Module 2 Séquence 3

Format de fichier

Fred de Lamotte - Montpellier
https://orcid.org/0000-0003-4234-1172

Julien Seiler - Strasbourg
@julozi

Format de fichier

Deux grandes catégories de formats : **textuels** et **binaires**.

Enjeu pour la préservation et l'exploitation des données

Formats « textuels »

- Suite d'octets représentant des caractères imprimables et affichables à l'écran
- Peuvent entre lus dans un éditeur de texte
- Mais souvent besoin d'un logiciel spécifique pour interpréter la structure interne, matérialisée par certains caractères, et en donner une représentation informatique exploitable

Ex. de format textuel : HTML
Contenu lisible dans un éditeur texte :
<html>
<head><head>
<body>
Bonjour tout le monde </body>
</html>
Mais « interprétable » par un logiciel dédié (navigateur web) :

Caractères ordinaires + caractères ayant une valeur spéciales : **< > /**, etc.
Mots ayant des valeurs spéciales en HTML (« balises ») si encadrés par < > ou </>: <body>, </body>, etc...

Ex. de format textuel : RTF (texte structuré) Contenu lisible dans un éditeur texte

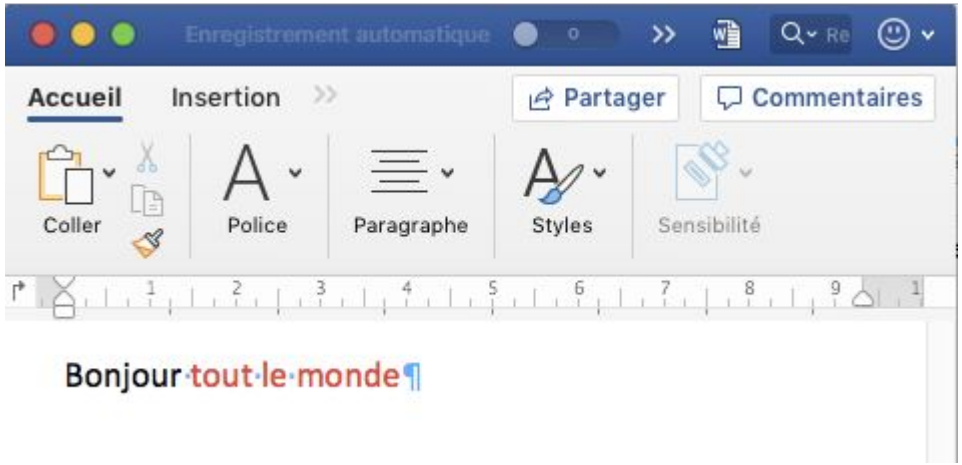
```
{\rtf1\adeflang1025\ansi\ansicpg1252\uc1\adeff0\deff0\stshfdbch37\stshfl
och37\stshfhich37\stshfbi0\deflang1036\deflangfe1036\themelang1036\theme
langfe0\themelangcs0{\fonttbl{\f0\fbidi \froman\fcharset0\fprq2{\*\panos
e 02020603050405020304}Times New Roman;}{\f34\fbidi \froman\fcharset0\fp
rq2{\*\panose 02040503050406030204}Cambria Math;}

\mlMargin0\mrMargin0\mdefJc1\mwrapIndent1440\mintLim0\mnaryLim1}{\info{\
author Mathieu Saby}{\operator Mathieu Saby}{\creatim\yr2018\mo6\dy10\hr
13\min44}{\revtim\yr2018\mo6\dy10\hr13\min44}{\version2}{\edmins1}{\nofp
ages1}{\nofwords3}{\nofchars19}

\fs24\lang1036\langfe1033\loch\af37\hich\af37\dbch\af37\cgrid\langnp1036
\langfenp1033 {\rtlch\fcs1 \af0 \ltrch\fcs0 \insrsid16651434 \hich\af37\
dbch\af37\loch\f37 Bonjour }{\rtlch\fcs1 \af0 \ltrch\fcs0 \cf6\insrsid16
651434\charrsid16651434

\hich\af37\dbch\af37\loch\f37 tout le monde}{\rtlch\fcs1 \af0 \ltrch\fcs
0 \insrsid16651434
```

Mais uniquement interprétable avec Word, Libre office ou autre traitement de texte



Formats « binaires »

- Suite d'octets non interprétables comme des caractères imprimables ou affichables
- Structure interne opaque
- Besoin de logiciel spécifique pour les lire et les interpréter

Ex. de format binaire : PNG (image)
Contenu illisible dans un éditeur texte (à part «?PNG » au début)

