Formats de fichiers adaptés à l'archivage électronique à moyen et long terme

Version	Date	Objet de la version
1.0	19.10.2011	Document validé par le Collège spécialisé des systèmes d'information (CSSI). Rédaction: Archives d'Etat.

1. Introduction

La mise en œuvre d'un système d'archivage électronique à long terme doit prendre en compte la pérennité des formats de documents et de données. La question des formats est critique, car elle constitue un point particulièrement difficile à résoudre dans la perspective de la conservation à long terme de l'information. Les évolutions techniques sont si rapides qu'il est difficile de présager la solution qu'il faudra adopter. C'est pourquoi il est raisonnable de conserver les données dans des formats répondant à des critères définis. Les documents numériques doivent être enregistrés dans des formats pérennes si possible dès leur production, qu'ils proviennent de suites propriétaires ou libres.

Les formats de données fixés dans ce document sont considérés comme étant compatibles avec l'archivage à moyen et long terme et respectent les exigences permettant de garantir une compréhension et une exploitabilité sur le long terme.

Le nombre de formats acceptés est restreint. Ce choix de limiter les formats d'archivage est effectué dans toutes les institutions pratiquant de l'archivage électronique. En effet, un petit nombre de formats soigneusement sélectionnés et contrôlables garantit de manière plus sûre leur lisibilité sur le long terme qu'une grande quantité de formats dont l'entretien sera plus complexe et onéreux.

Les formats choisis correspondent principalement aux critères d'ouverture et d'indépendance. Un format ouvert doit avoir une documentation complète et accessible à tous. C'est à partir de cette documentation qu'il sera possible d'écrire un programme pour lire les données ou les convertir vers un autre format.

2. Objectif et définitions

L'objectif de ce document est d'établir la liste des formats de fichiers adaptés à l'archivage électronique à moyen et long terme et exigés pour leur conservation à l'Etat de Genève.

Ce document sert de **recommandation** pour déterminer les formats à utiliser dans le cycle de vie du document au sein de l'administration (archivage électronique à moyen terme).

Ces formats sont **obligatoires** dans le cadre de l'archivage électronique à long terme.

On entend par **archivage électronique à moyen terme** la durée d'utilité légale des documents (DUL), à savoir la durée pendant laquelle le document doit être conservé et rester accessible pour des raisons légales (au terme de sa DUL, le document peut être soit détruit soit versé aux archives d'Etat).

On entend par **archivage électronique à long terme** les documents qui sont versés aux archives d'Etat pour être conservés définitivement, pour les besoins de la gestion et de la justification des droits des personnes physiques ou morales, publiques ou privées, et pour la documentation historique de la recherche. L'archivage à long terme des documents électroniques a pour objectif de faire en sorte que les documents électroniques remis aux AEG restent durablement compréhensibles et que leur authenticité, leur intégrité et leur accessibilité soient garanties.

3. Formats d'archivage exigés

Le choix des formats sélectionnés est expliqué dans les pages suivantes.

	Domaine d'utilisation	Formats acceptés	Remarques
1.	**Données textuelles**		Le format PDF/A est considéré comme le mieux adapté à l'archivage.
	Texte (non structuré)	**"Texte uniquement" ("plain text")** .txt	UTF-8 UTF-16 ISO-8859-1 ISO 8859-15 US-ASCII
	Texte (Documents "Office")	**PDF/A** .pdfa, .pdf	Correspond à PDF 1.4 avec restrictions
2.	**Bases de données**		
	Tableaux	**CSV** .csv	Comma separated values
	Bases de données relationnelles	**SIARD RDB DATA**	Pour l'archivage à long terme uniquement (prendre contact avec les AEG).
3.	**Données graphiques**		Le format bien éprouvé TIFF reste le mieux adapté à l'archivage.
	Format de fichier graphique Bitmap	**TIFF** .tif, .tiff	
	Plans (vectoriels ou non)	**PDF/A** **.pdfa, .pdf**	
4.	**Données audio**		
	Audio	**WAVE** .wav	

4. Documentation sur les formats d'archivage

Texte

Le format textuel peut être considéré comme le format le plus stable de l'informatique; le codage ASCII est déjà connu et utilisé depuis des décennies; les codages plus récents comme la famille ISO 8859 ainsi que les divers codages Unicode sont aussi rétrocompatibles. Ce type de fichiers très simples présente les meilleures qualités de conservation et de compréhension mais n'est quasiment pas utilisé dans l'administration.

