Recueil et structuration de corpus

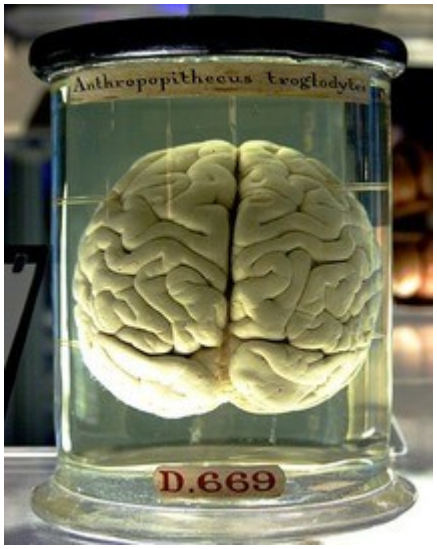
Achille Falaise − Alexandre Roulois

Contenu du cours

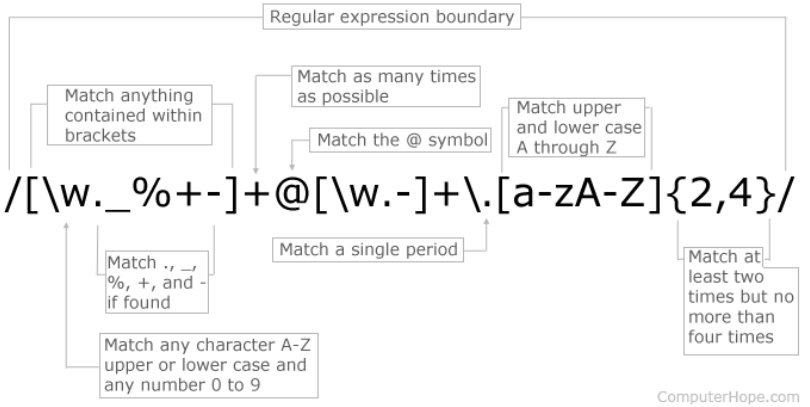
Choix des sources
Choix des textes
Licence, éthique

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

TXT, CSV, XML



Regular Expression E-mail Matching Example



Pourquoi utiliser des corpus ?

- Linguistique empirique : **sciences** du langage
  - Il faut des **données** quantifiables et contrôlées pour vérifier les hypothèses

