LES DONNÉES

SNT

IDENTIFIER – STRUCTURER DES DONNÉES

Définitions

Données : Informations se rapportant à un objet, une personne, un évènement.

Données personnelles : Données se rapportant à une personne identifiable directement (nom, adresse, numéro de carte d'identité,…) ou par recoupement de plusieurs informations (taille, âge, lieu de travail,…)

Données numériques : Informations codées en binaire (ou hexadécimal) présentes dans un fichier

Ces informations définiront le type de fichier (programme, image, musique, texte, etc)

On identifiera ce type de fichier (en ajoutant une extension au nom du fichier (.pdf, .mp3, .mp4, jpg, etc)

⚠ Ne pas confondre l'extension d'un fichier (format) qui indique son type de donnée avec l'extension d'un site qui indique son type de domaine (TLD : top level domain)

Catégorie	Format (extension)
Image	png, tiff, jpeg, gif…
Son	mp3,wav, wma …
Vidéo	mp4, avi, mpeg …
Texte	txt, odt, doc …
archives	zip, rar …

(niveau D d'une adresse : fichier)

Catégorie	TLD (extension)
Localisation	fr, uk, ua, ch, cn, bzh, …
Activité	com, biz, pro…
Institution	Edu, mil, gov …
Entreprise	Apple, amazon, beauty (loreal,)google …

(niveau B d'une adresse internet : domaine)

Exercice : Indiquer Le type de fichier présent aux adresses suivantes

http://passiondesgifs.p.a.pic.centerblog.net/bf5d6b63.gif	sur internet □	page web □
K:/Ambiance/jdr/320_Cultist_Cavern.mp3	sur internet □	page web □
debeir.fr/cours/internet/indexactivitees.html	sur internet □	page web □

Métadonnées : Données apportant des informations complémentaires sur la donnée principale (elles se trouvent dans les propriétés d'un fichier) Pour y accéder cliquer avec le bouton droit sur le nom du fichier, puis choisir propriété

Activités : Aller sur le réseau du lycée, dans le répertoire de votre classe, fichiers en consultation, puis lire les métadonnées des 4 fichiers proposés.

Vérifier les données suivantes : Nom, date, taille.

Modifier certaines propriétés (par exemple émettez un commentaire ou notez-le)

La protection des données est gérée, en Europe, par le RGPD (Règlement Général sur la Protection des Données). Accessible sur le site de la CNIL (https://www.cnil.fr/fr/reglement-europeen-protection-donnees)

Activités : De combien d'articles est composé le RGPD ?

Si vous avez fourni des données personnelles à une entreprise et que vous voulez vous rétracter, quel article vous permet le droit à l'oubli ?

Représentation numérique des données

Selon la manière dont elles sont codées, les données possèderont un format différent

Les caractères informatiques peuvent être codés en ASCII (créé en 1961) ou en UTF-8 (créé en 2006 et désormais utilisé à plus de 90%))

Exemple : le caractère M correspond au code (à la case) D4 en hexadécimal ou 11010100 en binaire

ASCII Code Chart

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

Exercices : écrire votre prénom en code ASCII (en hexadécimal, puis en binaire)

Données structurées

Lienmini.fr/3389-105

Afin de retrouver des données plus facilement, on utilisera des bases de données selon le modèle d'edgar Codd.
Ses données pourront être représentées par tableau, listes ou graphiques mais possèderont toujours les mêmes caractéristiques.

1923-2003

Descripteur

Valeur

Nom de Pays	Capitale	Hymne
Chine	Pékin	La marche des volontaires
États-Unis	Washington	Star Spangled Banner
France	Paris	La Marseillaise
Royaume-Uni	Londres	God save the queen

Le descripteur est le type de donnée. La donnée est la valeur liée au descripteur.

