Windows code page

From Wikipedia, the free encyclopedia

Windows code pages are sets of characters or code pages (known as character encodings in other operating systems) used in Microsoft Windows from the 1980s and 1990s. Windows code pages were gradually superseded when Unicode was implemented in Windows, although they are still supported both within Windows and other platforms.

There are two groups of code pages in Windows systems: OEM and ANSI code pages. Code pages in both of these groups are extended ASCII code pages.

Contents

ANSI code page

ANSI code pages (officially called "Windows code pages"[1] after Microsoft accepted the former term being a misnomer[2]) are used for native non-Unicode (say, byte oriented) applications using a graphical user interface on Windows systems. ANSI Windows code pages, and especially the code page 1252, were called that way since they were purportedly based on drafts submitted or intended for ANSI. However, ANSI and ISO have not standardized any of these code pages. Instead they are either supersets of the standard sets such as those of ISO 8859 and the various national standards (like Windows-1252 vs. ISO-8859-1), major modifications of these (making them incompatible to various degrees, like Windows-1250 vs. ISO-8859-2) or having no parallel encoding (like Windows-1257 vs. ISO-8859-4; ISO-8859-13 was introduced much later).[2] About twelve of the typography and business characters from CP1252 at code points 0x80–0x9F (in ISO 8859 occupied by C1 control codes, which are useless in Windows) are present in many other ANSI/Windows code pages at the same codes. These code pages are labelled by Internet Assigned Numbers Authority (IANA) as "Windows-*number*".[3]

OEM code page

The **OEM code pages** (original equipment manufacturer) are used by Win32 console applications, and by virtual DOS, and can be considered a holdover from DOS and the original IBM PC architecture. A separate suite of code pages was implemented not only due to compatibility, but also because the fonts of VGA (and descendant) hardware suggest encoding of line drawing characters to be compatible with code page 437. Most OEM code pages share many code points, particularly for non-letter characters, with the second (non-ASCII) half of CP437.

A typical OEM code page, in its second half, does not resemble any ANSI/Windows code page even roughly. Nevertheless, two single-byte, fixed-width code pages (874 for Thai and 1258 for Vietnamese) and four multibyte CJK code pages (932, 936, 949, 950) are used as both OEM and ANSI code pages. Code page 1258 uses combining diacritics, as Vietnamese requires more than 128 letter-diacritic combinations. This is in contrast to VISCII, which replaces some of the C0 (i.e. ASCII) control codes.

History

Initially, computer systems and system programming languages did not make a distinction between characters and bytes. This led to much confusion subsequently. Microsoft software and systems previous to the Windows NT line are examples of this, using the OEM and ANSI code pages, which do not make the distinction.

Since the late 1990s, software and systems are increasingly adopting more direct encodings of Unicode, in particular UTF-8 and UTF-16; this trend has been improved by the widespread adoption of XML, which provides a more adequate mechanism for labelling the encoding used.[4] Recent Microsoft products and application program interfaces use Unicode internally, but many applications and APIs continue to use the default encoding of the computer's *locale* when reading and writing text data to files or standard output. Therefore, though Unicode is the accepted standard, there is still backwards compatibility with the older Windows code pages.

The euro sign was added relatively recently to ANSI and OEM code pages (1998 in the case of Code page 858) and therefore obsolete versions of Windows are unable to use it with code pages.