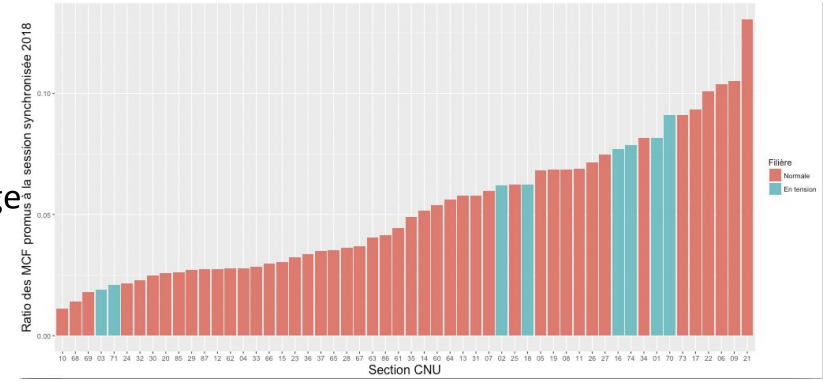
```
?PNG

?V4?????6n?I6?"?d??Θ??83???OEP|1?L?? (??>?/?
%?? (>???P苦?;3?i???e?|??{?g?蹟X????-2?s???=+?????WQ+]?L6O

w?[?C?{_???????F qb??

                    ???U?vz????Z?b?l@?/z??c??s>~?if?,?HUS
j??????F
```

Uniquement lisible et interprétable avec une visionneuse d'image



Formats d'archives : ensemble de fichiers (binaires ou textuels) regroupés dans un fichier unique, compressé (ZIP, 7Z, RAR, TAR.GZ…) ou non (TAR)

Ex : le format DOCX (Word récent) est en fait une archive contenant plusieurs fichiers textuels (XML). Il faut changer l'extension en .ZIP pour lister le contenu.

```
├── [Content_Types].xml
├── _rels
├── docProps
│   ├── app.xml
│   └── core.xml
└── word
    ├── _rels
    │   └── document.xml.rels
    ├── document.xml
    ├── fontTable.xml
    ├── settings.xml
    ├── styles.xml
    ├── theme
    │   └── theme1.xml
    └── webSettings.xml
```


```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<w:document xmlns:wpc="http://schemas.microsoft.com/office/word/2010/wordprocessingCa
nvas" xmlns:cx="http://schemas.microsoft.com/office/drawing/2014/chartex" xmlns:cx1="
http://schemas.microsoft.com/office/drawing/2015/9/8/chartex" xmlns:cx2="http://schem
as.microsoft.com/office/drawing/2015/10/21/chartex" xmlns:cx3="http://schemas.microso
ft.com/office/drawing/2016/5/9/chartex" xmlns:cx4="http://schemas.microsoft.com/offic
e/drawing/2016/5/10/chartex" xmlns:cx5="http://schemas.microsoft.com/office/drawing/2
016/5/11/chartex" xmlns:cx6="http://schemas.microsoft.com/office/drawing/2016/5/12/ch
artex" xmlns:cx7="http://schemas.microsoft.com/office/drawing/2016/5/13/chartex" xmln
s:cx8="http://schemas.microsoft.com/office/drawing/2016/5/14/chartex" xmlns:mc="http:
//schemas.openxmlformats.org/markup-compatibility/2006" xmlns:aink="http://schemas.mi
crosoft.com/office/drawing/2016/ink" xmlns:am3d="http://schemas.microsoft.com/office/
drawing/2017/model3d" xmlns:o="urn:schemas-microsoft-com:office:office" xmlns:r="http
://schemas.openxmlformats.org/officeDocument/2006/relationships" xmlns:m="http://sche
mas.openxmlformats.org/officeDocument/2006/math" xmlns:v="urn:schemas-microsoft-com:v
ml" xmlns:wp14="http://schemas.microsoft.com/office/word/2010/wordprocessingDrawing"
xmlns:wp="http://schemas.openxmlformats.org/drawingml/2006/wordprocessingDrawing" xml
ns:w10="urn:schemas-microsoft-com:office:word" xmlns:w="http://schemas.openxmlformats
.org/wordprocessingml/2006/main" xmlns:w14="http://schemas.microsoft.com/office/word/
2010/wordml" xmlns:w15="http://schemas.microsoft.com/office/word/2012/wordml" xmlns:w
16cid="http://schemas.microsoft.com/office/word/2016/wordml/cid" xmlns:w16se="http://
schemas.microsoft.com/office/word/2015/wordml/symex" xmlns:wpg="http://schemas.micros
oft.com/office/word/2010/wordprocessingGroup" xmlns:wpi="http://schemas.microsoft.com
/office/word/2010/wordprocessingInk" xmlns:wne="http://schemas.microsoft.com/office/w
ord/2006/wordml" xmlns:wps="http://schemas.microsoft.com/office/word/2010/wordprocess
ingShape" mc:Ignorable="w14 w15 w16se w16cid wp14"><w:body><w:p w:rsidR="00FE14AA" w:
rsidRDefault="00FE14AA"><w:r><w:t xml:space="preserve">Bonjour </w:t></w:r><w:r w:rsi
dRPr="00FE14AA"><w:rPr><w:color w:val="FF0000"/></w:rPr><w:t xml:space="preserve">tou
t le </w:t></w:r><w:bookmarkStart w:id="0" w:name="_GoBack"/><w:bookmarkEnd w:id="0"/
><w:r w:rsidRPr="00FE14AA"><w:rPr><w:color w:val="FF0000"/></w:rPr><w:t>monde</w:t></
w:r></w:p><w:sectPr w:rsidR="00FE14AA" w:rsidSect="00F22F18"><w:pgSz w:w="11900" w:h=
"16840"/><w:pgMar w:top="1417" w:right="1417" w:bottom="1417" w:left="1417" w:header=
"708" w:footer="708" w:gutter="0"/><w:cols w:space="708"/><w:docGrid w:linePitch="360
"/></w:sectPr></w:body></w:document>
```

