FUJITSU

White Paper
Unicode in BS2000

Key system and system support programs in BS2000 now also support UNICODE character sets.
The aim of the development project was to add the Unicode characters that are relevant in the Western European language area to the EBCDIC variants currently available for BS2000. Cyrillic can also be supported if necessary.
For internal communication within the BS2000 system, the supported character sets are supplemented by UTF-16 (in the scope of UCS-2). For sorting and normalization purposes, tables which support the characters within the code points U+0000 to U+2FFF are available as standard.
This ensures support for all European characters, including the special characters used in the signaling signals of public authorities, because they are fully contained within the supported value range.

Contents

Motivation 2
Objective 2
Nature of the development 2
Current situation in BS2000 2
Unicode in the BS2000 system environment 2
Unicode in the BS2000 application environment 4
Preparations for use of Unicode 4
Brief description of the Unicode-specific extensions in the individual products 4
- AID 4
- COBOL2000 4
- EDT 4
- ESQL-COBOL 4
- FHS 4
- IFG 4
- MT9750 - Terminal emulation 5
- OMNIS 5
- openFT 5
- ORACLE 5
- PERCON 5
- RSO 5
- SESAM/SQL 5
- SORT 5
- UDS/SQL 5
- VTSU 5
- XHCS 6

Cross-references 6

Motivation

Certain EU directives and their transposition into national legislation oblige our customers to adapt their electronic data processing (EDP) applications to take account of the following guidelines:

- Every EU citizen has the right to have their name written correctly in Latin script, diacritical characters included.
- International postal directive on the correct way of writing a foreign address
- BundOnline 2005 directive
- German Federal Ministry of the Interior directive on data exchange between authorities.

The aim of Unicode is to define a globally unique coding for all characters occurring worldwide. The BS2000 implementation achieves this for the Western European area.

For more details on Unicode (including Glossary), visit the Unicode homepage at http://www.unicode.org/.

Objective

With the introduction of Unicode support in BS2000, the EBCDIC character sets available in BS2000 systems will be extended by additional characters that will be required in the European language area in the future. This will be achieved through the use of selected Unicode code points in addition to the existing EBCDIC variants.
As a result of the Unicode support, the further use of BS2000 systems is ensured even under the more exacting requirements to be met by the code characters that are to be supported. To assure an evolutionary expansion, this will happen while providing extensive protection for the invested assets of customers' application resources.

Nature of the development

Extension of the EBCDIC character sets available in BS2000 systems by additional characters / glyphs. This will be brought about by the use of selected Unicode code points in addition to the current EBCDIC variants.
When reference is made in the following to Unicode support in BS2000, all that is meant in the present context is this extension of the codeset.

The primary consideration is: What do users require in order to extend their existing applications by Unicode data fields. It is assumed here that the number of fields needing to be converted to Unicode or to be inserted in addition is small. These will mainly be name and address fields.

A principal objective is that the customer only has to modify those applications/application parts that actually use the extended functionality. Mixed operation with the existing ("legacy") applications must be ensured.

Current situation in BS2000

When processing data, BS2000 uses a 7-bit EBCDIC character set as standard. The Fujitsi-defined international 7-bit EBCDIC table, which is used as the system standard character set, is called EDF03IRV. This makes 95 different displayable or printable characters available.
Using XHCS, BS2000 supports the conversion of a series of 8-bit ISO codes (ISO 8859-x variants) into the associated EBCDIC codes that the system uses during processing.

Unicode in the BS2000 system environment

Characters of the most important industry standard character sets, such as the ISO standards, have a 1:1 correspondence in Unicode (this means that the same result is obtained in a conversion from EBCDIC to Unicode and back).
Unicode is stored and transferred in different formats. UTF (Unicode Transformation Format) describes methods for mapping a Unicode character to a sequence of bytes. Each character of a Unicode string is represented in this case by one or more "code units".

- **UTF-8** is the most widely used encoding scheme for Unicode characters and the most space-saving method of mapping Unicode characters for all scripts based on the Latin alphabet. A code unit is 8 bits; each Unicode character needs 1, 2, 3 or 4 bytes.
- Alongside UTF-8, **UTF-16** is also very important, e.g. for character encoding in Java. A code unit is 16 bits; each Unicode character needs either 2 or 4 bytes. UTF-16 is largely identical to the 2-byte Unicode transformation format UCS-2.
- **UTF-32** is an encoding format with a constant number of bytes per Unicode character. UTF-32 has 32-bit code units; each Unicode character needs precisely 4 bytes. (Not supported in BS2000.)
- **UTF-EBCDIC** is a Unicode extension based on the proprietary EBCDIC format used on IBM mainframes. Code units are 8 bits; each Unicode character needs 1, 2, 3, 4 or 5 bytes.
- **UTF-E** is being newly introduced in BS2000; UTF-E in BS2000 is similar to IBM's UTF-EBCDIC format and is based on a modified UTF-8.