Un fichier texte non structuré convient pour présenter un contenu texte pur acceptant des possibilités de structure minimales et qui ne nécessite aucune autre information de structure ou de présentation (aucune instruction de présentation intégrée ou visible comme des caractères gras, des retraits, de la couleur, etc. ni information sur la structure comme des titres, paragraphes, sommaires, etc.).

Les fichiers qui se trouvent dans d'autres jeux de caractères que les formats acceptés doivent être convertis à la norme Unicode, de préférence UTF-8.

UTF-8 (*UCS transformation format* 8 bits) est un format de codage de caractères défini pour les caractères Unicode. C'est une extension du code ASCII utilisant le bit de poids fort. Chaque caractère est codé sur une suite de un à quatre octets. Il fait partie intégrante de la norme Unicode dans son chapitre 3 "*Conformance*", approuvé par l'ISO et la plupart des organismes de normalisation.

Unicode est un standard international qui fixe durablement un code numérique pour chaque signe ou élément de texte significatif de toutes les cultures écrites et systèmes de signes connus. On trouve dans Unicode les principaux jeux de caractères ISO tels que les normes de la série 8859 en reproduction à l'identique.

Il faut faire attention à la distinction fondamentale entre les formats basés sur les pages et ceux non basés sur les pages. Pour des documents dont le format est basé sur les pages, le PDF/A est optimal car le saut de page et la mise en page sont déterminés.

Le format Text convient pour des données texte simples, non structurées. Par exemple, de brèves descriptions.

PDF/A

Le Portable Document Format (PDF) est un format ouvert (c'est-à-dire publié) propriétaire qui permet de décrire des pages imprimées, créé par la société Adobe Systems Inc.

La spécification PDF/A-1 a été publiée comme norme ISO 19005-1 (*Electronic Document File Format for Long-Term Preservation, PDF/A-1*) sur la base de la version 1.4 de PDF. PDF/A est donc un "Portable Document Format" conçu pour l'archivage à long terme. Il comporte un certain nombre de restrictions par rapport au format PDF, mais contrairement à ce format, qui peut ne pas inclure les polices de caractère dans un fichier, il intègre au sein du format PDF/A tous les éléments nécessaires à la restitution du document, et notamment les polices de caractère et les informations colorimétriques dont il a besoin. Le volume des fichiers s'en trouve augmenté, mais ils sont totalement indépendants des plates-formes sur lesquelles on les utilise et sont auto-documentés.

Le contenu de la norme ISO 19005-1 est très complet. Il comprend la définition du format PDF/A-1, mais aussi la façon de développer un outil de visualisation de fichier conforme à ce format. Cela garantit ainsi la possibilité future de toujours disposer d'un outil de visualisation. La norme ISO 19005 contient également le document « PDF reference manual ».

Les restrictions définies par la norme ISO sont les suivantes:

- Non inclusion d'objets dynamiques de type audiogrammes ou vidéogrammes,
- Interdiction du lancement de codes scripts ou de fichiers exécutables,
- Inclusion de toutes les polices de caractères et leur utilisation sans contrainte légale et d'affichage,
- Palette des couleurs utilisée spécifiée de manière indépendante,
- Interdiction du chiffrement et de la sécurité, pas de codage ni de protection par mot de passe
- Utilisation obligatoire de quelques métadonnées standards

Il existe deux variantes de PDF/A-1 : PDF/A-1a , qui représente la forme complète de la norme ISO et PDF/A-1b, qui représente une forme allégée de la norme ISO. Cette deuxième version préserve la lisibilité du document et sa bonne présentation à l'affichage et à l'impression.

Le format PDF/A convient pour la conservation de documents dont le contenu peut être restitué de manière adéquate sur des feuilles imprimées (documents Word, Excel, etc.). Chaque document qui pourrait être imprimé et dont la version imprimée restitue de manière adéquate le contenu des documents peut être converti en PDF/A pour l'archivage. Il convient pour les documents "Office".

Le format PDF/A convient pour la conservation des plans (vectoriels ou non).

CSV (Comma Separated Values)

Les données CSV sont des données texte structurées. Chaque ligne présente des champs (colonnes) qui sont séparés par une virgule, un point virgule ou un autre caractère choisi. Les lignes des tableaux sont séparées par un saut de ligne. Ce format permet d'archiver très facilement des tableaux.

