

編碼標準 Encoding Standard

□ 電腦是美國人發明的

• ASCII (American Standard Code for Information Interchange)

□ 地方的電腦也要顯示中文

- Big5
- 台灣財團法人資訊工業策進會 在 1983 年為 五大中 文套裝軟體 設計的編碼系統
- 繁體中文中最常用的電腦中文字符集標準
- 萬年遺毒

編碼標準

ASCII

- 8 bits (理論上有 256 種可能)
- 0x00 ~ 0x7F 共 128 種字元
 - > 0x00 \sim 0x1F \rightarrow control characters
 - > 0x20 ~ 0x7F \rightarrow printable characters

🗆 Big5

- 使用 2 bytes 來存放中文字 (理論上有 65536 種可能)
- • 實際上為與 ASCII 相容,只能使用 19782 個

 [0x81 ~ 0xFE][0x40~0x7E, 0xA1 ~ 0xFE] = 126 * (63 + 94) = 126 * 157 = 19782

Ref: http://www.cns11643.gov.tw/AIDB/encodings.do



Big5的問題

□ 使用者造字區

每個人都可以自己造字
 於是自己造的字放到別人電腦上就看不到

□ 缺字

• 堃、煊、栢、喆

□ 延伸版本繁雜

- 倚天Big5延伸
- Code Page 950
- Big5+
- 族繁不及備載..

□許功蓋問題(\)

- 0x5C (\) 會有特殊意義
- 許(0xB35C)、功(0xA55C)、蓋(0xBB5C)

編碼標準 - Unicode

全世界共有上百種文字,因此有很多種不同的編碼系統

日本有 JIS,中國有 GB 2312,... etc

同樣的編碼在不同的編碼系統下顯示會不同

Unicode 組織就誕生了!!!!

Unicode VS ISO 10646

□1991年左右,同時有兩個組織著手規範世界字碼

- Unicode
- ISO 10646

□過不了多久,他們就互相體認到 「這個世界不需要兩套不同的單一字符集」

• 因此他們決定共用同樣的字碼

□ 現在這兩個組織各自存在,各自互相砥礪

ref:<u>http://zh.wikipedia.org/wiki/ISO_10646</u>



ISO10646 and Unicode (1)

🛛 Goal

- 集結全球通用字符集,成一大聯集
- UCS-4

> Universal multiple-octet coded Character Set

- 4 bytes encoding (2^31)
 - > 128 Groups
 - > 256 Planes each group
 - > 256 Rows each plane
 - > 256 Cells each row
- BMP (UCS-2)
 - > Basic Multilingual Plane
 - > 00 group, 00 plane
 - > 65536 encoding space



➢ 若所有字集都在 BMP 中, 就可以只使用 2 bytes, 否則就要用 4 bytes, 不能混用



編碼標準 -ISO10646 and Unicode (2)

BMP



Unicode 的問題

Big Endian & Little Endian

- U+4E59? (乙)
- U+594E? (奎)
- □ 編碼空間浪費
 - ASCII 字元通通都用 2byte 表示: 0x00 0x41 A」第一位永遠是0

Γ

Unicode Transformation Format

□ UTF: UCS/Unicode Transformation Format

- UTF-16(2 \ 4 bytes)
 - ▶將一個 32-bit ISO10646 字元轉成多個 16-bit Unicode
 ▶ Windows
- UTF-8(1~4 bytes)
 - ▶將一個32-bit ISO10646 字元轉成多個 8-bit Unicode
 - ▶將一個16-bit Unicode 字元轉成多個 8-bit Unicode
 - ▶ 128個US-ASCII字元只需1 bytes編碼
 - ▶ 帶有附加符號的拉丁文、希臘文、西里爾字母、亞美尼亞語、希伯 來文、阿拉伯文、敘利亞文及它拿字母則需要 2 bytes 編碼
 - ▶ 其他基本多文種平面(BMP)中的字元(這包含了大部分常用字)使用 3 bytes 編碼
 - ▶ 其他極少使用的 Unicode 輔助平面的字元使用 4 bytes 編碼
 - >Unix-like systems

非常經典的 UTF-8...

