МИНИСТЕРСТВО ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ И РАЗВИТИЯ КОММУНИКАЦИЙ

ФЕРГАНСКИЙ ФИЛИАЛ ТАШКЕНТСКОГО УНИВЕРСИТЕТА ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

Факультет «Компьютерной инженерии»

Отдел информационных технологий

По веб-программирования

РЕФЕРАТ

Подготовила: Г.Рахимжонова

Фергана - 2014

Что такое Unicode?

Unicode (**Юникод** или **Уникод**, англ. ***Unicode***) — стандарт кодирования символов, позволяющий представить знаки практически всех письменных языков. Юникод имеет несколько форм представления: UTF-8, UTF-16 (UTF-16BE, UTF-16LE) и UTF-32 (UTF-32BE, UTF-32LE). Была разработана также форма представления UTF-7 для передачи по семибитным каналам, но из-за несовместимости с ASCII она не получила распространения и не включена в стандарт. В MicrosoftWindows NT и основанных на ней системах Windows 2000 и Windows XP в основном используется форма UTF-16LE. В UNIX-подобных операционных системах GNU/Linux, BSD и Mac OS X принята форма UTF-8 для файлов и UTF-32 или UTF-8 для обработки символов в оперативной памяти.

Стандарт предложен в 1991 году некоммерческой организацией «Консорциум Юникода» (англ.*Unicode Consortium*), объединяющей крупнейшие IT-корпорации. Применение этого стандарта позволяет закодировать очень большое число символов из разных письменностей: в документах Unicode могут соседствовать китайские иероглифы, математические символы, буквы греческого алфавита и кириллицы, при этом становятся ненужными кодовые страницы.

Коды в стандарте Unicode разделены на несколько областей. Область с кодами от U+0000 до U+007F содержит символы набора ASCII с соответствующими кодами. Далее расположены области знаков различных письменностей, знаки пунктуации и технические символы. Часть кодов зарезервирована для использования в будущем. Под символы кириллицы выделены коды от U+0400 до U+052F (см. Кириллица в Юникоде).

Суть проблемы:

- Как правильно «Юникод» или «Уникод»?
- Как осуществляется поддержка Юникода в операционных системах?
- Двунаправленное письмо или поддержка арабских языков.

- Юникод - набор графических символов. Общие сведения.
  - Комбинируемые символы.
  - Представленные символы.
- Все символы, представленные в Юникоде.
  - Базовая многоязыковая плоскость.
  - Дополнительная многоязыковая плоскость.
  - Частная область.
- Проблемы Юникода.
- Версии юникода.
  - ISO/IEC 10646
- Управляющие символы

Посмотреть на форуме.

Как правильно «Юникод» или «Уникод»?

В русском языке слова с латинским элементом «uni-» традиционно писались через «уни-» (универсальный, униполярный, унификация, униформа). Однако для слова Unicode распространилось написание «Юникод» (видимо, изобретённое в компании «Майкрософт» при создании русской версии Windows 95). С пуристической же точки зрения предпочтительнее использовать написание «Уникод», так как в русском языке уже есть морфемы «уни-» и «код». До сих пор написание «юни-» использовалось только для собственных имён, заимствованных из английского языка (напр., «Юнилевер»). «Unicode» — международный термин, никак не привязанный к английскому языку, однако написание «Юникод» уже твёрдо вошло в русскоязычные тексты. Согласно «Яндексу», частота использования слова «Юникод» в 3,5 раза превышает «Уникод».

Как осуществляется поддержка Юникода в операционных системах?

Большинство современных операционных систем в той или иной степени обеспечивают поддержку Юникода.

В операционных системах семейства **Windows NT** для внутреннего представления имён файлов и других системных строк используется двухбайтовая кодировка UTF-16LE. Системные вызовы, принимающие строковые параметры, существуют в однобайтном и двухбайтном вариантах.

UNIX-образные операционные системы, в том числе, **Linux**, **BSD**, **Mac OS X**, используют для представления Юникода кодировку UTF-8. Большинство программ могут работать с UTF-8 как с традиционными однобайтными кодировками, не обращая внимания на то, что символ представляется как несколько последовательных байт. Для работы с отдельными символами строки обычно перекодируются в UCS-4, так что каждому символу соответствует машинное слово.

