

私立東海大學資訊工程與科學研究所  
碩士論文

指導教授：蔡清欉

繁簡體轉換在 ERP 系統的應用與研究  
The Transformation of Complex and  
Simplified Chinese in ERP System



研究生：許旭正  
中華民國九十五年一月

## 誌謝

在職進修的求學過程中，常常是學校、公司、家庭三頭忙，能夠順利的完成此篇論文，最感謝的是指導老師 蔡清欉教授的協助。老師除了教導學生在資料蒐集方面的訣竅外，最重要的是在研究論文的每個瓶頸點，皆能提供適當的建議及重要的思考的方向，讓學生能明確掌握每個細節及研究的重點。

另外特別感謝的是同窗好友也是公司同事的士哲，除了課業上問題協助解答外，更感謝他熱心教導我如何完成我的第一次期刊投稿，無論是投稿單位資料的搜尋或是如何構思期刊的內容，皆能不藏私的傾囊相授，真是一位難得的好同學。

最後感謝家人全力的支持，感謝老婆幫忙照顧家中常感冒的小寶貝，讓我在求學過程中無後顧之憂；感謝小兒子常在我的報告上塗鴉，讓我有機會重寫報告而更熟悉裡面的內容。謹以這份論文完成的榮耀與喜悅獻給我最愛的家人。

## 摘要

隨著全球經濟型態的轉變，國內許多企業也逐漸突破國界限制而在國外設立了許多據點，利用人力成本的降低或是藉助外來的高科技人才，來提高本身競爭力，並維繫整個企業的生存。協助企業電腦化的 ERP(Enterprise Resource Planning)系統已面臨到必須發展出滿足不同語系使用者的操作介面，能夠找到一套適合台灣及大陸使用者可以共用的繁簡體共存的 ERP 系統，實屬必要。

尤其近年海峽兩岸文化及商業互動日漸頻繁，彼此間許多文件及書信往來日漸增多，雙方使用者除了必須適應繁簡中文字的不同之外，另外還必須學習兩方用詞的差異，才能充分了解資料上的訊息。為了解決使用者未來在文字閱讀上的隔閡，增進兩岸資訊快速交流，繁簡自動轉詞的研究顯的更加重要。

本研究目標在於提出快速簡單的繁簡體轉換方法以及在 ERP 系統上的應用，期使能讓企業在舊系統移植上或新系統的開發上更能符合使用者的需求。本研究的方法乃運用中文字具有固定內碼的特性，並結合兩岸用詞差異所建立的詞庫，透過斷詞的技術，來達到 ERP 系統繁簡字或詞的快速轉換。研究中所提出的方法並不侷限應用於 ERP 系統，諸如繁簡網頁的運用，兩岸三地資料的傳遞皆可適用。能夠幫助企業花最少的成本，快速建置具有彈性且合用的繁簡 ERP 系統是本研究最終的目標。

關鍵詞：繁簡轉換、轉詞、ERP 系統、BIG5、GB2312

## ABSTRACT

With the global economic mode, internal enterprises are going to break through the limit of national boundary to many footholds abroad and reduce manpower cost or with the help of high-technology adventitious elitists to enhance their own competitive capacity and support the survival of the whole enterprises. Enterprise Resource Planning (ERP) system that bears a part in enterprise computerization is also facing must to develop operational interface for satisfying various languages users. It is necessity that we should find a set of ERP system that traditional and simplified Chinese could be coexisted to fit for users from Taiwan and China.

This paper focus on offering the fast and simple transformational way of complex and simplified Chinese and the application in the ERP system, and hope it could make enterprises more corresponding the require of user s in the old system's transplantation or new system's exploitation. The approach of this paper is to use the trait of Chinese in which has faxed inner-code, associate the lexicon in which is built by the difference of using words between Taiwan and the Chinese mainland and achieve the fast transform of complex and simplified Chinese in the ERP system. The method is not limited to the ERP system, it could be extended to the use of complex and simplified internet-page and for the transferring information among Taiwan, Hong Kong, and Mainland. Most importantly, this study is to be able to help enterprises spending the least cost and quickly building the elastic and suitable ERP system.

Key word: Transformation of Complex and Simplified Chinese 、

Word Translation 、 ERP system 、 BIG5 、 GB2312

# 目錄

摘要	i
英文摘要	ii
目錄	iii
圖目錄	v
表目錄	vii
第一章 緒論	1
1.1 研究動機	1
1.2 研究目標	1
1.3 論文大綱	2
第二章 相關研究	4
2.1 ERP 的概念	4
2.2 中文內碼介紹	5
2.3 中文字的概念	9
2.4 轉詞的概念	10
2.5 中文轉詞發展概況	12
第三章 系統開發與實例說明	15
3.1 ERP 系統的架構	15
3.2 繁簡轉換的對象	16
3.3 繁簡轉換在兩岸三地 ERP 系統的運用	21
3.4 兩岸 ERP 習慣用詞的比較	24
3.5 BIG-5 碼與 GB2312 碼轉換原理	26
3.6 BIG-5 碼與 GB2312 碼轉換元件	29
3.7 中文轉詞架構	30

3.8 系統模型.....	34
第四章 效能分析.....	50
4.1 字對字轉換.....	50
4.2 詞對詞轉換.....	51
4.3 資料內容轉換.....	52
第五章 結論.....	53
參考文獻.....	54

## 圖目錄

圖 3.1 WorkFlow ERP 系統 3-Tier 架構	15
圖 3.2 ERP 操作介面的轉換	17
圖 3.3 報表資料內容的轉換	18
圖 3.4 圖檔的轉換	19
圖 3.5 提示訊息的轉換	20
圖 3.6 兩岸三地系統整合架構-資料庫分散	22
圖 3.7 兩岸三地系統整合架構-資料庫集中	22
圖 3.8 兩岸三地系統營運模式	23
圖 3.9 單一中文字組成特性	26
圖 3.10 BIG-5 碼與 GB2312 碼對應陣列建立的流程圖	28
圖 3.11 轉換元件	29
圖 3.12 對應字碼維護工具	30
圖 3.13 轉詞架構圖	31
圖 3.14 建檔畫面詞彙轉換	32
圖 3.15 報表資料詞彙轉換	33
圖 3.16 訊息說明詞彙轉換	33
圖 3.17 使用者自行定義轉換詞句的工具	33
圖 3.18 TcodeConvert Class Diagram	36
圖 3.19 轉換流程，(a)字與字的轉換，(b)詞與詞的轉換	36
圖 3.20 登入畫面語系的選擇	37
圖 3.21 工具程式類組成架構	37
圖 3.22 ERP 工具類程式的畫面轉換	38
圖 3.23 ERP 圖形類的轉換	39

圖 3.24 建檔類繼承架構圖 .....	40
圖 3.25 建檔類程式架構圖 .....	40
圖 3.26 建檔類畫面與資料內容的轉換 .....	41
圖 3.27 報表類繼承架構圖 .....	42
圖 3.28 報表類程式架構圖 .....	42
圖 3.29 報表類畫面與資料內容的轉換 .....	43
圖 3.30 批次類繼承架構圖 .....	44
圖 3.31 批次類程式架構圖 .....	44
圖 3.32 批次作業類畫面與資料內容的轉換 .....	45
圖 3.33 查詢類繼承架構圖 .....	46
圖 3.34 查詢類程式架構圖 .....	46
圖 3.35 查詢作業類畫面與資料內容的轉換 .....	47
圖 3.36 憑證類繼承架構圖 .....	48
圖 3.37 憑證類程式架構圖 .....	48
圖 3.38 憑證作業類畫面與資料內容的轉換 .....	49
圖 4.1 轉字效能測試數據圖 .....	50
圖 4.2 轉詞效能測試數據圖 .....	51



## 表目錄

表 2.1 BIG-5 碼系統的內碼分佈	7
表 2.2 台灣地區所使用的 BIG-5 碼主要版本	8
表 2.3 GB2312 字符集的區位分布表	8
表 2.4 繁簡用語差異表	10
表 2.5 多對一關係的字根詞語	11
表 2.6 繁簡詞彙表達方式不同的詞語	12
表 2.7 繁簡轉換軟體對於轉詞的支援狀況	14
表 3.1 兩岸 ERP 習慣用詞比較表	25
表 3.2 兩岸 ERP 使用者操作畫面習慣用語	26
表 3.3 BIG-5 與 GB2312 字碼對照表	27
表 3.4 BIG-5 及 GB2312 標準及常用字碼分佈的範圍	27
表 3.5 BIG-5 及 GB2312 字碼唯一編號運算式	28
表 3.6 BIG-5 與 GB2312 字碼相互對應之陣列	28
表 3.7 鼎新 ERP 系統開發軟體及環境	35
表 3.8 ERP 外購元件廠商資訊	38

# 第一章 緒論

## 1.1 研究動機

隨著台灣與大陸三通腳步的逐漸加快，兩岸間不僅個人互動頻繁，企業如何依賴彼此間的經貿往來互相依存，更是考驗著兩岸企業主的經營智慧。隨著經濟型態及規模逐漸擴大，企業所面臨環境與運作的複雜性也相對提升不少，資訊科技幫助企業在競爭激烈的環境取得先機已是不爭的事實。但是要如何有效的規劃、管理與持續改善企業資訊運用卻是一項極大挑戰。許多企業選擇導入 ERP 系統來取代一些傳統人工作業方式，藉由電腦化的準確與效率，提升公司的對外競爭力。

近年來由於兩岸政策的開放，另一種兩岸三地的經濟模式慢慢受到大家的重視，許多整合的方案及想法也被熱烈的討論。為了因應此經濟發展趨勢，許多企業開始著手規劃內部舊有資訊系統的研發以及改良，目的除了要克服雙方使用者文字隔閡的問題之外，另外還必須滿足雙方使用者同時共用一套系統的需求，所以提供一個快而有效率的解決方案，是本研究主要的動機。

## 1.2 研究目標

由於全球資訊網的普及，許多人早已習慣由網際網路來獲取最新的資訊，除了英文網站外，中文網站是大家第二個最常接觸到網站。一些比較體貼的網站會為使用者提供不同語系的選擇，來滿足來自全球不同使用者資訊的取得。在商業應用軟體上，由於台灣企業橫跨兩岸三地的經營模式漸漸盛行，企業內部所使用的一些軟體也慢慢具備有多國語系的選擇。了解目前兩岸對於繁簡中文轉換的發展狀況及未來趨勢，以及在企業界運用的成果，是本研究首要的研究目標。

另一方面，為了進一步驗證理論上所提的方法是可以實際應用於實務上，研究過程中特別以台灣目前較為知名的 ERP 系統來作為實作的對象，透過中文字內碼的編碼原則，並結合現有的轉詞技術，實際分析整個系統模型運作的方式及效能，以及後續所顯現的效益。藉由實際案例的驗證，希望能達到拋磚引玉的效果，吸引更多人投入兩岸文化及經濟資訊交流方面的研究，這是本研究另一個重要的目標。

### 1.3 論文大綱

本文是從中文字內碼編碼的特性來探討繁體中文與簡體中文轉換的方法，並根據此方法建置企業 ERP 系統中文轉換模型。繁簡中文字對字單純轉換為目前比較常見的方式，但是因為兩岸用詞的差異，這種字對字的轉換方式並不能滿足使用者的要求，所以衍生出來的轉詞技術就相形變的重要了，因此字與字的互轉以及詞與詞的互轉是本文研究的兩大重點。在本文中，我們以台灣業界較為普遍的 ERP 系統來驗證我們的研究方法及成果，同時在整個系統模型架構完成後，我們更針對整個轉換過程中的效能及所耗用的資源做一個統計，藉以說明本研究所提出的方式是確實可行的方案，以及和其他研究作比較，展現我們研究的成果。

在第二章中，我們先回顧整個 ERP 的歷史以及概念，最後再說明目前在業界應用的狀況。接下來我們針對中文字的架構、內碼的概念做比較深入的介紹，因本研究是運用中文內碼的特性來做為繁簡轉換的基礎，所以此部份是本文比較重要的部份。另外在本章中我們也針對轉詞的部份作深入的探討，除了分析現在兩岸用詞方面的差距外，另外也導入轉詞的方法到我們實作的系統模型，藉以補強字對字轉碼的功能不足，強化我們研究的目標。

第三章是本文的重點。我們根據相關的學理以及目前兩岸中文字使用的狀況，提出我們建議的系統模型。在這個模型中，分別定義出

雙方中文內碼的分佈範圍及相互關係，進而得到繁簡互相轉換的公式。另外針對中文比較困難的轉詞部份，也參照兩岸的用語習慣，建立相對應的詞庫，藉由中文內碼的特性建立 Hash Index 來解決中文繁簡體在用詞方面的複雜對應關係。在本章最末我們再依照前述的理論及方法實際架設出整個系統模型，藉由系統的實作來分析所用的方法以及呈現出來的效果。

第四章內容主要是根據實驗的結果作一個效能評估，針對字轉字的效能以及詞轉詞的效能做一個分析圖，藉以驗證我們提出的方法是可行的。最後再利用資料庫內存的龐大資料量做壓力測試，得出整個轉換架構所能承載的資料量。

第五章將根據整個系統模型作一個檢討，並提出本研究的結論。

## 第二章 相關研究

關於繁簡中文轉換研究上主要分成兩個方向：第一是探討字對字的轉換，第二是探討詞對詞的轉換。第一部份研究著重在利用中文字內碼特性，嘗試找出其編排的規則性，進而利用此特性來構思轉換的方法。第二部份的研究則專注於兩岸中文用詞的差異性，根據其中文的相互關係找出對應邏輯，再運用此對應邏輯建立相互對照的詞庫，進而達成我們詞轉詞的目標。在下面的幾個小節中，我們將由 ERP 的介紹來揭開序幕，經由中文字的剖析來慢慢切入主題，最後再以目前轉詞的發展概況作為結尾。