Corpus écrit

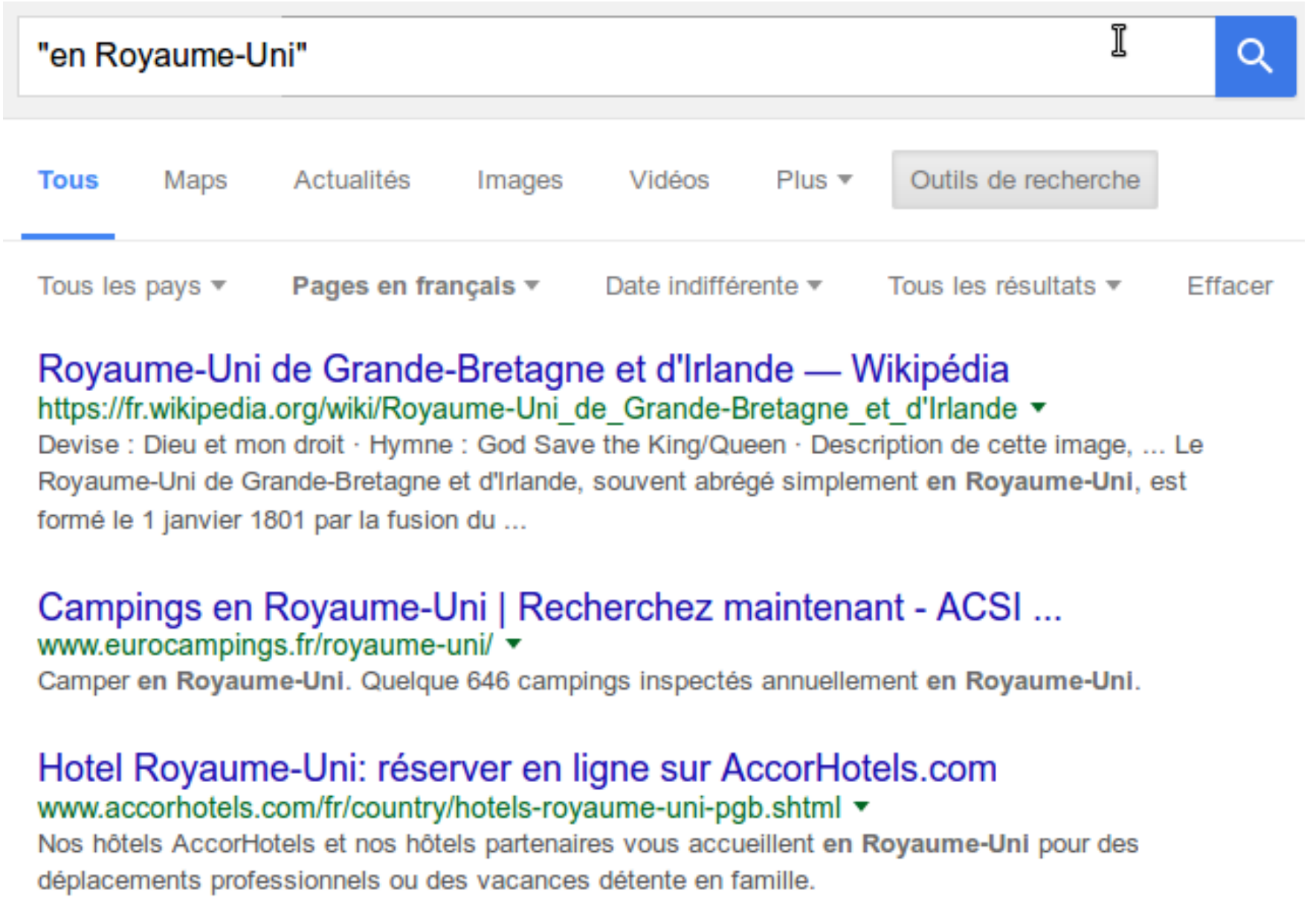
Corpus oral

Corpus multimodal

Qu'est-ce qu'un corpus ?

- Un corpus scientifique doit être quantifiable et contrôlé
  - Taille
  - Langue
  - Date
  - Genre textuel, conditions d'énonciation
  - Auteur
- Une notion relative
- Le Web est-il un corpus ?

Le Web est-il un corpus ?



Choix des textes

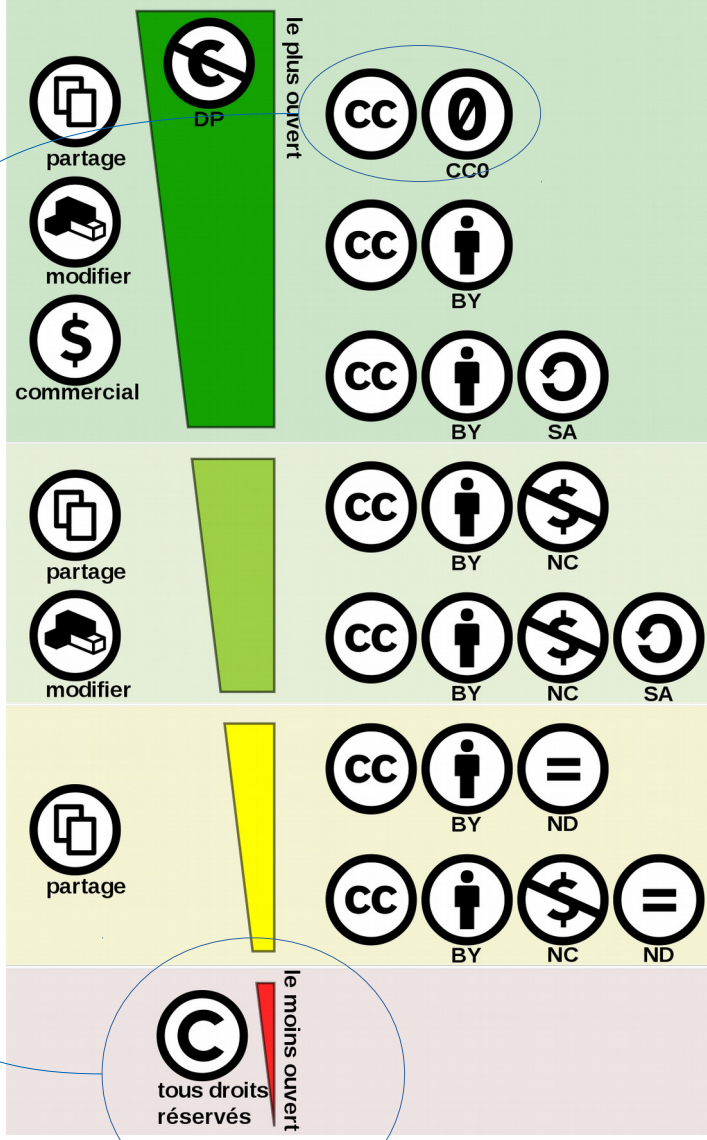
- Définir un cadre
  - Date(s)
  - Genre(s) textuel(s)
  - D'autres variables…
  - … et essayer de bien le couvrir !
- Un corpus témoin ?
- Nombre de mots
  - Plus il y en a, mieux c'est !
    - > 1M mots
  - Équilibrer
- Licence

