유니 코드

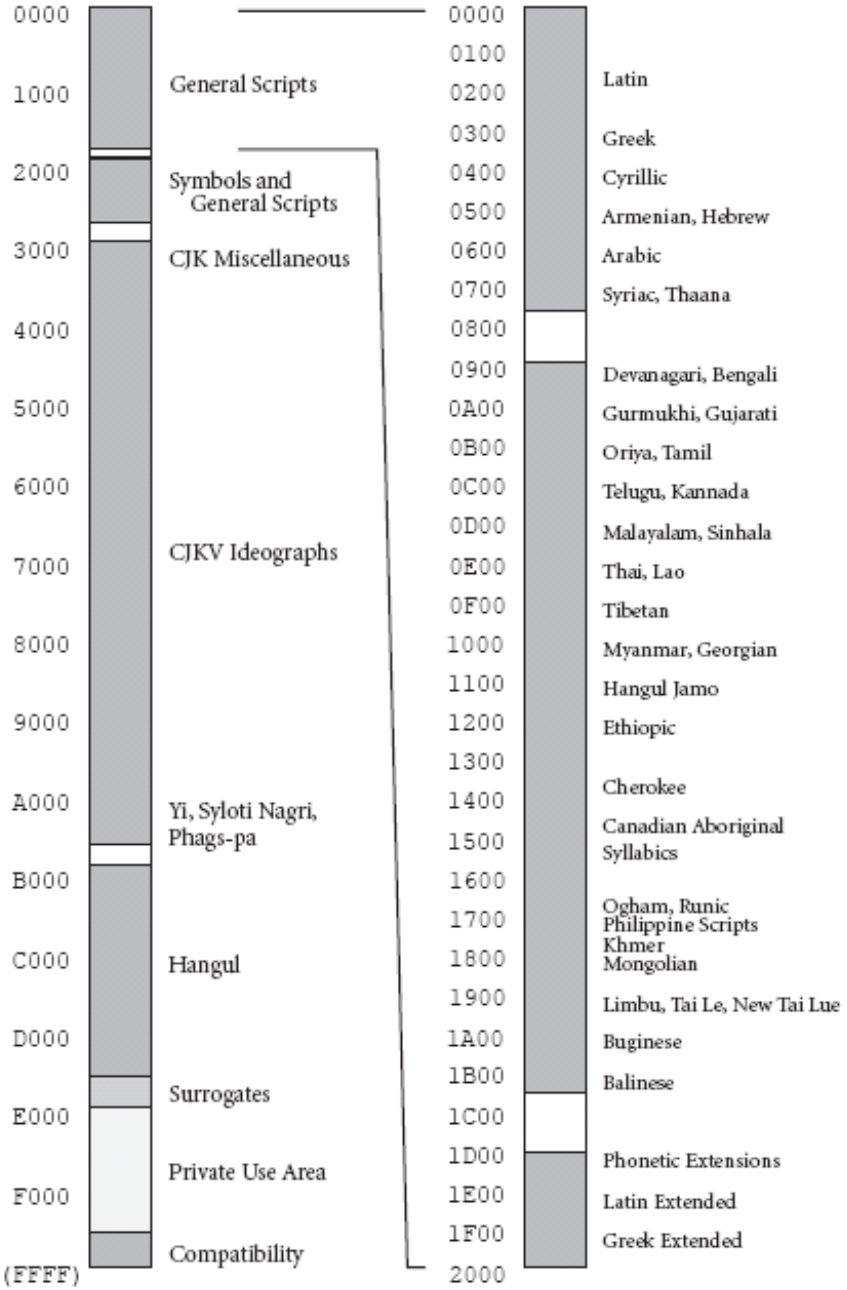
국민대학교 소프트웨어학부

강 승 식

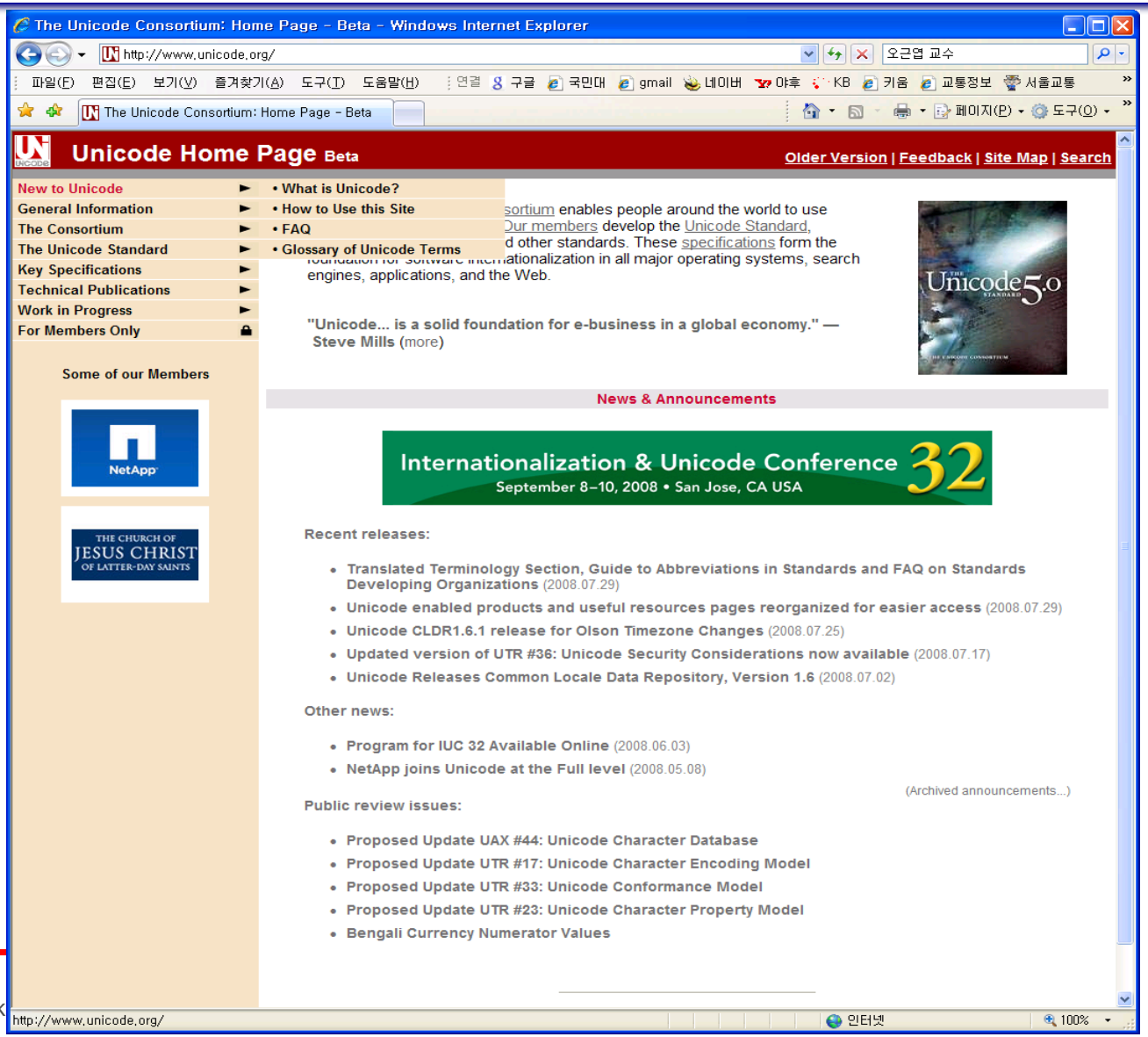
- What is unicode?
  - BMP(Basic Multilingual Plane), SMP
  - UCS-2, UCS-4
  - UTF-8, UTF-16, UTF-32
  - BOM(Byte Order Mark)
  - Big Endian, Little Endian