In contrast to UTF-8, multibyte encoding in UTF-EBCDIC and UTF-E only starts at Unicode code point U+A0. The first 256 code points in Unicode tally with the ISO8859-1 character set. In other words, the entire second control character block of ISO8859-1 is preserved as single-byte encoding. In the subsequent conversion from this modified UTF-8 to EBCDIC in BS2000, the conversion table EDF041 is used as a basis.
The single-byte encoding of UTF-E corresponds to the EDF03IRV character set.
Since UTF-E is just another representation of Unicode, any other Unicode encoding format can be mapped to UTF-E and vice versa.

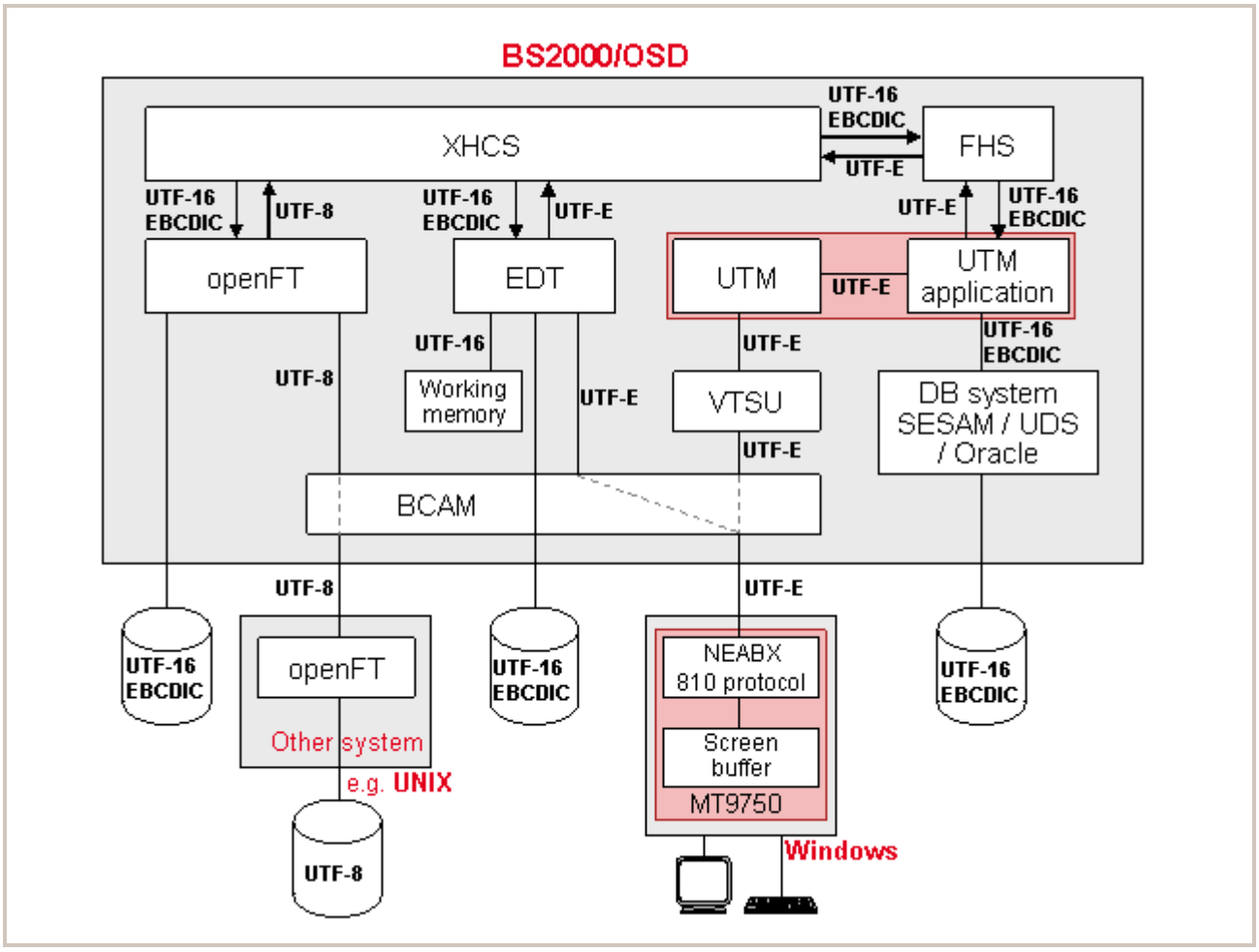
Important: Both ORACLE's UTF-E character set and SUN's UTF-EBCDIC are based on IBM's EBCDIC tables and differ from the UTF-E form used in BS2000.

