

Licence, département Informatique, CNAM
Multimédia & interaction humain-machine (2004-5)

Le texte

P. Cubaud <cubaud@cnam.fr>

- 1. Codes, structures**
- (2. Transport, compression, cryptage)**
- 3. Analyse**
- (4. Synthèse)**
- 5. Présentation**

Bibliographie

B. Habert, C. Fabre, F. Isaac
De l'écrit au numérique
InterEditions, 1998

I.H. Witten, A. Moffat, T.C. Bell
*Managing gigabytes. Compressing and indexing
documents and images*
Van Nostrand, 1994

Techniques de l'ingénieur
série H. section Document numérique
(en ligne au CNAM)

R. Laufer, D. Scavetta
Texte, hypertexte, hypermédia
Que-sais-je n°2629 (v2 1995)

1. Codes - Structures

Codage des caractères

Trois problèmes :

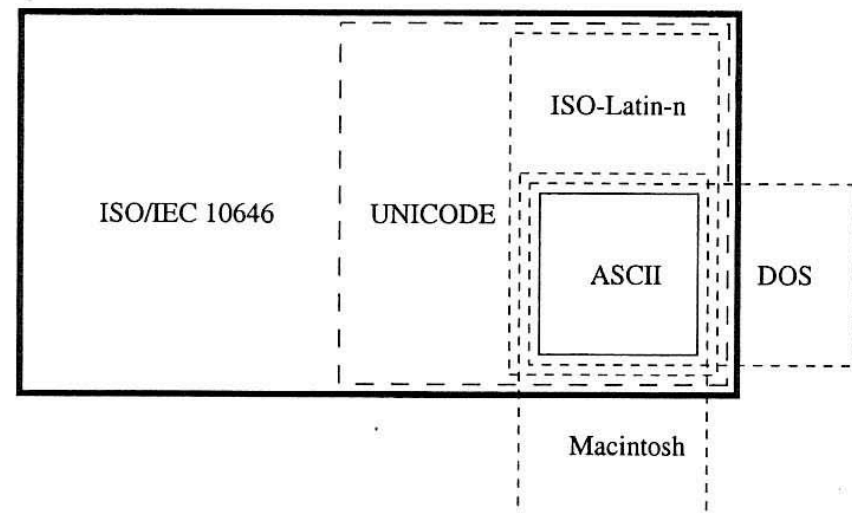
- caractère \neq glyphe
- coder \neq classer
- norme \neq standard

Aussi vieux que le télégraphe... et toujours non résolus.

Tableau 1 – Relation entre le nombre de moments et le nombre de caractères pour les principaux codages		
Nom du codage	Nombre de bits ou moments	Nombre de caractères
Telex	6	64
Ascii	7	128
ISO Latin-1	8	256
Unicode	16	65 536
ISO/IEC 10646	32	> 2 milliards

Nombre de bits

— 32
-- 16
--- 8
—— 7



7 bits : American Standard Code for Information Interchange (ASCII, 1967 - puis ISO646 en 1983)

Tableau 6 – Codage Ascii dans sa version finale de 1983 (ISO 646)								
	000	001	002	003	004	005	006	007
0	NUL	DLE	SP	0	@	P	'	p
1	STX	DC1	!	1	A	Q	a	q
2	SOT	DC2	"	2	B	R	b	r
3	ETX	DC3	#	3	C	S	c	s
4	EOT	DC4	\$	4	D	T	d	t
5	ENQ	NAK	%	5	E	U	e	u
6	ACK	SYN	&	6	F	V	f	v
7	BEL	ETB	'	7	G	W	g	w
8	BS	CAN	(8	H	X	h	x
9	HT	EM)	9	I	Y	i	y
A	LF	SUB	*	:	J	Z	j	z
B	VT	ESC	+	;	K	[k	{
C	FF	FS	,	<	L	\	l	
D	CR	GS	-	=	M]	m	}
E	SO	RS	.	>	N	^	n	~
F	SI	US	/	?	O	_	o	DEL

Les codes sont donnés en hexadécimal (par exemple Q a pour code 0051₁₆).

code	acronyme	nom anglais	nom français
0	<NUL>	NUL	Nul
1	<SOH>	Start Of Heading	Début d'en-tête
2	<STX>	Start of TeXt	Début de texte
3	<ETX>	End of TeXt	Fin de texte
4	<EOT>	End Of Transmission	Fin de transmission
5	<ENQ>	ENQuiry	Requête
6	<ACK>	ACKnowledgement	Accusé de réception
7	<BEL>	BELl	Sonnerie
8	<BS>	BackSpace	Espace arrière
9	<HT>	Horizontal Tabulation	Tabulation horizontale
10	<LF>	Line Feed	Interligne
11	<VT>	Vertical Tabulation	Tabulation verticale
12	<FF>	Form Feed	Présentation de feuille
13	<CR>	Carriage Return	Retour chariot
14	<SO>	Shift Out	Hors-code
15	<SI>	Shift In	En-code
16	<DLE>	DataLink Escape	Échappement à la transmission
17	<DC1>	Device Control 1	Contrôle de périphérique 1
18	<DC2>	Device Control 2	Contrôle de périphérique 2
19	<DC3>	Device Control 3	Contrôle de périphérique 3
20	<DC4>	Device Control 4	Contrôle de périphérique 4
21	<NAK>	Negative AcKnowledge	Accusé de reception négatif
22	<SYN>	SYNchronous idle	Synchronisation
23	<ETB>	End of Transmission Block	Fin de bloc de transmission
24	<CAN>	CANcel	Annulation
25		End of Medium	Fin de support
26	<SUB>	SUBstitute	Substitution
27	<ESC>	ESCape	Échappement
28	<FS>	File Separator	Séparateur de fichiers
29	<GS>	Group Separator	Séparateur de groupes de données
30	<RS>	Record Separator	Séparateur d'enregistrement
31	<US>	Unit Separator	Séparateur d'unités

TAB. 3.2 – Codes ASCII inférieurs à 32.

8 bits : ISO-Latin-XXX (ISO 8859-n)

Tableau 9 – Codage ISO 8859-1 (Latin-1), seconde partie (la première partie est équivalente au codage Ascii du tableau 6)								
	008	009	00A	00B	00C	00D	00E	00F
0	XXX	DCS	NBSP	°	À	Ð	à	ð
1	XXX	PU1	¡	±	Á	Ñ	á	ñ
2	BPH	PU2	¢	²	Â	Ò	â	ò
3	NBH	STS	£	³	Ã	Ó	ã	ó
4	IND	CCH	¤	´	Ä	Ô	ä	ô
5	NEL	MW	¥	µ	Å	Ö	å	ö
6	SSA	SPA		¶	Æ	Ø	æ	ø
7	ESA	EPA	§	·	Ç	×	ç	×
8	HTS	SOS	¨	¸	È	Ø	è	ø
9	HTJ	XXX	@	¹	É	Ù	é	ù
A	VTs	SCI	ª	º	Ê	Ú	ê	û
B	PLD	CSI	«	»	Ë	Û	ë	ü
C	PLU	ST	¬	¼	Ì	Ü	ì	ü
D	RI	OSC	-SHY	½	Í	Ý	í	ý
E	SS2	PM	@	¾	Î	Þ	î	þ
F	SS3	APC	–	¿	Ï	ß	ï	ÿ

Tableau 11 – Parties spécifiques des codages ISO 8859-2 (Latin-2 : Europe orientale) et 8859-7 (grec)												
ISO 8859-2							ISO 8859-7					
	00A	00B	00C	00D	00E	00F	00A	00B	00C	00D	00E	00F
0	NBSP	°	Ŕ	Ŗ	ŕ	ŗ	NBSP	°	ı	Π	ΰ	π
1	Ą	ą	Ą	Ń	ą	ń	•	±	Α	Ρ	α	ρ
2	ˆ	˙	Ă	Ń	â	ņ	•	²	Β		β	ς
3	Ł	ł	Ą	Ō	ā	ó	£	³	Γ	Σ	γ	σ
4	Ł	ˆ	Ą	Ō	ā	ō		ˆ	Δ	Τ	δ	τ
5	Ł	ł	Ł	Ō	í	o		ˆ	Ε	Υ	ε	υ
6	Š	š	Č	Ō	č	ö	ı	ˆ	Ζ	Φ	ζ	φ
7	š	ˆ	Č	×	č	+	š	ˆ	Η	Χ	η	χ
8	ˆ	˙	Č	Ř	č	ř	ˆ	ˆ	Θ	Ψ	θ	ψ
9	Š	š	É	Ů	é	ů	©	ˆ	Η	Ι	ι	ω
A	Š	š	Ę	Ů	ę	ú		ˆ	Κ	Ι	κ	ĩ
B	Ť	ť	Ě	Ů	ě	ú	«	»	Λ	Υ	λ	ü
C	Ž	ž	Ě	Ů	ě	ū	¬	ˆ	Ο	Μ	α	μ
D	-	ˆ	Ĭ	Ÿ	ı	ý	-	½	Ν	έ	ν	ó
E	Ž	ž	Ĭ	Ÿ	ı	ť		ˆ	Ξ	ή	ξ	ώ
F	Ž	ž	Đ	β	d'	ˆ	—	ˆ	Ω	Ο	ι	ο

Tableau 10 – Comparaison Latin-1/Latin-9			
Code hexadécimal	Glyphe Latin-1	Remplacé par le caractère de Latin-9	Glyphe Latin-9
00A4	¤	SYMBOLE EURO	€
00A6		LETTRE MAJUSCULE LATINE S CARON	Š
00A8	¨	LETTRE MINUSCULE LATINE S CARON	š
00B4	ˆ	LETTRE MAJUSCULE LATINE Z CARON	Ž
00B8	˙	LETTRE MINUSCULE LATINE Z CARON	ž
00BC	¼	DIGRAMME SOUDÉ MAJUSCULE LATIN OE	Œ
00BD	½	DIGRAMME SOUDÉ MINUSCULE LATIN OE	œ
00BE	¾	LETTRE MAJUSCULE LATINE Y TRÉMA	ÿ

Au delà : Unicode (1990, v3 en 2000) et ISO 10646

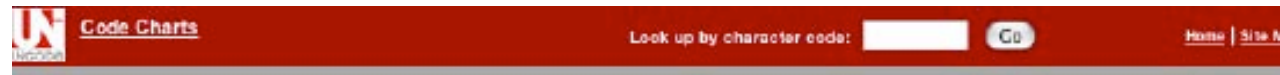
...	...
0041	LATIN CAPITAL LETTER A
...	...
A356	YI SYLLABLE ZHUT
...	...
FC78	ARABIC LIGATURE THEH WITH MEEM FINAL FORM
...	...

1 – **16-bits** : Unicode code donc les caractères sur deux octets, ce qui permet 65 536 valeurs, dont 63 486 sont affectées pour des caractères simples et 2 048 permettent de coder 10 448 544 autres caractères par l'emploi de paires d'octets additionnelles. Un sys-

İ = I + ı

Tableau 14 – Exemples de définitions de caractères par Unicode (le troisième exemple est en français, selon la norme)			
Code	Glyphe	Description	Commentaires
0026	&	AMPERSAND	nom officiel
002E	.	FULL STOP = PERIOD =dot, decimal point • may be rendered as a raised decimal point in old style numbers → 06D4 - arabic full stop → 3002 °ideographic full stop	nom officiel nom optionnel noms optionnels renseignement renvois
00E5	à	LETTRE MINUSCULE LATINE A ROND EN CHEF • danois, norvégien, suédois, wallon = 0061 a 030A °	décomposition canonique

Site web unicode: www.unicode.org



The Unicode Character Code Charts By Script

[SYMBOLS AND PUNCTUATION](#) | [NAME INDEX](#) | [HELP AND LINKS](#)

