Formats de fichiers

Introduction

Suivant le type de données enregistrées dans un fichier, le format est différent. La partie emergée de l'*iceberg* des formats est l'extension du fichier.

Tout fichier a un nom. Le plus souvent, ce nom est de la forme `foo.bar`. On voit deux parties, séparées entre elles par un point. Ici `bar` est l'extension.

> L'extension est donc la partie du nom située après le dernier point. Elle se compose en général de 3 lettres.

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 4 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. Les extensions doivent toujours être affichées.

Remarque : un répertoire (aussi appelé « dossier » ou « catalogue ») contient des fichiers et d'autres dossiers.

Aparté : les archives et la compression. Des données peuvent-être compressées pour réduire leur encombrement, *ie.* leur taille. Il y a deux types de compression : la compression sans perte (pour les fichiers de données) et la compression avec perte (typiquement pour les fichiers images, audio et vidéo où tous les détails que l'œil ou l'oreille humaine ne perçoivent pas peuvent être supprimés).

Une archive est un fichier contenant plusieurs fichiers ou dossiers. Une archive est souvent compressée (sans perte).

Du fichier au programme

Un fichier en lui-même n'a aucun intérêt. Un fichier contient des données mais il faut pouvoir accéder aux données (*qu'importe le flacon, pourvu qu'on ait l'ivresse !* dit l'adage). Pour accéder aux données stockées dans un fichier, il faut ouvrir le fichier. Un fichier s'ouvre avec une application *ie.* un programme *ie.* un logiciel.

Citons comme exemples de programmes MS Word (MS = *Microsoft*), OOo Writer (OOo = OpenOffice.org, du nom du site *Web* de la suite bureautique libre OpenOffice : `http://www.openoffice.org`), Internet Explorer, Firefox, Outlook, Thunderbird, RealPlayer, VLC, ...

Classons les programmes par usage :

- MS Word et OOo Writer sont des « logiciels de traitement de texte »,
- Internet Explorer et Firefox sont des « navigateurs *Web* » (butineurs, *browsers*, ...),
- Outlook et Thunderbird sont des « clients de courrier »,
- RealPlayer et VLC servent à lire des sons et des vidéos.

On pourrait citer encore des milliers de logiciels en les classant en centaines de catégories, suivant leur(s) usage(s) et les formats de données qu'ils sont capables de lire/écrire.

Programmes, formats et extensions

De plus, certains programmes lisent parfaitement certains formats mais sont incapables de créer des fichiers dans ces formats. Par exemple, un navigateur *Web* ne permet pas de modifier des images mais il les affiche. Pour jouer les sons et les vidéos, il fait appel à des *plugins* (aussi appelés *addons*, greffons ; par exemple les vidéos sur `http://YouTube.com` ou `http://DailyMotion.com` sont jouées avec un greffon Shockwave-Flash).

Formats ouverts versus *formats propriétaires*

Les formats sont définis par les informaticiens qui programment les logiciels.

Certains formats sont dits « ouverts », signifiant par là que les spécifications sont publiques. Ainsi, quiconque ayant un fichier d'un tel format peut consulter les spécifications pour écrire un logiciel qui sera capable de manipuler ce format.

À l'inverse, certains éditeurs de logiciels protègent leurs formats (par des brevets par exemple). On parle alors de formats « propriétaires ». Le concepteur du format garde alors le monopole de l'utilisation du format (de ce fait, ils sont le plus souvent payants).

Certains formats sont ouverts quoique soumis à des brevets (PDF, MP3) : le détenteur du brevet a choisi de divulguer les spécifications (pour pousser à l'utilisation de son format).

Accents et alphabets

L'informatique est née dans les pays anglo-saxons. L'anglais utilise l'alphabet latin, sans accent. Avec la ponctuation, les variantes de casse, les chiffres et quelques autres caractères, cela représente 128 caractères différents. Cet encodage de base est souvent dit « ASCII » (ou « 7 bits »), c'est l'un des tous premiers utilisés en informatique.

Pour pouvoir travailler sur des ordinateurs dans d'autres langues que l'anglais, il a fallut encoder différemment les « nouveaux » caractères ; par exemple pour le français, il faut les trois accents, le trema, la cédille le œ, les guillemets à la française « et » et dernièrement le symbole €. Les constructeurs informatiques ont souvent fait des choix indépendemment les uns des autres, ainsi un é ne sera pas encodé de la manière sur un *Macintosh*, sur un PC sous Windows et sur une station UNIX. Les temps ont changé et petit à petit tout cela est normalisé. La normalisation a commencé par les ISO-8859-X (X = 1 pour les langues d'Europe occidentale, X = 5 pour le cyrillique comme le russe et l'ukrainien, X = 6 pour l'arabe, X = 8 pour l'hébreu, *etc.*). Elle continue depuis avec la généralisation d'UTF (un seul codage pour tous les systèmes d'écriture de langues de la Planète, vivantes et mortes, basé sur la norme Unicode).

Le plus souvent, les « bons » logiciels sont capables de lire plusieurs jeux de caractères.

