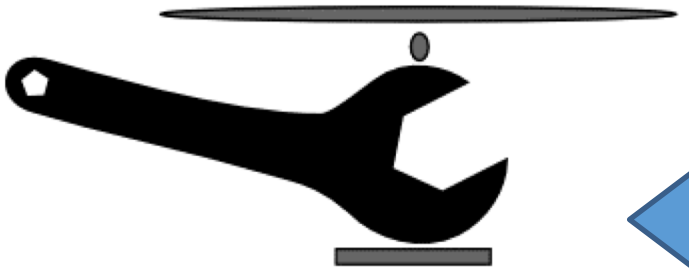
- Recensement des corpus de fichiers librement réutilisables
- Utilisation de Wikidata pour recenser les outils

Recensement des corpus de fichiers

- Première version
- 12 corpus génériques repérés

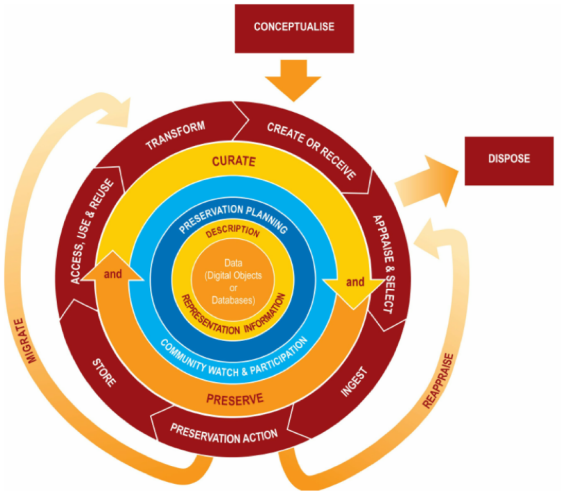
	A	B	C	D	E	F	G	H
1		**Recensement de corpus de fichiers librement utilisables**						
2	Code	**Nom**	**Organisation**	**URL**	**Type**	**Volumétrie**	**Licence**	**Repérage des fichiers valides ou non**
3								
4	CF_01	Format corpus	OPF	https://github.com/openpreserve/format-corpus	Tous		CC0	?
5	CF_02	veraPDF-Corpus	OPF, DualLabs	https://github.com/veraPDF/veraPDF-corpus	PDF		CC-BY 4.0	Oui
6	CF_03	Isartor	PDF Association	https://www.pdfa.org/resource/isartor-test-suite/	PDF/A		Freely downloadable and usable without restriction	Oui
7	CF_04	Bavaria	PDFLib	https://github.com/bfosupport/pdfa-testsuite	PDF/A		Creative Commons Public License	
8	CF_05	Jpylyzer test Suite	OPF	https://github.com/openpreserve/jpylyzer-test-files	JP2000		CC-BY	Oui
9	CF_06	Google image test suite		https://code.google.com/archive/p/imagetestsuite/	Images : TIFF, PNG, GIF, JPEG	143 Mo	CC-BY 3.0	