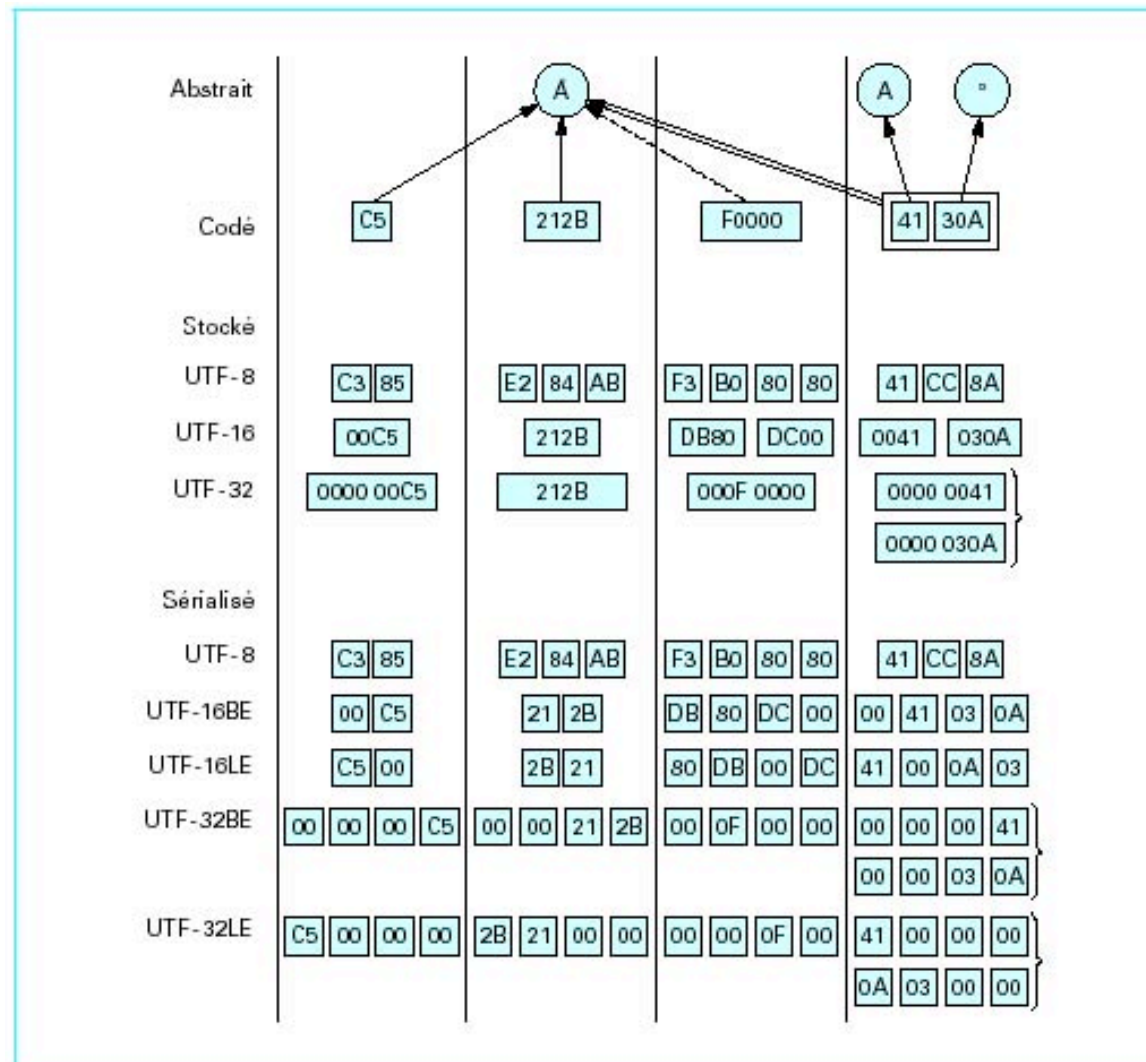
European Alphabets	African Scripts	Indic Scripts	East Asian Scripts	Central Asian Scripts
(see also Comb. Marks)	Ethiopic	Bengali	Han Ideographs	Kharoshthi
Armenian	Ethiopic	Devanagari	Unified CJK Ideographs (5MB)	Mongolian
Armenian Supplement	Ethiopic Supplement	Gujarati	CJK Ideographs Ext. A (2MB)	Phags-Pa (5.0)
Armenian Ligatures	Ethiopic Extended	Gurmukhi	CJK Ideographs Ext. B (13MB)	Tibetan
Coptic	Other African scripts	Kannada	Compatibility Ideographs (.5MB)	
Coptic	NKs (5.0)	Limbu	Compatibility Ideo. Suppl. (.5MB)	
Coptic in Greek block	Tifinagh	Malayalam	Kanbun	
Cyrillic	Middle Eastern Scripts	Orya	(see also Unihan Database)	Ancient Scripts
Cyrillic	Arabic	Sinhala	Radicals and Strokes	Ancient Greek
Cyrillic Supplement	Arabic	Syloti Nagri	CJK Radicals	Ancient Greek Numbers
Georgian	Arabic Supplement	Tamil	KangXi Radicals	Ancient Greek Musical
Georgian	Arabic Presentation Forms A	Telugu	CJK Strokes	Cuneiform
Georgian Supplement	Arabic Presentation Forms B		Ideographic Description	Cuneiform (5.0)
Greek	Hebrew	Philippine Scripts	Chinese-specific	Cuneiform Numbers (5.0)
Greek	Hebrew	Buhid	Bopomofo, Extended	Old Persian
Greek Extended	Hebrew Presentation Forms	Hanunoo	Japanese-specific	Ugaritic
(see also Ancient Greek)	Other ME Scripts	Tagalog	Hiragana	Linear B
Latin	Syriac	Tagbanwa	Katakana	Linear B Syllabary
Basic Latin	Thaana		Katakana Phonetic Ext.	Linear B Ideograms
Latin-1	American scripts	South East Asian	Malwidth Katakana	Other Ancient Scripts
Latin Extended A	Canadian Syllabics	Buginese	Korean-specific	Aegean Numbers
Latin Extended B	Cherokee	Balinese (5.0)	Hangul Syllables (4MB)	Counting Rod Num. (5.0)
Latin Extended C (5.0)	Deseret	Khmer	Hangul Jamo	Cypriot Syllabary
Latin Extended D (5.0)		Khmer Symbols	Hangul Compatibility Jamo	Gothic
Latin Extended Additional	Other Scripts	Leo	Malwidth Jamo	Old Italic
Latin Ligatures	Shavian	Myanmar	Yi	Ogham
Fullwidth Latin Letters	Osmanya	New Tai Lue	Yi (.6MB)	Runic
Small Forms	Olugalic	Tai Le	Yi Radicals	Phoenician (5.0)
(see also Phonetic Symbols)		Thai		

Ancient Greek Numbers

	1014	1015	1016	1017	1018
0	Ϟ	ϙ	Ϛ	ϛ	Ϝ
1	ϝ	Ϟ	ϟ	Ϡ	ϡ
2	Ϣ	ϣ	Ϥ	ϥ	Ϧ
3	ϧ	Ϩ	ϩ	Ϫ	ϫ
4	Ϭ	ϭ	Ϯ	ϯ	ϰ
5	ϱ	ϲ	ϳ	ϴ	ϵ
6	϶	Ϸ	ϸ	Ϲ	Ϻ
7	ϻ	ϼ	Ͻ	Ͼ	Ͽ
8	Ͽ	Ͽ	Ͽ	Ͽ	Ͽ
9	Ͽ	Ͽ	Ͽ	Ͽ	Ͽ
A	Ͽ	Ͽ	Ͽ	Ͽ	Ͽ
□	Ͽ	Ͽ	Ͽ	Ͽ	Ͽ

Exemple : on voit, figure 3, qu'en UTF-8, le scalaire $C5_{16}$, 1100 0101 en binaire se code (puisqu'il a plus de 7 chiffres) avec une suite de deux octets, à savoir 1100 0001 1000 0101, soit $C3_{16}$ et 85_{16} .

De même, au scalaire $212B_{16}$ correspond en UTF-8 la suite des trois octets $E2_{16}$, 84_{16} et AB_{16} .



Une table à garder en attendant...

	ISO-Latin-1	Macintosh	MS-DOS ¹⁰	MIME	SGML	L ^A T _E X
à	224	136	133	=E0	à	\‘a
â	226	137	131	=E2	â	\^a
é	233	142	130	=E9	é	\’e
è	232	143	138	=E8	è	\‘e
ê	234	144	136	=EA	ê	\^e
ë	235	145	137	=EB	ë	\"e
î	238	148	140	=EE	î	\^i
ï	239	149	139	=EF	ï	\"i
ô	244	153	147	=F4	ô	\^o
ù	249	157	151	=F9	ù	\’u
ü	252	159	129	=FC	ü	\"u
ç	231	141	135	=E7	ç	\c c
ÿ	255	216	152	=FF	ÿ	\"y
œ	-	207	-	=F7	œ	\oe
Œ	-	206	-	-	Œ	\OE
Ÿ	-	217	-	-	Ÿ	\"Y

TAB. 3.4 – Représentations de caractères accentués.

et un outil : iconv

```
[numer:~] pcutbaud% iconv -l
ANSI_X3.4-1968 ANSI_X3.4-1986 ASCII CP367 IBM367 ISO-IR-6 ISO646-US ISO_646.IRV:1991 US US-ASCII CSASCII
UTF-8
ISO-10646-UCS-2 UCS-2 CSUNICODE
UCS-2BE UNICODE-1-1 UNICODEBIG CSUNICODE11
UCS-2LE UNICODELITTLE
ISO-10646-UCS-4 UCS-4 CSUCS4
UCS-4BE
UCS-4LE
UTF-16
UTF-16BE
UTF-16LE
UTF-32
UTF-32BE
UTF-32LE
UNICODE-1-1-UTF-7 UTF-7 CSUNICODE11UTF7
UCS-2-INTERNAL
UCS-2-SWAPPED
UCS-4-INTERNAL
UCS-4-SWAPPED
C99
JAVA
CP819 IBM819 ISO-8859-1 ISO-IR-100 ISO8859-1 ISO_8859-1 ISO_8859-1:1987 L1 LATIN1 CSISOLATIN1
ISO-8859-2 ISO-IR-101 ISO8859-2 ISO_8859-2 ISO_8859-2:1987 L2 LATIN2 CSISOLATIN2
ISO-8859-3 ISO-IR-109 ISO8859-3 ISO_8859-3 ISO_8859-3:1988 L3 LATIN3 CSISOLATIN3
ISO-8859-4 ISO-IR-110 ISO8859-4 ISO_8859-4 ISO_8859-4:1988 L4 LATIN4 CSISOLATIN4
CYRILLIC ISO-8859-5 ISO-IR-144 ISO8859-5 ISO_8859-5 ISO_8859-5:1988 CSISOLATINCYRILLIC
ARABIC ASMO-708 ECMA-114 ISO-8859-6 ISO-IR-127 ISO8859-6 ISO_8859-6:1987 CSISOLATINARABIC
ECMA-118 ELOT_928 GREEK GREEK8 ISO-8859-7 ISO-IR-126 ISO8859-7 ISO_8859-7 ISO_8859-7:1987 CSISOLATINGREEK
HEBREW ISO-8859-8 ISO-IR-138 ISO8859-8 ISO_8859-8 ISO_8859-8:1988 CSISOLATINHEBREW
ISO-8859-9 ISO-IR-148 ISO8859-9 ISO_8859-9 ISO_8859-9:1989 L5 LATIN5 CSISOLATIN5
ISO-8859-10 ISO-IR-157 ISO8859-10 ISO_8859-10 ISO_8859-10:1992 L6 LATIN6 CSISOLATIN6
ISO-8859-13 ISO-IR-179 ISO8859-13 ISO_8859-13 L7 LATIN7
ISO-8859-14 ISO-CELTIC ISO-IR-199 ISO8859-14 ISO_8859-14 ISO_8859-14:1998 L8 LATIN8
ISO-8859-15 ISO-IR-203 ISO8859-15 ISO_8859-15 ISO_8859-15:1998
ISO-8859-16 ISO-IR-226 ISO8859-16 ISO_8859-16 ISO_8859-16:2000
KOI8-R CSKOI8R
KOI8-U
KOI8-RU
CP1250 MS-EE WINDOWS-1250
CP1251 MS-CYRL WINDOWS-1251
CP1252 MS-ANSI WINDOWS-1252
CP1253 MS-GREEK WINDOWS-1253
CP1254 MS-TURK WINDOWS-1254
CP1255 MS-HEBR WINDOWS-1255
CP1256 MS-ARAB WINDOWS-1256
CP1257 WINBALTRIM WINDOWS-1257
CP1258 WINDOWS-1258
850 CP850 IBM850 CSPC850MULTILINGUAL
862 CP862 IBM862 CSPC862LATINHEBREW
866 CP866 IBM866 CSIBM866
MAC MACINTOSH MACROMAN CSMACINTOSH
MACCENTRALEUROPE
MACICELAND
MACCROATIAN
MACROMANIA
MACCYRILLIC
MACUKRAINE
MACGREEK
MACTURKISH
MACHEBREW
MACARABIC
MACTHAI
HP-ROMAN8 R8 ROMAN8 CSHPROMAN8
NEXTSTEP
ARMSII-8
GEORGIAN-ACADEMY
GEORGIAN-PS
KOI8-T
MULELAO-1
CP1133 IBM-CP1133
ISO-IR-166 TIS-620 TIS620 TIS620-0 TIS620.2529-1 TIS620.2533-0 TIS620.2533-1
CP874 WINDOWS-874
VISII VISIII.1-1 CSVISII
TCVN TCVN-5712 TCVN5712-1 TCVN5712-1:1993
ISO-IR-14 ISO646-JP JIS_C6220-1969-RO JP CSISO14JISC6220RO
JISX0201-1976 JIS_X0201 X0201 CSHALFWIDTHKATAKANA
ISO-IR-87 JIS0208 JIS_C6226-1983 JIS_X0208 JIS_X0208-1983 JIS_X0208-1990 X0208 CSISO87JISX0208
ISO-IR-159 JIS_X0212 JIS_X0212-1990 JIS_X0212.1990-0 X0212 CSISO159JISX02121990
CN GB_1988-80 ISO-IR-57 ISO646-CN CSISO57GB1988
CHINESE GB_2312-80 ISO-IR-58 CSISO58GB231280
CN-GB-ISOIR165 ISO-IR-165
ISO-IR-149 KOREAN KSC_5601 KS_C_5601-1987 KS_C_5601-1989 CSKSC56011987
EUC-JP EUCJP EXTENDED_UNIX_CODE_PACKED_FORMAT_FOR_JAPANESE CSEUCPKDFMTJAPANESE
MS_KANJI SHIFT_JIS SHIFT_JIS SJIS CSSHIFTJIS
CP932
ISO-2022-JP CSISO2022JP
ISO-2022-JP-1
ISO-2022-JP-2 CSISO2022JP2
CN-GB EUC-CN EUCCN GB2312 CSGB2312
CP936 GBK
GB18030
ISO-2022-CN CSISO2022CN
ISO-2022-CN-EXT
HZ HZ-GB-2312
EUC-TW EUCTW CSEUCTW
BIG-5 BIG-FIVE BIG5 BIGFIVE CN-BIG5 CSBIG5
CP950
BIG5-HKSCS BIG5HKSCS
EUC-KR EUCKR CSEUCKR
CP949 UHC
CP1361 JOHAB
ISO-2022-KR CSISO2022KR
[numer:~] pcutbaud%
```

Structuration de documents textuels

Deux problèmes :

- séparation forme et fond
- changer / échanger

Une difficulté : qu'est-ce que décrire une structure ?

