Projekt UnicodeBuild

Implementacja i wyniki

Jarosław Kuchta

4.09.2025

Szczegóły implementacyjne są opisane w pliku Readme.md. W tym dokumencie opisano natomiast kluczowe zagadnienia implementacji i przedstawiono wyniki pracy.

# Kluczowe zbiory danych

Aplikacja przetwarza dane z bazy danych Unicode2.accdb obsługiwanej przez silnik MS Access. Jest to kopia lokalna bazy Unicode.accdb stworzona w celu ochrony podstawowej bazy danych przed przypadkowym zniszczeniem. Kluczowe są cztery zbiory danych:

* UcdCodePoints – zawierający dane poszczególnych punktów kodowych Unicode,
* UcdBlocks – zawierający definicje bloków danych Unicode,
* WritingSystems – zawierający definicje systemów pisma i związanych z nimi jednostek organizacyjnych.
* UcdCodePoints-WritingSystems – zawierający przypisania systemów pisma do punktów kodowych.

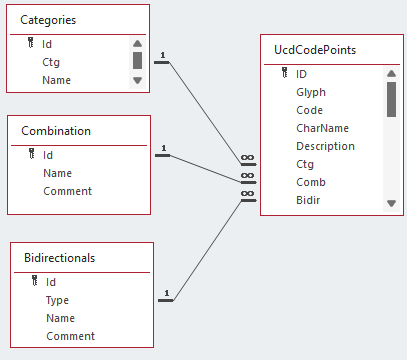
## UcdCodePoints

Zbiór danych UcdCodePoints zawiera kopię danych z pliku Unicode.txt opublikowanego przez konsorcjum. Plik ten został sparsowany i wczytany do tablicy bazy danych aplikacji MS Access. Poszczególne rekordy tablicy mają klucz główny ID typu integer, który zawiera wartości dziesiętne z kolumny z kodami heksadecymalnymi punktów kodowych. Dzięki temu punkty kodowe są przedstawiane w kolejności wartości numerycznej pola Code, a nie alfabetycznej.

Na podstawie wartości numerycznej punktów kodowych wygenerowano kolumnę Glyph, która umożliwia podgląd znaku w domyślnym kroju pisma (aktualnie jest to Segoe UI). Glyph przeważnie zawiera pojedynczy znak Unicode odpowiadający punktowi kodowemu. W przypadku znaków diakrytycznych, które normalnie są dodawane nad lub pod poprzedzającym znakiem i mają ujemny offset renderowania, Glyph zawiera dwa znaki: znak kółka przerywanego (o kodzie U+25CC) i dany znak diakrytyczny.

Dodano kolumnę CharName, która ma zawierać wynikowe nazwy znaków.

Dla każdej kolumny, która może zawierać jedynie ściśle określone wartości (jak np. kolumna Ctg zawierająca dwuliterowe kody kategorii znaków), została utworzona pomocnicza tabela z kluczem głównym jednobajtowym o kolejnych wartościach oraz z nazwą wartości dozwolonej. Sama kolumna stała się kluczem obcym do tabeli pomocniczej. Utworzono relację o wymaganej spójności wewnętrznej między tą kolumną a tabelą pomocniczą (rys. 1). Później w kodzie aplikacji zdefiniowano typ wyliczeniowy o nazwach i wartościach z tej tabeli pomocniczej. Przy wczytywaniu i zapisywaniu danych z wykorzystaniem mechanizmu Entity Framework, wartość klucza obcego jest przetwarzana na wartość wyliczaną i z powrotem.



Rys. . Tabele pomocnicze dla typów wyliczeniowych UcdCodePoints

Kolumnę Mirr, która może zawierać literę ‘Y’ lub ‘N’ przetworzono w kodzie aplikacji na typ Boolean.

