

DIN Committee Information Technology

**For visitors
in Berlin-Mitte:**
Burggrafenstr. 6
10787 Berlin

Vorsitzender:
Geschäftsführer:

Payments requested with reference
Dresdner Bank AG Deutsche Bank AG
Bank code 100 800 00 Bank code 100 700 00
Account no. 921676500 Account no. 130368400

Postbank AG
Bank code 100 100 10
Account no. 384 56-101

Member of the International Orga-
nization for Standardization (ISO)
and the European Committee for
Standardization (CEN)

Michael Y. Ksar
ISO/IEC JTC1/SC2/WG2 Secretariat
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
U.S.A.

Your reference:
Your date:
Our reference: htg
Our date:

Name: Peter Hortig
Telephone: +49 30 2601-2535
Fax: +49 30 2601-4 2535
Email: peter.hortig@din.de
Internet: <http://www.din.de>

Date: 2003-06-25

Dear Mr. Ksar

This German contribution to SC2/WG2 is for discussion. It is in preparation of a formal request to encode the characters "LATIN VARIATION SELECTOR UMLAUT" and "LATIN VARIATION SELECTOR TREMA"

Currently, a substantial amount of existing German data distinguishes between Umlaut and Trema. Both diacritics have a similar, but not necessarily identical representation, both have quite different properties e. g. with regards to sorting (cf. DIN 5007).

In particular, German library data is currently stored according to ISO 5426 "Extension of the Latin alphabet coded character set for bibliographic information interchange" which distinguishes between the two diacritics Umlaut (4/9) and Trema (4/8). However, in ISO/IEC JTC1/SC2 N3125 "Finalized Mapping between Characters of ISO 5426 and ISO/IEC 10646-1 (UCS)" both are mapped to the same UCS character, U0308. There is thus no standardized way to ensure roundtrip compatibility between the two standards.

For Germany and in particular for its national library (Deutsche Bibliothek) it is imperative for the integrity of German data that it be possible to maintain the distinction between Umlaut and Trema also in the UCS in a standardized way. Lack of ability to do so affects millions of bibliographic data records in the Deutsche Bibliothek alone (to be exact, 14 956 289 records as of October 2002) and about 110 million bibliographic data records in German and Austrian regional library networks.

We therefore request

a) the encoding of two new characters, LATIN VARIATION SELECTOR UMLAUT in position U0241 and LATIN VARIATION SELECTOR TREMA in position U0240 (the positions are suggestions only).

b) the insertion of the following text into informative Annex F "Alternate format characters" as F.2.6 "Latin selectors"

"LATIN VARIATION SELECTOR UMLAUT (U0241): Uniquely identifies the preceding character as using /being the Umlaut diacritic (cf. ISO 5426, code position 4/9)

LATIN VARIATION SELECTOR TREMA (U0240): Uniquely identifies the preceding character as using / being the Trema diacritic (cf. ISO 5426, code position 4/8)

In the absence of any variation selector, neither the character COMBINING DIAERESIS U0308 nor any of the Latin letters with diaeresis can be interpreted as representing uniquely the Umlaut or uniquely the Trema.

The LATIN VARIATION SELECTOR UMLAUT or the LATIN VARIATION SELECTOR TREMA should only be used directly following the Latin characters shown below:

00C4 LATIN CAPITAL LETTER A WITH DIAERESIS
00D6 LATIN CAPITAL LETTER O WITH DIAERESIS
00DC LATIN CAPITAL LETTER U WITH DIAERESIS
00E4 LATIN SMALL LETTER A WITH DIAERESIS
00F6 LATIN SMALL LETTER O WITH DIAERESIS
00FC LATIN SMALL LETTER U WITH DIAERESIS

U0308 COMBINING DIAERESIS

Neither the LATIN VARIATION SELECTOR UMLAUT nor the LATIN VARIATION SELECTOR TREMA carry a defined meaning when they follow any other character .

"

c) Change in ISO/IEC JTC1 SC2 N3125 (= ISO/TC46/SC4 WG1), section 3 "Mapping of Characters" the table to:

4/8 Trema, Diaeresis	0308 0240
4/9 Umlaut	0308 0241

Secretariat NI-29.01 (German part to SC02)
Peter Hortig
DIN German Institute for Standardization
DIN Committee Information Technology

Comment: Anschließend freie Texteingabe bzw. Dokumentgestaltung. Sie können dazu den Dokumentschutz entfernen.

Übrigens, um Eingaben in einmal geschützten Bereichen zu überarbeiten, erneuern Sie den Dokumentschutz am besten mit Hilfe der Tastenkombination Alt+Strg+Eingabe. Andernfalls könnten Ihre Eingaben verloren gehen.