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 la licence
  - ou en obtenir une
  - ou se contenter de la copie privée



creative commons

Attribution	Noncommercial	Non- Derivatives	Share alike
Compulsory - Must always **credit** me.	Use it but don't make **money**	Your version must **equal** mine - no changes	If I allow you to change it, **repeat** my CC **licence**

by @EduWells more info at EduWells.com

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

Contenu du cours

Choix des sources
Choix des textes
Licence, éthique

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

TXT, CSV, XML

Les documents électroniques :
langues,
scripts,
caractères,
types de documents

Pour définir un document

- Quelle langue, et quel symbole pour la langue ?
  - français, Français, francais, French, f, fr, fre, fra… ?
  - Norme : ISO 639
- Quel script, et quel symbole pour le script ?
  - chinois (traditionnel, simplifié, boponofo), serbo-croate (cyrillique, latin), turc, malais (arabe, latin)…
  - Norme : ISO 15924

Codage de la langue : ISO 639

language	639-1	639-2 (B/T)	639-3 type	639-3 code
English	en	eng	individual	eng
German	de	ger/deu	individual	deu
Arabic	ar	ara	macro	ara
			individual	arb + others
Chinese	zh	chi/zho[4][5]	macro	zho
Mandarin			individual	cmn
Cantonese			individual	yue
Minnan			individual	nan

Le meilleur !

https://en.wikipedia.org/wiki/ISO_639-3

- Quelques cas intéressants :
  - bavarois ≠ allemand
  - hindoustani -> hindi, urdu
  - serbo-croate -> serbe, croate, monténégrin, bosniaque (2008)