Do tabeli UcdCodePoints dodano kolumnę na klucz obcy do tabeli UcdBlocks. Klucz ten przyspiesza powiązanie znaków Unicode do bloków, gdyż dekodowanie zakresu bloku następuje tylko raz. Reasumując tabela UcdCodePoints zawiera następujące kolumny:

* ID: integer – zawiera wartość liczbową kodu danego znaku.
* Glyph: string[2] – zawiera prezentację danego znaku,
* Code: string[6] – zawiera kod szesnastkowy danego znaku (min. 4 cyfry),
* CharName: string – przeznaczone na wygenerowaną nazwę znaku,
* Description: string – opisowa nazwa znaku,
* Ctg: string[2] – dwuznakowy kod kategorii znaku (np. Lu, Ll, Mn), zawiera klucz obcy do pola Ctg w tabeli Categories,
* Comb: byte – zawiera klucz obcy do tabeli Combination określającej sposób łączenia znaku z innymi znakami,
* Bidir: string[3] – zawiera trójznakowy klucz obcy do tabeli Bidirectionals określającej dwukierunkowość znaku,
* Decomposition: string – zawiera opcjonalny łańcuch możliwej dekompozycji znaku złożonego na znaki składowe (składnia określona przez konsorcjum Unicode),
* DecDigitVal: string[1] – zawiera opcjonalną cyfrę odpowiadającej wartości dziesiętnej znaku cyfry,
* DigitVal: string[1] – zawiera opcjonalną cyfrę odpowiadającą temu znakowi (to samo do DecDigitVal, ale ma też wartości dla cyfr w indeksie górnym i dolnym),
* NumVal: string[15] – zawiera łańcuch znaków wyrażający wartość danego znaku (może to być liczba kilkucyfrowa lub wyrażenie z kreską ułamkową),
* Mirr: string[1] – zawiera literę ‘Y” oznaczającą, że dany znak w piśmie od prawej do lewej zmienia się na swoje lustrzane odbicie, albo ‘N’ – w przeciwnym wypadku,
* OldDescription: string[255] – przy niektórych znakach zawiera stary opis znaku (z wcześniejszych wersji standardu Unicode),
* Comment – przy niekrórych znakach zawiera dodatkowy komentarz,
* Upper: string[6] – przy literach małych zawiera szesnastkowy kod odpowiedniej litery wielkiej,
* Lower: string[6] – przy literach wielkich zawiera szesnastkowy kod odpowiedniej litery małej,
* Title: string[6] – przy części liter zawiera szesnastkowy kod odpowiedniej litery stosowanej w tytułach,
* Block: integer – zawiera klucz obcy do tabeli UcdBlocks.

Jak się okazało w czasie pracy badawczej, nie wszystkie pola zostały wykorzystane.

## UcdBlocks

Tabela UcdBlocks zawiera definicje bloków znakowych publikowanych w pliku Blocks.txt przez konsorcjum Unicode uzupełnione o informacje wczytane z angielskiej Wikipedii.

Plik Blocks.txt zawiera tylko dwie kolumny: Range i BlockName, które zostały wczytane do tabeli bazy danych Unicode. Dodano do nich klucz główny w postaci automatycznie wypełnianej wartości całkowitej. Dodano też pola Start i End (oba typu całkowitego), które zawierają wartości dziesiętne początku i końca bloku (odczytane z pola Range).

Do tego dodano pola Comment, Scripts, Major alphabets, Symbol sets, Size, Assigned i Plane odczytane z Wikipedii. Ostatnie trzy pola nie są dalej przetwarzane. Pola Scripts, Major alphabets i SymbolSets są używane do określenia głównego systemu pisma, do którego przypisana będzie większość punktów kodowych w danym bloku. Klucz obcy do tabeli WritingSystems jest wpisywany do pola WritingSystemID.

Powiązanie tabeli UcdBlocks z tabelą WritingSystems sprawia dwa problemy:

* Nazwa bloku często nie odpowiada nazwie żadnego systemu pisma, a wiele bloków może się odwoływać do jednego systemu pisma,
* Jeden blok może odwoływać się do wielu systemów pisma. Pole Scripts może zawierać kilka nazw systemów pisma oddzielonych przecinkami (z podaną liczbą odwołań do każdego z nich).

Do przypisania utworzono plik Block-WritingSystem.txt, który zawiera dwie kolumny: Range i WritingSystemName.

Zbiór UcdBlocks pełni funkcję pomocniczą. Przypisanie systemów pisma do zbioru UcdCodePoints następuje przez plik Scripts.txt.