N2593 Attachment - Proposal Summary Form

-- coding: utf-8 --

```

*****
*                                     ISO/IEC JTC 1/SC 2/WG 2                               *
*                                     *                                                   *
*                                     PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS        *
*                                     FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646      *
*                                     Please fill all the sections A, B and C below.       *
*                                     *                                                   *
* (Please read Principles and Procedures Document for guidelines and details             *
* before filling this form.) *
* See http://www.dkuug.dk/JTC1/WG2/docs/summaryform.html for latest Form. *
* See http://www.dkuug.dk/JTC1/WG2/docs/principles.html for latest Principles *
* and Procedures document. *
* See http://www.dkuug.dk/JTC1/WG2/docs/roadmaps.html for latest roadmaps. *
* *
* (Form number: N2352-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, *
* 1996-08, 1999-03, 2001-05, 2001-09) *
* *
*****

```

A. Administrative

1. Title: Variation Selectors for Umlaut and Trema
2. Requester's name: Deutsches Institut für Normung (DIN)
3. Requester type (Member body/Liaison/Individual contribution): Member body
4. Submission date: 2003-07-01
5. Requester's reference (if applicable): _____
6. (Choose one of the following:)

This is a complete proposal:	X (together with
the attached rationale)	
or, More information will be provided later:	_____

B. Technical - General

1. (Choose one of the following :)
 - a. This proposal is for a new script (set of characters): _____
Proposed name of script: _____
 - b. The proposal is for addition of character(s) to an existing block: X
Name of the existing block: Latin Extended-B
2. Number of characters in proposal: 2
3. Proposed category (see section II, Character Categories):

A - Contemporary

4. Proposed Level of Implementation (1, 2 or 3)
(see clause 14, ISO/IEC 10646-1: 2000): 3

Is a rationale provided for the choice? _____

If Yes, reference: _____

5. Is a repertoire including character names provided?

YES

a. If YES, are the names in accordance with the 'character naming guidelines in Annex L of ISO/IEC 10646-1: 2000?

YES

- b. Are the character shapes attached in a legible form suitable for review?

Not applicable

6. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard?

The Saphor GmbH will provide a True Type font with the two characters.

If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used:

Not applicable

7. References:

a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided?

The characters are currently encoded in ISO 5426 "Extension of the Latin alphabet coded character set for bibliographic information interchange" as well as in various non-standardized coding systems.

b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached?

Printed examples can be provided on request.

8. Special encoding issues:

Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)?

Not as such. The proposal is, however, an essential prerequisite for sorting, searching, indexing, and transliterating German texts and data correctly.

9. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script.

Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at <http://www.unicode.org> for such information on other scripts. Also see Unicode Character Database

<http://www.unicode.org/Public/UNIDATA/UnicodeCharacterDatabase.html> and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

Cf. attached rationale

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before?

NO

If YES explain _____

2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)?

YES

If YES, with whom?

German industry represented in DIN

German national library

German and Austrian regional library networks

If YES, available relevant documents:

ISO 5426 "Extension of the Latin alphabet coded character set for bibliographic information interchange"

DIN 5007 "Ordnen von Schriftzeichenfolgen (ABC-Regeln)"

3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included?

The proposal is relevant for the complete German speaking population in Austria, Germany and Switzerland (more than 100 million speakers). These countries have a long-standing tradition of the use of IT systems for text data storage and processing.

In particular, the requested characters are essential for use in the German library community.

Reference:

Cf. attached rationale

4. The context of use for the proposed characters (type of use; common or rare)

The proposed characters are very common especially in library-related data. More than 110 million data records are impacted in the German and Austrian regional library networks alone.

Reference:

Data from the German and Austrian regional library networks

5. Are the proposed characters in current use by the user community?

YES

If YES, where? Reference:

In all German library systems that use ISO 5426 "Extension of the Latin alphabet coded character set for bibliographic information interchange". They are potentially in use in all applications that build on ISO 5426.

In addition, there are other coding systems in use in Germany that distinguish between umlaut and trema.

6. After giving due considerations to the principles in Principles and Procedures document (a WG 2 standing document) must the proposed characters be entirely in the BMP?

YES

If YES, is a rationale provided?

YES

If YES, reference:

The characters are contemporary characters in common use. Furthermore, the characters to which the variation selectors will be applied are in the BMP

7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?

Preferably YES, but this is not essential

8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence?

NO

If YES, is a rationale for its inclusion provided? _____

If YES, reference: _____

9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters?

NO

If YES, is a rationale for its inclusion provided? _____
If YES, reference: _____

10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character?

NO

The LATIN VARIATION SELECTOR UMLAUT and LATIN VARIATION SELECTOR TREMA are used to distinguish different characters in conjunction with the characters U0308, U00C4, U00D6, U00DC, U00E4, U00F6 and U00FC, but they are not in themselves similar to any existing character.

Cf. attached rationale.

If YES, is a rationale for its inclusion provided? _____
If YES, reference: _____

11. Does the proposal include use of combining characters and/or use of composite sequences (see clauses 4.12 and 4.14 in ISO/IEC 10646-1: 2000)?

NO

If YES, is a rationale for such use provided? _____
If YES, reference: _____
Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?
If YES, reference: _____

12. Does the proposal contain characters with any special properties such as control function or similar semantics? _____

If YES, describe in detail (include attachment if necessary) _____

NO

The proposal asks for the encoding of two variation selectors to avoid compatibility issues with data that does not distinguish between trema and umlaut.

For details cf. the attached rationale.

13. Does the proposal contain any Ideographic compatibility character(s)?

NO

If YES, is the equivalent corresponding unified ideographic character(s) identified? _____
If YES, reference: _____
