Універсальний набір символів (UCS) - це стандартний набір символів, визначений Міжнародним стандартом ISO / IEC 10646, який є основою багатьох кодувань символів . Остання версія містить понад 136 000 абстрактних символів, кожен з яких ідентифікується однозначним ім’ям і цілим числом, що називається кодовою точкою. Цей стандарт ISO / IEC 10646 підтримується спільно зі стандартом Unicode ("Юнікод"), і вони ідентичні коду для коду.

Символи (букви, цифри, символи, ідеограми, логограми тощо) з багатьох мов, манускриптів та традицій світу представлені в UCS з унікальними кодовими точками. Інклюзивність UCS постійно покращується, оскільки додаються символи раніше непредставлених систем письма.

UCS має понад 1,1 мільйона можливих кодових точок, доступних для використання / розподілу, але лише перші 65 536 (Базова багатомовна площина(Basic Multilingual Plane, BMP)) увійшли до загального користування до 2000 року. Ситуація почала змінюватися, коли Китайська Народна Республіка (КНР) прийняла рішення у 2006 р., що все програмне забезпечення, що продається під їхньою юрисдикцією, повинне підтримувати GB 18030. Для цього потрібно було перенести програмне забезпечення, призначене для продажу в КНР, за межі BMP.

Система свідомо залишає багато кодових точок не заповненими навіть у BMP для майбутнього розширення набору і для того, щоб звести до мінімуму конфлікти з іншими формами кодування.

Кодування форм

ISO / IEC 10646 визначає кілька *форм кодування* символів для Універсального кодованого набору символів. Найпростіший, UCS-2, використовує одне значення коду (визначене як число, з яких одне чи більше представляє кодова точка загалом, але для UCS-2 це строго одне значення коду, яке представляє кодова точка ) від 0 до 65 535 для кожного символу і дозволяє рівно двом байтам (одне 16-бітове слово) представляти це значення. Таким чином, UCS-2 дозволяє двійкове представлення кожної точки коду в BMP, що представляє символ. UCS-2 не може представляти кодові точки поза BMP.

Перша поправка до оригінальної редакції UCS визначила UTF-16, розширення UCS-2, для представлення кодових точок за межами BMP. Діапазон кодових точок у S (спеціальній) зоні BMP залишається неприсвоєним символам. UCS-2 забороняє використання значень коду для цих кодових точок, але UTF-16 дозволяє використовувати їх парами. Unicode також прийняв UTF-16, але в його термінології елементи зони високої половини стають "високими сурогатами", а елементи зони низької половини - "низькими сурогатами".

Інше кодування, UCS-4, використовує чотири байти (загалом 32 біти) для кодування одного символу простору кодів. Навіть незважаючи на те, що Unicode обмежує простір коду верхньою межею 10FFFF. Стандарт ISO / IEC 10646 заявив, що всі майбутні призначення символів будуть виконуватися в діапазоні до 0x7FFFFFFFhex. UCS-4 дозволяє представляти кожне значення як рівно чотири байти (одне 32-бітове слово). Таким чином, UCS-4 дозволяє двійкове представлення кожної кодової точки в UCS, в тому числі поза BMP. Як і в UCS-2, кожен закодований символ має фіксовану довжину в байтах, що спрощує маніпуляції, але, звичайно, він вимагає вдвічі більше місця для зберігання, ніж UCS-2.

Зараз домінуючим кодуванням UCS є UTF-8, яке є кодуванням змінної ширини, призначеним для зворотної сумісності з ASCII, а також для уникнення ускладнень мінливості та позначок порядку байтів в UTF-16 та UTF-32. Понад 93% усіх веб-сторінок закодовано в UTF-8. Робоча група Інтернет-інженерії (IETF) вимагає від усіх протоколів Інтернету ідентифікації кодування, що використовується для даних символів, а підтримувані кодування символів повинні включати UTF-8. Інтернет-поштовий консорціум (IMC) рекомендує, щоб усі програми електронної пошти мали можливість відображати та створювати пошту за допомогою UTF-8. Він також все частіше використовується як кодування символів за замовчуванням в операційних системах, мовах програмування, API та програмних додатках.