Les Données peuvent être écrites sous plusieurs formats. Néanmoins 3 formats émergent

Le CSV : Les données sont présentées dans un fichier texte, dont les descripteurs sont séparées par des ,)

Chine ,Pékin ,La Marche des volontaires
… , … , ….
France ,Paris ,La Marseillaise

⚠ La virgule est en standard pour les données anglo-saxonnes, mais pas pour les données françaises (la virgule étant utilisée pour les nombres décimaux). C'est pourquoi on peut utiliser à la place de la virgule, un autre séparateur le point virgule ;
Mais certains tableurs gèrent mal ce séparateur en point virgule…
Le format CSV peut être converti facilement sous forme de tableau via un tableur (comme excel)

Le XML Les données sont présentées dans un fichier texte, dont les descripteurs utilisent des balises (comme en HTML)

```
<Pays>
    <nom de pays>Chine</nom de pays>
    <capitale>Pékin</capitale>
    <hymne>la marche des volontaires</hymne>
</Pays>
```

Le format XML est utilisé dans les pages web écrites en HTML

Le **JSON** : Les données sont présentées dans un fichier texte, dont les binômes descripteurs : valeurs utilisent la norme du Javascript (proche du css)

```
{   { "Pays" :  "Royaume uni" , "capitale" : "Londres" ,
    "hymne" : "god save the queen" }
}
```

Le format JSON est utilisé dans les pages web écrites en java script.

Avantages et inconvénients des formats CSV, XML et JSON

Le CSV et le JSON sont simples à écrire et/ou à lire

Le CSV est rigide. Tous les descripteurs doivent être renseignés

Le XML est facile à traiter par les machines (donc plus rapide à traiter)

Traitement de Données structurées

Les données constituent la matière première de toute activité numérique et par conséquent deviennent monnayable.

Les données possèdent un cycle de vie comme décrit ci-dessous

- La Collecte
- Le Traitement et partage
- L'analyse qui permet de donner du sens aux données (elles deviennent alors des informations)
- La sauvegarde et l'archivage (voir cours sur stockage de données)
- La destruction lorsqu'elles deviennent obsolètes.

Activité 1 : Tri de données via un tableur

Accéder à la page web debeir.fr

Télécharger les données présentes sur « la liste des différents articles »

Sous quel format sont stockées les données ?

Remarquer les filtres installés en tête de colonne grâce auxquels nous allons pouvoir trier les données.

Pour Trier les colonnes, il faudra

- Sélectionner la colonne sujette au tri
- Désélectionner tout (ou les titres cochés qui ne vous conviennent pas)
- Sélectionner les titres qui vous conviennent

Effectuer les tris nécessaires afin d'obtenir les données suivantes

Noter les 5 types d'articles présents sur le journal « le petit matin »

Noter les 4 titres des articles traitant des métiers sur tous les journaux

Noter les auteurs de tous les articles présents sur le journal « le petit poucet » traitant de géographie

Activité 2 : Opendatasoft © propose un logiciel hébergé en ligne permettant aux collectivités, services publics et entreprises[8] d'héberger et diffuser leurs données.

Accéder au site (https://public.opendatasoft.com/explore/dataset/correspondance-code-insee-code-postal/table/)

Télécharger le fichier CSV sur votre ordinateur

Ouvrir le fichier avec un tableur (excel) et ne laisser que les lignes des communes de Charente-maritime. (Données – Trier – colonne département –choisir la charente-maritime)

Garder les colonnes codes INSEE , Code postal , commune , superficie , population, geo_point_2d et supprimer les autres.

Enregistrer ce fichier sous le nom « commune_17.CSV » dans le répertoire « remise de devoir »

Trier les données afin de ne voir que votre commune d'habitation, noter leur population et leur géolocalisation.

Comparer, leurs géolocalisations avec votre voisin de table, En déduire si elles se situent plus au nord ou au sud, plus à l'est ou à l'ouest

Data.gouv.fr est une autre source de données fiable

Indiquer le(s) format(s) proposé(s) pour la liste des films sortis en France de 1945 à 2020

Indiquer le(s) format(s) proposé(s) pour l'index Egalité H/F des entreprises de 50 salariés ou plus