Recueil et structuration de corpus

Achille Falaise − Alexandre Roulois

Organisation des blocs méthodo « corpus »

- Trois blocs méthodo indépendants
  - S1 : SL4AY030 – Utilisation de corpus
  - **S2-début : SL4BY010 – Création de corpus**
  - S2-fin : SL4BY020 – Annotation de corpus
    - Prérequis pour SL4BY020 : informatique

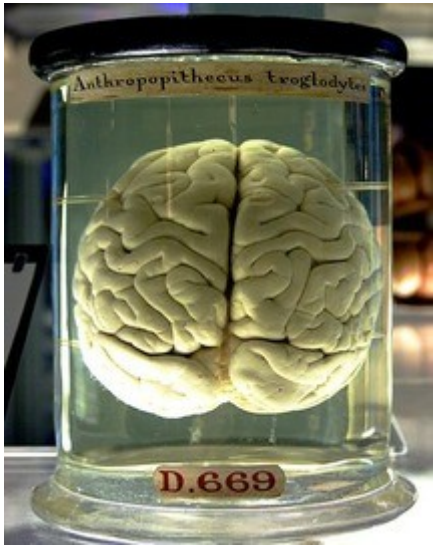
Contenu du cours

Définition d'un corpus → Collecte → Normalisation

Choix des sources
Choix des textes
Licence, éthique

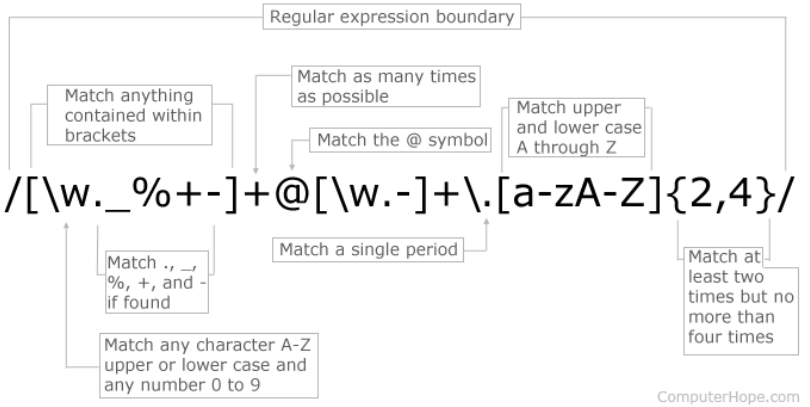
Corpus existants
Documents pdf, doc, odt...
Web

