Tableur et base de données

10 novembre 2018

Sommaire

1 Qu'est-ce qu'une donnée ?

2 Définitions

Une donnée est un élément brut, qui n'a pas encore été interprété, mis en contexte. La mise en contexte crée de la valeur ajoutée pour constituer une information (on peut définir l'information comme l'interprétation d'une source de données). Les données brutes peuvent être entrées dans un programme informatique ou utilisées dans des procédures manuelles comme l'analyse statistique d'une enquête par exemple.

Qu'entend-on maintenant par « données publiques » ? Ce sont les données qui figurent dans les documents communiqués ou publiés par les administrations. À partir de là, quand on parle d'*open data*, il s'agit de la mise à disposition de tous les citoyens sur internet des données publiques ayant vocation à être librement accessibles et gratuitement réutilisables. Ces données sont diffusées de manière structurée selon une licence ouverte garantissant leur libre accès et leur réutilisation par tous, sans restriction technique, juridique ou financière. Cela signifie que n'importe quel utilisateur peut utiliser ces données, les modifier ou les partager (même à des fins commerciales).

Il y a enfin une autre expression que vous avez certainement entendues : celle de *big data* qui désigne littéralement les « grosses données ». Qu'entend-on par là ? Les vibrations de tel tablier de pont, les sentiments exprimés sur tel réseau social, les achats ou recherches sur tel site... Toutes ces données, utiles pour la maîtrise de machines ou notre vie sociale, économique, voire sentimentale, laissent des traces, des scories, qui sont de plus en plus souvent conservées. C'est de cette profusion de données sur de nombreux domaines que résultent les Big Data ou mégadonnées. On le voit, ces Big Data désignent des ensembles de données tellement volumineux qu'il devient difficile, voire impossible, de les manipuler avec des outils classiques de gestion de base de données ou de gestion de l'information.

2.1 Formats

En fonction de ce que l'on veut faire, les données informatiques sont rangées dans des fichiers. Le choix de la méthode de rangement, c'est ce qu'on appelle le format. Pour l'utilisateur, le format est représenté par une extension. Le format est déterminé par le logiciel au moment de l'enregistrement d'un document. Comme vous l'avez certainement constaté, le système associe par défaut un logiciel à un format, raison pour laquelle, quand vous double-cliquez sur un document, il s'ouvre automatiquement sans que vous ayez à préciser quel logiciel doit l'ouvrir. Avant de vous présenter les différents formats que vous devez connaître, sachez que chaque format relève d'une logique, d'une structure. On en distingue trois.

les principales structures utilisées : voici les trois principales structures utilisées pour présenter les données :

1. la structure tabulaire : c'est la plus répandue. On organise les données dans des colonnes et des lignes. Voir par exemple les données concernant la fréquentation des musées italiens : `http://www.datiopen.it/opendata/Visitatori_musei_pubblici_e_similari_titolo_d_accesso#ui-tabs-3`. Un des formats qui relève de cette

structure est le format csv (Comma-separated values) ; il permet d'organiser des données en cellules afin qu'elles soient traitées par un tableur ou insérées dans une base de données. Les données dans un document csv sont le plus souvent encadrées par des guillemets et séparées par des points virgule.

```
"titre";"auteur";"date"
"Le lièvre de vatanen";"Arto Paasilinna";"1975"
"L'Abyssin";"Jean-Christophe Rufin";"1997"
"Hergé: biographie";"Pierre Assouline";"1996"
```

2. la structure hiérarchique : les données présentées ainsi montrent les rapports entre les points de données comme pour un arbre généalogique.
3. la structure en réseau : les données structurées ainsi permettent des rapports entre n'importe quelle combinaison d'éléments dans n'importe quelle direction. Le web en est un bon exemple puisque les pages web comportent des liens vers un nombre incalculable d'autres pages. Cf. le format xml (Extensible Markup Language) qui a pour objectif de faciliter l'échange automatisé de contenus complexes. Un document xml est constitué d'un prologue qui indique les informations de traitement (comme le jeu de caractères utilisé), et du corps du document, constitué d'un ensemble de balises qui décrivent les données (se présentant sous la forme d'une arborescence). Où trouve-t-on du xml ? Dans les pages web, les documents OpenOffice sont également des fichiers xml, les logiciels de dessin comme InkScape utilisent aussi ce format, etc.