Reasumując tabela UcdBlocks zawiera następujące kolumny:

* ID: autonumber – automatycznie tworzony klucz własny encji,
* Range: string[14] – łańcuch określający zakres bloku w formacie XXXX..XXXX,
* BlockName: string – nazwa bloku,
* Start: integer – wartość kodowa początkowego znaku w bloku,
* End: integer – wartość kodowa końcowego znaku w bloku,
* WritingSystemID: integer – klucz obcy do tabeli WritingSystems określający domyślny system pisma,
* Comment: text – opis bloku z Wikipedii,
* Scripts: string – łańcuch znaków określający pisma,
* Major alphabets: string – łańcuch znaków określający główne alfabety,
* Symbol sets: string – łańcuch znaków określający główne systemy pisma,
* Size: integer – rozmiar bloku (liczba pozycji w bloku),
* Assigned: integer – liczba zajętych pozycji w bloku,
* Plane: byte – numer planu Unicode (od 1 do 16)

W tym zbiorze również nie wszystkie pola zostały wykorzystane.

## WritingSystems

Tabela WritingSystem zawiera wszystkie systemy pisma. Pod tym pojęciem w aplikacji rozumie się nie tylko pismo i notację, ale też inne jednostki klasyfikujące znaki. Rekordy mają pole Type, które określa typ encji. Typy mogą być następujące:

* obszar (area) – zakres znaków, w którym poszczególne znaki nie są definiowane indywidualnie; zamiast tego Unicode jedynie definiuje początek i koniec obszaru.
* pismo (script) – zbiór znaków do zapisu języka naturalnego.
* język (language) – pismo może mieć kilka podzbiorów znaków o tej samej nazwie i roli, ale różnych glifach w zależności od języków, w których są stosowane.
* notacja (notation) – zbiór znaków stosowanych w określonej dziedzinie nauki lub techniki. Ma określone reguły użycia symboli.
* zbiór symboli (symbolSet) – zbiór znaków stosowanych w różnych dziedzinach; reguł użycia tych symboli nie ma lub są jedynie zaleceniami.
* podzbiór (subset) – wybrany podzbiór ze zbioru znaków określanego przez skrypt, notację lub zbiór symboli, może występować w wielu zbiorach.

Rekordy mają też pole Kind, które pozwala sprecyzować rodzaj w danym typie. Wyróżnia się:

* rodzaje pism:
  + alfabet – system pisma oparty na standardowym zestawie liter zapisywanych w celu reprezentowania określonych dźwięków w języku mówionym.
  + półsylabariusz – system pisma, który zachowuje się częściowo jak alfabet, a częściowo jak sylabariusz.
  + sylabariusz – system pisma oparty na zestawie symboli pisanych, które reprezentują sylaby, z których składają się słowa.
  + abugida – segmentowy system pisma, w którym sekwencje spółgłoska-samogłoska są zapisywane jako jednostki; każda jednostka jest oparta na literze spółgłoski, a notacja samogłosek jest drugorzędna, podobnie jak znak diakrytyczny.
  + abjad – System pisma, w którym reprezentowane są tylko spółgłoski, pozostawiając dźwięki samogłosek do wywnioskowania przez czytelnika.
  + logograficzne – system pisma oparty na zestawie znaków pisanych, który reprezentuje semantyczny składnik języka, taki jak słowo lub morfem.
  + ideograficzne – system pisma składający się z symboli, który reprezentuje idee lub koncepcje niezależne od określonego języka.
  + piktograficzne – system pisma oparty na zestawie symboli graficznych, które przekazują znaczenie poprzez wizualne podobieństwo do obiektów fizycznych.
  + piktofonetyczne – system pisma oparty na jednostkach składał się z elementów fonetycznych i semantycznych.
  + mieszany – system pisma, który ma mieszane cechy systemu logograficznego, sylabariusza i pisma fonetycznego.
  + hieroglificzne– starożytny system pisma, który łączy elementy ideograficzne, logograficzne, sylabiczne i alfabetyczne, z ponad 1,000 odrębnymi znakami.
  + klinowe – starożytny logo-sylabiczny system pisma, który był używany do zapisu kilku języków starożytnego Bliskiego Wschodu, nazwany tak od charakterystycznych odcisków w kształcie klina, które tworzą ich znaki.
  + inne – inne pismo, które nie kwalifikuje się do żadnego z powyższych rodzajów.
* rodzaje notacji:
  + matematyczna – dla formuł i wyrażeń matematycznych.
  + muzyczna – do wizualnej reprezentacji muzyki.
  + numeryczna – do zapisu liczb.
  + fonetyczna – wyrażająca dźwięki wymowy.
  + stenograficzna – skrócona metoda pisania symbolicznego, która zwiększa szybkość i zwięzłość pisania w porównaniu, bardziej powszechną metodą pisania języka.
  + metryczna – długości sylab w poezji (zwłaszcza łacińskiej i greckiej).
  + gestykularna – do zapisu języka gestów (dla osób niesłyszących)
  + ambliofatyczna – dla osób niedowidzących i niewidomych.
  + programistyczna – dla algorytmów komputerowych.
* rodzaje zbioru znaków:
  + sterujące – zestaw znaków sterujących przepływem tekstu.
  + graficzne – zestaw kształtów graficznych do użytku indywidualnego lub używany do konstruowania rysunków opartych na siatce.
  + symboliczne – zestaw symboli z określonej dziedziny.
  + niewidoczne – zestaw punktów kodu, które nie mają wizualnej prezentacji.
  + kodujące – zestaw znaków do kodowania innych znaków.
  + strukturalne – zestaw symboli, które odgrywają określoną rolę w systemie pisma (samogłoski, spółgłoski, sylaby, ligatury).
  + formy – różne formy prezentacyjne znaków.
* Podzbiory mogą określać:
  + kontekst – specjalną formę znaku zależną od jego kontekstu (czy występuje na początku, w środku, czy na końcu wyrazu),
  + szerokość – specjalną formę znaku o określonej szerokości (wąskie, szerokie)
  + rozmiar – rozmiar znaku (wielkie litery, małe litery, małe wielkie litery, małe znaki)
  + czcionkę – wybór czcionki o określonych cechach, np. wytłuszczenie, pochylenie

