Les formats de fichiers

Version 1.0
F. Lemainque
Septembre 2016

1 Introduction

Un fichier, au sens informatique du terme, peut contenir tout type de document (ou type de données) : logiciel, texte, son, vidéo, image animée ou non...et même bien sûr du code source quelconque. Toutefois, aussi diverses soit-elles, les données sont toutes représentées en définitive par des 0 et des 1 (représentation binaire) pour qu'elles puissent être interprétables par l'ordinateur.
Une fois représentées grâce à une première mise en forme (ou format), les données peuvent être stockées (enregistrées) sur support informatique *via* un fichier possédant lui aussi un format spécifique. Pour réaliser le fichier qui va permettre le stockage de ces données, il est défini un format de fichier. Autrement dit, chaque logiciel qui crée un fichier, quel qu'il soit, utilise un format d'enregistrement spécifique pour le créer. Ce format répond à une convention sur la représentation d'une donnée sur un support numérique, qui peut être :

- **Spécifié** : il existe une description de la convention utilisée pour représenter la donnée et celle-ci est suffisamment décrite pour en développer une implémentation complète.
- **Ouvert** : la convention est publique (sinon le format est dit fermé). Elle est donc sans restriction d'accès ni de mise en œuvre.
- **Normalisé** : la convention est adoptée par des organismes de normalisation (ISO, W3C). Par exemple, XML, PDF, PNG, OpenDocument.
- **Standardisé** : il n'existe pas de norme sur ce format mais son utilisation est tellement répandue qu'il est considéré comme un standard. Exemple : DOC, JPG.
- **Propriétaire** : si l'exploitation du format entre dans le cadre du droit privé, il dépend de l'existence du propriétaire. Il peut toutefois être publié. Exemple : PDF, DOC.

La notion de format étant un peu difficile à cerner car très abstraite, illustrons les choses à l'aide du tableau suivant :

A	B	C
B	C	B

Problème : Comment peut-on représenter de manière informatique ce tableau ainsi que les informations textuelles qu'il contient ?

Solution : La première étape consiste à choisir le format des données textuelles à utiliser. Il existe de nombreux formats dont les plus utilisés sont les formats ASCII et Unicode (que nous allons examiner plus loin). Le principe est simple : chaque caractère (lettre, nombre, symbole...) correspond à un nombre donné. Ce nombre est représenté de manière binaire sous forme de 0 et de 1.
Ainsi, en ASCII :

- pour le "A" on utilise le code binaire 1000001 (égal au nombre décimal 65)
- pour le "B" on utilise le code binaire 1000010 (égal au nombre décimal 66)
- pour le "C" on utilise le code binaire 1000011 (égal au nombre décimal 67)

Une fois le texte codé, il reste à écrire dans le fichier d'une manière ou d'une autre qu'il se trouve dans un tableau composé de deux lignes et trois colonnes avec le A dans la première case, le B dans la seconde et le C dans la troisième... Il faut donc recourir à une convention de représentation (format de fichier). Il en existe là aussi un grand nombre mais la plupart sont un peu complexes. Servons-nous d'un exemple simplifié : la virgule « , » représente une nouvelle colonne et le point virgule « ; » représente une nouvelle ligne. Nous obtenons alors le résultat final suivant :

```
; , 1000001 , 1000010 , 1000011 ; , 1000010 , 1000011 , 1000010
```

Et voilà, le tableau a été traduit dans un fichier qui pourra être reconstitué en un tableau de lettres, en appliquant l'opération inverse.

La standardisation des formats

Le codage que nous avons retenu dans l'exemple précédent aurait pu être effectué différemment. Les séparateurs pourraient être autres, on pourrait indiquer d'une façon ou d'une autre qu'il n'y a qu'une valeur dans chaque case, etc. Il est donc logique qu'au fil du temps se soient développés un grand nombre de formats souvent propriétaires dans une premier temps (car liés à un logiciel).
Un format est dit **propriétaire** s'il a été élaboré par une entreprise, dans un but essentiellement commercial. Un format propriétaire peut être ouvert (le format PDF d'Adobe par exemple) s'il est publié, ou fermé (le format Doc de Microsoft par exemple). Mais même lorsque des spécifications sont rendues publiques, les entreprises à l'origine de formats propriétaires tentent d'en conserver le contrôle soit en proposant régulièrement de nouvelles versions plus élaborées (contrôle par maintien d'une avance technologique), soit en utilisant des moyens juridiques comme le brevet. Ce type de pratiques anti concurrentielles *via* des outils juridiques est admis aux Etats-Unis, mais sujette à controverse en Europe.
Comment donc s'assurer que votre ordinateur puisse lire n'importe quel format alors que tout individu peut créer son propre format et imposer un logiciel pour le lire ? Pour résoudre le problème et couper court à l'anarchie en la manière, des formats standards ont été proposés, normalisés et validés par une institution publique ou internationale (ISO, W3C), tandis que d'autres s'imposaient comme standard de par leur omniprésence sur le marché. Certains ont été normalisés par la suite, comme PDF (respectivement ISO 19005-1 et ISO 15930) ou OpenDocument (ISO 26300:2006).