   NB : vous remarquerez que les mêmes données peuvent être présentées sous des formats différents.

les formats que vous devez savoir utiliser : on distingue les formats ouverts dont les spécifications sont publiquement accessibles des formats fermés qui sont souvent propriétaires (même quand un format propriétaire est ouvert, les entreprises qui le commercialisent tentent d'en conserver le contrôle en proposant de nouvelles versions plus élaborées ou en ayant recours aux brevets). Voici la liste des principaux formats que vous rencontrerez. Vous penserez à préciser quels sont ceux qui sont ouverts.

Sur le format txt : éditeur de textes et traitement de textes. Qu'est-ce qu'un éditeur de textes ? À quoi cela sert-il ? Est-ce la même chose qu'un traitement de texte ? Un éditeur de texte est un programme qui permet de modifier des fichiers de texte brut, sans mise en forme (gras, italique, souligné...). Sous Windows, on dispose d'un éditeur de texte très basique, le Bloc-Notes, mais il existe aussi NotePad++ (plus évolué). Sous Linux, on a le choix entre Nano, Vim, Emacs, et bien d'autres. Un traitement de texte, en revanche, est fait

```xml
<?xml version="1.0" encoding="utf8"?>

<bibliotheque>

    <roman>
        <titre>Le lièvre de vatanen</titre>
        <auteur>Arto Paasilinna</auteur>
        <date>1975</date>
    </roman>

    <roman>
        <titre>L'Abyssin</titre>
        <auteur>Jean-Christophe Rufin</auteur>
        <date>1997</date>
    </roman>

    </biographie>
        <titre>Hergé: biographie</titre>
        <auteur>Pierre Assouline</auteur>
        <date>1996</date>
    </biographie>


</bibliotheque>
```

pour rédiger des documents mis en forme. Word et LibreOffice Writer sont certainement les plus célèbres.

Exercice 1 Quand a-t-on besoin d'un éditeur de texte ? Chaque fois qu'on veut éditer un fichier de texte brut (au format .txt). Si les éditeurs de texte sont parfaits pour les programmeurs, ils sont aussi utiles pour retravailler du texte à l'aide de commandes puissantes, avant de le structurer dans un traitement de textes. Exemple : quand on récupère une œuvre ou un extrait d'œuvre depuis une bibliothèque numérique, il faut très souvent supprimer les retours à la ligne intempestifs qu'on appelle hard wrap. Il est très facile de le faire grâce à un éditeur comme Notepad++ : allez dans le menu TextFX, commande TextFXEdit, sous-commande Unwrap Text. Les retours à la ligne simples sont convertis en fin de ligne (mode soft wrap) tandis que les doubles retours subsistent. À ce problème, s'ajoute parfois aussi celui de caractères cabalistiques qui apparaissent à la place des caractères accentués.

Exercice 2

1. Récupérez au format txt sur Gutenberg *Le corbeau* de Poe et ouvrez-le dans LibreOffice ;
2. Faites apparaître au début du document le titre de l'œuvre en italique (comme il se

Extension	Définition	Comment l'afficher ?	Comment le modifier ?
pdf			
html			
docx, xls			
odt, ods			
zip, gz, tar			
jp(e)g, png, gif			
exe			
txt			

doit), puis enregistrez ce fichier au format natif d'OO (.odt). Enregistrez maintenant ce fichier au format txt et ouvrez-le avec un éditeur de textes par exemple : commentez la différence ;

3. Exportez-le enfin au format pdf et veillez à ce que l'ouverture de ce fichier soit protégé par un mot de passe que vous définirez.

2.2 Le problème des données textuelles : coder du texte

Encodage binaire C'est dans les années 60 qu'apparaît la nécessité de représenter chaque caractère en code traitable par l'ordinateur. Or la mémoire d'un ordinateur n'est capable d'enregistrer qu'une suite de 0 et 1 (encodage binaire) : à l'origine, les lettres de l'alphabet ont donc été encodées sous la forme d'une suite de 0 et de 1.

