L2/02-228

ISO/IEC JTC 1/SC 2/WG 2 N 2488

DATE: 2002-05-21

ISO/IEC JTC 1/SC 2/WG 2

Universal Multiple-Octet Coded Character Set (UCS) - ISO/IEC 10646

Secretariat: ANSI

TITLE:	Questionable Terms: Canonical Form and UTF-16
SOURCE:	Zhang Zhoucai, China
STATUS:	***Expert Contribution***
DUE DATE:	2002-05-21
DISTRIBUTION:	SC2/WG2 members and Liaison organizations
MEDIUM:	Electronic
NO. OF PAGES:	1

The term UCS-4 , CANONICAL FORM in ISO/IEC 10646 does NOT reflect the reality in IT world. The purely four octet form has never been , and will not be utilized by IT industry, however, it is named as CANONICAL FORM which confused, is confusing and would mislead users of the standards. (In Chinese, the word CANONICAL is very critical, even stronger than the term NORMAL!)

In the meantime, the ambiguity of UTF-16 does exist. For most applications, UTF-16 implies 16 bit fixed length form of encoded characters, for a few cases, it is 2 or 4 byte variable length encoding. Therefore, one of UTF-16 is WITHOUT surrogate while the another UTF-16 is WITH surrogate. They are represented in totally different structure in sense of data type. Therefore, there is a need to differentiate the two modes.

WG2 are expected to take the responsibility to revise the text or make additional notes to clarify the meaning of the terms in order to avoid possible misleading in using the standard.