The Unicode Standard Version 5.0
Edited by Julie D. Allen, Joe Becker, Richard Cook, Mark Davis, Michael Everson, Asmus Freytag, John H. Jenkins, Mike Ksar, Rick McGowan, Lisa Moore, Eric Muller, Markus Scherer, Michel Suignard, and Ken Whistler.

Addison-Wesley
October, 2006.
1417 pages + 49 front pages + CD.
ISBN-13: 978-0-321-48091-0 (ISBN-10: 0-321-48091-0)

The Unicode Consortium was formed late in the 1980's, to create a universal and unified character set and encoding. At about the same time, the International Organization for Standards (ISO), and its International Electrotechnical Commission (IEC) started work on a similarly targeted project, which became the ISO/IEC 10464 standard. At this point, the Unicode Consortium has sponsored the publication of all versions of the Unicode Standard, and represents the only conforming implementation of ISO/IEC 10646.

The Unicode Standard Version 5.0 comes with so many endorsements from the computerati famous that this publication of the Standard appears to be an all-time first, rather than a version "5"—those listed endorsing the book are: Shai Agassi, Sir Tim Berners-Lee, Brian E. Carpenter, Vint Cerf, Bill Gates, James Gosling, Tytti Granqvist, Kazuhiro Kazama, Brendan Kehoe, Donald E. Knuth, Tim Marsland, Steve Mills, Makoto Murata, James O'Donnell, Larry Page, Hugh McGregor-Ross, Bertrand Serlet, Richard Mark Soley, Joe Spolsky, Guy L. Steele Jr., Kazunori Ukigawa, Guido van Rossum, John Wells—and there is an invitation to see more at the Unicode web-site (http://www.unicode.org). They cannot all be wrong!

The intervening fifteen years have seen remarkable strides in coming to both an intellectual understanding of the problems of a universal character set, a compilation of rules for encoding it, and an exposure of some of the underlying questions of the ways human language is used for communication. It is, in fact, quite astounding that an effort that started out to encode a character set embodying all writing systems in current use has became the current enterprise to encode a character set containing all writing systems that have ever been used (or shall be used). It says something about the commonality of human writing systems that encoding symbols into a linear array of codes seems to work for all of them. Or maybe those for which it does not work have simply been omitted, but this really does not appear to be the case.

The newly published Version 5.0 of the standard appears to retreat a little from the sheer size of Version 4.0 (which, at 1462 pages, exceeded Version 3.0 by 422 pages). However, appearances are deceiving. Although Version 5.0 has fewer pages, and is in an octavo size (10 × 7.5 inches) rather than the previous American letter size (11 × 8.5 inches), the typeface and character tables are more compact. The result is a more manageable and more easily read book that contains considerably more data.

The real body of the book, the code charts, is now a chapter later, and occupies the middle third of the book (460 pages, *c.f.*, 776 in Version 4.0), describing 99,024 characters (*c.f.*, 96,248), attesting to the more compact printing. The new characters come from N'Ko, New Tai Lue, Buginese, Glagolitic, Coptic, Tifinagh, Syloti Nagri, Balinese, Phags-pa, Old Persian, Phoenician, Kharoshthi, and Sumero-Akkadian Cuneiform. Other additions include extensions for Arabic and Ethiopic scripts, and for Biblical Hebrew texts.

The first third of the book, some 562 pages, is about the history, structure, principles, and use of the Unicode character set and its encoding. This is where the difficult topics of writing direction, combining forms, decomposition, conformance, spacing, numerical use, casing, properties, *etc.*, of the characters themselves are presented, together with the discussions of the features of the scripts themselves. This part of the book by itself is a reference on the principles of written communication, and is a thorough education, even without going further.

The last third of the book, approximately 390 pages, contains the Han radical stroke index, essential to locating specific Han ideographic characters within the Han ideographic set, and all the appendices, including the notational conventions, reprints of the Technical Standards and Reports, the detailing of the relationship of the Unicode standard to the ISO/IEC 10646 Standard, and how the Versions have changed. In particular, Appendix F states the formal stability policies adopted by the Unicode Consortium. This latter work becomes increasingly important as the Standard grows, and many more organizations adopt the use of the Standard.

Even though there is now a large number of places where Unicode is used (in at least one of its transformational formats), it remains true that there is still a large number of computer and information technology professionals who are not familiar with the true range of the Unicode Standard encoding. “Isn’t it just a sort of big ASCII?” is an attitude that is still encountered. One really does not have to go far to step outside the range of ASCII. To just pose the question, “Where is Québec?” is sufficient to break an ASCII editor; and to drop the accent makes as much sense to the French reader as dropping the a’s out of “Alabama” does to a Southern reader. The prevalent ISO 8859-1 Latin-1 character set addresses that problem, yet it does not allow for the mention of Québec and Россия in the same sentence. The need for computer programs to communicate with users globally can no longer be considered a luxury.

The work of the Unicode Consortium continues in many areas. Under the heading of “Issues for Public Review” at the official website are currently listed three areas of concern to the processing of text, and one concerning programming languages, each addressed by proposed updates to Unicode Annexes. The text processing topics are *Line Breaking Properties*, *Text Boundaries*, and *Unicode Normalization Forms*. The latter is of deep concern to the question of identifying “equivalent” strings, under differing degrees of equivalence. In particular, attention is being paid to the idempotency of normalization and the stability of the Unicode definition in order that the normalization of assigned characters will not change in future versions of Unicode. The other area, *Identifier and Pattern Syntax*, concerns itself, amongst other matters, with a uniform and consistent approach within the programming languages that support the Unicode character set for identifiers

So, why buy this book? The Version 5.0 Standard itself answers this question, and lists, in addition to other advances, that “four fifths of the figures are new; two-thirds of the definitions are new; one-half of the Unicode Standard Annexes are new; one-third of the conformance clauses are new; one-fourth of the tables are new.” As good and great as the Version 4.0 Standard was at its time of publication, it is definitely now obsolete. And then there is the question this reviewer asked when Version 4.0 was released: “is there a reason to buy the ‘printed-on-paper’ version?” Although the online and electronic files are both essential and invaluable, you do need to know for what it is you are looking amongst that material. For education, for being able to see the BIG picture, and the whole context, there is still no adequate alternative to the printed book.

Bruce K. Haddon