Extension de fichier

Il demeure actuellement des centaines de formats standards pour tous les types de fichiers (vidéos, images, sons, texte, données structurées, archives...). Pour reconnaître le format d'un fichier et en connaître par conséquent le type, il suffit d'examiner son **extension** : tout fichier est identifié par deux parties distinctes, son nom et son extension sous la forme NOM.EXT.

L'extension est capitale. Elle identifie la nature du fichier, signale au système d'exploitation ce qu'il est capable d'en faire et quel logiciel il conviendra de lancer pour lire les informations contenues dans ce fichier. Par exemple, un fichier avec l'extension .doc lancera le logiciel Word et un fichier .xls, le logiciel Excel.
Certaines extensions plus ouvertes nécessitent une affectation manuelle dans la table de correspondance. Ce sera notamment le cas des fichiers images (.jpeg ou .jpg ; .gif ; .png, etc) qui peuvent être lus dans la plupart des visualiseurs d'images. Il faudra donc, dans ce cas, choisir un logiciel graphique ou encore le navigateur web associé à telle ou telle extension.

Astuce : sous Windows, par défaut, les extensions de fichiers sont masquées. Pour les afficher, ouvrir le « Poste de travail » ; dans le menu « Outils », cliquez sur « Options des dossiers ». Une fenêtre s'ouvre avec quatre onglets (en haut). Cliquez sur « Affichage ». Vérifiez que l'option « Masquer les extensions des fichiers dont le type est connu » est décochée puis cliquez au dessus sur « Appliquer à tous les dossiers ». Désormais, les extensions de tous les fichiers sont affichées. Mieux vaut toujours afficher les extensions.

Dans un envoi par courriel, il est essentiel de bien mettre l'extension : le protocole de transmission peut détériorer le fichier en l'absence d'indication sur la nature du format. Ce n'est pas systématique, mais cela arrive fréquemment. L'autre point à respecter est la règle suivante : ne jamais employer dans un nom de fichier d'espace ou caractère de ponctuation autre que le tiret (-), le souligné (_) et le point avant l'extension. Mieux vaut ne jamais employer de caractère accentué ni de caractère fantaisiste, type %, $, *, etc. Voici quelques exemples de noms de fichiers corrects et de noms pouvant poser problème :

Noms corrects :	Noms incorrects :	Erreur commise :
monfichier.doc	mon fichier.doc	espace
mon_image.jpg	mon_image	pas d'extension
accordeon.mp3	accordéon.mp3	caractère accentué
film003.mov	film:003.mov	ponctuation autre que le . avant l'extension
fichier-copie.doc	fichier.doc-copie	ne termine pas avec l'extension
animation_01_03.swf	animation/01/03.swf	caractère spécial
exemple-test.rft	exemple&test.rtf	caractère spécial

Si l'utilisation de minuscules ou/et de majuscules ne pose en principe pas de problème, sachez toutefois que si les PC sous Windows ne font pas la différence entre les minuscules et les majuscules, les Macintosh et les machines sous Unix ou Linux font cette différence. Il est donc plus prudent de vous en tenir à une convention stricte.

2 Les jeux de caractères

Au commencement était l'ASCII

A l'origine du système actuel de codage des ordinateurs se trouve le standard ASCII (*American Standard Code for Information Interchange*), créé dans les années soixante du siècle dernier et fixé en 1983. Il représente le codage numérique (et donc l'utilisation) de 128 signes (= 2^7), soient 33 signes pour les commandes (utiles essentiellement, à l'origine, pour les communications par télex), 52 signes pour les 26 lettres minuscules et majuscules, 10 pour les chiffres, 21 pour les signes de ponctuation et 12 autres qui pouvaient initialement être utilisés diversement selon les langues, mais l'usage américain dominant s'est imposé. Voici la liste des 128 signes graphiques de l'ASCII :

