# 字元集與編碼介紹

*Charset、Encoding、Unicode、UTF-8*

網路系統組　謝立益

2009.11.19

# 亂碼？
## 檔案名稱出現亂碼

# 亂碼？
## 網頁有亂碼，看不懂！

| 嶺臬? | 莞瑕　閩 ? |
| --- | --- |
| 銖　??餉　璈　　?蝳　恣?嗆　??odp | ?　風?脰???/a> |
| 鋻箏　?脣　璈　　?餉?鎢?媞　　).doc | ?　風?脰???/a> |
| 撒　?璈　蝙?刻牧??doc | ?　風?脰???/a> |
| 瘣賢　隢　蝙?典?雜　?.doc | ?　風?脰???/a> |
| 閱　?璈　??　?銖剖??餉　璈　　蝞∠?剴行?.doc | ?　風?脰???/a> |
| 閱　?璈　??　?銖剖??餉　璈　　蝞∠?剴行?_?　”銖.doc | ?　風?脰???/a> |
| 閱剖??脣　蝝?　”(媞　　).doc | ?　風?脰???/a> |
| 3F_door.pdf | ?　風?脰???/a> |

瘥　?嶺臬?

# 編碼
## 寫在網頁裡面

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xht
 strict.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="zh-TW" lang="zh-TW">
  <head>
    <meta http-equiv="content-type" content="text/html;charset=UTF-8" />
    <meta http-equiv="generator" content="WebSVN 2.0" /> <!-- leave this for stats
    <link rel="shortcut icon" type="image/x-icon" href="./templates/calm/images/fa
```

# 字元
Character

1. 自然語言的書寫單位