□ 與既有系統的相容性

- 只包含 ASCII 0-127 的字串是合法的 UTF-8 字串
- NULL-terminated 字串處理

□ 極高的辨識性

• UTF-8字串可以由一個簡單的演算法可靠地識別出來。

□ 可以容納所有 Unicode 字元

- UTF-8 理論值可以容納百萬個字元 (實際是 1112064 個)
- (2012 年發佈的 Unicode 6.2 也才十一萬個字元)

□ Unicode 與 UTF-8 之間的轉換很方便

中文環境(1)



- 中文訊息
- 中文顯示
- 中文輸入
- 中文列印
- 中文處理

困難

簡單



□ 中文化方式

- 直接修改程式
 - ▶ 套件以排山倒海之勢而來只有真強者才能改完了Orz

18 chars

- 國際化(InternationalizatioN, i18n)
 - > Multi-language architecture
 - 程式設計人員按照該架構的機制與準則寫程式, 便可支援各式各樣的語言
 - > Locale (LOCALization Environment database)
 - 程式根據使用者選擇的 locale 聯繫到不同資料庫, 進而提供該語言的支援
- 在地化(LocalizatioN, L10n)
 ▶ 在 i18n 的大架構下,加入「在地化」的特性
- 通常i18n只需做一次,而L10n要針對每個語言個別做

i18n & L10n

🖵 i18n + L10n

- 語言/翻譯
- 文化、書寫習慣
 - ▶ 名字和稱謂的位置
 - ▶ 電話號碼,位址和國際郵遞區號的格式
 - ➤ 貨幣單位
 - ▶ 度量衡
 - ➤ 日期時間
 - ≻ 時區
 - ➤ 數字格式

□ L10n only

- 內容在地化
- 道德在地化
- 文化價值
- 社會環境

中文環境(3)

□ locale in FreeBSD

- 地區性語言的資訊
 - > LC_ALL
 - > LC_CTYPE
 - 掌管該 locale 中所有字元的處理方式
 - > LC_MESSAGES
 - 掌管程式訊息輸出所用的語言
 - > LC_TIME
 - 時間格式
 - > LC_NUMERIC
 - 數字格式
 - LC_MONETARY
 - 貨幣格式
 - > LC_COLLATE
 - 字母順序與特殊字元比較
 - > LANG
 - 語言顯示
- 效力優先性:LC_ALL > LC_* > LANG

中文環境(4)

□ 設定 locale

- csh/tcsh shell
 - > setenv LC_CTYPE en_US.UTF-8
- Bourne Shell
 - > export LC_CTYPE=en_US.UTF-8

Note: 可以寫在 .tcshrc/.bashrc 中登入後自動載入

- /usr/share/locale/
 - ▶各國的 locale 資訊
 - ▶ 命名規則: 語言_地區名.字元編碼名稱
 - zh_TW.UTF-8
 - zh_CN.GBK

中文環境 (5)

中文 Terminal (Remote Login)

- M\$ Windows: putty, pietty, netterm, multi-term, telnet, ...etc.
- X Window: xterm, rxvt, aterm, mterm, roxterm...etc.
- 設定好中文支援,登入後即可看到中文
 - > setenv LC_CTYPE en_US.UTF-8 (csh/tcsh)
 - > export LC_CTYPE=en_US.UTF-8 (sh/bash)
 - ▶ 顯示為英文但支援 multibyte characters

中文 Xwindow

- 建立支援 L10n 中文環境
- > 安裝中文字型
- ▶ 設定 Shell locale 環境
- ▶ 安裝中文輸入法 (Ex. ibus)

Steps





□兩大中文字型種類

- 點陣字型 (Bitmapped Font)
 - ▶ BDF (Bitmap Distribution Format) 點陣分散格式
 - ▶ HBF (Hanzi Bitmap Font) 漢字點陣字體

>PCF (Portable Compiled Font)

● 曲線描邊字型 (Outline Fonts)
 > True Type Font (TTF)



Font Path

- % xset q
- % xset fp+ [directory]
- % xset fp rehash

Font Path: /usr/local/lib/X11/fonts/misc/, /usr/local/lib/X11/fonts/TTF/, /usr/local/lib/X11/fonts/Type1/, /usr/local/lib/X11/fonts/100dpi/, /usr/local/lib/X11/fonts/local/

□安裝字型

- 1. 直接從 Windows 下偷過去
- 2. 透過 ports 安裝字型檔案
- 使用 fc-cache 建立字型資料庫
- 修改各軟體設定使用別的字型





□ Fireflyttf

- 透過 ports 安裝的都會自己跑
 - >% ttfm.sh --add xttfm /usr/local/share/fonts/TrueType/fireflysung.ttf
 - >% fc-cache –f –v /usr/local/lib/X11/fonts/TrueType/
- portmaster chinese/fireflyttf