- par langage génératif
- par programme
- par contraintes ?

un
slide ?

titre

liste de points

une autre ?

Structuration de documents textuels

Deux (graves) problèmes :

- séparation forme et fond
- changer / échanger

Une difficulté : qu'est-ce que décrire une structure ?

- par langage génératif
- par règles, contraintes
- par programme

Approche par grammaire générative : les langages de balisage (« markup »)

Generalized Markup Language (GML : IBM, Goldfarb, 1971-3)

```
:h1.Chapter 1:  Introduction
:p.GML supported hierarchical containers, such as
:ol
:li.Ordered lists (like this one),
:li.Unordered lists, and
:li.Definition lists
:eol.
as well as simple structures.
:p.Markup minimization (later generalized and formalized in SGML),
allowed the end-tags to be omitted for the "h1" and "p" elements.
```

(extrait de www.sgmlsource.com/history/)

SGML : norme ISO 8879 en 1986

HTML : premiers jets en 1991

XML (eXtensible ML) : v1.0 en 1997

Un exemple de DTD : TEI (text encoding initiative, 1994)

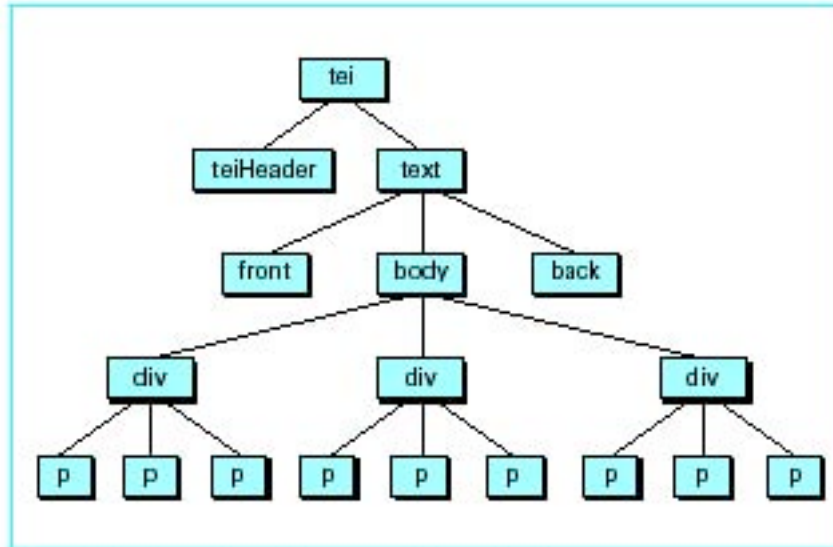


Figure 1 – Structure arborescente d'un texte en prose structuré conformément à la TEI

```
<TEI>
  <teiHeader> [informations contenues dans l'en-tête TEI]
  </teiHeader>,
  <text>
    <front> [textes préliminaires...] </front>,
    <body> [corps du texte...] </body>
    <back> [annexes...] </back>
  </text>
</TEI>
```

+ de 400 éléments
une version « Lite » existe

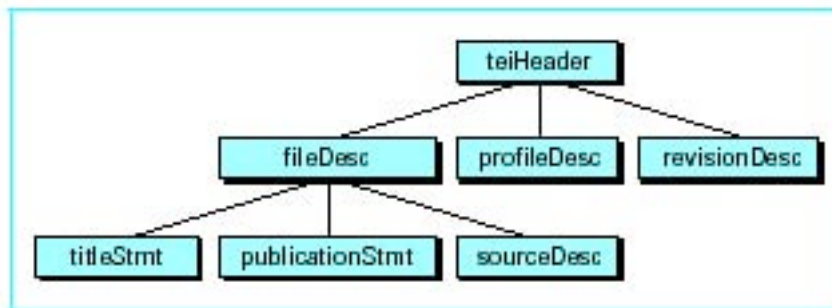


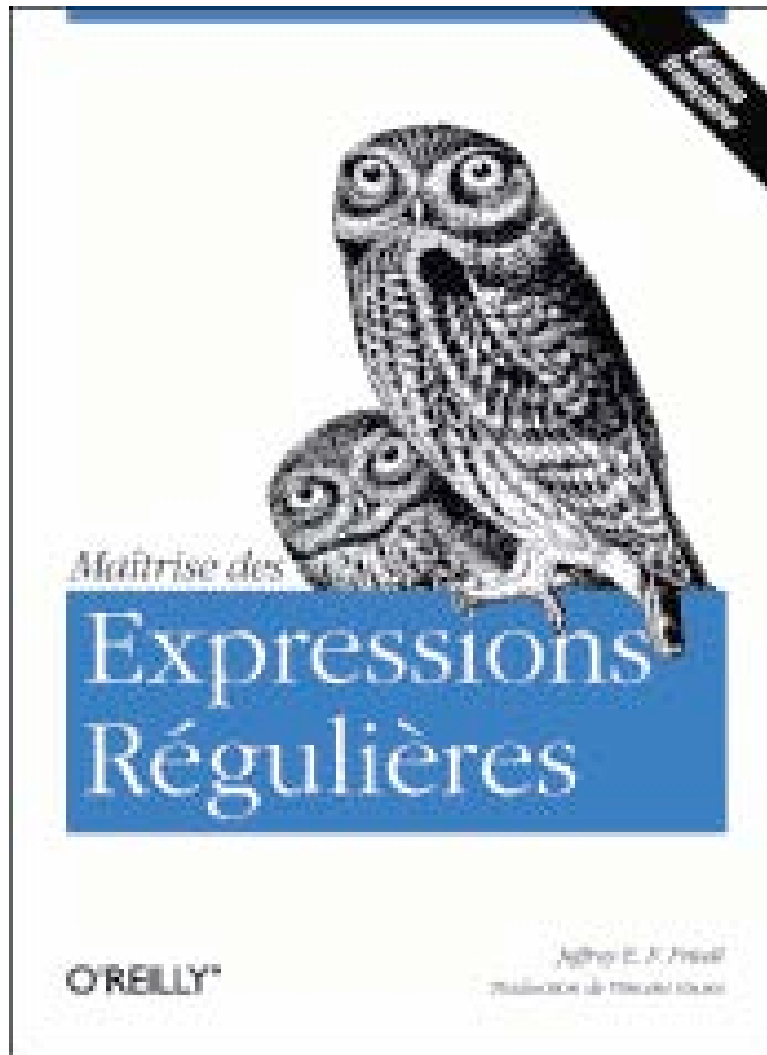
Figure 2 – Structure arborescente de l'en-tête TEI

D'une façon générale, l'utilisation massive de la TEI dépend de la capacité de ses promoteurs à tenir un équilibre entre la nécessité de structurer de façon rigoureuse pour assurer l'échange des informations et le besoin d'annotations personnalisées et très spécialisées éprouvé par chaque chercheur.

Text Encoding Initiative home page.
<http://www.uic.edu/orgs/tei/>.

2. Analyse de textes

Expressions régulières : utilisées par grep, sed, awk, perl, etc.



[abc...]

matches any of the characters *abc...*

[^abc...]

matches any character except *abc...*

r1|r2

matches either *r1* or *r2* (alternation).

r1r2

matches *r1*, and then *r2* (concatenation).

r+

matches one or more *r*'s.

r*

matches zero or more *r*'s.

r?

matches zero or one *r*'s.

(r)

matches *r* (grouping).

r{n,m}

matches at least *n*, *n* to any number, or *n* to *m* occurrences of *r* (interval expressions).

\b

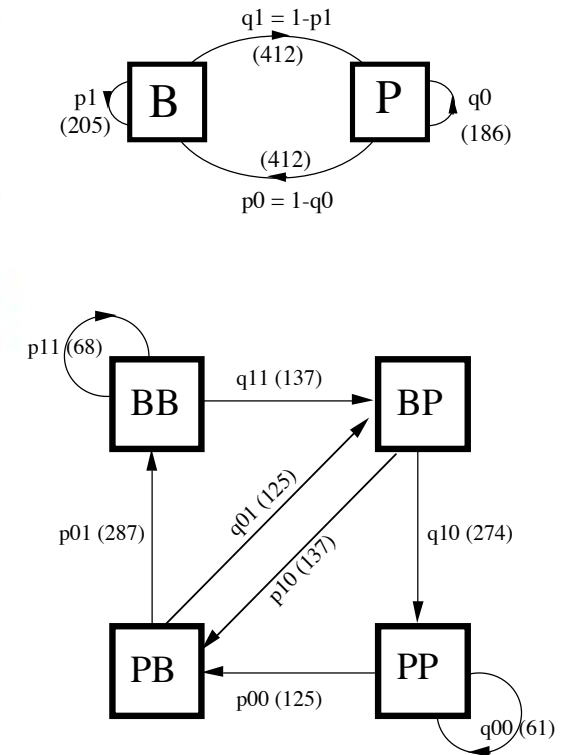
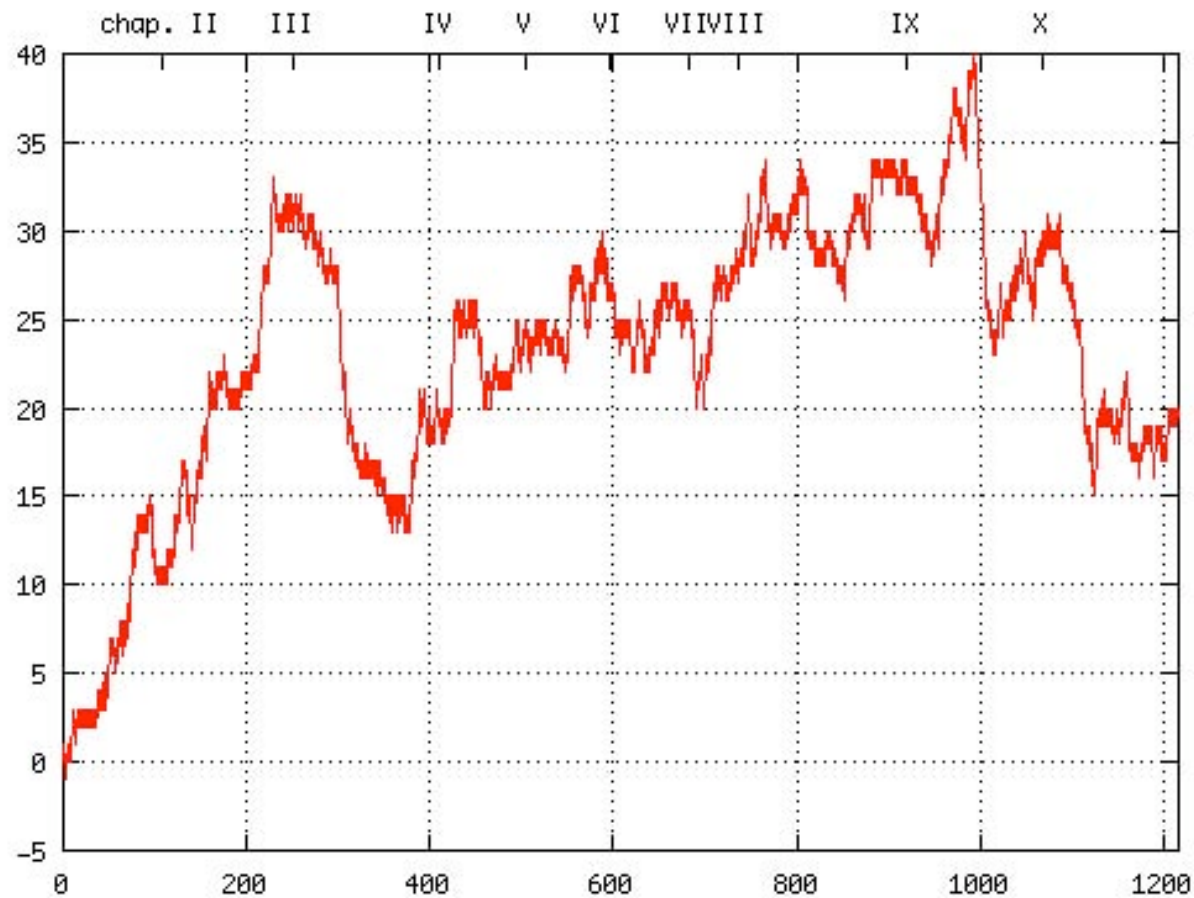
matches the empty string at either the beginning or the end of a word.

etc ...

