

# Comment choisir les bons formats ?

Philippe Martin, s'exprimant au nom de l'Aproged  
Administrateur de l'Aproged, responsable du Pôle Normalisation  
Directeur Associé au Bureau van Dijk Ingénieurs Conseils

## Introduction

Le choix des formats d'archivage numérique est un élément technique clé d'une politique d'archivage. Ce choix doit garantir la pérennité de l'accessibilité aux contenus archivés en minimisant les coûts des opérations d'archivage et de maintenance des archives sur le long terme.

Au cours des dernières années, une meilleure analyse des contraintes et l'arrivée de nouvelles options permettent aujourd'hui de proposer des orientations claires sur les choix possibles et les précautions à mettre en œuvre pour démarrer et réussir un projet d'archivage électronique à long terme.

Cette conférence présente successivement : les postulats de base pour le choix des formats d'archivage, les critères de choix, les options possibles. Un zoom est proposé sur les différents formats Adobe PDF qui sont déjà normalisés ou en voie de l'être. Quelques repères méthodologiques sont préposés à la fin de la présentation.

## Postulats de base

L'objectif d'une solution d'archivage électronique est de garantir la conservation et l'accessibilité des contenus archivés sur le long terme. Cette première assertion contient plusieurs idées qui doivent être explicitées.

L'accessibilité aux contenus implique d'offrir plusieurs services.

Tout d'abord un moyen de recherche qui reposera essentiellement sur les métadonnées associés à chaque document archivé. Sur le plan technique, ces métadonnées pourront être contenues dans chaque document archivé et/ou dans un référentiel externe (base de données, instrument de recherche). En complément, la recherche pourra porter sur le contenu textuel des documents archivés. Ce qui suppose qu'il soit possible d'indexer ce contenu, donc que le document soit codé en mode caractères ou que sa version image soit convertie en mode caractères au moyen d'un logiciel d'OCR (Reconnaissance Optique de Caractères).

Le document retrouvé doit pouvoir être lu sur écran ou imprimé pour qu'un humain puisse en exploiter le contenu. Cela suppose donc que le format d'archivage puisse être interprété par un système informatique pour restituer une version lisible.

Face à l'évolution constante des systèmes informatiques, la garantie d'une accessibilité à long terme implique d'utiliser des formats d'archivage dont toutes les spécifications sont publiées afin d'être capable soit de créer des outils d'affichage fonctionnant dans des environnements informatiques futurs inconnus aujourd'hui, soit de faire migrer les contenus archivés vers des formats nouveaux compatibles avec les environnements futurs. La création de normes internationales rendant publiques les caractéristiques des formats numériques est un facteur essentiel pour garantir le maintien de l'accessibilité à long terme.

L'utilisation de formats d'archivage normalisés permet de s'affranchir des caractéristiques propres aux systèmes de création des documents archivés. Il convient cependant de vérifier que les formats d'archivage choisis n'entraînent pas une dégradation de la qualité de la restitution ou des fonctionnalités offertes.

Le choix de formats d'archivage numérique pérennes renvoie au second plan le problème du choix du support de conservation des données archivées. Il est aujourd'hui admis que la conservation à long terme des données peut être réalisée sur des supports informatiques réinscriptibles à condition de mettre en œuvre des procédures techniques garantissant la sécurité des données : enregistrement des données sur plusieurs supports et sur plusieurs systèmes répartis sur des sites distants, contrôle des accès aux systèmes d'information.

Enfin, tout ce qui précède repose sur l'hypothèse implicite que les données sont conservées par des systèmes permettant la recherche et la consultation des documents sur un poste de travail informatique conventionnel actuel ou futur. Les solutions consistant à enregistrer les documents sous forme d'images sur des supports micrographiques ne sont pas prises en compte dans la présente conférence, quelles que soient leurs avantages particuliers.

## Critères clés pour le choix

Plusieurs points fondamentaux doivent être étudiés lors de la conception d'une solution d'archivage et, en particulier, pour le choix de formats d'archivage.

La nature de l'information à archiver (le contenu) est le premier point à analyser. Il peut s'agir : de texte, de graphiques statiques (illustrations, dessins industriels) ou de graphiques animés en 2D ou en 3D, d'images (photos à tons continus), de son (musique, discours, messages vocaux), de vidéo (analogique ou numérique) et de contenus mixtes. Pour chaque type d'information, un ou plusieurs formats numériques sont utilisables pour l'archivage.

Ces données peuvent être provenir de différentes sources qui déterminent le format des données disponibles pour l'archivage.

Actuellement, les outils bureautiques sont une source majeure de données susceptibles d'être archivées. La multiplicité des logiciels et des conditions d'utilisation créent une situation complexe dans la perspective d'un projet d'archivage organisé à partir des données produites.

