DOWNLOAD PDF THE UNICODE STANDARD, VERSION 3.0

Chapter 1 : CiteSeerX â€" Citation Query The Unicode Standard, Version

Formally, a version of the Unicode Standard is defined by an edition of the core specification, The Unicode Standard, together with the Code Charts, Unicode Standard Annexes and the Unicode Character Database.

As well as encoding characters used for written communication in a simple and consistent manner, the Unicode Standard defines character properties and algorithms for use in implementations. The previous versions of the Unicode Standard are: The Unicode Standard, Version 1. The general principles and architecture of the Unicode Standard, requirements for conformance, and guidelines for implementers precede the actual coding information. Useful ancillary information is given in the appendices. The accompanying CD-ROM contains tables of use to implementers and all technical reports published to date. Basic text processing, working with combining marks, encoding forms, and doing bidirectional text layout are all described. A special chapter on implementation guidelines answers many common questions that arise when implementing Unicode. Chapter 2 sets forth the fundamental principles underlying the Unicode Standard and covers specific topics such as text processes, overall character properties, and the use of combining marks. Chapter 3 constitutes the formal statement of conformance. This chapter also presents the normative algorithms for three processes: Chapter 4 describes character properties in detail, both normative required and informative. Chapter 5 discusses implementation issues, including compression, strategies for dealing with unknown and unsupported characters, and transcoding to other standards. Character Block Descriptions Chapters 6 through 13 contain the character block descriptions that give basic information about each script or collection and may discuss specific characters or pertinent layout information. Chapter 6 describes the general punctuation characters. Chapter 8 presents the Middle Eastern, right-to-left scripts: Hebrew, Arabic, Syriac, and Thaana. Chapter 12 presents symbols, including currency, letterlike and technical symbols, and mathematical operators. Chapter 13 describes special characters such as the Private Use Area, surrogates, and specials. Chapter 14 gives the code charts and the Character Names List. The code charts contain the normative character encoding assignments, and the names list contains normative information as well as useful cross references and informational notes. Appendices and Tables The appendices contain detailed background information on important topics: Appendix B gives instructions on how to submit characters for consideration as additions to the Unicode Standard. Appendix D lists the changes to the Unicode Standard since Version 2. The appendices are followed by a glossary of terms, a bibliography, and two indices: The Unicode Character Database and Technical Reports The Unicode Character Database is the name for a collection of files that contain character code values, character names, and character property data. It is described more fully in the file UnicodeCharacterDatabase. Updates and revisions will be made available online. The following Unicode Technical Reports are formally part of this standard: East Asian Width, Version 5. Unicode Newline Guidelines, Version 5. Line Breaking Properties, Version 6. Unicode Normalization Forms, Version For information on the latest available version, see http: In addition to the Unicode Character Database and Unicode Technical Reports that are part of this standard, the CD-ROM also contains additional technical reports covering topics such as compression, collation, and transformation formats , as well as property-based mapping tables for example, tables for case and transcoding tables for international, national, and industry character sets including the Han cross-reference table. Unicode character names contain only uppercase Latin letters A through Z, digits, space, and hyphen-minus; this convention makes it easy to generate computer-language identifiers automatically from the names. The names of Hangul syllables are generated algorithmically; for details, see Hangul Syllable Names in Section 3. Italics are also used to refer to a text element that is not explicitly encoded for example, pasekh alef or to set off a foreign word for example, the Welsh word ynghyd. The symbols used in the character names list are described at the beginning of Chapter 14, Code Charts. In the text of this book, the word "Unicode" when used alone as a noun refers to the Unicode Standard. In this book, unambiguous dates of the current common era, such as , are unlabeled. In cases of ambiguity, CE is used. Dates before the common era are labeled with BCE.

Chapter 2 : Unicode Standard, Version , The | InformIT

The Unicode Consortium is a non-profit organization founded to develop, extend, and promote the use of the Unicode Standard. The membership of the Consortium represents a broad spectrum of corporations and organizations in the computer and information processing industry.