List

> *This list is incomplete; you can help by expanding it (https://en.wikipedia.org/w/index.php?title=Windows_code_page&action=edit).*

The following Windows code pages exist:

ID	Names	Description	Type	Base	Encoding	Standard	Support DOS-based Windows	Support Windows NT family	Support Windows CE family	Comments
37	CP037, IBM037	IBM EBCDIC US-Canada	Other	EBCDIC derivation	8-bit SBCS	IBM CP037[5]	?	Yes		
437	CP437, IBM437	IBM PC US	OEM	ASCII derivation	8-bit SBCS	IBM CP437[6]	1.00-4.90	Yes		
1250	CP1250, Windows-1250	Latin 2 / Central European	ANSI	ASCII derivation	8-bit SBCS	Microsoft CP1250[7][8]	?	Yes		
1251	CP1251, Windows-1251	Cyrillic	ANSI	ASCII derivation	8-bit SBCS	Microsoft CP1251[9][10]	?	Yes		
1252	CP1252, Windows-1252	Latin 1 / Western European	ANSI	ASCII derivation	8-bit SBCS	Microsoft CP1252[11][12]	?	Yes		letter repertoire similar to CP850
1253	CP1253, Windows-1253	Greek	ANSI	ASCII derivation	8-bit SBCS	Microsoft CP1253[13][14]	?	Yes		
1254	CP1254, Windows-1254	Turkish	ANSI	ASCII derivation	8-bit SBCS	Microsoft CP1254[15][16]	?	Yes		
1255	CP1255, Windows-1255	Hebrew	ANSI	ASCII derivation	8-bit SBCS	Microsoft CP1255[17][18]	?	Yes		
1256	CP1256, Windows-1256	Arabic	ANSI	ASCII derivation	8-bit SBCS	Microsoft CP1256[19][20]	?	Yes		
1257	CP1257, Windows-1257	Baltic	ANSI	ASCII derivation	8-bit SBCS	Microsoft CP1257[21][22]	?	Yes		
1258	CP1258, Windows-1258	Vietnamese	OEM+ANSI	?	8-bit SBCS	Microsoft CP1258[23][24]	?	Yes		

- 500
- 708
- 720
- 737
- 775
- 850
- 852
- 855
- 857
- 858
- 860
- 861
- 862
- 863
- 864
- 865
- 866 - cp866
- 869 - IBM869
- 870 - IBM870
- 874 - Thai
- 875 - cp875
- 932 - Japanese
- 936 - Chinese (simplified) (PRC, Singapore)

- 949 - Korean
- 950 - Chinese (traditional) (Taiwan, Hong Kong)
- 1026 - EBCDIC Turkish
- 1047 - IBM01047
- 1140 - IBM01141
- 1141 - IBM01141
- 1142 - IBM01142
- 1143 - IBM01143
- 1144 - IBM01144
- 1145 - IBM01145
- 1146 - IBM01146
- 1147 - IBM01147
- 1148 - IBM01148
- 1149 - IBM01149
- 1200 - Unicode (BMP of ISO 10646, UTF-16LE)
- 1201 - Unicode (BMP of ISO 10646, UTF-16BE). Available only to managed applications [25]
- 1361 - Korean (KS C 5601-1992)
- 10000 - Apple Macintosh Roman
- 10001 - Apple Macintosh Japanese
- 10002 - Apple Macintosh Chinese (traditional) (BIG-5)
- 10003 - Apple Macintosh Korean
- 10004 - Apple Macintosh Arabic
- 10005 - Apple Macintosh Hebrew
- 10006 - Apple Macintosh Greek
- 10007 - Apple Macintosh Cyrillic
- 10008 - Apple Macintosh Chinese (simplified) (GB 2312)
- 10010 - Apple Macintosh Romanian
- 10017 - Apple Macintosh Ukrainian
- 10021 - Apple Macintosh Thai
- 10029 - Apple Macintosh Roman II / Central Europe
- 10079 - Apple Macintosh Icelandic
- 10081 - Apple Macintosh Turkish
- 10082 - Apple Macintosh Croatian
- 12000 - utf-32
- 12001 - utf-32 Big endian
- 20000 - x-Chinese-CNS
- 20001 - x-cp20001
- 20002 - x-x-Chinese-Eten
- 20003 - x-cp20003
- 20004 - x-cp20004
- 20005 - x-cp20005
- 20105 - IA5 IRV (DIN 66003)
- 20106 - IA6 (German) (DIN 66003)
- 20107 - IA6 (Swedish) (SEN 850200 B)
- 20108 - IA6 (Norwegian) (NS 4551-1)
- 20127 - US-ASCII (7-bit with no character larger than 127)
- 20261 - T.61 (T.61-8bit)
- 20269 - ISO-6937
- 20273 - EBCDIC Germany
- 20277 - EBCDIC Denmark/Norway
- 20278 - EBCDIC Finland/Sweden
- 20280 - EBCDIC Italy
- 20284 - EBCDIC Latin America/Spain
- 20285 - EBCDIC United Kingdom
- 20290 - EBCDIC Japanese

