Les Mémos du CR2PA

Les formats d'archivage

Guide de recommandations

Dans un projet d'archivage électronique, la pérennité des documents est un des points importants à prendre en compte.
Les formats des fichiers qu'un système d'archivage électronique accepte doivent donc être analysés précisément pour satisfaire ce critère : le contenu du document doit être accessible et lisible pendant toute la durée de conservation.

L'orientation vers l'utilisation de formats ouverts et standardisés s'impose mais la multitude de formats existants et la diffusion de formats propriétaires ne facilitent pas la tâche.
La politique d'archivage de l'entreprise doit prendre en compte ces situations spécifiques liées à ses activités et son environnement.

Document rédigé par le groupe de travail AMOA – décembre 2014
Ont participé à ce groupe de travail les responsables de politique et projets d'archivage de :
BnF, BNP Paribas, Renault, RFF, Systra, Thales Systèmes Aéroportés.

Les principes

Il n'existe pas de normes imposant des formats précis de fichiers électroniques à respecter pour un Système d'Archivage Electronique (SAE).
La norme NF Z 42-013:2009 (et sa version ISO 14641-1:2012) précise néanmoins qu'il faut utiliser « *de[s] formats normalisés ou standardisés et utilisables librement* ».
Elle détaille ensuite :
Les documents sous une forme numérique reçus pour archivage doivent être conservés dans un format de fichier normalisé ou standardisé dont la spécification doit être accessible librement tout au long de la durée de vie du document. Dans la politique d'archivage, doivent être mentionnés le référentiel normatif associé aux différents types de documents numériques concernés ainsi que les spécifications des formats, afin d'en assurer leur exploitation dans le temps ainsi que la stabilité de leur contenu informationnel.

L'AFNOR a également publié en juin 2010 un Guide d'Application de la NF Z42-013, GA Z42-019. Ce document identifie notamment une liste de formats (« 2.2.6 Choisir un format de fichier pour l'archivage »), basée sur le RGI (Référentiel Général d'Interopérabilité)

Méthodologie

Définition des formats autorisés

Lors de la mise en place d'un projet de SAE, il convient d'identifier et d'analyser les fonds documentaires afin de lister les différents formats des fichiers électroniques candidats à l'archivage. Il convient ensuite d'identifier les usages associés : est-ce un format pour de la relecture simple, pour être récupéré afin de définir une nouvelle version ou pour servir de preuve ?
Le référentiel de conservation est également un document à prendre en compte car la durée de conservation peut être un facteur important dans le choix du format.
Ainsi, pour toute durée inférieure à 5 ans, le risque de non-relecture étant assez faible on peut considérer que tout format est acceptable.
Le poids du fichier peut également intervenir dans la décision de retenir ou pas un format car les coûts du SAE peuvent être impactés (espace de stockage, licence du progiciel liée au volume archivé…). Et même pour un format donné, le poids peut varier suivant les paramètres retenus (taux de compression, gestion des images, version du format retenue…).

Contrôle en entrée & conversion

Le SAE peut contrôler ou pas les formats des archives versées.
Si un contrôle est effectué et si le format identifié ne fait pas partie de la liste établie, Il y a trois possibilités :

- Le versement est refusé
- Le format est soumis à approbation
- Le fichier est automatiquement transformé dans un format pérenne. La mise en place de ce dernier cas est délicate dans le sens où le SAE prend la responsabilité de la transformation. Or l'intégrité du document n'est pas conservée au sens informatique du terme (le fichier en entrée et le fichier transformé n'ont pas la même « empreinte », ils sont différents)

Si aucun contrôle n'est effectué, il y a risque, lors de la restitution, de ne pas pouvoir accéder au contenu.
Toutes ces décisions relèvent de décisions d'entreprise prenant en compte les situations spécifiques liées aux types d'activités.

