Section 7 - Formats de représentation de l'information

version 1

Françoise BANAT-BERGER
Claude HUC

22 novembre 2011

PIAF
PORTAIL INTERNATIONAL
archivistique
FRANCOPHONE

Table des matières

Section 7 - Formats de représentation de l'information

Chapitre 1. Objet de la section

La question des formats de représentation de l'information est celle de la transformation d'une ou plusieurs séquences de bits en une information intelligible. C'est une question fondamentale pour pouvoir relire nos documents dans le futur proche ou lointain.

Le choix du format des objets numériques constitue un point particulièrement sensible pour l'archivage numérique.

- Dans le passé, certaines données ont été perdues parce qu'on n'avait plus aucune connaissance sur leur format ni sur l'application qui les avait créées.
- D'autres données ont dû subir des migrations coûteuses parce qu'elles avaient été enregistrées dans un format propriétaire sans pérennité. Ces migrations ont impliqué le développement puis l'exploitation de logiciels permettant de relire ces données et de les enregistrer dans un format neutre, indépendant de tout système propriétaire.
- Pire, certains documents bureautiques ont été entièrement ressaisis manuellement parce qu'ils avaient été enregistrés dans des formats propriétaires totalement fermés définis par des entreprises qui, elles non plus, n'ont pas eu de pérennité.

L'objectif de cette section est de donner à l'archiviste un ensemble d'éléments et de repères pour faire face à ce problème.

Cette section permet aussi de proposer à l'informaticien une approche des logiciels et des fichiers qui ne lui est pas familière.

(Voir aussi le module 9 section 2 sur la numérisation).

Chapitre 2. Éléments de terminologie

La terminologie du domaine n'est pas amusante en soi. Il est cependant tout à fait indispensable d'être précis sur les différents termes que nous allons utiliser dans cette partie.

Un format, dans son sens le plus général, permet de définir les caractéristiques physiques ou logiques d'un support d'information. Les formats sont le plus souvent normalisés ou standardisés.

Le format peut définir :

- soit le support physique, on parlera alors de format de support ;

dans ce cas, il précisera les caractéristiques physiques de ce support : A4 est un format papier de dimensions 21cm x 29.7 cm,

- soit les caractéristiques logiques d'organisation de l'information, nous parlerons alors de format de données et c'est cela qui nous intéresse ici,
- soit l'ensemble des caractéristiques physiques et logiques qui peuvent être imbriquées (VHS, CD-Photo Kodak), situation peu propice à la pérennisation.

Nous verrons dans ce cours, toute une série d'exemples concrets sur les formats.

Le format va posséder de multiples caractéristiques. Certaines d'entres elles sont essentielles pour savoir si le format sera recevable ou non dans une perspective d'archivage numérique :

- Format fermé : un format fermé a une structure non documentée et a priori inconnue hors de ceux qui l'ont défini.
- Format normalisé : un format sera dit normalisé s'il est conforme à une norme émanant d'un organisme de normalisation (ISO, AFNOR…).

Attention : une norme décrivant un format de données peut n'être qu'un conteneur à l'intérieur duquel doivent être insérés des éléments qui peuvent ou non être normalisés, voire privés.

- Format propriétaire : c'est un format défini par une entreprise ou un propriétaire privé qui dispose des droits de propriété intellectuelle ou du copyright correspondant (par exemple PDF, TIFF, GIF…) ;

Deux cas de figure peuvent se présenter :

le format propriétaire n'a pas été publié *(par exemple les fichiers produits par Microsoft Word 97).*

- Format maison ou format projet : c'est un format de données défini spécifiquement par une application maison ou par un projet au sein d'une entreprise.
- Format publié : il s'agit d'un format dont les spécifications sont publiées et accessibles à tous sans restriction ; cela ne signifie pas que l'usage de ce format puisse se faire sans restriction.
- Format ouvert : format publié et libre de droit, sans restriction d'usage et de mise en œuvre ; c'est le cas des formats définis par le consortium W3C (par exemple HTML, PNG).
- Format standardisé : un format sera dit standardisé s'il est conforme à un standard

Chapitre 3. Retour sur le codage

Nous avions donné quelques rudiments de codage dans la section 3. Nous y revenons ici plus en détail en illustrant les différentes catégories d'information que nous pouvons coder :

• bien évidemment les caractères, mais dans le monde d'aujourd'hui, ce sont les caractères de toutes les langues du monde qu'il faut pouvoir coder, c'est aussi notre capacité à produire des documents multilingues que les techniques de codage doivent prendre en charge,

• les nombres de toute nature et de toute précision. Certains modes de codage des nombres seront plus adaptés à la réduction de l'espace de stockage occupé et à la facilité de manipulation pour les calculs,

• le codage des couleurs sera un autre exemple.

Remarque

Nous ne sommes pas obligés, pour suivre l'ensemble de la section, d'entrer dans le détail de ce chapitre sur le codage mais il est intéressant de voir combien les fichiers peuvent être volumineux.

3.1. Tour de Babel du codage des caractères

Jusque dans les années 80, chaque constructeur d'ordinateurs utilisait son propre codage.

Exemple

Control Data utilisait le « Display Code », codage sur 6 bits seulement.

IBM avait, de son côté, créé le code EBCDIC (Extended Binary Coded Decimal Interchange Code), mode de codage des caractères sur 8 bits créé par IBM.

Cette multiplicité de codes entraînait des incompatibilités multiples entre les différents ordinateurs.

Complément : Le code ASCII

Le code ASCII (American Standard Code for Information Interchange) est un code à 7 bits permettant de représenter tous les caractères anglo-saxons utilisés par un certain nombre de constructeurs d'ordinateurs. C'est également la variante américaine de la norme de codage de caractères ISO/CEI 646.

Les codes 0 à 31 ne sont pas des caractères visibles. On les appelle caractères de contrôle car ils permettent de faire des actions telles que :

- retour à la ligne (CR signifiant Carriage Return),
- Bip sonore (BEL).

Les codes 65 à 90 représentent les majuscules.

Les codes 97 à 122 représentent les minuscules.

	Représentation graphique	Représentation binaire
Display code (Control Data)	A	000 001
ASCII 7 bits	A	100 0001
EBCDIC (IBM)	A	1100 0001

Tableau 1 Les représentations de la lettre A en majuscule en Display code, ASCII et EBCDIC

Tous ces codes ont subi des variations au cours du temps mais aucun d'entre eux ne permettait de représenter des caractères latins, grecs, cyrilliques, etc.

Dans les années 1990, l'ISO a re-normalisé et étendu le code ASCII et a créé la norme ISO 8859 :

Le codage se fait systématiquement sur 8 bits.

• Les 128 premiers caractères sont ceux d'ASCII,

• Les 128 suivants sont spécifiques de la langue.

16 versions ont été créées pour toutes les langues européennes, l'hébreu, le cyrillique, l'arabe et quelques autr

Complément : Les 16 versions de la norme de codage ISO 8859

ISO 8859-1 (latin-1 ou européen occidental)	C'est la partie la plus largement utilisée de ISO 8859, couvrant la plupart des langues européennes occidentales : l'allemand, l'anglais, le basque, le catalan, le danois, l'écossais, l'espagnol, le féringien, le finnois (partiellement), le français (partiellement), l'islandais, l'irlandais, l'italien, le néerlandais (partiellement), le norvégien, le portugais, le rhéto-roman et le suédois, certaines langues européennes sud-orientales (l'albanais), ainsi que des langues africaines (l'afrikaans et le swahili). Le symbole de l'euro et la capitale Ÿ, qui manquaient, sont dans la version révisée ISO 8859-15 (latin-9). Le jeu de caractères correspondant ISO-8859-1, approuvé par l'IANA, est le codage par défaut des anciens documents HTML ou des documents transmis par messages MIME, tels que les réponses HTTP quand le type de média du document est « text » (par exemple les documents « text/html »).
ISO 8859-2 (latin-2 ou européen central)	Langues d'Europe centrale ou de l'Est basées sur un alphabet romain (le bosniaque, le croate, le polonais, le tchèque, le slovaque, le slovène et le hongrois).
ISO 8859-3 (latin-3 ou européen du Sud)	Le turc, le maltais, et l'espéranto ; supplanté par ISO 8859-9 pour le turc et par Unicode pour l'espéranto.
ISO 8859-4 (latin-4 ou européen du Nord)	L'estonien, le letton, le lituanien, le groenlandais et le sami.
ISO 8859-5 (cyrillique)	La plupart des langues slaves utilisant un alphabet cyrillique, y compris le biélorusse, le bulgare, le macédonien, le russe, le serbe et l'ukrainien (partiellement).
ISO 8859-6 (arabe)	Couvre les caractères les plus courants de l'arabe. Nécessite un moteur de rendu qui prend en charge l'affichage bidirectionnel et l'analyse contextuelle.
ISO 8859-7 (grec)	La langue grecque moderne (orthographe monotonique).
ISO 8859-8 (hébreu)	L'alphabet hébraïque moderne tel qu'il est utilisé en Israël.
ISO 8859-9 (latin-5 ou turc)	Proche de l'ISO 8859-1, où les lettres islandaises peu utilisées sont remplacées par des lettres turques. Il est aussi utilisé pour le kurde.
ISO 8859-10 (latin-6 ou nordique)	Réarrangement du latin-4. Considéré plus utile pour les langues nordiques. Les langues baltes utilisent plus souvent le latin-4.
ISO 8859-11 (thaï)	Contient la plupart des glyphes requis pour la langue thaï.
ISO 8859-12	Était supposé couvrir l'alphabet devanāgarī, mais ce projet a été abandonné en 1997. ISCII et Unicode/ISO/CEI 10646 couvrent le devanāgarī.
ISO 8859-13 (latin-7 ou balte)	Ajoute quelques caractères supplémentaires pour les langues baltes qui manquaient en latin-4 et latin-6.
ISO 8859-14 (latin-8 ou celtique)	Couvre des langues celtiques telles que l'irlandais (orthographe traditionnelle), le gaélique écossais, le mannois (langue disparue) et le breton (certaines anciennes orthographes).