Logiciels gratuits, libres, propriétaires

Il existe plusieurs types de logiciels : certains sont payants (typiquement, les logiciels *Microsoft* : Windows, MS Word, Outlook...), d'autres sont gratuits (Adobe Reader, Internet Explorer), d'autres encore sont libres (logiciels de la *Mozilla Foundation* : Firefox, Thunderbird, ...) *ie.* si vous avez les compétences, vous pouvez les modifier puisque les codes sources sont fournis. Souvent ces logiciels libres sont le fruit du travail de bénévoles. Attention néanmoins, un logiciel libre peut-être payant.

Le système d'exploitation : un ensemble logiciel particulier

Un ordinateur n'est qu'un assemblage complexe de matériaux divers (silicone, cuivre, aluminium, plastiques, ...). Quand on le branche sur une source d'électricité, un logiciel imprimé sur le matériel lance le système d'exploitation (OS, *operating system*) qui est un ensemble de logiciels enregistrés sur le disque. L'OS est responsable de l'interface entre l'utilisateur (humain) et le matériel, il traduit les actions de l'utilisateur (double-clic sur une icône) en instructions compréhensibles par la machine (ouvrir le fichier correspondant à l'icône).

Sur les ordinateurs de bureau du commerce, on peut trouver plusieurs systèmes d'exploitation différents : le très classique Windows (propriétaire) sur les ordinateurs de type PC, le fameux MacOS X (en partie propriétaire) sur les machines *Apple* (*Macintosh*), une version de GNU/Linux (libre, Ubuntu par exemple), ...

Un téléphone portable a son propre OS tout comme un modem ADSL, un appareil photo numérique ou un ascenseur.

Disponibilités des programmes

Suivant les plateformes (MacOS X, UNIX ou GNU/Linux, Windows), certains programmes sont disponibles ou non. Il faut donc toujours se poser la question de la pérennité des données dans un format :

- Si j'enregistre mes données dans le format truc, pourrais-je les lire et/ou les modifier sur un autre ordinateur ?
- Si mes fichiers sont sur clef USB, pourrais-je les lire sur un autre ordinateur ? Variante : si j'envoie mon fichier à Untel, pourra-t-il le lire ?

Extensions et logiciels à connaître

- TXT : texte seul. Pas de gras ni d'italique, pas de souligné, pas de tailles différentes de caractères, pas de changement de fontes, pas de tableaux complexes... uniquement des caractères alphanumériques. Il s'agit d'un format ouvert.
- RTF : texte enrichi (*rich text format*) : la mise en forme est basique. Il s'agit d'un format ouvert, utilisable avec tout bon logiciel de traitement de textes.
- DOC (et maintenant DOCX) : le format (propriétaire) de MS Word. Les fichiers DOC sont aussi lisibles avec OOo Writer (mais pas encore DOCX). Adapté à des textes longs, mis en forme, avec tableaux et schémas, index et tables des matières, ...
- ODT : le format (ouvert et normalisé) de OOo Writer. Adapté lui aussi à des textes longs, mis en forme, avec tableaux et schémas, index et tables des matières, ... Les fichiers sont lisibles et modifiables avec plusieurs logiciels libres dont OOo Writer mais aussi Abiword et KOffice.
- XLS : le format (propriétaire) de MS Excel, un logiciel de manipulation de tableaux (« tableur »). Les fichiers XLS sont aussi lisibles avec OOo Calc. Adapté aux tableaux complexes avec des calculs, par exemple pour faire de la comptabilité.
- ODS : le format (ouvert et normalisé) du tableur libre OOo Calc.
- PPT : le format (propriétaire) de MS PowerPoint, un logiciel pour faire des présentations. Les fichiers PPT sont aussi lisibles avec OOo Impress.
- ODP : le format (ouvert et normalisé) de OOo Impress, logiciel de mise en forme de présentations.

- PDF : format ouvert de documents imprimables, très utilisé dans le monde de l'édition et sur Internet. Il s'agit d'un format ouvert en cours de normalisation. Les fichiers sont lisibles avec des logiciels libres ou gratuits ; par contre ces logiciels (Adobe Reader, Foxit PDF Reader, GhostView, Xpdf, ...) ne permettent pas de modifier des fichiers PDF.

