

國家圖書館

「CMARC3-XML 轉換程式」

特殊字元處理原則

1.W3C 特殊控制字元包括如下：(轉成 XML 時須轉成「內建實體」entity)

& - 表示&字元

< -表示<字元

> - 表示>字元

" - 表示"字元

' - 表示'字元

2.特殊字元、控制碼用"&#" + 內碼 + ";"表示 Unicode (ex. &#a9; 表示 16 進位 a9 的 unicode 字元)

3.CCCII 字碼注意事項：考慮到與中文及 ASCII 夾雜的情況，所以在遇到中文字的時候，需要一個 Shift-in 的機制，而遇到正常的 ASCII 時候，就要做 Shift-out 的動作。Shift-in 與 Shift-out 均有三碼：Shift-in: (0x1B) + "\$1" ; Shift-out: (0x1B) + "(B"。而(0x1B)也就是 ESC 控制碼，在 XML 文件當中是不被允許的。因此在儲存 CCCII 至 XML 文件時，會把 (0x1B) 轉為 Entity <esc/> ;相反的，在解析 XML 文件時，會把 Entity <esc/> 轉為 (0x1B)。如此即可正確的轉換。