TN1010-20-n: Codepages and character sets in message content

Contents

1 Introduction 1

2 Body Property (Section Object) 1

2.1 Property Set 1

2.2 Property Get 1

3 BCC, CC, From, ReplyTo, Sender, Subject, To Properties (Message Object) 2

3.1 Property Set 2

3.2 Property Get 2

4 Codepage/character set mapping 2

4.1 Mapping a codepage to a character set 2

4.2 Mapping a character set to a code page 2

4.3 Internal Mapping Table 3

1 Introduction

The CSMail library supports internationalised email content in the `Body`, `BCC`, `CC`, `From`, `ReplyTo`, `Sender`, `Subject` and `To` properties.

This document describes the mechanisms used to encode the text and the implementation of those mechanisms in the library.

2 Body Property (Section Object)

The body is stored internally as a SBCS or MBCS string using the character set indicated by the `Characterset` property. The control converts the internal representation to and from the correct Unicode text when the property is accessed.

2.1 Property Set

The library will:

- Determine the system codepage (through the Windows `GetACP()` API call).
- Convert the Unicode text from the calling application to a SBCS or MBCS string using the active codepage (through the `WideCharToMultiByte()` API call).
- Map the codepage to a character set (see 4) and set the `Characteset` property accordingly.

2.2 Property Get

The library will:

- Map the value of the `Characterset` property to a codepage (see 4).
- Convert the SBCS or MBCS string to a Unicode text using the codepage (through the `MultiByteToWideChar()` API call). If no codepage could be found for the `Characteset` property then the system codepage is used.

3 BCC, CC, From, ReplyTo, Sender, Subject, To Properties (Message Object)

From CSMail V1.3 onwards the library supports internationalised content in the above properties through the mechanism defined in RFC 2047.

These properties are stored internally as RFC 2047 encoded strings in the Header object. The control converts the internal representation to and from the correct Unicode text when the properties are accessed.

The control will not encode or decode the headers when they are accessed directly through the Header property – this allows developers direct access to the encoded strings.

3.1 Property Set

The library will:

- Determine the currently active codepage on the system (through the Windows `GetACP()` API call).
- Convert the Unicode text from the calling application to a SBCS or MBCS string using the active codepage (through the `WideCharToMultiByte()` API call).
- If necessary encode the SBCS or MBCS string according to RFC 2047.

3.2 Property Get

The library will:

- Decode the RFC 2047 encoded header if necessary.
- Map the value of the character set in the RFC 2047 encoded header to a codepage (see 4).
- Convert the SBCS or MBCS string to a Unicode text using the codepage (through the `MultiByteToWideChar()` API call). If no codepage could be found for the `Characteset` property then the system codepage is used.

4 Codepage/character set mapping

4.1 Mapping a codepage to a character set

The library will:

- Lookup the codepage in the registry under the key: `HKEY_CLASSES_ROOT\MIME\Database\Codepage\` and use the corresponding `BodyCharset` value.
- If no codepage is found in the registry then iterate through an internal table of mappings (see 4.3) to find the codepage and use the corresponding character set.

4.2 Mapping a character set to a code page

The library will:

- Lookup the character set in the registry under the key: `HKEY_CLASSES_ROOT\MIME\Database\Charset\` and use the corresponding `InternetEncoding` value.
- If no codepage is found in the registry then iterate through an internal table of mappings (see 4.3) to find the character set and use the corresponding codepage.

4.3 Internal Mapping Table

Codepage	Character Set
1250	"iso-8859-2"
1251	"koi8-r"
1252	"iso-8859-1"
1253	"iso-8859-7"
1254	"iso-8859-9"
1255	"iso-8859-8-i"
1256	"iso-8859-6"
1257	"iso-8859-4"
1258	"windows-1258"
20866	"koi8-r"
21866	"koi8-ru"
28592	"iso-8859-2"
28593	"iso-8859-3"
28594	"iso-8859-4"
28595	"iso-8859-5"
28596	"iso-8859-6"
28597	"iso-8859-7"
28598	"iso-8859-8"
38598	"iso-8859-8-i"
50220	"iso-2022-jp"
50225	"iso-2022-kr"
51932	"euc-jp"
51949	"euc-kr"
52936	"hz-gb-2312"
65000	"utf-7"
65001	"utf-8"
708	"ASMO-708"
720	"DOS-720"
852	"ibm852"
862	"DOS-862"
866	"cp866"
874	"windows-874"
932	"iso-2022-jp"
936	"gb2312"
949	"euc-kr"
950	"big5"
20105	"us-ascii"
28585	"iso_8859-5"
28599	"iso-8859-9"