Niektóre rekordy zostały oznaczone jako grupy. To znaczy, że mają one bardziej ogólne znaczenie i umożliwiają grupowanie innych rekordów. Encje składowe grupy mają wypełnione pole GroupID.

Inna hierarchia jest określana przez pole ParentID. Dla pism to pole oznacza system pisma, z którego dane pismo się wywodzi (informacja zebrana z Wikipedii) – ułatwia to grupowanie pism. Dla języków – może zawierać identyfikator pisma, przy którym dany język jest rozpoznawany. Dla zbioru znaków – może zawierać identyfikator nadrzędnego zbioru, w którym dany zbiór się zawiera.

Pole GroupID jest stosowane dla klasyfikacji systemów pisma przy wyświetlaniu ich w drzewie. Pole ParentID służy bardziej do analizy semantycznej systemów pisma.

Pole KeyPhrase jest używane do rozpoznawania powiązania punktów kodowych z danym systemem pisma (zbiorem znaków). Może zawierać jedno lub kilka słów, może rozpoczynać lub kończyć się gwiazdką. Rozpoznanie polega na poszukiwaniu danej frazy w nazwie opisowej znaku. Jeśli na końcu frazy jest gwiazdka, to fraza jest poszukiwana tylko na początku opisu. Jeśli gwiazdka jest na końcu frazy, to fraza musi występować na końcu opisu. Każda fraza kluczowa (jeśli występuje) musi być unikatowa, ale może być tak, że jedna fraza wchodzi w skład drugiej. Wówczas o przynależności decyduje dłuższa fraza.

Niektóre zbiory znaków mogą być rozpoznawane po kategorii znaków. Wówczas jest wykorzystywane pole Ctg, które może zawierać dwuliterowy kod kategorii. Na drugiej pozycji może wystąpić gwiazdka, co oznacza, że kwalifikowane są wszystkie kategorie, których kod rozpoczyna się od pierwszej litery pola Ctg.

Pisma i języki zarejestrowane w systemie ISO mają wypełniane pole ISO. Z reguły pismo ma identyfikator 4-znakowy, a język – 3-znakowy.

Pole Abbr służy do identyfikacji danego systemu pisma w nazwie wynikowej znaku. Nazwy wygenerowane przez aplikację muszą być unikatowe, ale mogą zawierać skróty identyfikujące systemy pisma. Sekwencje znaków należących do tego samego systemu mogą być skrócone do nazw własnych i poprzedzone skrótem systemu pisma. Jako podstawę przyjęto system ISO, ale nie wszystkie systemy są tam reprezentowane.

