

I. LES DONNEES ET LEUR FORMAT

I.1. LES PRINCIPAUX FORMATS DE DONNEES

Pour assurer la persistance des données, ces dernières sont stockées dans des fichiers.

Il y a deux types de formats de fichiers dont voici les principaux :

- Ceux lisibles par des éditeurs de texte comme *Notepad++*, *Geany*, *Bloc-Notes*, (on dit alors "de type texte") qui ne contiennent que des caractères alphanumériques : **CSV**, **XML**, **JSON** et **vCard** (extension .vcf) pour une collection de contacts. Les caractères sont codés en mémoire en respectant une **norme d'encodage**, telle que le [Latin1 \[html, fr\]](#) (spécialisé pour coder efficacement les caractères latins) ou l'[UTF-8 \[html, fr\]](#).
- Les autres que les éditeurs de texte ne pourront pas lire correctement (on dit "de type binaire"). Ce sont principalement les formats des tableurs (LibreOffice Calc, Microsoft Office Excel, Numbers...) : **ODS**, **XLS**, **XLSX** ...

et ceux des logiciels de gestion de bases de données (Dbase pour n'en citer qu'un): **DBF**
...

- **Le format CSV** (*Comma Separated Values*, données avec des séparateurs) est à privilégier ; c'est le plus **universel et le plus simple**. C'est aussi très flexible : on peut troquer la virgule contre un point-virgule ou une tabulation comme séparateur de données, si celles-ci contiennent elles-mêmes des virgules (en comme français où la virgule est utilisée pour écrire des nombres décimaux).

- **Le format de Google Sheets**
Google Sheets est le tableur en ligne de Google. Les données s'affichent dans un navigateur, et sont en HTML, format de type texte de description de pages WEB. Les données en HTML sont lisibles par un traitement de texte mais sont polluées par les métadonnées et sont difficilement exploitables en dehors de Google Sheets. Fort heureusement, ce tableur offre, comme les autres, des exports dans les formats usuels.

- **Affichage des extensions de fichiers**
En informatique, une [extension de nom de fichier \[html, fr\]](#) (ou simplement extension de fichier) est un suffixe de nom de fichier fait pour identifier son format. Il est séparé du nom par un point.

Par exemple, une collection au format ODS pourra s'appeler MaCollectionDeTimbres.ODS. **Les extensions des formats XML, CSV et JSON les mêmes que leur nom : XML, CSV et JSON.**

C'est grâce à cette extension qu'un système d'exploitation tel que Linux ou Windows sait qu'il faut lancer le programme LibreOffice Calc, pour lire ce fichier lorsqu'on clique dessus.

Par défaut, Windows n'affiche pas les extensions de fichiers. Cela peut être source d'erreur quand on code un programme informatique, par exemple en langage Python, dans lequel on doit lire le contenu d'un fichier. Il faut alors indiquer dans le code le nom complet du fichier, avec son extension, sans quoi le programme ne trouvera pas le fichier. Pour cette raison, il est conseillé, lorsque vous codez, de régler Windows de façon à faire apparaître les extensions de fichiers. Tapez dans un moteur de recherche les mots "Windows afficher extension" pour trouver comment faire.

▪ **Lecture d'un fichier binaire avec un traitement de texte**

Quand on ouvre un fichier binaire avec un traitement de texte basique tel que Pluma Text Editor, voici le genre de résultat qu'il affiche :

?m'ýôã\$SoŽâÄn0'H"ò'ãÝsw|S..oS:Ù`^?—??ê9?œ!?',Y8??~ëž;—Ëi.X??,Ã`juŠ3á?âŽâb"...³"4?
gÍ³?Á,?a?S\,,?...ÇY%?ÚèP?eZ'²±â?IK |+Æ

L'explication est qu'au moment de la lecture, le traitement de texte décode les données du fichier en utilisant les règles qu'il connaît, c'est à dire les encodages de caractères, comme le latin1 ou l'UTF8. Comme le fichier source n'a pas été codé de cette façon, on obtient cette suite de caractères cabalistiques. Les caractères affichés correspondent à la traduction des octets lus par le logiciel en caractères, en utilisant l'encodage des caractères par défaut du logiciel (UTF-8 ou Latin-1 souvent).

II. RECONNAITRE LES FORMATS DE DONNEES AU FORMAT TEXTE

II.1. RECONNAITRE LES FORMATS CSV, XML ET JSON

Voici la même table codée sous trois formats de type texte différents.

CSV	XML	JSON
Nom, Prénom, Profession, Décès Hugo, Victor, Écrivain, 1885 Camus, Albert, Écrivain, 1960	<pre> <Auteurs> <Auteur> <Nom>Hugo</Nom> <Prénom>Victor</Prénom> <Décès>1885</Décès> </Auteur> <Auteur> <Nom>Camus</Nom> <Prénom>Albert</Prénom> <Décès>1960</Décès> </Auteur> </Auteurs> </pre>	<pre> [{ "Nom" : "Hugo", "Prénom" : "Victor", "Décès" : 1885 }, { "Nom" : "Camus", "Prénom" : "Albert", "Décès" : 1960 }] </pre>

- [CSV \[html, fr\]](#) : les données sont séparées par une virgule pour cet exemple. La première ligne du format CSV contient la liste des descripteurs de la table, séparés eux aussi par une virgule. Chaque ligne correspond aux données d'un individu de la collection. Les séparateurs n'étant pas standardisés, ils peuvent être dans d'autres cas un point-virgule ou une tabulation par exemple on parle alors de format TSV (*Tabulation-separated values*)
- [XML \[html, fr\]](#) (*Extensible Markup Language* langage de balisage extensible) : c'est un langage "à balises". Chaque enregistrement est ici écrit sous la forme <descripteur>donnée</descripteur> , une donnée pouvant être elle-même une table (données hiérarchisées).
- [JSON \[html, fr\]](#) (*JavaScript Object Notation*) : cet exemple montre que JSON est similaire au XML. Les enregistrements, séparés par des virgules, sont écrits de la forme descripteur : donnée. Chaque un individu de la collection est délimité par des accolades { ... }.

III. DONNEES NUMERIQUES OUVERTES

III.1. LES DONNEES OUVERTES

Les **données ouvertes** (*Open Data* en anglais) sont des informations **accessibles librement et gratuitement**, sous la forme de fichiers respectant des formats **interopérables** [html, fr] ⁽⁶⁾.

La finalité est de **donner la possibilité à tout citoyen, toute entreprise ou association d'utiliser ces données numériques à ses propres fins d'analyse pour en extraire l'information désirée.**

Les données publiques sont considérées comme ouvertes si elles répondent à ces 8 principes (2007, Open Government Data, USA)

- **Complètes** : toutes les données doivent être rendues disponibles sauf les données pouvant porter atteinte à la vie privée des citoyens ou à la sécurité ;
- **Primaires** : les données doivent être brutes, telles qu'elles ont été collectées à la source, non agrégées, non modifiées ;
- **Récemment et actualisées** : elles doivent être rendues disponibles aussi vite que possible afin de préserver leur valeur ;
- **Accessibles** : les données sont disponibles au plus large spectre d'utilisateurs ;
- **Exploitable** : elles doivent être structurées et documentées afin de permettre un traitement informatisé ;
- **Accès non discriminatoire** : elles sont disponibles à tout le monde de façon anonyme ne nécessitant pas d'enregistrement ;
- **Format non propriétaire** : elles doivent être rendues disponibles au moins dans un format sur lequel aucune entité ne détient le monopole (ex : non PDF, non Excel) ;