TXT, CSV, XML

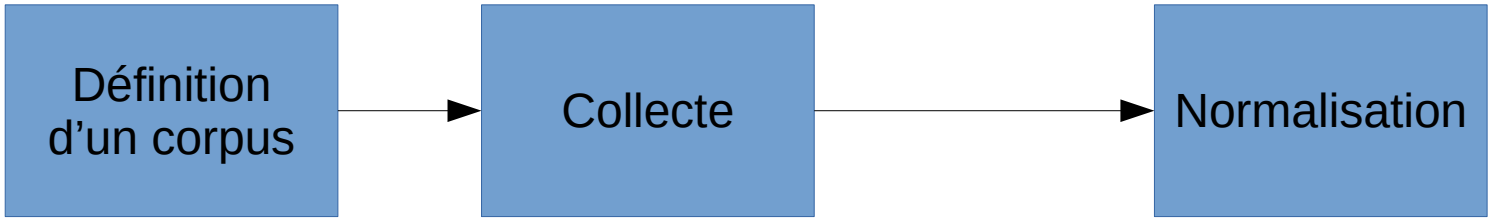


BASH
THE BOURNE-AGAIN SHELL

python



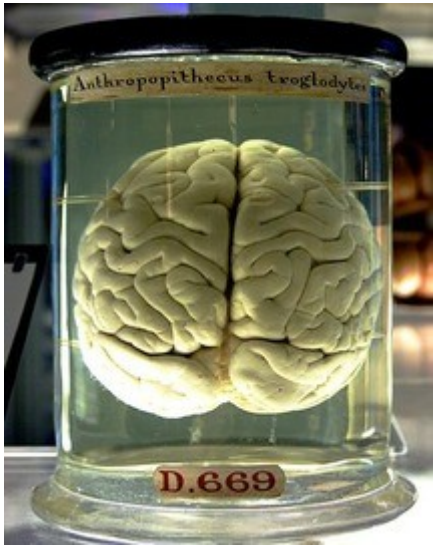
Contenu du cours



Choix des sources
Choix des textes
Licence, éthique

Corpus existants
Documents pdf, doc, odt...
Web

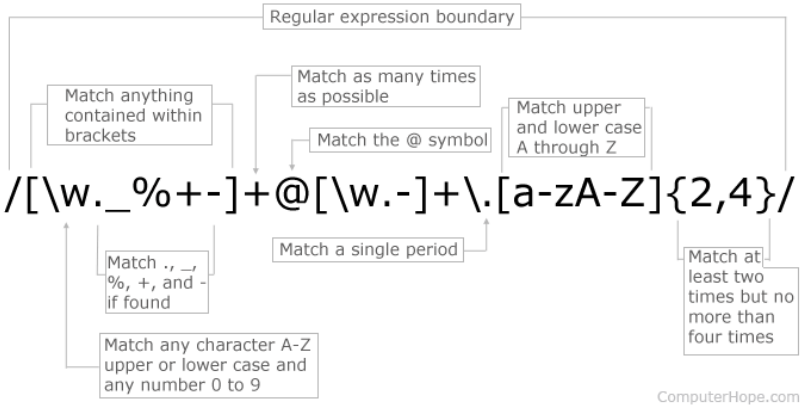
TXT, CSV, XML



LibreOffice
The Document Foundation

Sublime Text

python



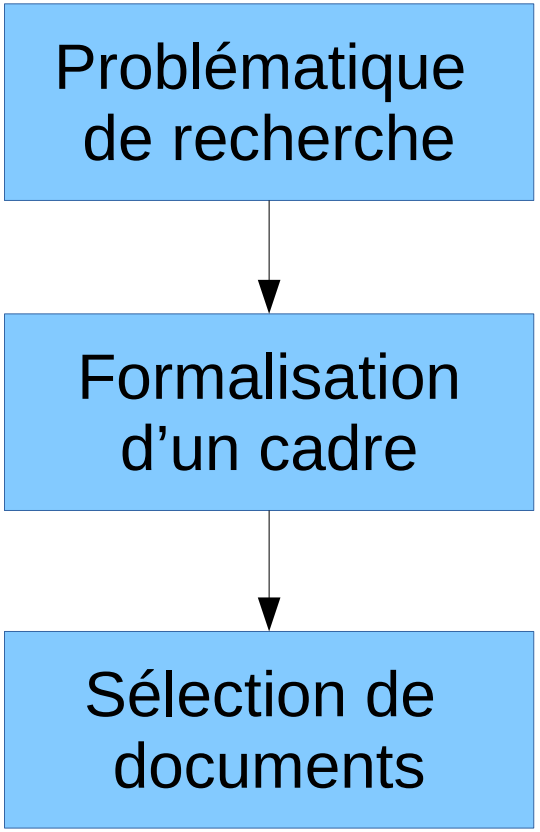
Plan du bloc méthodo

- Planification d’un corpus
  - Types de corpus
  - Codage des caractères, format CSV, licences
  - Planification pratique d’un corpus
    - *Exemples : Scientext, Presto, COLAJE*
    - Pratique sur un thème au choix en CSV
- Formatage XML
  - Étude et extraction à partir d’un corpus existant
  - Formatage en XML TEI-Lite
- Scrapping Web + normalisation
  - Semi-automatisation en Python

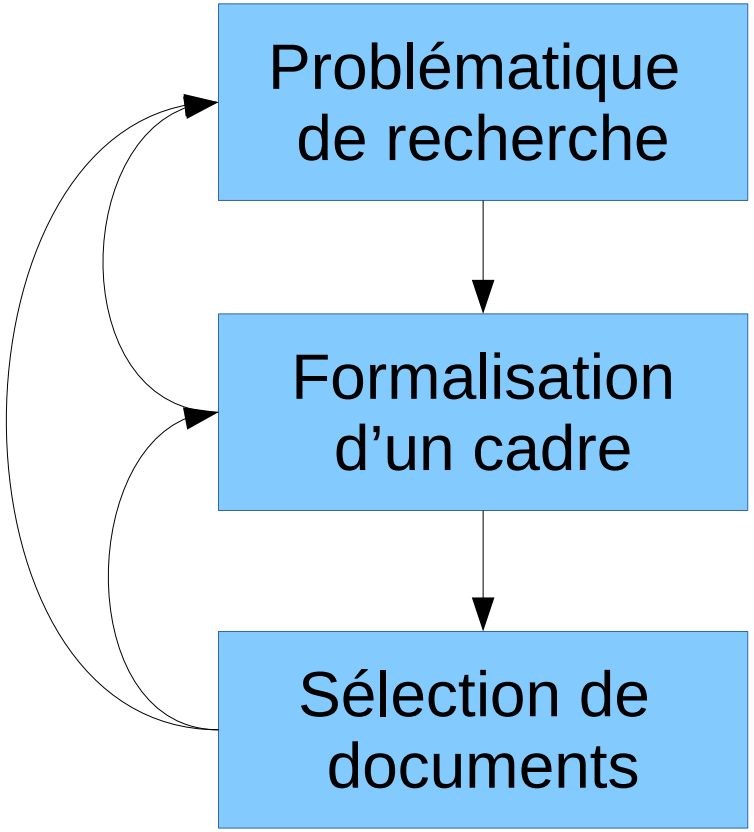
Types de corpus

- Les corpus existent dans la plupart des disciplines
- Sur une finalité / un thème donné
  (≠ pas de « corpus général de la langue »)
  - Collection d'anecdotes
    - Permet de prouver que qqch existe
  - Collection de documents
    - Permet d'étudier la fréquence de qqch
  - Collection structurée de documents
    - Permet d'étudier la répartition de qqch

