ISO/IEC JTC1/SC2/WG2 N28xx
L2/04-269
2004-06-20

Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation Internationale de Normalisation
Международная организация по стандартизации

Doc Type: **Working Group Document**
Title: **Response to DIN request regarding *umlaut* and *tréma.***
Source: **United States National Body**
Status: **National Body Contribution**
Date: **2004-06-20**

Introduction

The U.S. National Body recognizes the problem summarized in JTC1/SC2/WG2 N2766 regarding the representation of umlaut in German bibliographic data in library networks. However, the particular solution to the problem proposed in that document would, we believe, lead to worse data representation problems than the problem it is attempting to fix.

In this document we present an alternative solution which we believe addresses the requirements expressed by DIN, die Deutsche Bibliothek, and the Consortium for German and Austrian Library Networks, while not introducing further irreconcilable problems in interchange of German umlauts represented in ISO/IEC 10646 and Unicode.

Restatement of the problem

WG2 N2766 correctly states the problem: a distinction is systematically made between umlaut and tréma in German bibliographic data and in the ISO standard used for the representation of German bibliographic records within the relevant library networks, ISO 5426. This distinction is not preserved by an encoded character difference in ISO/IEC 10646, and that represents a barrier to the adoption of ISO/IEC 10646 by German library networks, as well as an interoperability problem between such networks and other data networks implementing ISO/IEC 10646 as the basis of their data representation.

The U.S. National Body stipulates that the German National Body assertion that a distinction must be observed between umlaut and tréma in German bibliographic data is a correct assessment of the data processing requirement. Some way to maintain the relevant distinction in ISO/IEC 10646 must be found.

However, the proposal to add U+0358 COMBINING UMLAUT would, at this stage, result in massive data representation ambiguities for German data and would exacerbate, rather than eliminate, data interoperability issues.

In existing Unicode implementations of ISO/IEC 10646, **ä** is canonically equivalent to **<ä, ◌̈>** and is interpreted as a-umlaut in massive amounts of data, as well as being hard-wired into mapping tables for other pre-existing German character sets including, of course, ISO/IEC 8859-1 and Windows Code Page 1252. These are facts that cannot be changed now, and which must be taken into account when suggesting new encodings that will impact German data.

Existing **ä** characters *will* be equated to **<ä, ◌̈>** <A, COMBINING DIAERESIS>. Even if a COMBINING UMLAUT character were encoded, it would not and could not be treated as generating a canonical equivalent for **ä** interpreted as a-umlaut. This would create a significant data mapping problem at the interface between German bibliographic systems and other representations of German data, since umlauted characters would be incorrectly mapped to trémas, and the representation of umlauts from German bibliographic data would, on conversion out, be represented by 10646/Unicode sequences which would appear, in rendering, to be umlauted letters, but which would in fact *not* be treated as equivalent to umlauted letters. This would create a de-facto situation where German data would be represented one way inside bibliographic systems using 10646/Unicode and in an incompatibly different way outside such systems. We do not believe that such a situation would, in the long run, benefit any of the users of German information systems.

Alternative solution

While recognizing the drawbacks to all of the alternatives to encoding a new COMBINING UMLAUT character outlined in WG2 N2766, we believe that there is a workable alternative solution which has, to date, been overlooked. The solution consists, essentially, of using U+034F COMBINING GRAPHEME JOINER (CGJ), in its intended semantics in 10646/Unicode, to make the relevant sorting, searching, and data mapping distinctions required for umlaut *versus* tréma. In particular, the distinction we propose is:

U+0308 ◌̈	→	umlaut
<CGJ, ◌̈>	→	tréma
<**a**, ◌̈>	→	a umlaut
<**a**, CGJ, ◌̈>	→	a tréma

The sequences <**a**, ◌̈> and <**a**, CGJ, ◌̈> are not canonically equivalent. this means that the distinction will not be normalized away on conversion in and out of bibliographic systems. This eases the interoperability problem. Both sequences will *display* as **ä**, as they should. Furthermore, the semantics of CGJ are such that it should impact only searching and sorting, for systems which have been tailored to distinguish it, while being ignored in other respects in interpretation.

The reason for treating the existing sequence <**a**, ◌̈> as representing the *umlaut* in German bibliographic systems, despite the name of U+0308 COMBINING DIAERESIS, is that this is the unmarked case, representing the vast majority of extant data. The marked form <**a**, CGJ, ◌̈> should be utilized for the marked case in the data, namely the *tréma*, which is far, far less frequent in German bibliographic data. This minimizes the conversion and data rectification issues, and also guarantees that representations including CGJ will be uncommon in data converted out of the German bibliographic records.

The existence of separate representations for umlaut and for tréma, which are not canonically equivalent (and thus not neutralized by normalization processes in the data) enables German implementations which need to distinguish the two for searching and sorting, to systematically maintain weighting distinctions to do the right thing. <**a**, ◌̈> = <**ä**> can be treated as equivalent to <**a**, **e**> for sorting purposes, while the tréma <**a**, CGJ, ◌̈> can be weighted as a secondary variant of <**a**> thus resulting in the desired behavior for such systems. *Existing* collations which do not distinguish tréma and umlaut in German data will continue to work exactly as they currently do, since in default collation tables CGJ is ignored in weighting.

We believe that this proposed solution has the correct mix of technical attributes to enable the German library networks to make the required distinction, to correctly convert existing ISO 5426 bibliographic records, and to implement the desired sorting and searching behavior for German data represented directly in 10646/Unicode.

At the same time, this solution does not introduce incompatibilities or non-interoperability issues for other existing implementations of 10646/Unicode which handle German data.