2. A b α β 1 2 3 € お ザ ㄅ ㄆ ㊣ 獻 献
3. 字元可以組成字串（Text String）

# 字元集
## Character Set

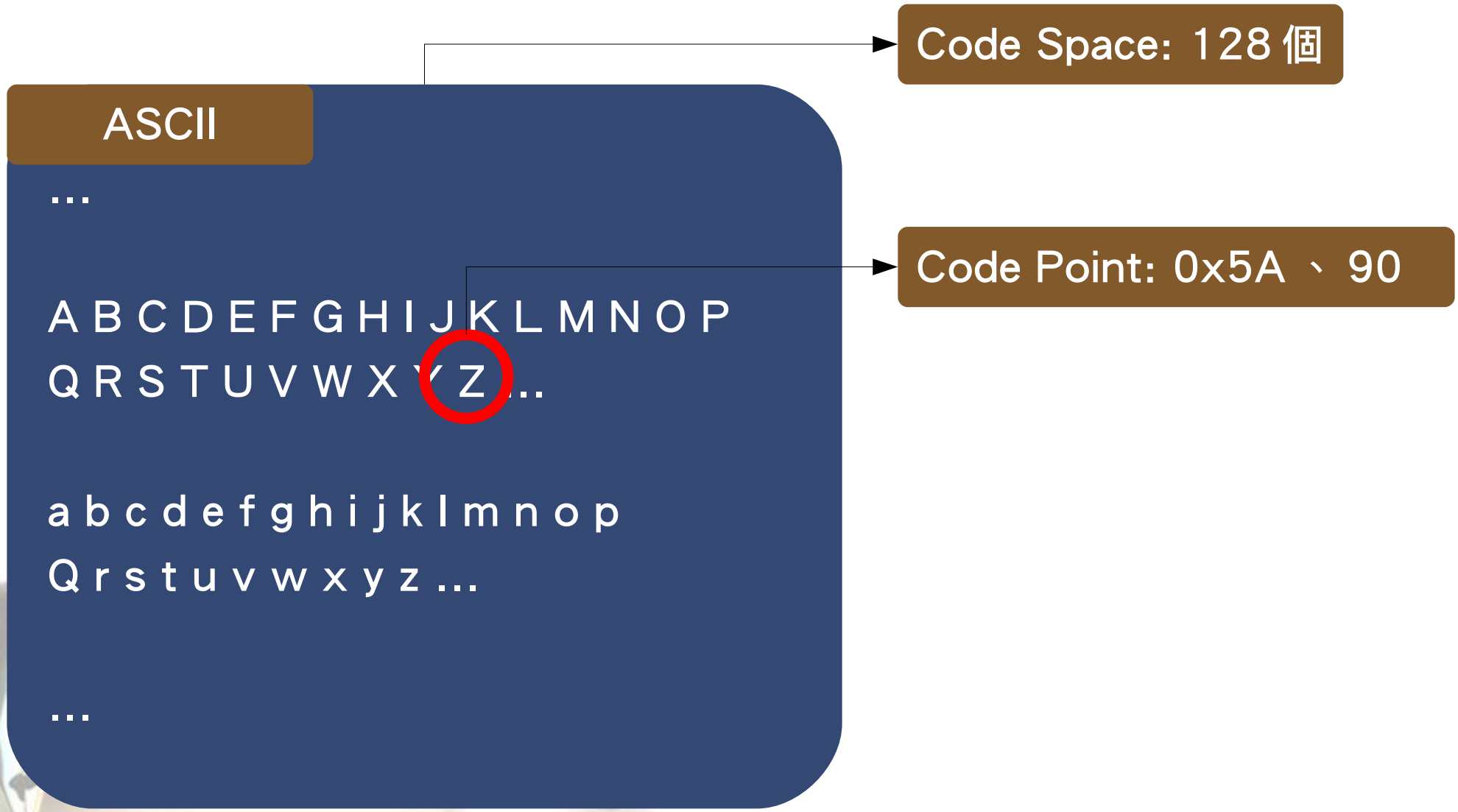
1. 一群字元的集合
2. 經常與 Character Encoding 名詞混用
3. 有 Universal Character Set (UCS-2 、 UCS-4)

- *Unicode 不等於 UTF-8 ， UTF-8 只是眾多 Unicode 編碼之一。*

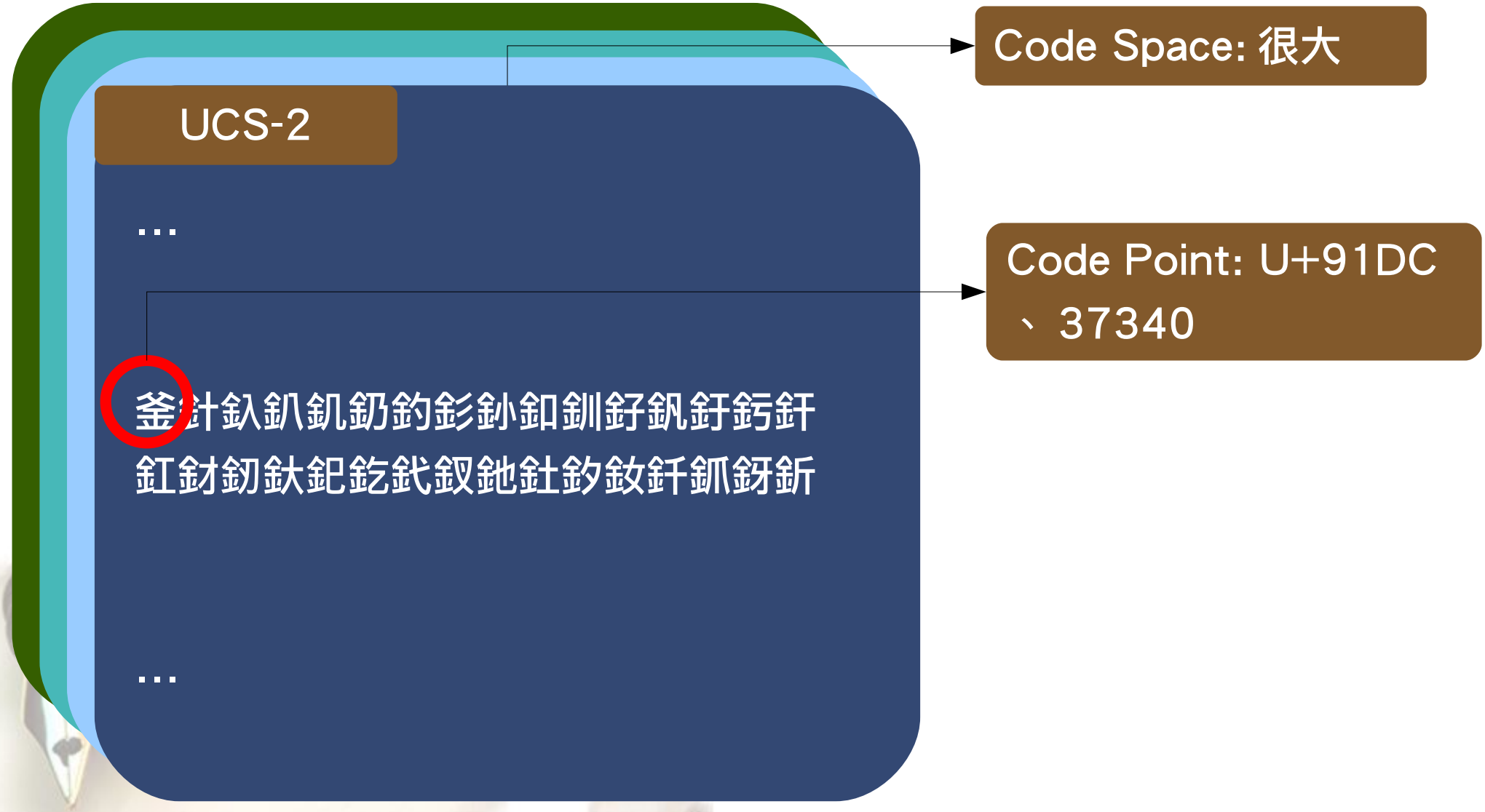
# 字元編碼
## Character Encoding

1. 給每個字元一個數值（octet），固定或變動長度
2. 大寫 Z 的實際編碼是 1011010 (0x5A 、 90)
3. 常見編碼有 ASCII 、 Big5 、 UTF-8 、 UTF-16