Démarche scientifique

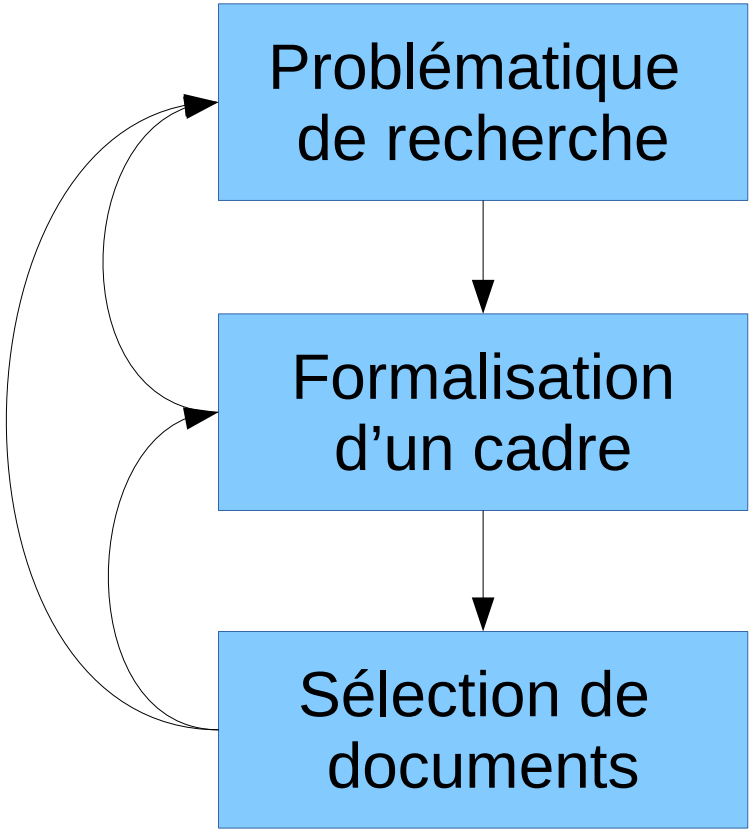


Démarche scientifique



Démarche scientifique

Ex. corpus *Scientext*

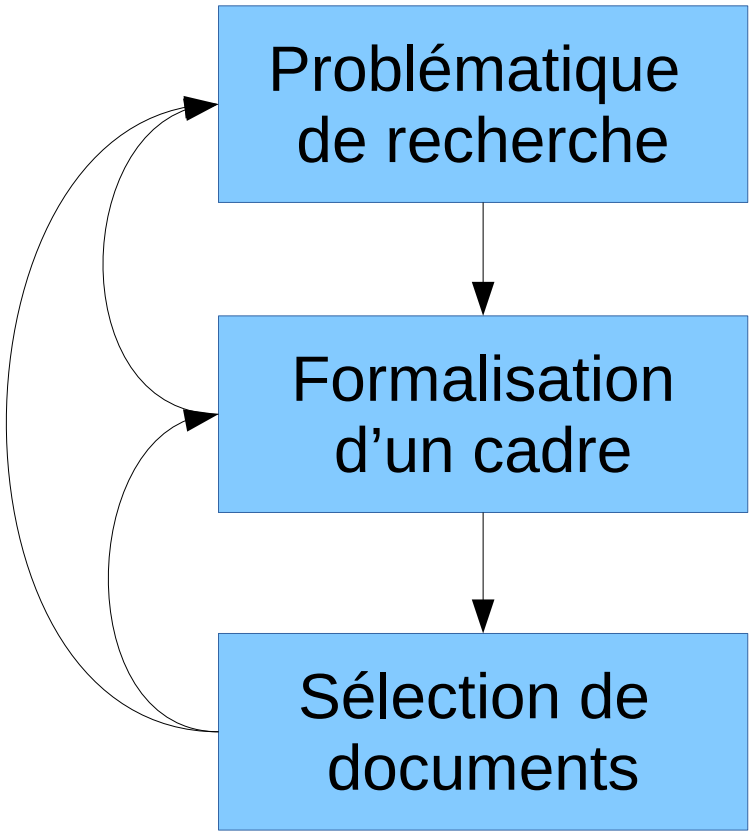


La phraséologie des écrits scientifiques.
Y a-t-il des spécificités de l'écrit scientifique ?
Est-ce un genre homogène ?

Écrits relevant de plusieurs sous-genres et disciplines.
Structuration des textes (intro, développement, conclusion, notes, annexes, titres, citations...) et de la mise en forme (gras, italique).

Démarche scientifique

Ex. corpus *Presto*

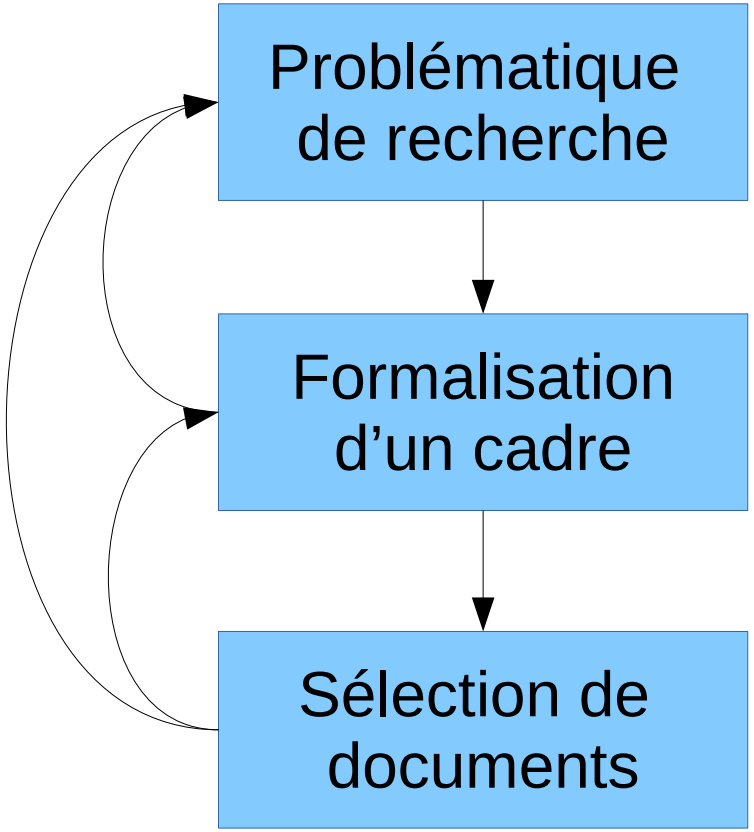


L'évolution des prépositions en français.
Comment la distribution (voisinage) des prépositions en français a-t-elle évolué entre le XVIe et le XXIe siècle ?