Syriac, containing 71 characters used for writing in Syriac script, was added. Thaana, containing 49 characters used for writing in Thaana script, was added. Sinhala, containing 80 characters for the Sinhala script, was added. Myanmar, containing 78 characters for the Burmese script, was added. Ethiopic, containing syllables and punctuation marks for the Ethiopic script, was added. Cherokee, containing 85 syllables for the Cherokee script, was added. Unified Canadian Aboriginal Syllabics, containing syllables and punctuation marks for writing in aboriginal languages of Canada, was added. Ogham, containing 29 characters for the ancient Ogham script, was added. Runic, containing 81 characters for the Germanic runes, was added. Khmer, containing characters for the Khmer script, was added. Mongolian, containing characters for the classical Mongolian script, was added. Braille Patterns, containing Braille letters, was added. Kangxi Radicals, containing radicals from the Kangxi dictionary, was added. Ideographic Description characters, used to describe a Han ideograph not available in the font, was added. Bopomofo Extended, containing 24 characters used for phonetic transcription of minority languages of Taiwan, was added. Yi Syllables, containing 1, syllables of the modern Yi script, was added. Yi Radicals, containing 50 radicals of Yi Syllables, was added. Extended blocks[edit] Additional precomposed characters, letters and capital letters of lowercase-only letters total 30 characters were added to Latin Extended-B. Extensions for disordered speech total 5 characters were added to IPA Extensions. Some additional modifier letters total 6 characters were added to Spacing Modifier Letters. Lowercase versions of archaic letters and the Kai symbol total 5 characters were added to Greek and Coptic. Nonstandard letters for Macedonian, combining numeral signs and three letters for Kildin Sami total 12 characters were added to Cyrillic. Combining hamza and maddah and nine additional Arabic characters total 12 characters were added to Arabic. Additional letters and religious symbols total 25 characters were added to Tibetan. A narrow no-break space and 6 additional punctuation marks total 7 characters were added to General Punctuation. An enclosing screen and an enclosing key total 2 characters were added to Combining Diacritical Marks for Symbols. The information symbol and a rotated Q total 2 characters were added to Letterlike Symbols. Some additional arrows total 9 characters were added to Arrows. Some additional technical symbols, including common keys on a keyboard total 33 characters were added to Miscellaneous Technical. Two additional control pictures total 2 characters were added to Control Pictures. Squares and circles with quadrants total 8 characters were added to Geometric Shapes. Two Syriac crosses and a signature mark total 3 characters were added to Miscellaneous Symbols. Three additional control characters for ruby markup total 3 characters were added to Specials. It encoded 94, characters and mainly focused on blocks outside of the Basic Multilingual Plane. Old Italic, containing 35 letters for the Etruscan script, was added. Gothic, containing 27 letters for the Gothic script, was added. Deseret, containing 76 letters for the constructed Deseret script, was added. Byzantine Musical Symbols, containing symbols for musical notation in Byzantine, was added. Musical Symbols, containing characters for current musical notation, was added. Tags, containing 97 language tags, was added.

Chapter 3 : Unicode character encoding

Of course this version, , is already out-of-date. But updates and corrections are easily available from the official Unicode website where data for Beta appears as I write this. My book bulges with interleaved additions and changes.

A fundamental aspect of computer systems is that displaying and signing a digital document are separate and unlinked processes. In addition, the same digital document can be displayed differently on different systems. As a consequence it is difficult to determine what exactly has been signed, both f This paper discusses how confusion about the meaning of digitally signed documents can occur, and proposes some mitigation strategies. Maruf Hasan, Yuji Matsumoto " Electronically available multilingual information can be divided into two major categories: The information available in non-English alphabetic langu The information available in non-English alphabetic languages as well as in ideographic languages especially, in Japanese and Chinese is growing at an incredibly high rate in recent years. Due to the ideographic nature of Japanese and Chinese, complicated with the existence of several encoding standards in use, efficient processing representation, indexing, retrieval, etc. In this paper, we propose a Han Character Kanji oriented Interlingua model of indexing and retrieving Japanese and Chinese information. We report the results of mono- and cross-language information retrieval on a Kanji space where documents and queries are represented in terms of Kanji oriented vectors. We also employ a dimensionality reduction technique to compute a Kanji Conceptual Space KCS from the initial Kanji space, which can facilitate conceptual retrieval of both mono- and cross-language information for these languages. Similar indexing approaches for multiple European languages through term association e. The Interlingua approach investigated here with Japanese and Chinese languages, and the term or concept association model investigated with the European languages are similar; and these approaches can be easily integrated. Therefore, the proposed Interlingua model can pave the way for handling multilingual information access and retrieval efficiently and uniformly. Show Context Citation Context This offers us an opportunity to represent Japanese and Chines Maruf Hasan , " With the advent of the Internet and digital libraries, as well as the proliferation of multilingual information, sophisticated methods of representation and indexing, and the retrieval of such information is essential. In recent years, the amount of electronically available information has escalated In recent years, the amount of electronically available information has escalated. The non-English information information in Asian and European languages is growing rapidly. In this thesis, I concentrate on these two languages, which are quite different from European languages in the sense that semantically rich ideographic Han-characters hereafter, Kanji are used in the writing systems of both languages. I explore several strategies using the Kanji and Kanji-derived-semantics for indexing and retrieval of Japanese and Chinese information. The Kanji-based Interlingual framework proposed in this thesis for Japanese-Chinese information retrieval is a flexible vector-space framework. Therefore, projection and dimensionality reduction techniques, such as the singular value decomposition SVD , are easy to incorporate with this framework. As explained above, it can be concluded that effective cross-language informat

Chapter 4 : Unicode Emoji List

Version has been superseded by the latest version of the Unicode Standard. This page summarizes the important changes for the Unicode Standard, Version The core specification was not republished for Version Thus the chapters of the core specification use the Version PDF files.