Історія

Міжнародна організація зі стандартизації (ISO) мала намір скласти універсальний символ, встановлений у 1989 р., і опублікувала проект ISO 10646 у 1990 р.. Х'ю Макгрегор Росс був одним з головних архітекторів. Цей стандарт помітно відрізнявся від поточного. Він визначав:

128 груп

256 площин

256 рядків

256 комірок,

для загальної суми 21474834848 символів, але насправді стандарт міг кодувати лише 679477248 символів, оскільки політика забороняла значення байтів контрольних кодів C0 та C1 (від 0x00 до 0x1F та 0x80 до 0x9F, у шістнадцяткових позначеннях) у будь-якому з чотирьох байтів із зазначенням групи, площини, рядка та комірки. Наприклад, латинська велика літера A мала розташування в групі 0x20, площині 0x20, рядку 0x20, комірці 0x41.

Можна кодувати символи цього первинного стандарту ISO 10646 одним із трьох способів:

1. UCS-4, чотири байти для кожного символу, що забезпечує просте кодування всіх символів;
2. UCS-2, два байти для кожного символу, що дозволяє кодувати першу площину, 0x20, Базову багатомовну площину, що містить перші 36 864 кодових точок прямолінійно та інші площини та групи, переключившись на них із послідовностями виходу з ISO 2022;
3. UTF-1, який кодує всі символи в послідовності байтів різної довжини (від 1 до 5 байт, кожен з яких не містить контрольних кодів).

Тому в 1990 році існували дві ініціативи щодо універсального набору символів: Unicode з 16 бітами для кожного символу (65 536 можливих символів) та ISO 10646. Компанії, що виробляють програмне забезпечення, відмовились прийняти вимогу щодо складності та розміру стандарту ISO і змогли переконати низку національних органів ISO проголосувати проти. Стандартизатори ISO зрозуміли, що не можуть продовжувати підтримувати стандарт у його поточному стані і домовились про уніфікацію свого стандарту з Unicode. Відбулися дві зміни: скасування обмеження на символи (заборона значень контрольного коду), таким чином відкриваючи кодові точки, такі як 0x0000101F для розподілу і синхронізація репертуару Базового багатомовного плану з Unicode.

Тим часом із плином часу ситуація змінилася і в самому стандарті Unicode: 65 536 символів виявилося недостатнім, а стандарт з версії 2.0 і далі підтримує кодування 1112 064 кодових точок з 17 площин за допомогою сурогатного механізму UTF-16 . З цієї причини ISO 10646 містив стільки символів, скільки може кодувати UTF-16, і не більше, тобто, трохи більше мільйона символів, замість понад 679 мільйонів. Кодування UCS-4 за стандартом ISO 10646 було включено до стандарту Unicode з обмеженням до діапазону UTF-16 та під назвою UTF-32, хоча воно майже не використовує поза внутрішніми даними програм.

Роб Пайк і Кен Томпсон, розробники операційної системи Plan 9, розробили нове, швидке та добре розроблене кодування змішаної ширини UTF-8, яке нині найпопулярніше кодування UCS.

Відмінності від Unicode

ISO 10646 та Unicode мають однакове кодування та номери - однакові символи з однаковими номерами існують в обох стандартах, хоча Unicode випускає нові версії та додає нові символи частіше. Unicode має правила та специфікації, що виходять за рамки ISO 10646. ISO 10646 - це проста карта символів, розширення попередніх стандартів, таких як ISO 8859. На відміну від цього, Unicode додає правила зіставлення, нормалізації форм та двонаправленого алгоритму для рукописів справа наліво, таких як арабський та іврит. Для взаємодії між платформами, особливо якщо використовуються двонаправлені рукописи, недостатньо підтримувати ISO 10646; Необхідно реалізувати Unicode.

Для підтримки цих правил та алгоритмів Unicode додає багато властивостей до кожного символу в наборі, таких як властивості, що визначають стандартний двонаправлений клас символу та властивості, щоб визначити, як символ поєднується з іншими символами. Якщо символ представляє числове значення, таке як європейське число «8» або звичайний дріб «¼», то це числове значення також додається як властивість символу. Unicode має на меті підтримувати властивості взаємодії обробки тексту із сумішшю мов.