Quelques recommandations :

- Mettre en place l'outillage nécessaire à la vérification du format en entrée du système
- Conserver les outils permettant d'accéder au contenu des documents (cela peut conduire à conserver une plate-forme spécifique avec les logiciels associés…)
- Stocker également le fichier source en cas de conversion
- Essayer d'avoir un format d'entrée directement pris en charge par le SAE. Pour cela, l'original produit par l'application versante doit être dans un format pérenne. Par exemple, pour un document bureautique (de type Microsoft Word), l'étape de validation devrait se faire directement sur un format PDF ou PDF/A plutôt que sur le format source

Conversion de format durant le cycle de vie

Pour des documents dont la durée de conservation est importante (plusieurs dizaines d'années), Il n'y a pas de garantie que le format utilisé soit toujours lisible durant toute la conservation du document.
Pour se prémunir contre cela, il peut être nécessaire de mettre en place un processus de conversion vers un nouveau format considéré comme plus pérenne.
Là encore, l'intégrité, au niveau informatique, n'est pas conservée.
La norme NF Z 42-013 (« 10.5 Conversion de formats ») recommande que « *La réalisation et les caractéristiques de la conversion doivent être contrôlées et enregistrées dans le journal des évènements* ».
Les fichiers au format d'origine et au format converti doivent être conservés.

L'OAIS (ISO 14721:2012) parle de migration/transformation/émulation pour garantir la continuité d'accès au contenu.

Recommandations de formats

Précisions

La liste des formats présentée par la suite n'est qu'une recommandation. Il existe de nombreuses autres références publiées par différents organismes. Et bien qu'il y ait consensus sur certains formats, des divergences peuvent apparaître sur d'autres. La date de publication de ces choix est à prendre en compte car c'est un domaine qui évolue rapidement (cf. les dernières versions du PDF/A).
Ce « registre de formats » doit être régulièrement révisé par l'entreprise pour prendre en compte de nouveaux formats, interdire des formats obsolètes, revoir la politique de conversion…

Les principaux formats

Pour chaque typologie d'usage, un format recommandé est mis en avant (**EN GRAS, ROUGE ET SOULIGNÉ** dans le texte) dans la suite du document.

Bureautique

Le PDF/A est un format dédié à l'archivage (sous-ensemble du PDF avec un contenu autonome et sans fonction dynamique). Ses différentes versions sont des normes ISO (publiée en 2005 pour le PDF/A-1, 2010 pour le PDF/A-2 et 2012 pour le PDF/A-3).
C'est le format recommandé pour tout document non structuré de type bureautique.

Nom du format	Périmètre d'utilisation	Organisation responsable	Référence précise et détails	Commentaires
TXT	Texte			Format le plus simple possible mais également le plus pauvre !
XML	Texte structuré	W3C	http://www.w3.org/XML	Les Schémas ou DTD associés doivent être le plus standard possible et documentés. Et également archivés. Le préférer au format CSV

				(information non typée, non auto-porteuse)
PDF/A PDF	Texte (fond et forme)	ISO	La norme PDF/A est un sous-ensemble du PDF adapté à l'archivage. Il existe 3 versions (avec différents niveaux de conformité dans chacune d'elle) : - PDF/A-1 (ISO 19005-1:2005) est basée sur le format PDF v1.4 - PDF/A-2 (ISO 19005-2:2010) est basée sur le format PDF v1.7 (lui-même ISO 32000-1:2008). Permet d'utiliser la transparence, d'insérer d'autres fichiers PDF/A, supporte le JPGEG2000 - PDF/A-3 (ISO 19005-3:2012) est basée sur le format PDF v1.7. Il permet d'insérer des fichiers de tout format	Eviter le format Office (même OpenXML). Le format **PDF** peut être éligible (plus léger en poids qu'un PDF/A) en prenant des précautions (informations statiques, polices standards). La recommandation pour l'archivage est d'aller vers le **PDF/A-2.** Attention, le PDF/A a quelques restrictions.