L'archivage numérique des documents disponibles uniquement sur support physique (papier, film, etc.) implique la numérisation des documents avec un numériseur adapté aux caractéristiques physiques du support disponible. Des traitements complémentaires des images obtenues peuvent être nécessaires pour en réduire le volume (compression) et en extraire le contenu textuel (OCR / LAD).

Certains outils spécialisés produisent des données qui vont appeler des solutions spécifiques et des formats d'archivage adaptés : messagerie électronique, logiciels de CAO et DAO, appareils de photo numérique, appareils de prise de sons et caméra vidéo.

Enfin l'archivage des bases de données pose un problème majeur.

Pour ce dernier cas, les bases de données, mais également pour les autres, il faudra choisir entre conserver le contenu (un flux de données représentant toute l'information utile) ou conserver le document présentant l'information sous une forme plus habituelle (mise en page, logo, etc.). Par exemple, faut-il archiver le flux de données permettant d'imprimer un ensemble de factures ou les factures elles mêmes avec toutes données redondantes (mentions légales, fonds de page, etc.). Le volume archivé peut varier d'un facteur 10 selon l'option choisie.

La portée des services à rendre en matière d'accessibilité est également un point structurant. S'agit-il simplement de permettre la lecture des documents à l'écran ou en sortie d'imprimante ou faut-il répondre à d'autres besoins tels que la recherche sur le contenu textuel. Dans le premier cas, une simple numérisation en mode image sera satisfaisante, dans le second cas, il faudra faire un traitement OCR, applicable aux documents imprimés, mais pas aux documents manuscrits.

La gestion des métadonnées dans le processus d'archivage est une question clé. Certains formats permettent l'intégration des métadonnées directement dans le fichier des documents. C'est le cas des formats reposant sur le codage en XML. Si cette intégration n'est pas possible, l'accès aux documents passera par la consultation d'une base de données externe.

Le tableau ci-dessous résume les combinaisons possibles.

Contenu Source	Texte	Graphique	Image	Son	Vidéo	Mixte
Bureau- tique	Documents bureautiques			Insertion de contenus audiovisuels dans les documents bureautiques		
Numéri- sation	Numérisation en mode image de tout support physique			Conversion de sources analogiques		
Outils spécialisés	Messagerie électronique	Plans CAO 2D et 3D	Appareils photos	Appareils de prise de sons et caméras vidéo		
Base de données	Fichiers de texte	Données CAO	Contenus binaires			

## Formats candidats pour les documents bureautiques

Plusieurs formats d'archivage sont utilisables pour des documents produits avec des outils bureautiques. Leurs caractéristiques répondent plus ou moins bien aux attentes définies dans la section précédente.

PDF/A est basé sur le format Adobe PDF 1.4. Les spécifications de ce format ont été publiées sous forme de norme ISO validée en Septembre 2005. Les caractéristiques fonctionnelles et techniques de ce format en font une solution sans équivalent pour le moment. Ce format est présenté en détail ci-dessous.

XPS est le format concurrent développé par Microsoft (XML Paper Specifications). Ce n'est pas encore une norme ISO et sa portabilité est encore limitée à l'environnement Windows.

TIFF G4 est le format image dérivé du format des télécopies. La compression G4 est réversible sans perte de données. Sans être une norme ISO, il constitue un standard de fait largement utilisé pour la numérisation des documents. Le contenu textuel des images retraitées par un logiciel OCR peuvent être indexées. D'autres formats plus modernes, tels que JBIG2, apportent des taux de compression plus importants.

JPEG est un format de numérisation et de compression conçu pour les photos et largement utilisé depuis l'essor de la photo numérique. Le format est normalisé (ISO 10918). Des artefacts de numérisation peuvent apparaître lorsque ce format est utilisé pour des graphiques au trait et pour du texte. JPEG 2000 (ISO 15444) apporte une compression plus importante.

.DOC est le format natif du logiciel de traitement de texte Word développé par Microsoft depuis de nombreuses années. Le logiciel est très répandu et bien connu, ce qui ne signifie pas que les documents produits sont homogènes en terme de codage. Les versions diffèrent au fil du temps et des plates-formes de création. De ce fait, ce format est diamétralement opposé aux attentes exprimées pour définir un format d'archivage pour le long terme. En revanche, de nombreuses solutions existent pour convertir un fichier natif Word vers des formats plus adaptés au problème de l'archivage : PDF/A, XPS, voire XML.

XML (Extended Markup Language) est dérivé de la norme SGML ISO 8879 :1986. Cette norme définit une grammaire permettant de construire un système de balisage logique. En pratique, XML peut servir de format d'échange de données structurées entre deux systèmes. XML peut également être utilisé comme format d'archivage de contenus structurés quelconques. Cependant, pour conserver la présentation physique d'un document (polices de caractères, mise en page, etc.) il est indispensable d'appliquer une feuille de styles aux données XML. En d'autres termes, considéré de façon isolée, le format XML est adapté pour l'archivage d'un contenu structuré, mais pas pour la restitution d'un document mis en page.

