

Unicode Support

Chapter 2:

SYS-ED/Computer Education Techniques, Inc.

Ch 2: 1

Objectives

You will learn:

- Unicode features.
- How to use literals and data items.
- Coding commands to manipulate Unicode fields.

Unicode: What is it?

- It is the industry standard for the coded character set:
 - It is defined by Unicode Consortium and ISO.
- Covers all commonly used characters in the world in one code page.
 - As compared to one "language" per ASCII, EBCDIC, or EUC code page.
- Characters: text, digits, special characters, symbols, control characters, ...
- Multiple Unicode encoding formats:
UTF-8, UTF-16, UTF-32.

Unicode: Why

- Support global e-business environment:
 - Applications for multi-cultural/multi-geographic businesses.
 - Networks of heterogeneous systems.
- Enables a common implementation for global applications versus a separate code page for each geographic area or system platform.
- Supported by all key operating system and middleware platforms.
- Required by: XML, HTML, Java, and other software.

Unicode Support in Enterprise COBOL for z/OS

- Enable Unicode processing for COBOL applications.
- Consistent with COBOL 2002 standard.
- Interoperate with:
 - DB2 Unicode support.
 - Java.
 - COBOL XML support.

Unicode Support: Overview

- Unicode literal and value clause.
- Unicode data type.
- New compiler options:
 - CODEPAGE()
 - NSYMBOL()
- Collation: binary.
- Implicit conversions for EBCDIC data assigned to or compared with Unicode data.
- Explicit conversions via intrinsic functions.

Unicode Literals

- N'αβγ', N'Straße'
 - Literal value in source is encoded in some EBCDIC code pages.
 - Value is converted to UTF-16 for execution.
 - Value limited to those represented by the source program code page.
- NX'03B103B203B3'
 - Can be used for characters:
 - not supported by editor.or
 - not in code page of source program.

Unicode Data Type

- USAGE NATIONAL, Picture character N:
01 Japan pic N(20) usage national .
- One UTF-16 encoding unit (2 bytes) per PICTURE N character.
- "Character" defined in terms of PICTURE symbol positions, for reference modification, character counts, etc.

Compiler Options

- CODEPAGE (ccsid)
 - Specifies EBCDIC CCSID for:
 - literals in the source program.
 - contents of alphanumeric and DBCS data items.
- Shipped default is 1140 (Latin-1 with Euro)
 - NSYMBOL (DBCS NATIONAL).
- 01 X PIC NN. and N'. ' are ambiguous:
Unicode or DBCS?
 - NSYMBOL option controls default interpretation.
 - PICTURE G and G'...' are treated as DBCS regardless of NSYMBOL.

Assignment

- NATIONAL, DISPLAY or DISPLAY-1 item may be assigned to NATIONAL item:
 - 01 Country pic N(20) usage national.
 - 01 USA pic X(13) value 'United States'.
 - 01 Greece pic X(6) value 'Ελλάδα'.
 - Move USA to Country.
 - Move Greece to Country.
- Numeric integer may be assigned to NATIONAL.
- Padding with Unicode space character: X'0020'.
- Truncation by 2-byte encoding units:
 - Application logic responsible for avoiding partial character truncation when dealing with characters represented in two encoding units.

Unicode Compares

- National item may be compared with:
 - national, alphanumeric, DBCS, or numeric integer operand.

If Country = N'日本' ...

 - Non-Unicode operand is converted to Unicode.
 - Shorter operand values are padded with Unicode blanks.
- Byte for byte comparison in binary order:
 - No culturally sensitive compares:
 - e.g. **N'ç'** is not equal **N'c'**, regardless of locale
 - No normalization:
 - e.g. **á** (composed) is not equal to **a´** (decomposed)sd

Other Language Syntax Supporting Unicode

- Statements involving comparisons:
 - EVALUATE, IF, INSPECT, PERFORM.
 - SEARCH, STRING, UNSTRING.
 - SORT, MERGE, indexed file keys.
- Class condition on Unicode data:
 - NUMERIC, ALPHABETIC, ALPHABETIC-LOWER, ALPHABETIC-UPPER, class-name.
- Unicode arguments for CALL or INVOKE.
- INITIALIZE ... REPLACING NATIONAL ...
- Reference modification.

Intrinsic Conversion Functions

- **FUNCTION DISPLAY-OF(*national-data* [*ccsid*])**
 - Returns alphanumeric (EBCDIC) representation of NATIONAL argument.
- **FUNCTION NATIONAL-OF(*ebcdic-data* [*ccsid*])**
 - Returns UTF-16 representation of EBCDIC (DISPLAY or DISPLAY-1) argument.
- If *ccsid* omitted, it defaults to value from CODEPAGE() compiler option.
- *ccsid* may represent an EBCDIC, ASCII, EUC or UTF-8 code page.
 - Recommendation:
Use only one EBCDIC code page in a program.

Unicode: Using in DB2 COBOL Programs

- Consistent support for Unicode in DB2 and COBOL:
 - Same Unicode conversion facility.
 - Binary collation.
- CCSID information for NATIONAL, DISPLAY, or DISPLAY-1 host variables is automatically coordinated between COBOL and DB2.

e.g.

```
EXEC SQL DECLARE:X VARIABLE CCSID 1140  
END-EXEC
```

- Is no longer required.

COBOL Unicode Support and XML Processing

- COBOL now contains built-in syntax for processing XML documents.
- XML PARSE statement parses XML documents, drives processing procedure for each event.
- XML documents may be encoded in UTF-16 Unicode.
- XML documents encoded in UTF-8 may be converted to UTF-16 using the NATIONAL-OF function, then parsed.
- XML-NTEXT special register returns to the program the Unicode content from the document, that is associated with each event.

Unicode Example

```
IDENTIFICATION DIVISION.  
PROGRAM-ID. NATSAMP.  
DATA DIVISION.  
WORKING-STORAGE SECTION.  
01  f1                pic n(5)  usage national.  
01  f2                pic x(5)  value '12345'.  
PROCEDURE DIVISION.  
    move 'abc' to  f1  
    display f1  
    move function NATIONAL-OF(f2) to f1  
    display f1  
    move function DISPLAY-OF(f1) to f2  
    display f2  
    goback  
    .
```