Registre des outils de préservation



COPTR

Utilisation des travaux précédents



Le modèle du DCC sur le cycle de vie de la préservation numérique

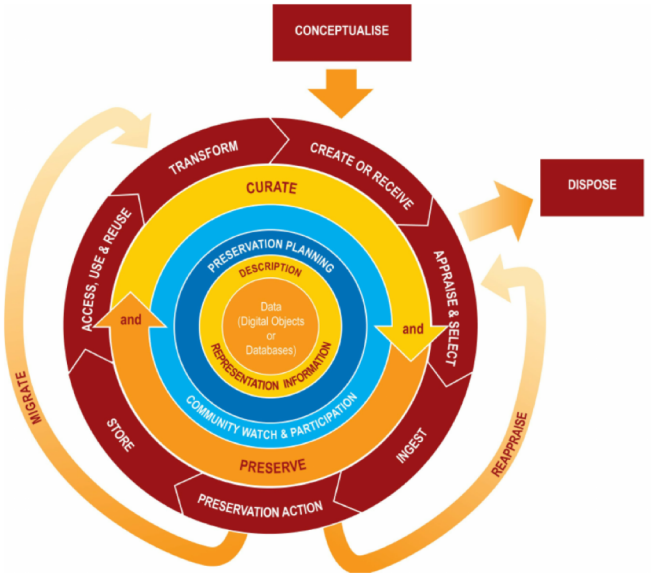
Accès multilingue

Wikidata

Bénéficie de

Wikipedias

Etape 1 : créer les catégories du registre



Cycle de vie du DCC

Chaque étape du cycle de vie est modélisé par une entité WD

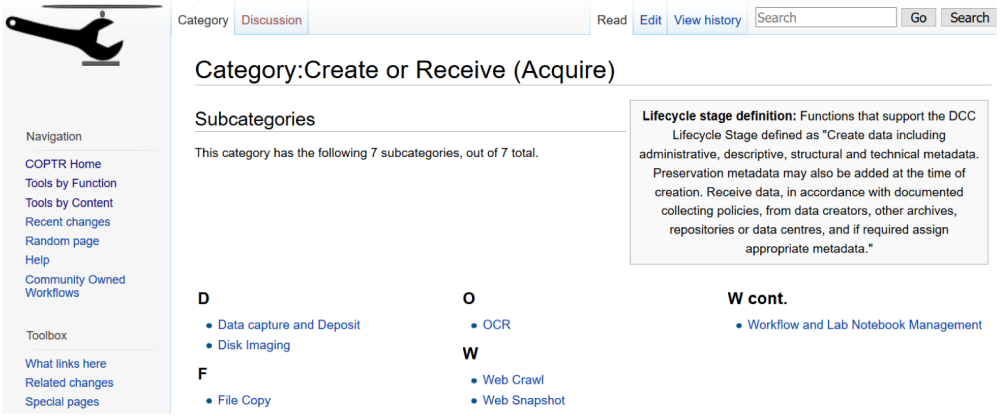
Wikidata

P31 (instance of - *nature de*) : Q714737 (category - *catégorie*)
P361 (part of - *partie de*) : COPTR
P1545 (series ordinal - *rang dans la série*) = 1

rang	qid	etape	stage
1	wd:Q90878156	Accéder, Utiliser et Réutiliser	Access, Use and Reuse
2	wd:Q90887043	Créer ou Réceptionner (Acquérir)	Create or Receive (Acquire)
3	wd:Q89021580	Agir transversalement	Cross-Lifecycle Functions
4	wd:Q90889758	Éliminer	Dispose
5	wd:Q90885460	Verser	Ingest
6	wd:Q90890255	Agir pour préserver	Preservation Action
7	wd:Q90890401	Planifier la préservation	Preservation Planning
8	wd:Q90890485	Stocker	Store

https://w.wiki/N5r

Etape 2 : lier les processus aux catégories



Lier les processus aux catégories

Activités:

- Chercher une entité WD préexistante appropriée
- En créer une nouvelle, le cas échéant

Constat : Très souvent une entité représentant le résultat du processus existe mais pas le processus lui-même.
Par exemple : “disk image file” (Q592312) au lieu de “disk imaging”

Wikidata

NOUS AVONS BESOIN DE VOUS !

t of - facette de) : ?etape

		process	processus
	741970	file copying	Copie de fichier
	4929239	data collection	collecte de données
	d:Q61466324	web crawling	exploration du Web
(Acquire)	wd:Q167555	optical character recognition	reconnaissance optique de caractères

https://w.wiki/N6M

Etape 3 : donner accès aux outils

P31 (instance of - *nature de*)/P279(subclass - *sous-classe*)* Q7397 (software)
P366 (use - *usage*) : ?processus

https://w.wiki/S4q

Les outils sont classifiés
De nouveaux outils sont identifiés et découverts

Wikidata

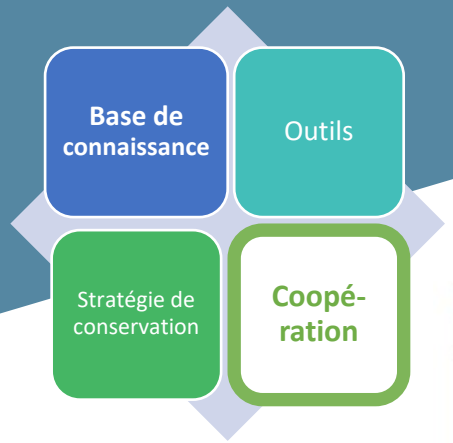
stage	etape	process	processus	tool	toolLabel
Ingest	Verser	file format identification	identification du format de fichier	wd:Q24305922	Digital Record Object IDentification
Ingest	Verser	file format identification	identification du format de fichier	wd:Q16927990	JHOVE
Ingest	Verser	file format identification	identification du format de fichier	wd:Q2858088	Apache Tika
Ingest	Verser	file format identification	identification du format de fichier	wd:Q181820	file
Ingest	Verser	data integrity	intégrité	wd:Q6802825	Md5deep
Ingest	Verser	data integrity	intégrité	wd:Q1932013	Md5sum