	symboles peu utilisés, les remplaçant avec le symbole de l'euro € et les lettres Š, š, Ž, ž, Œ, œ, et Ÿ, ce qui complète la couverture du français, du finnois et de l'estonien.
ISO 8859-16 (latin-10 ou européen du Sud-est)	Prévu pour l'albanais, le croate, le hongrois, l'italien, le polonais, le roumain et le slovène, mais aussi le finnois, le français, l'allemand et l'irlandais (en nouvelle orthographe). Cette police mise plus sur les lettres que les symboles. Le signe de monnaie est remplacé par le symbole de l'Euro.

Les normes de codages doivent évoluer car le langage évolue et de nouveaux besoins apparaissent.

Exemple

Le symbole € et les Œ, œ et Ÿ qui manquaient pour l'écriture du français ont été ajoutés à la norme 8859-15 qui constitue une révision de l'ISO latin 1 (8859-1).

Comme le nombre de position sur un octet est limité à 256, un certain nombre de symboles peu utilisés ont été abandonnés au profit de quelques nouveaux

Complément : De l'ISO latin-1 à l'ISO latin-15 : Différences ISO 8859-15 --- ISO 8859-1

Position	0*A4	0*A6	0*A8	0*B4	0*B8	0*BC	0*BD	0*BE
8859-1	¤	¦	¨	´	¸	¼	½	¾
8859-15	€	Š	š	Ž	ž	Œ	œ	Ÿ

De son coté, Windows utilise le codage Windows-1252 Identique à ISO 8859-1 sauf dans la plage 80-9F.

Conclusion

Mais alors, comment faire pour écrire un texte mixte français – grec ?

D'une manière générale comment encoder des documents multilingues ?

Comment prendre en compte les langues asiatiques ?

Comment prendre en compte le sens de l'écriture ?

3.2. Codage universel des caractères

A partir de 1993 s'est constitué le consortium Unicode afin d'apporter des réponses à ces questions et essayer de régler définitivement ce problème de la représentation des textes sur un ordinateur. Un jeu universel des caractères a ainsi été défini.

Au cours des années 1990, le consortium Unicode et le comité technique mixte ISO/CEI JTC 1, Technologies de l'information, sous-comité SC 2, Jeux de caractères codés ont coordonné leurs efforts. Le jeu universel de caractères codés sur plusieurs octets, défini par la norme internationale ISO 10646 et par le standard Unicode sont identiques. Par contre, la norme ISO ne précise ni les règles de composition de caractères, ni les propriétés sémantiques des caractères, ce que fait Unicode. La norme ISO 10646 existe en anglais et en français.

Complément

L'ensemble du jeu de caractères ainsi défini doit être considéré comme formé de 128 groupes de 256 plans. Chaque plan est formé de 256 rangées de caractères, chaque rangée contenant 256 cellules. Chaque caractère du jeu complet de caractères codés doit être représenté par une suite de quatre octets qui est appelée point de code.

Octet de groupe (octet G)	Octet de plan (octet P)	Octet de rangée (octet R)	Octet de cellule (octet C)

Tableau 2 Chaque caractère est représenté par une suite de 4 octets

Cet espace de codage est immense. Le jeu de caractère prévoit des zones spécifiques pour les usages privés ainsi que de multiples possibilités d'extensions.

Le plan 00 du groupe 00 constitue le **plan multilingue de base** (PMB). Ce PMB peut être utilisé comme jeu de caractères codés à deux octets (l'octet de rangée et l'octet de cellule, sachant implicitement que l'octet de groupe et l'octet de plan ont la valeur hexadécimale 00). Pour cette raison le PMB sera alors appelé UCS-2 (Universal Character Set 2). **Ce plan nous intéresse particulièrement car il permettra de répondre à l'essentiel des besoins.**

Rangée	code
00 à 02	Latin de base, Supplément Latin-1, Latin étendu A, Latin étendu B, Alphabet phonétique international
03 à 05	Grec et copte, Cyrillique, Arménien Hébreu

06	Arabe
07	Syriaque, Thâna
09 à 12	Dévanâgarî, Bengali, Gourmoukhî, Tamoul, Thaï, Tibétain, Birman, Éthiopien....
1E	Latin étendu additionnel
1F	Grec étendu
20 à 22	Ponctuation générale, Exposants et indices, Symb. Monétaires, Formes numériques, Flèches, Opérateurs mathématiques
...	...
28	Combinaisons Braille
2F	Clés chinoises (K'ang-hsi ou Kangxi)
.etc.	

Tableau 3 Rangées 00 à 2F du PMB

Chaque caractère graphique est identifié par un nom unique normalisé. Par contre, le glyphe représentant le caractère n'est pas normalisé par l'ISO 10646. Les symboles graphiques sont considérés comme des représentations visuelles types des caractères, représentations qui vont dépendre de la fonte utilisée.

Point de code hexadécimal (normalisé)	Nom (normalisé)	Glyphe (non normalisé)
0041	LETTRE MAJUSCULE LATINE A	A
00C7	LETTRE MAJUSCULE C CEDILLE	ç
03B1	LETTRE MINUSCULE GRECQUE ALPHA	?

Tableau 4 Normalisation des points de codes et des noms de caractère

Ce jeu de caractère et l'ensemble des mécanismes associés permettent :

• de traiter des textes multilingues,

• de traiter toutes les écritures quels que soient les symboles et quel que soit le sens d'écriture.

Compatibilité :

Le point de code d'Unicode est compatible

- Avec l'ASCII sur les 128 premiers index,
- Avec ISO Latin-1 sur les 128 suivants.

Cela rend les chaînes Unicode manipulables par les langages standards de l'Internet.

Mise en oeuvre :

En pratique, le jeu de caractères universel, sous sa forme canonique à 4 octets constitue un référentiel à partir duquel plusieurs encodages différents seront définis en fonction des besoins :

- UTF-8, qui est l'encodage par défaut de XML. UTF-8 peut être considéré aujourd'hui comme l'encodage standard de l'Internet. UTF signifie : « UCS transformation format ». UTF-8 permet la représentation de tous les caractères du jeu universel. C'est le plus utilisé, c'est un encodage de taille variable, chaque caractère est codé sur un ensemble variant de 1 à 6 octets.
- UTF-16 est un encodage fixe sur 2 octets.
- UTF-32 est un encodage fixe sur 4 octets. C'est le seul pour lequel le point de code correspond à la représentation en mémoire. C'est le plus gourmand en ressources.

UTF-8 a été inventé pour permettre à la fois :

- La représentation de tous les caractères existant dans le jeu universel,
- La compatibilité avec l'ASCII, ce qui offre le double avantage d'un gain en performance (le codage des caractères ASCII reste toujours sur un seul octet) et la compatibilité avec les textes existants.

Caractère	Code point	Encodage UTF8
A	41	01000001
é	00E9	11000011 10101001
€	20AC	11100010 10000010 10101100

Tableau 5 Exemple d'encodage UTF-8

3.3. Où spécifier le codage ?

La façon de spécifier le codage que l'on veut utiliser va dépendre du logiciel.

Dans Word nous pouvons procéder ainsi :

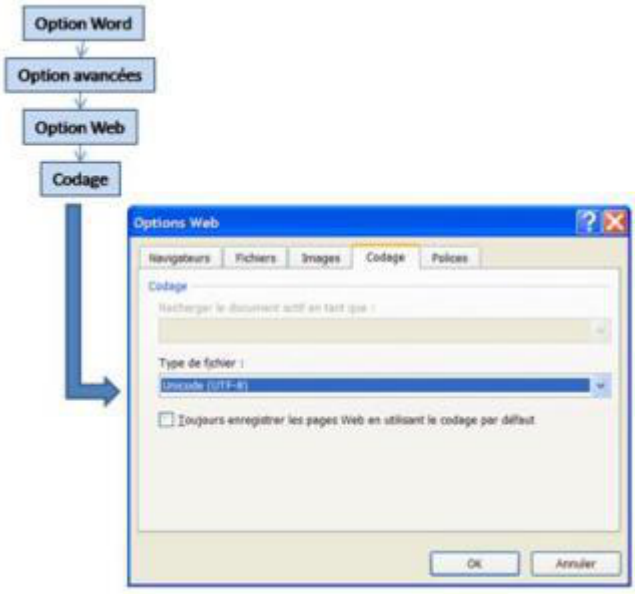

Image 1 Comment spécifier le codage sous Word

Nous pouvons ajouter qu'en l'absence de spécification du codage, Word s'appuiera sur un codage par défaut de Windows qui est un codage non normalisé.