We introduce a lazy XSLT interpreter that provides random access to the transformation result. This allows e - cient pipelining of transformation sequences. Nodes of the result tree are computed only upon initial access. As these computations have limited fan-in, sparse output coverage propagates ba As these computations have limited fan-in, sparse output coverage propagates backwards through the pipeline. Managing the organizational and software complexity of a comprehensive open source digital library system presents a significant challenge. The challenge becomes even more imposing when the interface is available in di erent languages, for enhancements to the software and changes to the interfac The challenge becomes even more imposing when the interface is available in di erent languages, for enhancements to the software and changes to the interface must be faithfully reflected in each language version. This paper describes the solution adopted by Greenstone, a multilingual digital library system distributed by UNESCO in a trilingual European version English, French, Spanish , complete with all documentation, and whose interface is available in many further languages. Show Context Citation Context Unicode is an ISO standard providing every character in every language with a unique number. For example, in Unicode, a Western Y with acute accent has the code while a Western dot-less i has th Marc Laukien, Matthew Newhook, Bernard NormierMany of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book and ZeroC was aware of the trademark claim, the designations have been printed in initial cap Where those designations appear in this book and ZeroC was aware of the trademark claim, the designations have been printed in initial caps or all caps. The authors and publisher have taken care in the preparation of this book, but make no expressed or implied warranty of any kind and assume no responsibility for errors or omissions. No liability is assumed for incidental or consequential damages in connection with or arising out of the use of the information or programs contained herein. Strings are not NUL-terminated. An empty string is encoded with a size of zero. However, XML-based applications often use XML interfaces to legacy software which in many cases is not capable of dealing with the full Unicode character repertoire. We therefore propose a schema lang We therefore propose a schema language for XML which is capable of limiting the character repertoire of XML documents. In many application scenarios, only certain classes of XML documents are considered useful. In the classical XML world, these classes are defined by a D This manual is provided under one of two licenses, whichever you prefer:

Chapter 5 : The Unicode Standard, Version 0 by The Unicode Consortium

This book, The Unicode Standard, Version , is the authoritative source of information on the Unicode character encoding standard, the international character code for information processing that includes all major scripts of the world and is the foundation for develop-.

Bidirectional Behavior Improvements This new version updates the Unicode Bidirectional Algorithm to ensure that pairs of parentheses and brackets have consistent layout and to provide a mechanism for isolating runs of text. The updated Bidirectional Algorithm together with five newly introduced bidi format characters will improve the display of text for hundreds of millions of users of Arabic, Hebrew, Persian, Urdu, and many others. The display and positioning of parentheses will better match the normal behavior that users expect. By using the new methods for isolating runs of text, software will be able to construct messages from different sources without jumbling the order of characters. The new bidi format characters correspond to features in markup such as in CSS. Overall, these improvements bring greater interoperability and an improved ability for inserting text and assembling user interface elements in these languages. The improvements come with new rigor: Other Enhancements In a major enhancement for CJK usage, this new version adds standardized variation sequences for all 1, CJK compatibility ideographs. Using the new standardized variation sequences allows authors to write text which will preserve the specific required shapes of these CJK ideographs, even under Unicode normalization. Synchronization Two other important Unicode specifications are maintained in synchrony with the Unicode Standard, and have updates for Version 6. See Sections D through H below for additional details regarding the changes in this version of the Unicode Standard, its associated annexes, and the other synchronized Unicode specifications. Version Information Version 6. The core specification gives the general principles, requirements for conformance, and guidelines for implementers. The code charts show representative glyphs for all the Unicode characters. The Unicode Standard Annexes supply detailed normative information about particular aspects of the standard. The Unicode Character Database supplies normative and informative data for implementers to allow them to implement the Unicode Standard. The Unicode Standard, Version 6. The Unicode Consortium, The citation and permalink for the latest published version of the Unicode Standard is: That page also provides the recommended reference format for Unicode Standard Annexes. Code Charts Several sets of code charts are available. They serve different purposes: The latest set of code charts for the Unicode Standard are available online. Those charts are always the most current code charts available, and may be updated at any time. The charts are organized by scripts and blocks for easy reference. An online index by character name is also provided. A set of delta code charts showing the blocks in which bidirectional format controls were added for Unicode 6. Those characters are visually highlighted in the relevant chart. These delta code charts also include blocks which contain significant glyph changes to fix errata. A set of archival code charts that represent the entire set of characters, names and representative glyphs at the time of publication of Unicode 6. The delta and archival code charts are a stable part of this release of the Unicode Standard. They will never be updated.

Chapter 6 : The Unicode Standard, Version - PDF Free Download - Fox eBook

The Unicode Standard, Version is THE authoritative source of information on the Unicode character-encoding standard, which makes it possible to create global software and share data across languages, nations, and locales worldwide.

Chapter 7 : Unicode/Versions - Wikibooks, open books for an open world

A book/CD-ROM technical guide to the Unicode character encoding standard, the international character code for information processing that includes all major scripts of the world and is the foundation for development of software for worldwide use.

DOWNLOAD PDF THE UNICODE STANDARD, VERSION 3.0

Chapter 8 : CiteSeerX â€” Citation Query Unicode Standard, Version

Note: Citations are based on reference standards. However, formatting rules can vary widely between applications and fields of interest or study. The specific requirements or preferences of your reviewing publisher, classroom teacher, institution or organization should be applied.

Chapter 9 : Unicode Standard

We're upgrading the ACM DL, and would like your input. Please sign up to review new features, functionality and page designs.