Traductions

Sous-groupe « Traduction »

Composé de la BnF, du CD Moselle, de l'ENC, de Mintika et du Ministère des Armées (SHD)

- Traduction de la grille d'évaluation de la NDSA
- Traduction du Rapid Assessment Model du DPC
- Traduction du Handbook sur la préservation numérique élaboré par le DPC



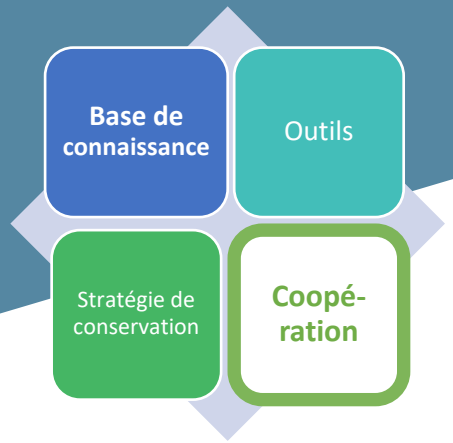
Document 1 : le NDSA Level of Preservation

NDSA

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.

Grille d'évaluation des niveaux de préservation avec les domaines fonctionnels en ligne

https://hal-bnf.archives-ouvertes.fr/hal-02552208v1 et https://hal-bnf.archives-ouvertes.fr/hal-02551807v1



Document 2 : le Rapid Assessment Model de DPC

Grille d'évaluation rapide de la Digital Preservation Coalition

Annexe 1 – tableur DPC RAM

Organisation	
Responsable de l'évaluation	
Date de l'évaluation	
Observations sur le périmètre de l'évaluation	

Niveau stratégique				
	Niveau actuel	**Justification**	**Niveau cible**	**Observations**
A - Viabilité de l'organisation : Gouvernance, structure organisationnelle, dotation en personnel et en ressources des activités de préservation numérique				
B - Politique et stratégie : Politiques, stratégies et procédures qui régissent le fonctionnement et la gestion des archives numériques				

https://www.dpconline.org/digipres/implement-digipres/dpc-ram/ram-french

Document 3 : le Handbook de DPC

Digital Preservation **Handbook**

Explore the Handbook

- Home
- Contents
- Introduction
- Digital preservation briefing
- Getting started
- Institutional strategies
- Organisational activities
- Technical solutions and tools

Digital Preservation Handbook

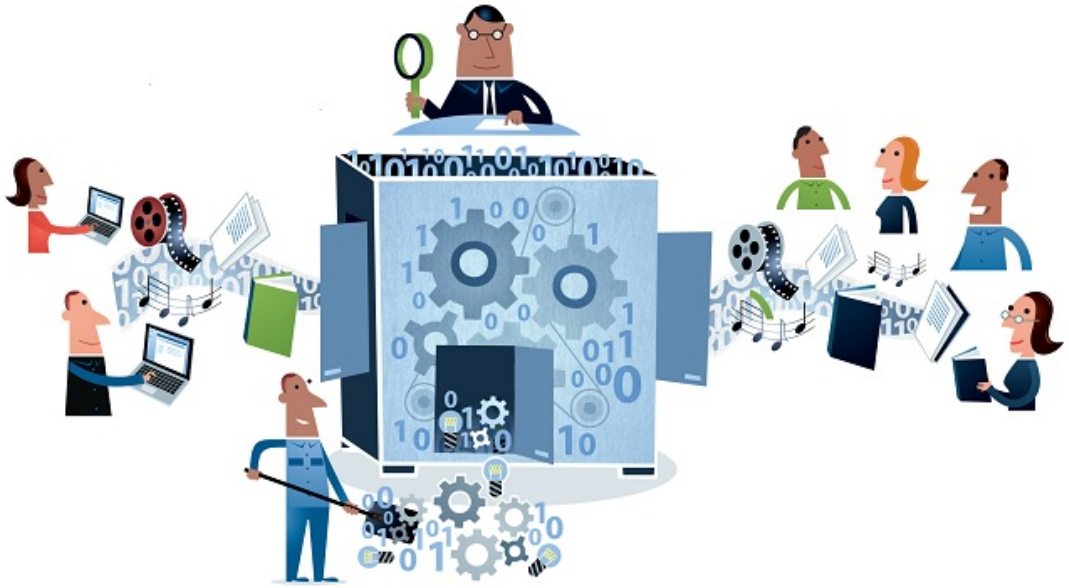


Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Welcome to the revised 2nd edition of the Digital Preservation Handbook. A key knowledge base for

Merci de votre attention !

Contact : archivage.numerique.siaf@culture.gouv.fr