Quelles conséquences pour vous ?
Et pour ceux qui arriveront plus tard ?

Formats et logiciel ?

Allez à : https://scrumblr.ethibox.fr/format

et listez les formats que vous connaissez dans les colonnes idoines

Les logiciels nécessaire pour traiter les formats cités sur scrumblr :

Fonctionnent-ils en ligne ou après installation sur un ordinateur ?

Fonctionnent-ils avec un système d'exploitation particulier (Windows, Mac, Linux) ?

Sont-ils lies à un type d'ordinateur ou à un instrument particulier (ex : microscope) ?

Sont-ils gratuits ou payants ? Qui paye ?

S'ils n'existaient plus ou si vous n'y avez plus accès, pourriez-vous continuer à travailler ?

L'éditeur du logiciel est il en bonne santé ?

? Que proposez vous pour garantir la pérennité de l'accès à vos données ?

Recommandations sur le format des fichiers

Privilégiez les formats ouverts afin de faciliter le partage des données

Format ouvert	Format fermé
Spécifications publiques et gratuites	Spécifications non publiques
Aucune restriction légale pour l'utiliser	Des restrictions légales s'opposent à son utilisation (droit d'auteur, copyright, brevet)
Format indépendant du logiciel utilisé qui assure l'interopérabilité des données	Format lisible qu'avec un logiciel particulier
Maintenu par une organisation à but non lucratif	Format propriétaire

Recommandations sur le format des fichiers

Type	Format conseillé	Format non conseillé
Document texte	PDF, TXT, ODT	MS Word, RTF
Feuille de calcul	ODS, CSV	MS Excel, PDF, OOXML
Base de données	SQL, SIARD, DB tables (.CSV)	MS Access, dBase (.dbf), HDF5
Données statistiques	SPSS Portable, STATA, XML, CSV, TXT	SAS et R
Images	JPEG, TIFF, PNG	DICOM
Audio	BWF, MXF, Matroska (.mka), FLAC, OPUS	WAVE, MP3, AAC, AIFF, OGG
Video	MXF, MKV	MPEG-4, MPEG-2, AVI, QuickTime (.mov, .qt)
Information géographique	GML, MIF/MID	ESRI Shapefiles, MapInfo, KML
Images géoréférencées	GeoTIFF (.tif, .tiff)	TIFF World File
Raster	ASCII GRID (.asc, .txt)	ESRI GRID