Le tableau ci-dessous permet de comparer les propriétés des formats qui viennent d'être présentés.

	PDF/A	XPS	TIFF G4	JPEG	.DOC	XML
Norme ISO	✓	X	X Standard de fait	✓	X	✓
Fidélité restitution	✓	✓	✓	✓	X	X Feuille de styles
Polices caractères	✓	✓	✓ Pixels	✓ Pixels	X	X Feuille de styles
Texte indexable	✓	✓	X Avec OCR	X	✓	✓
Données structurées	✓	✓	X	X	✓	✓
Multi plate-forme	✓	X	✓	✓	✓ Disponibilité polices	✓
Afficheur gratuit	✓	✓ Windows uniquement	✓	✓	X	✓

## Formats utilisables pour l'audiovisuel

Les informations fournies pour ce domaine sont volontairement limitées car ces questions sont largement développées dans l'atelier 2 consacré aux supports d'archivage.

## Autres formats utilisables

### Texte seul

Ce format est utilisable pour archiver des contenus textuels sans aucune mise en forme. Sont exclus : les changements de polices de caractères, les mises en valeur (gras, souligné, italique), les effets de mise en page (alinéa, retrait, pied de page), les tableaux, etc.

Trois normes ISO définissent des alphabets de codage : ISO 646 pour l'Ascii 7 bits, ISO 8859 pour l'Ascii 8 bits (apports des lettres accentuées et des signes diacritiques), ISO 10646 pour l'Unicode (Ascii 16 à 32 bits).

### Formats spéciaux

Signalons, à titre d'exemples, quelques formats spéciaux

EDIFACT définit un format pour l'échange de données constituant des factures électroniques (ISO 9735)

La TEI (Text Encoding Initiative) est un projet international visant à mettre au point des directives pour l'élaboration et l'échange de documents électroniques à des fins de recherche érudite, et pour répondre aux besoins les plus variées des industries de la langue en général. Initialement basé sur SGML, le modèle a évolué vers XML. Ce n'est pas une norme ISO.

ALTO est un format conçu pour la numérisation des documents anciens en vue de leur diffusion sur Internet. Ce format très avancé combine une description de la structure physique de chaque page sous forme de données XML, une image de la page complète et de ces composants en PDF transparent et le contenu textuel des composants. Cette solution permet l'affichage de l'image de la page, l'indexation du contenu textuel, et la sélection directe des contenus dans l'image affichée.

## Et les métadonnées ?

Les métadonnées contiennent des informations sur les documents : identification, origines du contenu, descripteurs pour la recherche, etc.

Certains formats permettent d'enregistrer certaines métadonnées directement dans le fichier du document numérique. C'est le cas des formats PDF/A et XPS, mais aussi du format TIFF. Les photos provenant d'un appareil de photos numérique contiennent des métadonnées IPTC (conditions de prise de vue, date, géo positionnement). Cette intégration facilite la gestion des métadonnées.

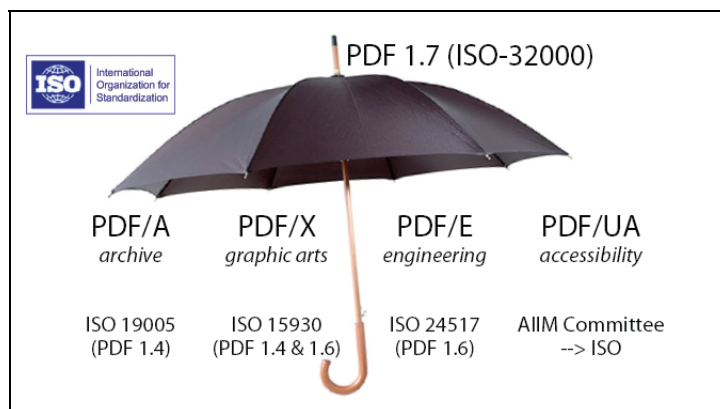
Lorsqu'elles ne sont pas intégrées dans les documents, les métadonnées doivent être gérées dans une base de données de type catalogue. Les fichiers des documents archivés sont accessibles via un identifiant retrouvé dans la base de données.

Dans les deux cas, les métadonnées doivent être soigneusement conservées et archivées avec autant de précautions que les documents eux mêmes.

## Zoom sur la normalisation des formats PDF

La société Adobe, éditrice du format de description de page PostScript, a développé le format PDF (Portable Document Format) à partir de 1993. Le but était de fournir un format de description de page garantissant la restitution fidèle de la mise en page originale des documents créés avec divers logiciel sur des plates-formes différentes. La mise à disposition gratuite d'un afficheur – Acrobat Reader – a été un élément essentiel de cette stratégie éditoriale.