Exemples de traitements

6	Quand ils furent arrivés au milieu du boulevard, ils s'assirent à la même minute, sur le même banc.	0	0
7	Pour s'essuyer le front, ils retirèrent leurs coiffures, que chacun posa près de soi ; et le petit homme aperçut écrit dans le chapeau de son voisin : Bouvard_1 ; pendant que celui-ci distinguait aisément dans la casquette du particulier en redingote le mot : Pécuchet_1 .	0	2
8	-- "Tiens !" dit-il "nous avons eu la même idée, celle d'inscrire notre nom dans nos couvre-chefs."	0	2
9	-- "Mon Dieu, oui ! on pourrait prendre le mien à mon bureau !"	0	2
10	-- "C'est comme moi, je suis employé."	0	2
11	Alors ils se considérèrent.	0	2
12	L'aspect aimable de Bouvard_2 charma de suite Pécuchet_2 .	0	4
13	Ses yeux bleuâtres, toujours entreclos, souriaient dans son visage colore. Un pantalon à grand-pont, qui godait par le bas sur des souliers de castor, moulait son ventre, faisait bouffer sa chemise à la ceinture ; -- et ses cheveux blonds, frisés d'eux-mêmes en boucles légères, lui donnaient quelque chose d'enfantin.	0	4
14	Il poussait du bout des lèvres une espèce de sifflement continu.	0	4
15	L'air sérieux de Pécuchet_3 frappa Bouvard_3 .	0	6
16	On aurait dit qu'il portait une perruque, tant les mèches garnissant son crâne élevé étaient plates et noires. Sa figure semblait tout en profil, à cause du nez qui descendait très bas. Ses jambes prises dans des tuyaux de lasting manquaient de proportion avec la longueur du	0	6

Y. Leclerc, P. Cubaud. *Bouvard : 618, Pécuchet : 598. Etude de statistique textuelle.*
 Revue Flaubert, n°3, 2003 <http://www.univ-rouen.fr/flaubert/>



Y. Leclerc, P. Cubaud. *Bouvard* : 618, *Pécuchet* : 598. *Etude de statistique textuelle*.
 Revue Flaubert, n°3, 2003 <http://www.univ-rouen.fr/flaubert/>

AWK (Aho, Weinberger, Kernighan)

```
BEGIN {n=0;r=0}  
/Bouvard/{n++;printf "B %d",n;  
           if (n>r) {printf "*";r=n}; printf "\n";}   
/Pécuchet/{n--;print "P",n}
```

- Autres exemples (voir les manpages) :

```
length($0) < 200
```

affiche les lignes de longueur < 200 caractères

```
BEGIN { FS = ",[ \t]*| [ \t]+ »; s=0 }  
{ print $2, $1; s += $1 }  
END { print "sum is", s, " average is", s/NR }
```

sépare des champs par espace, tab, ou virgule

inverse les deux premiers

calculs statistiques sur le premier

Practical Extraction Report Language (PERL, L. Wall)

```
#!/usr/bin/perl
```

fabrication de la sequence BP

```
$sum = 0;
```

```
open(F, "/Users/pcubaud/FLAUBERT/bouvard2-mac.txt");
```

```
while (<F>) {
```

```
$ligne = $_;
```

```
chop($ligne);
```

```
@lesmots = split(/ /,$ligne);
```

```
foreach $mot (@lesmots) {
```

```
if ($mot =~ m/Bouvard/) {
```

```
$sum++;
```

```
print "B";
```

```
if ($sum % 40 == 0) {print "\n";}
```

}

```
elseif ($mot =~ m/Pécuchet/) {
```

```
$sum++;
```

```
print "p";
```

```
if ($sum % 40 == 0) {print "\n";}
```

}

}

}

```
close(F);
```

```
print "\n\n\n";
```

[illegible]

```
### comptage par chapitre
```

```
$c = $nb = $np = $tb = $tp = $xtic = 0;
open(F, "/Users/pcubaud/FLAUBERT/bouvard2-mac.txt");
while (<F>) {
    $ligne = $_;
    chop($ligne);
    @lesmots = split(/ /,$ligne);
    foreach $mot (@lesmots) {
        if ($mot =~ m/CHAPITRE/) {
            if ($c > 0) {print "$c : $nb $np ($xtic) \n";}
            $tb += $nb; $tp += $np; $xtic += $nb + $np;
            $nb = $np = 0;
            $c++;
        }
        elsif ($mot =~ m/Bouvard/) {
            $nb++;
        }
        elsif ($mot =~ m/Pécuchet/) {
            $np++;
        }
    }
}
print "$c : $nb $np ($xtic)\n";
    $tb += $nb;
    $tp += $np;
print "TOTAL : $tb $tp \n";
close(F);
print "\n\n\n";
```

Chapitre	nB	nP	paragraphes	mots
I	60	50	145	6000
II	81	60	292	11503
III	74	86	431	12277
IV	50	45	352	9786
V	46	44	266	6227
VI	43	44	360	8309
VII	28	25	135	3251
VIII	94	89	506	13928
IX	73	77	410	11461
X	69	78	387	10531
total	618	598		

Fabrication du HTML

```
#!/usr/bin/perl
$n = $nb = $np = $sum = $para = $c = 0;

open(F,"bouvard2.txt"); open(T,">BPcol.html");
print T "<HTML><BODY bgcolor=White text=Black>\n";
while (<F>) {
    $ligne = $_; chop($ligne);
    if ($ligne =~ m/CHAPITRE/) {
        $c++;
        print "CHAPITRE $c VU\n";
        if ($c >1) {
            print T "</TABLE>\n";
        }
        print T "<H1>Chapitre $c</H1>\n";
        print T "<TABLE CELLPADDING=5 CELLSPACING=0 BORDER=0 >\n";
        print T "<TR VALIGN=TOP>\n";
        print T "<TD BGCOLOR=#A0B8C8>PARA</TD>\n";
        print T "<TD></TD>\n";
        print T "<TD BGCOLOR=Silver>nB-nP</TD>\n";
        print T "<TD BGCOLOR=Gray>nB+nP</TD>\n";
        print T "</TR>\n";
    }
}
```

```

else {
    @lesmots = split(/ /,$ligne);
    foreach $mot (@lesmots) {
        if ($mot =~ m/Bouvard/) {
            $n++;$nb++;
            $mot =~ s#Bouvard#<FONT color=Blue><B>Bouvard_$nb</B></FONT>#;
        }
        elsif ($mot =~ m/Pécuchet/) {
            $n--;$np++;
            $mot =~ s#Pécuchet#<FONT color=Red><B>Pécuchet_$np</B></FONT>#;
        }
    }
    $nligne = join(" ",@lesmots);
    if ($nligne ne "") {
        $sum = $nb + $np;
        $para ++;
        print T "<TR VALIGN=TOP>\n";
        print T "<TD BGCOLOR=#A0B8C8><A name=\" $para\">$para</A></TD>\n";
        print T "<TD>$nligne</TD>\n";
        print T "<TD BGCOLOR=Silver>$n</TD>\n";
        print T "<TD BGCOLOR=Gray>$sum </TD>\n";
        print T "</TR>\n";
    }
}

}

print T "</TABLE>\n";
print T "</BODY></HTML>";
close(T);
close(F);

```


Enchaîner les traitements : l'art du « pipe »

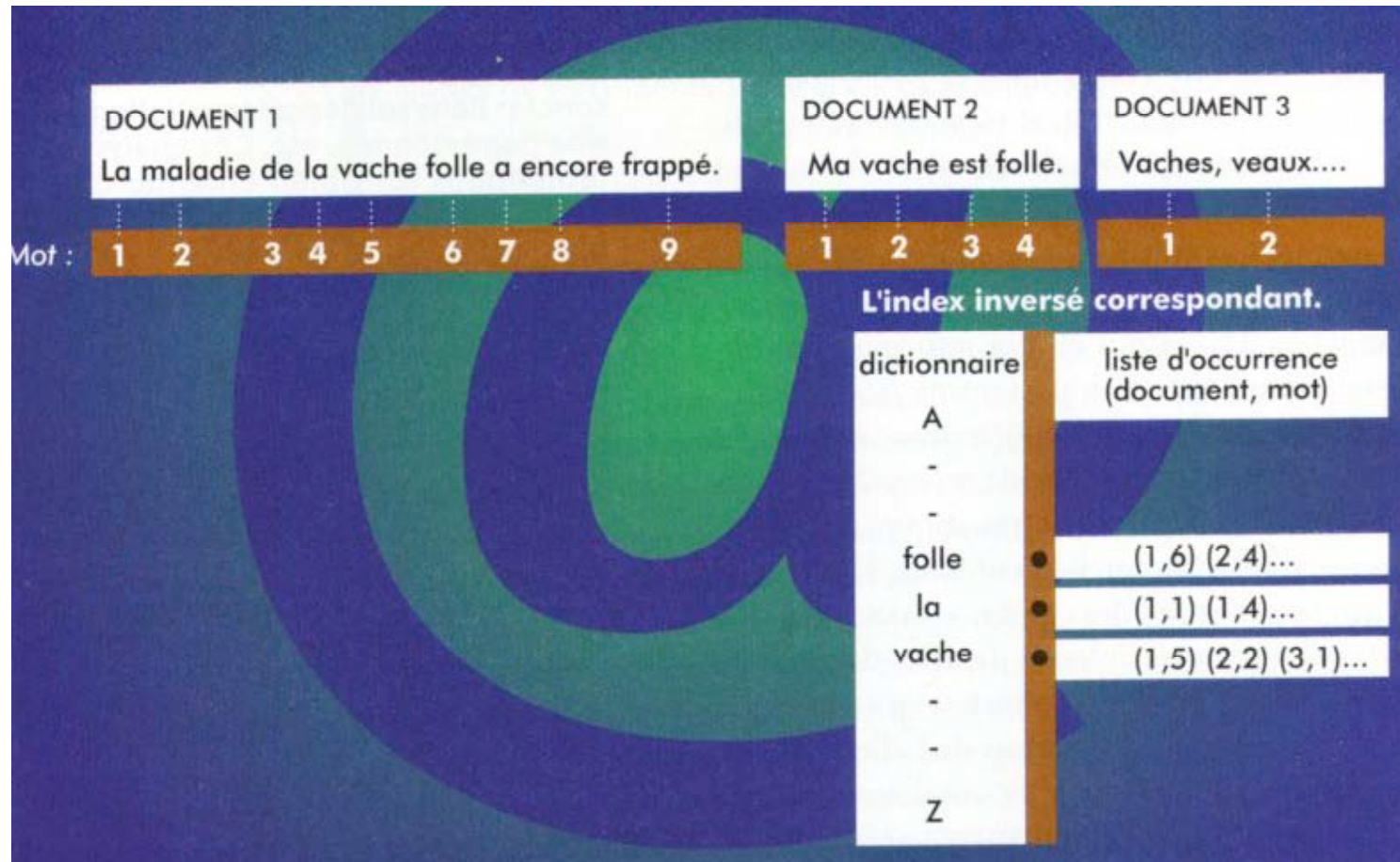
```
07/30/97 15:36:23 DONNE cocuage2 POUR pento.esil.univ-mrs.fr (139.124.4.66) AVEC OmniWeb/2.6-beta-4 OWF/1.0
07/30/97 15:47:22 DONNE jacques1 POUR prithivi (163.173.136.11) AVEC Mozilla/3.01 (X11; I; SunOS 5.5 sun4c)
07/30/97 15:52:04 DONNE commercel POUR dial066.ppp.lrz-muenchen.de (129.187.24.66) AVEC OmniWeb/2.5-beta-6 OWF/1.0
07/30/97 16:07:35 DONNE conscrit2 POUR ns.cci.dk (194.239.27.66) AVEC Mozilla/2.0 (compatible; MSIE 3.0;
    Windows 95) via Harvest Cache version 1.4pl3
07/30/97 16:09:00 DONNE rldased3 POUR rkawv247.kau.roche.com (145.245.98.98) AVEC Mozilla/3.01 (Win16; I)
07/30/97 16:11:04 DONNE DICO/prenoms POUR ifp.ifp.fr (156.118.212.2) AVEC Mozilla/3.01 (X11; I; SunOS 5.5.1 sun4u)
07/30/97 16:13:09 DONNE DICO/pays POUR ifp.ifp.fr (156.118.212.2) AVEC Mozilla/3.01 (X11; I; SunOS 5.5.1 sun4u)
```

...

```
awk '{print $6}' logdon | awk -F. '{print $7,$6,$5,$4,$3}' |
awk '{print $1}' | grep -v '^[0123456789]' | sort | uniq -c | sort -nr
```