Dec	Hx	Oct	Char	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL (null)	32	20	040		Space	64	40	100	@	@	96	60	140	`	`
1	1	001	SOH (start of heading)	33	21	041	!	!	65	41	101	A	A	97	61	141	a	a
2	2	002	STX (start of text)	34	22	042	"	"	66	42	102	B	B	98	62	142	b	b
3	3	003	ETX (end of text)	35	23	043	#	#	67	43	103	C	C	99	63	143	c	c
4	4	004	EOT (end of transmission)	36	24	044	$	$	68	44	104	D	D	100	64	144	d	d
5	5	005	ENQ (enquiry)	37	25	045	%	%	69	45	105	E	E	101	65	145	e	e
6	6	006	ACK (acknowledge)	38	26	046	&	&	70	46	106	F	F	102	66	146	f	f
7	7	007	BEL (bell)	39	27	047	'	'	71	47	107	G	G	103	67	147	g	g
8	8	010	BS (backspace)	40	28	050	(	(	72	48	110	H	H	104	68	150	h	h
9	9	011	TAB (horizontal tab)	41	29	051	)	)	73	49	111	I	I	105	69	151	i	i
10	A	012	LF (NL line feed, new line)	42	2A	052	*	*	74	4A	112	J	J	106	6A	152	j	j
11	B	013	VT (vertical tab)	43	2B	053	+	+	75	4B	113	K	K	107	6B	153	k	k
12	C	014	FF (NP form feed, new page)	44	2C	054	,	,	76	4C	114	L	L	108	6C	154	l	l
13	D	015	CR (carriage return)	45	2D	055	-	-	77	4D	115	M	M	109	6D	155	m	m
14	E	016	SO (shift out)	46	2E	056	.	.	78	4E	116	N	N	110	6E	156	n	n
15	F	017	SI (shift in)	47	2F	057	/	/	79	4F	117	O	O	111	6F	157	o	o
16	10	020	DLE (data link escape)	48	30	060	0	0	80	50	120	P	P	112	70	160	p	p
17	11	021	DC1 (device control 1)	49	31	061	1	1	81	51	121	Q	Q	113	71	161	q	q
18	12	022	DC2 (device control 2)	50	32	062	2	2	82	52	122	R	R	114	72	162	r	r
19	13	023	DC3 (device control 3)	51	33	063	3	3	83	53	123	S	S	115	73	163	s	s
20	14	024	DC4 (device control 4)	52	34	064	4	4	84	54	124	T	T	116	74	164	t	t
21	15	025	NAK (negative acknowledge)	53	35	065	5	5	85	55	125	U	U	117	75	165	u	u
22	16	026	SYN (synchronous idle)	54	36	066	6	6	86	56	126	V	V	118	76	166	v	v
23	17	027	ETB (end of trans. block)	55	37	067	7	7	87	57	127	W	W	119	77	167	w	w
24	18	030	CAN (cancel)	56	38	070	8	8	88	58	130	X	X	120	78	170	x	x
25	19	031	EM (end of medium)	57	39	071	9	9	89	59	131	Y	Y	121	79	171	y	y
26	1A	032	SUB (substitute)	58	3A	072	:	:	90	5A	132	Z	Z	122	7A	172	z	z
27	1B	033	ESC (escape)	59	3B	073	;	;	91	5B	133	[	[	123	7B	173	{	{
28	1C	034	FS (file separator)	60	3C	074	<	<	92	5C	134	\	\	124	7C	174			|
29	1D	035	GS (group separator)	61	3D	075	=	=	93	5D	135	]	]	125	7D	175	}	}
30	1E	036	RS (record separator)	62	3E	076	>	>	94	5E	136	^	^	126	7E	176	~	~
31	1F	037	US (unit separator)	63	3F	077	?	?	95	5F	137	_	_	127	7F	177	&#127;	DEL

Source: www.LookupTables.com

Il est assez évident que ce nombre réduit de signes, s'il suffit pour le codage des caractères usuels de l'anglo-américain, ne permet pas le codage des graphèmes spécifiques d'autres langues européennes, ni même d'une seule (le français nécessite 92 signes rien que pour les lettres et les chiffres).

A partir du moment où les logiciels de traitement de texte se sont développés et diffusés dans le monde, il a fallu l'étendre à 256 signes (2^8). Mais les numéros de code de 128 à 255 n'ont pas été attribués de la même façon sur les différentes plateformes (Windows, Unix ou Mac), d'où l'existence de plusieurs standards (avec des règles de conversion d'un standard à l'autre et éventuellement quelques petits problèmes). Et de toute façon, 256 signes ne suffisent pas non plus à coder les seules langues à alphabet latin (de l'anglais au vietnamien).

La solution adoptée a consisté à définir des jeux de caractères (en anglais *character set*, en abrégé *charset*) correspondant aux besoins spécifiques de groupes de langues : c'est le standard connu sous le nom de ISO 8859). Ainsi, le premier de ces jeux, « Latin-1 » (8859-1), contient les caractères nécessaires pour l'écriture des langues à écriture latine d'Europe occidentale (langues germaniques, romanes – à l'exception du roumain pour lequel manquent quelques signes –, celtiques, basque, albanais). D'autres jeux ont été définis ensuite pour les autres langues à écriture latine (notamment pour les langues slaves – « Latin-2 » = 8859-2 –, les langues baltes, le turc), et également pour corriger quelques défauts du jeu « Latin-1 » (dans lequel, manquaient les signes œ et Œ nécessaires pour le français) et pour regrouper différemment les langues en fonction de préoccupations économiques, politiques ou culturelles (exemple : vaut-il mieux avoir un jeu permettant à la fois l'écriture des langues occidentales et du turc ou des langues occidentales et de l'islandais ?).
Etaient également définis des standards pour les langues à écriture alphabétique non latine : cyrillique (8859-5), arabe (8859-6), grec (8859-7), hébreu (8859-8).
Mais cette norme ne règle que partiellement le problème :

- les regroupements de langues correspondant aux différents jeux de caractères latins sont plus ou moins pertinents... et plus ou moins arbitraires ;
- on ne peut utiliser qu'un seul jeu de caractères à la fois dans un même texte (par exemple une page web) : comment faire si je veux avoir un texte bilingue vietnamien et russe ?