Pour renforcer la position de ses solutions, la société Adobe a entrepris de publier les spécifications de ses formats sous forme de normes ISO à partir de 2003. Ce mouvement est maintenant très avancé comme le montre le schéma ci-dessous.



## PDF/A ISO 19005

Basé sur PDF 1.4., ce format est conçu pour l'archivage des documents bureautiques. Le format peut être produit à partir de sources diverses (bureautique, numérisation). Il supporte plusieurs méthodes de compression normalisées et intègre des métadonnées dont la signature.

Pour assurer une réelle pérennité à long terme, la norme exclue tous les éléments susceptibles de rendre le fichier incomplet ou inaccessible. Sont donc exclus : les scripts, les liens hypertextes, les contenus multimédia, le chiffrement et les codes de sécurité. L'algorithme JPEG 2000 est également exclu. En outre, pour garantir une réelle portabilité, les polices de caractères doivent être intégrées dans le fichier, ce qui en augmente considérablement le volume.

La norme actuelle définit deux niveaux de conformité.

- PDF/A-1a implique le respect de la norme complète, y compris de la structure logique et des métadonnées
- PDF/A-1b est une forme allégée qui garantit la préservation de la lisibilité et la restitution à l'affichage et à l'impression.

Cette norme fait l'objet d'importants travaux dans les groupes de travail de l'ISO. Des évolutions significatives sont attendues dans les prochaines années.

En 2009-2010, PDF/A-2 devrait apporter la compatibilité avec les formats PDF 1.5 et 1.7 et autoriser la compression JPEG 2000.

Par la suite (2012 ?), PDF/A-3 devrait intégrer la signature électronique et la gestion des ressources externes autorisant la mutualisation des polices de caractères.

La conception d'un plan de migration vers PDF/A est un projet informatique à part entière. Il convient de choisir les outils de conversion et de validation adaptés en fonction des sources disponibles et des volumes à traiter.

*Bibliographie : le livre intitulé « PDF/A : L'essentiel » co publié en octobre 2008 par l'Aproged et le PDF/A Competence Center présente de façon détaillée tous les aspects de ce format.*

## Les autres normes PDF

### PDF/X

Le format PDF/X fait est publié dans la norme ISO 15930. Il concerne les échanges de fichiers d'impression dans l'industrie des arts graphiques. Ce standard couvre les aspects suivants : gestion des images hautes résolutions transmises avec les mises en page, gestion des profils colorimétriques et de la transparence, gestion des polices de caractères comme des ressources externes. Cette dernière fonctionnalité devrait faire partie de la norme PDF/A-3.

### PDF/E

Ce format concerne les échanges de plans numériques. Il a été publié dans la norme ISO 24517-1 validé en mars 2008. Ce format intègre des fonctionnalités spécifiques des logiciels de CAO : gestion des couches multiples (layers), gestion des versions, gestion des éléments en 3D. Il est important de souligner que cette nouvelle norme vise uniquement l'échange de plans numériques et non pas l'échange de données CAO.

## ISO 32000

La norme ISO 32000 résulte de la publication des spécifications techniques du format PDF 1.7. Ce format offre de nombreuses possibilités, en particulier sur les composants qu'il peut gérer : animations 2D et 3D, multimédia. Dans son état actuel, la norme définit plutôt un conteneur pouvant intégrer différents types d'objets dont les spécifications ne sont pas encore toutes normalisées.

## Repères méthodologiques

Pour terminer cet exposé technique, il convient de souligner que le choix des formats doit s'insérer dans une réflexion plus large de conception d'un projet d'archivage électronique qui devra répondre aux exigences d'une politique d'archivage définie au niveau de l'entreprise ou de l'institution.

Cette réflexion sur les formats doit passer par les étapes suivantes :

- Recenser et caractériser les contenus à archiver : typologie des documents et des flux à conserver, sources des données et mode d'obtention de ces données ;
- Définir les fonctionnalités attendues en termes de services à rendre : simple consultation sur écran avec quelles fonctions, impression, indexation des contenus textuels, gestion des métadonnées intégrées dans les documents, extraction de contenu pour récupération, possibilité de charger les données archivées pour créer de nouveaux documents ;
- Choisir et spécifier les formats retenus pour l'archivage : versions à utiliser, paramétrage des options, contrôles à prévoir ;
- Rechercher et qualifier les outils de conversion et de vérification parmi les produits offerts sur le marché ;
- Organiser la migration : choix des acteurs, planification des opérations, pilotage du chantier.

Contact            Philippe Martin,  
[phm@bvdic.com](mailto:phm@bvdic.com)  
01 45 24 29 23