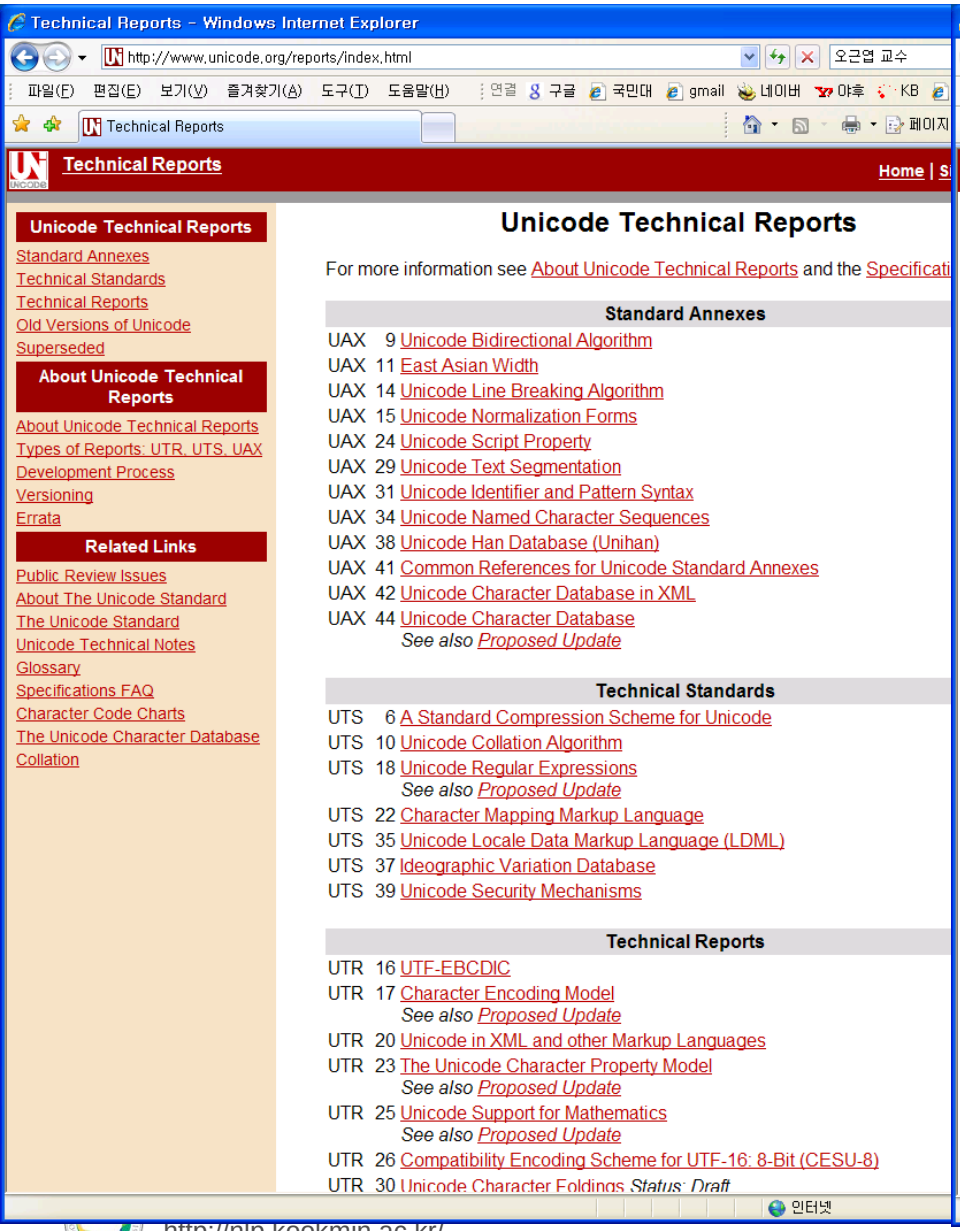
- 유니코드 문자셋 정의 영역: 언어별로
  - 각 언어별 문자 인식 방법
    - 한글/한자/아랍어/러시아어 등 어떤 영역의 문자인지 확인 방법
  - 다국어 문자가 혼합되어 있을 때 언어별로 tokenization