# Code Point (1/2)
## 字元集位置、空間



# Code Point (2/2)
## 字元集位置、空間



# 名詞釋疑
## 字元集？字元編碼 ?UCS-2?UTF-16

### 字元集？字元編碼？

- 字元集就是一群字元的集合，而字元編碼就是給這字元集內的每個字元一個整數數值，來加以表示。
- 一種特定的字元編碼可以容納一群字元集合。

### UCS-2?UTF-16?

- UCS-2 一種固定長度的編碼方式，每個字元需要 16 bits (2 bytes 或 2 octets)。
- UTF-16 與 UCS-2 類似，部份為變動長度編碼，每個字元可能是 16 bits 或 32 bits 。

# 固定長度、變動長度的編碼
## Fixed-length 、 Variable-length

### 固定長度

- 每個字元都用固定長度的數值來表示，如 1 bytes 、 2 bytes 、 3 bytes 、 4 bytes 或更多。
- 很簡單、方便計算字元數目。

### 變動長度

- 不同字元可能使用不同長度的數值來表示，可能有 1 bytes 、 2 bytes 、 3 bytes 甚至 4bytes 或更多。UTF-8 就是如此。
- 很複雜、不方便計算字元數目，需要特別去辨別與計算。
- 對於網路傳輸、檔案儲存較節省時間與空間。

- *固定與變動長度編碼是相對的，UTF-? 是變動、UCS-? 是固定。*

# Text String vs Binary String

## Characters or Bytes ?

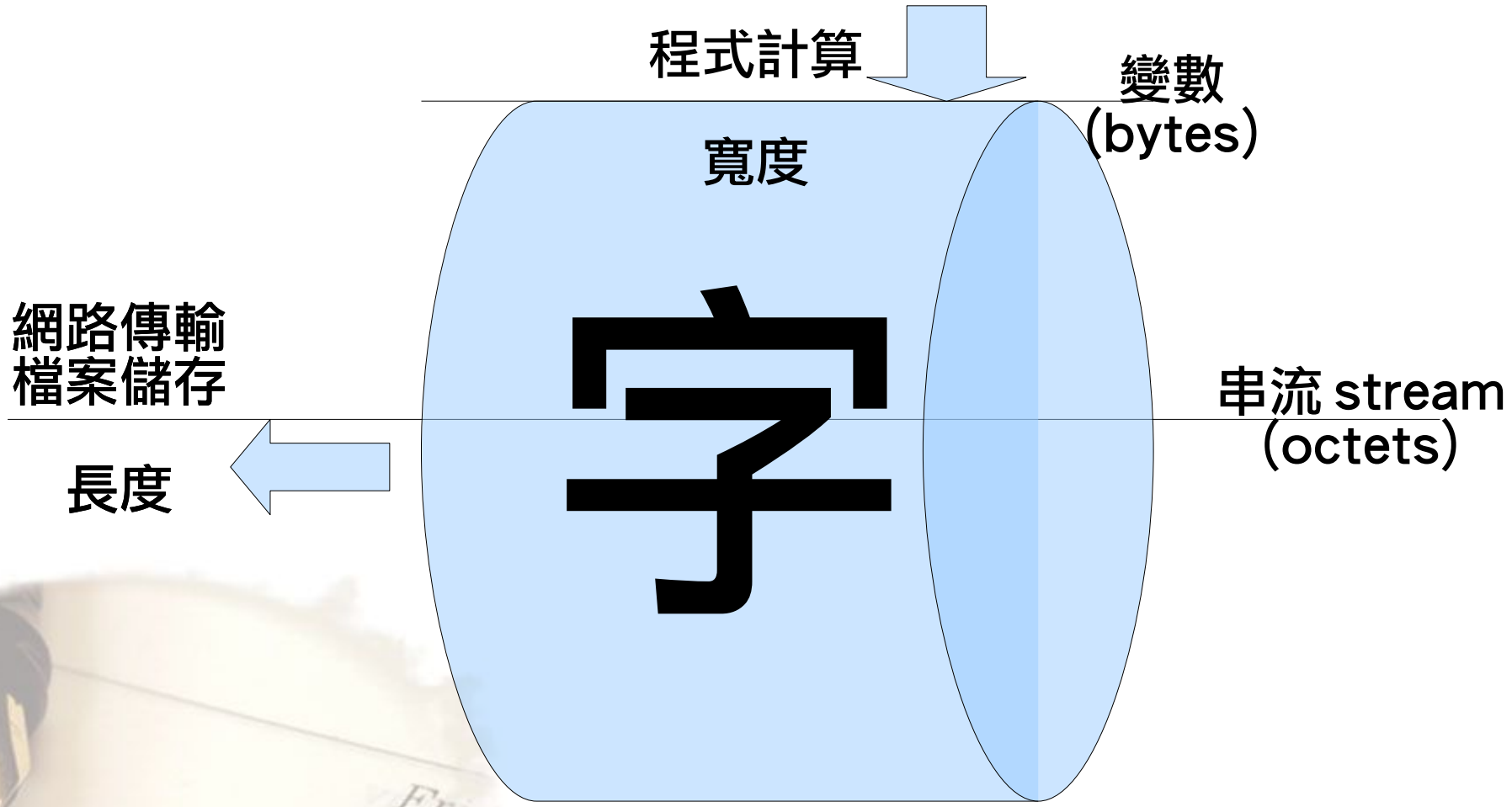
Text String (Characters)

| U+6E05 | U+83EF | U+5927 | U+5B78 | ..... |
|---|---|---|---|---|

Binary String (Bytes)

| 0xE6 | 0xB8 | 0x85 | 0xE8 | 0x8F | 0xAF | 0xE5 | 0xA4 | 0xA7 | 0xE5 | 0xAD | 0xB8 | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|---|