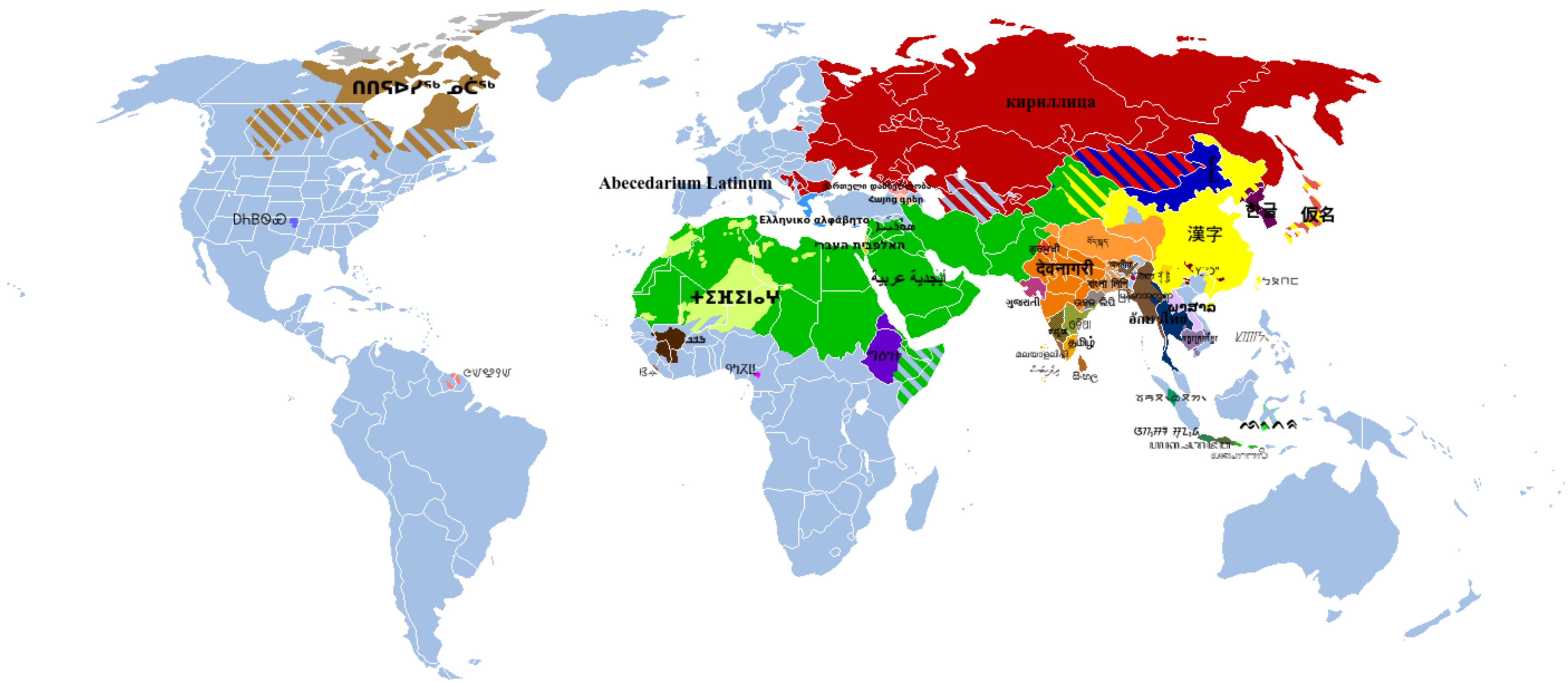
Codage du pays : ISO 3166

https://en.wikipedia.org/wiki/List_of_ISO_3166_country_codes

- Quand le pays est connu/pertinent
  - Souvent utilisé avec ISO 639-1 :
    - fr-fr, fr-ch, fr-be, fr-re

Codage du script : ISO 15924

https://en.wikipedia.org/wiki/ISO_15924



Intercontinental
- Abecedarium Latinum — Alphabet latin
- Кириллица — Alphabet cyrillique
- أبجدية عربية — Alphabet arabe

Amériques
- Syllabaire Inuktitut et Cri
- DhBᎾᏍ — Syllabaire cherokee
- Afaka

Europe
- Ελληνικο αλφάβητο — Alphabet grec

Afrique
- Tifinagh et Néo-Tifinagh
- N'ko
- Bamoun
- Alphasyllabaire gue'ez
- Syllabaire Vaï

Moyen-Orient/ Caucase
- ქართული დამწერლობა — Alphabet géorgien
- Հայոց գրեր — Alphabet arménien
- האלפבית העברי — Alphabet hébreu
- Alphabet syriaque

Chine et Asie de l'Est
- Mongol bitchig
- Alphasyllabaire tibétain
- Syllabaire yi
- 漢字 — Sinogrammes
- Bopomofo
- 한글 — Hangeul
- Syllabaire Miao
- 仮名 — Hiragana + Katakana

