

Retour sur l'étude sur la préservation des archives audiovisuelles des AD34

Assemblée plénière PIN – 8 octobre 2024

Le contexte : Phare 34

- plateforme Héraultaise d'ARchivage électronique
- Dédicée à l'archivage définitif du ressort des AD34
- Projet basé sur les solutions Vitam et Vas

=> Volets « reprise de données » et « préfiguration de la politique de pérennisation » du projet

Un existant volumineux

- Poids des archives audiovisuelles dans l'existant

Typologie	Volume estimé à 2020
archives audiovisuelles	65 To
Archives sonores	1,3 To
Presse	549 Go
archives électroniques administratives	350 Go
TOTAL	67,2 To

- Archives numérisées / archives nativement numériques

Les ambitions de l'étude

- Consolider les processus de versement et de traitement des archives audiovisuelles
- Préfigurer la politique de préservation des AD

La démarche

Audit des formats des archives audiovisuelles déjà collectées

Constitution de l'échantillon
représentatif

Identification et sélection des outils de
validation des formats audiovisuels

Diagnostic des fichiers de l'échantillon

Fiabilisation des processus de versement et traitement des archives audiovisuelles

Méthodologie et modes opératoires pour le traitement et le versement des archives audiovisuelles

Scénarios de pérennisation

Préconisations pour pérenniser les archives audiovisuelles
déjà collectées

Feuille de route associée

Préfiguration de la politique de pérennisation

Stratégies de migration des formats

Rédaction de la V1 de la Politique de pérennisation

Constitution de l'échantillon

- Critères pour l'échantillon représentatif
 - Fichiers ayant posant problème après une première analyse par les AD avec l'outil DROID
 - Au moins 1 fichier de chaque format
 - En provenance de producteurs différents
 - En privilégiant les fichiers plus lourds
- Critères de sélection non retenu : date de création, de modification ou de versement car pas fiable

=> 51 fichiers retenus

Les outils testés

- Identification, sélection et test d'outils d'identification, de caractérisation et de validation des formats adaptés aux archives audiovisuelles
- En privilégiant les solutions Open source, pré-identification de 14 outils parmi les outils utilisés par la communauté des archivistes :
- ❖ Les outils retenus et testés sont de 2 types :
 - Des outils de type « logiciel »
 - DROID
 - JHOVE
 - MediaInfo
 - MP3 Validator
 - Siegfried
 - Des outils de type « boîte à outils » qui reprennent un ou plusieurs logiciels et/ou qui permettent d'aller plus loin dans la préparation d'un versement
 - Facile (DROID, ImageMagick, Jhove etc...)
 - Fits (JHOVE, MediaInfo, DROID, ExifTool, Apache Tika...)
 - Octave (DROID)
- ❖ Outils non retenus et non testés dans le cadre de cette étape :
 - FFMPEG : outil de manipulation et transcodage : utilisation « détournée » par le CINES dans Facile pour faire de la validation
 - Fido (difficultés pour l'installer)
 - Mkvalidator et MediaConch : formats analysés non présents dans les fonds des AD (Matroska, FFv1 et LPCM, WebM)
 - MPlayer : outil d'encodage
 - Tika (mêmes fonctionnalités que MediaInfo mais périmètre moins large d'analyse et utilisation en lignes de commande)

Critères d'évaluation des outils

- Evaluation des outils selon différents critères :
 - ❖ Généralités : version, OS, type de licence, éditeur/développeur, communauté, documentation
 - ❖ Fonctionnalités :
 - Identifie / caractérise / valide
 - Formats couverts
 - Métadonnées produites/récupérées
 - Exports disponibles pour les rapports d'audit
- Performance et capacité à traiter des gros fichiers ou un grand nombre de fichiers
- Ergonomie : langue d'utilisation, interface graphique / ligne de commande, facilité d'installation, facilité d'utilisation

Constats sur les outils

- Il n'existe pas un outil qui couvre toute la chaîne (identification, validation, caractérisation) pour tous les formats :
 - ❖ JHOVE couvre toute la chaîne mais uniquement pour les formats de l'échantillon Wave et MP3
 - ❖ DROID et Siegfried permettent d'identifier tous les formats testés
 - ❖ Les outils de validation sont dédiés à quelques formats (MP3 Validator, Jhove, Facile) et certains formats ne peuvent pas être validés
 - ❖ MediaInfo et Fits permettent de caractériser tous les formats testés
- Les outils ne renseignent jamais sur la qualité du contenu d'un fichier
- L'utilisation de certains outils n'est pas ergonomique (lignes de commande, lenteur, analyse unitaire, lourdeur ...)
- L'exploitation des résultats d'analyse n'est pas toujours adaptée aux besoins des AD (pas d'export possible)
- => Nécessité de combiner les outils entre eux et de concevoir des workflows de traitement