### 2.1 ERP 的概念

企業資源計劃，即 ERP (Enterprise Resources Planning)，是運用資訊整合的技術，以系統化的管理概念，建立企業決策資源及資源分配管理的操作平台。ERP 系統結合電腦的資訊技術與先進的管理概念，建立現代企業營運的運作模式，幫助企業能掌握公司內最新的經營狀況，進而能更合理調配資源，滿足創造社會財富的要求。所以 ERP 系統已成為企業在資訊時代下取得生存、發展不可或缺的重要基石。在 ERP 領域的範疇上，我們可以從管理概念、軟體產品、管理系統三個層次來定義它：

- 1) 是由美國著名的電腦技術諮詢和評估集團 Garter Group Inc. 提出了一整套企業管理系統體系標準，以 MRPII (Manufacturing Resources Planning) 為基礎，進一步發展而成的雙向供應鏈 (Supply Chain) 的管理概念。
- 2) 是整合了客戶服務體系、關聯式資料庫結構、資料交換技術、圖形用戶界面、網際網路通訊等資訊產業成果，應用層面涵蓋了『製

『製造管理』、『品質管理』、『配銷管理』、『存貨管理』、『財務管理』、『人力資源管理』及『決策支援』等等多項範疇，以 ERP 管理思想為基礎的軟體產品[1]。

- 3) 是結合了企業管理理念、業務流程、基礎數據、人力物力、電腦硬體和軟體於一體的企業資源管理系統。導入 ERP 觀念及系統後，以電腦自動化為核心的企業級的管理系統更為成熟，並增加了包括銷售預測、製造配銷能力、調整資源調度等方面的功能，有效促成企業全面流程自動化，即時掌握企業內經營資訊。

## 2.2 中文內碼介紹

無論在台灣或大陸，中文是目前彼此溝通的主要工具，雖然大陸採用簡體中文而台灣則採取繁體中文，但是中文字基本架構是相同的。如果要對中文內碼下一個簡單地定義，可以說就是：中文在資訊處理系統內部最基本一種表達形式，可供作為資料的儲存、傳送或處理等用途。例如：臺灣業界標準內碼—BIG-5 大五碼，目前通用於個人電腦、終端機、印表機等[2]。關於內碼方面，因為需要配合不同平台、不同電腦系統、與不同的應用環境，故目前並沒有一套是真正的國家標準。但是目前在整個資訊產業界中，PC 的使用比率還是佔了絕大的消費市場，所以通用於 PC 間的中文 BIG-5 碼就變為使用最廣的一種內碼，故我們也可以說 BIG-5 碼是中文內碼在台灣의 業界標準。

當在談到中文碼時，常常會提到一些專有名詞，比如字面、字集、字數、字碼、字序、字形等等，針對上述這些名詞，簡要說明如下：

- 1) 字面：整個中文字庫的結構可以想像成一個二維矩陣，由 X、Y 兩軸線(代表高、低位元組)分別代表每個字所在字庫的位置，每一

個矩陣單位內皆存在一個中文字。一般七位元環境下的一個矩陣，共有 94x94 個編碼矩陣單位(扣除控制字元)。而雙八位元組環境下的一個字面則有 256x256 個編碼矩陣單位。

- 2) 字元：即俗稱的「字」，跟英文字元一樣，是一組已獲認可且通行於電腦間之表達符號，集合多個字元可構成一句話或一篇文章。字元可以是字母、數字、標點及其他的符號，常以分開或連接在一起的形態來表示，不同的排列組合可產生不同溝通效果。
- 3) 字集：由不同字元所組成之有限集，它們係一完整且已被認同的字元集，不同的地區使用不同的字集。現階段電腦系統均採用已具共識地相同字集，再依不同需求訂定不同的字碼。例如：BIG-5、EUC、TCA 碼的字集均相同，都採用 CNS 11643 中文標準交換碼的字集。目前全世界存在許多不同的字元集，每個國家皆有自己標準共通的字元集。
- 4) 字數：即每個完整字元集中的全部字元的總數。例如：CNS 11643-1992 中文標準交換碼共有 441 個符號和 48,027 個中文字；BIG-5 碼共有 441 個符號和 13,053 個中文字。不同字元集的字數不一定相同，繁體(BIG-5 碼)的字數就多於簡體(GB2312)的字數。
- 5) 字碼：表示每個字元在字集中的編碼。依照既定之編碼範圍，將一固定之字集依照某種訂定的方法排列之，此編碼具有唯一性，即每個字皆有唯一的編碼方式。如 217E 表示某字元內部存放形式的高位元組為第 2 區，低位元組為第 94 位。
- 6) 字序：不同的字集中每個字元排列順序所採用的方法。例如：CNS 的每個字面均依先筆畫，後部首為序；電信碼則依先部首，後筆畫的字序排列。
- 7) 字形：字的外觀或是字元表現的形態，也就是以電腦能懂的方式，

將字的形體表現在週邊設備上。例如早期的 DOS 作業系統大多使用點矩陣字形，現今視窗環境的 Window 作業系統則多使用向量外框字形。

目前繁體字碼最常使用的是 BIG-5 大五碼，簡體字碼最常使用的是 GB2312 碼，針對此兩種字碼簡單介紹如下：

1) BIG-5 碼：係由資策會於 1984 年策劃制定，宗旨原是儘量不使用到控制碼範圍，並能配合國人自制的五大(BIG-5)套裝軟體。由於剛開始是委託民間設計，導致初期的 BIG-5 碼並不能完全使用於五大套裝軟體。雖然如此，市面上絕大多數的套裝軟體都是在 BIG-5 內碼系統發展出來的，因此目前市面上有 2 至 3 個 BIG-5 碼版本，對使用者來說很難明白其中差異，所以在 2003 年由財團法人中文數位化技術推廣基金會接受經濟部標準檢驗局委託，召集國內業者代表、專家和學者，就 BIG-5 編碼字元表原始版本和各主要業界版本予以重整之最新版本[3][4]。BIG-5 碼系統為兩位元組之內碼系統，共可定義 19782 個字碼，其高、低位元組的範圍如(表 2.1)。台灣地區所使用的 BIG-5 碼主要版本如(表 2.2)，其中以微軟的 CP950 較廣為業界使用。

表 2.1 BIG-5 碼系統的內碼分佈

字碼	資料範圍(16進位)	可表達的符號數	可表達的中文字
高位元	81-FE	126	126*157=19782
低位元	40-7E,A1-FE	63+94=157	



表 2.2 台灣地區所使用的 BIG-5 碼主要版本(引自[5])

版本	說明
BIG5-1984	早期由資策會所定義的版本
BIG5-Eten	倚天所使用的版本
CP950	微軟所使用的版本
BIG5-2003	2003年由財團法人中文數位化技術推廣基金會接受經濟部標準檢驗局委託，集合國內業者代表、專家和學者，由原來的BIG-5編碼字元表加上各主要業界版本予以重整得到的最新版本
BIG5-IBM	IBM所使用的版本

2) GB2312 碼：此編碼是漢字字元集和編碼的代號，全稱《信息交換用漢字編碼字元集》，由中華人民共和國國家標準總局發布，1981年5月1日實施，通行於大陸，新加坡等地也使用此編碼[6]。GB 乃是“國標”二字的漢語拼音縮寫。GB2312 收錄簡化漢字及符號、字母、日文假名等共 7445 個圖形字符，其中漢字占 6763 個，字母和符號 682 個。GB2312 將代碼表分為 94 個區，對應第一字節；每個區 94 個位，對應第二字節，兩個字節的值分別為區號值和位號值加 32, 因此也稱為區位碼。01-09 區為符號、數字區，16-87 區為漢字區，10-15 區、88-94 區是有待進一步標準化的空白區(表 2.3)。

表 2.3 GB2312 字符集的區位分佈表(引自[6])

區號	字數	字元類別
01	94	一般符號
02	72	順序號碼
03	94	拉丁字母
04	83	日文假名
05	86	Katakana
06	48	希臘字母
07	66	俄文字母
08	63	漢語拼音符號
09	76	圖形符號
10-15		備用區
16-55	3755	一級漢字，以拼音為序
56-87	3008	二級漢字，以筆劃為序
88-94		備用區

### 2.3 中文字的概念

中文的字數繁多，光是康熙字典就收集了四萬多個字。以繁體中文來說，根據教育部公佈的常用字有 5401 個，次常用字有 7652 個，共計有 13053 個字。在電腦系統內一般以一個位元組來表示一個字元，俗稱 ASCII 碼，但是一個位元組最多只能表達 256 種不同字元，無法滿足所有中文字的顯示，所以中文字必須由兩個 ASCII 碼組成，例如資訊的「資」字，是由 B8 及 1A 兩個碼組成。這種以 2 個位元組來顯示一個中文字的編碼方法，稱之為雙位元組的編碼系統 (DBCS, Double Byte Code System)。這種編碼系統是將原來的 ASCII 碼重新定義，使它能在中文系統下，同時表示 ASCII 碼與中文字，因此我們把這種編碼系統稱為「中文內碼系統」。

由於台灣資訊業界最先並無所謂的「國家標準中文內碼」的制定，所以不同的應用領域各自發展出「BIG-5 碼」、「公會碼」、「IBM 5550 碼」等等不同的中文內碼系統。目前在 PC 上所使用的繁體中文內碼系統，都屬於 BIG-5 碼的編碼系統。目前大陸地區使用的是簡體中文系統，屬於 GB2312 碼的編碼系統。

為了能有足夠位置以容納全世界各種語言的字元和符號，以及為了配合微處理器以 8、16、32 甚或 64 位元為一個運算處理單位的趨勢，ISO 的一些會員國於 1984 年發起制定新的國際字元集編碼標準，把一個字元碼長度規定為 4 個八位元。1988 年初，美國 Xerox 公司的 Joe Becker 提議以新的編碼結構，另外編訂世界性字元編碼標準：將電腦字元集編碼的基本單位由現行的 7 或 8 個位元一舉擴充為 16 個位元，並且充分利用 65,536 個編碼位置以容納全世界各種語言的字元和常用符號。新的字元集編碼標準被命名為“Unicode” [7]。由於新的字元集編碼採取 4 個八位元方式，其可以表示的字數多達

2, 147, 418, 112 個。這套字集將會涵蓋康熙字典全部用字、漢語大字典全部用字及大陸、台灣、韓國、越南、香港所提的國家標準字集，在現今海峽兩岸交流互動日漸頻繁，Unicode 倒不失為兩岸中文碼共通標準的解決方案之一。

#### 2.4 轉詞的概念

由於台灣與中國大陸對於一些物件的描述措詞不盡相同(如表 2.4)，所以除了考量到字碼對字碼轉換之外，還必須對雙方的用詞加以轉換，使轉換後詞句更富有可讀性。在討論轉詞之前，必須要先對一句完整句子做斷詞的動作，以便切割出須要轉詞的名詞，所以首先必須先對斷詞作一番深入的研究[8][9][10]。

表 2.4 繁簡用語差異表

繁體中文名詞	簡體中文名詞
雷射	激光
軟體	軟件
資料庫	數據庫
電腦	計算機
資訊	信息
寬頻	寬帶
程式	程序
計程車	德士

所謂的斷詞即是以一個句子為單位，使用者可以根據自己的需求，而選擇不同的詞典，根據此詞典將句子拆解多個名詞的組合。目前常見的幾種自動斷詞的方法計有法則式、統計式、混合式斷詞方法，在關鍵詞的擷取上也可分為詞庫比對法、文件剖析法、統計分析

法[11]。中文斷詞轉換系統目前以中央研究院所推出的中文斷詞系統[12]較為完整，在其展示系統當中，您可以輸入一篇文章(最簡單的方法是複製一篇新聞)，系統就會做未知詞擷取以及包含未知詞的斷詞標記動作。最後會把結果顯示出來，其結果不但包含了未知詞列表以及包含未知詞的斷詞標記結果，還有程式運作的過程，包含演算法中每一個步驟所相對應的結果。

中文的轉詞如果要做到百分之百的精確，實際上是一個非常複雜的工作。根據轉詞的精確性，大概可以分成下列三個等級：

1) 多對一關係的字根詞語識別轉換：

如表 2.5 所示，其中頭髮的”髮”字及出發的”發”字，千里的”千”字及鞦韆的”韆”字，在簡體的字分別為”发”及”千”，諸如此類的詞皆是屬於此種的轉詞類型。

表 2.5 多對一關係的字根詞語

繁體中文名詞	簡體中文名詞
頭髮	头发
出發	出发
千里	千里
鞦韆	秋千

2) 繁簡詞彙表達方式不同之轉換：

此類的用詞基本上詞與詞間無任何相關連的關係存在，純粹是依照各別使用者的用詞習慣而有所不同，如表 2.6 所示。此類的詞句通常是依照地方的不同而對事物表達的名詞不同，例如台灣系統操作最常用使用的”密碼”詞彙，大陸則習慣用”口令”來表達，諸如此類的狀況不勝枚舉，皆屬於表達方式的轉換。

表 2.6 繁簡詞彙表達方式不同的詞語

繁體中文名詞	簡體中文名詞
資訊	信息
密碼	口令
印表機	打印機
軟體	軟件

3) 根據上下文對詞彙翻譯：

有一些詞需要根據上下文才能夠準確地決定如何翻譯的，比如在中國大陸的電腦用語習慣中，「程序」通常是表示台灣電腦用語的「程式」，但是也可以表示一般對事情的處理「程序」。在繁體使用者的語言習慣中，「程式」與「程序」這兩個名詞是代表不一樣的事情，所以要轉換簡體「程序」的名詞到繁體來，還必須要根據上下文的意思來進行斷句或分析，進而選出較適合的轉換名詞來。因此，此種轉換類型是屬於繁簡轉詞中最難做的，而且消耗的系統資源也最大。