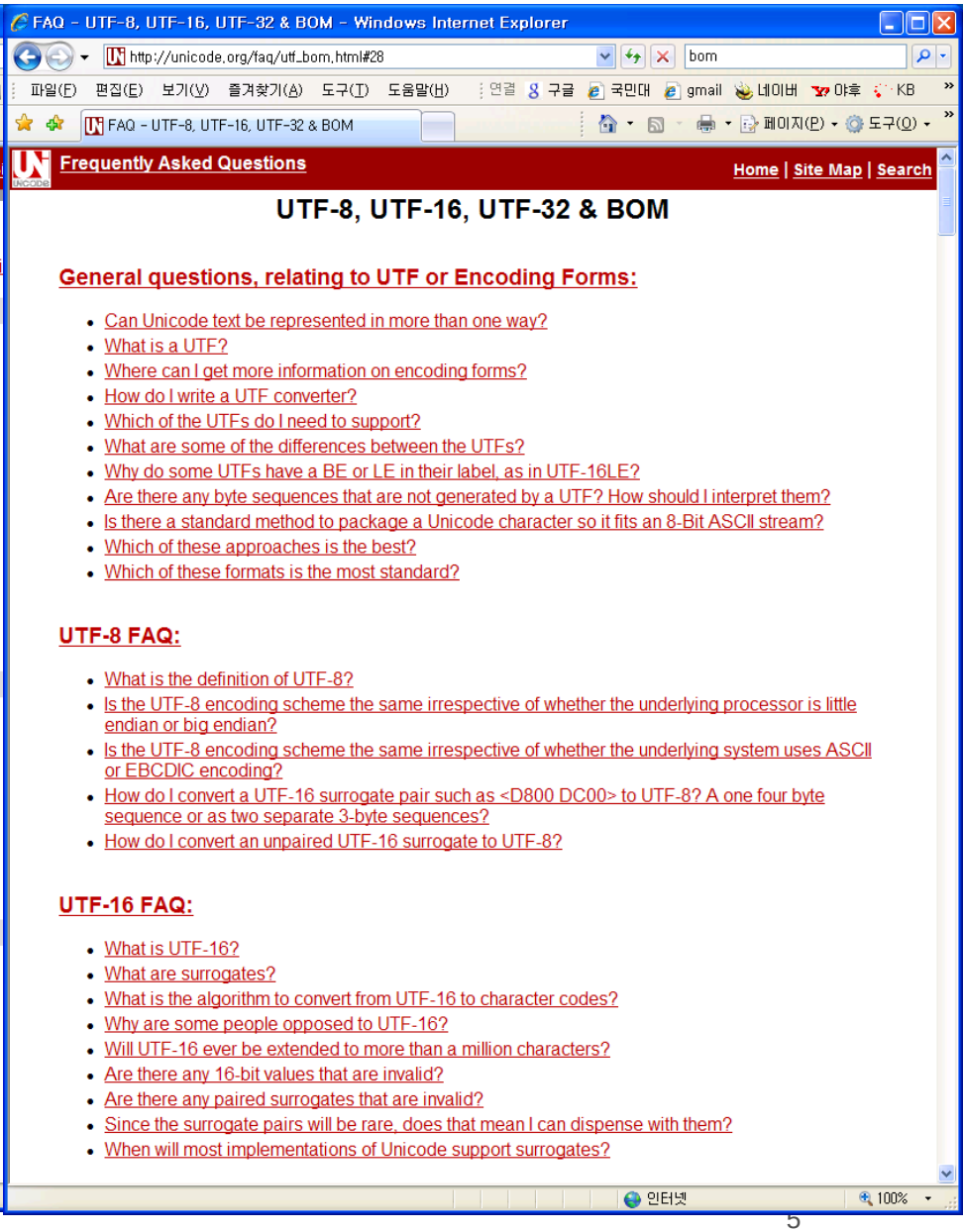
Figure 2-14. Allocation on the BMP

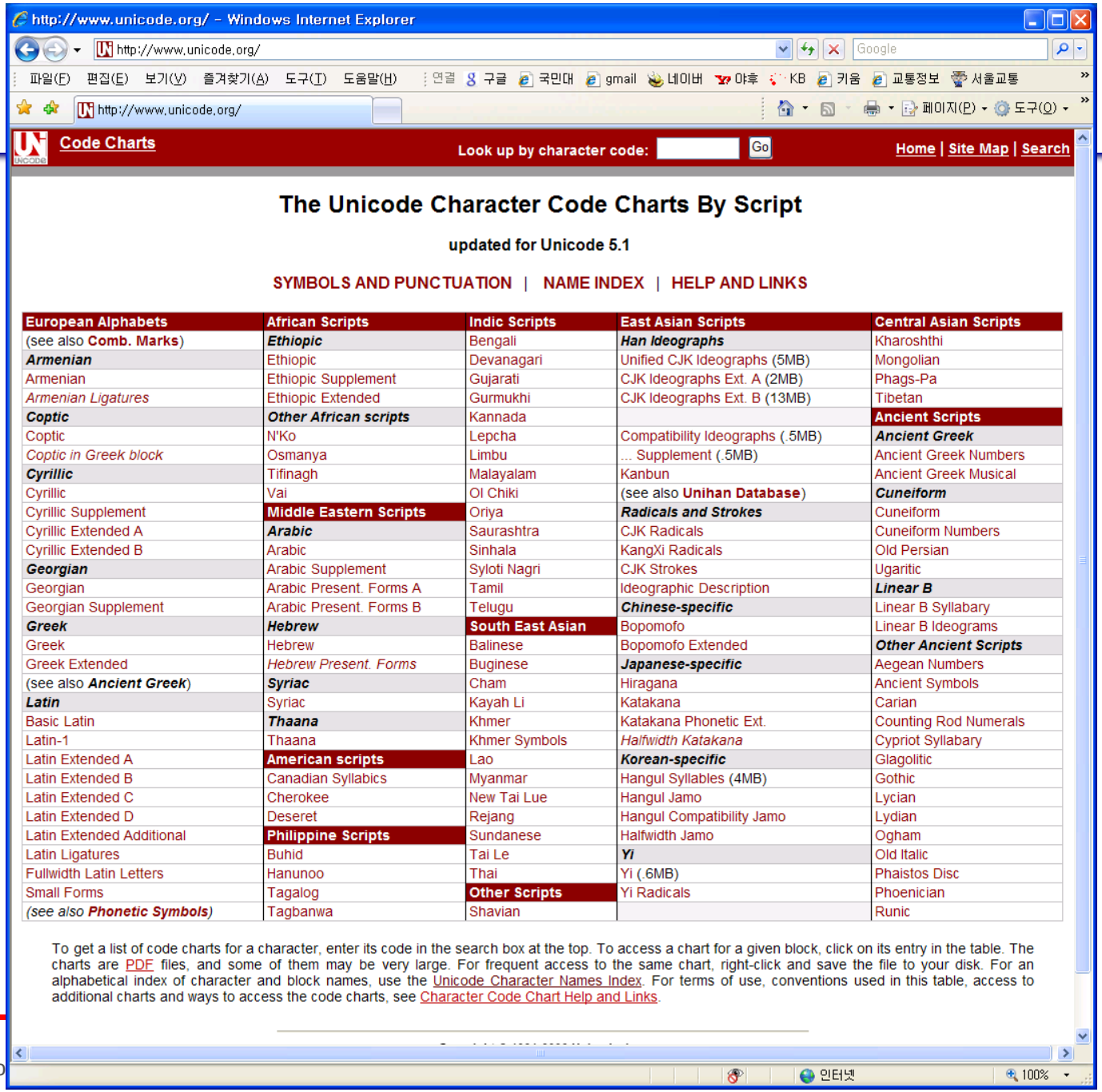


http://www.unicode.org/




Technical Reports - Windows Internet Explorer
http://www.unicode.org/reports/index.html
Technical Reports
Home
Unicode Technical Reports
Standard Annexes
Technical Standards
Technical Reports
Old Versions of Unicode
Superseded
About Unicode Technical Reports
About Unicode Technical Reports
Types of Reports: UTR, UTS, UAX
Development Process
Versioning
Errata
Related Links
Public Review Issues
About The Unicode Standard
The Unicode Standard
Unicode Technical Notes
Glossary
Specifications FAQ
Character Code Charts
The Unicode Character Database
Collation
Unicode Technical Reports
For more information see About Unicode Technical Reports and the Specificati
Standard Annexes
UAX 9 Unicode Bidirectional Algorithm
UAX 11 East Asian Width
UAX 14 Unicode Line Breaking Algorithm
UAX 15 Unicode Normalization Forms
UAX 24 Unicode Script Property
UAX 29 Unicode Text Segmentation
UAX 31 Unicode Identifier and Pattern Syntax
UAX 34 Unicode Named Character Sequences
UAX 38 Unicode Han Database (Unihan)
UAX 41 Common References for Unicode Standard Annexes
UAX 42 Unicode Character Database in XML
UAX 44 Unicode Character Database
See also Proposed Update
Technical Standards
UTS 6 A Standard Compression Scheme for Unicode
UTS 10 Unicode Collation Algorithm
UTS 18 Unicode Regular Expressions
See also Proposed Update
UTS 22 Character Mapping Markup Language
UTS 35 Unicode Locale Data Markup Language (LDML)
UTS 37 Ideographic Variation Database
UTS 39 Unicode Security Mechanisms
Technical Reports
UTR 16 UTF-EBCDIC
UTR 17 Character Encoding Model
See also Proposed Update
UTR 20 Unicode in XML and other Markup Languages
UTR 23 The Unicode Character Property Model
See also Proposed Update
UTR 25 Unicode Support for Mathematics
See also Proposed Update
UTR 26 Compatibility Encoding Scheme for UTF-16: 8-Bit (CESU-8)
UTR 30 Unicode Character Foldings Status: Draft