Inde
- देवनागरी — Devanagari
- ਗੁਰਮੁਖੀ — Gurmukhi
- ગુજરાતી — Alphasyllabaire gujarātī
- ಕನ್ನಡ — Alphasyllabaire kannada
- മലയാളലിപി — Alphasyllabaire malayalam
- Thâna
- සිංහල — Alphabet cingalais
- தமிழ் — Alphasyllabaire tamoul
- Télégou
- Utkala Lipi
- বাংলা লিপি — Alphasyllabaire bengali
- অসমীয়া — Alphabet assamais
- Metei Mayek
- Écriture birmane

Asie du Sud-Est
- อักษรไทย — Alphasyllabaire thaï
- Alphasyllabaire khmer
- Alphabet Lao
- Alphabet Hanuno'o
- Écriture javanaise
- Alphabet soundanais
- Écriture balinaise
- Alphasyllabaire batak
- Alphasyllabaire lontara

Codage du script : ISO 15924

https://en.wikipedia.org/wiki/ISO_15924

注音符號 srpskohrvatski-hrvatskosrpski
注音符号 српскохрватски-хрватскосрпски
ㄅㄆㄇㄈ

हिन्दुस्तानी Bahasa Melayu Türkçesi
ہندوستانی بهاس ملايو تُرکجۂ
تركجه

Un peu de maths

Complétez !

10	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
8	0	1	2	3	4	5	6	7	10	11	12	13	14	15	16	17
16	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
2	0	1	10	11	100	101	110	111	1000	1001	1010	1011	1100	1101	1110	1111

Les caractères

≠ Polices de caractères

http://alexandre.roulois.fr/data/supports/html/HTML150.html

Les jeux de caractères codés

Un charset ?

- définition

Charset : *Character set*, association d'un caractère abstrait avec une représentation numérique (décimale, octale, hexadécimale…)

Indispensable à l'échange d'informations sur Internet

Le Morse, l'ASCII ou l'UTF-8 sont des jeux de caractères

Chaque système (serveur, BDD, système de fichiers…) doit savoir dans quel format sont échangées les informations

Interopérabilité

Les jeux de caractères codés

Un charset ?

- définition
- encodage

Cerveau décode plus ou moins bien les caractères :

小さな猫は牛乳を飲んでいます。

Mały kot pije mleko.

Des traits forment des glyphes qui correspondent à des caractères

Processus décodage par blocs pour obtenir mots et phrases

Lepetitchatboitdulait.

Segmentation + décodage (français) =

Le petit chat boit du lait.

Les jeux de caractères codés

Un charset ?

- définition
- encodage

Informatique : deux caractères pour tout coder !

Bit : unité informatique qui revêt deux formes, `0-1`

Un bit = un caractère ? Si a = 0 et b = 1

?? ????? ??0? 1??? ?? ?0??.

Impossible car bien plus de deux caractères nécessaires !

Deux bits = un caractère ? Encore insuffisant…

- 00 = a
- 01 = b
- 10 = c
- 11 = d

?? ????? 10?00? 01??? 11? ?00??.

Les jeux de caractères codés

Un charset ?

- définition
- encodage

Question : Combien de bits pour afficher tous les caractères nécessaires ?

Inventaire du besoin :

- [Leptichabodul.] = 15 caractères (ponctuation incluse)

Combien de bits pour représenter 14 caractères ?

- 1 bit = 2 caractères (2^1)
- 2 bits = 4 caractères (2^2)
- 3 bits = 8 caractères (2^3)
- 4 bits = 16 caractères (2^4)

Les jeux de caractères codés

Un charset ?

- définition
- encodage

Charset personnalisé :

- 0000 = a
- 0001 = b
- 0010 = c
- …
- 1101 = espace
- 1110 = point

Encodage binaire :

```
1100 0100 1101 1001 0100 1010 0110 1010 1101 0010 0101
0000 1010 1101 0001 1000 0110 1010 1101 0001 1011 1101
0111 0000 0110 1010 1110
```

Résultat : segmentation en blocs de 4 bits, décodable grâce au *charset* défectif défini plus haut

Remarque : en informatique, stockage sur 8 bits (octet)