Dans les fichiers XML dont nous parlerons un peut plus loin, nous trouvons en tout début de chaque fichier XML :

<?xml version="1.0" encoding="UTF-8"?>

3.4. Codage des nombres

Pour pouvoir représenter des nombres, on attribue des valeurs aux positionnements des bits.

Complément : Cas des nombres entiers – représentation dite binaire

On décompose l'entier en puissances de 2 et on attribue 0 ou 1 à chaque puissance

183 = 1x128 + 0x64 + 1x32 + 1x16 + 0x8 + 1x4 + 1x2 + 1x1

183 pourra être représenté par

10110111 ou par 11101101 en fonction du sens qui sera utilisé pour lire cette petite séquence de bits.

Dans le premier cas, on a mis les bits de poids fort (ceux affectés aux puissances supérieures de 2) à gauche,

Dans l'autre, on a mis les bits de poids fort à droite :

- La première approche est appelé gros-boutiste (Big Endian en anglais), elle est utilisée sur les processeurs Intel (sous Windows par exemple),
- La seconde approche est appelé petit-boutiste (Little Endian en anglais), elle est utilisée par les processeurs Motorola (sous MacOS par exemple)

Tableau 6 ***Un problème d'indiens !!!!***

Selon la grandeur du nombre entier, il faut plus ou moins de bits pour le représenter :

- avec 8 bits, on représente des nombres entiers positifs de 0 à 255
- avec 16 bits on ira jusqu'à 65536
- avec 32 bits jusqu'à 4 294 967 296
- avec 64 bits jusqu'à 1 844 674 073 709 551 616

En fait les valeurs sont plus faibles car il faut réserver un bit pour le signe.

Cas des nombres entiers – représentation dite codée

On peut utiliser une autre méthode très simple pour représenter un nombre entier :

Le nombre 183 peut être représenté en codant successivement :

- le caractère « 1 » du codage universel des caractères (identique ici à l'ISO latin-1) (0110001)
- le caractère « 8 » de ce même codage (0111000)

- le caractère « 3 » de ce même codage (0110010)

Cette représentation présente un avantage et un inconvénient :

- l'avantage est que cette valeur sera immédiatement lisible sur n'importe quel éditeur de texte qui interprète les caractères ;
- l'inconvénient est que pour représenter « 183 » sous une forme codée, il faut 3 octets, soit 24 bits alors qu'il ne faut que 8 bits pour représenter ce nombre sous une forme binaire ; cet inconvénient sera marginal dans de nombreux cas mais sera problématique dès que l'on à affaire à des grands volumes de données ; il sera plus rapide et moins coûteux de stocker 1 Go sous une forme binaire que 3 Go sous une forme codée.

Nous pourrons donc dire que le nombre entier positif 183 peut être représenté par :

- la séquence de bits 10110111 (mode binaire),
- la séquence de bits 0110001 0111000 0110010 (mode codé)

Ces séquences de bits ne pourront être interprétées correctement que si nous disposons par ailleurs d'une information précisant le mode de représentation du nombre, cette information constituant elle-même une partie de l'Information de représentation au sens du modèle OAIS.

Codage des nombres rationnels et plus généralement des nombres réels :

En informatique, on parlera aussi de nombre en virgule flottante.

En théorie, un nombre réel est formé de 4 éléments :

- la mantisse (nombre entier positif),
- le signe du nombre réel,
- l'exposant,
- le signe de l'exposant.

Là encore, nous pouvons définir une représentation dite binaire et une représentation dite codée.

La représentation codée suivra les mêmes règles que pour les nombres entiers.

> Le nombre -523,12 peut être codé avec 7 caractères : le signe « - », le séparateur « virgule » entre la partie entière et la partie fractionnaire et les chiffres nécessaires à la composition du nombre.
>
> Le nombre 1,6327-4 pourra être en pratique codé sous la forme +1,6327E-04 (le caractère « E » vient préciser que le nombre qui suit est un exposant. 11 octets seront nécessaires dans ce cas.

La représentation binaire des nombres réels sera plus souvent utilisée pour effectuer des calculs. Un artifice technique de normalisation entre la mantisse et l'exposant permet d'avoir un exposant systématiquement positif et par conséquent, de ne pas avoir à stocker le signe de l'exposant.

En général, on utilise, en fonction des ordinateurs et des besoins en précision, des représentations d'une longueur de 32, 64 ou parfois 128 bits. Cela signifie que la longueur cumulée de la mantisse, de l'exposant et du signe doivent correspondre à ces longueurs.

Là encore, il y a un certain nombre de conventions différentes en fonction des systèmes d'exploitation et des constructeurs. L'une de ces conventions est standardisée et nous en recommandons l'utilisation :

Il s'agit du standard 754-2008 IEEE Standard for Floating-Point Arithmetic.

Selon ce standard, un nombre flottant simple précision est stocké sur 32 bits : 1 bit de signe, 8 bits pour l'exposant et 23 bits pour la mantisse.

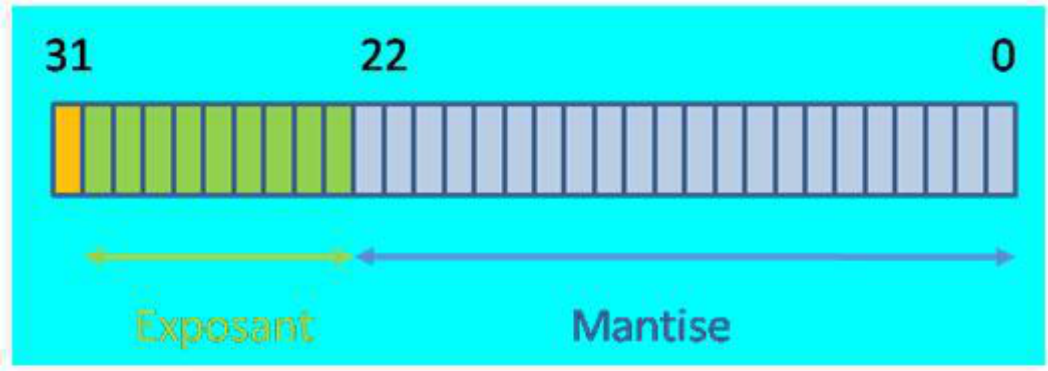

Image 2 représentation IEEE d'un nombre réel sur 32 bits

Un nombre flottant double précision est stocké sur 64 bits :1 bit de signe, 11 bit pour l'exposant et 52 bits pour la mantisse

3.5. Autres codages

Pour le codage des couleurs, il s'agit là encore de définir une convention qui permettra de définir la couleur d'un pixel dans une image ou d'un caractère dans un texte, etc. Le pixel est la plus petite unité adressable de l'écran.

Plusieurs possibilités existent :

- le codage d'une palette de 256 couleurs, dans ce cas, un octet suffira pour définir la couleur,
- le codage sur 3 octets (24 bits) : 8 bits sont consacrés à la teinte primaire rouge, 8 bits sont consacrés à la teinte primaire vert, 8 bits sont consacrés à la teinte primaire bleu,
- Le codage sur 4 octets (32 bits) : 24 bits sont utilisés comme pour le codage précédent, le dernier octet est soit inutilisé, soit utilisé à coder par exemple une information de transparence.

Il existe des codages spécifiques pour le son, d'autres pour la vidéo vue comme un ensemble d'images.

Chapitre 4. Quels critères pour évaluer les formats ?

Quels sont les critères qui vont nous aider à évaluer les formats de données par rapport à la perspective de pérennisation ?

Ces critères doivent permettre :

- d'identifier les formats qui vont faciliter la pérennisation des informations,
- d'éliminer les formats qui poseront à court ou à long terme, des difficultés sérieuses.

Garantir un accès aux données le plus longtemps possible implique de pouvoir conserver les informations dans leurs formats d'origine et d'avoir les moyens de les faire migrer vers un autre format si cela s'avère nécessaire. Apporter ces garanties seules ne suffit pas, il faut également se garder les moyens de pouvoir utiliser les données.

Compte tenu de la rapidité des évolutions techniques, il est extrêmement difficile de présager de la solution qu'il faudra adopter. Aussi, il est raisonnable de conserver les données dans des formats répondant à des critères définis dans l'optique de permettre la réalisation de chacune des solutions possibles.