Et pour les milliers de caractères du chinois, le japonais et le coréen ? Ils peuvent être codés sur deux octets (soient 2 fois 8 bits), mais là aussi se pose la question d'une standardisation : certains caractères sont spécifiques au japonais, au coréen et au chinois, d'autres communs, etc.

Unicode

La réponse à ces problèmes est l'attribution d'un code unique à tous les caractères utilisés dans les différentes langues du monde et donc la définition d'un jeu unique, universel, de caractères : c'est le standard Unicode (*Universal Character Set*, UCS). Après des initiatives séparées menées notamment par Xerox et Apple, le consortium Unicode a été créé en 1991, il regroupe de nombreux (et les plus gros) partenaires privés et publics du monde de l'informatique. La dernière version du standard Unicode, 9.0 est disponible depuis 2016. Elle dépasse les 100 000 caractères, et des projets nombreux existent d'ores et déjà pour le codage d'autres écritures.
Le standard Unicode est identique à la norme internationale ISO 10646. Il repose sur deux principes essentiels :

1. Distinction entre écriture et langue.
   Unicode définit les caractères indépendamment de leur utilisation spécifique dans les langues. C'est, en quelque sorte, le même principe que celui régissant l'alphabet, qui sert pour l'écriture de multiples langues. Les signes Unicode ne correspondent donc pas aux graphèmes d'une langue donnée, ce sont les signes avec lesquels on peut saisir ces graphèmes. Le standard contient un certain nombre de redondances : ainsi, pour le

hangul (coréen), Unicode contient les signes des sons individuels (jamos) et les signes (plus de 10 000) correspondant aux syllabes formées à partir de ces sons. De même, les signes diacritiques sont répertoriés pour eux-mêmes, ainsi que les combinaisons entre lettres de l'alphabet et ces mêmes signes diacritiques (voir ci-dessous l'exemple de a).

2. Distinction entre caractère (« abstrait ») et glyphes (formes géométriques d'un caractère).
   Unicode attribue un code à chaque caractère, les caractères étant conçus comme des entités abstraites : « les plus petites composantes de la langue écrite auxquelles on peut faire correspondre une valeur linguistique ». Chaque caractère représente une classe de glyphes, c'est-à-dire de représentations possibles selon la police, le corps, le style, la graisse, etc., sans oublier les réalisations manuscrites.

Abstrait par rapport aux graphèmes qu'il peut représenter, le caractère Unicode l'est aussi par rapport aux glyphes auxquels il peut correspondre. Et cette double distinction est évidemment problématique. Trois exemples :

- l'écriture gothique, dans la mesure où elle contient les mêmes caractères que l'écriture romaine, n'a pas de codes spécifiques (sauf pour quelques signes utilisés en mathématiques), mais dans certains cas, la distinction gothique/romain peut avoir une pertinence linguistique – de la même façon que d'autres distinctions typographiques (italique, gras, etc.) ;
- l'écriture gothique, à la différence de l'écriture en romains, contient deux "s" différents selon qu'il est à l'intérieur d'un mot ou non : comment assurer cette distinction si le standard Unicode ne contenait qu'un seul caractère "s" minuscule ?
- les guillemets ouvrant et fermant possédant – globalement – la même fonction en français, en anglais ou en allemand, il serait donc cohérent qu'ils soient représentés par deux caractères seulement (guillemet ouvrant et guillemet fermant), mais leur forme usuelle est différente – d'où plusieurs caractères différents spécifiés partiellement par rapport à la langue.

UTF-8

En théorie, Unicode c'est parfait. Sauf qu'en Unicode, un caractère est codé sur deux octets : deux fois plus que pour de l'ASCII. Alors que, en anglais la totalité (et en français la grande majorité) des caractères utilisent seulement le code ASCII. Seuls quelques rares caractères nécessitent l'Unicode. C'est donc un gaspillage de place incontestable.
L'UTF-8 répond à ce problème, grâce à une astuce simple : il est partout en ASCII, sauf lorsqu'un caractère Unicode est nécessaire. Il est alors inséré un caractère spécial signalant « attention, le caractère suivant est en Unicode ». Et pour prévenir qu'un fichier est codé en UTF-8, on ajoute au début du fichier des caractères spéciaux.
L'UTF-8 rassemble le meilleur de deux mondes : l'efficacité de l'ASCII et l'étendue de l'Unicode. Il a été adopté comme norme pour l'encodage des fichiers XML. La plupart des navigateurs récents prennent en charge l'UTF-8 et le détectent automatiquement dans les pages HTML.

Et pour une page Web ?