Les jeux de caractères codés

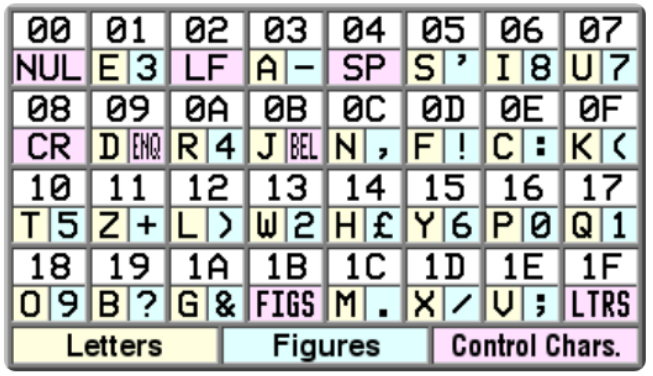
Un charset ?

- définition
- encodage
- code Baudot

1832 : code Morse

- caractère associé à signal (lumière, son, geste)
- communication souvent chiffrée

1874 : code Baudot



00	01	02	03	04	05	06	07
NUL	E 3	LF	A −	SP	S ’	I 8	U 7
08	09	0A	0B	0C	0D	0E	0F
CR	D ENQ	R 4	J BEL	N ,	F !	C :	K (
10	11	12	13	14	15	16	17
T 5	Z +	L)	W 2	H £	Y 6	P 0	Q 1
18	19	1A	1B	1C	1D	1E	1F
O 9	B ?	G &	FIGS	M .	X /	V ;	LTRS

Letters | Figures | Control Chars.

Les deux jeux de caractères du code Baudot US

- codage sur 5 bits
- 2^5 soit 32 caractères
- 2 jeux de caractères, soit $2 \times 32 = 64$ caractères
- caractères spéciaux `LTRS` et `FIGS` pour basculer
- système probabiliste

Les jeux de caractères codés

Un charset ?

- définition
- encodage
- code Baudot

Exemple : *On boit le thé à 16 heures.*

Encodage Baudot :

```
18 0C 04 19 18 06 10 04 12 01 04 10 14 01 04 03 04 1B 17 15
04 1F 14 01 07 0A 01 05 1B 1C
```

Transcription :

```
o n SP b o i t SP l e SP t h e SP a SP FIGS 1 6 SP LTRS
h e u r e s FIGS .
```

Remarques :

- `LTRS` et `FIGS` pour basculer entre jeux de caractères
- message commence probablement par une lettre
- certains caractères communs aux deux jeux (`SP` `LF` …)
- perte majuscules et diacritiques
- poids : $30 \text{ signes} \times 5 \text{ bits} = 150 \text{ bits}$ (232 en ASCII)

Les jeux de caractères codés

Un charset ?

- définition
- encodage
- code Baudot
- entités HTML

Tout document est stocké dans un système de fichiers (bit/octet) selon un jeu spécifique

Un logiciel peut encoder différemment du jeu de caractères déclaré dans le document

Le navigateur, le serveur et le document peuvent utiliser des jeux différents

Les entités de caractères abolissent ces frontières

Syntaxe simple : `&entite;`

Quelques entités : `é` , `à` , `¨` , `ζ` , `€` ...

Liste complète des entités HTML

Les jeux de caractères codés

Un charset ?

Vers l'uniformisation

- ASCII

ASCII : *American Standard Code for Information Interchange* (1968)

- 62 caractères : A-Z, a-z, 0-9
- 33 contrôles : sauts, tabulations…
- 33 signes de ponctuation

128 caractères codés sur 7 bits. Or octet = 8 bits

Lettre	Base décimale	Base binaire
C	67	01000011
a	97	01100001
t	116	01110100

Exemple de codage sur 7 bits, le 1er étant 0

Avantage : économie de stockage

Limite : prolifération de systèmes *ad hoc*

Les jeux de caractères codés

Un charset ?