À partir de ce paradigme, nous pouvons retenir deux principaux critères, l'ouverture et l'indépendance, auxquels nous ajouterons des critères complémentaires d'un niveau d'importance moindre par rapport aux précédents.

4.1. Ouverture : formats publiés et ouverts

Un format publié doit disposer d'une documentation complète et accessible. Cette documentation doit être valide, à jour et suffisamment détaillée pour permettre l'écriture de programme pour lire les données ou les convertir vers un autre format. C'est un gage essentiel de sécurité pour le futur.

Il est préférable que le format soit ouvert, c'est-à-dire libre de droits mais ce point n'est pas bloquant. Il convient cependant de vérifier les contraintes légales, l'usage de certains formats pouvant être payant.

Un format propriétaire largement diffusé (comme PDF ou TIFF) sera préférable à un format ouvert peu utilisé.

Plus un format est diffusé, plus il existe des outils qui sont développés pour l'exploiter. La large diffusion d'un format, à elle seule, n'est cependant pas un critère qui apporte la garantie de pouvoir utiliser ce format.

Exemple

Le format DWG (**D**ra**W**in**G**, littéralement dessin) est un format fermé lié au logiciel AUTOCAD, propriété de Autodesk. Ce format est très utilisé par les géomètres, les géographes, les architectes, les urbanistes. Or malgré une utilisation internationale très forte et diverse, la pérennisation de données au format DWG pose de sérieuses difficultés.

Attention

Attention à ce qu'on appelle les formats « enveloppes » ou encore « conteneur » qui spécifient une structure mais qui autorisent plusieurs algorithmes de compression différents : c'est le cas pour les formats d'image, les formats audio et vidéo.

En effet, le format peut être libre de droit, mais l'algorithme de compression sera libre de droit ou non, voire payant suivant le choix que l'on aura retenu.

4.2. Indépendance

Pour être pérenne, un format doit être totalement indépendant. Cette indépendance doit se caractériser :

- vis-à-vis des autres formats : certains formats peuvent paraître ouverts mais font appel à d'autres formats qui peuvent être fermés ou soumis à des brevets qui limitent le champ de leurs utilisations ; ils peuvent aussi faire appel à d'autres éléments comme des jeux de caractères qui seront normalisés ou propriétaires suivant les cas ;
- vis-à-vis des systèmes d'exploitation : lorsque les formats de données sont liés à un système d'exploitation, nous sommes dans le cas d'une forme cachée de fermeture ;
- au plan économique : même dans le cas des formats ouverts, les coûts de développement des outils de manipulation doivent être raisonnables pour permettre aux organisations ou à une communauté restreinte d'utilisateurs d'en assumer l'élaboration ;

- au plan matériel : il s'agit de s'assurer que le format choisi n'est pas lié à un périphérique ou un support de stockage spécifique non contrôlé.

4.3. Autres critères à considérer

D'autres critères peuvent utilement être pris en compte dans les cas où plusieurs solutions satisfaisant aux critères précédents se présentent. Ainsi :

- la disponibilité et le coût des outils et des facilités de création des données, ainsi que des outils de transformation des formats et de présentation des données,
- la possibilité de vérifier automatiquement qu'un fichier de données respecte les spécifications du format et respecte également les règles d'utilisation et les restrictions qui auront été définies pour la pérennité,
- les conséquences du choix en matière de volume de données : l'usage de formats inutilement volumineux sera évité,
- la complexité : un format simple est préférable à un format complexe,
- la structure du format : plus le contenu et le style seront mélangés dans le format, plus il sera difficile de transcoder l'un sans modifier l'autre ou d'adapter un autre style au même contenu,
- la disponibilité et les potentialités de développements de services à valeur ajoutée comme l'extraction de sous-ensembles, les changements de format pour la diffusion,…

Naturellement, ces critères complémentaires ne pourront jamais être tous satisfaits. Ils peuvent même être contradictoires entre eux : tel format sera complexe mais tel autre, plus simple, conduira à des volumes de données plus importants. Ils sont donc à apprécier en fonction des caractéristiques et du contexte de l'Archive.

Attention : Recommandation

Plus restreint sera le nombre de formats de documents acceptés et gérés par l'**Archive**, plus le risque de difficultés pour restituer l'information de manière intelligible sera réduit.

4.3. Exemples de recommandations existantes

Dans le cadre de développement de la plate-forme PIL@E, la Direction des Archives de France (DAF) a défini une stratégie basée sur la distinction entre formats d'entrée et formats d'archivage.

Le format d'entrée est le format des fichiers en entrée du système d'archivage alors que le format d'archivage ou format cible est le format retenu pour l'archivage à long terme des documents dans le système d'archivage.

L'approche retenue par la DAF repose sur les règles de base suivantes :

- nombre restreint de formats cibles (trois ou quatre formats au maximum pour chaque domaine : images, textes, messagerie électronique, fichiers comprimés),
- faible nombre de formats acceptés en entrée (les formats pris en compte doivent largement couvrir les besoins de l'administration sans toutefois être trop nombreux),
- tests des formats en entrée grâce à un logiciel testeur afin de s'assurer de la conformité de ces formats à leurs spécifications,
- conversion des formats d'entrée vers les formats d'archivage grâce à un logiciel convertisseur ; cette conversion est réalisée lors de l'entrée des fichiers dans le système d'archivage si le format en entrée n'est pas un format cible,
- archivage dans un journal des opérations de test et de conversion.

Le principe de sélection des formats cibles d'archivage est défini comme suit :

- le format doit être très largement répandu et/ou disposant d'une norme européenne ou internationale,
- dans le cas où le format ne dispose pas d'une norme, les spécifications de ce format doivent être publiques et facilement accessibles,
- la stabilité du format doit être raisonnable : le renouvellement des versions ne doit pas s'effectuer trop rapidement (2 à 3 ans est une périodicité acceptable),
- il doit exister au moins deux logiciels, d'éditeurs différents, disponibles sur le marché français ou européen qui exploite ce format ou il doit exister un logiciel en Open Source (dont le code source est public) qui gère ce format ; ces logiciels doivent au moins permettre une interprétation des documents qui rend compréhensible toute l'information contenue pour la communauté d'usagers visée.

Le Référentiel Général d'Interopérabilité (RGI) publié en mai 2009 dans une version non encore officielle, émet de son côté un ensemble de recommandations sur les formats d'image, les séquences sonores, les séquences vidéo, les objets graphiques en deux dimensions ou trois dimensions, les dessins techniques et les formats composites qui incluent en particulier toute la bureautique.

4.4. Registres de format

La collecte d'information et de documentation sur les formats numériques et sur les logiciels qui permettent de créer ou lire des données organisées selon ces formats est un lourd travail (caractéristiques, type, disponibilité de la documentation, droits de propriété applicables).

Les changements de version sont fréquents. L'évaluation d'un format par rapport aux critères énoncés ci-avant est consommatrice de ressources humaines significatives.

Il est donc illusoire d'imaginer qu'une institution pourra toujours seule suivre l'évolution de formats numériques de c

de référencement et d'évaluation pour une communauté d'utilisateurs.

Exemple

Les deux principales initiatives dans ce domaine, PRONOM au Royaume-Uni et Global Digital Format Registry (GDFR) (initiative de la « Digital Library Federation », DLF) ont décidé en avril 2009 de joindre leurs efforts pour constituer ensemble l'Unified Digital Formats Registry (UDFR).

Cette nouvelle initiative est soutenue par un certain nombre d'archives nationales et de grandes bibliothèques.

Les Informations rassemblées pour chaque format dans un registre doivent inclure au moins :

- les noms canoniques du format et ses variantes :
  - par exemple PDF, Adobe PDF, Portable Document Format
- les « signatures » internes et externes
  - extension = .pdf
- les spécifications du format
  - http://partners.adobe.com/public/developer/en/pdf/PDFReference16.pdf
- les auteurs, titulaires de droits, chargés de la maintenance
  - Société Adobe
- les relations avec d'autres formats dérivés, les versions
  - PDF 1.7, PDF 1.4, PDF/A, PDF/X…
- les systèmes, services et outils pour la création, la lecture, la validation de documents conformes à ce format
  - Adobe Acrobat Reader, Adobe Acrobat Distiller

Les registres peuvent aussi fournir des évaluations sur l'aptitude des formats à répondre à tel ou tel usage.

Autre initiative utile dans ce domaine : la Bibliothèque du Congrès américain a mis en place un site informatif sur les formats. Ce site propose un ensemble d'informations, de publications, de ressources sur les formats et sur leur pérennité probable. Le site présente une description des formats classée par type (texte, image, audio, vidéo). Les critères d'évaluation des formats rejoignent fortement ceux énoncés dans ce cours.

Chapitre 5. Tour d'horizon des formats de données

Nous avons vu jusqu'ici comment représenter des entités simples comme les nombres ou les caractères.

Il s'agit maintenant de représenter des objets numériques plus complexes et pouvant contenir du texte, des images, des graphiques, du son, de la vidéo. Ces différentes catégories d'informations pouvant par ailleurs être rassemblées, organisées, combinées au sein de documents appelés multimédias.