Vous le savez probablement, écrire directement des caractères accentués dans une page web, ce n'est pas bien. Vous devez impérativement choisir une des trois solutions suivantes :

- recourir aux entités HTML, et donc mettre `&eacute;` à la place de « é ».
- laisser le « é » tel quel mais préciser le charset employé dans l'élément `head` du fichier HTML, dans un élément `meta` :

```
<meta http-equiv="Content-type" content="text/html; charset=ISO-8859-1">
```

  ISO-8859-1 est le jeu de caractère latin habituel sous Windows.
- soit travailler directement en UTF-8 dans votre éditeur HTML (la plupart le permettent) et ajouter dans l'élément meta du head :

```
<meta http-equiv="Content-type" content="text/html; charset=UTF-8">
```

L'ISO-8859-1 convient pour la plupart des langues latines ou occidentales (anglais, français, allemand, espagnol...), et l'UTF-8 vous sera indispensable pour les autres langues (japonais, hébreu, etc.). Toutefois, pour une sécurité et une adaptabilité maximale je vous recommande fortement d'employer systématiquement UTF-8 dans votre éditeur, et ce pour tous les types de code.

3 Transmission de fichiers

Vous voudrez transmettre un fichier à quelqu'un pour deux raisons :

- Permettre à l'interlocuteur uniquement de les lire ;
- ou permettre à l'interlocuteur de les modifier.

Souvent, dans le second cas cet interlocuteur sera d'ailleurs vous-même : vous voudrez transférer un fichier d'un logiciel dans un autre…

Fichiers destinés uniquement à la lecture

Désormais, sauf exception rarissime, tout fichier écrit ou imprimable destiné à une simple lecture doit être exporté au format PDF (*Portable Document Format*), lisible sous Windows avec le client gratuit Acrobat Reader, sous MacOS avec Aperçu et sous Linux avec xpdf. La plupart des logiciels permettent désormais d'exporter directement au format pdf, sinon faites appel à un logiciel comme Foxit Reader, capable de générer des fichiers PDF grâce à une pseudo-impression.

Le format PDF est un format de diffusion et de conservation normalisé par l'ISO (Organisation Internationale de Normalisation) sous les normes PDF/A-1 et PDF/X (respectivement ISO 19005-1 et ISO 15930). C'est désormais un format ouvert qui devrait permettre de pérenniser l'archivage électronique, idéal pour transmettre des documents : sa mise en forme est conservée quel que soit le logiciel ou le système d'exploitation de l'usager.

Les fichiers .pdf ne sont pas modifiables aisément, de par le statut même du format pdf : c'est un format de diffusion et non d'édition, ce qui peut aussi présenter un intérêt certain. D'ailleurs, il est hautement recommandé, si vous avez rédigé un document nécessitant une mise en page soignée, de l'exporter en vue d'une impression sur un autre système au format pdf : les règlages d'imprimantes, même avec le même logiciel, pourraient être différents (ils le sont le plus souvent) et remettraient en cause toute votre mise en page. Un fichier pdf conservera en revanche exactement la mise en page définie lors de sa création. Le seul inconvénient dans ce cas est de ne pas pouvoir (le plus souvent encore) retoucher éventuellement le document.

Fichiers devant (ou pouvant) subir une modification ou un réemploi

Tout dépend du type de fichier concerné et du logiciel qui a servi à le créer ainsi que du logiciel cible. Bien sûr, si vous savez que votre interlocuteur possède exactement le même logiciel, dans la même version, que celui qui a permis de créer le fichier, aucun problème ne se pose : vous lui transmettez simplement le fichier tel que vous l'avez enregistré.
En revanche, les choses se compliquent lorsqu'il s'agit de logiciels différents, voire de versions différentes. Attention : les listes présentées ci-dessous n'ont rien d'exhaustives !

Fichiers texte

C'est le cas le plus simple. Les fichiers texte peuvent être importés dans pratiquement tout logiciel, et lu par tous les éditeurs de texte.

Extension	Description	Lu par
.txt	Format standard. Texte brut (en ASCII, perte d'une partie du formatage)	Tout éditeur de texte ou traitement de texte
.rtf	(*Rich Text Format*) Format texte mis en forme. Conserve l'essentiel du formatage du texte. A très longtemps constitué la seule passerelle employée entre Windows et Mac. Désormais obsolète.	Tout éditeur de texte ou traitement de texte
.htm et .html	(*Hyper Text Markup Language*) Texte au format Web	Tout navigateur, éditeur et traitement de texte.
.csv	Texte délimité. Employé pour les tableurs et les bases de données.	Tout éditeur, traitement de texte, tableur, base de données.
.doc	Format propriétaire Microsoft Word. Peu fiable pour de gros fichiers.	Word et d'autres traitements de texte (Ooo).
.docx	Format Office Open XML. Les formats Open XML sont théoriquement une norme internationale, mais aucun logiciel ne respecte actuellement l'intégralité de la norme.	Word (depuis Office 2007) et d'autres traitements de texte (Ooo)

.odf	(*Open Document Format*) Le format OpenDocument est une norme internationale pour la bureautique (ISO 26300:2006).	Ooo, Word récents.
.pub	Format propriétaire Microsoft de publication assistée par ordinateur. Désormais obsolète.	Publisher

Désormais, depuis pratiquement tout logiciel vous pouvez exporter un fichier selon l'un de ces formats.