## 2.5 中文轉詞發展概況

為了排除兩岸之間的文件傳遞及訊息交換的文字隔閡，許多繁簡轉換的軟體也開始蓬勃發展。台灣目前較常用的轉換工具約可分成兩類，分為使用者端的文件轉換、頁面轉換，以及主機端的線上轉換。使用者端的軟體供應者以「兩岸文件通」[13]較常見，除了提供不同的作業平台及不同的開發語言有不同的解決方案外，更進一步整合微軟的產品，針對 Office 家族、IE、SQL Server 等等軟體提供轉換服務；另外也可從網站下載南極星[14]免費提供的轉換軟體，直接轉換電腦畫面文字。主機端繁簡互換的軟體，則是網業家代理的「信使網路繁

簡通」[15]，該軟體是直接安裝在網路公司內部主機，進行網頁資訊即時轉換、資料庫搜尋結果轉換，甚至連結出去的網頁，也可一併完成繁簡體互換。現今網際網路上也有許多網站提供線上繁簡轉換服務，滿足一般使用者在瀏覽繁簡網站資訊時，不受文字侷限而更能暢行無阻，其中比較常見的有「簡繁轉換網」[16]、以及蕃薯藤提供的網頁服務「簡體網頁轉成繁體網頁」[17]。

除了中央研究院在中文詞庫上有多項研究及開發成果外，業界也有一些較具代表性繁簡轉換的軟體。針對這些軟體在轉詞發展支援的程度，我們統一整理資料如下：

#### 1) 信使網路繁簡通

系統內提供一個轉換詞庫，可以支援繁簡轉換一對多的映射，習慣用語的轉換及用戶自定義詞典的功能。尤其是詞庫新增功能可以讓用戶自己發明的詞或者是沒有收錄進去的詞收錄到轉換詞庫中。運用此功能，該產品所建立的詞庫將更趨於完善、更有彈性，尤其在個人化用詞方面的超強功能。

#### 2) 華文館(兩岸文件通)

華文館可以當做瀏覽器(IE)的外掛程式，會在瀏覽器上多加三組按鈕：「簡轉繁」、「繁轉簡」、「我要轉詞」，其中我要轉詞按鈕即可轉換字詞。另外他也提供詞彙管理之功能，讓用戶可以新增、編輯、載入常用的詞彙轉換表與詞彙轉換項目組、並鎖住專有名詞，使得用戶端瀏覽的內容會依詞彙管理之設定而立即作口語化的詞彙轉換。該系統最大特色是可以鎖住專有名詞的轉譯，這項功能對於比較專業的系統(比如：ERP系統)，系統內眾多專有名詞的翻譯可說助益非淺。

### 3) 南極星

南極星(NJStar Chinese Word Processor)南極星中文文書處理系統是一個能夠在所有語言文字版本的微軟作業系統下獨立執行的軟體。其內建有 20 多種常用中文輸入法、多種南極星中文全真字體及便利的中英文雙向詞典和漢字資訊庫，並具有繁簡自動多內碼識別/轉換功能。南極星並沒有提供華文館的三大核心技術：「轉文字碼、轉詞口語化、並鎖住專有名詞」，所以在轉詞的表現上較其他產品為弱。

### 4) Word2000

Word2000 是微軟 OFFICE 家族內的一套文書編輯軟體，內建繁簡字體的轉換，目前除了提供字對字一對一的轉換，也提供簡單的轉詞功能，但是使用者無法自行定義字庫。

針對這四大產品在轉詞方面可以支援的變化性上，我們整理出來的比較表如表 2.7。表中所提中文圖的置換係指該軟體有提供圖形內中文字詞的換圖方案，另外動態轉詞指的是該軟體支援一些動態性的內容詞的轉換，比如 Web 上的聊天系統就是最好的範例。

表 2.7 繁簡轉換軟體對於轉詞的支援狀況

繁簡轉換軟體	字對字的 轉換	中文詞的 轉換	中文圖的 替換	動態轉詞
信使網路繁簡通	有	有	有	有
華文館(兩岸文件通)	有	有	無	有
南極星	有	有	無	有
Word2000	有	無	無	無

### 第三章系統開發與實例說明

本研究之實作對象乃採用企業內使用最普及的 ERP 系統，實際樣本則採用由鼎新電腦開發的 ERP 企業資源規劃系統[18]。鼎新電腦 ERP 系統使用的客戶群橫跨台灣與大陸，因此可以算是一個非常合適的繁簡轉換實作案例。本研究中利用實際在兩岸運行的 Workflow ERP 系統，來說明該系統如何透過繁簡字詞轉換的技術，達到兩岸使用者跨越用字與用詞的障礙來共用一套 ERP 系統。

#### 3.1 ERP 系統的架構

鼎新電腦 ERP 企業資源規劃系統係利用 Borland Delphi 的開發工具結合微軟的 SQL Server 資料庫，以 Windows 為作業系統平台所開發出來的商業套裝軟體。其中為了讓前端使用者能跨越整個網際網路操作該系統，特別運用了 Delphi MIDAS 技術，發展出適合窄頻使用者的 3-Tier 架構(圖 3.1)[19]。圖中 MIDAS 係 Borland 應用於 3-Tier 的技術，BDE 則是應用程式聯結資料庫的引擎。

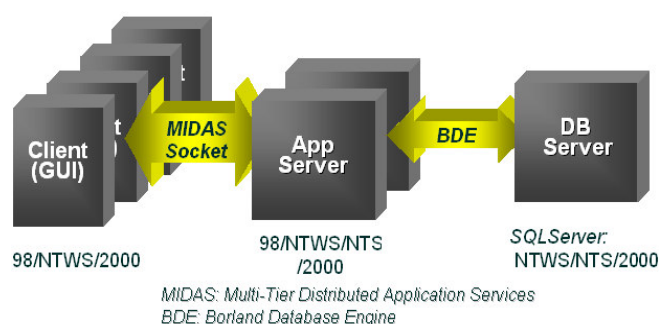


圖 3.1 Workflow ERP 系統 3-Tier 架構

一套完整的 ERP 系統涵蓋『配銷管理』、『存貨管理』、『製造管理』、『財務管理』、『人力資源管理』、『品質管理』及『決策支援』等範疇。如果依照不同的使用需求來看，其作業型態又分為建檔類、報表類、憑證類及批次執行類。前端使用者依不同的職能別使



用不同的作業。各類型作業所負責的功能說明如下：

- 1) 建檔類：負責 ERP 資料的新增、修改、刪除與查詢，如「客戶資料建立作業」。
- 2) 報表類：負責把資料庫的資料以特定的格式顯示出來，用於一些數據統計或資料分析，如「客戶銷售統計表」。
- 3) 憑證類：主要在於產生交易的憑單，並在其上加上流程簽核的控管，完整留存交易的紀錄，如「客戶訂單憑證」。
- 4) 批次執行類：負責整批大量資料的新增、修改或刪除，如「傳票整批過帳作業」。

### 3.2 繁簡轉換的對象

依據上述 ERP 架構及作業類型說明，我們可以把需要做繁簡轉換的對象鎖定在 3-Tier 的 Client(GUI)端及使用者會接觸到的作業類型上。至於需做繁簡轉換的控管部分，我們可分為下面五個方面來做分析：

- 1) 資料庫：為了達到資料庫資料編碼的一致性，在系統上線初期就必須先定義資料庫資料的中文編碼原則，即存進資料庫的中文需統一規定為 BIG-5 碼或 GB2312 碼。
- 2) 操作介面：使用者操作的使用介面，如建檔畫面上的中文字、報表的選項條件中文字等等(圖 3.2)。
- 3) 資料內容：包含使用者所存入資料庫的資料及儲存在一般文字檔的資料，甚至是報表經過運算後所產生出來的中文結果，皆屬於資料內容的一部份(圖 3.3)。
- 4) 圖形檔案：系統中所用到的一些圖形檔(如公司的 LOGO)，操作流程關係圖。因圖形上的中文字屬於圖形，無法用繁簡轉換技術進行轉換，所以此部份只能製作 2 個不同圖檔內容來解

決，即一個是繁體圖檔，一個是簡體圖檔（圖 3.4）。

5) 訊息說明：系統操作過程中，會應用到的一些提示或警告訊息，如『單別不存在』或『該品號庫量不足』（圖 3.5）。

項目	值
客戶代號	1102
客戶簡稱	和大家俱行
客戶全名	和大家俱行
負責人	陳大和
連絡人	陳純祥
TEL_NO (一)	02-22324567
TEL_NO (二)	02-22345673
FAX NO	
E-Mail	
統一編號	04016004
資本額	250 萬
年營業額	1000 萬
員工人數	0 人
總店號	
總公司請款	<input type="checkbox"/>
發票號碼依總公司控管	<input type="checkbox"/>
分店數	0
交易幣別	NT\$
部門別	3500 營業國內部
業務人員	0558 朱南安
收款業務員	0558 朱南安
開業日期	<input type="checkbox"/>
歇業日期	<input type="checkbox"/>

圖 3.2a ERP 操作介面的轉換(繁)

項目	值
客戶代號	1102
客戶簡稱	和大家俱行
客戶全名	和大家俱行
負責人	陳大和
連絡人	陳純祥
TEL_NO (一)	02-22324567
TEL_NO (二)	02-22345673
FAX NO	
E-Mail	
統一編號	04016004
資本額	250 萬
年營業額	1000 萬
員工人數	0 人
總店號	
總公司請款	<input type="checkbox"/>
發票號碼依總公司控管	<input type="checkbox"/>
分店數	0
交易幣別	NT\$
部門別	3500 營業國內部
業務人員	0558 朱南安
收款業務員	0558 朱南安
開業日期	<input type="checkbox"/>
歇業日期	<input type="checkbox"/>

圖 3.2b ERP 操作介面的轉換(簡)

佇列工作控制台[系統管理者][測試用公司] - [閱覽報表(品號品名對照表-工作代號:20060117000001-200...]

佇列工作管理員(Y) 視窗(Z)

品號品名對照表

製表日期: 95/01/17 第1頁

類別一	品號 條碼	品名 規格	建立日期 修改日期	單位 小單位	主供應商 品號屬性	庫存管理 保稅品	主要庫別	批 補
電子零件	A	A	89/06/03 89/09/11	PCS		Y N		不 依
	L11002	電阻	88/12/19	PCS	自製件 佳佳工業	N Y	1A	不 依
	L21001	50歐阻 機構	89/09/11 88/12/19	PCS	採購件	N Y	成品倉 1A	不 依
電阻	L11001	IC	89/09/11	PCS	採購件	N	成品倉	不 依
			88/12/19		佳佳工業	Y	1A	不 依
包裝材料	35001001	普通泡綿 500座	88/05/24	PCS	總合泡綿	Y	1C	不 依
			89/09/11		採購件	N	原料倉	不 依
大型辦公椅	111001	辦公椅SL 500	88/05/24	PCS	自製件	N	1A	依 需
			89/09/11					
	111002	辦公椅BL 500KL	88/06/24	SET	自製件	Y	1A	不 依
			89/09/11					
	91001	辦公椅 600	88/06/24	SET	自製件	N	1A	依 需
89/09/11								
一般休閒椅	121001	休閒椅 ML	88/06/04	SET	佳佳工業	Y	1A	依 需
			89/09/11		採購件	N	成品倉	不 依
			88/07/03	張		Y	1A	不 依
			89/09/11		自製件	N	成品倉	依

座標(英寸)(3.77, 13.62) (X: 362, Y: 1308)

圖 3.3a 報表資料內容的轉換(繁)

佇列工作控制台[系統管理者][測試用公司] - [閱覽報表(品號品名對照表-工作代號:20060117000001-200...]

佇列工作管理員(Y) 視窗(Z)

品號品名對照表

製表日期: 95/01/17 第1頁

類別一	品號 條碼	品名 規格	建立日期 修改日期	單位 小單位	主供應商 品號屬性	庫存管理 保稅品	主要庫別	批 補
電子零件	A	A	89/06/03 89/09/11	PCS		Y N		不 依
	L11002	電阻	88/12/19	PCS	自製件 佳佳工業	N Y	1A	不 依
	L21001	50歐阻 機構	89/09/11 88/12/19	PCS	採購件	N Y	成品倉 1A	不 依
電阻	L11001	IC	89/09/11	PCS	採購件	N	成品倉	不 依
			88/12/19		佳佳工業	Y	1A	不 依
包裝材料	35001001	普通泡綿 500座	88/05/24	PCS	總合泡綿	Y	1C	不 依
			89/09/11		採購件	N	原料倉	不 依
大型辦公椅	111001	辦公椅SL 500	88/05/24	PCS	自製件	N	1A	依 需
			89/09/11					
	111002	辦公椅BL 500KL	88/06/24	SET	自製件	Y	1A	不 依
			89/09/11					
	91001	辦公椅 600	88/06/24	SET	自製件	N	1A	依 需
89/09/11								
一般休閒椅	121001	休閒椅 ML	88/06/04	SET	佳佳工業	Y	1A	依 需
			89/09/11		採購件	N	成品倉	不 依
			88/07/03	張		Y	1A	不 依
			89/09/11		自製件	N	成品倉	依

座標(英寸)(7.33, 4.66) (X: 704, Y: 447)

圖 3.3b 報表資料內容的轉換(簡)