# 長度、寬度
## 觀看角度不同




- *In computing, an octet is a grouping of eight bits.*
- *Computer networking standards almost exclusively use octet.*

# 實際應用
## 瀏覽器網址列

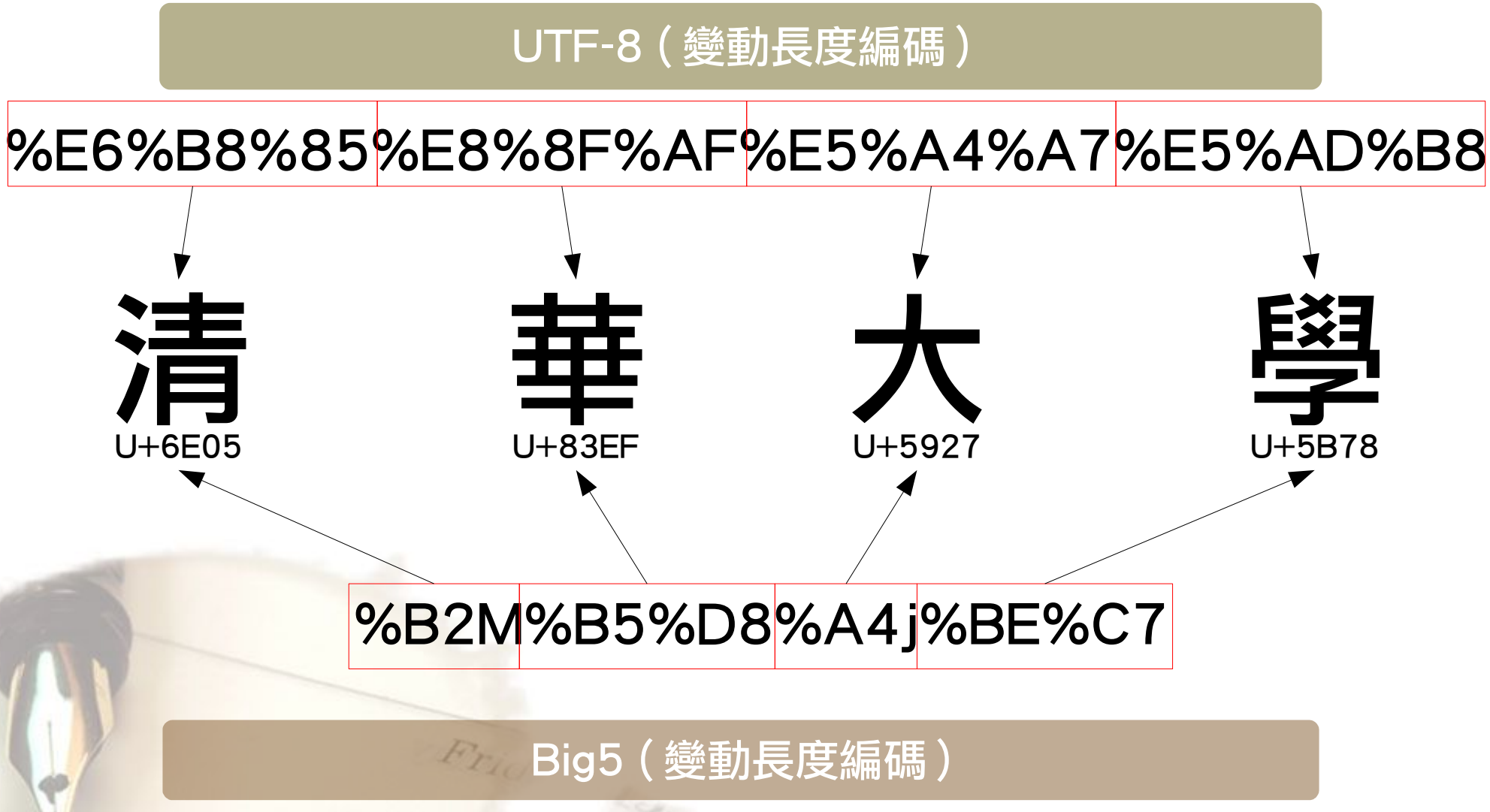
### 未編碼


清華大學

### 已編碼（UTF-8）


%E6%B8%85%E8%8F%AF%E5%A4%A7%E5%AD%B8

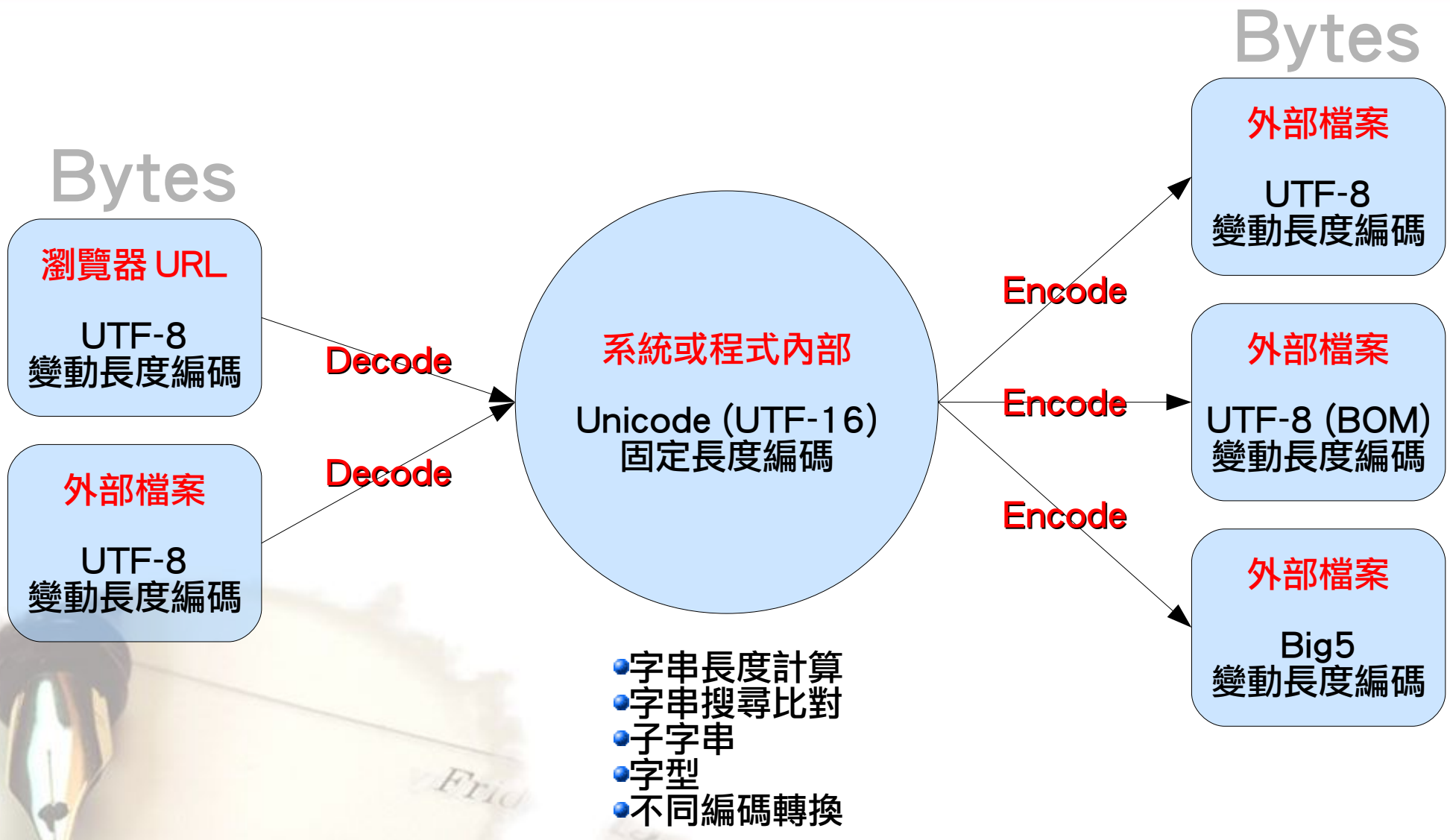
# 實際應用
## URL 的編碼與解碼



# 實際應用
## 檔案儲存

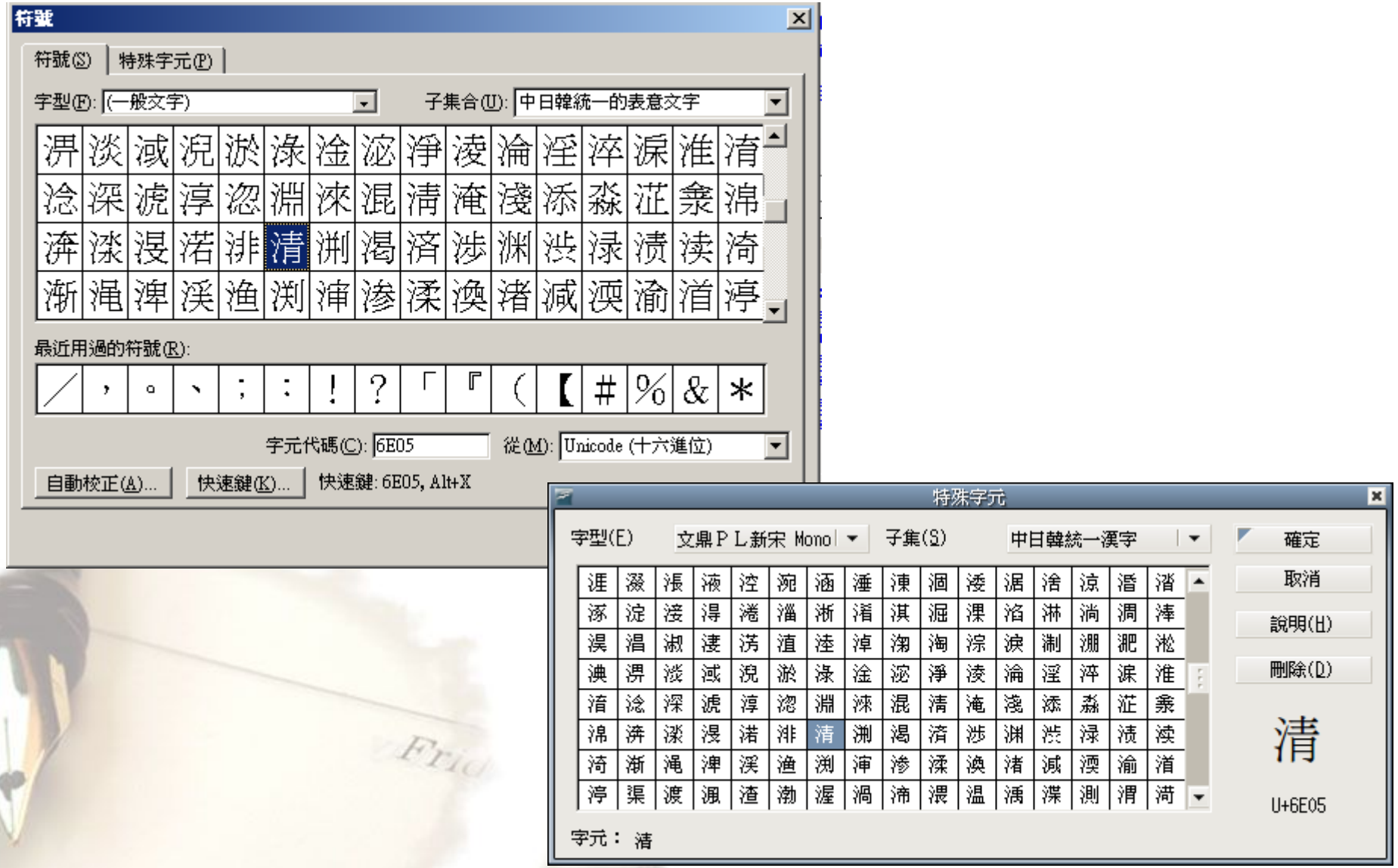