Vers l'uniformisation

- ASCII

ASCII TABLE

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(	72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29	)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[SHIFT IN]	47	2F	/	79	4F	O	111	6F	o
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P	112	70	p
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q	113	71	q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R	114	72	r
19	13	[DEVICE CONTROL 3]	51	33	3	83	53	S	115	73	s
20	14	[DEVICE CONTROL 4]	52	34	4	84	54	T	116	74	t
21	15	[NEGATIVE ACKNOWLEDGE]	53	35	5	85	55	U	117	75	u
22	16	[SYNCHRONOUS IDLE]	54	36	6	86	56	V	118	76	v
23	17	[ENG OF TRANS. BLOCK]	55	37	7	87	57	W	119	77	w
24	18	[CANCEL]	56	38	8	88	58	X	120	78	x
25	19	[END OF MEDIUM]	57	39	9	89	59	Y	121	79	y
26	1A	[SUBSTITUTE]	58	3A	:	90	5A	Z	122	7A	z
27	1B	[ESCAPE]	59	3B	;	91	5B	[	123	7B	{
28	1C	[FILE SEPARATOR]	60	3C	<	92	5C	\	124	7C	|
29	1D	[GROUP SEPARATOR]	61	3D	=	93	5D	]	125	7D	}
30	1E	[RECORD SEPARATOR]	62	3E	>	94	5E	^	126	7E	~
31	1F	[UNIT SEPARATOR]	63	3F	?	95	5F	_	127	7F	[DEL]

Table de correspondance ASCII

Les jeux de caractères codés

Un charset ?

« *Le petit chat boit du lait.* » en ASCII :

Vers l'uniformisation

- ASCII

```
01001100 01100101 00100000 01110000 01100101 01110100
01101001 01110100 00100000 01100011 01101000 01100001
01110100 00100000 01100010 01101111 01101001 01110100
00100000 01100100 01110101 00100000 01101100 01100001
01101001 01110100 00001010
```

Sur un terminal :

```
$ echo "Le petit chat boit du lait" | xxd -b
```

Remarques :

- message ne tient pas compte des caractères de contrôle
- codage sur 7 bits
- 1er bit = 0
- poids information supérieur en binaire

Les jeux de caractères codés

Un charset ?

Vers l'uniformisation

- ASCII

Comparaison poids de la chaîne « *chat* » :

- manuscrit : *chat* (4 signes)
- ASCII binaire : `01100011 01101000 01100001 01110100` (32 signes)
- ASCII hexadécimal : `63 68 61 74` (8 signes)

Hexadécimal économique pour stockage 8 bits

Lettre	ASCII binaire	ASCII décimal	ASCII hexadécimal
c	01100011	99	63

Convertir en décimal ?

$$01100011 = 0 \times 2^7 + 1 \times 2^6 + 1 \times 2^5 + 0 \times 2^4 + 0 \times 2^3 + 0 \times 2^2 +$$
$$01100011 = 0 + 64 + 32 + 0 + 0 + 0 + 2 + 1 = 99$$

Convertir en hexadécimal ? Octet séparé en deux blocs de 4 bits `0110` et `0011`

$$0110 = 0 \times 2^3 + 1 \times 2^2 + 1 \times 2^1 + 0 \times 2^0 = 0 + 4 + 2 + 0 = 6$$
$$0011 = 0 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 = 0 + 0 + 2 + 1 = 3$$

Les jeux de caractères codés

Un charset ?

Évolution du codage des protocoles de communication Internet sur 8 bits (soit 1 octet).

Vers l'uniformisation

ASCII étendu à 256 caractères (2^8) :

- développement de l'ISO-8859-1 ou Latin1 pour les langues européennes, puis ISO-8859-15 (Latin9) qui introduit le symbole €
- langues asiatiques évoluent de leur côté
- documents illisibles à l'international, dès que l'on change les systèmes

- ASCII

Les jeux de caractères codés

Un charset ?

Vers l'uniformisation

- ASCII
- Unicode

1991 : Unicode 1.0 par le consortium Unicode

Représenter tout caractère (110 000 env.), peu importe le système d'écriture (alphabétique, syllabique, logographique…) :

- nom
- identifiant numérique

Compatible avec la norme ISO/CEI 10646 dont il est un sous-ensemble.

Chaque caractère dispose d'un *point de code* :

- préfixé U+
- base hexadécimale sur 4 à 6 caractères selon le plan

Les jeux de caractères codés

Un charset ?