Reasumując tabela WritingSystems ma następujące kolumny:

* ID: autonumber – automatycznie ustalany identyfikator encji,
* Name: string – nazwa systemu pisma, notacji, zbioru lub podzbioru,
* Aliases: string – przy niektórych rekordach lista aliasów oddzielonych przecinkami,
* IsGroup: boolean – czy jest to grupa,
* Type: byte – typ encji, klucz obcy do tabeli WritingSystemTypes,
* Kind: byte – rodzaj w danym typie rekordu, klucz obcy do tabeli WritingSystemKinds,
* ParentID: integer – identyfikator nadrzędnego systemu pisma, klucz obcy do tej samej tabeli,
* GroupID: integer – identyfikator grupy, klucz obcy do tej samej tabeli,
* KeyPhrase: string – fraza kluczowa do rozpoznawania danego systemu pisma w nazwach opisowych punktów kodowych Unicode,
* Ctg: string[2] – kod kategorii znaków do alternatywnego przypisywania znaków do danego zbioru znaków,
* ISO: string[4] – identyfikator pisma lub języka w standardach ISO,
* Abbr: string[10] – skrót nazwy pisma, języka, notacji lub zbioru znaków w nazwie wynikowej,
* Description: text – opis systemu pisma odczytany z Wikipedii,
* RTL: boolean – czy pismo jest dwukierunkowe,
* Bicameral: boolean – czy pismo odróżnia wielkie i małe litery,
* Extinct: boolean – czy system pisma jest już wymarły,
* Invented: boolean – czy system pisma został sztucznie stworzony.

## UcdCodePoints-WritingSystems

Ponieważ relacje wiążące punkty kodowe Unicode z systemami pism są typu wiele-wiele, więc powiązania muszą być zapisane w osobnej tabeli.

Powiązania między punktami kodowymi i systemami pisma są rozpoznawane na podstawie analizy nazw opisowych punktów kodowych. Zapis wyników analizy do tabeli służy do sprawdzenia poprawności rozpoznawania fraz kluczowych.

W przypadku znaków złożonych, których nazwy opisowe zawierają słowo kluczowe WITH i ewentualnie dalej AND, frazy kluczowe poszukiwane są osobno w każdej części nazwy.

Reasumując tabela zawiera rekordy składające się z czterech pól:

* ID: autonumber – identyfikator wewnętrzny powiązania (dla przetwarzania przez model związków encji),
* Part: byte – numer części nazwy, 0 dla części początkowej, 1 dla części po WITH, 2 i kolejne dla części po AND,
* CodePoint: integer – zawiera klucz obcy do tabeli UcdCodePoints,
* WritingSystem: integer – zawiera klucz obcy do tabeli WritingSystems,

# Klasyfikacja systemów pisma

Dla łatwiejszego przeglądania systemów pisma stworzono drzewo systemów pisma Na najwyższym poziomie znajduje się sześć grup klasyfikujące typy systemów pisma:

* Obszary,
* Pisma,
* Języki,
* Notacje,
* Zbiory znaków,
* Podzbiory.

## Obszary

Obszary są trzy:

* Obszar prywatnego użycia (Private Use),
* Obszar kodów zastępczych (Surrogates),
* nie-znaki (Noncharacters).

## Pisma

Unicode odwołuje się do około 170 pism. Liczba ta jest nieprecyzyjna, bowiem część pism jest traktowanych jako języki. Pisma zostały podzielone w trakcie pracy badawczej na grupy obejmujące:

* Pisma antyczne (Ancient scripts). Są to pisma stosowane w starożytności. Stanowią podstawę dla pism współczesnych. Należą tu hieroglify egipskie, hieroglify hetyckie (inaczej anatolijskie), pismo klinowe, pisma liniowe A i B.
* Pisma euro-kaukaskie (Euro-Caucasian scripts). Do tej grupy należą alfabety pochodzące od pisma fenickiego: łacińskie, greckie, koptyjskie, cyrylica, armeńskie, gruzińskie. Należą też pisma wymarłe: staroitalskie, cypryjskie, głagolica etc.
* Pisma bliskowschodnie (Near East scripts). Należą tu starożytne pismo aramejskie i inne pisma stosowane na Bliskim Wschodzie: hebrajskie, arabskie, syryjskie.
* Pisma azjatyckie (Asian scripts). Ta wyjątkowo duża grupa zawiera pisma stosowane w Azji środkowej i południowo-wschodniej:
  + rodzinę pism brahmickich (Brahmi scripts) - do której należy starożytne pismo brahmickie i inne pochodzące od niego alfabety sylabiczne (abugidy),
  + grupę pism wynalezionych współcześnie (w XX wieku).
* Pisma mongolskie (Mongolian scripts) – stanowią odrębną rodzinę pism stosowanych dla języka mongolskiego,
* Pisma CJK (Chinese-Japan-Korea), które się dzielą na:
  + pisma chińskie,
  + pisma japońskie,
  + pisma koreańskie.
* Pisma afrykańskie (African scripts). Zawiera głównie pisma wynalezione w XX wieku dla języków afrykańskich i jedno starożytne pismo berberyjskie.
* Pisma północnoamerykańskie (Nort American scripts). Ta grupa zawiera pisma wynalezione współcześnie dla zapisywania języków ludów zamieszkujących Amerykę Północną.
* Inne pisma (Other scripts) – należą tutaj trzy pisma wynalezione współcześnie o stosowaniu na całym świecie lub w różnych regionach świata.

Do pism Unicode zalicza też dwa określenia:

* Common – na znaki stosowane w różnych pismach.
* Inherited – na pismo określane przez kontekst.

Nie są to właściwe pisma, ale w celu zapewnienia zgodności z Unicode, określenia te są przypisywane do znaków tak, jak pisma.

## Języki

W Unicode języki są traktowane osobno tylko wtedy, gdy w jakimś systemie pisma znajdują się znaki przeznaczone specjalnie do danego języka. Zdefiniowano 34 języki, wśród których są:

* Grupa języków dla pisma rdzennych Kanadyjczyków (Canadian Aboriginal Languages). Obejmuje ona kilkanaście pozycji, takich jak Blackfoot, Ojibway, Athapascan.
* Grupa języków mongolskich (Mongolian languages). Obejmuje ona cztery pozycje: Galik, Manchu, Todo, Xibe. Właściwie Todo jest opisywany w Wikipedii jako osobne pismo (o nazwie Clear Script), ale w Unicode jest to podzbiór pism mongolskich.
* Grupa trzech języków używających pisma arabskiego (Arabic languages). Są to: język ujgurski (Uighur), kirgiski (Kirghyz) i kazachski (Kazakh). Ludy mówiące tymi językami stosują też cyrylicę, ale w Unicode są one wyróżnione dla pisma arabskiego.
* Grupa języków pisma birmańskiego (Myanmar languages) obejmująca dwie pozycje: Western Pwo Karen i Eastern Pwo Karen.
* Grupa dwóch języków dla cyrylicy (Cyrillic languages). Są to język ukraiński (Ukrainian) i białoruski (Byelorussian). Język rosyjski ani inne języki wschodnioeuropejskie nie są wyróżnione w Unicode.
* Język Wizygotów (Visigothic) jako wyróżniony podzbiór pisma łacińskiego,
* Język Rohingja (Rohingya) jako wyróżniony podzbiór pisma arabskiego. Rohingjowie mają własne pismo Hanifi Rohingya, ale posługują się też pismem arabskim.
* Język Gurage jako wyróżniony podzbiór pisma etiopskiego.

## Notacje

Notacje podobnie jak pisma definiują zbiory znaków i mają reguły zapisu. Odróżnia je to, że nie wyrażają języków naturalnych, lecz pewne umowne sposoby zapisu w różnych dziedzinach. Wyróżnia się 29 notacji:

* notację matematyczną,
* grupę notacji numerycznych obejmującą systemy zapisu liczb i liczebników,
* grupę notacji muzycznych (np. Byzantine Musical Notation),
* grupę notacji metrycznych (do oznaczania metryki poezji lub śpiewu),
* grupę notacji fonetycznych (IPA, UPA),
* grupę notacji programistycznych (APL, Z-notation),
* notację stenograficzną (shorthand).
* notację do zapisu języka gestów (signwriting),
* notację Braille’a przeznaczoną dla osób niewidomych i niedowidzących.