Toutefois, vu l'impossibilité de conserver des relations, des métadonnées et des informations structurelles dans ce format, seuls des ensembles de données en format CSV bien documentés garderont leur valeur.

Le format CSV convient pour des données organisées en forme de tableau; les fichiers Excel si le contenu du tableau est important (contrairement à la présentation); divers tableaux de petites banques de données (MS Access, MySQL, etc).

SIARD RDB-DATA

SIARD permet d'enregistrer dans un codage XML simple des structures (schémas, tableaux etc.) et le contenu de bases de données relationnelles. Les archives SIARD consistent en un fichier de contenu et un fichier de métadonnées comprenant des métadonnées de tous les niveaux. SIARD est basé sur des normes ISO (SQL:1999 et XML 1.0) et permet de conserver des bases de données relationnelles en provenance de différents systèmes, notamment MS Access, Oracle et MS SQL.
La spécification originale de SIARD a été publiée par les Archives fédérales suisses.

Le format SIARD convient pour l'archivage à long terme des données bases de données relationnelles (prendre contact avec les AEG pour l'utilisation de ce format au moment du versement).

TIFF (Tagged Image File Format)

Le format TIFF est un format ouvert dont les spécifications appartiennent à la société Adobe. Ce format sauvegarde des images raster (également appelées bitmap) sous la forme d'une grille de pixels. Les images raster peuvent être produites dans différents types de résolutions et de profondeur.

Certaines limitations sont nécessaires pour garantir la compréhension à long terme:
- pas de compression (exception: images noir-blanc)
- les fichiers TIFF doivent être validés TIFF 6.

L'intégration de plusieurs pages dans un seul fichier TIFF (Multipage-TIFF) n'est pas autorisée. Si la même image doit être archivée en diverses résolutions ou profondeurs en bits, il faut créer divers fichiers dans divers répertoires afin de permettre de commander facilement chacune des diverses versions. Si un document doit être archivé avec diverses pages, il faut utiliser le format PDF/A.

Le format TIFF convient pour les données d'image. Les images numériques dans d'autres formats (par ex. JPEG, GIF) doivent être converties au format TIFF.

WAVE

Le format WAVE (ou WAV) est un format conteneur destiné à l'enregistrement sans pertes de données audio. Il est fondé sur le «Resource Interchange File Format» (RIFF) mis au point par Microsoft et IBM pour le système d'exploitation Windows.
Le format audio PCM (Pulse Code Modulation) contenu dans le format WAVE garantit l'enregistrement et la lecture de signaux acoustiques de la qualité la plus élevée. Le format WAVE n'offre pas de compression des données, mais peut contenir des données audio comprimées (par exemple des signaux comprimés ADPCM ou encore MP3). Il fonctionne avec des profondeurs d'échantillonnage de 8 et 16 bits et un taux d'échantillonnage atteignant 44,1 kHz, ce qui correspond à une quantité de données de 88,2 ko par seconde.
Le format WAVE est très proche du format CD-Audio (CDA). Toutefois, la documentation le concernant est d'un accès extrêmement difficile. Des programmes courants (Nero, etc.) convertissent le CDA en WAVE.

Le format WAVE convient pour les données audio numériques.

5. Références

1. Documentation sur les formats et leurs versions : PRONOM

Les archives nationales anglaises développent depuis plusieurs années un système d'information en ligne sur les formats de fichiers dénommé PRONOM. Conçu à l'origine pour répondre aux besoins propres des archives nationales anglaises en matière d'archivage à long terme des documents électroniques, PRONOM s'est par la suite ouvert au public. De fait, PRONOM est aujourd'hui un outil de référence pour toutes les informations relatives aux formats.

La documentation détaillée des formats d'archivages spécifiés dans ce document se trouve donc sur la section PRONOM du site des archives nationales anglaises:

http://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=new

2. Définition des formats acceptés

Archives fédérales suisses, *Formats de fichiers adaptés à l'archivage. Normes et standards pour l'archivage de documents numériques*, juillet 2007.

CECO, *Catalogue des formats de données d'archivage* (Cfa, v.2), mai 2010.

Direction générale de la modernisation de l'Etat (DGME - France), *Référentiel Général d'Interopérabilité RGI*, version 1.0, 12 mai 2009.

Françoise Banat-Berger, Laurent Duplouy, Claude Huc, *L'archivage numérique à long terme, les débuts de la maturité?*, Direction des Archives de France, 2009.