La représentation de ces objets pose d'autres problèmes qui se superposent aux précédents.

5.1. Rôle particulier de XML

Origine

Dès les années 1970 et au début des années 1980, IBM étudie la possibilité de stocker des textes sur ordinateur.

Trois de ses ingénieurs conçoivent un langage de balisage destiné à séparer les instructions de style et le contenu des documents ; Charles Goldfarb, Edward Mosher et Raymond Lorie nomment ce langage de leurs initiales : GML.

Les travaux d'IBM sont repris par l'ISO et conduisent à la norme ISO 8879/1986 baptisée SGML.

SGML signifie Standardized General Markup Language.

Concepts

Deux concepts essentiels sont introduits par SGML :

- la séparation du contenu et du style : le contenu est ce qu'il faut pérenniser, un même contenu peut être utilisé avec différents styles
  - papier
  - écran
  - téléphone portable
- la structure du document est aussi importante à pérenniser que son contenu : le balisage est le moyen d'y parvenir ; cela permet d'affecter une véritable sémantique aux différents éléments de la structure (titre de document, titre de chapitre, glossaire, note de bas de page,…).

De SGML à XML

En dehors de grands centres de documentation technique, il n'y aura que fort peu d'applications de SGML jusqu'en 1990. La mise en œuvre reste complexe, les outils logiciels sont coûteux.

En 1990, Tim Berners Lee crée le Web avec HTML (Hypertext Markup Language) qui est un langage construit sur les principes de SGML. A la suite des évolutions anarchiques de HTML entre 1993 et 1996, le W3 Consortium crée un groupe de travail pour dépoussiérer SGML et en faire une version allégée pour les réseaux.

Cela aboutit à la Recommandation du W3C : **Extensible Markup Language** (XML) 1.0 du 10 Février 1998.

Principes

XML est un métalangage (ou une grammaire) permettant d'écrire des langages qui auront les propriétés suivantes :

- ils seront tous conformes à XML, c'est à dire qu'ils partageront la même syntaxe, les mêmes règles et pourront donc être manipulés par des outils génériques;
- ils seront interopérables c'est à dire qu'on pourra créer des documents composites où coopèreront plusieurs langages XML ;
- ils seront balisés de telle sorte qu'ils matérialiseront la structure du document ; le contenu et le style seront séparés.

Ainsi tout document XML pourra être :

- vérifié par un analyseur syntaxique (appelé aussi « parseur »),
- mis en forme : cette mise en forme peut être réalisée par les langages CSS (Cascading Style Sheets : feuilles de style en cascade) ou XSL-FO (eXtensible Stylesheet Language - Formatting Objects). CSS et XSL-FO sont deux standards du W3C,
- transformé par le langage de transformation XSLT (eXtensible Stylesheet Language Transformations), ce qui permettra de composer des documents à partir de plusieurs autres documents, ou encore de produire plusieurs versions du même document en fonction des destinataires, etc.

Langages et formats

Lorsqu'un document a été structuré par le langage XML, on connaît en pratique l'ensemble des règles d'organisation de l'information au sein de ce document. A ce titre, XML peut donc être considéré comme un format. C'est un format ouvert, lisible par les ordinateurs et les humains. XML utilise le jeu de caractères d'Unicode et permet l'utilisation de différents encodages dont UTF-8 qui est le codage par défaut. XML est standardisé par le W3C.

Il existe à ce jour des centaines, peut-être des milliers de langages XML dédiés à des applications métier particulières. Un grand nombre de formats de données et de métadonnées dans les domaines les plus divers s'appuient sur la syntaxe XML. Certains sont assez génériques pour être largement employés

• SMIL (Synchronized Multimedia Interface Language) pour les documents multimédia,

• SVG (Scalable Vector Graphics) pour les graphismes vectoriels 2D,

• Xforms pour les formulaires.

XML n'est pas normalisé par l'ISO mais de nombreuses normes ISO s'appuient sur lui. Il bénéficie d'un grand nombre d'outils pour toutes plates-formes.

Attention : Important

Il faut ajouter que si on veut archiver des documents XML, il faut pouvoir valider ces documents à l'entrée de l'Archive et par conséquent pouvoir disposer des modèles auxquels se réfèrent les fichiers et archiver ces modèles. Il pourra s'agir de DTD (Définition de type de document), de schémas XML ou de tout autre type de grammaire standardisée.

L'adoption de modèles tels que DocBook, TEI (Text Encoding Initiative), textML, ALTO, est essentielle dans la perspective de l'archivage.

XML, un atout pour la pérennisation des documents numériques

- XML est un format ouvert standardisé par le W3 Consortium et mondialement utilisé dans tous les secteurs d'activité,
- XML est lisible par les humains et les ordinateurs et peut donc être transcodé sans difficultés. Il est syntaxiquement vérifiable,
- XML dissocie le contenu et le style et permet donc d'associer différents styles à un même contenu,
- XML est interopérable et ne dépend donc pas de la plateforme informatique utilisée,
- XML ne dépend que de l'encodage des caractères qui est assuré par UTF-8

5.2. Formats texte

Documents en texte intégral

On parle aussi de texte but, de texte simple (Plain Text), ce sont les documents qui portent une extension .txt sous Windows.

C'est un format ouvert dont le contenu va dépendre de l'encodage des caractères. En pratique, le document se réduit à une suite de caractères, d'espaces et de retours à la ligne.

Il existe de très nombreux outils d'édition, de manipulation, de conversion sur tous les systèmes d'exploitation.

Exemple

- Sous Windows : le bloc note, Textedit, Notepad++, Ultraedit…
- Sous Linux : Vi, Emacs…

Les documents en texte intégral ne posent pas de difficulté quant à leur pérennisation dès lors qu'on a mémorisé le codage utilisé. Cependant, ce sont des documents pauvres contenant du texte sans structure et sans style.

Formats bureautiques :

Il existe un certain nombre de formats utilisés par les suites logicielles bureautiques.
Nous ne citerons que les deux principaux

- ODF (Open Document Format)
- et OOXML (Office Open XML).

Complément

Format	Description
ODF – Open Document Format	Le format ODF (Open Document Format) est un format ouvert basé sur un langage normalisé de définition de schéma de document RELAX NG, lui même construit sur le langage XML. ODF a été standardisé par le consortium OASIS (Organization for the Advancement of Structured Information Standards) en 2005 puis est devenu la norme ISO 26300 en 2006 (Technologies de l'information - Format de document ouvert pour applications de bureau, OpenDocument v1.0). ODF a été retenu par nombre d'organismes publics nationaux et internationaux Plusieurs suites logicielles libres utilisent ce format. C'est le cas de la suite Open Office qui est distribué sous la licence GNU LGPL (Lesser General Public Licence1) qui fonctionne sur plusieurs plates-formes dont Microsoft Windows, Linux, Sun Solaris, ou encore Apple Mac OS. Le code source d'Open Office et sa documentation sont accessibles. D'autres suites bureautiques libres comme Lotus Note Symphony d'IBM ou encore Koffice permettent d'enregistrer des données au format ODF, ce qui renforce encore le poids de ODF dans la perspective de pérennisation des documents.
OOXML – Office Open XML	Les formats proposés par la suite logicielle bureautique Office de Microsoft (.doc, .xls, .ppt pour les versions anciennes et .docx, .xlsx et .pptx à partir de la version 2007) ont été des formats fermés non publiés jusqu'au début des années 2000. Ces formats, ainsi que la suite logicielle Microsoft Office ont évolué au rythme soutenu d'une version tous les deux ans depuis 1990. **Attention** **Très important !** **La compatibilité ascendante, permettant de lire, avec une version récente de la suite Office, un fichier créé avec une version plus ancienne n'est pas assurée au-delà de 10 ans (voir la section 12 de ce module sur les retours d'expérience).** Depuis le début des années 2000, la structure des formats, basée sur le langage XML, a été publiée. Ce format est complexe et sa documentation souvent peu explicite. Microsoft a ensuite été moteur dans la normalisation par le consortium ECMA en 2006 du format OOXML (Office Open

	XML), format qui a été proposé à l'ISO. La normalisation ISO d'OOXML a été obtenue en mars 2008. Par ailleurs, Microsoft avait annoncé son intention d'intégrer les formats ODF et PDF 1.5 à la suite Office 2007. La mise en application du format OOXML ne sera quant à elle effective qu'avec Office 2010. La documentation du format OOXML reste trop volumineuse (6000 pages), à la mesure de la complexité du format. Cette complexité n'est pas un avantage par rapport à la problématique de conservation à long terme. Le coût de développement d'une solution logicielle alternative à celle de Microsoft serait naturellement extrêmement élevé.

La situation à l'égard des formats des documents bureautiques a évolué de façon spectaculaire au cours des dernières années. ODF parait être actuellement la meilleure base pour l'archivage long terme **dès que l'on souhaite archiver des documents sous une forme révisable**. Nous ne saurions prédire ce qu'il adviendra dans le futur. Les épisodes passés nous incitent à rester prudents dans ce domaine.