Écrits relevant de plusieurs genres et périodes.

Démarche scientifique

Ex. corpus *Colaje*



Évolution du langage chez le jeune enfant.

Corpus multimodal (vidéo/audio/transcription).
Identification des enfants et des adultes (rôle).

Types de corpus

- Très nombreux types
- En linguistique :
  - Modalité : corpus écrits, oraux, vidéo, multimodaux, discours/interaction…
    - Codage/transcription de tout/partie de ce qui est non-écrit
  - Langue : corpus monolingues, multilingues comparables, multilingues alignés…

En pratique

- Choisissez un thème, pour lequel vous aurez à collecter des documents pour un corpus
  - 4 genres textuels imposés : articles de presse, texte encyclopédique, texte scientifique, blog
  - Noter : le titre, l'URL, le genre textuel, le codage, le format, la licence, la taille et 2 variables au choix
  - À organiser dans un tableau au format CSV (1 ligne par document)

Codage des caractères
ASCII, ISO-8859, UTF-8 et tous leurs amis

- Les ordinateurs fonctionnent avec des nombres, pas avec des caractères
  - Table de caractères, nombre → caractère

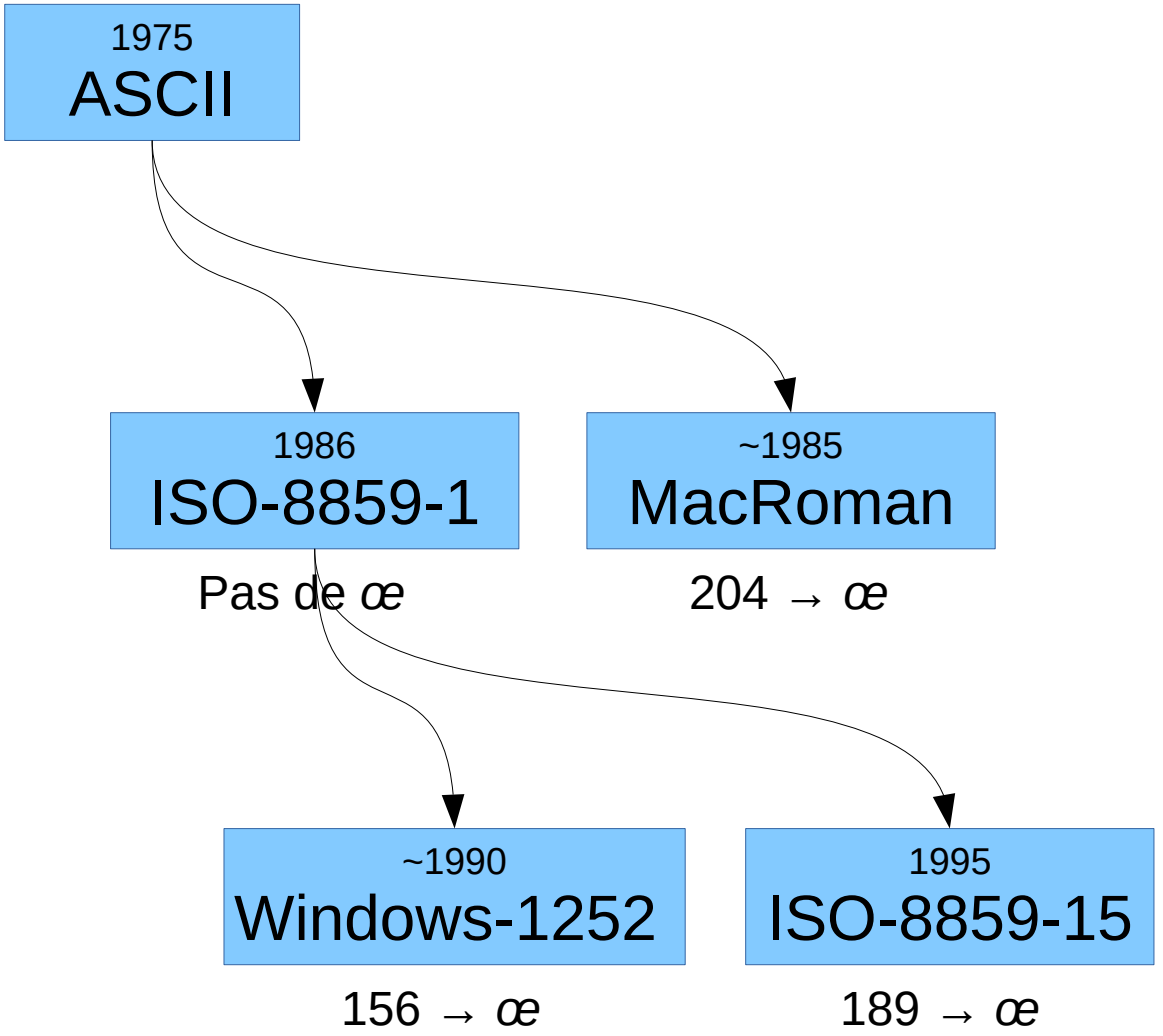
Co
AS

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
32	20	[SPACE]	64	40	@	96	60	`
33	21	!	65	41	A	97	61	a
34	22	"	66	42	B	98	62	b
35	23	#	67	43	C	99	63	c
36	24	$	68	44	D	100	64	d
37	25	%	69	45	E	101	65	e
38	26	&	70	46	F	102	66	f
39	27	'	71	47	G	103	67	g
40	28	(	72	48	H	104	68	h
41	29	)	73	49	I	105	69	i
42	2A	*	74	4A	J	106	6A	j
43	2B	+	75	4B	K	107	6B	k
44	2C	,	76	4C	L	108	6C	l
45	2D	-	77	4D	M	109	6D	m
46	2E	.	78	4E	N	110	6E	n
47	2F	/	79	4F	O	111	6F	o
48	30	0	80	50	P	112	70	p
49	31	1	81	51	Q	113	71	q
50	32	2	82	52	R	114	72	r
51	33	3	83	53	S	115	73	s