La table ASCII Mais de quel alphabet parle-t-on ? Tout commence par une constatation très simple : les premiers informaticiens parlaient anglais. Et l'anglais s'écrit avec peu de choses : deux fois 26 lettres, 10 chiffres, une trentaine de signes de ponctuation, de signes

mathématiques, sans oublier le symbole dollar. : avec 95 caractères au total on peut se débrouiller. À l'époque dont je parle, on ne pouvait utiliser que la moitié des octets, soit 128 valeurs. On en a pris 33 comme caractères de « contrôle » (comme le retour à la ligne par exemple), plus les 95 dont on avait besoin pour écrire l'anglais. On a donc attribué des numéros à toutes ces valeurs : le code ASCII (American Standard Code for Information Interchange) était né. Voir la figure 1.

Character	Decimal Number	Binary Number	Character	Decimal Number	Binary Number
blank space	32	0010 0000	^	94	0101 1110
!	33	0010 0001	-	95	0101 1111
"	34	0010 0010	`	96	0110 0000
#	35	0010 0011	a	97	0110 0001
$	36	0010 0100	b	98	0110 0010
A	65	0100 0001	c	99	0110 0011
B	66	0100 0010	d	100	0110 0100
C	67	0100 0011	e	101	0110 0101
D	68	0100 0100	f	102	0110 0110
E	69	0100 0101	g	103	0110 0111
F	70	0100 0110	h	104	0110 1000
G	71	0100 0111	i	105	0110 1001
H	72	0100 1000	j	106	0110 1010
I	73	0100 1001	k	107	0110 1011
J	74	0100 1010	l	108	0110 1100
K	75	0100 1011	m	109	0110 1101
L	76	0100 1100	n	110	0110 1110
M	77	0100 1101	o	111	0110 1111
N	78	0100 1110	p	112	0111 0000
O	79	0100 1111	q	113	0111 0001
P	80	0101 0000	r	114	0111 0010
Q	81	0101 0001	s	115	0111 0011
R	82	0101 0010	t	116	0111 0100
S	83	0101 0011	u	117	0111 0101
T	84	0101 0100	v	118	0111 0110
U	85	0101 0101	w	119	0111 0111
V	86	0101 0110	x	120	0111 1000
W	87	0101 0111	y	121	0111 1001
X	88	0101 1000	z	122	0111 1010
Y	89	0101 1001	{	123	0111 1011
Z	90	0101 1010	|	124	0111 1100
[	91	0101 1011	}	125	0111 1101
/	92	0101 1100	~	126	0111 1110
]	93	0101 1101			

Figure 1 – La table ASCII

L'unicode Mais au bout d'un certain temps est apparue la nécessité de taper du français ou de l'allemand : on a donc utilisé les valeurs laissées de côté par l'ASCII et il a été possible de caser les caractères accentués et divers autres symboles utilisés par les langues d'Europe de l'ouest. Dans ces 128 valeurs, il n'y a hélas pas eu de place pour les caractères des langues occidentales et l'alphabet cyrillique et l'alphabet grec et l'alphabet hébreu. Pour pouvoir taper plusieurs langues sur un même ordinateur et pour que les ordinateurs puissent communiquer entre eux, des organismes de standardisation ont créé des

tables de correspondance, comme l'ISO-8859-1, qui propose un jeu de caractères pour les langues occidentales, l'ISO-8859-5 qui offre du cyrillique, l'ISO-8859-7, qui propose du grec, etc. Mais, malgré tout, il n'a pas été possible de faire rentrer les 1945 idéogrammes du japonais officiel dans un octet, ni les 11 172 syllabes coréennes, ni les dizaines de milliers d'idéogrammes chinois qu'on arrive à recenser… Pour résoudre durablement tous ces problèmes de langues, au début des années 2000, s'est formé un consortium regroupant des grands noms de l'informatique et de la linguistique : le consortium Unicode. Sa tâche : recenser et numéroter tous les caractères existant dans toutes les langues du monde. Est donc né un jeu universel de caractères, acceptant plusieurs encodages, l'unicode. En 2007, le standard publié comportait environ 60 000 caractères. Prenons, par exemple, le sigma majuscule : il a été encodé avec le point de code `U+03A3` (voir la figure 2).