Tableurs

Extension	Description	Lu par
.wks	Format propriétaire Microsoft. Désormais obsolète.	Works.
.ods	(*Open Document Format*) Le format OpenDocument est une norme internationale pour la bureautique (ISO 26300:2006).	Ooo, Excel récent.
.xls	Format propriétaire Microsoft Excel. Peu fiable pour de gros fichiers.	Excel et d'autres tableurs (Ooo). Access.
.xlsx	Format Office Open XML. Les formats Open XML sont théoriquement une norme internationale, mais aucun logiciel ne respecte actuellement l'intégralité de la norme.	Excel (depuis Office 2007) et d'autres tableurs de texte (Ooo)

Bases de données

Extension	Description	Lu par
.cvs	Format texte délimité par des virgules	La plupart des logiciels de base de données
.sql	Format texte	SGBD compatibles SQL.
.mdb	Format propriétaire Microsoft Access	Access, Excel

Images

Tous les systèmes d'exploitation courants disposent d'un utilitaire permettant d'ouvrir les fichiers .jpeg, .gif, .tif, .png et .bmp, voir d'en éditer certains.

Extension	Description	Lu par
.bmp	(*Windows bitmap*)	Ne provoque aucune altération de l'image. Format très lourd (par définition, puisqu'en mode point ou *bitmap*).
.gif	(*Graphical Interchange Format File*)	Gére la transparence, léger et donc pratique pour le web. Offre des possibilités d'animations. Limité à 256 couleurs, en perte de vitesse depuis la normalisation de PNG.
.jpg/.jpeg	(*JPEG File Interchange Format*) ISO/CEI 10918-1	Très répandu sur le Web et dans le monde de la photographie. Ne gère pas la transparence. Dégradation de l'image à chaque sauvegarde.
.png	(*Portable Network Graphic*) ISO 15948	Gère la transparence. Pas de dégradation de l'image, mais fichiers plus lourds.
.tiff/.tif	(*Tagged Image File Format*)	Ne provoque aucune altération de l'image. Format lourd.

A propos des images

Il y a parfois de quoi se perdre dans la jungle des unités de mesure employées tant par les dispositifs d'acquisition d'image (scanneurs, appareils photos et caméras numériques) et de restitution (imprimantes, tables traçantes, flasheuses). Le terme le plus employé est celui de dpi (*dots per inch*) : c'est le nombre de points sur un pouce, soit 2,54 cm. Ce terme est employé pour les scanneurs et les imprimantes. De nos jours, une résolution d'impression de 600 dpi est un minimum, une résolution de 2 400 dpi n'offrant plus rien d'exceptionnel. Sachez cependant que les limites sont proches : les professionnels de l'imprimerie ne dépassent que rarement 3 600 dpi pour des reproductions d'art, en employant des papiers spéciaux fort onéreux, une résolution supérieure n'apportant rien et pouvant même dégrader le rendu final.

Les imprimantes noir et blanc (y compris les imprimantes laser N&B) fonctionnent en tons de gris : chaque pixel de l'image de départ est traduite en carré de 16 pixels (4 × 4) afin de pouvoir afficher des teintes différentes. La résolution est donc en fait divisée par 4 par rapport à un dessin au trait (bien qu'elle puisse être augmentée de façon logicielle par l'imprimante). Cette organisation se faisant horizontalement, l'appellation correcte devrait être lpi (ligne par pouce) au lieu de dpi.

En revanche, les appareils photos numériques parlent le plus souvent en nombre total de pixels (jusqu'à 20 millions pour les modèles les plus perfectionnés), tandis que les écrans mélangent la notion de diagonale de l'écran (17, 19, 21 pouces) et de résolution (1280 × 960, 1280 × 1024 etc.). Pas facile de comparer !

Mieux vaut donc convertir cette résolution théorique en nombre de points : soit une photo normale de format 24 × 18 cm, soit 9,6 × 7,2 pouces, avec une résolution de 1200 points par pouce vous avez imprimé (9,6 × 1200) × (7,2 × 1200) = 99 532 800 points : plus de 99,5

millions de points ! En affichant cette photo numérique à l'écran, une telle photo occuperait dans la plupart des cas une superficie théorique équivalente à plusieurs écrans. Elle est donc totalement inexploitable en l'état sur une page Web.

Résolution	Nombre total de points	Résolution en dpi (écran 19')
1024 × 768	786 432	73 dpi
1152 × 864	949 248	82 dpi
1280 × 1024	1 310 720	91 dpi
1600 × 1200	1 900 000	114 dpi

Le problème est similaire lorsque vous voulez placer sur votre site une image provenant d'un scanner à plat. Si la numérisation est effectuée dans le but d'une impression, vous devez choisir une résolution d'entrée égale au produit de la résolution de sortie par le rapport de taille souhaité par rapport à l'original. Par exemple, avec une imprimante réglée sur 600 dpi et un rapport d'impression du double de l'original, vous aurez :

600 × 2 = 1200 dpi

En revanche, en vue d'un affichage écran, une résolution d'entrée de 90 dpi est largement suffisante, voire exagérée, puisqu'il est exceptionnel qu'une image occupe tout l'écran : tout dépend encore une fois du rapport de taille entre l'original et ce que vous souhaitez obtenir à l'écran.