Одной из первых успешных коммерческих реализаций Юникода стала среда программирования**Java**. В ней принципиально отказались от восьмибитного представления символов в пользу шестнадцатибитного. Сейчас большинство языков программирования поддерживают строки Unicode, хотя их представление может различаться в зависимости от реализации.

Двунаправленное письмо или поддержка арабских языков.

Стандарт Юникод поддерживает языки как с направлением написания слева-направо (англ. left-to-right, **LTR**) так и с написанием справа-налево (англ. right-to-left, **RTL**), как **иврит** и**арабский** язык. Кроме того, Юникод поддерживает комбинированные тексты, содержащие одновременно **RTL** и **LTR** фразы. Данная возможность называется **двунаправленность** (англ. bidirectional, Bidir). Простые реализации Юникода могут не иметь поддержки двунаправленности.

Юникод - набор графических символов.

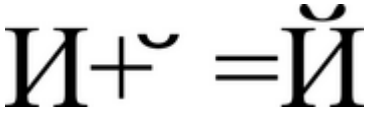
Универсальная система кодирования (Юникод) представляет собой набор графических символов и способ их кодирования для компьютерной обработки текстовых данных. Графические символы — это символы, имеющие видимое изображение. Графическим символам противопоставляются управляющие символы и символы форматирования.
Графические символы включают в себя следующие группы:

- буквы, содержащиеся хотя бы в одном из обслуживаемых алфавитов;

- цифры;
- знаки пунктуации;
- специальные знаки (математические, технические, идеограммы и пр.);
- разделители.

Юникод — это система для линейного представления текста. Символы, имеющие дополнительные надстрочные или подстрочные элементы, представляются в виде последовательности кодов, составленной по определённым правилам (декомпозированный вариант) или единого символа (композированный вариант).

Комбинируемые символы.



Представление символа «Й» (U+0419) в виде базового символа «И» (U+0418) и комбинируемого символа «̆» (U+0306).

Графические символы в Юникод подразделяются на протяжённые и непротяжённые (бесширинные). Непротяженные символы при отображении не занимают места в строке. К ним относятся ударения, диакритические знаки и т. п. При кодировании в Юникоде, как протяжённые, так и непротяжённые символы имеют собственные коды. Протяжённые символы иначе называются базовыми, а непротяжённые — комбинируемыми, потому что они не могут встречаться самостоятельно. Например, символ «á» будет представлен как последовательность базового символа «a» (U+0061) и комбинируемого символа «́» (U+0301) или как отдельный символ «á» (U+00C1).

Представленные символы.

Юникод включает практически все современные письменности, в том числе: арабскую, армянскую, бенгальскую, бирманскую, греческую, грузинскую, деванагари, иврит, кириллицу, коптскую, кхмерскую, латинскую, тамильскую, хангыль, хань (Китай, Япония, Корея), чероки, эфиопскую, японскую (катакана, хирагана, кандзи) и другие.

С академической целью добавлены многие исторические письменности, в том числе: древнегреческая, египетские иероглифы, клинопись, письменность майя, этрусский алфавит.
В Юникоде представлен широкий набор математических и музыкальных символов, а также пиктограмм.

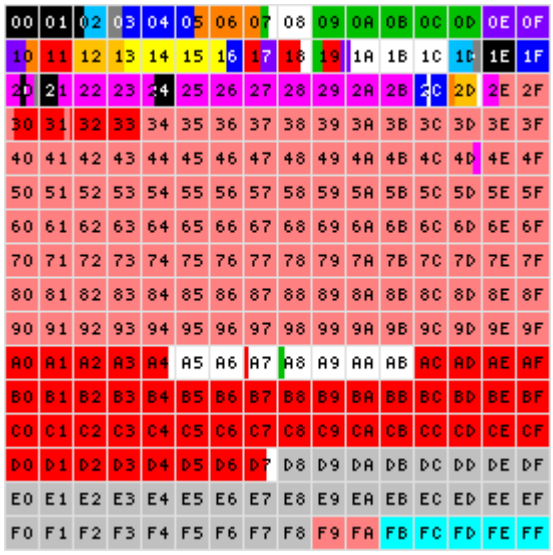
Все символы, представленные в Юникоде.