Mais l'unicode prend beaucoup plus de place que l'ASCII. Or, pour prendre l'exemple du français, la grand majorité des caractères utilisent seulement le code ASCII. On a donc imaginé l'UTF-8 (Unicode Transformation Format) : un texte en UTF-8 est partout en ASCII et dès qu'on a besoin d'un caractère appartenant à l'unicode on utilise un caractère spécial pour l'indiquer.

Figure 2 – Aperçu de la table de codage unicode pour l'alphabet grec

Cela dit, comment faire pour saisir une citation en espagnol, chinois, arabe ou grec ancien au milieu d'un texte en français ? Il faut non seulement disposer d'une police unicode (comme Gentium), mais encore d'un clavier virtuel qui vous permet de savoir où se trouvent les caractères. Ainsi, pour être en mesure de saisir du texte dans une langue autre que le français, vous devez attribuer à votre clavier la langue de saisie souhaitée. Par exemple, pour taper οὐκ ἔλαβον πόλιν, vous devez configurer votre clavier de façon à saisir π quand

vous tapez sur la touche P. Pour ce faire, il suffit de cliquer du droit sur l'icône FR, puis de choisir « Paramètres » et « Ajouter ». Il vous est aussi possible d'utiliser des claviers virtuels en ligne, comme celui disponible à l'adresse suivante : http://www.lexilogos.com/clavier/multilingue.htm

2.3 Coder des images

Une image se décompose en points appelés pixels (premier critère de qualité d'une image). À chaque pixel est associée une couleur décomposée en trois composantes, rouge, vert et bleu, chacune étant notée par un nombre entre 0 et 255. Exemple : le code pour le bleu ciel est (119, 181, 254), chaque nombre représentant le dosage nécessaire de chacune des couleurs primaires pour obtenir la couleur désirée. C'est ce qu'on appelle le code RVB (Red Green Blue). Notez que le poids d'une image correspond à 3*nombre de pixels.

3 Le tableur : fonctionnalités simples

Un classeur permet de stocker des données numériques en vue de calculs ou d'affichages graphiques (par opposition à l'affichage texte qu'on vient de voir avec le format csv). Chaque classeur peut contenir de nombreuses feuilles qu'on sélectionne avec des onglets. Chaque feuille de calcul permet de saisir, contrôler, répertorier et analyser des données (textuelles, numériques, fonctionnelles, etc.). Elle contient des cellules éventuellement regroupées en plages.

3.1 Mise en forme

Chaque cellule peut être mise en forme avec une palette complète d'outils. Il est possible de reproduire la mise en forme à une autre cellule ou plage (voir le pinceau brosse).

Dans un tableur, on peut insérer des graphiques dont on règle les dimensions, les axes, les légendes et titres. En fonction du type de données, on pourra privilégier le graphique en histogramme, en lignes et courbes (pour représenter des tendances ou une évolution dans le temps de valeurs numériques), à nuage, en secteurs (ou camemberts).

3.2 Fonctions de calcul

Voir la moyenne.

3.3 Tri, filtre et conversion de données

Tri Il est possible de trier des données en fonction de textes (tri croissant ou décroissant), de nombres, de dates. Plusieurs critères peuvent être définis (ex. : classement d'une classe par ordre alphabétique des noms puis des notes obtenues). Pensez à cliquer sur l'onglet Options pour déterminer les options de tri : vous pourrez ainsi indiquer que la plage contient des étiquettes de colonne afin d'éviter que les en-têtes de colonne soient triés avec les autres données.

Filtre Le filtre automatique (Données>AutoFiltre) permet de faciliter la recherche d'informations au sein d'une plage de données. L'utilisateur peut ainsi choisir des informations qu'il souhaite afficher ou masquer.

Conversion Une fonction permet de diviser une colonne de données texte en plusieurs colonnes (Données>Texte en colonnes). Ex. : à partir d'un nom complet, vous voulez une colonne nom et une colonne prénom.