- **實際字串**
  清華大學 1234
- **Big5 (12 bytes)**
  4db2 d8b5 6aa4 c7be 3231 3433
- **Unicode (UTF-16 LE，18 bytes)**
  feff 6e05 83ef 5927 5b78 0031 0032 0033 0034
- **Unicode (UTF-8 with BOM，19 bytes)**
  bbef e6bf 85b8 8fe8 e5af a7a4 ade5 31b8 3332 0034
- **Unicode (UTF-8，16 bytes)**
  b8e6 e885 af8f a4e5 e5a7 b8ad 3231 3433

# 示意圖
## 內部與外部

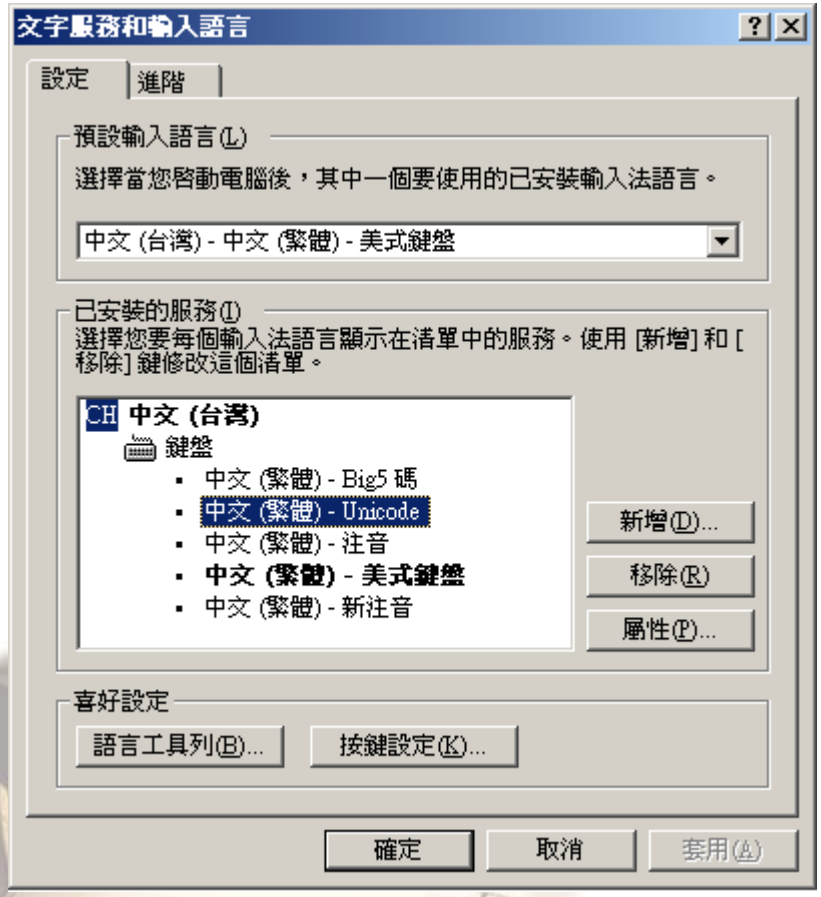

Bytes
瀏覽器 URL
UTF-8
變動長度編碼
外部檔案
UTF-8
變動長度編碼
Decode
Decode
系統或程式內部