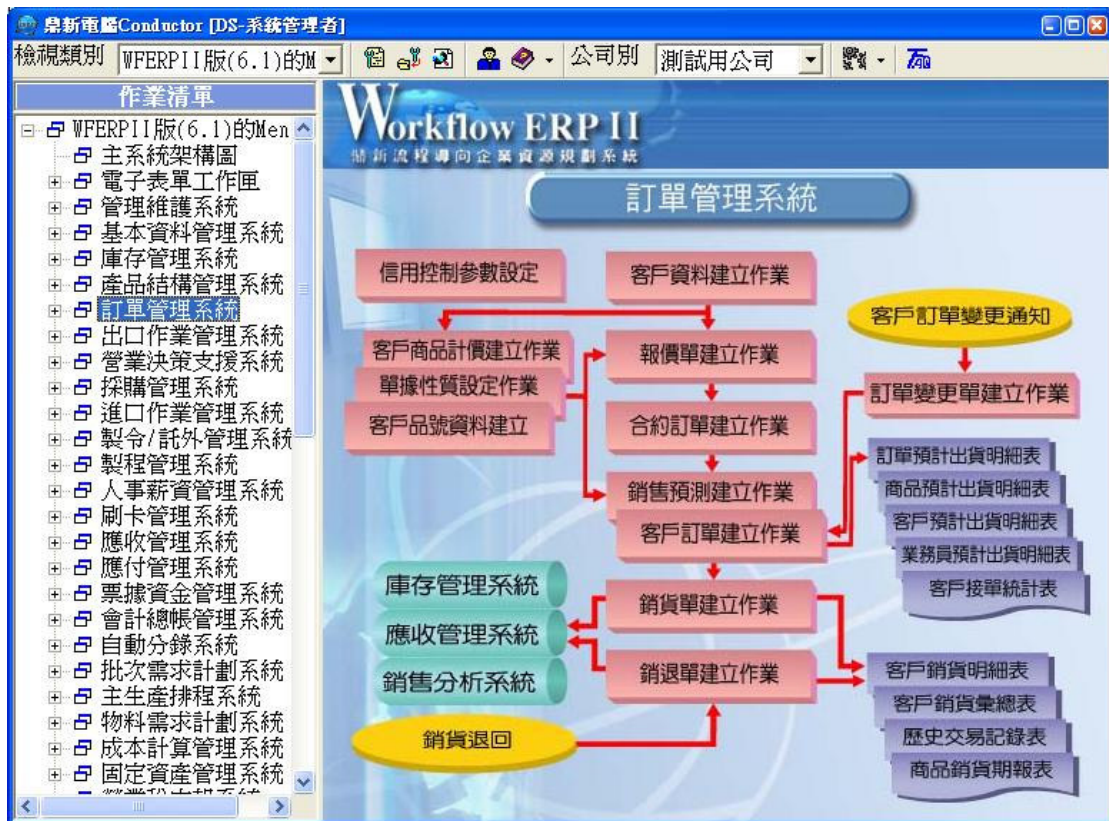


圖 3.4a 圖檔的轉換(繁)

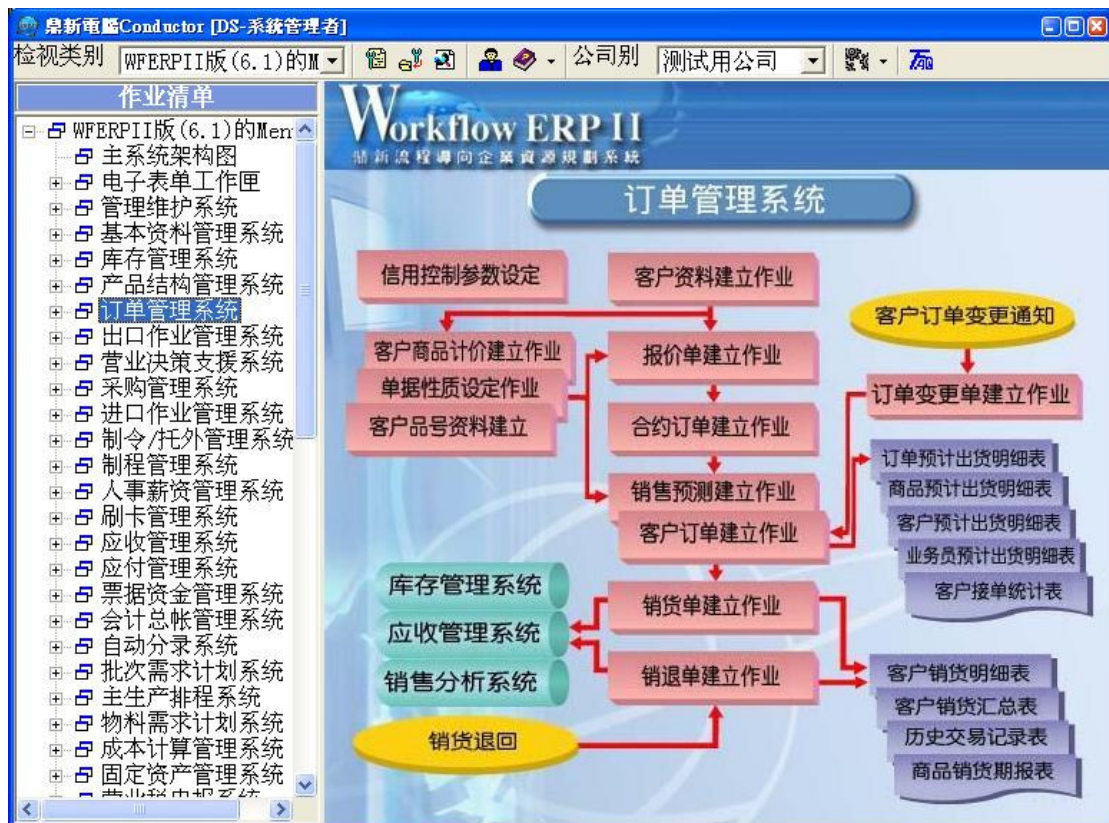


圖 3.4b 圖檔的轉換(簡)

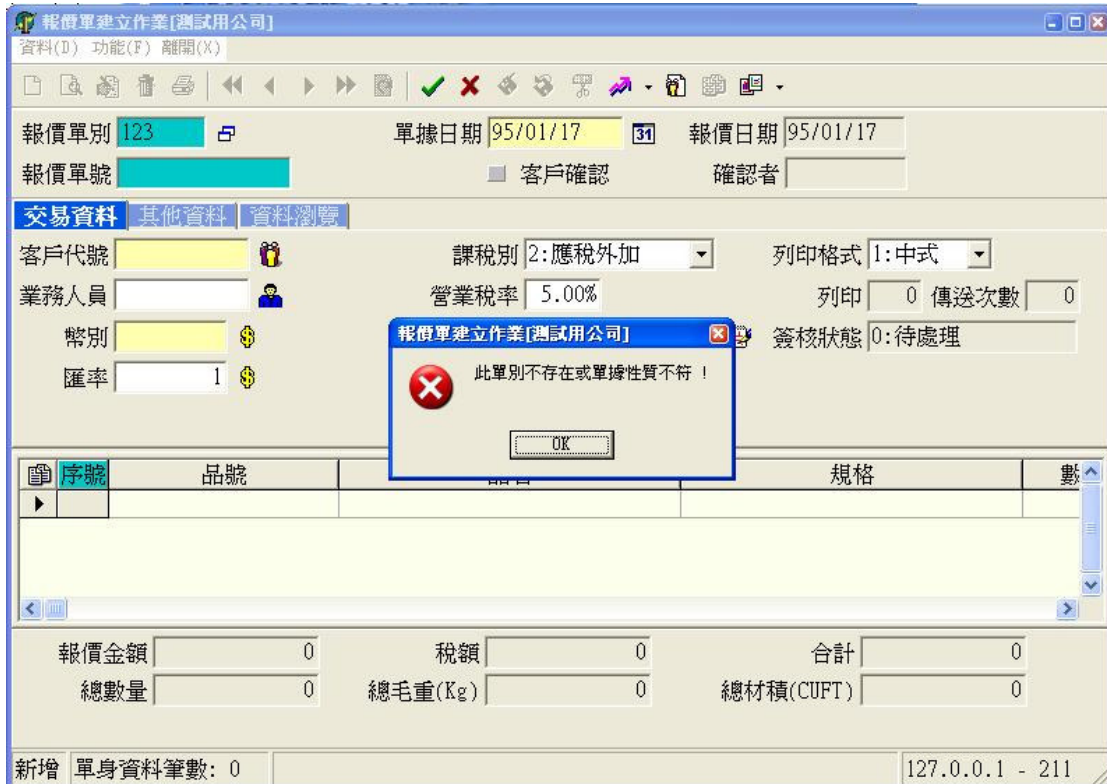


圖 3.5a 提示訊息的轉換(繁)

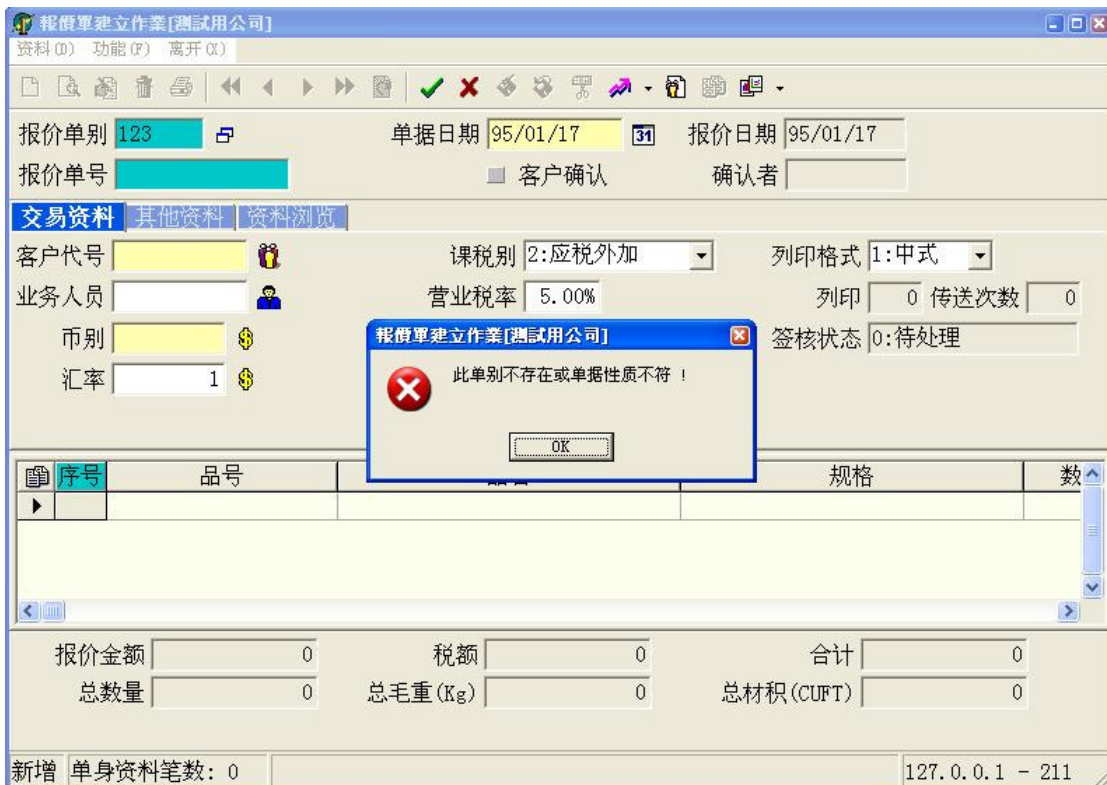


圖 3.5b 提示訊息的轉換(簡)

### 3.3 繁簡轉換在兩岸三地 ERP 系統的運用

目前台灣企業西進風潮日熾，企業導入 e 化多半就近選擇當地系統服務業者，然大陸業者提供之服務無法與台灣母公司系統整合，很難針對台商個別需求規劃，常遭遇到兩岸資訊建設上的不平衡與跨國技術支援的不足，導致 e 化建構成本增加與資料傳輸的遲緩等困擾。

以 WorkFlow ERP 產品開發所架構的兩岸三地為範例，其系統整合架構共可分為二種：

- 1) 第一種方式：為採取資料庫分散架構，如圖 3.6 所示，其中三地預設為台灣、香港及大陸，資料的流通藉由雙邊的 AP Server 透過網際網路互相同步。因三地皆有自己的資料庫，所以在資料同步之前，上游公司會先判斷下游公司資料庫語言別，如異於自己資料庫，則資料在轉出前會先經過轉詞轉字的處理，然後再組成可執行的 SQL 指令傳送至下游資料庫執行。
- 2) 第二種方式為採取資料庫集中架構，如圖 3.7 所示，圖中的 DB 及 AP 主機可選擇性的放在台灣或香港或大陸，唯其存入資料庫的語言必須統一為繁體或是簡體。至於各用戶端的電腦可以散佈於台灣、香港或大陸，透過網際網路連線至 AP 主機，藉由 AP 主機再把資料存放至 DB 主機。同樣的，用戶端在儲存資料之前，必須先判斷資料庫的語言別，如有異於目前用戶端使用的語言別則必須先行經過轉詞轉字的動作，進而做資料送出存檔的動作。因資料庫採取集中方式處理，所以在資料庫主機內就會存在有台灣公司的資料、香港公司的資料及大陸公司的資料，資料的拋轉完全仰賴 AP 主機的運作，但因用戶端在資料送出前已做好資料繁簡體的轉換，所以此階段 AP 主機就不再插手資料繁簡體的處理動作。

不論是資料庫分散或是資料庫集中，只要存在有 ERP 系統使用者橫跨兩岸的情況，勢必皆要處理資料的繁簡轉換。轉換的過程除了要求轉換結果要正確之外，轉換的速度也是重點之一。一般 ERP 開發廠商如果是搭配其他協力廠商繁簡轉換的軟體（如兩岸文件通），除了必須架設一台繁簡轉換的主機之外，更要負擔轉換資料要往返於網路的時間成本。本研究所提出的轉換元件（檔案大小：376 KB）或是轉字對照字庫（檔案大小：501 KB）皆是安裝在使用者端，所以除了節省網路傳輸資料的時間外，同時利用記憶體預先載入字庫的方式減少硬體 I/O 所需要的時間，所以轉換的速度表現足以應付兩岸三地大量資料的同步。