3.4 Importation d'une source de données

On peut vouloir importer une source de données dans un tableur, par exemple une liste au format texte (qu'on peut visualiser dans un traitement de textes en affichant les caractères non imprimables) : chaque caractère tabulation délimite le champ d'une cellule et chaque pied de mouche indique qu'il faut passer à la ligne. Commençons par ouvrir cette source de données (Fichier>Ouvrir) : il faut alors indiquer que le séparateur est la tabulation.

4 Le tableur comme base de données

Un document Calc peut constituer une base de données simplifiée. Dans une base de données, un enregistrement est un groupe d'éléments de données liés entre eux et traités comme une seule unité d'information. Chaque élément dans l'enregistrement est appelé un champ. Une table est un ensemble d'enregistrements. Chaque enregistrement, à l'intérieur d'une table, a la même structure. Une table peut être vue comme une série de lignes et de colonnes. On le voit, une feuille d'un document Calc a une structure similaire à une table de base de données.

Nous allons définir une plage de base de données de façon à trier, grouper, rechercher et effectuer des calculs avec la plage comme si c'était une base de données (Données>Définir la plage).

4.1 Utiliser des critères de recherche pour trouver des données

Exemple n°1 Comment compter toutes les cellules d'une plage de données dont le contenu correspond à des critères de recherche que nous aurons définis. Hypothèse de travail : nous recherchons le nombre d'étudiants de la base dont la moyenne est égale ou supérieure à 10 OU dont l'âge est inférieur ou égal à 17. Voici ce qu'on doit saisir dans la cellule H5 :

```
=BDNB(A9:G51;0;A1:G3)
```

Commentaire Le nom de la fonction est suivi d'une parenthèse dans laquelle figurent :

- la plage de cellules contenant les données : A9:G51
- le champ de la base (colonne) utilisé pour les critères de recherche : 0
- la plage de cellules contenant les critères de recherche : A1:G3

Exemple n°2 Comment déterminer le contenu de la cellule d'une plage de données correspondant aux critères de recherche. Hypothèse de travail : nous recherchons le nom de l'étudiant qui a obtenu 5/20 au devoir n°1 :

```
=BDLIRE(A9:G51;"PRÉNOM";A1:C2)
```

4.2 Utiliser des formules pour trouver des données

Ne prenons qu'un exemple : la fonction RECHERCHEV (pour recherche verticale) : il s'agit de récupérer des données issues d'une feuille différente de la feuille de travail grâce à une "clé" commune aux deux feuilles.

1. dans la feuille n°1, j'ai les noms de tous mes étudiants et leur n° d'étudiant ;
2. dans la feuille n°2, j'ai les noms d'un seul groupe d'étudiants et leur note.

−> Comment faire pour récupérer dans ma deuxième feuille les n° d'étudiant du seul groupe concerné ?

```
=RECHERCHEV(A2;$Feuille1.$A$1:$B$120;2;0)
```

Commentaire Notez que le premier argument est la valeur cherchée dans la feuille 1 : il s'agit, dans notre exemple, de chercher l'étudiant Charles-Daniel (cellule A2).

Le deuxième argument identifie les cellules où effectuer la recherche.

Le troisième argument identifie la colonne à renvoyer : dans notre exemple, celle des n° d'étudiants est la deuxième colonne de notre feuille 1.

Le dernier argument est facultatif. La valeur par défaut est 1 ou VRAI, ce qui indique que la première colonne est triée dans l'ordre croissant. Une valeur de 0 ou FAUX indique que les données ne sont pas triées.

Une fois que la recherche a abouti pour le premier étudiant, il faut étendre la recherche à toutes les données de notre feuille 2 : pour cela, il suffit de faire un copier/coller en ayant pris la précaution de protéger la formule en encadrant les numéros des colonnes du chiffre $.

4.3 Utiliser des formules pour traiter des données textuelles

Plusieurs fonctions permettent de travailler sur du texte. En voici quelques exemples :

- Mettre en majuscule la première lettre de chaque mot : =NOMPROPRE(A1)
- Supprimer les espaces en trop dans le texte de la cellule A1 : =SUPPRESPACE(A1)
- Extraire le premier mot d'un texte saisi dans la cellule A1 : =GAUCHE(A1 ;CHERCHE(" ";A1 ;1)-1)