] 增加 Font Path

• Edit /etc/X11/xorg.conf

>/usr/local/share/fonts/TrueType/fireflysung.ttf

>/usr/local/lib/X11/fonts/TrueType/fireflysung.ttf

• Restart xwindow

Section "Files" ModulePath "/usr/local/lib/xorg/modules" FontPath "/usr/local/lib/X11/fonts/misc/" "/usr/local/lib/X11/fonts/TTF/" FontPath "/usr/local/lib/X11/fonts/OTF" FontPath "/usr/local/lib/X11/fonts/Type1/" FontPath "/usr/local/lib/X11/fonts/100dpi/" FontPath "/usr/local/lib/X11/fonts/75dpi/" FontPath **EndSection**

symbolic link



安裝中文 Terminal

rxvt-unicode

/usr/ports/x11/rxvt-unicode

🗅 aterm

/usr/ports/chinese/aterm

🛛 eterm

/usr/ports/chinese/eterm

ROXterm

- /usr/ports/x11/roxterm
- 🗅 mlterm
 - /usr/ports/x11/mlterm

ROXterm

□ X11/rxvt-unicode □ roxterm-config

実際用	程式選単 ▋ [● 1a < *・	··· •• •• •• •• •• ••	RO 🔄 ro 🔄 ro	. 🔄 ro 📄 [wi 📓 […			16:03 Charlie Root
<u> </u>				ROXTerm			+_∂×
<u>F</u> ile <u>E</u> d	it <u>V</u> iew <u>S</u> earch <u>P</u> referen	nces Ta <u>b</u> s <u>H</u> elp					
💥							
【板主	: starblue/a28124209	/saMing1371/Mod]		〔板〕開學囉!!!	« NCTU	>	<u> </u>
[←]離日	<u>用 [→] 因 讀 [^P] 被 衣</u> 日 曲 作 老	[D] 進 板 量 面 [TGL] 授 等 立 音 栖 厢	[V]投票 [S] 速達有板		課・○ 人気・367	_	
53977	+ 10/01 james1224		碟		計.0 八乘.307		
53978	+ 1 10/01 sandraLJ	◇ [交換] 全家海賊					
53979	2 10/01 riko193	○ [急微] ASUS小筆	電充電器				
53980	+ 9 10/01 monkeynea	□ ◇ [1)育報] 日本語既 ◇ 「明晰」 車座 0	調比費 02 「 棚 小能後尼倫駅園」				
53982	+12 10/01 kevinlee8	→ [周期] 単件9 10 ◇ [失物] 九金十金	交誼廳的椅子				
53983	+ 10/01 qwer1234	◇ [家教] 高二 數					
53984	+ 1 10/01 jenjye 01	Re [失物] 九舍十舍	交誼廳的椅子				
53985	+24 10/01 lin50513	○ 一年一度的闌陽」	周來囉!!!	~			
53980	+33 10/01 Openhouse	◇【網旦】UN研設省 ate △【網官】101巻年1	代位弗兰迪场头 对位达不元 64星藤委正式開駒曜!!!				
53988	+ 7 10/01 error	◇ [情報] English	Lunch Time 回來囉!!!!!!				
53989	+ 7 09/20 abc5558168	8 🛇 [++++ 男綱女綱材	交隊選拔賽 ++++				
53990	10/01 ltrc	◇ [徵人]交通大學	英教所誠徵國科會兼任助理				
53991	+ 4 10/01 e40614	◇ 【黒符】13舍四樓2	光衣機 1月				
53003	10/02 pp012345	○ [黑行] 厚変唱な ○ [譜間] 請問NCTI	/笏 J大·袖				
53994	+27 10/02 JirenAdmin	n ◇ ○★▽及人服務	團◆期中隊徵才▽★⊙				
53995	+ 1 10/02 mT∨Tm	◇ [協尋] 01560212	學生證				
53996	+ 3 10/02 Opethmist	◇ [請問] 賣腳踏車	的地方	*			
53997	+ 3 10/02 WanChinCh:	1en 🛇 してご ★ 10/4 既 に 🔿 「重生四…なた油」	上 工四102 手語社課第二第				
53000	+35 10/02 blue1203	· ∨ [里王·□~又八碗) ○ [徼遲] 値科系』	■ 業列/A 動 ■ 10/04(A) 隆重 ■ 業 製 作 演 昌	47F (FR)			
54000	+ 4 10/02 starblue	◇ [~ ☆★*:.・汪注	王社迎新茶會 就在今晚・.:	*★☆~			
54001	+ 1 10/02 rmjheng	(BetaSite]IPv6	暫停服務說明				
54002	+ 4 10/02 brian40820	002 ◇ [情報] 二餐極麵	道徵工讀生				
54003	+ 2 10/02 tony41004	100 ◇ 【 追款】 借用 溜 ぶ	場 海 / 本 半 干 积 织 新 戓 ウ ! !				
54004	+ 3 10/02 596387799	✓ [炳旦] < ◇ [情報] 聽說禮拜	四有個大活動				
54006	+ 0 10/02 wouldrath	er 🛇 [贈送] 烤箱和電	磁爐(已送出)				
54007	+ 3 10/02 OpenHouse	◇ [尋人] OH座談會	餐盒				
54008	+ 10/02 文書組公告	; ◇ [文書組公告]宿;	全郵件領取通知	2 300 AND DE AA Die AD - Adr Jon July 1			
54009	+ 10/02 nyp.	〇 【選課公告】10, ○ 伸伸用可無照対	F1-5日学生書面週期加退選 , 社調問始醒, 毎日四晩上	() 速 課 敢 後 憎 段 - 請 把 握 。			
54010	$+ \frac{10}{02} n5565294$	◇ 师研研可列回社 ◇ 【情報】活動中心	·吐林用加强,每周四晚上 理髮部 新生煎髮體驗半價決	自動 開 狗			
54012	10/02 thisissta	r 🔾 [誠徽] ASUS筆電	充電器				
重要	b53 05/23 ensky	◆ [公告] BS2實用	看板介紹				
重要	b14 05/20 ensky	○ [公告] 板規 V3.	2 Not noted 646, 187, once million 166, only shift 64- 187, 64-	877 1.1 \			
● 里 受 → 音 列	▶ (i)呼叫小鏨毛 (────────────────────────────────────	實驗恢物突擊駕宣報告(附 (Tab/z)進入請華區(h)迨	照月)			
~ 7	(*/) ()	/1x 5/1/ A2 (•/1x L A2 4	(1)01	//			······
I	T A A			1 📻 🖉 📃 II			