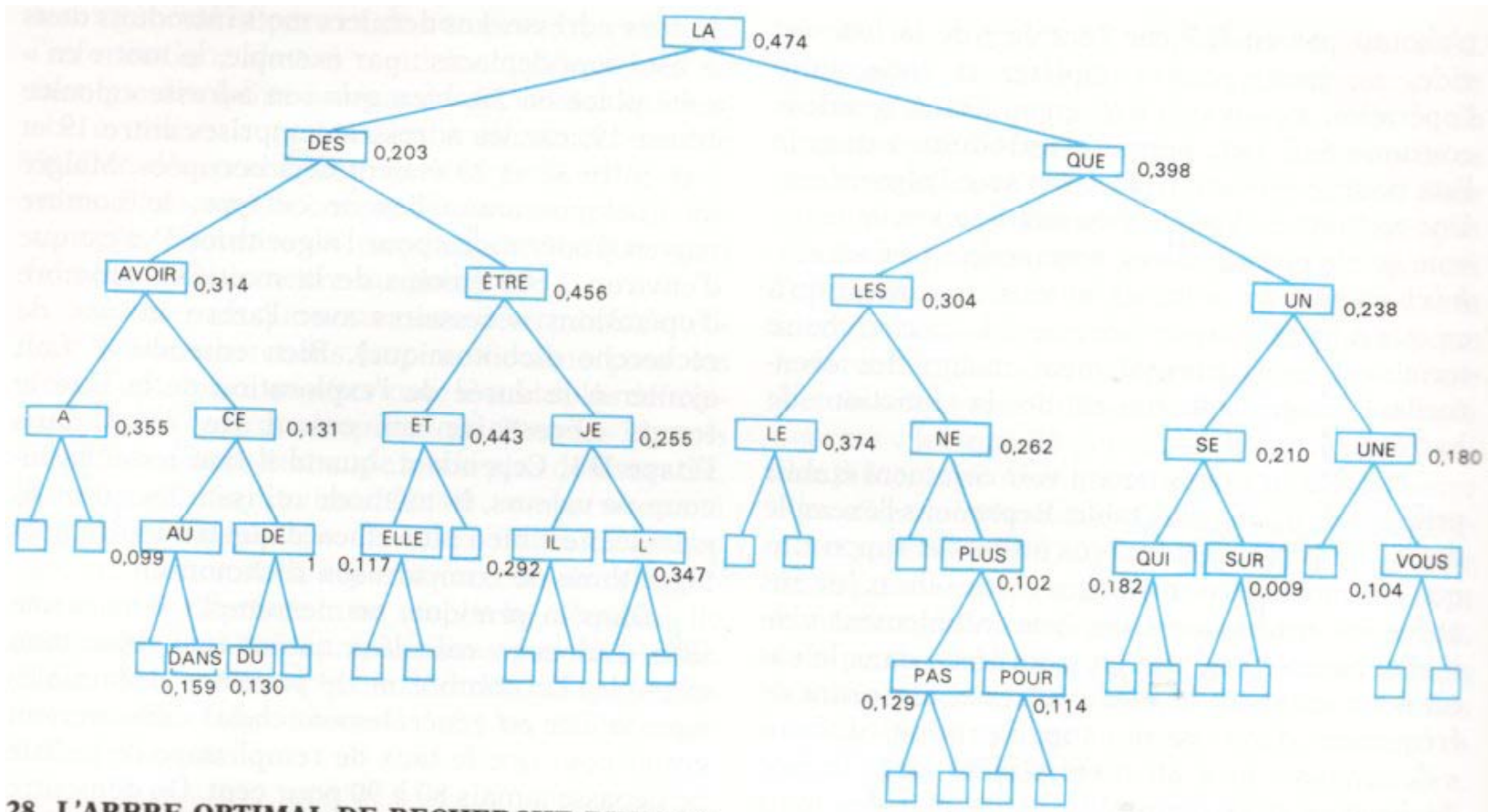
```
209479 fr
63549 com
61941 net
26769 ca
21707 be
16277 ch
13267 it
10855 de
8088 jp
7863 es
6803 edu
3743 uk
3197 br
3085 ru
```

Les index ("inverted files" ≠ index inversé)



F. BOURDONCLE, P. BERTIN. *Recherche d'aiguilles dans une botte de liens*. Numéro spécial "L'avenir du Web" La Recherche n°328. Février 2000. p. 66

Arbres de recherche : tenir compte des fréquences



28. L'ARBRE OPTIMAL DE RECHERCHE

D. KNUTH. *L'élaboration des algorithmes*. in *Le calcul intensif*. Belin, 1989.

Arbre binaire de profondeur optimale

... didn't appear in the index at all! The idea of KWIC indexing is due to
 Luhn, *Amer. Documentation* **11** (1960), 288–295. (See W. W. Youden,
ibid. **10** (1963), 583–646, where the full KWIC index appears.)

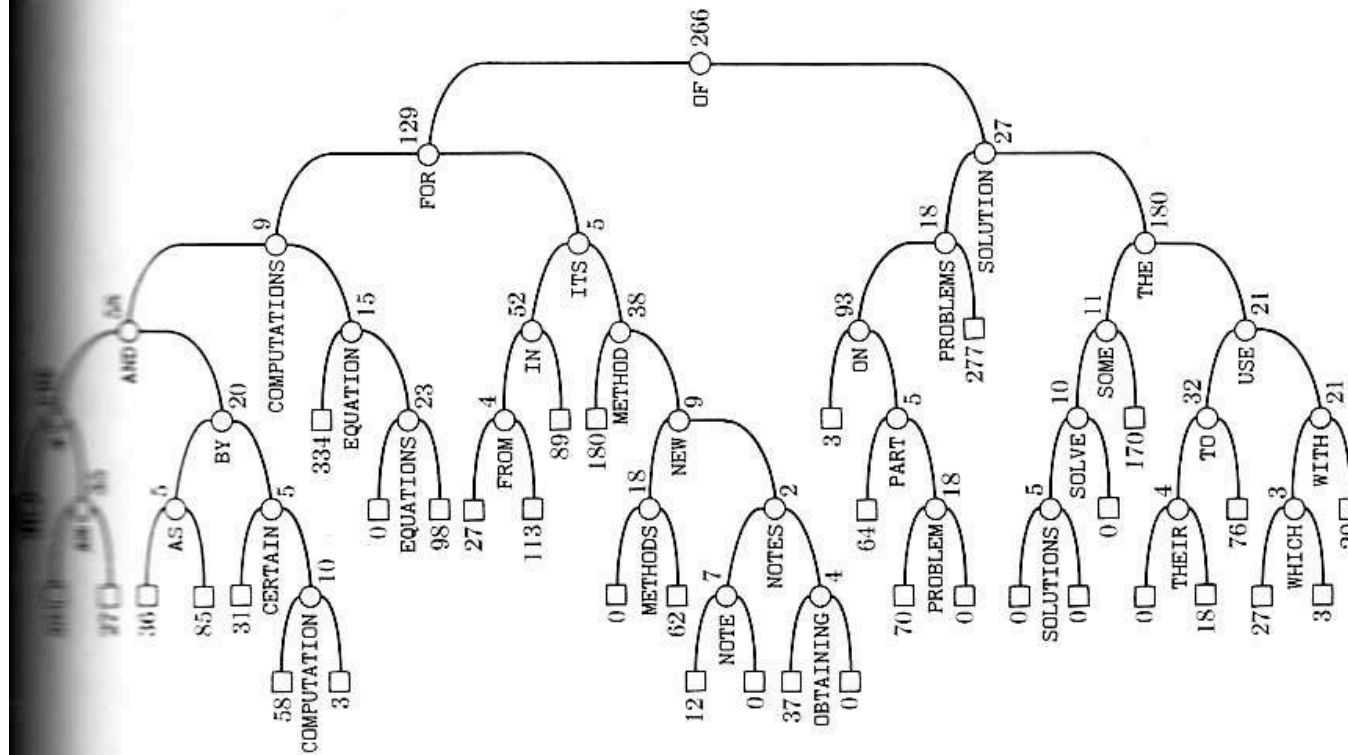


Fig. 15. An optimum binary search tree for a KWIC indexing application.

Indexeurs : exemple de Glimpse <http://webglimpse.net>

Speed and Power

Com Edu

Features included

- ✓ ✓ *** Fast Searching:** Glimpse builds a keyword index in advance for very fast searching (though it can also access individual files for complex boolean queries). Uncommon words will be found rapidly even in a very large fileset, up to several Gigabytes. Common words (with 100's or 1000's of matches) will take longer, but if the number of hits returned can be limited, even those will be very fast. The core index and search programs are written in C.

Glimpse uses a small index - typically less than 5% of the total data size - so it can usually be entirely loaded into memory.

Sample Search Times on a dual 400MHz processor running Linux (run on command line to not include page load time)


dataset size	index size	number of times keyword appears	Search Time
12608 files, 1.5Gb	3.1Mb	4	1 sec
12608 files, 1.5Gb	3.1Mb	606	31 sec to return all hits; 1 sec to return only top 20 hits
499 files, 64Mb	300Kb	49	1 sec

* Edu version will search fast for small numbers of returned hits; but when searching for common words the performance may be degraded.

- ✓ ✓ **Large Data Sets:** Glimpse is used to handle data sets up to 9 Gigabytes, to our knowledge. Because of the two-level search design that localizes keywords to a 'block' of data, the index footprint is quite small, typically less than 5% of the total data set size. The speed of the search scales with the number of matches to the keyword and only secondarily with the total size of the indexed data.

- ✓ **Result Caching/Large # of Hits:** Webglimpse maintains a cache of recent searches, which allows even those searches returning a very large number of hits to return quickly if they are commonly performed. The cache also allows rapid navigation using a **Next Hits** toolbar, which lets the user jump to any page of hits returned.

Un autre : mg (I. Witten et al.)

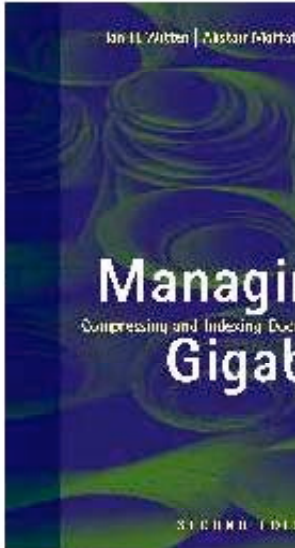

Adresse :  <http://www.cs.mu.oz.au/mg/>

Managing Gigabytes

Compressing and Indexing Documents and Images

Second Edition, 1999

The second edition of *Managing Gigabytes: Compressing and Indexing Documents and Images* by [Ian H. Witten](#), [Alistair Moffat](#), and [Timothy C. Bell](#), is now available (May 1999), published by [Morgan Kaufmann Publishing](#), San Francisco, ISBN 1-55860-570-3.



Information available:

- [Publicity brochure for the second edition](#) (PDF file, 15kB)
- [Table of contents of the second edition](#)
- [Preface to the second edition](#)
- [Errata listing for the second edition](#)
- [Buy the second edition from Amazon.com](#)
(Yes, we will receive a small cut from Amazon if you use this link and make a purchase).
- [Morgan Kaufmann catalog entry](#)
- The current version of the software, [mg-1.2.1](#), dated August 1999

Concordances (KWIC : keyword in context)

Tableau 2.7
Concordance de la forme *enfants* dans le corpus
des réponses à la question *Enfant*

L 304	c' est pas drole d' avoir des enfants a l' heure actuelle a
L 420	la peur de l' avenir pour les enfants a quelle place peuvent
L2415	onomique* risques d' avoir des enfants anormaux* avenir incer
L2281	s raisons medicales(risque d' enfants anormaux)* manque de c
L1760	est pas la peine de mettre des enfants au monde dans un clima
L2307	c' est bien beau de mettre des enfants au monde mais si c' es
L1879	chomage, on ne peut mettre les enfants au monde sans leur ass
L 429	affection entre le mari et les enfants* aucune garantie d' av
L 710	as* les gens n' aiment pas les enfants* aucune raison n' est
L1461	l' age, l' argent, la peur des enfants aussi, le confort a de
L1295	ou d' argent* ne pas aimer les enfants, avoir peur d' en avoi
L2165	ur, elles sont tributaires des enfants c' est astreignant* le
L 498	and on n' en a pas assez, deux enfants c' est le maximum* les
L1009	il y en a qui n' aime pas les enfants, c' est le travail, la
L 846	desirent pas, un mariage sans enfants, c' est pas un mariage
L2013	ve au chomage avec beaucoup d' enfants c' est plus difficile*
L1903	la vie et le chomage pour les enfants* c' est trop dur d' el
L1647	er une meilleure education aux enfants, ca coute cher actuell
L 285	st difficile de s' occuper des enfants* ca depend de la menta
L1720	roblemes pour faire garder les enfants, ca permet au couple d
L2170	ieres, on choisit d' avoir des enfants, choix delibere de la
L 68	financieres, ne pas aimer les enfants* chomage* difficultes
L1839	probleme financier, avenir des enfants (chomage)* le cout de
L1196	rs, peur de l' avenir pour les enfants* crise actuelle peur d
L 921	fants, difficile d' elever des enfants de nos jours car les e
L 441	l' avenir du chomage pour les enfants* des femmes ne sont pa
L 921	REP* peur du chomage pour ses enfants, difficile d' elever d
L 289	les, materielles, la garde des enfants difficile due au trava
L2072	P* la peur du chomage pour ses enfants* difficulte d' assumer
L 966	eulent plus se priver pour les enfants* economique, pas de de
L 744	couples qui n' aiment pas les enfants* egoisme* de peur d' a
L 612	mage* la femme ne veut plus d' enfants, elle veut sa liberte,
L1650	structures d' accueil pour les enfants en bas age dans le cas
L1467	actuelle n' est fait pour les enfants en bas age les infrast
L 166	nconvenients presentes par des enfants en bas age(trop absor

"Le Concordeur" (MacOS)

http://omega.crm.umontreal.ca/~rand/CC_fr.html

Lemmatisation ("stemming")

Difficile ...

est : le point cardinal ou le verbe ?