## Zbiory znaków

Zbiory znaków (symbol sets) łączą znaki o pewnej wspólnej charakterystyce, ale bez określonych zasad stosowania. Zbiory znaków mogą być samodzielne, mogą być grupowane, mogą wchodzić w skład innych zbiorów znaków (w tym definiowanych przez pisma i notacje). Stosując różne kryteria klasyfikacji wyróżniono 110 zbiorów znaków podzielonych na:

* grupę strzałki i harpuny (Arrows and Harpoons). Grupa ta obejmuje:
  + strzałki (arrows),
  + groty strzałek (arrowheads),
  + brzechwy strzałek (arrow tails),
  + harpuny (harpoons) - harpun przypomina strzałkę, ale ma grot jednoostrzowy,
* grupę symboli technicznych i naukowych (Technical and science). Należą tu:
  + symbole alchemiczne (alchemic),
  + symbol astrologiczne (Astrological symbols),
  + znaki zodiaku (Zodiac symbols),
  + symbole elektroniczne
* grupę symboli stosowanych w grach (Gaming symbols):
  + w szachach (Chess symbols),
  + w domino (Domino Tiles),
  + w Mahjong (Mahjong Tiles)
  + w grach karcianych (Playing Card Suits),
  + w szachach chińskich (Xiangqi figures),
  + w kostkach do gry (Die faces),
  + w warcabach,
* grupę symboli ideograficznych systemu pisma CJK, do której należą:
  + zbiór ideogramów kompatybilnych z CJK (CJK Compatibility Ideographs),
  + zbiór elementów systemu ideograficznego CJK (CJK Radicals),
  + zbiór kresek systemu ideograficznego CJK (CJK Strokes),
  + zunifikowane ideogramy CJK w kwadratach (Squared CJK Unified Ideographs),
  + zunifikowane ideogramy CJK w nawiasach „żółwiowych” (Tortoise Shell Bracketed CJK Unified Ideographs),
  + inne symbole CJK (CJK symbols),
* grupę symboli stosowanych w piśmie koreańskim Hangul,
  + Choseong – nagłos sylaby (Hangul Choseong),
  + Jungseong – ośrodek sylaby (Hangul Jungseong),
  + Jongseong – wygłos sylaby (Hangul Jongseong),
  + litery Hangul (Hangul letters),
  + znaki Hangul w nawiasach (Parenthesized Hangul),
  + znaki Hangul w kółku (Circled Hangul),
  + znaki Hangul połowicznej szerokości (Halfwidth Hangul)
  + ideograficzne adnotacje fonetyczne (Kanbun),
  + znaki opisujące symbole ideograficzne (ideographic description characters),
  + telegraficzne symbole ideograficzne (Ideographic Telegraph Symbols),
  + elementy systemu ideograficznego Kangxi (Kangxi Radicals),
  + elementy systemu ideograficznego Yi (Yi Radicals),
  + formy pionowe znaków (Presentation forms for vertical ideographic),
  + formy szerokości połówkowej i pełnej znaków ideograficznych podzielone na dwa podzbiory,
* zbiór wielkich liter w alfabecie gruzińskim stosowanych w tytułach i inskrypcjach (Mtavruli),
* zbiór znaków wspólnych dla pism japońskich Katakana i Hiragana,
* grupę form liczbowych:
  + ułamki zwykłe (Vulgar Fractions),
  + ułamki północno-indyjskie (North Indic Fractions),
* symbole stosowane w systemach rozpoznawania pisma (OCR Symbols),
* grupę znaków diakrytycznych (Diacritical Symbols). Są to symbole, które dodaje się do podstawowych znaków zmieniając ich wymowę lub znaczenie. Grupa ta dzieli się na dwa podzbiory:
  + kombinatoryczne symbole diakrytyczne (Combining Diacritic Symbols), które dodaje się nad lub pod znakiem bazowym, niektóre symbole diakrytyczne otaczają symbol bazowy,
  + litery modyfikujące (Modifier Letters), czyli litery, które umieszcza się obok znaku podstawowego (najczęściej przed, ale czasami po), najczęściej mają formę indeksu górnego, a czasami dolnego,
  + symbole modyfikujące (Modifier Symbols), działające tak samo jak litery modyfikujące, ale nie są literami,
* znaki kontrolne (Control Chars), ten zbiór obejmuje znaki ASCII o kodach od 0 do 1F, 7F oraz od 80 do 9F,
* symbole waluty (Currency Symbols),

# Składnia nazwy opisowej

* Symbole astrologiczne
  + <planet name> [FORM <digit word>]
  + ASTRONOMICAL SYMBOL FOR <planet name>
* Znaki zodiaku:
  + <zodiac name>