ABS Tech Note: AS01 AppleSearch Character Support (8/93)

Revised: 5/21/96
Security: Everyone

ABS Tech Note: AS01 AppleSearch Character Support (8/93)

==========

Article Created: 31 August 1993

TOPIC --

As with most information retrieval applications, AppleSearch deals with a variety of language and character rules in order to function properly. This note will outline the AppleSearch character support for version 1.0.Note that our search engine is licensed from Personal Library Software.

DISCUSSION --

Summary

For version 1.0 of AppleSearch, we are supporting the US English Character Set, with some mapping exceptions as defined below. Technically, we are 7-bit compatible, but 8-bit aware.

We've defined our character support based on the Inside Mac Volume VI definition of the Roman Character set; Appendix E: The Standard Roman Character Set.

The server software "handles" accented and double characters in a limited fashion. Accented characters are mapped to their US English counterpart, for example, é to e, å to a. AppleSearch makes no distinction between accented and non-accented characters. This mapping is consistent between indexing and searching. For example, while indexing the word "élève" would be changed to "eleve". When you search for the word "élève", it would be changed to "eleve", before searching the index.

Double characters are mapped to the first letter, for example, œ to o, æ to a. This may seem odd, but our options were to:

- Store the character
- Map it to a space
- Map it to a 7-bit equivalent

We have chosen to map characters with accent marks and double characters to their 7-bit equivalent for the following reasons:

- Staying 7-bit clean today, will ensure compatibility between future 8-bit releases and version 1.0 index files.

- Users may get false hits when using languages where diacritical marks change the meaning of a word. However, for languages where the grammatical rules for diacritical marks are not very rigid, it is better to compare words after stripping diacritical marks. You may get false hits, but you'll not miss any documents.

Note

Download this document, open with your favorite word processor, and change the font to either Courier, Palatino, or Times to read all the characters below.

AppleSearch v1.0 Character Mapping

0x30	0	0	0x66	f	F	0x95	ï	I
0x31	1	1	0x67	g	G	0x96	ñ	N
0x32	2	2	0x68	h	H	0x97	ó	O
0x33	3	3	0x69	i	I	0x98	ò	O
0x34	4	4	0x6a	j	J	0x99	ô	O
0x35	5	5	0x6b	k	K	0x9a	ö	O
0x36	6	6	0x6c	l	L	0x9b	õ	O
0x37	7	7	0x6d	m	M	0x9c	ú	U
0x38	8	8	0x6e	n	N	0x9d	ù	U
0x39	9	9	0x6f	o	O	0x9e	û	U
			0x70	p	P	0x9f	ü	U
0x41	A	A	0x71	q	Q	0xa7	ß	S
0x42	B	B	0x72	r	R	0xae	Æ	A
0x43	C	C	0x73	s	S	0xaf	Ø	O
0x44	D	D	0x74	t	T	0xbe	æ	A
0x45	E	E	0x75	u	U	0xbf	ø	O
0x46	F	F	0x76	v	V	0xcb	À	A
0x47	G	G	0x77	w	W	0xcc	Ã	A
0x48	H	H	0x78	x	X	0xcd	Õ	O
0x49	I	I	0x79	y	Y	0xce	Œ	O
0x4a	J	J	0x7a	z	Z	0xcf	œ	O
0x4b	K	K				0xd8	ÿ	Y
0x4c	L	L	0x80	Ä	A	0xd9	Ÿ	Y
0x4d	M	M	0x81	Å	A	0xde	fi	F
0x4e	N	N	0x82	Ç	C	0xdf	fl	F
0x4f	O	O	0x83	É	E	0xe5	Â	A
0x50	P	P	0x84	Ñ	N	0xe6	Ê	E
0x51	Q	Q	0x85	Ö	O	0xe7	Á	A
0x52	R	R	0x86	Ü	U	0xe8	Ë	E
0x53	S	S	0x87	á	A	0xe9	È	E
0x54	T	T	0x88	à	A	0xea	Í	I
0x55	U	U	0x89	â	A	0xeb	Î	I

0x56	V	V	0x8a	ä	A	0xec	Ï	I
0x57	W	W	0x8b	ã	A	0xed	Ì	I
0x58	X	X	0x8c	å	A	0xee	Ó	O
0x59	Y	Y	0x8d	ç	C	0xef	Ô	O
0x5a	Z	Z	0x8e	é	E	0xf1	Ò	O
			0x8f	è	E	0xf2	Ú	U
0x61	a	A	0x90	ê	E	0xf3	Û	U
0x62	b	B	0x91	ë	E	0xf4	Ù	U
0x63	c	C	0x92	í	I	0xf5	ı	I
0x64	d	D	0x93	ì	I			
0x65	e	E	0x94	î	I			

Tech Info Library Article Number:13125