52	34	4	84	54	T	116	74	t
53	35	5	85	55	U	117	75	u
54	36	6	86	56	V	118	76	v
55	37	7	87	57	W	119	77	w
56	38	8	88	58	X	120	78	x
57	39	9	89	59	Y	121	79	y
58	3A	:	90	5A	Z	122	7A	z
59	3B	;	91	5B	[	123	7B	{
60	3C	<	92	5C	\	124	7C	|
61	3D	=	93	5D	]	125	7D	}
62	3E	>	94	5E	^	126	7E	~
63	3F	?	95	5F	_	127	7F	[DEL]

Ce que « voit » l'ordinateur.

Ce que voit l'utilisateur.

Codage des caractères
ASCII, ISO-8859, UTF-8 et tous leurs amis



Codages utilisés pour représenter du français.

Codage des caractères
ASCII, ISO-8859, UTF-8 et tous leurs amis

- ISO-8859-1 à -15 : langues européennes + arabe, hébreu, thaï, turc.
- KOI-7 et KOI-8 : latin + cyrillique
- CJK : Big5, EUC-JP...

Le caractère n°198 selon la norme...												
ISO-8859-...												KOI-8
...-1	...-2	...-3	...-4	...-5	...-6	...-7	...-8	...-9	...-10	...-11	...-13	
Æ	Ć	Ĉ	Ć	X	۶	Z		Æ	Æ	ฦ	Ę	ф

Mojibake

ウィキペディア
フリー百科事典

æ–‡å—åŒ–ã ‘

å‡ºå…¸: ãƒ•ãƒªãƒ¼ç™¾ç§‘äº‹å…¸ã€Žã‚¦ã‚£ã‚ãƒšãƒ‡ã‚£ã‚¢ï¼ˆWikipediaï¼‰ã€

W ã‚¦ã‚£ã‚ãƒšãƒ‡ã‚£ã‚¢ã§ã¯ ®æ–‡å—åŒ–ã ‘ã «ã ¤ã „ã ¦ã ¯ã€ Help:ç‰¹æ®Šæ–‡å—ã,'ã "è¦§ã ã ã •ã „ã€,

æ–‡å—åŒ–ã ‘ï¼ˆã„ã ¯ã °ã ‘ï¼‰ã ¨ã ¯ã€ ã‚³ãƒ³ãƒ”ãƒ¥ãƒ¼ã‚¿ ã§æ–‡å—ã,'è¡¨ç¤ºã ™ã,‹éš›ã «ã€ æ£ã —ã è¡¨ç¤ºã •ã,Œã ªã „ç ¾è±¡ã ®ã "ã ¨ã€,

- ä¾‹ï¼šã€Œæ–‡å—åŒ–ã ‘ã€ ã Œã€ ã€Œ Ã¦â€“‡å—â€"Ã¥Â'â€"Ã£Â â€¯ã€ ã„ã€Œè ï¿½ï¿½æ€œå–§ç„¡ï¿½ã€ ã ¨è¡¨ç¤ºã •ã,Œã,‹ ªã ©ã€,

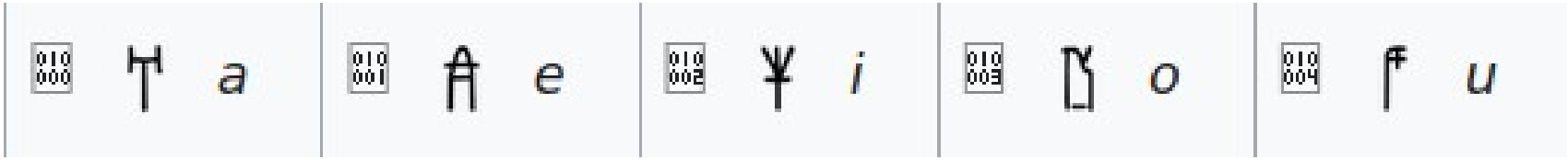
ã€Œæ–‡å—åŒ–ã ‘ã€ ã ¨ã „ã †è¨€è‘‰ã ¯ã€ ã,³ãƒ³ãƒ”ãƒ¥ãƒ¼ã,¿ç'°å¢ƒã §åŽŸå‰‡ ¨ã —ã ¦ãƒžãƒ«ãƒ ã,¤ãƒˆæ–‡å—ã,'ä½¿ç"¨ã —ã ªã „æ¬§ç±³ç‰ ®ãƒ©ãƒ†ãƒ³ã,¢ãƒ«ãƒ•ã,¡ãƒ™ãƒƒãƒˆä½¿ç"¨è¨€èªžã «ã Šã „ã ¦ã ¯è©²å½"ã ™ã,‹ç"¨èªžã Œå¯åœ¨ã —ã ªã ‹ã £ã Ÿã "ã ¨ã ‹ã,‰ã€ æ—¥æœ¬èªž ã ® â€œMojibakeâ€ ã ¨ã „ã †è¨€è‘‰ã Œã ã ®ã ¾ã ¾é€šç"¨ã ™ã,‹ã,ˆã †ã «ã ªã £ã Ÿã€,â†'#Mojibake