Le format PDF (Portable Document Format) et sa version PDF/A dédiée à l'archivage

Format	Description
PDF 1.7	PDF est un format propriétaire publié. Il appartient à la société Adobe. C'est un format conteneur. Il permet de contenir d'autres types de format de données tels que des images couleur compressées en JPEG, du son, de la vidéo, etc. Il existe de nombreux outils pour manipuler ce format. Aussi bien des outils issus du monde du logiciel libre que des outils propriétaires. La politique d'Adobe est de distribuer gratuitement les outils de lecture et de vendre les outils de création. Le fait qu'il soit un format conteneur oblige à vérifier rigoureusement que les fichiers PDF destinés à l'archivage ne contiennent **que** les éléments attendus et **tous** les éléments attendus Le format peut inclure des métadonnées au format XMP (*Extensible Metadata Platform*). En juillet 2008, la version 1.7 de PDF est devenue la norme ISO 32000-1 2008 (Gestion de documents - Format de document portable - Partie 1: PDF 1.7). Cette normalisation ne change rien à la nécessité de prendre les précautions indispensables définies ci-avant.
PDF/A	La version 1.4 de PDF a été la base sur laquelle a été définie en 2005, la norme ISO 19005-1 : *Electronic Document file format for long-term preservation*, PDF/A-1. PDF/A comporte un certain nombre de restrictions par rapport à PDF mais il intègre au sein du format tous les éléments nécessaires à la restitution du document et notamment les polices de caractères dont il a besoin. Il s'ensuit une augmentation du volume des fichiers mais en contrepartie, une indépendance de ces fichiers par rapport aux plates-formes sur lesquelles on les utilise. Les restrictions imposées par la norme sont susceptibles d'entraîner une perte d'informations (cas d'utilisation du mode de transparence, existence d'audiogrammes ou de

	s'assurer, avant toute conversion de PDF en PDF/A, que le fichier PDF origine ne fait pas appel à des fonctionnalités de PDF non supportées par PDF/A.

Attention : A savoir

Le format PDF est destiné aux documents non révisables. Il présente l'avantage de pouvoir restituer la présentation originale du document de façon fidèle alors que cette garantie ne peut pas être totalement assurée par les formats bureautiques révisables.

Il ne faut pas pour autant croire qu'un document PDF ne pourra pas être modifié de façon malveillante.

5.3. Formats image et graphiques vectoriels

Il existe deux types de description pour les images :

- Les images à description de pixel. Chaque pixel de l'écran est affecté individuellement d'un ou de plusieurs nombres entiers représentant sa luminosité, sa couleur ou son opacité. Ces images sont faciles à mettre en œuvre mais d'une précision limitée si l'on veut rester avec des volumes raisonnables. Il pourra s'agir d'images obtenues par numérisation ou d'images nativement numériques comme c'est le cas pour les photographies. Le nombre de formats de ce type est très important (plusieurs centaines),
- Les images à description vectorielle. Les objets de la scène sont décrits de façon mathématique dans un espace orthonormé. La description est alors aussi précise que nécessaire. Il s'agira par exemple de graphiques mathématiques ou statistiques.

Formats d'image à description de pixels

Abréviation	**Nom et statut**	**Principales caractéristiques**
GIF	Graphics Interchange Format Format propriétaire publié Contrainte liée au brevet	Ce format d'image, très utilisé au début du Web, est assez peu performant. Il est frappé d'un brevet détenu par Compuserve.
TIFF	Tagged Image File Format Format publié Propriété de la société Adobe Pas de licence d'utilisation	C'est un format conteneur : Il définit une structure. Il permet d'inclure les profils ICC ((*International Color Consortium*) dans le fichier. Le profil ICC permet une gestion des couleurs indépendante des plates-formes et des périphériques. L'image peut être enregistrée selon différents algorithmes de compression selon le choix de l'utilisateur. Les images peuvent aussi être enregistrées sans compression. Le succès de ce format est dû à deux raisons principales : • Il permet l'enregistrement des images noir et blanc avec l'algorithme ITU T6 qui avait été conçu pour la transmission par télécopie : cet algorithme est très efficace et offre des taux de compression élevés sans perte, • Ce format offre aussi la possibilité d'enregistrer dans son en-tête des métadonnées techniques très complètes. TIFF est fréquemment utilisé comme format de numérisation.
JPEG et JPEG2000	JPEG (Joint Photographic Expert Group) est un format publié et	La norme JPEG décrit l'algorithme de compression et

	ouvert Norme ISO/IEC IS 10918-1. JPEG2000 est un format publié et ouvert Norme ISO/CEI 15444-1	les informations minimales pour l'utiliser. La raison de son succès est qu'il est largement implémenté en Open Source et qu'il a été adopté par tous les navigateurs Internet. Son défaut est qu'il est un algorithme de compression avec perte (suivant le paramétrage retenu, ce taux de compression peut être très faible). JPEG n'est pas à proprement parler un format de fichier, JPEG s'appuie sur le format JFIF (*JPEG File Interchange Format)*. Avec JPEG2000, la compression est l'une des améliorations importantes de ce format : par rapport au JPEG, à qualité de rendu égale, elle est beaucoup plus importante (surtout valable dans les forts taux de compression).
JBIG	Joint Bi-level Image experts Group Format publié ouvert. Norme ISO/IEC IS 11544 Limitation au niveau du brevet sur l'algorithme de compression	JBIG utilise un algorithme de compression sans perte. Il peut également être employé pour le codage à niveau de gris et les images de couleur avec un nombre limité de bits par pixel. L'inconvénient est que l'algorithme de compression sur lequel il repose est soumis à un brevet détenu par IBM, Mitsubishi et Lucent. C'est probablement l'une des raisons de sa faible diffusion
PNG	Portable network Graphics Format publié ouvert Standard du W3C Norme ISO/IEC 15948:2003	Format performant qui supporte les images en niveaux de gris ou en couleur. Il est accepté par la plupart des navigateurs modernes.

Formats à description vectorielle

Abréviation	**Nom et statut**	**Principales caractéristiques**
SVG	Scalable Vector Graphics Standard ouvert du W3C	Ce format ouvert et performant est doté de nombreux outils gratuits. Les fichiers au format SVG sont lisibles sur la plupart des navigateurs. Ce format est notamment utilisé en cartographie.
DWG	Abréviation de **D**ra**W**in**G** (Dessin) Format fermé, propriété de la société Autodesk, distributeur du logiciel Autocad	Ce format est très utilisé par les architectes, les géomètres, les géographes, etc. Pourtant, son caractère fermé et non publié ne le rend pas adapté à une conservation à long terme, prisonnière de Autocad

5.4. Formats audio et vidéo

Les formats audiovisuels et les formats sonores constituent un domaine techniquement complexe. Cependant, les critères d'évaluation des formats par rapport à l'archivage numérique s'appliquent entièrement.

Une distinction systématique sera faite entre la spécification du format qui définit le mode d'encapsulation du contenu, c'est à dire l'organisation de l'information au sein d'un conteneur et l'encodage proprement dit de l'information qui utilisera souvent un algorithme de compression des données visant à réduire leur volume. Nous ne donnons ici que quelques indications générales sur une sélection de formats existants.

Attention : Remarque

Il est fréquent ici de séparer les formats d'archivage, c'est à dire ceux qui permettent de conserver toute l'information utile, des formats de diffusion correspondant à un usage donné de ces informations.

Formats Audio

Les formats audio sont des formats « enveloppe », appelés aussi format « conteneur ». Au sein de ces formats, l'information audio peut être codée et compressée de différentes manières. Pour définir entièrement la représentation audio, il convient donc à chaque fois de préciser le type de codage qui sera utilisé.

Des recommandations sur les formats sonores sont émises par l'IASA (*International Association of Sound and Audiovisual Archives*). Il n'y a pas malheureusement pas d'équivalent pour la vidéo.

Nous présentons ici une courte synthèse des principaux formats.

Abréviation	Nom et statut	Principales caractéristiques
Wave	Format publié, propriété de Microsoft	Format très répandu. C'est l'encodage PCM (Pulse code modulation), encodage sans compression, qui est le plus utilisé, notamment pour le « disque compact ». Il peut aussi recevoir d'autres encodages comme MP3.
AIFF	Audio Interchange File Format Format publié, propriété d'Apple	Format de fichier audio développé par Apple et utilisé sur les ordinateurs Macintosh. Les données sont codées en PCM big-endian sans compression
MP3	MPEG Audio Layer 3 Algorithme de compression normalisé (ISO/CEI IS 11172-3 et ISO/CEI IS 13818-3) mais soumis à des redevances	Format audio doté d'un algorithme de compression capable de réduire très fortement la quantité de données nécessaire pour restituer de l'audio avec une perte de qualité sonore acceptable pour l'oreille humaine. La mise en œuvre de l'encodeur est dans le champ d'un brevet détenu conjointement par un ensemble de sociétés.
OGG	Format publié et ouvert	Format ouvert promu par la fondation Xiph.org. Ce format doit être utilisé avec le codage Vorbis, défini également par cette fondation. Vorbis est un algorithme de compression et de décompression audio numérique, ouvert et libre, plus performant en termes de qualité et taux de compression que le format MP3.