Une boîte à outils (BAO) évolutive

Nom de l'outil	Actions de l'outil		
	Identification	Validation	Caractérisation
MP3 Validator	non	oui	non
MediaInfo	non	non	oui
JHOVE	oui	oui	oui
Siegfried	oui	non (uniquement contrôle de l'extension)	non
Facile	oui	oui	non
DROID	oui	non (uniquement contrôle de l'extension)	non
Octave	oui	non	non
Fits	oui	non	oui

➤ Déjà dans la BAO des AD34 :

- ✓ DROID
- ✓ Siegfried
- ✓ Octave

➤ Ajouté à la BAO suite à l'étude :

- ✓ MediaInfo
- ✓ Fits

Réflexion sur les métadonnées issues de la phase de caractérisation

- Métadonnées techniques plus précises que le format conteneur et qu'il serait nécessaire de conserver en vue des futures migrations
- Sélection des métadonnées à prendre en compte (à partir des exports csv de Mediainfo réalisés lors de l'analyse de l'échantillon)
- Implémentation de ces métadonnées
 - ✓ dans le SEDA (absence de balise correspondante)
 - ✓ dans le contexte Vas (ontologie)

Préfiguration de la stratégie de préservation

- Des formats complexes :
 - Formats conteneurs d'un ou plusieurs flux vidéos et / ou sonores
 - Formats compressés avec des codecs à privilégier à d'autres
 - Formats orientés exploitation / diffusion / archivage (conservation)
 - Qualité / Compression
- Des travaux publiés plutôt en langue anglaise
- Des formats qui peuvent être au choix et selon les institutions et leur stratégie :
 - Privilégiés / préférés
 - Acceptés (tolérés)
 - Non acceptés

Catégorisation des formats et engagement

Catégorie du format	Détail	Engagement des AD34
Accepté sans traitement préalable	Les AD acceptent les archives numériques dans leur format d'origine, sans traitement du fichier originel et mettent en œuvre les opérations nécessaires à leur versement dans la plateforme d'archivage électronique.	Les AD s'engagent auprès du service producteur à réaliser des audits réguliers de ces archives et à mettre en œuvre de manière prioritaire et pour toute la durée de la conservation , les actions nécessaires à la préservation des archives : maintien de la lisibilité, de la forme, du contenu et de ses caractéristiques, des sons et couleurs. Ces formats doivent être privilégiés autant que possible par les services producteurs.
Accepté avec traitement préalable	Les AD acceptent les archives numériques sous réserve d'appliquer un traitement au fichier originel avant le versement (le plus souvent il s'agira d'une migration de format). Elles mettent en œuvre les opérations nécessaires à leur versement dans la plateforme d'archivage électronique.	
Accepté en privilégiant...	Les AD acceptent les archives numériques sous réserve que le format identifié réponde à certaines caractéristiques techniques attendues. Si les caractéristiques ne sont pas respectées, le format passe au statut « toléré » (exemple : pour des archives audiovisuelles, si le codage n'est pas le bon, le format devient « toléré »)	
Toléré	Les AD acceptent les archives numériques dans leur format d'origine et mettent en œuvre les opérations nécessaires à leur versement dans la plateforme d'archivage électronique.	Les archives collectées sont conservées en l'état, sans engagement sur leur maintenabilité dans le temps.
Refusé	Le format est identifié de manière certaine mais les AD ont choisi de le refuser, souvent pour des raisons techniques ou d'intérêt intrinsèque du format.	Les AD n'acceptent pas les archives numériques et ne les versent pas dans la plateforme d'archivage électronique. Des opérations de migration en amont du versement dans un format appartenant aux catégories « Accepté » ou « Toléré » peuvent être étudiées au cas par cas, en fonction de l'intérêt archivistique, des possibilités des technologies et des coûts.
Inconnu	Le format n'est pas identifié dans le registre Pronom.	Les AD n'acceptent pas les archives numériques et ne les versent pas dans la plateforme d'archivage électronique.
Non étudié	Le format n'a pas encore fait l'objet d'une étude par les AD, et n'est donc pas rattaché à une catégorie.	Les AD collectent les archives numériques mais ne les versent pas dans la plateforme d'archivage électronique. Elles réalisent une étude complémentaire afin de déterminer la catégorie du format.

Nos critères de catégorisation des formats

- Selon les volumétries présentes dans le corpus :
 - Format très fréquent : > 150 archives présentes parmi les archives déjà collectées
 - Format fréquent : > 10 archives présentes parmi les archives déjà collectées
 - Format rare : entre 1 et 10 archives présentes parmi les archives déjà collectées
 - Format exceptionnel : 1 archive présente parmi les archives déjà collectées
- Selon l'ouverture du format, c'est-à-dire définir si le format est :
 - Propriétaire : les spécifications sont contrôlées par des organismes privés. Concrètement pour l'utilisateur, cela signifie qu'il est obligé d'avoir le logiciel adéquat pour ouvrir le fichier. Cet aspect propriétaire n'empêche pas que les spécifications soient pour autant publiées.
 - Spécifié : il existe une description de la convention utilisée pour représenter la donnée.
 - Ouvert : cette convention est publique.
 - Normalisé : la convention est adoptée par des organismes de normalisation (ISO, W3C).
 - Standardisé : son utilisation est tellement répandue que le format est considéré comme un standard.
- Selon que l'éventuelle compression soit avec ou sans perte
- Selon les politiques des autres institutions de référence : BNF, Archives nationales, INA pour les archives audiovisuelles
- Selon les erreurs / anomalies rencontrées lors de l'analyse du corpus déjà collecté