В Юникоде зарезервировано 1 114 112 (= $2^{20} + 2^{16}$) позиций символов, из которых сейчас используется около 90000. Первые 256 знакомест совпадают с кодовой таблицей **ISO 8859-1**(Latin-1).
Хоть формы записи **UTF-8** и UTF-32 позволяют кодировать до 2^{31} (2 147 483 648) кодовых позиций, принято решение использовать лишь $2^{20}+2^{16}$ (1 114 112) для совместимости с **UTF-16**. Впрочем, даже и этого более чем достаточно — на сегодняшний день используется чуть больше 96 000 кодовых позиций.

Кодовое пространство разделено на 17 «плоскостей» по 65536 (= 2^{16}) символов. Нулевая плоскость называется **базовой**, в ней расположены символы наиболее употребительных письменностей. Первая плоскость используется, в основном, для исторических письменностей. Плоскости 16 и 17 выделены для частного употребления.

Для обозначения символов Unicode используется запись вида «U+xxxx» или «U+yyyyyyyy», где xxxx и yyyyyyyy — шестнадцатеричная запись номера символа. Например, символ «я» (U+044F) имеет код $044F_{16} = 1103_{10}$.

Базовая многоязыковая плоскость.



Базовая плоскость UNICODE:

Чёрный	— расширенный латинский алфавит;
Голубой	— лингвистические символы международного фонетического алфавита IPA;
Синий	— другие европейские алфавиты;
Оранжевый	— письменности Ближнего Востока;
Светло-оранжевый	— письменности Африки;
Зелёный	— письменности Южной Азии;
Фиолетовый	— письменности Юго-восточной Азии;
Красный	— письменности Восточной Азии;
Розовый	— унифицированные китайско-японско-корейские символы;
Жёлтый	— письменности аборигенов Северной Америки;
Пурпурный	— символы;
Тёмно-серый	— диакритики;
Светло-серый	— суррогатные пары UTF-16 и области для частного использования;
Сине-зелёный	— другие знаки;
Белый	— не используется.

Плоскость 0 (Основная многоязыковая плоскость, англ. Basic Multilanguage Plane, **BMP**) содержит символы практически для всех современных письменностей и большое число специальных символов. Большая часть таблицы занята китайско-японско-корейскими иероглифами.

О том, какие символы представлены в основной плоскости в **Unicode 4.1**, можно посмотретьздесь.

Дополнительная многоязыковая плоскость.

Плоскость 1 (дополнительная многоязыковая плоскость, англ. Supplementary Multilingual Plane,**SMP**) отведена, в первую очередь, для исторических письменностей, но также включает музыкальные и математические символы.

Частная область.

Некоторые регионы Unicode выделены для частного использования и экспериментов.
Частная область включает:

- Регион в Базовой плоскости U+E000…U+F8FF
- Расширенные плоскости 15 (U+F0000…U+FFFFF) и 16 (U+100000…U+10FFFF)

Проблемы Юникода.

Как любая изобретённая человеком система, Юникод не свободен от недостатков.

- Многие системы письма всё ещё не представлены в Юникоде. Например, письменность церковнославянского языка содержит много дополнительных графических элементов (такие как титлы и надстрочные буквы). Они не могут быть должным образом представлены в системе Юникод, хотя отдельные элементы для этого имеются. Изображение «длинных» надстрочных символов, простирающихся над несколькими буквами, пока в принципе не предусмотрено.
- Тексты на китайском, корейском и японском языке имеют традиционное написание сверху вниз, начиная с правого верхнего угла. Данная возможность не отражена в Юникоде (впрочем, она и не должна быть

отражена, поскольку это относится к форматированию текста, а не к кодированию символов).