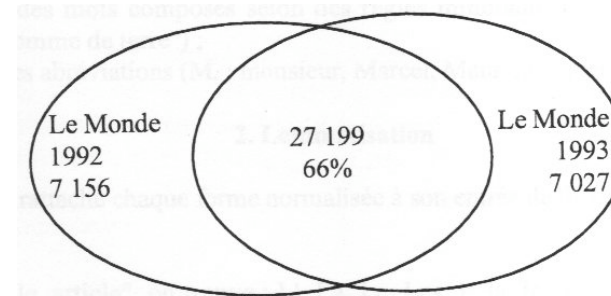
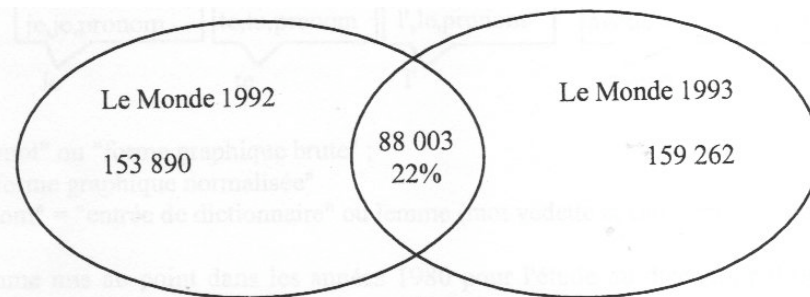
été : la saison ou le verbe ?

être : l'être humain ou le verbe ?

fût : le tonneau ou le verbe ?

plus d'un tiers des mots sont "homographes" dans un texte français
(une graphie et ≥ 2 entrées dans le dictionnaire)

Mais utile :



D. Labbé. *La lemmatisation des grandes bases de textes.*

Colloque édition électronique en littérature et dictionnaire. Univ. Rouen, juin 2002.

A ESSAYER : algorithme de Martin PORTER pour l'anglais (1980)

<http://www.tartarus.org/~martin/PorterStemmer/index.html>

Removing suffixes by automatic means is an operation which is especially useful in the field of information retrieval. In a typical IR environment, one has a collection of documents, each described by the words in the document title and possibly by words in the document abstract. Ignoring the issue of precisely where the words originate, we can say that a document is represented by a vector of words, or \terms\. Terms with a common stem will usually have similar meanings, for example:

CONNECT
CONNECTED
CONNECTING
CONNECTION
CONNECTIONS

remov suffix by automat mean is an oper which is espec
us in the field of inform retriev. in a typic ir environ,
on ha a collect of document, each describ by the word in the
document titl and possibl by word in the document abstract. ignor the
issu of precis where the word origin, we can sai that a document is
repres by a vetor of word, or \term\. term with a common stem will
usual have similar mean, for exampl:

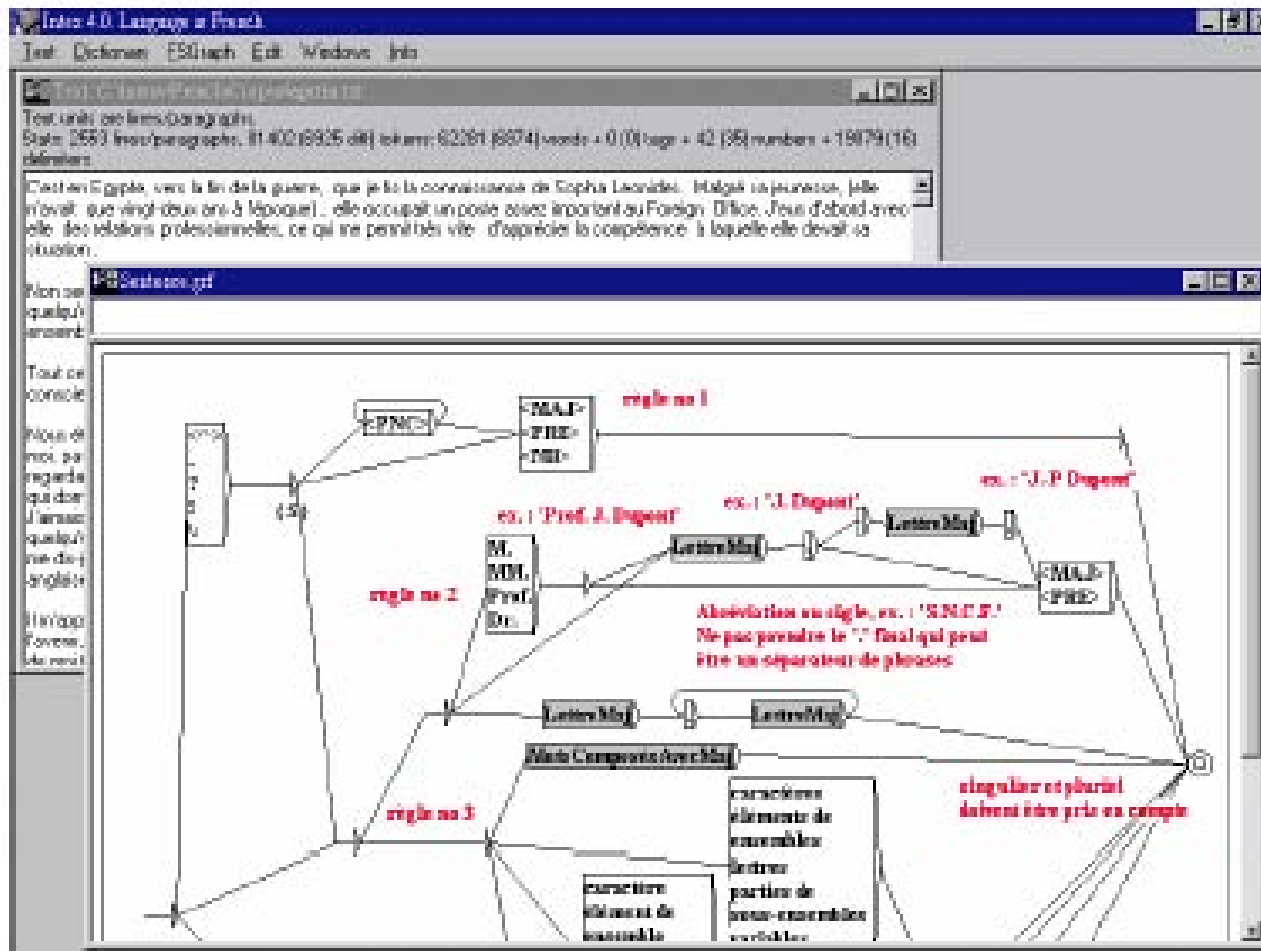
connect
connect
connect
connect
connect

A voir du même :
<http://snowball.tartarus.org/>

Segmentation (des mots, des phrases)

INTEX - Max Silberztein (>1993)

<http://laseldi.univ-fcomte.fr/intex/>



Le classement (« ranking »)

- dépasser les réponses « oui/non » sur requêtes booléennes
- fournir une mesure de similarité entre les documents et la requête

Ex. : le *cosinus directeur* entre vecteurs requête R et document D

$$\cos(R, D) = \frac{R \cdot D}{|R||D|} = \frac{1}{W_D W_R} \sum_{t \in R} w_{D,t} \bullet w_{R,t}$$

$$W_D = \sqrt{\sum_{t \in R} w_{D,t}^2} \text{ avec } w_{D,t} = f_{d,t} \left(\log \frac{N}{f_t} \right)$$

$$W_R = \sqrt{\sum_{t \in R} w_{R,t}^2} \text{ avec } w_{R,t} \text{ arbitraire}$$

N : nbre de documents

f_{d,t} : occurrences du terme t ds le doc d

f_t : nombre de documents contenant le terme t

Le classement interactif (« relevance feedback »)

L'utilisateur retient parmi l'ensemble des documents en réponse ceux qui l'intéressent le plus (S) et en demande d'autres du même type

On exécute une séquence de requêtes $\{R_i\}$ et on veut faire tendre cette séquence vers une requête optimale.

Ex. d'heuristique :

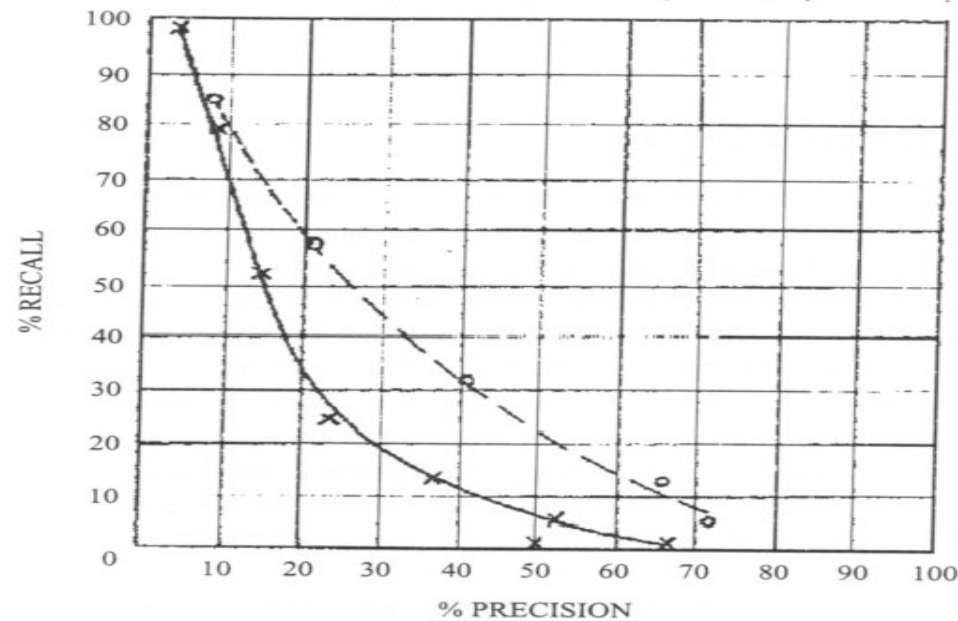
$$R_{i+1} = R_i + \sum_{d \in S} D_d - D_n$$

avec D_n : document de plus grand poids non retenu

Le rapport « recall / precision »

Recall : Capacité d'un système de recherche à fournir le plus de documents possibles

Precision : Capacité d'un système de recherche à sélectionner uniquement des documents pertinents (« relevants ») - et éliminer les autres



M. LESK *Digital libraries* (...) p. 169

ex : 50 doc. pertinents pour une requête. Le système en signale 40 dont 20 pertinents
 $\text{recall} = 20 / 50 = 40\%$
 $\text{precision} = 20 / 40 = 50\%$

3. La présentation

A visiter : musée de l'imprimerie de Lyon

<http://www.imprimerie.lyon.fr/>

Sauvez le patrimoine !!!

