The ASCII and "ANSI" Character Sets

Douglas A. Kerr, P.E.

Issue 2
July 18, 2004

ABSTRACT

In connection with personal computer technology we often hear reference to the "ASCII" and "ANSI" character sets. This paper describes what these are and explains the basis of the acronyms used to identify them.

The ASCII character set

ASCII refers to the *American National Standard Code for Information Interchange*, a standardized 7-bit coded character set for general use in representing character-based information. The 128 code positions in this set are assigned to 95 *graphic characters* (letters, numbers, SPACE, and symbols) and 33 *control characters*, which give coded indications and commands pertaining to document formatting, information transmission, and other such areas.

This character set was the prototype of an international standard character set which we can identify (using a "short name") as ISO-646. Initially, that character set had some very minor differences from ASCII, but those were resolved in 1991 for the "baseline" version (The International Reference Version –IRV). There are indeed "national variants" of the character set (legitimatized by the ISO-646 document) which cater to the special alphabetic needs of various countries.

Its identifier

The ASCII character set was originally developed by committee X3 of the American Standards Association (ASA), the organization now known as the American National Standards Institute (ANSI). The first form of the character set was approved in 1963 under the title *American Standard Code for Information Interchange*, and it soon became common to speak of it by the acronym of its title, "ASCII". This first form, incidentally, included only a single "case" of the alphabetic characters (considered to be the "upper case").

The next edition of the standard, approved in 1967, included both upper- and lower-case instances of the alphabetic characters, and was essentially the ASCII we know today.

By this time work on the 1967 version neared completion, the name of the organization had changed to *The USA Standards Institute*, and so the title of the new edition of the standard would become *The USA Standard Code for Information Interchange*. This led to the prospect that a new acronym, "USASCII", would likely be adopted.

Concerned that further changes in the name of the sponsoring organization would probably occur, leading to further confusing changes in the identifier, the principal editor[1] of the 1967 standards document included a clause establishing a permanent short designation for the standard, independent of future changes in the title of the document: "ASCII". It had been granted a title in perpetuity.

The current designation of the ASCII standards document is:

ANSI INCITS 4-1986 (R2002)

ANSI identifies the "publishing" organization, the *American National Standards Institute*. INCITS identifies the developing organization, the newly-formed *InterNational Committee for Information Technology Standards*, an autonomous body operating under ANSI rules and accredited by ANSI as a standards developing organization (SDO). It is the successor to ANSI internal committee X3 (formerly USASI X3, earlier yet ASA X3), under which ASCII had been developed and sustained. The "4" indicates that it was the fourth standard created by that body (as ASA X3, in 1963), the "1986" indicates the year of approval of the current version, and the "(R2002)" means that it was reaffirmed, without change, in 2002.

Happy birthday, ASCII

Let me interrupt this discussion to wish happy birthday to ASCII, which reached the big 4-0 on June 17, 2003, and to recognize the insightful work of all those who led to its development, a process which not only faced formidable technical challenges but also the opposition (often covert) of many organizations with vested interests they saw as being foiled by the introduction of an industry-wide standard coded character set.

The "ANSI" character set

The character set generally characterized as the "ANSI" character set is an 8-bit character set, a superset of ASCII, most accurately

---

[1] Coincidentally, the author of this paper

described as "Windows Code Page 1252". Despite its common "moniker", it is in fact not specified by an ANSI standard, not in fact by **any** national nor international industry standard.

Its first 128 code positions (the "lower half") are identical to the 128 positions of ASCII. The additional 128 code positions contain further graphic symbols.

This character set, originally devised by Microsoft, is the normal character set used in modern Windows systems operating in an environment in which the language of choice is represented in the "Western European Latin" alphabet (as distinguished, for example, from the Cyrillic alphabet, the Greek alphabet, the Katakana syllabary, and so forth, and in fact from other Latin alphabets).

The "ISO-8859-1" character set

The International Organization for Standardization (ISO), in concert with the International Electrotechnical Commission (IEC), standardized the *8-Bit Single Byte Coded Graphic Character Sets - Part 1: Latin Alphabet No. 1*, ISO/IEC 8859-1. (Note that 8859-2, etc., cover character sets for other alphabets.)

This is based on the character set of Windows Code Page 1252, but with a modest difference. In ISO-8859-1 (as I will call it for convenience), the 32 character positions in the "upper half" of the code table with codes 0x00 through 0x9F, which mirror the locations in the "lower half" devoted to control characters, are kept clear (so as to accommodate the possibility that interpretation of control characters might be done on only a 7-bit basis), whereas in Code Page 1252, 24 useful graphic characters are placed in that area. These include the symbols opening and closing single and double quotes, baseline single and double quotes, the en dash and em dash, single and double dagger, and so forth.

ANSI makes available in the US many standards of the ISO, which become, by endorsement, a special type of ANSI standards. The ISO-8859-1 character set standard, as such an "embraced" ANSI standard, was designated:

ANSI ISO/IEC 8859-1-1998

Since standards in this area are now in the purview of INCITS, the current formal designation of the 8-bit character set, as published by ANSI, is:

INCITS/ISO/IEC 8859-1-1998

Note the loss of the "ANSI" and the difference between the current designation of this standard and the current designation of the ASCII standard. There, the "family name" still includes the ANSI prefix, since this is in fact a standard developed by an ANSI-accredited organization (whereas ISO-8859-1 is a "godchild").

How' s that again?

Now, let' s get this straight. The character set known as ASCII:

- Was developed by ANSI (albeit under an earlier name).
- Is a full-fledged native ANSI standard.
- Is the only character set entitled to the short name "ASCII", with which it was invested in perpetuity (through the foresight of the undersigned).

The character set widely known as "ANSI" (Windows Code Page 1252):

- Is not defined by any ANSI standard (including any ISO/IEC standard embraced by ANSI).
- Is similar, but not identical, to a character set:
  - Which was developed by ISO (with IEC).
  - Which was approved by ISO and IEC.
  - Which is an ANSI standard only by "embrace".
  - Which has recently lost the "ANSI" part of is designation as a result of its stewardship being taken on by INCITS.
  - Is only one of numerous character sets covered by ANSI or ANSI-embraced standards (including ASCII and several others in the "ISO-8859- " series).

So why is "ANSI" a good short name for that character set? It isn' t at all – it' s a dumb name. Where did that name come from? I don' t know – I sure didn' t do it. Maybe somebody who works for Bill Gates didn' t know a lot of things.

A much better name for this character set is its real name: "Windows Code Page 1252" (or just "CP1252" for short).

A good short name for the related industry standard character set (with due respect to IEC and INCITS) is "ISO-8859-1".

#