DIN Deutsches Institut für Normung e.V.
Normenausschus Informationstechnik (NI) im DIN
Geschäftsstelle D-10772 Berlin
Tel.: +49 30 26 01-25 35
Fax: +49 30 26 01-4 25 35
E-Mail: cord.wischhoefer@din.de

NI-29.01 N 0128

Title: Definition of a Code Position for German Umlauts in ISO/IEC 10646 / Unicode Request of the German NB committee NI-29.01 "Codierte Zeichensätze" to JTC 1/SC 2/WG 2
Date assigned: 2004-05-28
Source: DIN Information Technology Standards Committee (NI) Subcommittee 29.1 "Codierte Zeichensätze"
For submission to: ISO/IEC JTC 1/SC 2/WG 2
Status: • National member body contribution • For consideration and decision by ISO/IEC JTC 1/SC 2/ WG 2
Medium: Electronic; pdf-file
Number of pages: 1 cover page + 3 pages

Definition of a Code Position for German Umlauts in ISO/IEC 10646 / Unicode

Request of the German NB committee NI-29.01 "Codierte Zeichensätze" to JTC 1/SC 2/WG 2

Introduction

The DIN-Subcommittee NI-29.01 "Codierte Zeichensätze" (Coded Character Sets), the "Die Deutsche Bibliothek" (the German National Library) and the Consortium for German and Austrian Library Networks put forward this request to ISO/IEC JTC 1/SC 2/WG 2 for defining a code position for German umlauts in ISO/IEC 10646.

The solution proposed here is intended not to interfere with national practices in other countries/cultures which may well be different from German usage of the umlaut or the trema.

The Consortium of German and Austrian Library Networks is a bi-national cooperation of the Library Networks of both countries which integrates academic and public libraries into one system. Die Deutsche Bibliothek acts as the German National Library and is the national bibliographic centre of Germany. The DIN-Subcommittee NI-29.01 is in charge of mirroring all work done in JTC 1/SC 2 and CEN/TC 304.

The German umlauts

Umlauts are an important characteristic of the German language at both the phonetic and morphologic/semantic level. They are i.a. used for

- forming the plural of nouns:
  - Vater - Väter
  - Mutter - Mütter
  - Wolf - Wölfe
  - Maus - Mäuse
- forming the comparative and superlative forms of adjectives:
  - kalt - kälter / kältest
  - groß - größer / größt
- forming the subjunctive:
  - hat - hätte
  - durfte - dürfte
  - schlug - schlüge

The umlauts are characters in the German language which are frequently used, indeed.

The umlauts Ä, Ö, Ü, ä, ö and ü, besides being pronounced differently than A, O, U, a, o and u, are in addition handled differently in library catalogues, bibliographies and indexes of names when it comes to

- sorting
- indexing
- and searching

Here umlauts are always spelt using two letters:

Ä	Ae
Ö	Oe
Ü	Ue
ä	ae
ö	oe
ü	ue

It is only in German that this equivalence between a special character and a combination of characters occurs. In the eyes of the German library community this current practice in presenting and handling/manipulating umlauts is indispensable.

Description of the problem

In ISO/IEC 10646 / Unicode the following code is defined among the "Combining Diacritical Marks":

U+0308 COMBINING DIAERESIS
= double dot above, umlaut
= Greek dialytika
= double derivative

By grouping umlaut and diaresis under one code there is no way to distinguish between a German umlaut e.g "ä" and a diaresis, e.g. a small latin character "a" with a trema "ä".

For the library systems of both Germany and Austria this ambiguity creates problems in the transition process from other character sets to ISO/IEC 10646 / Unicode. When, many years ago, electronic data processing was introduced in German libraries much energy and money was spent on automatically converting and precisely identifying existing script and language characters. By introducing ISO/IEC 10646 / Unicode without being able to differentiate between umlaut and diaresis a considerable amount of data would be lost.