<http://www.garamonpatrimoine.org/petition.html>

La typographie : vocabulaire

Meuble pour le typographe : deux *casses*

Capitales et Bas de casse

Ex. français

A	B	C	D	E	F	G
H	I	K	L	M	N	O
P	Q	R	S	T	V	X
À	É	Î	Ô	Ù	Y	Z
É	È	Ê		W		
À	È	Î	Ô	Ù	()	*
U	J	j				

1	2	3	4	5	6	7	8
s			f	g	h	9	0
o	p	q		fi	k	1/2	cad.
a	r	.	.				Cadral

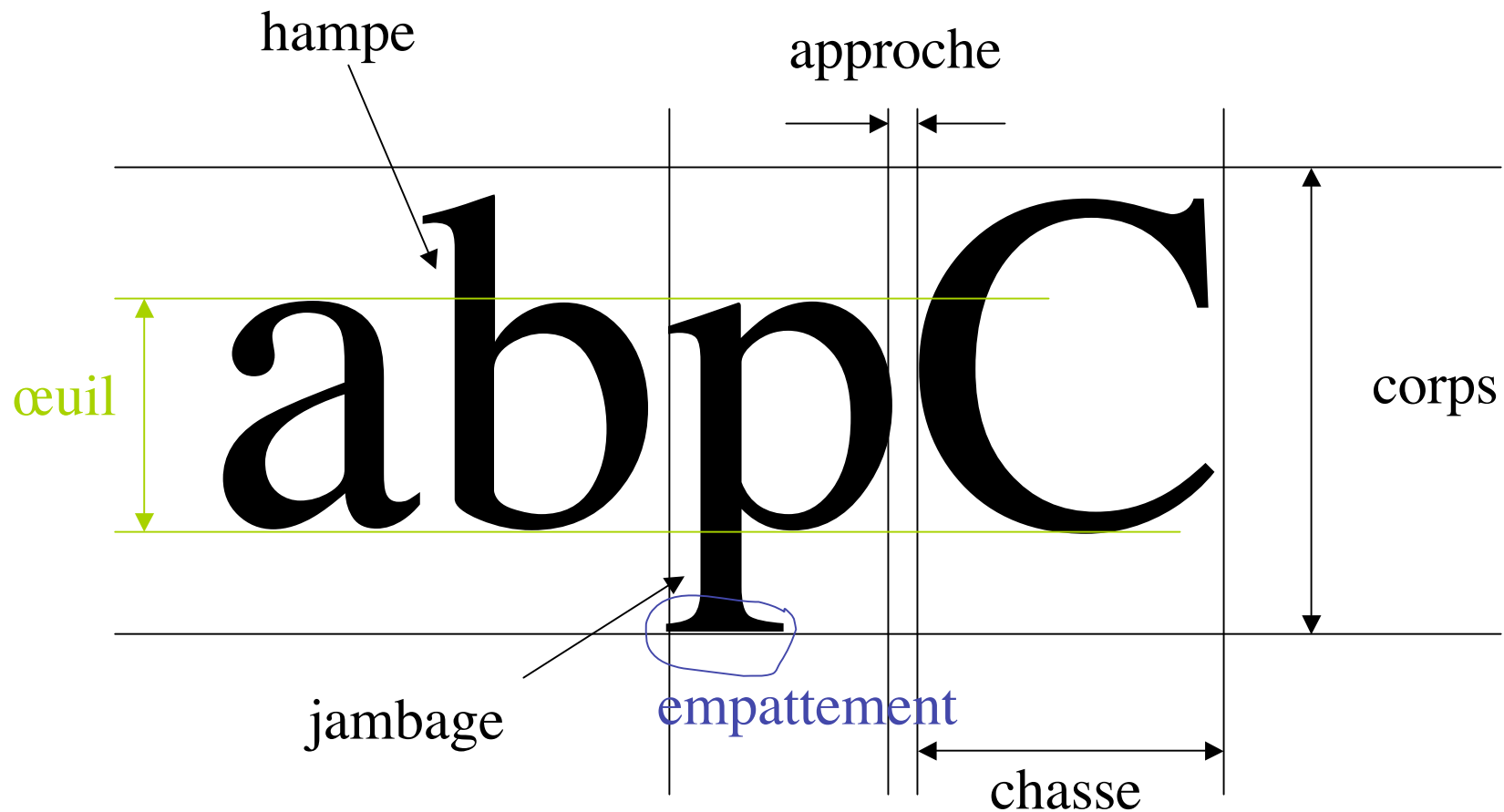
<http://www.textesrares.com/>

Ex. anglo-saxon

A	B	C	D	E	F	G
H	I	K	L	M	N	O
P	Q	R	S	T	V	W
X	Y	Z	Æ	Œ	U	J
ä	ë	ï	ö	ü	£	
1	2	3	4	5	6	7
8	9	0	1/4	1/2	3/4	k

Thin Sp.)	?	!	;	...	fi
i	s	f	g			ff
o	y	p	,	w		fi
a	r	q	:			Large Quads

<http://www.collectionscanada.ca/presses/>



Unité de mesure : le point typographique

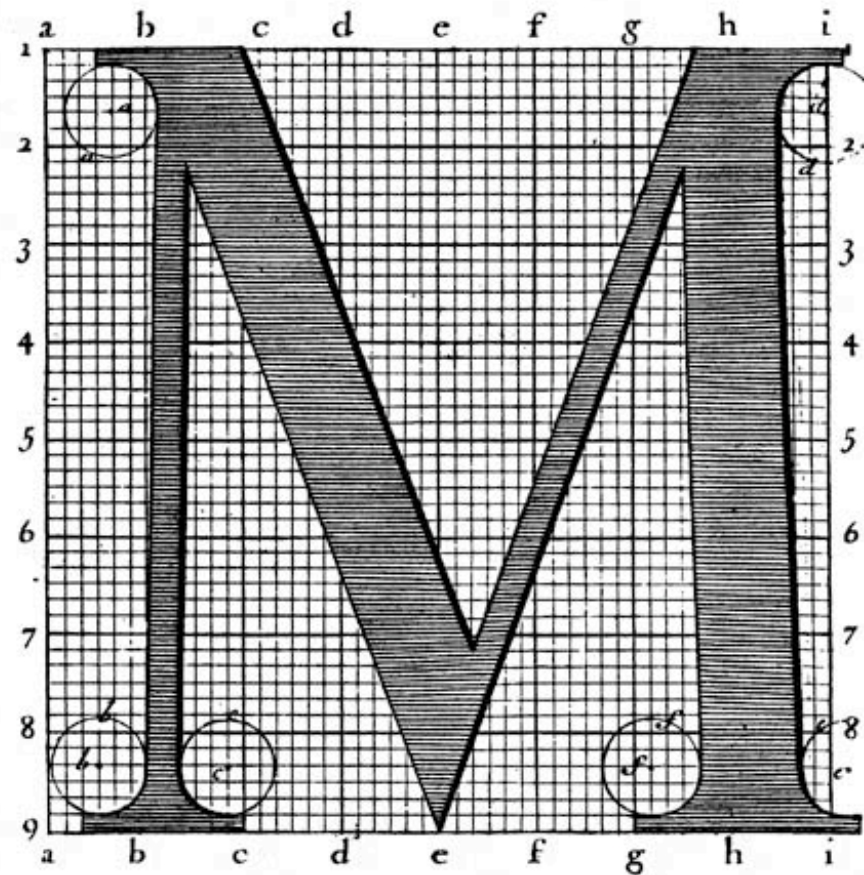
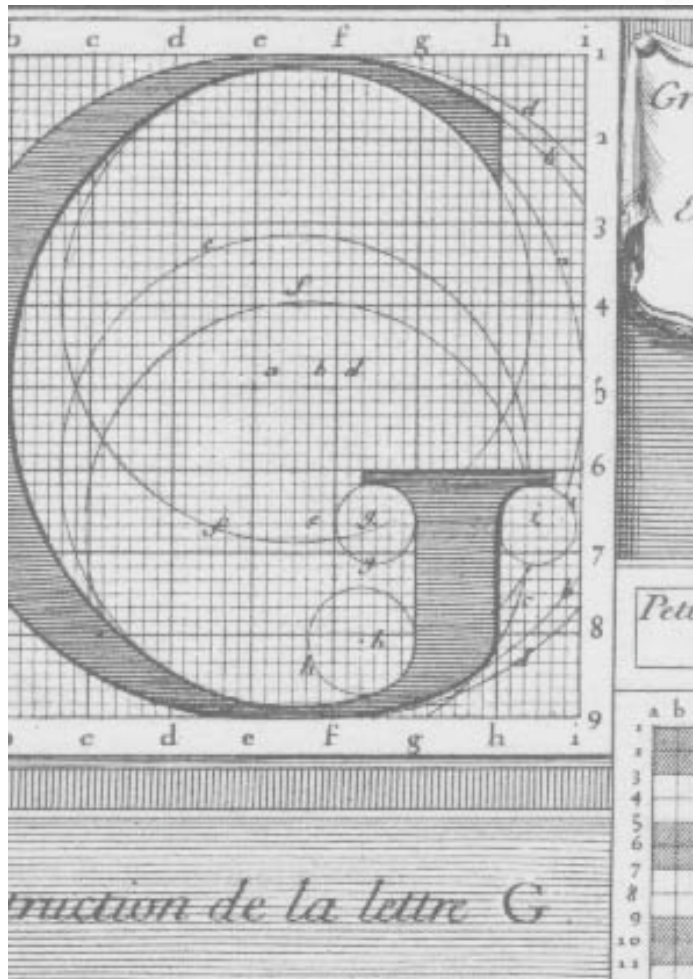
Point Didot = 0,3579mm

Point pica = 1/72 pouce = 0,35135mm

Point métrique = 0,4 mm

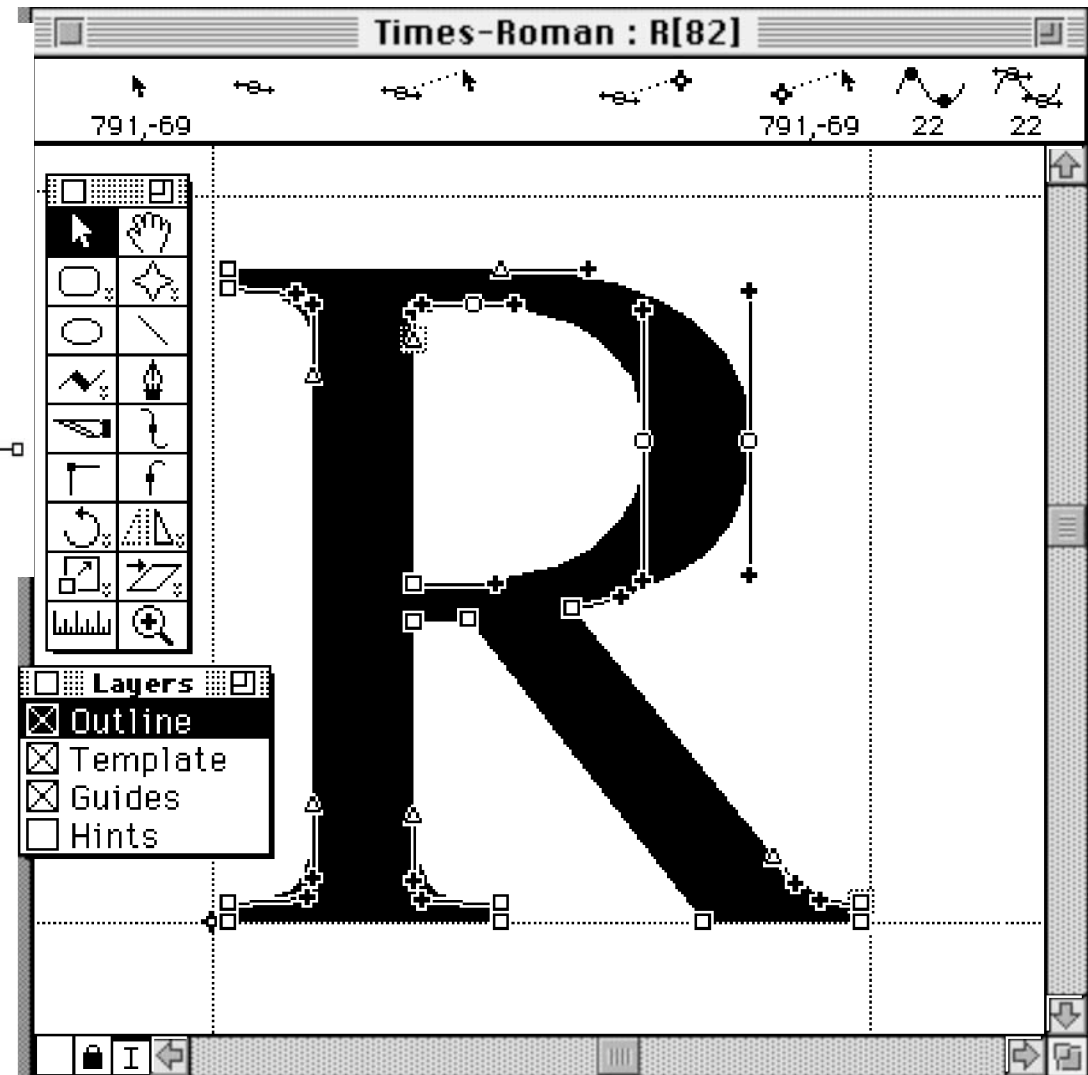
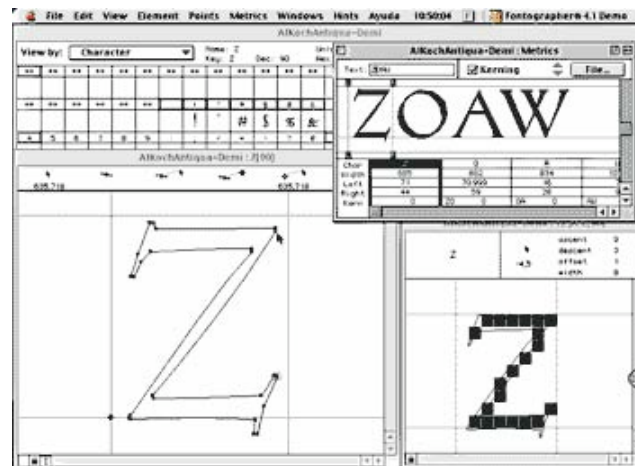
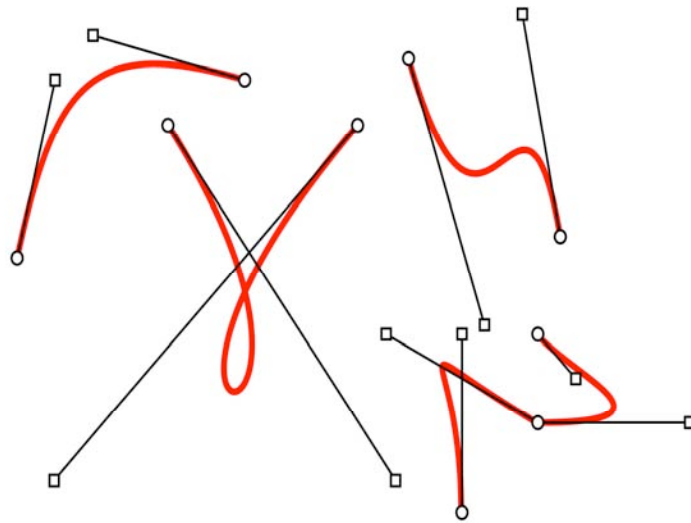
Le dessin des caractères : ex. du "romain du roi" (Truchet et Grandjean, 1702)

Grille à 2300 dpi !