ç›®æ¬¡ [é ž表示]

1 ä¸»ã ªåŽŸå›
- 1.1 æ–‡å—ã,³ãƒ¼ãƒ‰ã ®å¤§ã ã •
- 1.2 æ–‡å—ã,³ãƒ¼ãƒ‰ã ®é •ã „
- 1.3 ã,¨ãƒ³ã,³ãƒ¼ãƒ‡ã,£ãƒ³ã,°
- 1.4 æ–‡å—ãƒ•ã,©ãƒ³ãƒˆã ®é •ã „
- 1.5 ç‰¹å®šã ®æ©Ÿèƒ½æŒ‡å®š
- 1.6 æ–‡å—åˆ‡ã,Œ

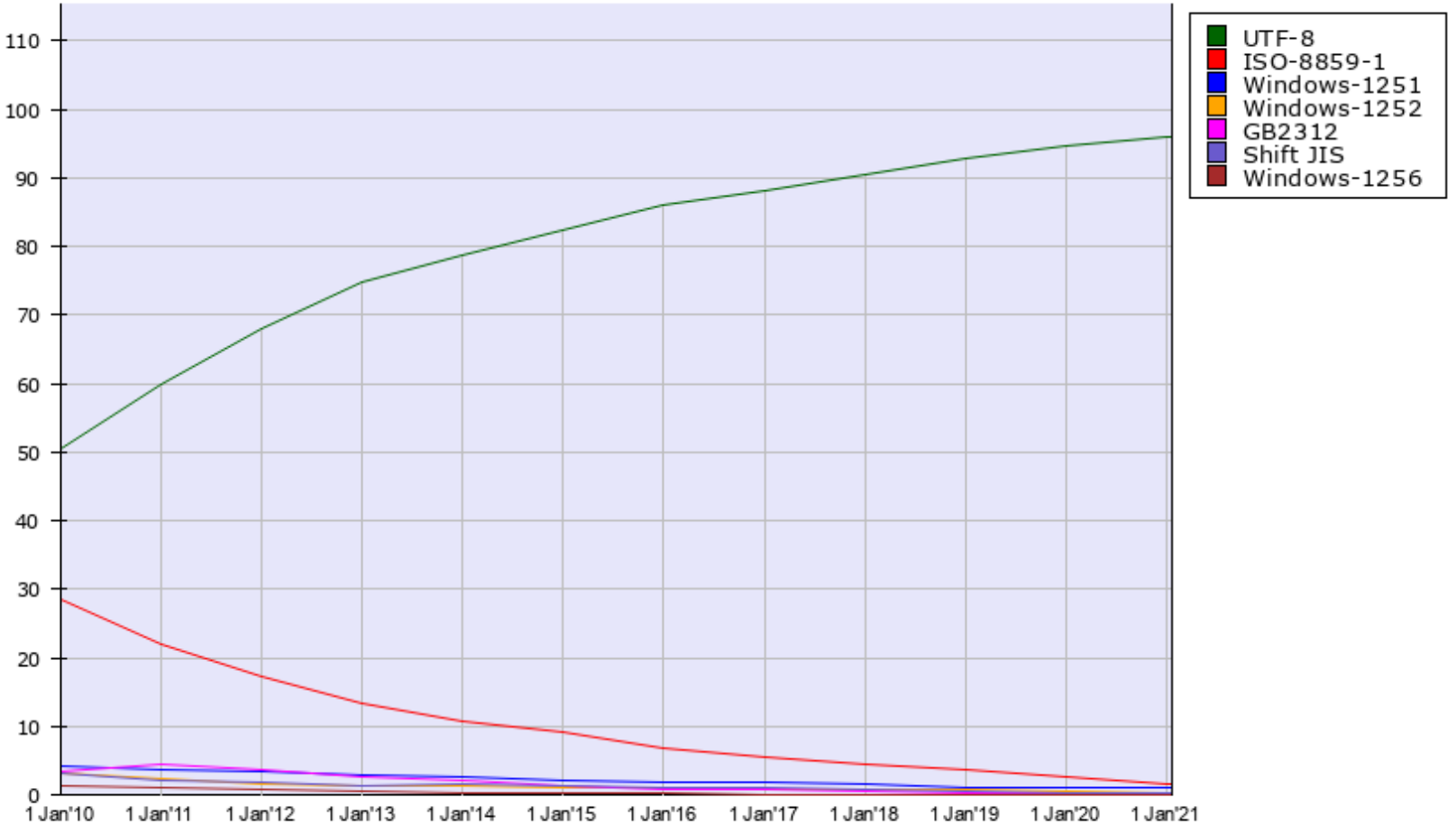
Unicode / UTF-8
Les sauveurs de l’humanité

- Unicode
  - Liste numérotée de *tous* les caractères des langues humaines.
- UTF-8
  - Une façon efficace de noter des numéros souvent petits (caractères latins), mais parfois très grands (hiéroglyphes).
- /!\ Les polices de caractères (*Arial*, *Times*, *etc*.)