  OOo permet d'exporter les documents en PDF. PDFCreator permet de transformer (sous Windows) tout fichier imprimable en un fichier PDF.
- BMP : format ouvert d'images fixes non-compréssées. Les principaux logiciels de manipulation d'images le connaissent. Transformer l'image en un format compressé avant de l'insérer dans un document.
- GIF, JPEG et PNG : formats ouverts d'images fixes compressées. Préferer PNG ou JPEG. Les principaux logiciels de manipulation d'images les manipulent sans souci. Pour les retoucher à la manière d'Adobe Photoshop (propriétaire), le logiciel The GIMP est libre.
- AIF, WAV: formats ouverts de fichiers audio non-compressés. Les logiciels audio savent les lire, par exemple SongBird.
- MP3 : format ouvert de fichiers audio compressés (avec perte). Les bons logiciels audio savent lire les fichiers MP3, par exemple MPlayer et VLC.
- BZ2, GZ, ZIP et 7Z: formats ouverts de fichiers compressés sans perte. ZIP est très courant (MacOS X et Windows XP les lisent nativement), BZ2 et GZ sont les plus utilisés sous UNIX (MacOS X et GNU/Linux).
- RAR : format propriétaire d'archives compressées sans perte.
- RAM : format propriétaire (*Real*) audio et vidéo. Utiliser RealPlayer ou RealPlayer Alternative pour les lire.
- AVI et DIVX : formats ouverts de vidéo, lisibles par exemple avec MPlayer et VLC.
- MOV :format ouvert de vidéo (prévu pour QuikTime), lisible par exemple avec VLC.
- ISO et DMG : format d'images de disques. ISO est le format des CD de données. DMG est exclusivement utilisé sous MacOS X.
- SWF : fichiers Flash, lisibles dans un navigateur *Web* disposant du greffon Shockwave-Flash. Ce greffon n'est pas disponible sur toutes les plateformes.
- HTML : pages *Web*, lisibles avec un navigateur *Web*. C'est évidemment un format ouvert ; c'est même plus qu'un format mais cela sort du cadre de ce document...

Réencodage

Une erreur classique consiste par exemple à renommer un fichier DOC `file.doc` en `file.pdf` pour espérer l'ouvrir ensuite avec Adobe Reader. Le nom d'un fichier devrait toujours correspondre à son format mais rien n'y contraint, l'utilisateur a toute lattitude pour nommer ses fichiers comme bon lui semble.

Pour reprendre l'exemple précédent, une méthode consiste à ouvrir le fichier DOC avec MS Word puis à l'« imprimer » avec PDFCreator (PDFCreator se comporte comme une imprimante virtuelle). Mais PDFCreator n'est disponible que sous Windows. Sur UNIX, pas de MS Word non plus. Une autre méthode consiste donc à ouvrir le fichier avec OOo Writer puis à l'exporter en PDF.

De même, renommer un fichier WAV en MP3 ne va pas le compresser : autant essayer de transformer du plomb en or en le peignant en jaune... On peut multiplier les exemples

pour chaque format... Et pour chaque format il faut trouver une (ou plusieurs) méthodes.

Enregistrements de fichiers depuis Internet

On veut parfois garder sous la main un fichier disponible sur le *Web* pour le cas où la connexion réseau serait coupée. Le plus souvent, il suffit de demander « Enregistrer sous... » (dans le menu fichier). Parfois il faut ruser en utilisant d'autres méthodes (attention au droit d'auteur, ce n'est pas parce qu'un fichier est disponible sur Internet qu'il est librement enregistrable et/ou redistribuable mais c'est une toute autre question).

1. Les flux audio/vidéo. Ces données sont disponibles sur un serveur mais non pas sous forme de fichier mais de flux (comme un programme à la télévision ou à la radio par opposition à une cassette ou un DVD). Il faut donc enregistrer le flux au fur et à mesure. MPlayer fait ça très bien quel que soit le format, RAM par exemple (chercher sur le *Web* avec les mots clefs `mplayer+dumpfile`).
2. Les vidéo en Flash disponibles sur les sites de partage de vidéos comme `http://YouTube.com` ou `http://DailyMotion.com`. Enregistrer le flux *via* le site `http://keepvid.com/` dans un fichier FLV puis utiliser un *player* comme MPlayer ou VLC.

Avoir ses outils sous la main

Pour disposer de sa boîte à outils où que l'on aille, sur quelque machine que l'on travaille, il existe deux méthodes reposant sur l'usage de clefs USB (voire de CD ou de DVD).

La première est popularisée sous le nom de *FramaKey* (du nom du site FramaSoft : `http://www.framasoft.net/`) : il s'agit de réserver une partie de la clef USB aux fichiers de données et installer les logiciels que l'on utilise sur le reste. Voir la clef USB distribuée par le Conseil régional de l'Île de France ou encore : `http://www.framakey.org/Main/Details`

La seconde méthode, plus brutale encore consiste à avoir son propre système d'exploitation avec les logiciels dont on a besoin sur une clef USB (de bonne capacité donc plus chère), typiquement l'OS sera un système UNIX ou GNU/Linux, utilisant les techniques de *LiveCD*.

Conclusion

On vient de voir que plusieurs notions étaient imbriquées :

- Les données sont enregistrées dans des fichiers.
- Ces fichiers sont d'un format différent suivant le type de données contenues et le logiciel ayant permis de les enregistrer.
- Suivant le format d'un fichier, on ne choisit pas nécessairement le logiciel pour le manipuler, ce qui peut s'avérer coûteux ou bloquant.

On a aussi vu le problème de l'interopérabilité :

- Si j'enregistre mes données dans le format truc, pourrais-je les lire / modifier à l'avenir (si l'entreprise qui me fournit le logiciel vient à disparaître, que ferai-je de mes fichiers) ?
- Si mes fichiers sont sur clef USB, pourrais-je les lire sur un autre ordinateur ? Variante : si j'envoie mon fichier à Untel, pourra-t-il le lire facilement ?