Деякі програми підтримують символи ISO 10646, але не повністю підтримують Unicode. Одне з таких додатків, Xterm, може належним чином відображати всі символи ISO 10646, які мають індивідуальне відображення символів у гліфи та єдину спрямованість. Він може обробляти деякі комбінуючі знаки простими методами, що вражають, але не може відображати іврит (двонаправлений), деванагарі (один символ на багато гліфів) або арабський (обидві функції). Більшість програм графічного інтерфейсу використовують стандартні процедури малювання тексту ОС, які обробляють такі сценарії, хоча самі програми все ще не завжди обробляють їх правильно.

Посилання на Універсальний набір символів

ISO 10646 - загальне, неформальне цитування для сімейства стандартів ISO / IEC 10646 є прийнятним у більшості прозаїчно. І хоча це окремий стандарт, термін Unicode часто використовується так само неофіційно під час обговорення UCS. Однак, будь-які нормативні посилання на UCS , як публікацію, повинні вказувати рік видання у формі ISO / IEC 10646: {рік}, наприклад: ISO / IEC 10646: 2014.

Відносини з Unicode

З 1991 року Консорціум Юнікод та ISO розробили The Unicode Standard ("Unicode") та ISO / IEC 10646 в тандемі. Репертуар, імена символів та кодові точки Unicode Версії 2.0 точно відповідають вимогам ISO / IEC 10646-1: 1993 з першими сімома опублікованими поправками. Після публікації Unicode 3.0 у лютому 2000 р. Відповідні нові та оновлені символи увійшли до UCS через ISO / IEC 10646-1: 2000. У 2003 році частини 1 та 2 ISO / IEC 10646 були об'єднані в одну частину, яка з тих пір мала ряд поправок, що додають символи до стандарту приблизно в синхронній ситуації зі стандартом Unicode.

ISO / IEC 10646-1: 1993 = Unicode 1.1

ISO / IEC 10646-1: 1993 плюс поправки 5-7 = Unicode 2.0

ISO / IEC 10646-1: 1993 плюс поправки 5-7 = Unicode 2.1, за винятком символу євро та символу заміни об'єкта, які включені до Поправки 18

ISO / IEC 10646-1: 2000 = Unicode 3.0

ISO / IEC 10646-1: 2000 та ISO / IEC 10646-2: 2001 = Unicode 3.1

ISO / IEC 10646-1: 2000 плюс поправка 1 та ISO / IEC 10646-2: 2001 = Unicode 3.2

ISO / IEC 10646: 2003 = Unicode 4.0

ISO / IEC 10646: 2003 плюс поправка 1 = Unicode 4.1

ISO / IEC 10646: 2003 плюс поправки 1 до 2 = Unicode 5.0, виключаючи літери Деванагарі GGA, JJA, DDDA та BBA, які включені до поправки 3

ISO / IEC 10646: 2003 плюс поправки від 1 до 4 = Unicode 5.1

ISO / IEC 10646: 2003 плюс поправки від 1 до 6 = Unicode 5.2

ISO / IEC 10646: 2003 плюс поправки 1 до 8 = ISO / IEC 10646: 2011 = Unicode 6.0, за винятком знаку індійської рупії

ISO / IEC 10646: 2012 = Unicode 6.1

ISO / IEC 10646: 2012 = Unicode 6.2, за винятком знака турецької ліри, який включений до поправки 1

ISO / IEC 10646: 2012 = Unicode 6.3, за винятком знака турецької ліри, який входить до поправки 1 та п’яти двонаправлених контрольних символів (арабська літерна позначка, ізоляція зліва направо, ізоляція справа наліво, перший сильний ізолятор, Pop Directional Isolate), які включені до поправки 2

ISO / IEC 10646: 2012 плюс поправки 1 та 2 = Unicode 7.0, за винятком знаку рубля

ISO / IEC 10646: 2014 плюс поправка 1 = Unicode 8.0, за винятком знака ларі, дев'яти уніфікованих ідеографів CJK та 41 символу смайликів

ISO / IEC 10646: 2014 плюс поправки 1 та 2 = Unicode 9.0, за винятком Adlam, Newa, японських символів телебачення та 74 смайлів та символів

ISO / IEC 10646: 2017 = Unicode 10.0, за винятком 285 символів Хентаігани, 3 символів квадрата Занабазара та 56 символів смайлів