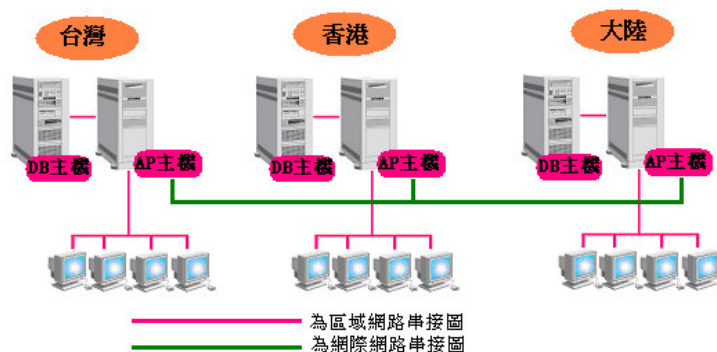


圖 3.6 兩岸三地系統整合架構-資料庫分散

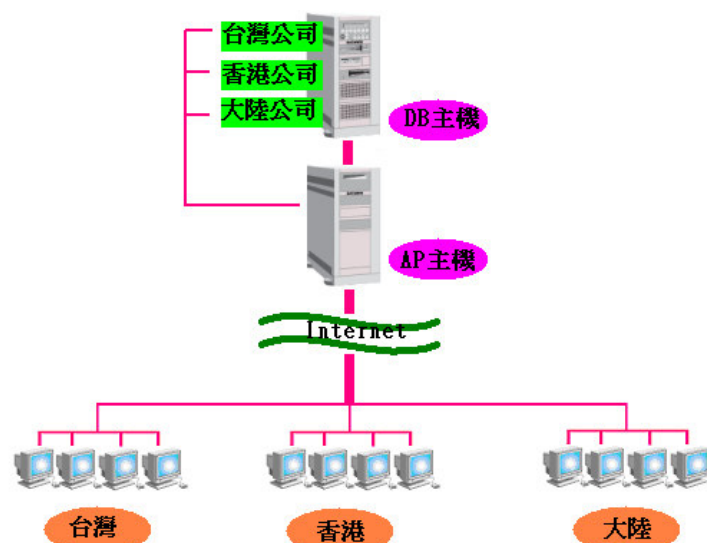


圖 3.7 兩岸三地系統整合架構-資料庫集中

從兩岸三地 ERP 資料流通的方式來探討，可以整理出營運模式如圖 3.8。圖中以台灣為接受客戶訂單的窗口，香港為轉運口，而生產產品的工廠則設在大陸。整個營運模式為由台灣接受客戶的訂單，自動產生一筆對香港的採購單，香港公司也產生相對應的訂貨單及採購單，最後只於大陸公司的訂單。同理，大陸公司接到訂單以後備完貨就會產生有銷貨動作的銷貨單，隨即香港及台灣也會產生相對應的進貨單及銷貨單。最後再以產生結帳單及應付憑單來作為整個流程的結尾。整個採購及銷貨流程中，有用紅點標示的地方表示資料在同步出去前，必須先識別是否下游的資料庫語言別有異於自己要轉出的資料，如果相異的話就必須先經過轉字的處理後，再行同步轉至下游資料庫。

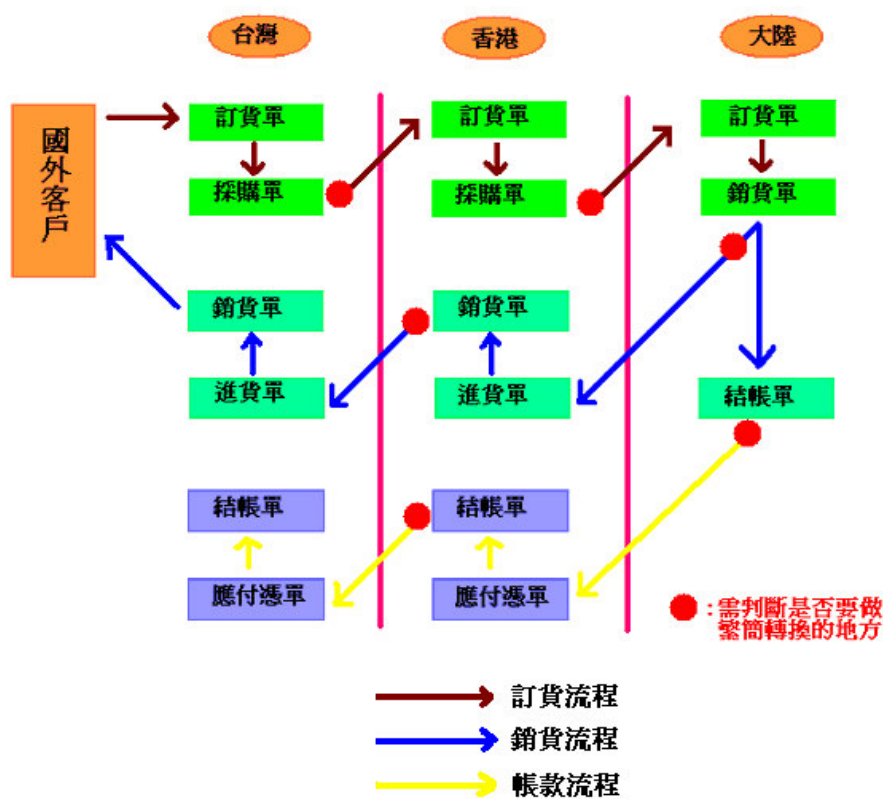


圖 3.8 兩岸三地系統營運模式



### 3.4 兩岸 ERP 習慣用詞的比較

目前兩岸面臨溝通上的一大問題就是「用詞不同」。幾十年來兩岸三地長久分離造成使用詞句的不同，台灣的繁體字與大陸的簡體字雖然都是中文字，中間不僅有字與詞極大的差異，更有許多不同的專有名詞，造成兩岸文字與思想上的溝通障礙，此問題實為兩岸間急需解決的重要問題[20]。在 ERP 系統使用上，也同樣面臨到兩岸習慣用詞的問題(表 3.1)及使用者操作畫面習慣用語(表 3.2)。由表 3.1 及表 3.2 內容來看，我們可以得知用詞的不同造成前端操作使用者困擾的程度，為了避免溝通上的障礙，使用者必須試著再學習兩岸用詞的差異處，習慣單一事物有著兩樣說法的狀態。針對這些問題，本研究也蒐集了兩岸在 ERP 系統上一些常用的習慣用詞，再針對這些對應關係建立一個完整的詞庫，所以研究中除了提出轉字的方式之外，另外也針對轉詞的部份，藉由建立詞庫的方式再搭配詞彙搜尋的公式，提出一種快而有效率的轉詞方法，其主要目的在於期使兩岸的資料溝通上能更具親合力。

表 3.1 簡單列舉出使用者在使用 ERP 系統時產業用語的相異處，這是系統在設計的時候必須特別留意的部份，其中如台灣地區報稅常用的統一編號，大陸則稱為稅號；買賣業常用的『通路』一詞，大陸則習慣稱為『渠道』等等。表 3.2 則列舉出使用者在操作 ERP 畫面時一些操作用詞的相異處，如台灣在使用 ERP 系統時會做『登入』與『登出』的動作，大陸的使用者習慣稱為『登錄』及『註銷』。使用者在登入系統時常常需要輸入『密碼』，大陸用詞則習慣稱做『口令』等等。

表 3.1 兩岸 ERP 習慣用詞比較表

台灣 ERP 習慣用詞	大陸 ERP 習慣用詞
生產線別	工作中心
本國幣別	記賬本位幣
課稅別	稅種
通路	渠道
加工廠商	委外廠商
片語	常用語
出售利得	銷售收益
立沖帳目	核算項目
單品結帳	按項結算
單據憑證	憑證
傳票	核算項目
郵遞區號	郵編
業主權益	所有者權益
投資抵減	投資抵扣
統一編號	稅號
登錄者	用戶
營業稅	增值稅
西元年	公曆
行事曆	假日表
資本額	註冊資金
母工單	源工單
底薪	基本工資
註記	頁腳
申報公司	納稅公司

表 3.2 兩岸 ERP 使用者操作畫面習慣用語

台灣 ERP 使用者操作用語	大陸 ERP 使用者操作用語
區域印表機	本地打印機
數據機	調制解調器
光碟機	光盤驅動器
執行	運行
密碼	口令
登入	登錄
登出	註銷
視窗	窗口
進階	高級
傳統	常規
滑鼠	鼠標
資訊	信息
游標	光標
首筆	第一個
尾筆	最後一個
預設	缺省

### 3.5 BIG-5 碼與 GB2312 碼轉換原理

不論是繁體中文字或簡體中文字，每個字皆有一個相同的特性，即每個字皆是由 2 個位元組成，這兩個位元分別稱為高位元及低位元 [21](如圖 3.9 所示)。

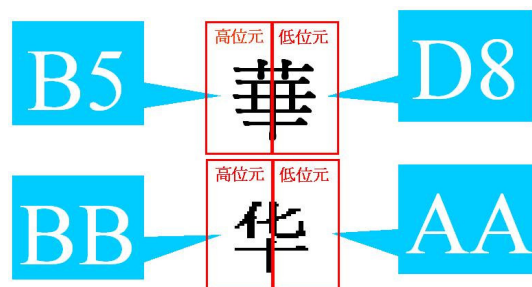


圖 3.9 單一中文字組成特性

依照繁簡中文組成份子的架構，我們可以發現 2 個特性：(1)不論是大 5 碼或是 GB2312 碼，每個字碼皆由 2 個位元組所構成，(2)不同的字其字碼具有唯一性。利用此兩特性我們可以把這兩種字碼的關係做成一字碼對照表(表 3.3)。另外利用兩字型標準及常用字碼分佈的範圍(表 3.4) [22][23]，我們可以把這 2 種字碼每個字依小至大給予一個唯一編號(表 3.5)，最後運用此唯一編號製作出 BIG-5 與 GB2312 字碼對照的矩陣關係表(表 3.6)，以方便由應用程式直接來讀取。

表 3.3 BIG-5 與 GB2312 字碼對照表

BIG5 字型(字碼)	GB2312 字型(字碼)
風(ADB7)	风(B7E7)
飛(ADB8)	飞(B7C9)
閩(ADAC)	闽(E3C5)
頁(ADB6)	页(D2B3)
韋(ADB3)	韦(CEA4)
軍(AD78)	军(BEFC)
軌(AD79)	轨(B9EC)
計(AD70)	计(BCC6)
訂(AD71)	订(B6A9)
貞(AD73)	贞(D5EA)
負(AD74)	负(B8BA)
倆(ADC7)	俩(C1A9)
限(ADAD)	限(CFDE)

表 3.4 BIG-5 及 GB2312 標準及常用字碼分佈的範圍

		高位元組	低位元組
BIG5 字碼	16 進位	A1~FE	40~7E, A1~FE
	10 進位	161~254	64~126, 161~254
GB2312 字碼	16 進位	A1~FE	A1~FE
	10 進位	161~254	161~254

表 3.5 BIG-5 及 GB2312 字碼唯一編號運算式

	轉換運算式	範例
BIG5 碼	低位元在 40~7E: (高位元-161)*157+(低位元-64) 低位元在 A1~FE: (高位元-161)*157+(低位元-161)+63	風(ADB7)→1969
GB2312 碼	(高位元-161)*94+(低位元-161)	风(B7E7)→2138

表 3.6 BIG-5 與 GB2312 字碼相互對應之陣列

BIG5[1967]=D2F4	GB[2136]=AE70
BIG5[1968]=D2B3	GB[2137]=BE57
<b>BIG5[1969]=B7E7</b>	<b>GB[2138]=ADB7</b>
BIG5[1970]=B7C9	GB[2139]=BAC6
BIG5[1971]=CAB3	GB[2140]=B26C
BIG5[1972]=CAD7	GB[2141]=B37B
BIG5[1973]=CFE3	GB[2142]=B6BE
BIG5[1974]=B3CB	GB[2143]=C15F
BIG5[1975]=D9F1	GB[2144]=BFD8
BIG5[1976]=D9C4	GB[2145]=A95E
BIG5[1977]=B1B6	GB[2146]=BBF1
BIG5[1978]=B7C2	GB[2147]=A6F2
BIG5[1979]=B8A9	GB[2148]=A75F
BIG5[1980]=BEEB	GB[2149]=A4D2

表 3.6 中，BIG-5 碼與 GB2312 碼對應陣列建立的流程圖如圖 3.10 所示，整個流程可分為以下四個步驟：

- 1) 取得繁體字元的字碼，如『華』字的字碼為 B5D8。
- 2) 取得簡體字元的字碼，如『华』字的字碼為 BBAA。
- 3) 建立矩陣關係式。
- 4) 運用表 3.5 運算式轉換矩陣內容為矩陣位置，如 B5D8→3250。



圖 3.10 BIG-5 碼與 GB2312 碼對應陣列建立的流程圖

### 3.6 BIG-5 碼與 GB2312 碼轉換元件

BIG-5 與 GB2312 碼對應陣列表建立好以後，為了讓一般應用程式能快速使用此轉換功能，我們在對應陣列表、轉詞詞庫及應用程式之間建立了一個轉換元件(圖 3.11)，提供轉換的介面以及負責整個轉換過程流程串接的動作，其功能包含有：

- 1) 載入對應陣列表、轉詞詞庫至記憶體，以增快資料的搜尋時間。
- 2) 接收應用程式需要轉換的文字並過濾屬於非中文字碼。
- 3) 對於需轉詞的詞句先做相對應詞句轉換。
- 4) 藉由搜尋陣列表內的資料，找出對應的 BIG-5 碼或 GB2312 碼。
- 5) 傳回轉換過的中文字給應用程式。
- 6) 提供對應陣列表的維護功能，包含新增新字對應內容及修改原對應表資料內容。
- 7) 提供對應陣列表的維護功能，包含新增新詞對應及修改原對應表資料內容。由於轉換元件不具有操作介面，所以我們另外撰寫一支使用者端的維護工具(圖 3.12)來呼叫該元件進行整個陣列或詞庫的維護動作。