Cas des emails

Il n'existe pas de format standardisé pour l'archivage des mails. Un email est un format complexe composé d'une partie structurée (Emetteur, Destinataire, Objet…), d'un corps et d'éventuelles pièces jointes.

Nom du format	Périmètre d'utilisation	Organisation responsable	Référence précise et détails	Commentaires
EML	Courrier		Compatible de RFC 822 (https://tools.ietf.org/html/rfc822)	Pas de format pérenne standard idéal. Le format EML est composé de textes ainsi que de pièces jointes. Il est nécessaire de transformer chaque pièce jointe dans un format pérenne. Eviter le .MSG, format aux spécifications publiques mais de technologie Microsoft

Cas des sites Web

L'archivage d'un site Web peut être complexe car les contenus peuvent être hétérogènes (textes dans des fichiers HTML, photos, sons, films…), riches et dynamiques (javaScript, Flash, formulaires, base de données…) d'où des limites techniques lors d'un archivage d'un site. Néanmoins, les principes restent le même : valider le périmètre des informations à archiver, utiliser des formats pérennes.

Il est à noter que la BnF a un service d'archivage de sites (« dépôt légal des sites Web »).

Nom du format	Périmètre d'utilisation	Organisation responsable	Référence précise et détails	Commentaires
WARC	Site Web	ISO	ISO 28500:2009, Information and documentation -- WARC file format	Le format Web ARChive permet d'avoir une représentation d'un site Web

Cas des bases de données

Avant d'archiver une base de données, il est important de se poser la question de l'objectif de l'archivage. Une base de données étant exploitée par une application, il est nécessaire d'étudier cette application au niveau fonctionnel afin d'identifier les besoins d'archivage.
Dans certains cas (une base de résultats par exemple), le format SIARD est un bon candidat pour la conversion d'un modèle de données dans un format pérenne (XML). Il est cependant indispensable d'archiver le contexte associé pour être capable d'exploiter ces données lors d'une restauration.

Nom du format	Périmètre d'utilisation	Organisation responsable	Référence précise et détails	Commentaires
XML		W3C	http://www.w3.org/XML	Se poser la question sur l'archivage de l'application utilisant cette base de données. Aller plutôt vers une extraction lisible (=PDF) d'informations élémentaires via un processus contrôlé. Le modèle de données (DTD ou schéma) doit être documenté
SIARD		Archives fédérales suisses	SIARD est l'abréviation de Software-Independent Archival of Relational Databases. http://www.ech.ch/vechweb/page?p=dossier&documentNumber=eCH-0165	D'origine Suisse, ce format est une représentation XML "standardisée' d'un modèle de données relationnel.

Image

Nom du format	Périmètre d'utilisation	Organisation responsable	Référence précise et détails	Commentaires
TIFF	Image bitmap image scannée	Adobe	Tag(ged) Image File Format V6 ISO 12639:1998 http://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf	Version Gr 4 pour les scans N&B. Attention au poids des fichiers
PNG	Image bitmap dessin	ISO	Portable Network Graphics ISO/IEC 15948:2004	Alternative au format GIF (Graphics Interchange Format) plus limité
JPEG	Image bitmap photo	ISO	Joint Photographic Experts Group ISO/IEC 10918, ISO/IEC 14495	Attention au niveau de compression qui peut détériorer le contenu
JPEG 2000	Image bitmap photo	ISO	Joint Photographic Experts Group SO/IEC 15444-1:2004	Compression avec ou sans perte, Plus performant (qualité/poids) que le JPEG mais moins répandu et plus complexe
SVG	Image vectorielle	W3C	Scalable Vector Graphics 1.1 http://www.w3.org/TR/2011/REC-SVG11-20110816/	Format de référence pour le vectoriel
CGM	Image vectorielle	ANSI	Computer Graphics Metafile ISO 8632:1992	Conservé à titre historique

Audio-vidéo

Un fichier audio ou vidéo est constitué de 2 parties : un conteneur (l'enveloppe) + un codec (partie codage/décodage du contenu proprement-dit).