Le tableau ci-dessous présente une comparaison de la taille des fichiers d'une image exemple mesurant 800 × 600 pixels, selon le format de fichier employé. Sa taille mémoire théorique, à 24 bits par pixel (soit en 16 millions de couleurs) sera donc de :

800 × 600 × 24 = 11 524 000 bits = 1 440 000 octets = 1 406 Ko

Type	**Nbre bit/pixel**	**Couleurs**	**Taille en mémoire**	**Taille réelle du fichier**
JPG	24	16 millions	1 406 Ko	70 Ko
PNG	24	16 millions	1 406 Ko	532 Ko
TIF	24	16 millions	1 406 Ko	1 254 Ko
BMP	24	16 millions	1 406 Ko	1 407 Ko
TIF	16	64 000	703 Ko	551 Ko
PNG	16	64 000	703 Ko	378 Ko
TIF	8	256	469 Ko	194 Ko
PNG	8	256	469 Ko	163 Ko
GIF	8	256	469 Ko	251 Ko
TIF	4	16	234 Ko	59 Ko
PNG	4	16	234 Ko	49 Ko
GIF	4	16	234 Ko	49 Ko

Cette résolution ne permet pas le format GIF, qui n'accepte pas plus de 256 couleurs. En revanche, le format BMP la reconnaît : comme vous le voyez, il n'effectue aucune compression. Ce format n'est d'ailleurs pas reconnu sur le Web, de même que le format TIF : ils ne sont là que pour permettre la comparaison. Dans cet exemple, le fichier JPEG est de loin le plus petit.

La première possibilité pour réduire la taille de ce fichier sans modifier sa taille consiste à réduire le nombre de couleurs, pour passer à 64 000 couleurs. Cela élimine le format JPEG, obligatoirement en 16 millions de couleurs. Nous restons toutefois loin de la taille du fichier JPEG, malgré cette baisse de qualité. Poursuivons en passant en 256 couleurs, pour pouvoir tester le format GIF.

Si les fichiers restent de taille supérieure à celui d'un fichier JPEG en 16 millions de couleurs, le format GIF se révèle le plus gourmand des autres. Poussons le raisonnement à son terme en réduisant radicalement le nombre de couleurs à 16 (soit 4 bits par couleur).
La taille des fichiers poursuit sa réduction, passant cette fois en dessous de celle du fichier JPEG. Il est toutefois permis de s'interroger sur la pertinence d'une telle réduction de qualité, pour ne gagner que 30 % en taille !
Sachez toutefois que les chiffres présentés ici sont propres à l'image employée : une autre image pourrait donner des résultats différents, le format GIF pouvant s'avérer plus intéressant en 256 couleurs. Il reste cependant intéressant de noter qu'inversement la conversion d'un fichier GIF en JPEG apporte parfois une grande amélioration !
Lorsque vous voulez afficher une image sur le Web ou l'envoyer à quelqu'un, posez-vous toujours la question du but ultime et de la taille réelle souhaitée. L'expérience prouve que les images sont souvent exagérement surdimensionnées par rapport à l'usage attendu (autrement dit, hors impression).

Présentations/diaporamas

Extension	Description	Lu par
.ppt	Format propriétaire Microsoft PowerPoint.	Powerpoint, Ooo.
.swf	Flash	Plugin FlashPlayer.
.odp	(*Open Document Format*) Norme internationale pour la bureautique (ISO 26300:2006).	Ooo, PowerPoint récent.

Sons

Extension	Description	Lu par
.mp3	(*Moving Pictures Expert Group*). Très répandu sur l'Internet, surtout en raison de son historique. Format "destructeur", dont la taille dépend fortement de la qualité (mesurée en kbps). Format compressé devenu un standard. Permet d'obtenir des fichiers peu volumineux tout en conservant une bonne qualité de son.	Le codec est intégré nativement sur tous les systèmes d'exploitation modernes.
.wav	En réalité un encapsuleur (ou conteneur) de différents formats. La plupart du temps, contient du son au format PCM, format non-compressé et sans perte. Fichiers lourds.	La plupart des lecteurs audio.
.ogb	Format OGG Vorbis. Alternative ouverte du MP3, mais reste aujourd'hui nettement moins répandu.	Certains lecteurs.
.wma	Assez répandu, notamment en raison de la prise en charge native du système de gestion des droits (DRM), permettant de vérifier la légalité d'un téléchargement.	Nombreux lecteurs

.mid	Format MIDI. Très employé dans les années 1990, entre autre pour sa légèreté. Il est néanmoins impossible d'enregistrer un audio complexe (voix, par exemple) au format MIDI.	Nombreux lecteurs
.aif .aiff	(*Audio Interchange File Format*)	Nombreux lecteurs
.au/.snd	Sons de qualité très moyenne mais de petite taille	Nombreux lecteurs
.ra .ram	Real Player. Sons et vidéo sur l'Internet de qualité satisfaisante en tolérant un débit limité.	Nombreux lecteurs