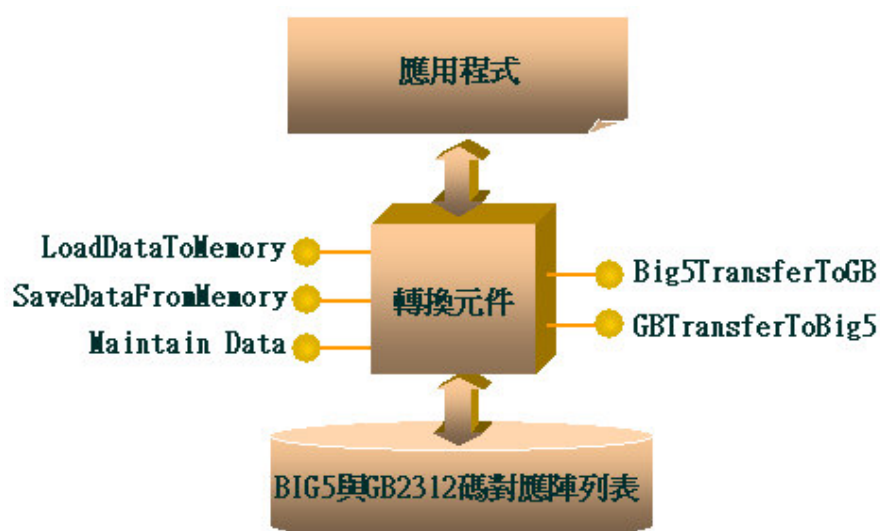


圖 3.11 轉換元件



圖 3.12 對應字碼維護工具

### 3.7 中文轉詞架構

本研究在轉詞方法係參照 Xiaohong Huang 等人所提出的斷詞方法[24]，其研究中提出一個有規則排列的斷詞詞庫，並對此詞庫建立一個索引來加快搜尋資料的時間，所以該研究所提出的方法具有詞庫空間小、斷詞速度快等特性。本研究延伸其所提出的詞庫架構，配合中文字內碼的特性，改良用陣列來儲存詞庫，並以內碼來取代索引的方式，加快詞庫的搜尋。

整個轉詞的核心在於詞庫的建立，詞庫建立方式如下：

- 1) 取每個轉詞詞句的第一個字元經過公式運算(如表 3.5)得出一維的矩陣位置值。
- 2) 相同的一維矩陣位置下建置二維矩陣位置以放置第一個字元相同的詞句。
- 3) 二維矩陣內詞句依據詞句長短由大排到小。

以『國家資訊管理中心』詞句為例(如圖 3.13)，轉詞步驟如下：

- 1) 載入詞庫的詞句到記憶體的二維陣列，如圖 3.13 所示，『資』字

開頭的詞句有 4 組，所以二維陣列的資料分布分別為：

BIG5[3747][0]：資訊網=万维网

BIG5[3747][1]：資訊額=注册资金

BIG5[3747][2]：資料庫=数据库

BIG5[3747][3]：資訊=信息

- 2) 對於要轉換的詞句由左至右一次取出一個字元，經過公式運算(表 3.5)得出矩陣位置值 X(比如：資->3747)。
- 3) 如果 BIG5[X]內有二維陣列值，表示此字元可能有相對應的字詞要轉換，再由 BIG5[3747][0]~BIG5[3747][4] 逐一取出內存字詞與『國家資訊管理中心』做比對，如果內存的字詞包含於『國家資訊管理中心』內，則終止比對動作且進行換詞動作(資訊→信息)。
- 4) 重複步驟 2~3，一直到比對『國家資訊管理中心』字詞至最後一個字元『心』，然後完成字詞轉換動作。

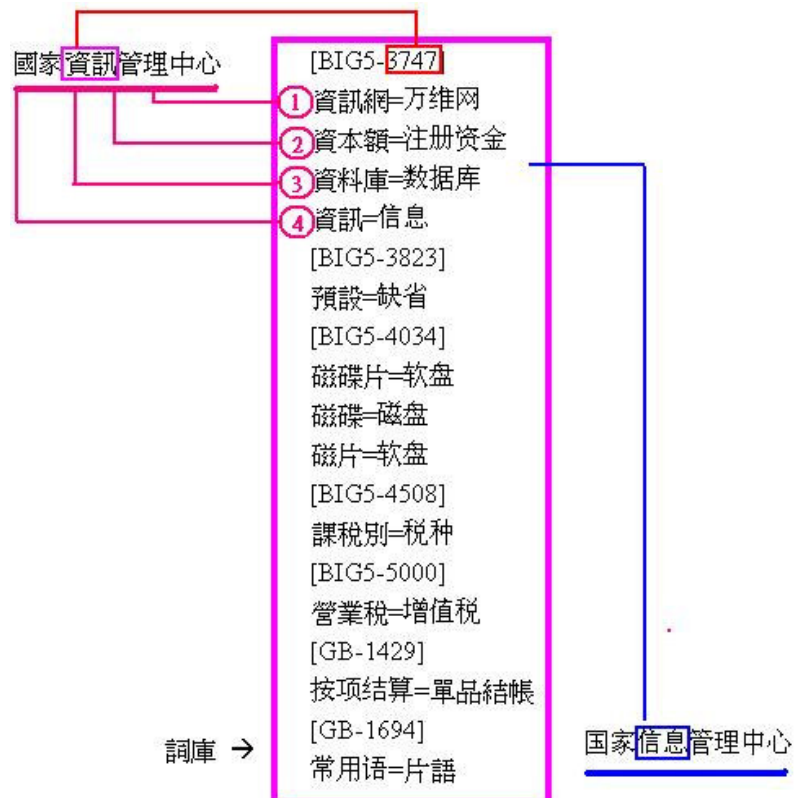


圖 3.13 轉詞架構圖



依據前述 ERP 架構及作業類型說明，我們可以把需要做字詞轉換的對象鎖定在 3-Tier 的用戶端或是使用者會接觸到的作業類型上。至於需做字詞轉換的控管部分，我們可分為下面五個方面來做進行：

- 1) 資料庫：與轉字同樣的考量，我們為了達到資料庫資料編碼的一致性，在系統上線初期就必須先定義資料庫資料的中文編碼原則，即存進資料庫的中文需統一規定為 BIG-5 碼或 GB2312 碼。
- 2) 操作介面：使用者操作的使用介面，如建檔畫面上的詞彙、報表的選項條件詞句等等(圖 3.14)。
- 3) 資料內容：包含使用者所存入資料庫的資料及儲存在一般文字檔的資料，甚至是報表經過運算後所產生出來的中文詞彙，皆屬於資料內容的一部份(圖 3.15)。
- 4) 訊息說明：系統操作過程中，會應用到的一些提示或警告訊息，如『使用者密碼錯誤』或『統一編號輸入錯誤』(圖 3.16)。
- 5) 使用者自行定義轉換字詞：詞庫內容開放使用者可依據個別需求自行新增、修改、刪除(圖 3.17)。

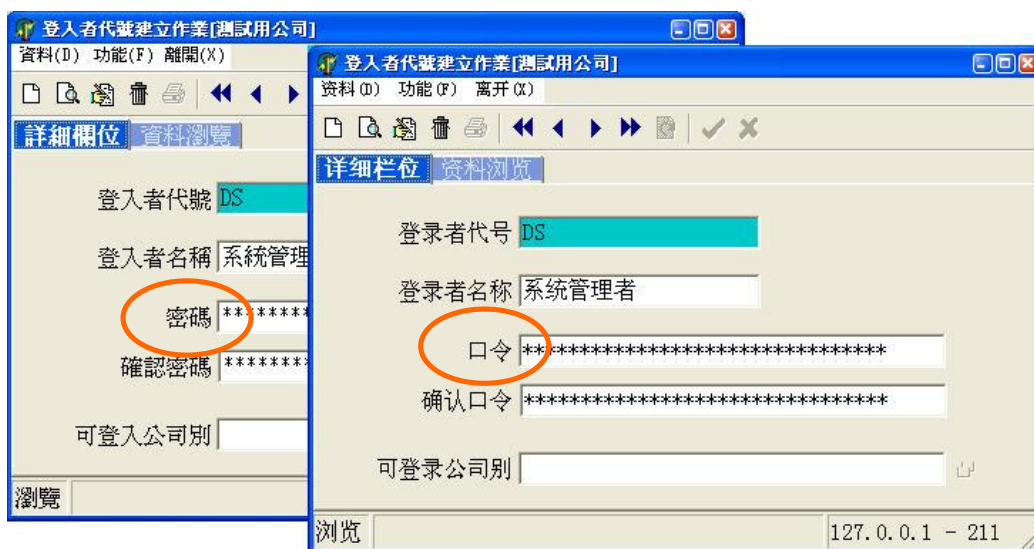


圖 3.14 建檔畫面詞彙轉換

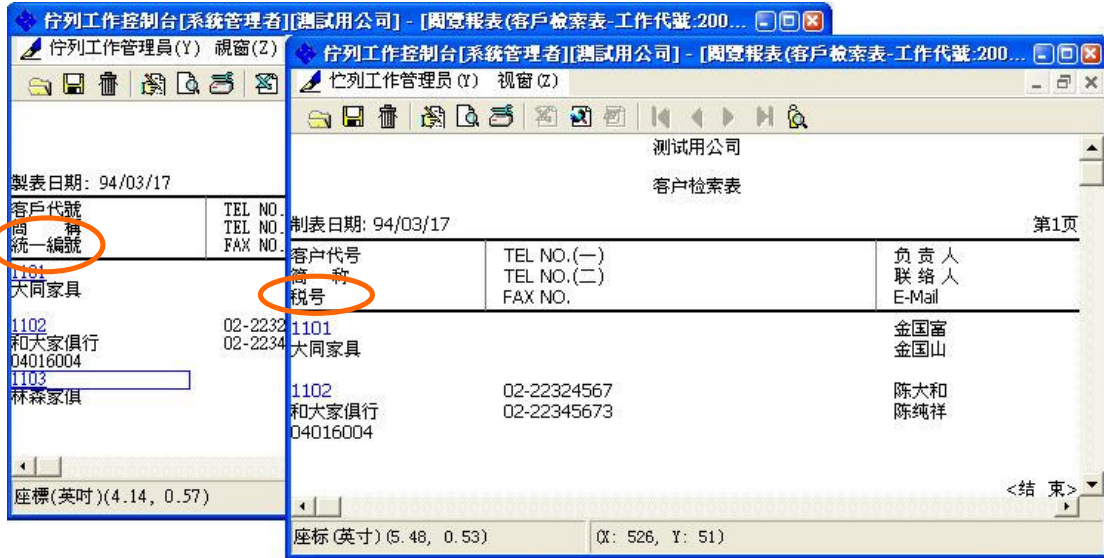


圖 3.15 報表資料詞彙轉換



圖 3.16 訊息說明詞彙轉換

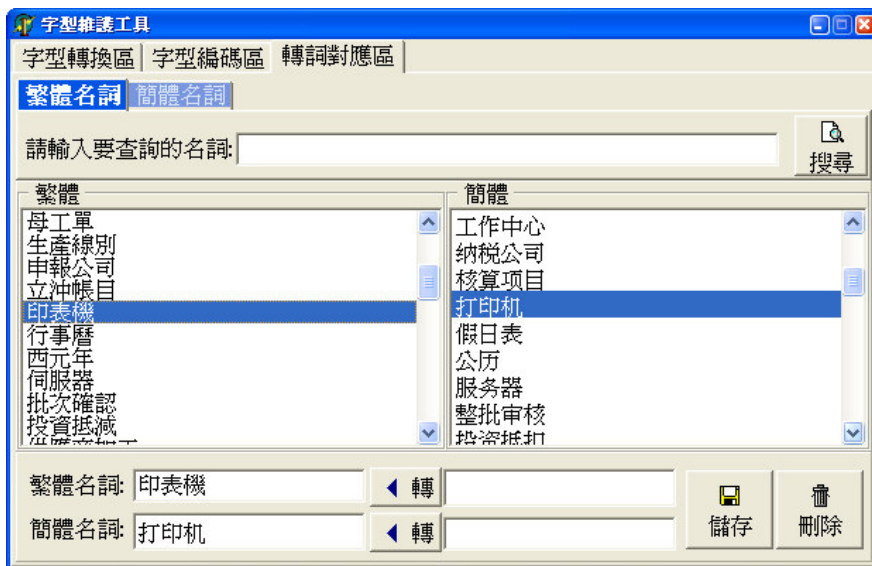


圖 3.17 使用者自行定義轉換詞句的工具

### 3.8 系統模型

在一個功能完備的 ERP 系統內，通常提供許多不同類型的作業供使用者操作，以鼎新電腦 ERP 系統為範例，我們大致可將作業區分為以下幾種類型：建檔類、報表類、批次作業、資料查詢作業、憑證類、工具類與流程圖類。以上這些分類就是我們做繁簡轉換的對象，在研究規劃的內容中，我們不但是要轉換操作畫面上的中文字，進一步對於資料庫所查詢到的資料也嘗試做轉換。

鼎新 ERP 系統是一套標準 3-Tier 架構的商業應用軟體，其 3-Tier 指的是資料庫層(Database)、應用層(Application)與使用者層(Client)，其中應用層與使用者層之間的連線方式是可以橫跨網際網路；運用此種架構，ERP 用戶可以是台灣地區的繁體中文的使用者，也可以是遠在對岸的簡體中文使用者，本研究所提供繁簡轉換的架構與方法即是針對此種需求應運而生。

鼎新 ERP 系統開發軟體及環境如表 3.7 所示，其程式開發工具採用 Borland 公司[25]所研發的 Delphi5.0 版本，此工具目前尚未支援萬國碼 Unicode 字集；資料庫是採用微軟所出的 SQL Server[26]英文版，資料存放字集宣告 CP950(即繁體中文字集)；報表資料庫字集宣告是採用 Paradox Taiwan 950，也是繁體中文；另外在搭配的作業系統是採用微軟的 Windows 平台，台灣地區使用的是繁體中文版，大陸地區使用的是簡體中文版。由上述的產品條件我們整理出在研究中做繁簡中文轉換的幾項規則：