FAQ - UTF-8, UTF-16, UTF-32 & BOM - Windows Internet Explorer
http://unicode.org/faq/utf_bom.html#28
FAQ - UTF-8, UTF-16, UTF-32 & BOM
Frequently Asked Questions
Home | Site Map | Search
UTF-8, UTF-16, UTF-32 & BOM
General questions, relating to UTF or Encoding Forms:
Can Unicode text be represented in more than one way?
What is a UTF?
Where can I get more information on encoding forms?
How do I write a UTF converter?
Which of the UTFs do I need to support?
What are some of the differences between the UTFs?
Why do some UTFs have a BE or LE in their label, as in UTF-16LE?
Are there any byte sequences that are not generated by a UTF? How should I interpret them?
Is there a standard method to package a Unicode character so it fits an 8-Bit ASCII stream?
Which of these approaches is the best?
Which of these formats is the most standard?
UTF-8 FAQ:
What is the definition of UTF-8?
Is the UTF-8 encoding scheme the same irrespective of whether the underlying processor is little endian or big endian?
Is the UTF-8 encoding scheme the same irrespective of whether the underlying system uses ASCII or EBCDIC encoding?
How do I convert a UTF-16 surrogate pair such as <D800 DC00> to UTF-8? A one four byte sequence or as two separate 3-byte sequences?
How do I convert an unpaired UTF-16 surrogate to UTF-8?
UTF-16 FAQ:
What is UTF-16?
What are surrogates?
What is the algorithm to convert from UTF-16 to character codes?
Why are some people opposed to UTF-16?
Will UTF-16 ever be extended to more than a million characters?
Are there any 16-bit values that are invalid?
Are there any paired surrogates that are invalid?
Since the surrogate pairs will be rare, does that mean I can dispense with them?
When will most implementations of Unicode support surrogates?


http://www.unicode.org/ - Windows Internet Explorer
http://www.unicode.org/
Code Charts
Look up by character code:
Go
Home | Site Map | Search
The Unicode Character Code Charts By Script
updated for Unicode 5.1
SYMBOLS AND PUNCTUATION | NAME INDEX | HELP AND LINKS
European Alphabets
(see also Comb. Marks)
Armenian
Armenian
Armenian Ligatures
Coptic
Coptic
Coptic in Greek block
Cyrillic
Cyrillic
Cyrillic Supplement
Cyrillic Extended A
Cyrillic Extended B
Georgian
Georgian
Georgian Supplement
Greek
Greek
Greek Extended
(see also Ancient Greek)
Latin
Basic Latin
Latin-1
Latin Extended A
Latin Extended B
Latin Extended C
Latin Extended D
Latin Extended Additional
Latin Ligatures
Fullwidth Latin Letters
Small Forms
(see also Phonetic Symbols)
African Scripts
Ethiopic
Ethiopic
Ethiopic Supplement
Ethiopic Extended
Other African scripts
N'Ko
Osmanya
Tifinagh
Vai
Middle Eastern Scripts
Arabic
Arabic
Arabic Supplement
Arabic Present. Forms A
Arabic Present. Forms B
Hebrew
Hebrew
Hebrew Present. Forms
Syriac
Syriac
Thaana
Thaana
American scripts
Canadian Syllabics
Cherokee
Deseret
Philippine Scripts
Buhid
Hanunoo
Tagalog
Tagbanwa
Indic Scripts
Bengali
Devanagari
Gujarati
Gurmukhi
Kannada
Lepcha
Limbu
Malayalam
Ol Chiki
Oriya
Saurashtra
Sinhala
Syloti Nagri
Tamil
Telugu
South East Asian
Balinese
Buginese
Cham
Kayah Li
Khmer
Khmer Symbols
Lao
Myanmar
New Tai Lue
Rejang
Sundanese
Tai Le
Thai
Other Scripts
Shavian
East Asian Scripts
Han Ideographs
Unified CJK Ideographs (5MB)
CJK Ideographs Ext. A (2MB)
CJK Ideographs Ext. B (13MB)
Compatibility Ideographs (.5MB)
... Supplement (.5MB)
Kanbun
(see also Unihan Database)
Radicals and Strokes
CJK Radicals
KangXi Radicals
CJK Strokes
Ideographic Description
Chinese-specific
Bopomofo
Bopomofo Extended
Japanese-specific
Hiragana
Katakana
Katakana Phonetic Ext.
Halfwidth Katakana
Korean-specific
Hangul Syllables (4MB)
Hangul Jamo
Hangul Compatibility Jamo
Halfwidth Jamo
Yi
Yi (.6MB)
Yi Radicals
Central Asian Scripts
Kharoshthi
Mongolian
Phags-Pa
Tibetan
Ancient Scripts
Ancient Greek
Ancient Greek Numbers
Ancient Greek Musical
Cuneiform
Cuneiform
Cuneiform Numbers
Old Persian
Ugaritic
Linear B
Linear B Syllabary
Linear B Ideograms
Other Ancient Scripts
Aegean Numbers
Ancient Symbols
Carian
Counting Rod Numerals
Cypriot Syllabary
Glagolitic
Gothic
Lycian
Lydian
Ogham
Old Italic
Phaistos Disc
Phoenician
Runic
To get a list of code charts for a character, enter its code in the search box at the top. To access a chart for a given block, click on its entry in the table. The charts are PDF files, and some of them may be very large. For frequent access to the same chart, right-click and save the file to your disk. For an alphabetical index of character and block names, use the Unicode Character Names Index. For terms of use, conventions used in this table, access to additional charts and ways to access the code charts, see Character Code Chart Help and Links.
인터넷
100%