Politique des formats audiovisuels

Type de données	Nom du format	Extension	PUID	Type Mime	Acceptés, tolérés ou refusés par les AD34
Contacter les Archives départementales de l'Hérault pour tous les formats non cités ici : archives@herault.fr					
Fichiers audiovisuels	A noter que les Archives départementales de l'Hérault ne prennent pas en charge les fichiers correspondant aux fichiers "projets" issus de logiciels de production audiovisuelles (typologie de fichiers intermédiaires de travail). En plus d'être des formats propriétaires, ces fichiers ne présentent pas d'intérêt historique à être archivés.				
	Audio/Video Interleaved Format	avi	fmt/5	video/x-msvideo	Acceptés sans traitement préalable
	MPEG-4 Media File	mp4	fmt/199	video/mp4 ou application/mp4,	Accepté pour le codage vidéo H264
		m4v, _m4a, _m4p, _m4b, _m4r			Toléré pour les autres codages
	Quicktime	mov	x-fmt/384	video/quicktime	Accepté avec traitement préalable
	MPEG-2 Program Stream	mpeg mpg mod	x-fmt/386	video/mpeg	Accepté pour le codage vidéo H264
	Windows Media Video	wmv	fmt/133	video/x-ms-wmv	Toléré pour les autres codages
	Material Exchange Format	rmf	fmt/200	application/rmf	Toléré
	MPEG-1 Program Stream	mpeg mpg	x-fmt/385	video/mpeg	Accepté pour le codage vidéo H264
		mpg mpeg m1v			Toléré pour les autres codages
	MPEG-1 Elementary Stream	mpg mpeg m1v	fmt/649	video/mpeg	Refusés
	M2TS	m2ts, .MTS	fmt/1055	video/MT2S	Refusés
Fichiers son	Video Object File	VOB	fmt/425	video/mpeg pour Fits	Refusés
	Flash	flv	x-fmt/382	video/x-flv	Refusés - fournir le support dvd
	3GPP Audio/Video File	3gpp	fmt/357	audio/3gpp	Refusés
	MPEG 1/2 Audio Layer 3	mp3	fmt/134	audio/mpeg	Refusés
	Waveform Audio	wav ; .wave	fmt/6	audio/x-wav	Acceptés sans traitement préalable
			fmt/141 fmt/142		Acceptés pour le codage PCM non compressé
	Broadcast WAVE	wav	fmt/703 fmt/704	audio/x-wav	Toléré pour les autres codages
					Acceptés pour le codage PCM non compressé
	Windows Media Audio	wma	fmt/132		Toléré pour les autres codages
	Audio Interchange File Format	aif ou .aiff	fmt/414		Tolérés
					Acceptés pour le codage PCM non compressé
					Toléré pour les autres codages

Les conséquences de l'étude

- Consolidation des processus de versement de préparation mais modification des modes opératoires en attente des choix des métadonnées supplémentaire implémentées dans le SEDA
- Rédaction d'une politique et stratégie de préservation pour les autres typologies à partir de l'étude de préservation des archives audiovisuels
- Boîte à outils complétés par des nouveaux outils d'identification et de caractérisation (mediaInfo, Fts) et par des outils de migration de format (ffmpeg et imageMagick)
- Conversion des formats refusés mais déjà collectés => Mode opératoire pour l'extraction des données et la conversion des formats déjà collectés et refusés

Des réflexions toujours en cours

- Double contrainte technique de traitement liée à la capacité des ordinateurs mis à disposition

- ☐ espace de stockage sur le disque dur
- ☐ Puissance

=> Réflexion sur la méthodologie du processus de versement dans le SAE

Le cas particulier des applications interactives

- Présence dans le corpus de 7 fichiers exécutables (.exe)
 - Identifiant PRONOM fmt/899
 - Type Mime application/vnd.microsoft.portable-executable
 - Format 32 bits
 - Toujours accompagnés d'un ou plusieurs dossiers contenant des fichiers audiovisuels, sonores, iconographiques, bureautiques et/ou flash (swf)
 - Provenant de 5 producteurs différents (1 fonds privés et 5 publics)
- => Des applications interactives Macromédia flash pour 6 d'entre eux et une application quicktime

Merci de votre attention

Evelyne Secondy

Chargée d'archivage numérique

Archives départementales de l'Hérault

Direction des archives publiques, de la conservation et des systèmes d'information

Service Archives de la Réglementation et de la Protection - Archivage électronique



herault.fr

SOLIDARITÉ AU QUOTIDIEN,
ÉCOLOGIE EN ACTIONS