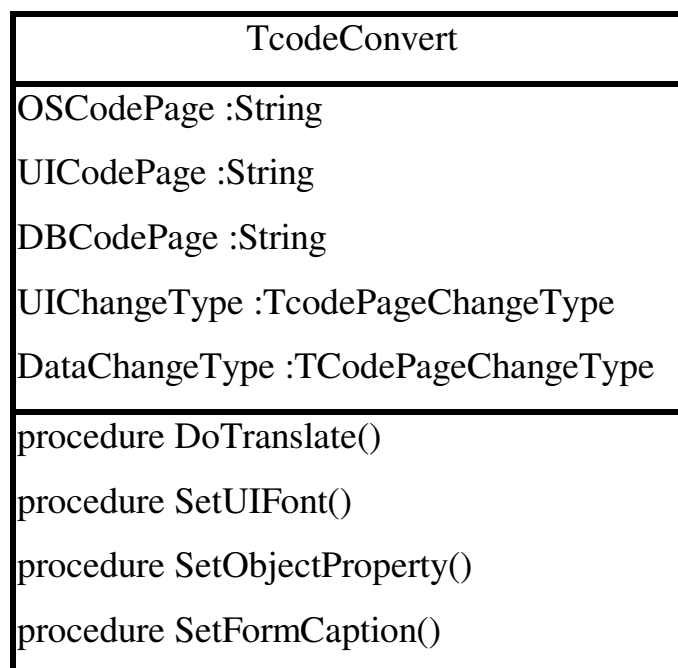
- 1) 轉換字集捨棄 Unicode 轉碼的方式，採用 BIG-5(繁)碼與 GB 碼(簡)字集。
- 2) 資料庫資料存放字碼全部轉成 BIG-5 碼(繁)，前端使用者在取用資料時依選取的語言別做轉換。
- 3) 繁(簡)體中文輸入法由使用者作業系統提供。

表 3.7 鼎新 ERP 系統開發軟體及環境

項目	說明
開發工具	Delphi5.0
資料庫	SQL Server7.0 , SQL Server2000
報表資料庫	Paradox
連結資料庫介面	BDE 5.11 版
作業系統	微軟 Windows 平台
Third Party 元件	DreamDesigner , QuickReport InfoPower , TeeChart

在本章節裡，我們先介紹整個轉換的流程，接下來再針對系統內各類型作業逐一整理出 Class Diagram 圖或架構圖，在系統展示方面我們整理出畫面中文轉換比對圖，藉由這些圖示的說明來驗證整個研究實作的成果。

系統整個轉換核心為一支轉換元件 TCodeConvert，負責讓所有作業呼叫以及提供轉換的動作，其 Class Diagram 如下圖所示：



```

function StrTransCode():Variant
function ConvertUIString():Variant
function ConvertDataString():Variant
function Big5ToGB():String
function GBToBig5():String
function DsCodeMessageDlg():Integer

```

圖 3.18 TcodeConvert Class Diagram

整個系統模型轉換的流程圖如下圖所示(圖 3.19)，其中左邊圖示說明字對字的轉換流程，而右邊圖示說明詞與詞轉換流程。

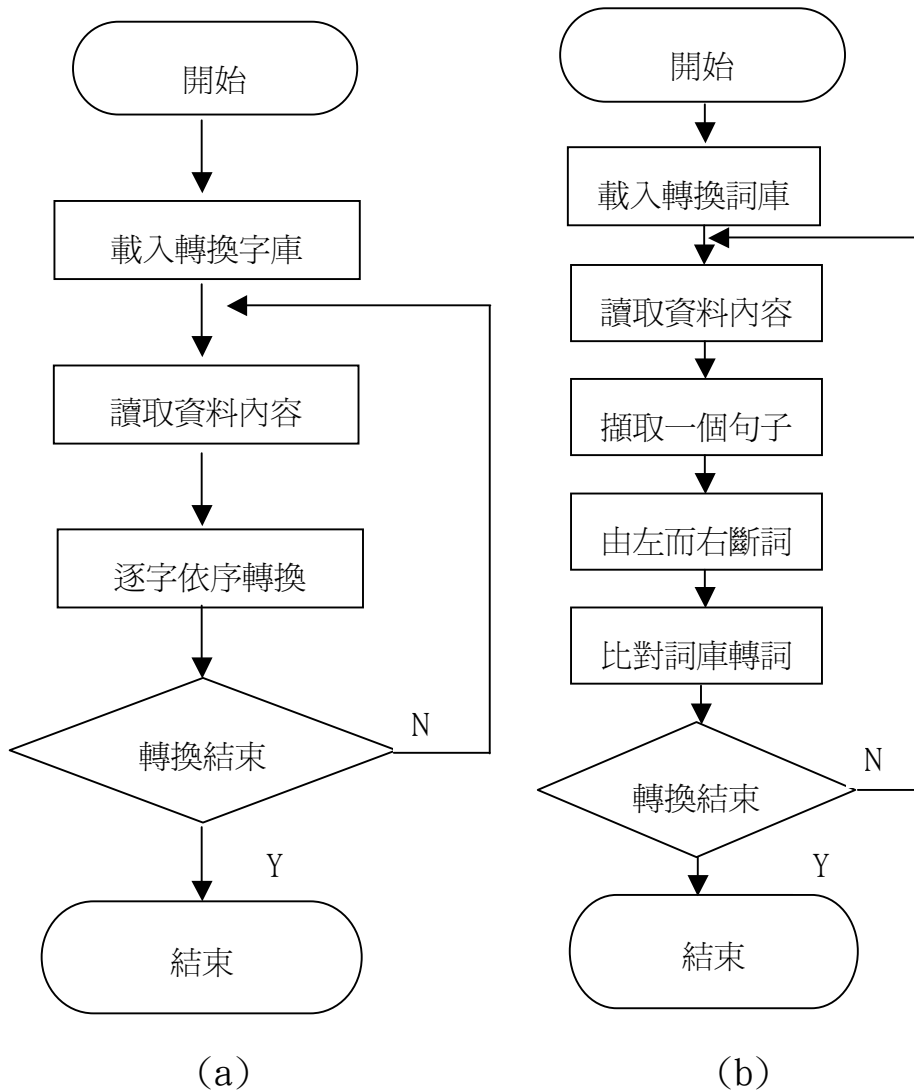


圖 3.19 轉換流程，(a)字與字的轉換，(b)詞與詞的轉換

整個系統模型設計在登入畫面來選擇語系(如圖 3.20)，後續的所有的作業皆依照該語系做外顯中文字的轉換處理。由於原始程式碼及資料是由繁體中文所開發，所以系統只針對使用者選擇簡體中文的部份做處理。



圖 3.20 登入畫面語系的選擇

系統登入後，接下來我們依照前述的分類逐一列出轉換的效果。

### 3.8.1 工具類與流程圖類

圖 3.21 為工具類程式的組成架構，其中 TDSForm 為所有外顯畫面的祖先，其他工具類的畫面全是繼承該祖先往下發展各別的程序畫面。圖中 TcodeConvert 內容請參考上圖 3.18 說明，TDSForm 內 TCodePageChangeType 是一個型態集合，內容如下：

TCodePageChangeType = (dtNONE, dtBTG, dtGTB)

dtNONE：表示不用轉換

dtBTG：表示 BIG-5 碼轉 GB2312 碼

dtGTB：表示 GB2312 碼轉 BIG-5 碼

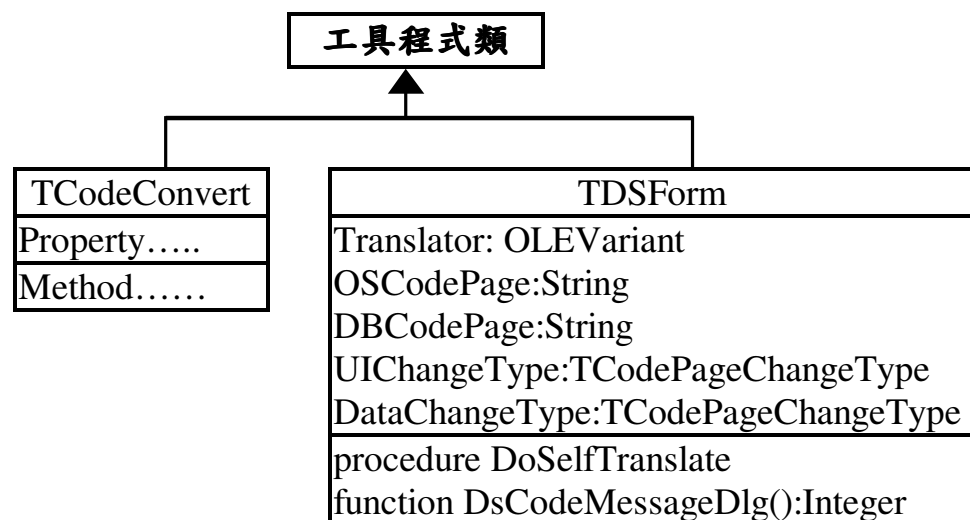


圖 3.21 工具程式類組成架構

圖 3.22 為 ERP 所有作業的管理中心，諸如此類的作業是屬於工具類的程式之一，其主要功能和管理 ERP 系統上的一些服務及周邊產品的整合。除此之外，有些工具是來自協力廠商的產品，比如與外商購買元件 DreamDesigner[27]、QuickReport[28]、InfoPower[29]、TeeChart[30](表 3.8 為外購元件廠商資訊)等等皆是。這部份的處理方式則必須與原來廠商購買原始的程式碼，配合核心元件 TcodeConvert 來做適當的修改。

表 3.8 ERP 外購元件廠商資訊

外購元件	原廠名稱	參考網址
DreamDesigner	Dream Company	http://www.dream-com.com/
QuickReport	QuSoft	http://www.qusoft.com/
InfoPower	woll2woll	http://www.woll2woll.com/
TeeChart	Steema Software SL compan	http://www.steema.com/

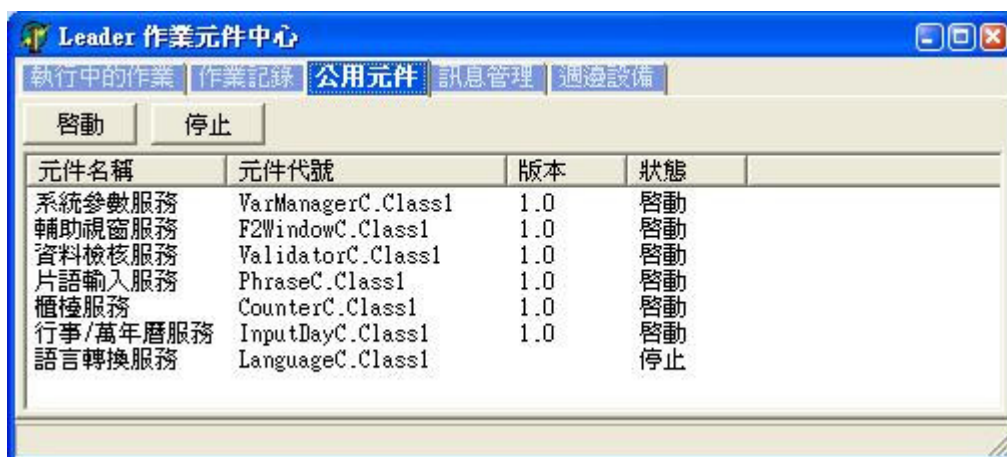


圖 3.22a ERP 工具類程式的畫面轉換(繁)

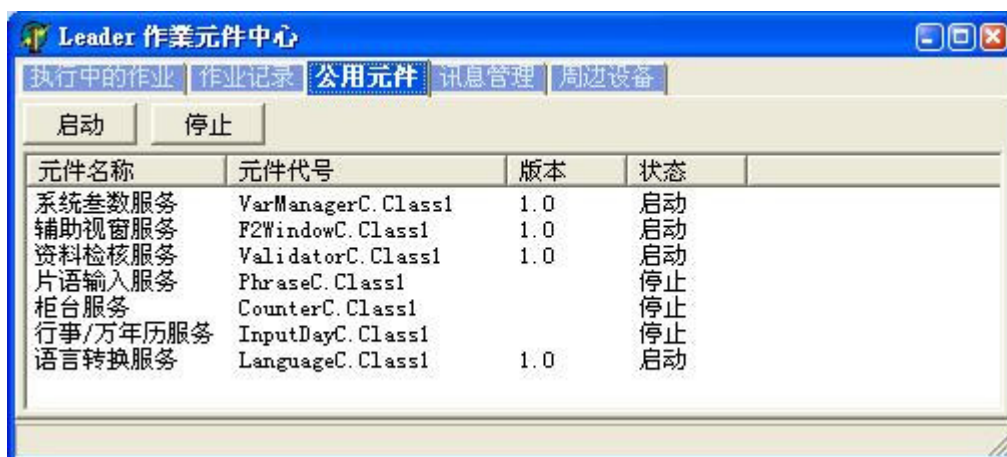


圖 3.22b ERP 工具類程式的畫面轉換(簡)

圖 3.23 是 ERP 系統輔助使用者建立資料的流程圖，使用者依照圖示內容建議的流程來依序建立資料。因圖形資料無法做轉換，所以我們製作兩份不同的圖檔來做動態的切換，比如：繁體庫存管理系統流程圖(圖形檔名：CHT\_INVFrame.jpg)與簡體庫存管理管流程圖(圖形檔名：CHS\_INVFrame.JPG)。



圖 3.23a ERP 圖形類的轉換(繁)



圖 3.23b ERP 圖形類的轉換(簡)