https://facile.cines.fr/ service de validation des formats

File formats for digital content: Probability for full long-term preservation

Content type	High	Medium	Low
Text	• Plain text (encoding: USASCII, UTF-8, UTF-16 with BOM) • XML (includes XSD/XSL/XHTML, etc.; with included or accessible schema) • PDF/A-1 (ISO 19005-1) (*.pdf)	• Cascading Style Sheets (*.css) • DTD (*.dtd) • Plain text (ISO 8859-1 encoding • PDF (*.pdf) (embedded fonts) • Rich Text Format 1.x (*.rtf) • HTML (include a DOCTYPE declaration) • SGML (*.sgml) • Open Office (*.sxw/*.odt) • OOXML (ISO/IEC DIS 29500) (*.docx) • Microsoft Word 2007 or newer (*.docx)	• PDF (*.pdf) (encrypted) • Microsoft Word 2003 or older (*.doc) • WordPerfect (*.wpd) • DVI (*.dvi) • All other text formats not listed
Raster image	• TIFF (uncompressed) • JPEG2000 (lossless) (*.jp2) • PNG (*.png)	• BMP (*.bmp) • JPEG/JFIF (*.jpg) • JPEG2000 (lossy) (*.jp2) • TIFF (compressed) • GIF (*.gif) • Digital Negative DNG (*.dng)	• MrSID (*.sid) • TIFF (in Planar format) • FlashPix (*.fpx) • PhotoShop (*.psd) • RAW • JPEG 2000 Part 2 (*.jpf, *.jpx) • All other raster image formats not listed
Vector graphics	• SVG (no Java script binding) (*.svg)	• Computer Graphic Metafile (CGM, WebCGM) (*.cgm)	• Encapsulated Postscript (EPS) • Macromedia Flash (*.swf) • All other vector image formats not listed
Audio	• AIFF (96kHz 16bit PCM) (*.aif, *.aiff) • WAV (96kHz 24bit PCM) (*.wav)	• SUN Audio (uncompressed) (*.au) • Standard MIDI (*.mid, *.midi) • Ogg Vorbis (*.ogg) • Free Lossless Audio Codec (*.flac) • Advance Audio Coding (*.mp4, *.m4a, *.aac) • MP3 (MPEG-1/2, Layer 3) (*.mp3)	• AIFC (compressed) (*.aifc) • NeXT SND (*.snd) • RealNetworks 'Real Audio' (*.ra, *.rm, *.ram) • Windows Media Audio (*.wma) • Protected AAC (*.m4p) • WAV (compressed) (*.wav) • All other audio formats not listed
Video	• Motion JPEG 2000 (ISO/IEC 15444-4)??*.mj2) • AVI (uncompressed/native, motion JPEG) (*.avi) • QuickTime Movie (uncompressed/native, motion JPEG) (*.mov)	• Ogg Theora (*.ogg) • MPEG-1, MPEG-2 (*.mpg, *.mpeg, wrapped in AVI, MOV) • MPEG-4 (H.263, H.264) (*.mp4, wrapped in AVI, MOV)	• AVI (others) (*.avi) • QuickTime Movie (others) (*.mov) • RealNetworks 'Real Video' (*.rv) • Windows Media Video (*.wmv) • All other video formats not listed

Recommandations sur le format des fichiers

Définition légale du **format ouvert** en France (loi no 2004-575 du 21 juin 2004) :
On entend par standard ouvert tout protocole de communication, d'interconnexion ou d'échange et tout format de données **interopérable** et dont les **spécifications techniques sont publiques** et **sans restriction d'accès** ni de **mise en œuvre**.
-> format bien documenté et utilisable sans demander d'autorisation
Ex : DOCX, ODT, XLSX, ODS, XML ; DOC et XLS depuis 2006

Format **fermés (ou propriétaires)** : format non ou mal documenté et/ou dont l'utilisation est payante ou soumise (au moins en théorie) à autorisation
Ex : PSD, JPEG, logiciels SPSS, STATA, SAS, NVIVO, Altas.ti

En pratique, on peut souvent travailler avec un format fermé populaire et le **convertir** en format ouvert. **Mais il faut vérifier si la conversion altère les informations, et prendre des mesures de compensation si nécessaire.**
Ex : la conversion XLSX -> CSV perd les mises en forme.

Formats standardisés

La documentation d'un format peut devenir une norme officielle nationale ou internationale ou un standard de facto.

Ex :

PDF/A1 est une version standardisée (ISO 19005) du format PDF. Les autres versions de PDF ne sont pas standardisées

Les formats Libre office (ODS, ODT…) sont standardisés (ISO/IEC 26300)

Le format XML est standardisé par une « recommandation » du W3C (équivaut à une norme)

Le format CSV est décrit dans la RFC 4180 de l'IETF, mais n'est pas réellement standardisé (la RFC est un document indicatif), plusieurs versions existent

Les formats bureautique Microsoft (XLSX, DOCX…) sont standardisés (ISO/IEC 29500). Mais les logiciels semblent parfois s'écarter du standard

http://dotwhat.net

dot What?
The leading File Extension resource

HOME ADD EXTENSION EDIT EXTENSION ARTICLES FAQ CONTACT US

J'aime 285 Tweet

Search...

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z OTHER

WELCOME TO DOTWHAT? ... THE LEADING FILE EXTENSION RESOURCE

Thanks to years of research and help from our loyal visitors, we now have one of the world's largest and most detailed databases of file extension information, covering multiple operating systems from Microsoft's Windows, Apple's OS X and all variations of Unix to those used on the latest mobile devices and phones.

EVERYTHING YOU NEED TO KNOW! IF NOT, JUST ASK!

We try to provide as much information on each file extension as possible and we encoura visitors to contact us if they have any additional information on an extension or if they think a new file extension should be added to the database. Alternatively, each entry can be edited and visitors have the option of adding a comment, question or tip!

Sections


Software Developers


Software Products


Common File Extensions

Categories


3D/CAD Files


Audio Files


Backup Files


Compressed Files


Configuration Files

Data Files