- 20297 - EBCDIC France
- 20420 - EBCDIC Arabic
- 20423 - EBCDIC Greek
- 20424 - x-EBCDIC-KoreanExtended
- 20833 - Korean
- 20838 - EBCDIC Thai
- 20866 - Russian - KOI8-R
- 20871 - EBCDIC Icelandic
- 20880 - EBCDIC Cyrillic
- 20905 - EBCDIC Turkish
- 20924 - IBM00924
- 20932 - EUC-JP
- 20936 - x-cp20936
- 20949 - x-cp20949
- 21025 - EBCDIC Cyrillic
- 21027 - Japanese
- 21866 - Ukrainian - KOI8-RU
- 28591 - ISO-8859-1
- 28592 - ISO-8859-2
- 28593 - ISO-8859-3
- 28594 - ISO-8859-4
- 28595 - ISO-8859-5
- 28596 - ISO-8859-6
- 28597 - ISO-8859-7
- 28598 - ISO-8859-8
- 28599 - ISO-8859-9
- 28600 - ISO-8859-10
- 28601 - ISO-8859-11
- (28602 - ISO-8859-12)
- 28603 - ISO-8859-13
- 28604 - ISO-8859-14
- 28605 - ISO-8859-15
- 28606 - ISO-8859-16
- 38596 - ISO-8859-6
- 38598 - ISO-8859-8
- 65000 - Unicode (BMP of ISO 10646, UTF-7)
- 65001 - Unicode (BMP of ISO 10646, UTF-8)

Problems arising from the use of code pages

Microsoft strongly recommends using Unicode in modern applications, but many applications or data files still depend on the legacy code pages.

- Programs need to know what code page to use in order to display the contents of files correctly. If a program uses the wrong code page it may show text as mojibake.
- The code page in use may differ between machines, so files created on one machine may be unreadable on another.
- Data is often improperly tagged with the code page, or not tagged at all, making determination of the correct code page to read the data difficult.
- These Microsoft code pages differ to various degrees from some of the standards and other vendors' implementations. This isn't a Microsoft issue *per se*, as it happens to all vendors, but the lack of consistency makes interoperability with other systems unreliable in some cases.
- The use of code pages limits the set of characters that may be used.
- Characters expressed in an unsupported code page may be converted to question marks (?) or other replacement characters, or to a simpler version (such as removing accents from a letter). In either case, the original character may be lost.

See also

- AppLocale — a utility to run non-Unicode (code page-based) applications in a locale of the user's choice.