Unicode / UTF-8
Les sauveurs de l’humanité

- https://w3techs.com/technologies/history_overview/character_encoding/ms/y

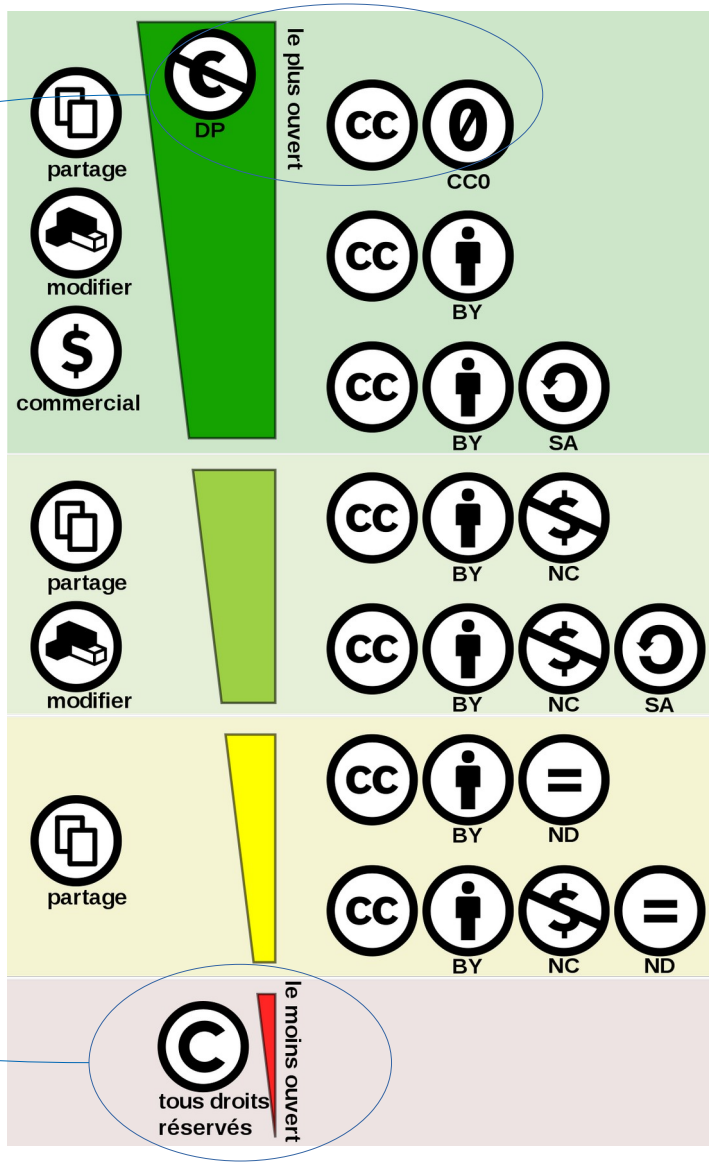


Licences

C'est un tarif, pas une licence !

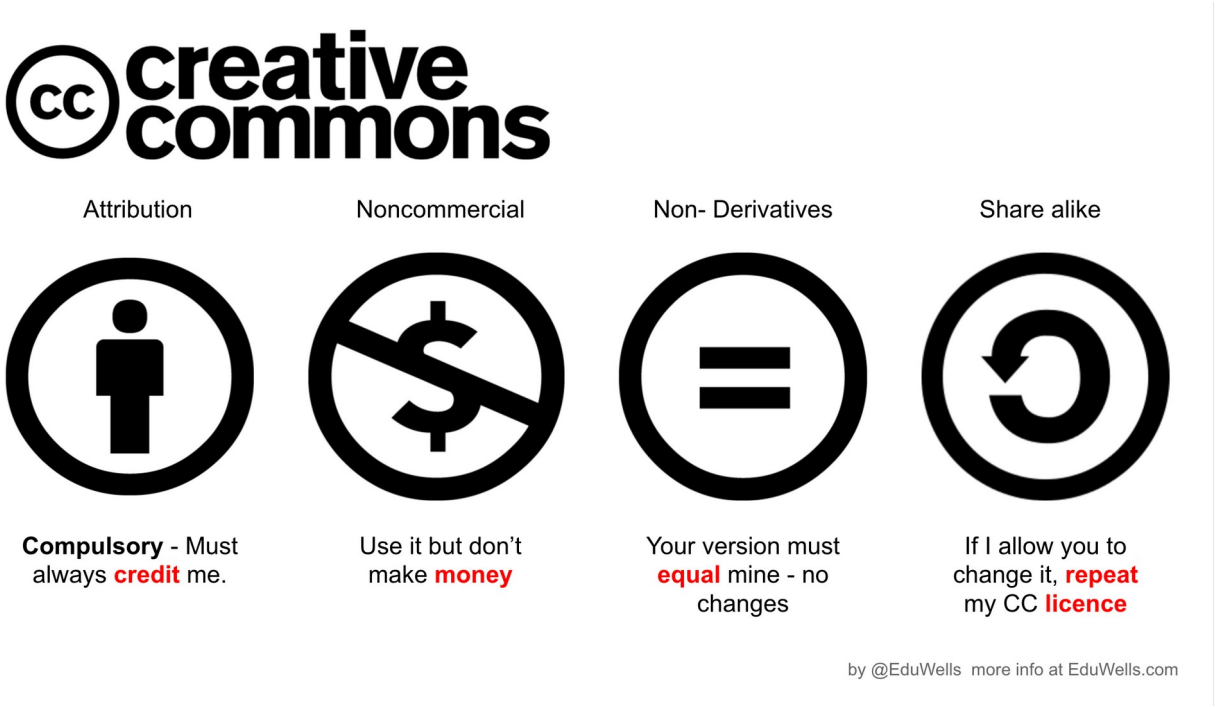
C'est une licence, mais qui n'existe pas en France

- Gratuit ≠ « Libre de droits »
  - « Libre de droits » n'existe pas dans le droit français
- Vérifier que la licence est libre
  - ou en obtenir une
    - demande écrite et précise au titulaire des droits (pas toujours l'auteur)
  - ou se contenter de la copie privée
    - donc pas de repartage !



