Proposed Updates to ANSI/NIST-ITL 1-2005
UTF8, GPS, Tracking Subcommittee

Scott Hills, Aware, Inc.
Rob Mungovan, Aware, Inc.
Dale Hapeman, DOD-BFC
Tony Misslin, Identix
Bonny Scheier, Sabre
Ralph Lessman, Smiths-Heimann

UTF-8 for User Defined Fields

Change to Section 6.1 File Format

The second paragraph restricts text or character data in Type 2 and Type 9 through Type 16 records to 7-bit ASCII code. This should be modified to allow 8-bit UTF-8 characters in all user defined fields within these records.

Proposed wording change to paragraph 2:

After the first sentence:
"The text or character data in Type-2, and Type-9 through Type-16 records will normally be recorded using 7-bit ASCII code in variable length fields with specified upper limits on the size of the fields."
… *Add this…*
"Eight bit UTF-8 characters shall be allowed to support international character sets for all user defined fields in all record types. By definition this excludes record types 1, 3, 4, 5, 6 and 8."

The third paragraph should be changed to suggest UTF-8 as the preferred way of storing data that cannot be represented in 7-bit ASCII.

Proposed wording change to paragraph 3:

After the first sentence:
"For data interchange between non-English speaking agencies, character sets other than 7-bit ASCII may be used in textual fields contained in Type-2 and Type-9 through Type-16 records."
… *Add this…*
"UTF-8 is the preferred method of storing textual data that cannot be represented as 7 bit ASCII."

Change to Section 7.2.3 International character sets

… at the very end of this section add this text…

"Usage of UTF-8 is allowed as an alternative to the technique that requires the usage of the ASCII "STX" and "ETX" characters to signify the beginning or end of of international characters. UTF-8 is only allowed to be used for user defined fields, for example, record type 2, and the UDF fields (200-998) of record types 10, 13, 14, and 15."

… and add this text…

"Even though there is no overlap within the character sets used with UTF-8, UTF-8 should be registered in the type 1 record within the DCS field 1.15 (Directory of Character Sets). "

Rationale:
NIST ITL-2000 reads "…The DCS is an ordered list of 3 information items containing an identifying code, the name of an international character set, and its version."

Because UTF-8 supports the full range of character sets, but the terminals (UNIX or Windows based workstations) have only the fonts for a subset of all character sets installed, the used character set should also be registered at the DCS field.

Change to Section 8.1.15: Directory of character sets(DCS)

After the second sentence:"This field shall contain one or more subfields, each with three information items. The first information item is the three-character identifier for the character set index number that references an associated character set throughout the transaction file. The second information item shall be the common name for the character set associated with that index number, the optional third information item is the specific version of the character set used."

… add this text…

"In the case of the use of UTF-8, the third optional information item can be used to hold the specific version of the character set used with UTF-8, so that the human terminal can be switched to the right font family.

Example

1.015:003<US>UTF-8<US>Chinese BIG5<GS>"

Change to Table 4, page 16
Add a new character set index.

003, UTF-8, 8-bit (addition)
004-127, reserved for ANSI/NIST future use (modification)