Vidéo

Extension	Description	Lu par
.asf/.avi	L' extension avi recouvre divers types de fichiers vidéo, mais sert aussi pour les fichiers vidéo compressés au format DivX permettant d'obtenir des fichiers peu volumineux facilement échangeables tout en ayant une bonne qualité.	Windows MediaPlayer
.mpg/.mpeg	(*Moving Pictures Experts Group*) MPEG-1 utilisé pour les diffusions sur CD-Rom (format quart d'écran) , MPEG-2 est utilisé pour les DVD vidéo, et sert de base à la diffusion sur les chaînes numériques	
.qt/.mov	Apple QuickTime	Lecteur QuickTime.
.ra/.ram	Real Player	Lecteur RealPlayer

Fichiers compressés

Il s'agit de fichiers « compressés » afin de réduire leur taille et de regrouper plusieurs fichiers en un seul. Cela est utile aussi bien pour l'archivage et le classement que pour l'envoi *via* Internet). Windows, depuis sa version XP, sait compresser et décompresser des fichiers : il gère même désormais des dossiers compressés (qui ne sont en réalité rien d'autre que des fichiers compressés, directement accessibles). Il existe aussi de nombreux logiciels libres ou graticiels qui le permettent, dont 7Zip et iZarc.

Extension	Description	Lu par
.ZIP	Format de compression et d'archivage la plus utilisée	Nativement par Windows. Winzip, 7zip, iZarc, etc.
.gz/.tar/.Z/.arc /.rar/.sit/.ace/. arj/.lha/lzh/.lzx /.zoo	Autres formats de compression parfois rencontrés.	Divers logiciels.
.exe, .sfx, .cab	Archive compressée auto-extractible.	Se lance s'elle-même. Attention : comme il s'agit de fichier exécutable, peut cacher un virus. Se méfier d'une pièce jointe de type .exe.

Les formats particuliers :

De façon générale, mieux vaut ne pas toucher à ces fichiers sauf lorsqu'il est mentionné autrement.

BAK : Fichier de sauvegarde. Peut être supprimé.	INI : Fichier de configuration
BAT : fichier batch DOS	LNK : Raccourci vers un document
CAB : fichier compressé Microsoft d'installation de logiciels	OLD : Souvent utilisé pour la sauvegarde de fichiers système
DAT : fichiers de la base de registre, essentiels au fonctionnement de votre ordinateur	REG : Fichier de données de la base de registre.
DLL : Bibliothèque de liens partagés. Fichier système	SYS : Fichier de pilote d'un périphérique. Seul CONFIG.SYS peut être édité
FON : Police de caractères	TMP : Fichier temporaire. A supprimer.
HLP : Fichier d'aide	SCR : Economiseur d'écran Windows

4 En guise de conclusion

Autant que possible, dans vos échanges avec d'autres, évitez les formats propriétaires. Servez-vous de préférence de logiciels libres : OpenOffice permet d'enregistrer et/ou exporter dans des centaines de formats différents dont les formats standards PDF, RTF TXT ainsi que dans une multitude d'autres formats, compatibles avec de nombreux logiciels « concurrents » (dont les produits Microsoft).
Toutefois, lors d'échanges de fichiers, il est recommandé d'utiliser des formats « standards » compatibles avec tout logiciel ou des formats qui s'ouvrent avec des lecteurs gratuits comme pdf et swf.
A titre d'exemple, voici ce qui est recommandé au niveau français pour les marchés publics (http://www.marche-public.fr/Marches-publics/Definitions/Entrees/Dematerialisation/Format-de-fichier.htm)
« L'acheteur public peut, par exemple, utiliser les formats de fichiers suivants :

Typologie des fichiers	Extensions correspondantes
PDF (mode non révisable)	.pdf
Texte universel (mode révisable)	.rtf
Bureautique ouvert ODF (mode révisable format ouvert, normalisé ISO)	.odt pour les textes .ods pour les feuilles de calcul .odp pour les diaporamas .odg pour les dessins et graphiques
Bureautique propriétaire de Microsoft (mode révisable)	.doc/.docx pour les textes .xls/xlsx pour les feuilles de calcul .ppt/pptx pour les diaporamas
CAO « OpenDWG » (mode révisable) pour les plans ou dessins techniques ou PDF 1.7 (mode non révisable, normalisé ISO, conservation des calques)	.dxf
Propriétaire DWG (mode révisable) pour les plans ou dessins techniques ou propriétaire DWF (mode non révisable)	.dwg
Images JPEG, PNG ou TIFF/EP pour les photographies et images	.jpg/.png/.tif
Audio MP3 (format compressé - qualité ordinaire) ou WAV (format non compressé - haute qualité) pour les fichiers sonores	.mp3/.wav
vidéo MPEG-4	.mp4

»