Proposal for a solution

To solve the problems addressed above the DIN-Subcommittee NI-29.01 "Codierte Zeichensätze" requests the definition of a new code position in the range of U+0300 to U+036F "Combining Diacritical Marks" i.e., for example:

U+0358 COMBINING UMLAUT

Justification

As noted above the umlauts are frequently used in the German language which is spoken and written as native language by about 100 million people. Giving up the possibility of separately coding the umlauts in order to align with international practice (Categorizing the umlauts as diacritic characters, ignoring them in sorting, indexing and searching) would entail considerable cultural damage.

There is agreement between ISO/IEC JTC 1/SC 2 and ISO/TC 46/SC 4 that all relevant bibliographic 7- and 8-bit character sets are to be reflected in ISO/IEC 10646 / Unicode. ISO 5426 "Extension of the Latin alphabet coded character set for bibliographic information interchange" has separate code positions for trema (12/08) and umlaut (12/09). The said agreement between SC 2 and TC 46/SC 4 has not been fully implemented -- even though TC 46/SC 4/WG 1 accepted that the mapping of ISO 5426 to ISO/IEC 10646 / Unicode is complete (see ISO/TC 46/SC 4 N 476 and ISO/TC 46/SC4/WG1 N 240).

In the libraries and library networks of Germany and Austria there are considerable amounts of data which consistenly show the differentiation between trema and umlaut in accordance with ISO 5426. In day-to-day library practice a factual difference is being made between umlaut and trema in cataloguing, exchange and application specific use of data. Representing both umlauts and tremas by "COMBINING DIAERESIS" would lead to a loss of data, information and functionality.

One example for a "minimal pair" is "Säul" and "Säul", i.e. two identical character strings with two dots above, which could either be umlaut or trema. The only possibility to distinguish between the two would be a differing encoding of "ä" as umlaut and "ä" as trema.

Other approaches (which do not solve the problem)

- An informal agreement among the German-speaking countries (ü = u + umlaut, but u +¨ = u + trema) is no solution to the problem. A distinction using pre-combination for the umlaut and post-combination for the trema does not work because ISO/IEC 10646 / Unicode reduces the two forms to one. Existing software treats both forms as equivalent, too, and converts them into a "COMBINING DIAERESIS".

- Chosing other characters as replacement (e.g. U+030E COMBINING DOUBLE VERTICAL LINE ABOVE or U+0364 COMBINING LATIN SMALL LETTER E) does not solve the problem as new ambiguities are created.
- Using code positions in the "Private Use Area" only works for bilateral agreements and would contradict the idea of a wide exchange of bibliograhic data in library networks.
- Identifying character strings by language codes ("language tagging") to achieve unambiguous identification is not a realistic proposition as tagging would have to be at least by data field, if not by word.
- Differentiating the code point "COMBINING DIAERESIS" by using "Variation Selectors", as proposed by DIN in June 2003 (ISO/IEC JTC1 SC2/WG2 N 2593) is not possible since variation selectors are applied exclusively to single basic characters and not to diacritic signs.
- Using control sequences from the alternative control-1-set specified in ISO 6630 "Documentation -- Bibliographic control characters"

  9/5 SIB Sorting interpolation, beginning

  9/6 SIE Sorting interpolation, end

  with the interpolated "e" would not work either because international exchange in ISO/IEC 10646 / Unicode always assumes the use of the standard-control-set from ISO 6429. See also the mapping of ISO 6630

  8/8 NSB Non-sorting character(s), beginning

  8/9 NSE Non-sorting character(s), end

  to the two control postions in ISO 6429, which are roughly equivalent in function

  U+0098 START OF STRING

  U+009C STRING TERMINATOR,

  MARC 21 Specifications for Record Structure, Character Sets, and Exchange Media, CHARACTER SETS: Part 3 Code Table Extended Latin [ANSEL], January 2000, http://lcweb2.loc.gov/cocoon/codetables/1.html).