Points de code encodés selon un format (UTF-8, UTF-16…)

Vers l'uniformisation

- ASCII
- Unicode

Caractère	Nom	Point de code	Représentation binaire UTF-8
ἐ	Lettre minuscule grecque epsilon esprit doux	U+1F10	11100001 10111100 10010000
地	Marque d'annotation idéographique de la terre	U+319E	11100011 10000110 10011110
ܬ	Lettre syriaque taw	U+072C	11011100 10101100

Attention ! Absence de représentation si le caractère ne figure pas dans la police d'écriture

Les jeux de caractères codés

Un charset ?

Vers l'uniformisation

- ASCII
- Unicode
- UTF-8

1992 : UTF-8 (*Universal character set Transformation Format 8 bits*) par Kenneth Thompson

Système de codage sur 4 octets maximum

Rétrocompatible avec les anciens systèmes (1 octet)

Potentiel de 2^{32} caractères : capable de représenter tous les systèmes d'écriture

Proposé en 1996 au consortium Unicode, il est universel deux ans après

Les jeux de caractères codés

Un charset ?

Vers l'uniformisation

- ASCII
- Unicode
- UTF-8

UTF-8 est-il 4 fois plus lourd que l'**ASCII** ? (4 octets vs 1)

Caractère	ASCII	UTF-8 ?
C	01000011	01000011 00000000 00000000 00000000
a	01100001	01100001 00000000 00000000 00000000
t	01110100	01110100 00000000 00000000 00000000

Cat en ASCII pèserait 3 octets contre 12 en UTF-8 ?

Faux ! UTF-8 à la fois compatible ASCII et Unicode !

Caractère	ASCII	UTF-8
C	01000011	01000011
a	01100001	01100001
t	01110100	01110100

Règle d'économie de l'UTF-8

Les jeux de caractères codés

Un charset ?

Vers l'uniformisation

- ASCII
- Unicode
- UTF-8

Caractéristiques :

- 4 octets maximum
- si 1er bit vaut `0` : caractère ASCII donc 1 octet
- si 1er et 2e bits valent `11` : 1er octet d'une suite
- si 1er et 2e bits valent `10` : octet d'une suite

Exemple :

Mały kot pije mleko.

UTF-8 :

01001101 01100001 **11**000101 **10**000010 01111001 00100000
01101011 01101111 01110100 00100000 01110000 01101001
01101010 01100101 00100000 01101101 01101100 01100101
01101011 01101111 00001010

Les jeux de caractères codés

Un charset ?

Vers l'uniformisation

- ASCII
- Unicode
- UTF-8

Remarques :

- ł = Unicode (`11000101 10000010`)
- autres = ASCII (1 octet, premier bit vaut `0`)

Conclusion : UTF-8 cumule les avantages de l'ASCII et de l'Unicode

Les jeux de caractères codés

Un charset ?

Vers l'uniformisation

MIME

MIME : *Multipurpose Internet Mail Extension*

A l'origine, ensemble de codages supplémentaires à l'ASCII pour e-mails + supports de fichiers binaires

Aujourd'hui, permet d'affiner le dialogue entre le navigateur et le serveur Web

Syntaxe de la déclaration : `type/sous-type`

Fonctionnement :

1. serveur Web détermine le type du fichier téléchargé grâce à l'extension
2. information transmise au navigateur via l'en-tête HTTP
3. navigateur anticipe l'utilisation qui sera faite du fichier

Les jeux de caractères codés

Un charset ?

Vers l'uniformisation

MIME

Comment définir, pour le navigateur, l'en-tête HTTP ?

- via une directive dans le fichier `.htaccess` à la racine du site :

```
AddDefaultCharset UTF-8
```

- via une instruction PHP :

```php
<?php
    header('Content-Type: text/html;charset=UTF-8');
?>
```

- via l'élément meta dans la section `head` du code source de la page HTML :

```
<meta charset="UTF-8">
```

Les caractères

- Les caractères bizarres
  - 
  - 

Les caractères

- Dans les documents HTML
  - En spécifiant un charset
  - Entités HTML :
    - œ
    - œ
    - œ

https://en.wikipedia.org/wiki/List_of_XML_and_HTML_character_entity_references