- В стандартах Юникода не было зафиксировано, когда вводится отдельная кодовая позиция для готового (Precomposed) символа, а когда его необходимо набирать из базового и диакритического. Например, русские буквы **Ё** (U+0401) и **Й** (U+0419) существуют в виде отдельных символов, хотя могут быть представлены и набором базового символа плюс диакритика (Decomposed): Е+¨ (U+0415 U+0308), И+˘ (U+0418 U+0306). В то же время, множество символов из языков с алфавитами на основекириллицы не имеют precomposed форм.
- Юникод предусматривает возможность разных начертаний одного и того же символа в зависимости от языка. Так, китайские иероглифы могут иметь разные начертания в китайском, японском (кандзи) и корейском (ханджа), но при этом в Юникоде обозначаться одним и тем же символом (так называемая CJK-унификация), хотя упрощённые и полные иероглифы всё же имеют разные коды. Часто возникают накладки, когда, например, японский текст выглядит «по-китайски». Аналогично, русский и сербский языки используют разное начертание курсивных букв *п* и *т* (в сербском они выглядят как *и* и *ш*). Поэтому нужно следить, чтобы текст всегда был правильно помечен как относящийся к тому или другому языку.
- Файлы с текстом в Юникоде занимают больше места в памяти, так как один символ кодируется не одним байтом, как в различных национальных кодировках, а последовательностью байтов (исключение составляет UTF-8 для языков алфавит которых укладывается в ASCII). Однако с увеличением мощности компьютерных систем и удешевлением памяти и дискового пространства эта проблема становится всё менее существенной.
- Хотя поддержка Юникода реализована в наиболее распространённых операционных системах, не всё прикладное программное обеспечение поддерживает корректную работу с ним. В частности, не всегда обрабатываются метки BOM и плохо поддерживаются диакритические символы. Проблема является временной

и есть следствие сравнительной новизны стандартов Юникода (в сравнении с однобайтовыми национальными кодировками).

Версии юникода.

По мере изменения и пополнения таблицы символов системы Юникода и выхода новых версий этой системы — а эта работа ведётся постоянно, поскольку изначально система Юникод была представлена в **ISO** в недоработанном виде — выходят и новые документы ISO. Система Юникод существует в общей сложности в следующих версиях:

- 1.1 (соответствует стандарту ISO/IEC 10646—1:1993),
- 2.0, 2.1 (тот же стандарт ISO/IEC 10646—1:1993 плюс дополнения: «Amendments» с 1-го по 7-е и «Technical Corrigenda» 1 и 2),
- 3.0 (стандарт ISO/IEC 10646—1:2000).
- 3.2 (стандарт 2002 года)
- 4.0 (стандарт 2003)
- 4.01 (стандарт 2004)
- 4.1 (стандарт 2005)
- 5.0 (стандарт 2006)

ISO/IEC 10646

Консорциум Юникода работает в тесной связи с рабочей группой **ISO/IEC/JTC1/SC2/WG2**, которая занимается разработкой международного стандарта **10646** (ISO/IEC 10646). Между стандартом Юникода и ISO/IEC 10646 установлена синхронизация, хотя каждый стандарт использует свою терминологию и систему документации.

Сотрудничество Консорциума Юникода с Международной организацией по стандартизации (англ. International Organization for Standardization, **ISO**) началось в 1991 году. В 1993 году ISO выпустила стандарт DIS 10646.1. Для синхронизации с ним, Консорциум утвердил стандарт Юникода версии 1.1, в который были внесены дополнительные символы из DIS 10646.1. В результате, значения закодированных символов в Unicode 1.1 и DIS 10646.1 полностью совпали.

В дальнейшем сотрудничество двух организаций продолжилось. В 2000 году стандарт Unicode 3.0 был синхронизирован с ISO/IEC 10646-1:2000. Предстоящая третья версия ISO/IEC 10646 будет синхронизирована с Unicode 4.0. Возможно, эти спецификации даже будут опубликованы как единый стандарт.

Аналогично форматам UTF-16 и UTF-32 в стандарте Юникода, стандарт ISO/IEC 10646 также имеет две основные формы кодирования символов: **UCS-2** (2 байта на символ, аналогично UTF-16) и **UCS-4** (4 байта на символ, аналогично UTF-32). UCS значит **универсальный многооктетный** (многобайтовый) **кодированный набор символов** (англ. Universal Multiple-Octet Coded Character Set). Как уже упоминалось, UCS-2 можно считать подмножеством UTF-16 (UTF-16 без суррогатных пар), а UCS-4 является синонимом для UTF-32.

Управляющие символы Unicode

Дополнительную информацию по управляющим символам Юникода вы можете посмотреть здесь.