All encodings with a variable number of bytes per character are structured in such a way that no more than 4 bytes must be inspected at any point in the string in order to find all the bytes for a UNICODE character. In other words, they do not always have to be examined from the beginning in order to find the character boundaries at a given point ("non-overlap").

Extensions in the XHCS subsystem (XHCS is part of the openNet Server product) form the basis for Unicode support in BS2000.
XHCS

- supports the encoding variants UTF-8, UTF-16 and UTF-E (variable length per character).
- includes conversion tables for the supported EBCDIC codes and 8-bit ISO codes (ISO8859-1/2/3/4/5/7/9/15) to Unicode and their reverse mapping.
- supports other characters of the Unicode character set that are used as special characters in the signaling systems of public authorities - in addition to the characters contained in these ISO codes.
- supports UTF-16 strings from U+0000 to U+2FFF for normalization and sorting.

Unicode-based characters are permitted only for the texts or user data to be processed/administered. For commands and keywords, only the characters in the previous code set (EBCDIC) are supported.



Unicode in the BS2000 system environment
(Example schematic for data input, status 2007. Certain technical details have been omitted to improve clarity.)

Unicode in the BS2000 application environment

Two types of applications need to be distinguished in the application environment:

- System support applications offered and supplied by Fujitsu, and
- Customer applications developed specifically for the particular deployment situation.

The system support applications developed by Fujitsu have also been extended to allow the use of Unicode characters.

Unicode support for the programming and runtime environment extends to the following functional areas:

- Programming: COBOL2000, ESQL-COBOL, IFG (Unicode data fields), AID
- Data storage: SESAM/SQL, ORACLE (Unicode data fields), UDS/SQL (National data)
- Data processing: EDT, SORT, PERCON, XHCS
- Input/output: Terminal support (VTSU, MT9750, FHS), printer support (RSO, Spool), file transfer (openFT)

Preparations for use of Unicode

In addition to upgrading the BS2000 system base, the system support applications and the programming environment to the requisite versions, other preparations also have to be made on the customer side. Particular attention should be given here to the following points:

- The data and files must be brought to a defined state.
- The CCS (Coded Character Set) names used in the BS2000 system are a key component in the handling of data. It must be ensured that the content and the identification of the data by means of CCS names are consistent.
- The applications concerned must be extended to allow for the Unicode data fields that are to be processed in addition.

Brief description of the Unicode-specific extensions in the individual products

Detailed information on the Unicode-specific extensions to the individual products can be found in the relevant manuals.

AID

AID provides low-level support for the UTF-16 data type, which is also supported as a new data type by COBOL.
In addition, in connection with the support by COBOL, AID also provides the UTF-16 data type in the LSD (List of symbolic debugging). This means that this data type can also be debugged symbolically.

COBOL2000

To support the Unicode functionality in BS2000, the data type NATIONAL (PIC N) is being introduced in its various forms in COBOL2000 along with the procedural use of this data in compiler and runtime system. The characters are represented in UTF-16. The character set is limited to UCS-2. In other words, only the two-byte representation is supported by the compiler, without surrogates.

EDT

Starting with Version 17.0, EDT can work with several different character sets simultaneously. All the data that EDT reads into or outputs from its screen buffer is subsequently present in UTF-E format. Internally, EDT processes the work file in main memory in UTF-16. Files using the character sets EBCDIC (as today), UTF-16 and UTF-E can be processed.

ESQL-COBOL

The precompiler supports the new data types NCHAR and NVARCHAR in the functional scope of SESAM.
ESQL-COBOL permits the COBOL data type NATIONAL in its Declare Sections and allows its use as a host variable in SQL statements for passing on to the ICSQLE interface.

FHS

If UTF-16 fields are to be read into a mask in addition to the present EBCDIC fields, FHS can no longer request the screen buffer in EBCDIC coding, but must request it in UTF-E.
A format is converted and output as a Unicode-encoded message (UTF-E) by FHS provided a field of the FHS format is defined as a field permitted for Unicode inputs or outputs. At input, the input string is analyzed and converted into the coded character set (EBCDIC or UTF-16) defined for each field by IFG. XHCS is used for the code conversion.