Nom du format	Périmètre d'utilisation	Organisation responsable	Référence précise et détails	Commentaires
FLAC	Audio	Xiph.org	Codec FLAC (Free Lossless Audio Codec) https://xiph.org/flac/index.html	Compression sans perte. Attention, peu répandu
WAV/LPCM	Audio	IBM/Microsoft	Conteneur WAV avec codec audio LPCM (Linear Pulse Code Modulated Audio)	Format sans compression. Le WAV étant limité à 4Go, utilisez RF64 si besoin
MPEG-4/AAC	Audio	ISO	Conteneur MPEG-4 (Moving Picture Experts Group) avec codec audio AAC (Advanced Audio Coding) ISO/IEC 14496-3:2001 - Partie 3	Compression avec perte mais meilleure qualité que le MP3
MPEG-4/AVC	Vidéo	ISO	Conteneur MPEG-4 (Moving Picture Experts Group) avec codec vidéo AVC (Advanced Video Coding) ISO/IEC 14496-10:2003 - Partie 10	AVC est également connu sous le nom H.264
AVI/H.264	Vidéo	Microsoft (AVI)	Conteneur AVI (Audio Video Interleave) et codec vidéo H.264	
MXF/MJPEG 2000	Vidéo	SMPTE (Society of Motion Picture and Television Engineers) pour MXF	Conteneur MXF (Material eXchange Format) et codec vidéo MJPEG 2000 (Motion JPEG 2000)	Récent et efficace
Ogg/Theora	Vidéo	Xiph.org	Conteneur Ogg avec codec vidéo Theora	Formats libres, ouverts, sans brevets mais peu répandus

Documents techniques/métier

Il est difficile de préconiser un format unique sur ce thème :

- Les formats sont très hétérogènes
- Les besoins ne sont pas les mêmes

Le principe est de faire une analyse détaillée avec les experts du métier afin d'identifier les formats candidats.
Il peut arriver de se trouver devant une difficulté car le format propriétaire n'a pas d'équivalent ouvert et/ou normalisé où un équivalent existe mais avec une perte d'information. Dans ce dernier cas, différentes options sont possibles :

- Ne pas archiver et conserver le fichier source dans l'environnement de production (risque à mesurer car il est nécessaire de maintenir son environnement de production ou passer par des systèmes d'émulation ou de virtualisation)
  Une variante est d'archiver le format source et les moyens associés
- Archiver dans le format pérenne (transformation avec perte) avec les informations manquantes dans un autre fichier complémentaire

Voici une illustration autour des dessins techniques :

Nom du format	Périmètre d'utilisation	Organisation responsable	Référence précise et détails	Commentaires
DXF	Dessin technique	Autodesk	Format d'échange des fichiers de dessins AutoCAD, logiciel édité par AutoDesk	Le préférer au format propriétaire DWG bien qu'il existe des initiatives libres pour lire/écrire des fichiers à ce format
STEP AP242	Dessin technique	ISO	ISO/DIS 10303-242:2014 STEP (STandard for the Exchange of Product model data)	Norme sortie le 1er décembre 2014

Sites/Documents de référence

- France : Formats et supports, Archives de France, 2010-2012
- France : Groupe PIN, Pérennisation des Informations Numériques
- France : CINES, Centre Informatique National de l'Enseignement Supérieur
- Suisse : Formats de fichiers adaptés à l'archivage électronique à moyen et long terme, état de Genève, octobre 2011
- Canada : Registre local de formats numériques, Bibliothèque et Archives Canada, octobre 2010
- Royaume-Uni : Selecting file formats for long-term preservation, National Archives, 2008
- Etats-Unis : Recommended Format Specifications, Library Of Congress, juin 2014
- Etats-Unis : Revised Format Guidance for the Transfer of Permanent Electronic Records, National Archives and Records Administration, mars 2014
- PDF, PDF/A : PDF Association, PDF Association