References

1. Code Pages (http://msdn.microsoft.com/en-us/goglobal/bb964653.aspx), MSDN
2. MSDN: Glossary of Terms (http://msdn.microsoft.com/en-us/goglobal/bb964658.aspx#a)
3. IANA list of Character Sets (http://www.iana.org/assignments/character-sets)
4. http://www.w3.org/TR/xml11/#charencoding
5. IBM. "SBCS code page information document - CPGID 00037" (http://www-01.ibm.com/software/globalization/cp/cp00037.html). Retrieved 2014-07-04.
6. IBM. "SBCS code page information document - CPGID 00437" (http://www-01.ibm.com/software/globalization/cp/cp00437.html). Retrieved 2014-07-04.
7. Microsoft. "Windows 1250" (http://msdn.microsoft.com/en-us/goglobal/cc305143). Retrieved 2014-07-06.
8. IBM. "SBCS code page information document CPGID 01250" (http://www-01.ibm.com/software/globalization/cp/cp01250.html). Retrieved 2014-07-06.
9. Microsoft. "Windows 1251" (http://msdn.microsoft.com/en-us/goglobal/cc305144). Retrieved 2014-07-06.
10. IBM. "SBCS code page information document CPGID 01251" (http://www-01.ibm.com/software/globalization/cp/cp01251.html). Retrieved 2014-07-06.
11. Microsoft. "Windows 1252" (http://msdn.microsoft.com/en-us/goglobal/cc305145). Retrieved 2014-07-06.
12. IBM. "SBCS code page information document CPGID 01252" (http://www-01.ibm.com/software/globalization/cp/cp01252.html). Retrieved 2014-07-06.
13. Microsoft. "Windows 1253" (http://msdn.microsoft.com/en-us/goglobal/cc305146). Retrieved 2014-07-06.
14. IBM. "SBCS code page information document CPGID 01253" (http://www-01.ibm.com/software/globalization/cp/cp01253.html). Retrieved 2014-07-06.
15. Microsoft. "Windows 1254" (http://msdn.microsoft.com/en-us/goglobal/cc305147). Retrieved 2014-07-06.
16. IBM. "SBCS code page information document CPGID 01254" (http://www-01.ibm.com/software/globalization/cp/cp01254.html). Retrieved 2014-07-06.
17. Microsoft. "Windows 1255" (http://msdn.microsoft.com/en-us/goglobal/cc305148). Retrieved 2014-07-06.
18. IBM. "SBCS code page information document CPGID 01255" (http://www-01.ibm.com/software/globalization/cp/cp01255.html). Retrieved 2014-07-06.
19. Microsoft. "Windows 1256" (http://msdn.microsoft.com/en-us/goglobal/cc305149). Retrieved 2014-07-06.
20. IBM. "SBCS code page information document CPGID 01256" (http://www-01.ibm.com/software/globalization/cp/cp01256.html). Retrieved 2014-07-06.
21. Microsoft. "Windows 1257" (http://msdn.microsoft.com/en-us/goglobal/cc305150). Retrieved 2014-07-06.
22. IBM. "SBCS code page information document CPGID 01257" (http://www-01.ibm.com/software/globalization/cp/cp01257.html). Retrieved 2014-07-06.
23. Microsoft. "Windows 1258" (http://msdn.microsoft.com/en-us/goglobal/cc305151). Retrieved 2014-07-06.
24. IBM. "SBCS code page information document CPGID 01258" (http://www-01.ibm.com/software/globalization/cp/cp01258.html). Retrieved 2014-07-06.
25. Code page identifier list [1] (http://msdn.microsoft.com/en-us/library/dd317756(VS.85).aspx)

External links

- Code page information from Microsoft (http://msdn.microsoft.com/goglobal/bb964653).
- Blog about Microsoft code pages (http://blogs.msdn.com/shawnste/pages/code-pages-unicode-encodings.aspx).
- Windows Code Page reference chart (http://msdn.microsoft.com/goglobal/bb964654)
- IANA Charset Name Registrations (http://www.iana.org/assignments/charset-reg)
- Unicode mapping table for Windows code pages (http://www.unicode.org/Public/MAPPINGS/VENDORS/MICSFT/WINDOWS)
- Unicode mappings of windows code pages with "best fit" (http://www.unicode.org/Public/MAPPINGS/VENDORS/MICSFT/WindowsBestFit)

Retrieved from "http://en.wikipedia.org/w/index.php?title=Windows_code_page&oldid=633651289"

Categories: Windows code pages

- This page was last modified on 13 November 2014, at 10:34.

- Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.