<http://www.irisa.fr/faqtypo/Welcome.html>

Le dessin par ordinateur (CAO) : courbes paramétrées



Fontographer

Exemple de famille de caractères (parmi des milliers)



<http://www.porchez.com/>

-sans (serif) : bâton

Helvetica

-à empattement

Times

-cursives

Lucida

La graisse:

maigre

Gras

ultralight

Le langage PostScript

```
/Times-Roman findfont  
15 scalefont setfont  
100 500 moveto  
(I love PostScript!) show  
showpage
```

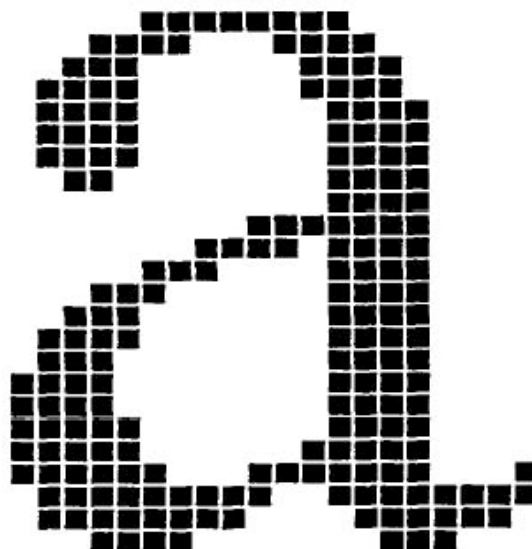
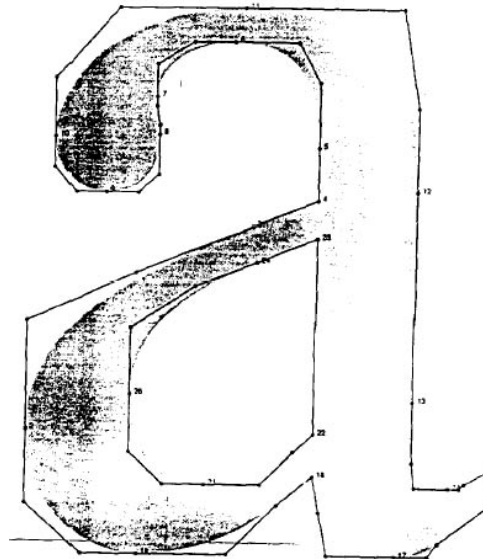
Les 20 ans de PostScript en 2004 :

<http://www.adobe.com/products/postscript/main.html>

Manuel de réf. en ligne :

<http://www.adobe.com/products/postscript/pdfs/PLRM.pdf>

« rasterisation »



```
% Char 97
/a {userdict begin 153 -12 moveto
95 -12 37 39 37 115 curveto
37 227 146 277 264 325 curveto
330 352 lineto
330 407 lineto
330 474 308 516 243 516 curveto
202 516 165 493 165 459 curveto
165 449 168 430 168 420 curveto
168 378 147 360 114 360 curveto
83 360 61 385 61 417 curveto
61 478 126 551 253 551 curveto
415 551 431 449 431 362 curveto
431 147 lineto
431 85 434 59 468 59 curveto
479 59 484 64 512 80 curveto
512 43 lineto
512 43 512 43 512 43 curveto
458 2 448 -12 415 -12 curveto
346 -12 337 32 330 69 curveto
294 39 244 -12 153 -12 curveto
closepath
143 153 moveto
143 94 178 60 220 60 curveto
277 60 310 95 330 113 curveto
330 313 lineto
269 288 lineto
207 263 143 221 143 153 curveto
closepath fill
end}def
```

J. André, J. Bur
Métrie des fontes PostScript
 Cahier GUTenberg n°8 mars 1991

Le traitement de texte, la PAO : Exemple de TeX

gives the Fibonacci sequence: 1, 1, 2, 3, 5, 8, 13, (The study of *difference equations* concerns the problem of going from recursive definitions to algebraic definitions. The Fibonacci sequence is give by the algebraic definition

$$f(x) = \frac{\sqrt{5}}{5} \left(\frac{1 + \sqrt{5}}{2} \right)^{x+1} - \frac{\sqrt{5}}{5} \left(\frac{1 - \sqrt{5}}{2} \right)^{x+1} .)$$

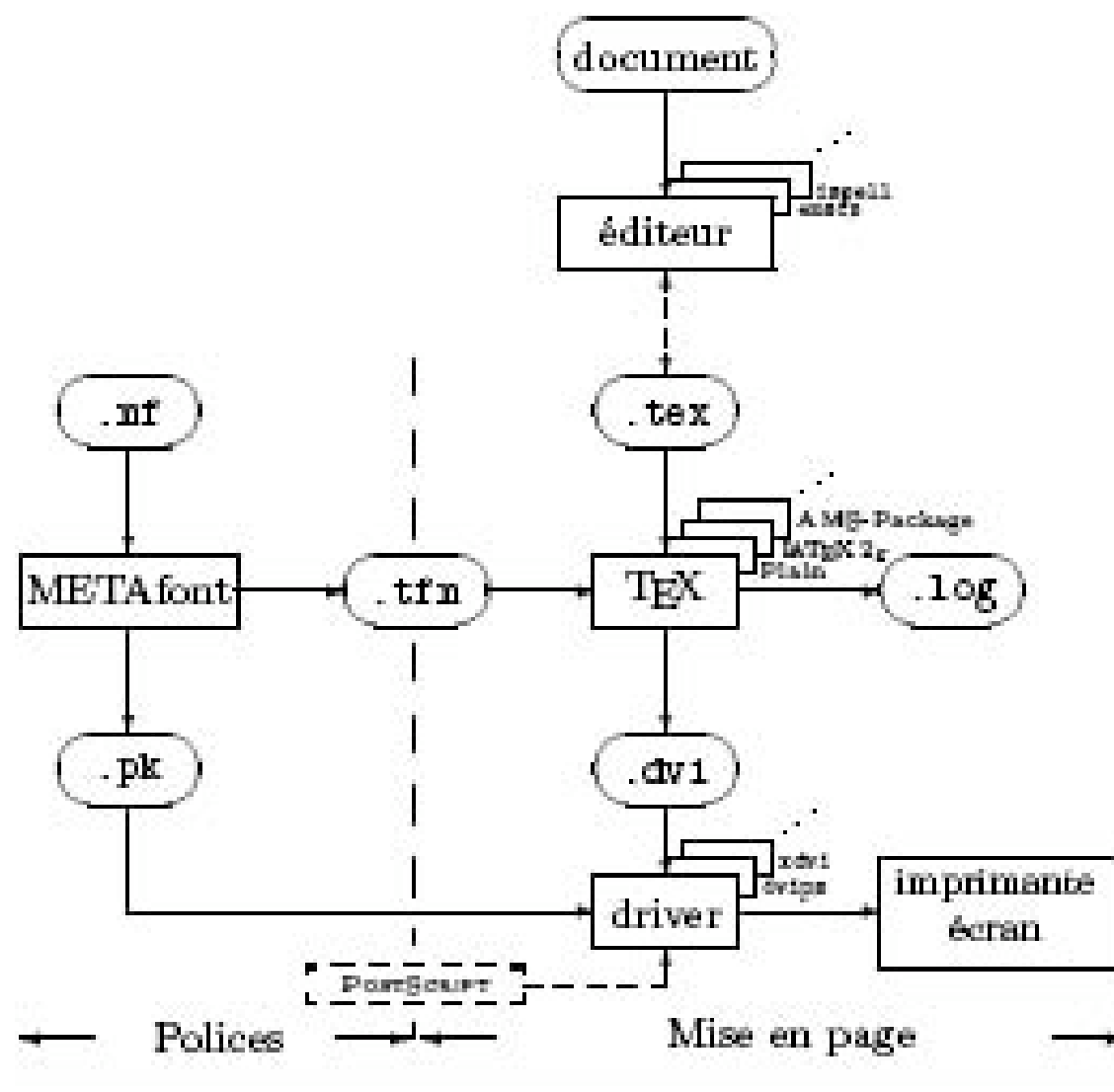
The *primitive recursive functions* are an example of a broad and interesting class of functions that cam be obtained by such a formal characterization.

Definition The class of *primitive recursive functions* is the smallest class \mathcal{C} (i.e., intersection of all classes \mathcal{C}) of functions such that

- i. All *constant functions*, $\lambda x_1 x_2 \cdots x_k [m]$ are in \mathcal{C} , $1 \leq k$, $0 \leq m$;

<http://www.ctan.org/>

A VOIR : <http://www.tug.org/texshowcase/>



LaTeX : le bon côté

```
\documentstyle[twocolumn]{article}
\begin{document}
\title{\Large\bf Performance Evaluation of non-Markovian Stochastic Event Graphs}
\author{Pierre H. Cubaud \\\
\\
CEDRIC - Conservatoire National des Arts et M\`etiers \\\
292 Rue Saint-Martin, F-75003 Paris \\\
{\tt cubaud@cnam.fr}}

\maketitle

\subsection*{\centering Abstract}
{\em
We describe a low polynomial complexity algorithm that provides bounds and approximation of
any firing date using properties of integral orderings among random variables.
We test the algorithm on some examples and discuss its use for the computation of the network cycle time.
}

\section{Introduction}
An event graph with  $T < \infty$  transitions is then entirely defined by its  $T \times T$ 
adjacency matrix  $M$ , where  $M_{ij}$  is the initial marking of the place connecting
transitions  $\theta_i$  and  $\theta_j$ .
If no such place exists, we shall write by convention  $M_{ij} = -\infty$ .
```

Et le mauvais...

```
\long\def\@makecaption#1#2#3{
\ifx\figurestring#3
  \vskip 5pt
  \setbox\@tempboxa\hbox{\small #1.~ #2}
  \ifdim \wd\@tempboxa >\hsize
    \setbox\@tempboxa\hbox{\small #1.~ }
    \setlength\captionindent{\wd\@tempboxa} \divide\captionindent by 2
    \parbox[t]{\hsize}{\small \hangindent \captionindent \hangafter=1%
    \unhbox\@tempboxa#2}
  \else \hbox to\hsize{\small\hfil\box\@tempboxa\hfil}
  \fi
\else
  \begin{center}{\sf #1}:{\sf #2}\end{center}
  \vskip 8pt
\fi}
```

Lire à l'écran

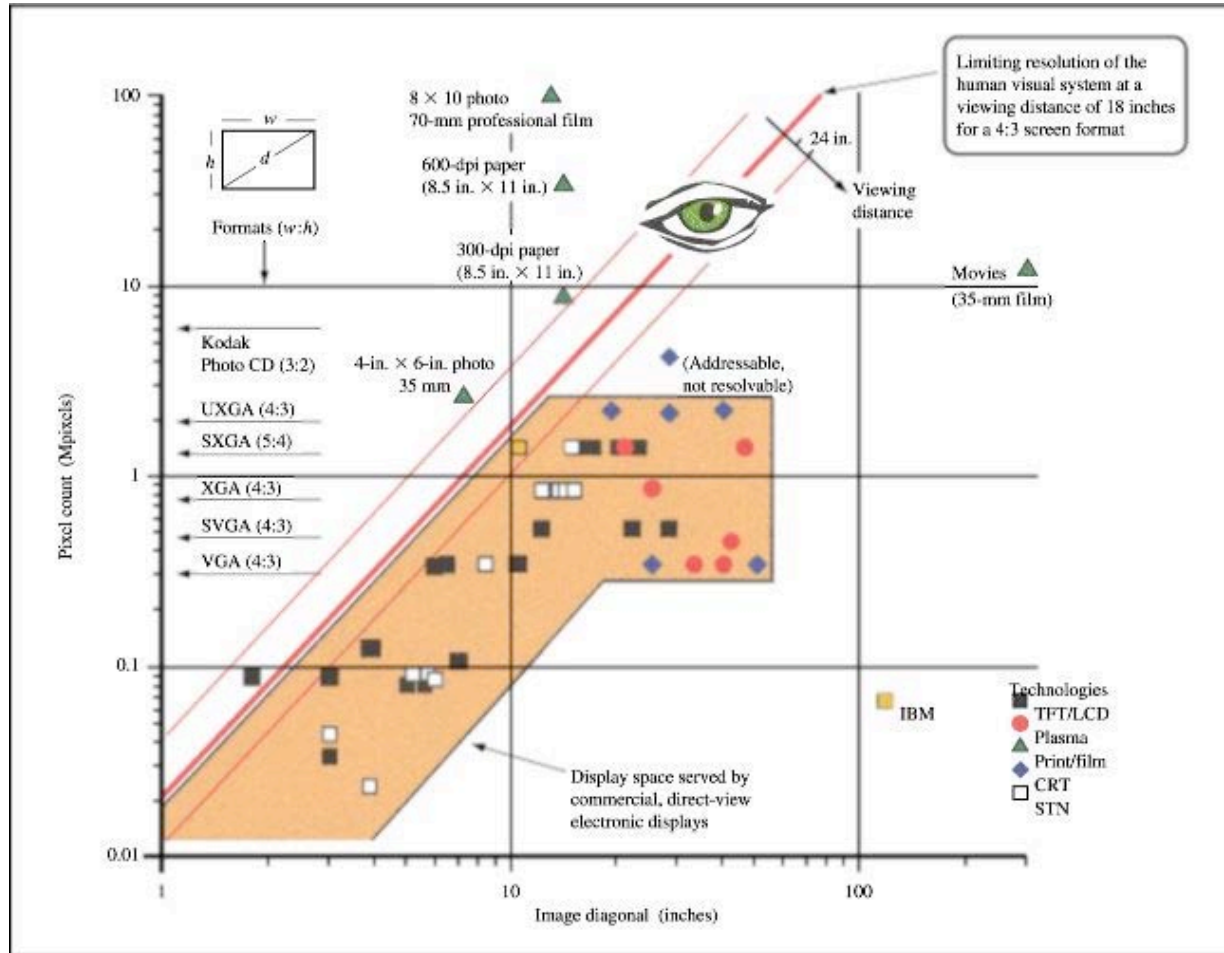


Figure 1

A view of display space showing existing electronic, paper, and film displays.

P. M. Alt, K. Noda. *Increasing electronic display information content: An introduction*. IBM J. of Res. & Dev. Volume 42, Numbers 3/4, 1998.