Formats Vidéo

Comme pour les formats audio, on distinguera les formats « conteneur » des algorithmes de codage et décodage des données vidéo. Ces algorithmes sont appelés « codec » (pour codage-décodage).

Les données vidéo font l'objet d'une organisation complexe et le choix d'un format pour la conservation mérite systématiquement une analyse spécifique. Nous proposons ici quelques éléments comme une base de réflexion préalable et non exhaustive.

Abréviation	Nom et statut	Principales caractéristiques
MPEG	Moving Pictures Expert Group Format ouvert Ensemble de normes ISO	Ce format correspond à une famille de normes ISO : MPEG-1 : les premiers films de l'Internet (ISO/CEI 11172-1 à 5) MPEG-2 : la télévision numérique actuelle MPEG-4 : la Télévision Numérique Terrestre

		MPEG-7, MPEG-21 : futures normes de composition de scènes, très riches en métadonnées. La plus performante actuellement est MPEG-4 qu'il convient d'utiliser avec un codage nommé H264. H.264, ou MPEG-4 AVC (Advanced Video Coding), est une norme de codage vidéo développée conjointement par l'UIT (Union Internationale des télécommunications) et l'ISO. La norme UIT-T H.264 et la norme MPEG-4, Part 10 (ISO/CEI 14496-10) sont techniquement identiques,
MJPEG2000	Motion JPEG2000 Partie 3 de la norme JPEG2000 Format publié et ouvert Norme ISO/CEI 15444-1	Chaque image de la vidéo est codée au format JPEG 2000. Une vidéo MJPEG 2000 n'est qu'une simple concaténation d'images au format JPEG 2000, incluant quelques modifications sur les en-têtes. Ce format est bien adapté pour l'indexation ou le montage vidéo. Il bénéficie de toutes les propriétés de JPEG 2000 en particulier le codage sans perte.
OGG	Format publié et ouvert	Format ouvert promu par la fondation Xiph.org. Ce format doit être utilisé avec le codage Theora. Il s'agit du mode de compression vidéo libre et sans brevets promu par la même fondation.
Matroska (extension mkv)	(Матрёшка ou Poupée russe en russe) Format ouvert	Matroska est un format qui peut regrouper au sein d'un même fichier plusieurs pistes vidéo et audio ainsi que des sous-titres et des chapitres. On peut, comme pour MPEG 4, recommander le codage H264.
AVI	Audio Video Interleave Format propriétaire (Microsoft) Format conteneur publié	Ce conteneur peut accueillir n'importe quel codec. En mode non comprimé, les fichiers sont rapidement très volumineux.
WMV	Windows Media Video Format propriétaire (Microsoft)	Peu recommandé pour l'archivage

5.5. Formats des fichiers produits par les applications « maison »

Jusqu'à maintenant, nous avons essentiellement parlé des formats de représentation ouverts ou propriétaires, qui sont disponibles sur le marché et pour lesquels il existe des logiciels d'écriture, de lecture, de conversion, etc.

De nombreuses entreprises ou institutions développent leurs propres applications logicielles afin de répondre à leurs besoins. Les données numériques produites par ces applications ont donc un format propre à ces applications.

Au lieu de s'appuyer sur une documentation descriptive du format disponible dans un organisme de normalisation ou chez son propriétaire, il sera nécessaire ici, de réaliser sa propre documentation descriptive des données. Cette description devra impérativement :

- être complète,
- être précise,
- avoir été validée de façon rigoureuse.

Attention : Essentiel

La validité et la complétude de cette description du format des données issues de l'application « maison » sont des éléments déterminants pour l'archivage des données issues de cette application. Toute non-conformité de cette documentation entraîne immédiatement un risque majeur de perte ou d'interprétation fausse de l'information archivée.

La présentation des méthodes de description des formats de données sort du cadre du présent cours mais il est bon de savoir :

- que des méthodes de description formelle existent. Ces méthodes permettent de produire une description du format qui sera interprétable aussi bien par des personnes que par des logiciels,
- que des outils permettent alors de s'assurer de la cohérence entre un document numérique et la description formelle de son format.

Chapitre 6. Ce fichier est-il au bon format ?

L'extension du nom de fichier (.pdf, .doc, .xml…) permet a priori de rapidement de connaître le type de données. Cette extension, surtout utilisée dans le monde windows est insuffisante car plusieurs types de données partagent la même extension. Par exemple pour PDF, nous pouvons avoir :

- Systems Management Server (SMS) Package Description File (Microsoft Corporation)
- ArcView Preferences Definition File (ESRI)
- Netware Printer Definition File
- **Acrobat Portable Document Format (Adobe Systems Inc.)**
- P-CAD Database Interchange Format (Altium Limited)
- Package Definition File
- etc.

Complément

Le magic number permet une identification du format d'un fichier par analyse des premières données de ce fichier. Cette technique est utilisée par les systèmes d'exploitation MacOS et Unix. Le magic number offre une fiabilité un peu supérieure à l'extension dont la valeur peut être changée volontairement ou non très facilement.

Par exemple, le magic number aura pour valeur %PDF-1 pour les fichiers PDF

TIFF aura pour magic number MM.* ou II* suivant que le fichier sera constitué en gros-boutiste ou en petit-boutiste.

Cependant, c'est seulement l'analyse complète des données qui nous permettra de nous assurer que le fichier est bien conforme aux spécifications du format auquel il prétend être.

Attention : Essentiel

L'opération de validation des formats des fichiers entrant dans l'Archive constitue une opération critique. Si les fichiers transférés par le producteur ne respectent pas complètement les spécifications du format et si de surcroît, les non-conformités ne sont pas explicitement mises en évidence par les outils actuels de lecture, on induit alors un risque important pour l'Archive qui devient responsable de la pérennisation dès lors que le fichier transféré a été accepté.

Il est fréquent par exemple que des fichiers PDF ne soient pas conformes à la spécification publiée par Adobe sans que les anomalies soient signalées par l'outil de lecture gratuit Acrobat Reader. La validation des formats en entrée implique l'usage et éventuellement le développement d'outils de contrôle de ces formats et la mise en œuvre de procédures très rigoureuses.

Exemple : C'est une vraie question à se poser: mon fichier est-il au bon format ?

Voici quelques exemples réels, rencontrés dans le cadre de l'expérimentation de la plate-forme pilote d'archivage électronique PIL@E (direction des Archives de France), qui montrent la difficulté de cette question.

• les fichiers HTML, versés par les services de l'administration, comportent pratiquement tous de très nombreuses erreurs de syntaxe qui devraient tous les faire rejeter par l'outil de validation, ce qui évidemment est problématique. À l'inverse, si on les accepte, un certain nombre de ces erreurs pourront rendre difficiles à opérer les migrations de format à venir,

• la validation des fichiers XML implique que l'on se réfère aux schémas ou DTD externes. Ceci impose donc d'avoir au préalable récupéré ces modèles de documents et de les conserver dans le système d'archivage,

• la validation des fichiers vidéo au format MPEG se heurte à la grande permissivité des logiciels de lecture de ce format,

• pour la conversion des fichiers graphiques (de type autocad), aucune solution satisfaisante n'a pu être mise en œuvre.

Un autre exemple est celui du CINES sigle à développer qui a analysé environ 150 000 fichiers PDF du serveur HAL (Hyper articles en ligne) en vue de leur archivage. Les résultats sont éloquents :

• plus de 11% des fichiers ne sont pas recevables pour l'archivage, soit parce que leur structure est incorrecte, soit par ce que leur structure est correcte mais non conforme au modèle attendu pour les fichiers PDF,

• toutes les versions du format PDF sont présentes depuis la version 1.0 à la version 1.7 (8 versions),

• une vingtaine d'outils logiciels de génération de fichiers PDF ont été identifiés,

• sur cet ensemble, 14 logiciels ont généré des fichiers invalides de façon non systématique. Parmi ces outils, on trouve même Acrobat Distiller distribué par Adobe, propriétaire du format PDF.

Un certain nombre de logiciels de validation des formats ont été développés.

Exemple : Citons en particulier :

• JHOVE (JSTOR/Harvard Object Validation Environment) qui permet de valider la conformité des fichiers par rapport à un certain nombre de formats parmi lesquels AIFF (Audio Interchange File Format, format audio de Apple), GIF, HTML, PDF, TIFF, JPEG (Joint Photographic Experts Group), XML, WAVE (format audio de Microsoft), etc. JHOVE est un logiciel libre sous licence GNU GPL,

• DROID (Digital Record Object Identification) est un outil Open source fourni par les Archives nationales du Royaume-Uni. Il s'appuie sur le registre de format PRONOM et relie l'identification du format aux documents techniques correspondants qui sont disponibles dans le registre,

Une autre initiative utile est celle du CINES qui a consisté à mettre en ligne un service de validation de formats basé sur JHOVE, DROID et autres outils. Ce service, nommé « Facile » permet de ne pas avoir à installer de logiciels de validation de format.