IFG

IFG supports an identifier for Unicode for individual fields of a format. A new "Unicode" attribute is being introduced for this purpose. In these fields, the contents are then interpreted by FHS as 2-byte UTF-16 characters. The generation of the FHS formats and addressing aids are adapted accordingly for this purpose. Formats containing Unicode fields can only be output by terminal emulations that support Unicode (UTF-E).

MT9750 - Terminal emulation

The MT9750 terminal emulation supports Unicode for European characters and in the character set from U+0000 to U+2FFF. With this feature it is possible to display characters from various character sets in one form. In Unicode mode characters not available on the keyboard can be created via the compose button. In addition, Unicode characters can be inserted via the clipboard.

OMNIS

OMNIS as of V8.5A supports the communication in the Unicode character set.

openFT

openFT for BS2000 supports the transfer of DMS and POSIX files using the Unicode variants UTF-16 (big-endian) and UTF-8. The Unicode variant is specified either via a parameter value in the TRANSFER-FILE command or via an XHCS codeset entry in the file catalog (for DMS files only).

ORACLE

Oracle-BS2000 provides the option to store UTF-16 data in the database, i.e. in fields containing specific NCHAR data types. Except for Pro*Cobol, these fields can also be processed in BS2000. Particular attention should be paid to the restriction to the exclusive use of Unicode characters in BS2000 to the textual content.
Oracle Database 10g allows Unicode characters to be stored in NCHAR, NVARCHAR2 and NCLOB data types using UTF8 or UTF16 format and to be processed by means of SQL or PL/SQL. The utilities DataPump, Export, Import, SQL*Loader are able to process Unicode data as well as applications using the Pro*Cobol, Pro*C and OCI interfaces. For the conversion to Unicode and vice versa Oracle offers several ASCII, EBCDIC and Unicode character sets.

PERCON

In PERCON extensions to allow conversion between Unicode data and other data (with compatible CCS) are realized. The normalization of Unicode data (composed) is supported in addition.

RSO

Support in RSO for Unicode printers which accept UTF8-encoded data as network printers. The following encoding forms are supported:
UTF-8 (including VTSU and RSO font control characters)
UTF-E (including VTSU and RSO font control characters)
UTF-16 (text files with and without form feed characters)

Cases 1 and 2 partially concern the interface to UTM; the extension supports the printing of Unicode text. The 3rd case permits the printing of Unicode documents on local (IPDS) and RSO printers.

SESAM/SQL

In the SESAM/SQL DB system, the introduction of the NCHAR and NVARCHAR data types enables Unicode characters to be stored as well and to be processed using SQL tools. As well as support for the new data types by the DML languages, their use in the different service functions (e.g. LOAD, UNLOAD, IMPORT, EXPORT) is also allowed here. To convert from e.g. CHAR to NCHAR and vice versus, SESAM/SQL makes use of the call functions provided by XHCS. The SESAM-internal data comparison is always performed in binary form for national data as well.

SORT

SORT currently supports a three-level sorting of UTF-16 fields according to the Unicode standard (see http://www.unicode.org/reports/tr10/tr10-9.html). To each Unicode character a sort item is assigned. The sorting elements are defined by means of a table supplied by XHCS (Unicode ollation Default Table).

UDS/SQL

UDS/SQL from V2.5 allows the storage and retrieval of Unicode characters in the database of UDS/SQL.
In UDS/SQL databases can define fields of data type NATIONAL CHARACTER and NATIONAL CHARACTER VARYING and therein Unicode data can be stored in UTF-16 format. Metadata such as names of record types, sets and realms are further specified and stored in EBCDIC.
As part of a restructuring, newly created fields of the data type NATIONAL are initialized with the national space.

VTSU

Extension of the functionality for detecting Unicode support by partners, Unicode text recognition, processing and conversion with the aid of XHCS.

XHCS

XHCS is implemented as a subsystem. It has been extended by the Unicode transformation formats UTF-8, UTF-16, UTF-E.
It provides information on the compatibility of the codes and the possibility of conversion. The functions for converting the character sets with the necessary sorting tables, for converting lowercase/uppercase and for supporting encoding formats in Unicode are also available to application programs.

Cross-references

- White Paper on BS2000/OSD-BC V7.0
- Overview manual on 'Unicode in BS2000/OSD' – see Manual Server at http://manuals.ts.fujitsu.com/
- Unicode homepage http://www.unicode.org/
- The Unicode Standard: A Technical Introduction http://www.unicode.org/standard/principles.html

Contact:
Fujitsu
Barbara Stadler
Mies-van-der-Rohe-Str. 8, 80807 Munich
Germany
Telephone: +49 (0) 89 62060-1978
Email: Barbara.stadler@ts.fujitsu.com
Web site: de.fujitsu.com
August, 13 2015 EM EN