Electronic Edition

This file is part of the electronic edition of *The Unicode Standard, Version 5.0*, provided for online access, content searching, and accessibility. It may not be printed. Bookmarks linking to specific chapters or sections of the whole Unicode Standard are available at

http://www.unicode.org/versions/Unicode5.0.0/bookmarks.html

Purchasing the Book

For convenient access to the full text of the standard as a useful reference book, we recommend purchasing the printed version. The book is available from the Unicode Consortium, the publisher, and booksellers. Purchase of the standard in book format contributes to the ongoing work of the Unicode Consortium. Details about the book publication and ordering information may be found at

http://www.unicode.org/book/aboutbook.html

Joining Unicode

You or your organization may benefit by joining the Unicode Consortium: for more information, see Joining the Unicode Consortium at

http://www.unicode.org/consortium/join.html

Library of Congress Cataloging-in-Publication Data

The Unicode Standard / the Unicode Consortium ; edited by Julie D. Allen ... [et al.]. — Version 5.0.
p. cm.
Includes bibliographical references and index.
ISBN 0-321-48091-0 (hardcover : alk. paper)
1. Unicode (Computer character set) I. Allen, Julie D.
II. Unicode Consortium.
QA268.U545 2007
005.7'22—dc22

2006023526

ISBN 0-321-48091-0
Text printed in the United States on recycled paper at Courier in Westford, Massachusetts.
First printing, October 2006

Foreword

Without much fanfare, Unicode has completely transformed the foundation of software and communications over the past decade. Whenever you read or write anything on a computer, you're using Unicode. Whenever you search on Google, Yahoo!, MSN, Wikipedia, or many other Web sites, you're using Unicode. Unicode 5.0 marks a major milestone in providing people everywhere the ability to use their own languages on computers.

We began Unicode with a simple goal: to unify the many hundreds of conflicting ways to encode characters, replacing them with a single, universal standard. Those existing legacy character encodings were both incomplete and inconsistent: Two encodings could use the same internal codes for two different characters and use different internal codes for the same characters; none of the encodings handled any more than a small fraction of the world's languages. Whenever textual data was converted between different programs or platforms, there was a substantial risk of corruption. Programs were hard-coded to support particular encodings, making development of international versions expensive, testing a nightmare, and support costs prohibitive. As a result, product launches in foreign markets were expensive and late—unsatisfactory both for companies and their customers. Developing countries were especially hard-hit; it was not feasible to support smaller markets. Technical fields such as mathematics were also disadvantaged; they were forced to use special fonts to represent arbitrary characters, but when those fonts were unavailable, the content became garbled.

Unicode changed that situation radically. Now, for all text, programs only need to use a single representation—one that supports all the world's languages. Programs could be easily structured with all translatable material separated from the program code and put into a single representation, providing the basis for rapid deployment in multiple languages. Thus, multiple-language versions of a program can be developed almost simultaneously at a much smaller incremental cost, even for complex programs like Microsoft Office or OpenOffice.

The assignment of characters is only a small fraction of what the Unicode Standard and its associated specifications provide. They give programmers extensive descriptions and a vast amount of data about how characters *function*: how to form words and break lines; how to sort text in different languages; how to format numbers, dates, times, and other elements appropriate to different languages; how to display languages whose written form flows from right to left, such as Arabic and Hebrew, or whose written form splits, combines, and reorders, such as languages of South Asia; and how to deal with security concerns regarding the many "look-alike" characters from alphabets around the world. Without the proper-

ties, algorithms, and other specifications in the Unicode Standard and its associated specifications, interoperability between different implementations would be impossible.

With the rise of the Web, a single representation for text became absolutely vital for seamless global communication. Thus the textual content of HTML and XML is defined in terms of Unicode—every program handling XML must use Unicode internally. The search engines all use Unicode for good reason; even if a Web page is in a legacy character encoding, the only effective way to index that page for searching is to translate it into the lingua franca, Unicode. All of the text on the Web thus can be stored, searched, and matched with the same program code. Since all of the search engines translate Web pages into Unicode, the most reliable way to have pages searched is to have them be in Unicode in the first place.

This edition of *The Unicode Standard, Version 5.0*, supersedes and obsoletes all previous versions of the standard. The book is smaller in size, less expensive, and yet has hundreds of pages of new material and hundreds more of revised material. Like any human enterprise, Unicode is not without its flaws, of course. This book will help you work around some of the "gotchas" introduced into Unicode over the course of its development. Importantly, it will help you to understand which features may change in the future, and which cannot, so that you can appropriately optimize your implementations. You will also find a wealth of other information on the Unicode Web site (www.unicode.org). If you are interested in having a voice in determining directions for future development of Unicode, or want to follow closely the ongoing work, you will find information there on joining the Consortium.

What you have in your hands is the culmination of many years of experience from experts around the globe. I am sure you will find it very useful.

Mark Davis, Ph.D.
President
The Unicode Consortium