Choices

- ibus-chewing(chinese/ibus-chewing)
- ibus-pinyin(chinese/ibus-pinyin)



安裝 ibus 中文輸入程式 (1)

ibus

- Intelligent Input Bus
 - 1. % cd /usr/ports/textproc/ibus-chewing ; make install clean
 - 2. setenv LC_CTYPE zh_TW.UTF-8 (csh/tcsh) export LC_CTYPE=zh_TW.UTF-8 (sh/bash)
 - 3. Edit .xinitrc(或是可以 setenv in .cshrc/.bashrc)

XIM=ibus GTK_IM_MODULE=ibus QT_IM_MODULE=xim XMODIFIERS='@im=ibus' XIM_PROGRAM="ibus-daemon" XIM_ARGS="--daemonize --xim"



安裝 ibus 中文輸入程式 (2)

□ ibus相關設定

- % ibus-setup (UTF-8)
- 可以加入 Chewing

\$	lBus 偏好設定	+ _ □ ×			
通用 輸入法 進階					
選取輸入法	~	♣ 加入(<u>A</u>)			
輸入法		移除(<u>R</u>)			
1111 (1111) (11111) (11111) (1111) (1111) (1111) (1111) (1111) (1111) (1111) (wing	▲ 向上(U)			
		● 何下(□)			
		☆ 關於(<u>A</u>)			
		■偏好設定(<u>P</u>)			
清單内最頂端的 悠可以使用向。	9輸入法為預設輸入法。 上/向下 按鈕進行變更。				
		₩開(<u>C</u>)			



安裝 ibus中文輸入程式 (3)

4. Switch to chinese input: Ctrl -Space



References

□ 中文碼介紹

http://www.cns11643.gov.tw/web/word.jsp

□ FreeBSD Chinese HOWTO

• http://netlab.cse.yzu.edu.tw/~statue/freebsd/zh-tut/index.html

□ Introduction to i18n

http://www.debian.org/doc/manuals/intro-i18n/

□ Unicode 介紹

<u>http://www.csie.ntu.edu.tw/~p92005/Joel/Unicode.ht</u>
 <u>ml</u>