Licences

- Attention, il y a différents types de licences libres
  - Et pour *Creative Commons*, il y a plusieurs paramètres



Données personnelles

Toute donnée qui permet d'identifier une personne.

Art. 2, al. 2 de la loi n° 78-17 du 6 janv. 1978 « *(...) Constitue une donnée à caractère personnel toute information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres. Pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens en vue de permettre son identification dont dispose ou auxquels peut avoir accès le responsable du traitement ou toute autre personne (...)* »

Souvent, on n'en a pas besoin, le plus simple est donc...
... de ne pas en avoir.

Sinon, voir :
http://ct3.ortolang.fr/download/corpus_droit.guide.nmp.pdf

Format CSV

- Comma-Separated Values
  - Un tableur, où les colonnes sont séparées par des virgules (comma)

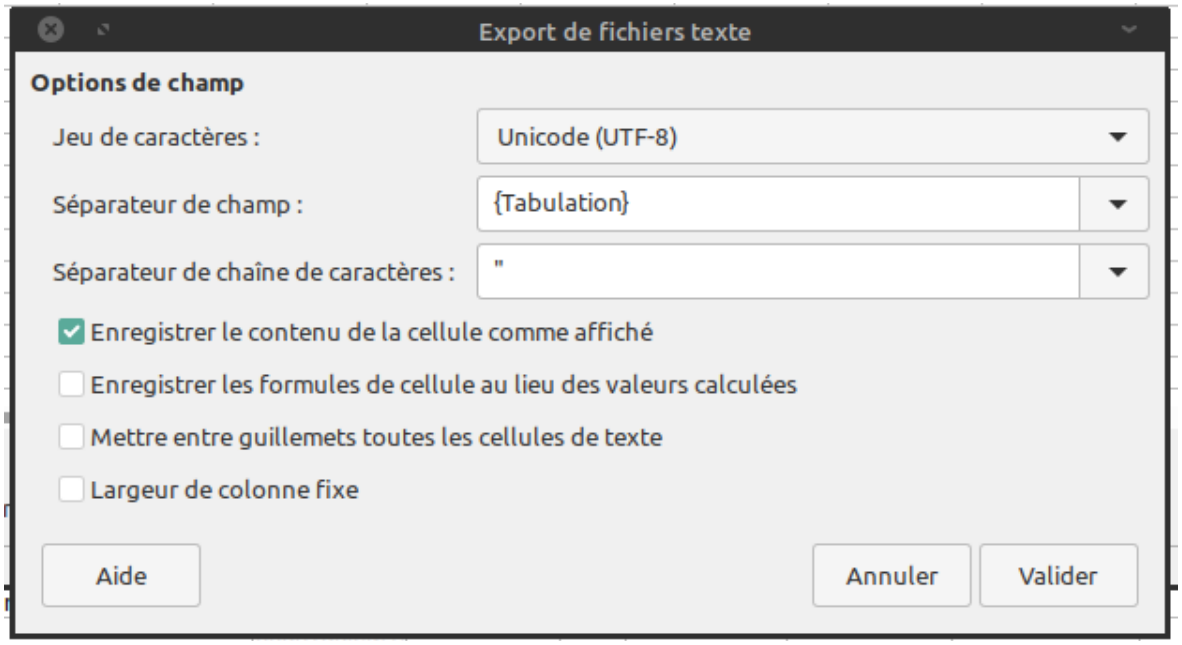
Year	Make	Model	Description	Price
1997	Ford	E350	ac, abs, moon	3000.00
1999	Chevy	Venture "Extended Edition"		4900.00
1999	Chevy	Venture "Extended Edition, Very Large"		5000.00
1996	Jeep	Grand Cherokee	MUST SELL! air, moon roof, loaded	4799.00

The above table of data may be represented in CSV format as follows:

```
Year,Make,Model,Description,Price
1997,Ford,E350,"ac, abs, moon",3000.00
1999,Chevy,"Venture ""Extended Edition""","",4900.00
1999,Chevy,"Venture ""Extended Edition, Very Large""",,5000.00
1996,Jeep,Grand Cherokee,"MUST SELL!
air, moon roof, loaded",4799.00
```

Format CSV

- Mais en pratique...
  - Un tableur, où les colonnes sont séparées par des point-virgules (Excel)
  - Un tableur, où les colonnes sont séparées par des tabulations (informaticiens)



En pratique 1/2

- Choisissez un thème, pour lequel vous aurez à collecter des documents pour un corpus
  - 4 genres textuels imposés : articles de presse, texte encyclopédique, texte scientifique, blog
  - Noter : le titre, l'URL, le genre textuel, le codage, le format (HTML, PDF...), la licence, la taille (nb mots) et 2 variables au choix
    (→ 8 colonnes en tout)
  - À organiser dans un tableau au format CSV
    (1 ligne par document, séparateur = tabulation)

En pratique 2/2

- Trouver sur le Web 5 documents pour chaque genre textuel imposé
  - Ils peuvent venir du même site, mais pour chaque genre, essayez d'avoir au moins :
    - un document sous droit d'auteur
    - un document sous licence libre
  - Vérifier la qualité
    - Le document est-il représentatif ? Pertinent pour votre thème ? Rédigé par un humain ?
  - Interdiction de piocher dans un corpus existant ;-)
  - Document CSV à rendre au + tard mercredi soir !