Attention : Conclusions sur les formats

La représentation de l'information numérique est la clé de voûte de la pérennisation, elle contrôle l'accès à toute l'information : données et métadonnées.

Mais c'est aussi le lieu d'enjeux commerciaux, la source possible sinon probable de nombreuses difficultés techniques, d'où la nécessité de se regrouper pour partager l'expérience et les outils, pour peser sur les choix de formats.

Questions : FORMATS DE REPRESENTATION DE L'INFORMATION

Objectifs

Avez-vous compris tout ce qui vient de vous être enseigné ?

Si vous voulez le vérifier, faites les exercices proposés ci-dessous.

Si vous ne savez pas répondre, ne regardez pas trop vite le corrigé, travaillez à nouveau la (les) section(s) précédente(s) où vous découvrirez les solutions.

Bien sûr, si vous n'y arrivez vraiment pas, vous pouvez consulter les réponses. Ne les lisez pas avec précipitation mais avec une grande attention et surtout essayez de comprendre.

Remarque sur la limite de nos exercices

Nous vous proposons des exercices sur le numérique avec toutes les réserves que cela comporte en raison notamment des évolutions technologiques. Nous sommes en effet dans une discipline récente qui est en constante évolution. Les principes qui ont été posés dans ce module resteront sûrement durablement valides, par contre le nombre de solutions aujourd'hui valides ne le seront plus que partiellement demain ou plus du tout.

Par exemple : dans un contexte organisationnel donné, ce que nous pouvons recommander en matière de formats ou de moyens de stockage, pourra être remis en cause dans un ou deux ans.

Il est donc nécessaire que les partenaires d'un projet d'archivage numérique s'approprient les principes mais ne procèdent à l'analyse de la situation et des contraintes qu'au moment où le projet se met en marche et cela avec les partenaires concernés ; en effet, alorsqu'un archiviste peut se trouver seul au sein de son organisation, l'archivage numérique ne doit pas être une activité individuelle : ce sera toujours la mise en commun d'un ensemble de compétences complémentaires. Donc rien ne remplacera les exercices en vraie grandeur.

Nous rappelons que les exercices du PIAF sont à l'usage d'une auto-formation. Nous proposons à cet effet des types d'exercices de mémorisation, d'accompagnement, de positionnement afin de permettre à l'utilisateur de vérifier l'acquisition d'une culture minimale précise.

Les exercices ci-dessous porteront surtout sur les principes puisqu'il n'est pas possible d'aller trop loin en matière de solution.

Exercice 1

[*Solution n°1 p 24*]

La norme internationale ISO 10646 (correspondant au standard Unicode) permet de normaliser la représentation numérique et la représentation graphique des caractères.

- ○ Vrai
- ○ Faux

Exercice 2

[*Solution n°2 p 24*]

Un format propriétaire est un format défini par une entreprise ou un propriétaire privé et qui n'a pas été rendu public.

- ○ Vrai
- ○ Faux

Exercice 3

[*Solution n°3 p 24*]

L'extension du nom d'un fichier sous Windows suffit pour savoir de façon sûre de quel type de fichier il s'agit

- ○ Vrai
- ○ Faux

Exercice 4

[*Solution n°4 p 25*]

Le premier critère est prendre en compte dans le choix d'un format d'archivage est :

- ☐ • Le format doit être adapté aux besoins des utilisateurs
- ☐ • Le format doit être indépendant, ouvert et si possible normalisé
- ☐ • Le format doit permettre une compression efficace pour réduire le volume de l'archive
- ☐ • Le format permet la vérification automatique de sa conformité par rapport à ses spécification

Solution des exercices

> Solution n°1 *(exercice p. 22)*

○ Vrai

la norme internationale ISO 646 permet de normaliser la représentation numérique des caractères, elle normalise également le nom de ces caractères (par exemple « Lettre minuscule grecque alpha », mais elle ne normalise pas la forme graphique correspondante qui sera en fait différente d'une police de caractère à une autre.

◉ Faux

La norme internationale ISO 646 permet de normaliser la représentation numérique des caractères, elle normalise également le nom de ces caractères (par exemple « LETTRE MINUSCULE GRECQUE ALPHA », mais elle ne normalise pas la forme graphique correspondante qui sera en fait différente d'une police de caractère à une autre.

> Solution n°2 *(exercice p. 22)*

○ Vrai

◉ Faux

Le format propriétaire est un format défini par une entreprise ou un propriétaire privé qui dispose des droits de propriété intellectuelle ou du copyright correspondant (par exemple PDF, TIFF, GIF…), mais suivant les cas, le format aura été rendu public ou non :

• Le format PDF est un format propriété de la société Adobe. Les spécifications de ce format sont néanmoins publiques,

• Le format DWG, propriété de la société Autodesk est un format propriétaire fermé non publié.

> Solution n°3 *(exercice p. 23)*

○ Vrai

◉ Faux

Plusieurs raison à cela. D'abord parce qu'il peut exister des extensions qui ont plusieurs significations (voir le cas de PDF coté dans le chapitre), en outre, il est possible, volontairement ou non, de modifier cette information sans modifier le fichier pour autant. Enfin, même si l'extension nous apporte une certaine information, elle ne permettra pas de savoir si le ficher en question respecte vraiment les spécifications du format. Pour cela, ce sont des outils de validation de format qu'il faudra utiliser.

> Solution n°4 *(exercice p. 23)*

☐	• Le format doit être adapté aux besoins des utilisateurs *faux : il sera toujours possible de transformer le format d'archive en un format adapté aux besoins des utilisateurs*
☑	• Le format doit être indépendant, ouvert et si possible normalisé
☐	• Le format doit permettre une compression efficace pour réduire le volume de l'archive *faux : la préoccupation sur le volume ne concerne que les très grandes Archives et passe toujours après le besoin d'avoir un format indépendant, ouvert et si possible normalisé*
☐	• Le format permet la vérification automatique de sa conformité par rapport à ses spécification *faux : critère de second ordre*

Glossaire

Format de données, ou format de fichier ou format de représentation de l'information :

le format de données peut être défini par l'ensemble des règles et algorithmes permettant d'organiser l'information dans un objet numérique.

Par exemple, le format de données permettra de :

- spécifier le codage des couleurs des pixels d'une image, définir un algorithme de compression des données et l'organisation de ces données dans un fichier (formats PNG, TIFF…),
- spécifier l'organisation et la structuration d'informations textuelles à partir de l'encodage élémentaire des caractères (formats SGML, XML) ;

en réalité, SGML et XML sont en premier lieu des langages comportant un ensemble de règles, une syntaxe, des mots clés permettant de constituer des documents structurés ; lorsqu'un document a été structuré par le langage XML, on connaît en pratique l'ensemble des règles d'organisation de l'information au sein de ce document ; à ce titre, XML (comme SGML) peut donc être considéré comme un format,

- définir comment les quatre informations élémentaires que sont la mantisse (nombre entier positif), l'exposant (nombre entier positif), le signe de l'exposant et le signe de la mantisse (caractères + et -) sont organisées pour représenter un nombre réel sous forme numérique (cf. standard ANSI/IEEE 754-1985).

Format de données, ou format de fichier ou format de représentation de l'information :

le format de données peut être défini par l'ensemble des règles et algorithmes permettant d'organiser l'information dans un objet numérique.

Par exemple, le format de données permettra de :

* spécifier le codage des couleurs des pixels d'une image, définir un algorithme de compression des données et l'organisation de ces données dans un fichier (formats PNG, TIFF...),

* spécifier l'organisation et la structuration d'informations textuelles à partir de l'encodage élémentaire des caractères (formats SGML, XML) ;

en réalité, SGML et XML sont en premier lieu des langages comportant un ensemble de règles, une syntaxe, des mots clés permettant de constituer des documents structurés ; lorsqu'un document a été structuré par le langage XML, on connaît en pratique l'ensemble des règles d'organisation de l'information au sein de ce document ; à ce titre, XML (comme SGML) peut donc être considéré comme un format,

* définir comment les quatre informations élémentaires que sont la mantisse (nombre entier positif), l'exposant (nombre entier positif), le signe de l'exposant et le signe de la mantisse (caractères + et -) sont organisées pour représenter un nombre réel sous forme numérique (cf. standard ANSI/IEEE 754-1985).

Bibliographie

[Premier ouvrage de synthèse sur l'archivage numérique en langue française.] • BANAT-BERGER F., HUC C., DUPLOUY L., L'Archivage numérique à long terme, les débuts de la maturité? Paris, La Documentation française, 2009.

Webographie

[Norme de référence essentielle pour comprendre le problème posé par l'archivage numérique] http://public.ccsds.org/publications/archive/650x0b1(F).pdf