Unicode (UTF-16)
固定長度編碼
字串長度計算
字串搜尋比對
子字串
字型
不同編碼轉換
Encode
Encode
Encode
Bytes
外部檔案
UTF-8
變動長度編碼
外部檔案
UTF-8 (BOM)
變動長度編碼
外部檔案
Big5
變動長度編碼

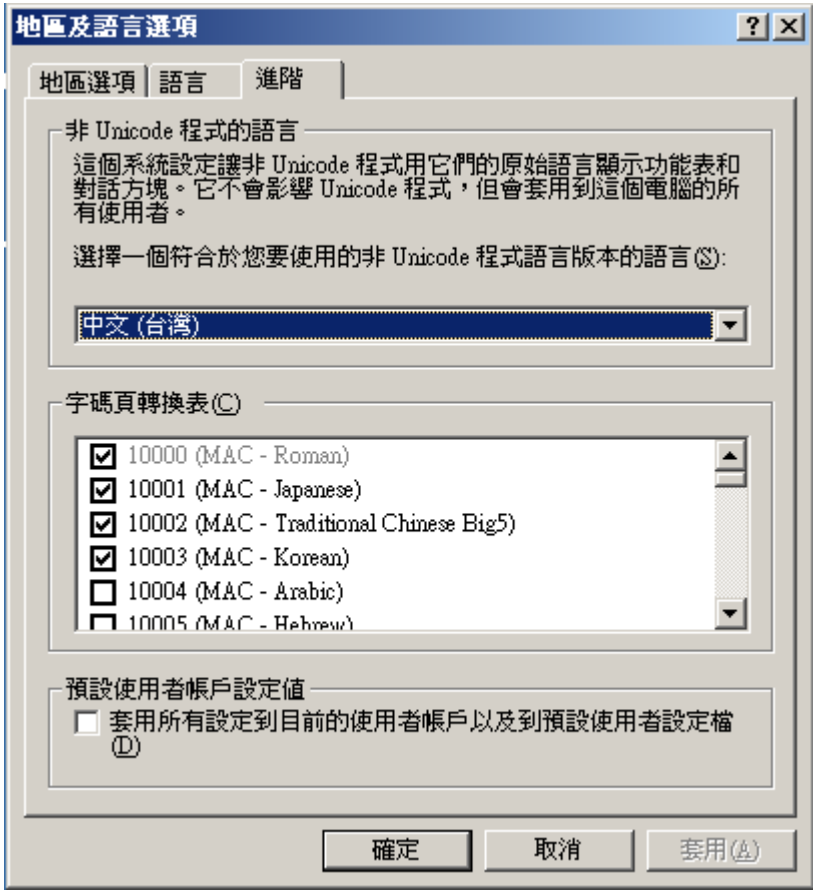
# 實際應用

## Word & OpenOffice.org 插入符號



# 實際應用
## 輸入法、 Code Page



# 與編碼相關的應用
## 儲存、傳輸

- 檔案內容、檔案名稱
- 網址、網頁內容
- 郵件
- Terminal
- FTP 傳輸
- GUI 環境
- 輸入法

# 常見字元編碼
## ASCII、CJK

1. 最基本的 ASCII
2. 中文漢字，CP950 / Big5、CP936 / GBK
3. UTF-8、UTF-16 / UCS-2、UTF-32 / UCS-4

# Thank you !