유니코드 2.0

- 한글 11,172자를 순서대로 코드 부여
  - 코드 범위 : '가'(0xAC00) ~ '힣'(0xD7A3)
- 자모 코드(0x31xx 영역)
  - 자음 : 0x3131 ~ 0x314E
  - 모음 : 0x314F ~ 0x3163
  - 채움 코드 : 0x3164
  - 옛글 자모 : 0x3165 ~ 0x318E
- 초성-중성-종성 코드(0x11xx 영역)

- CJKV 한자 코드
  - 한-중-일 공통 한자는 1개의 코드 부여
  - 음가 2 이상인 한자에 하나의 코드만 부여

유니코드에서 초성/중성/종성 인식

- 초성

  ((코드값 - 0xAC00) / 28) / 21) % 19

- 중성

  ((코드값 - 0xAC00) / 28) % 21

- 종성

  (코드값 - 0xAC00) % 28

유니코드 <–> KS완성형 변환

- 유니코드 <-> KS완성형 변환
  - 한글 11,172자
  - 자모
  - 문장부호: 가운데점 등
  - 한자
- 코드변환 테이블 작성
  - Windows의 notepad(메모장) 이용

- Java, Perl에서 유니코드-완성형 변환 방법
- “iconv” 라이브러리 사용

Topics

1. Unicode란 무엇인가?

- Unicode의 특징, EUC-KR과 다른 점 등
- Unicode의 한글영역: 음절, 자음과 모음, 특수문자(가운데점 등)

2. Unicode의 국가별 문자셋 영역(한글 제외)

- 아시아 언어 영역 -- 일본어 영역, 중국어(한자) 영역 등
- 유럽어 영역 -- 독일어의 움라우트, 러시아어 등
- 아랍어 영역
- 기타 언어 영역

3. Tokenization program 작성

한/중/일/아랍어/러시아어/문장부호 등이 혼합된 입력 문자열에 대해 각 언어 영역별로 문자열을 분리하는 프로그램

4. Unicode Encoding Forms

UTF-8, UTF-16, UTF-32

5. Unicode 관련 용어 설명

http://www.unicode.org/의 "Glossary of Unicode Terms" 참조

- Big Endian, Little Endian
- BMP Code Point, Supplementary Code Point, Surrogate Code Point, High-Surrogate Code Point
- BOM(Byte Order Mark)
- CJK
- Code Page, Code Point, Codespace
- DBCS, SBCS
- Consonant, Vowel, Diphthong, Jamo, Leading Consonant
- Syllable block, Hangul Syllable Block