### 3.8.2 建檔類

鼎新 ERP 系統建檔類運用範圍、繼承架構及程式架構如下：

#### 1) 運用範圍

類型包括：參數設定(OneRec)、單據性質設定(Single)、單檔(Single)、雙檔(Double)。對於一般對話方塊有 DsDialogCF 可供使用，另有特殊結構之程式架構如一頭多身或有子單身之程式架構則有 SimpleBody 可供使用，對整體系統資料進行新增、刪除、修改、查詢等功能。

#### 2) 繼承架構

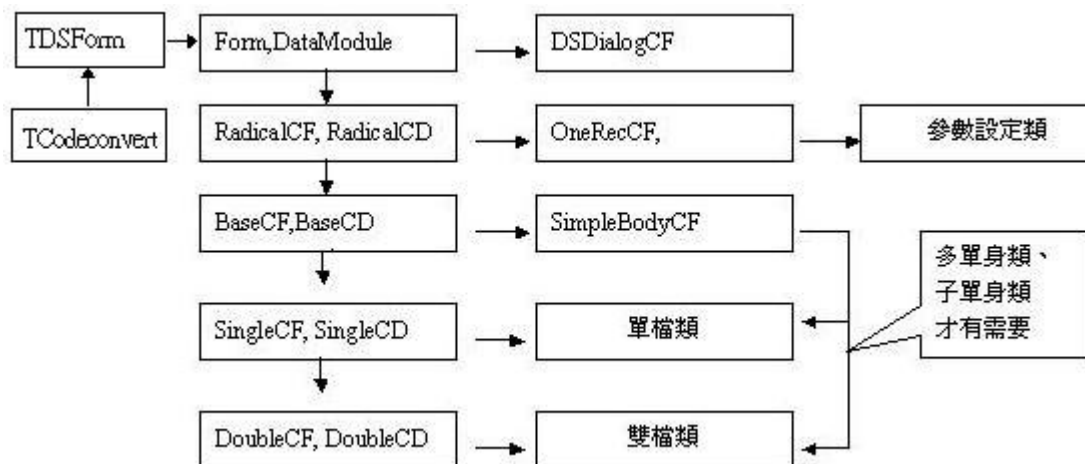


圖 3.24 建檔類繼承架構圖

#### 3) 程式架構

程式依照 3-Tier 架構分為 Client 端及 Application 端如圖 3.25 所示，其中 PURI01C.DLL 即是該支建檔的作業名稱。

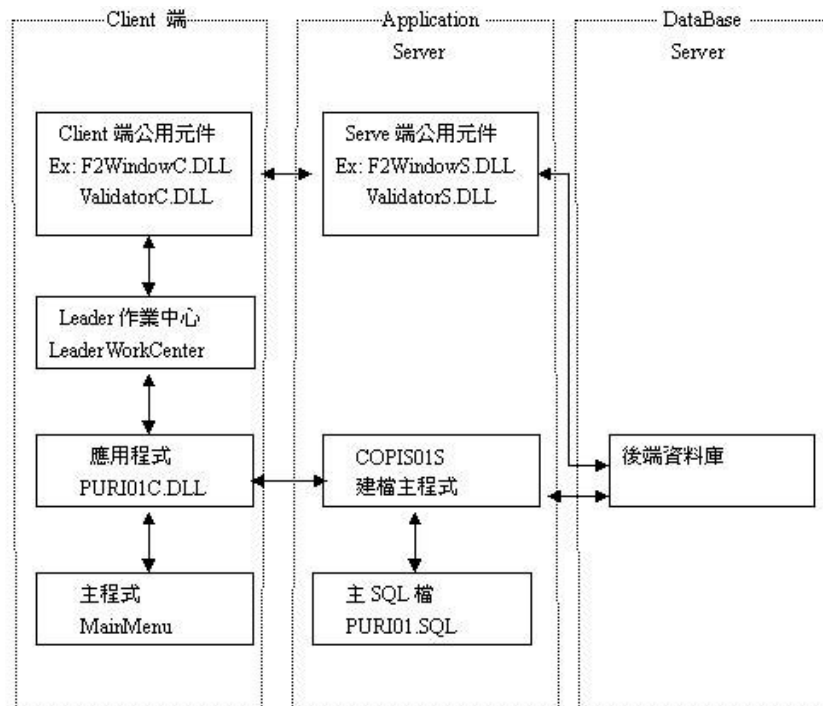


圖 3.25 建檔類程式架構圖

圖 3.24 為建檔類的組成架構，使用者可以透過建檔類的作業建立及儲存他們想要記錄的資料於資料庫，所以此類作業不但畫面上的操作介面需要轉換，進而對於輸入的文字或是從資料庫搜尋出來的資料也需一併做轉換。如圖 3.26 是 ERP 系統一支批號資料建立作業，畫面上我們可以觀察到外顯欄位如【品號】【品名】及內存資料欄位【辦公椅】【成品倉】即是一例。外顯欄位由程式啟動時轉換一次即可，但是內存資料欄位會隨使用者操作而不斷做轉換動作，所以此類作業大部份的轉換成本集中在資料欄位的轉換。

就我們實作的 ERP 系統而言，大概有 546 支程式屬於此類的作業，這表示幾乎所有系統操作者皆會有機會接觸到此類作業，所以架構在其上的繁簡文字轉換動作的順暢性與正確性額外顯得特重要。

異動日期	異動單別	異動單號	異動序號	庫別	庫別名稱	I/O別	異動別
88/05/24	232	88052400001	0001	1A	成品倉	0	2:銷貨
88/06/01	118	88060100001	0001	1A	成品倉	I	1:入庫

圖 3.26a 建檔類畫面與資料內容的轉換(繁)

异动日期	异动单别	异动单号	异动序号	库别	库别名称	I/O别	异动别
88/05/24	232	88052400001	0001	1A	成品仓	0	2:销货
88/06/01	118	88060100001	0001	1A	成品仓	I	1:入库

圖 3.26b 建檔類畫面與資料內容的轉換(簡)

### 3.8.3 報表類

鼎新 ERP 系統報表類運用範圍、繼承架構及程式架構如下：

#### 1) 運用範圍

類型包括：清單類報表(frmBaseQR)、明細類報表、統計類報表、彙總類報表(frmBaseCount)。

#### 2) 繼承架構

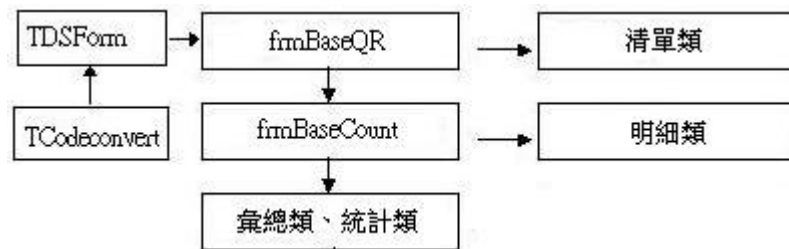


圖 3.27 報表類繼承架構圖

#### 3) 程式架構

報表作業同樣依照 3-Tier 架構分為 Client 端及 Application 端如圖 3.28 所示，其中 Invr04C.DLL 即是該支報表的作業名稱。

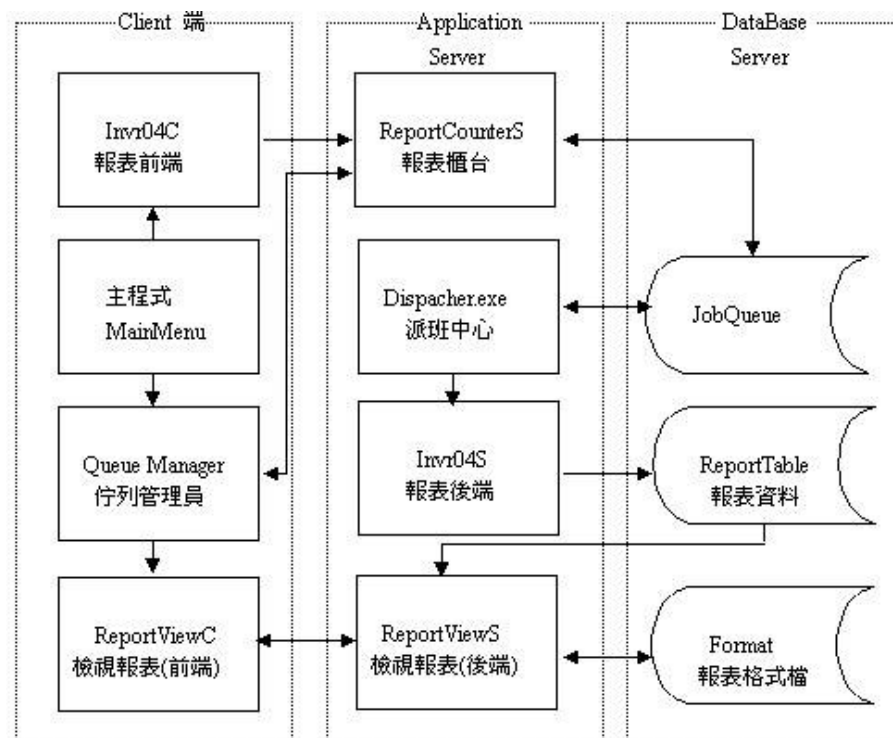


圖 3.28 報表類程式架構圖

報表類的作業主要在把存在資料庫的資料，依據商業性的法則整理成具有規則性的資料，再按照使用者下的條件選項篩選出想看的資料。由上面報表定義可得知，此類型的作業所呈現的資料幾乎全部由資料庫搜尋出來，所以所有外顯的畫面全部皆需要經過繁簡的轉換。

報表除了用來觀看之外，另外也提供電子檔案的轉出功能，比如轉出【Word】【Excel】【PDF】【網頁】【Mail】等等，硬體輸出方面又可分為傳真機及印表機。由於輸出繁簡文字的選擇皆視為與原操作畫面相同，所以不再經過第二次轉換，所見即為所輸出。

佇列工作管理員(Y) 視窗(Z)  
測試用公司  
科目編號資料清單  
製表日期: 94/11/10 第1頁

科目編號	科目名稱 科目別名	科目性質 報表借貸	資產損益別 科目類別	部門管理 慣用幣別	貨幣性科目 科目有效	財務比率分析類別 備註
1	資產類 Assets	帳戶 借方	資產負債 分類	N NT\$	Y Y	其他
11	流動資產 Current Assets	帳戶 借方	資產負債 分類	N NT\$	Y Y	其他
110	現金 Cash	帳戶 借方	資產負債 分類	N NT\$	Y Y	其他
1101	庫存現金 庫存現金	帳戶 借方	資產負債 獨立帳戶	N NT\$	Y N	其他
1102	銀行存款 Cash in Bank	帳戶 借方	資產負債 獨立帳戶	N NT\$	Y Y	其他
1104	匯撥中現金 匯撥中現金	帳戶 借方	資產負債 獨立帳戶	N NT\$	Y Y	其他
1105	零用金及週轉金 零用金及週轉金	帳戶 借方	資產負債 獨立帳戶	N NT\$	Y Y	其他
113	短期投資 Temporary Investment Tempor	帳戶 借方	資產負債 分類	N NT\$	Y Y	其他

座標 點

圖 3. 29a 報表類畫面與資料內容的轉換(繁)

佇列工作管理員(Y) 視窗(Z)  
測試用公司  
科目編號資料清單  
製表日期: 94/11/10 第1頁

科目編號	科目名稱 科目別名	科目性質 報表借貸	資產損益別 科目類別	部門管理 慣用幣別	貨幣性科目 科目有效	財務比率分析類別 備註
1	資產類 Assets	帳戶 借方	資產負債 分類	N NT\$	Y Y	其他
11	流動資產 Current Assets	帳戶 借方	資產負債 分類	N NT\$	Y Y	其他
110	現金 Cash	帳戶 借方	資產負債 分類	N NT\$	Y Y	其他
1101	庫存現金 庫存現金	帳戶 借方	資產負債 獨立帳戶	N NT\$	Y N	其他
1102	銀行存款 Cash in Bank	帳戶 借方	資產負債 獨立帳戶	N NT\$	Y Y	其他
1104	匯撥中現金 匯撥中現金	帳戶 借方	資產負債 獨立帳戶	N NT\$	Y Y	其他

座標 點

圖 3. 29b 報表類畫面與資料內容的轉換(簡)

### 3.8.4 批次作業

鼎新 ERP 系統批次類作業運用範圍、繼承架構及程式架構如下：

#### 1) 運用範圍

類型包括：逐筆執行類批次、整批執行類批次。

#### 2) 繼承架構

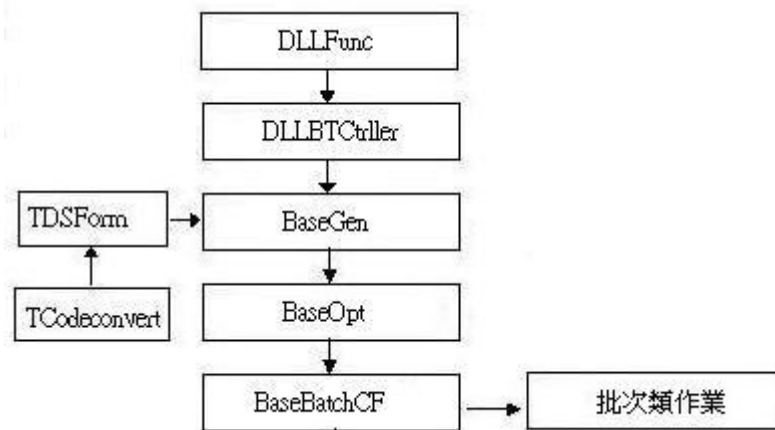


圖 3.30 批次類繼承架構圖

#### 3) 程式架構

批次類程式與報表類架構圖大同小異，差異不大，如圖 3.31 所示，其中 Purb04C.DLL 即是該支批次的作業名稱。

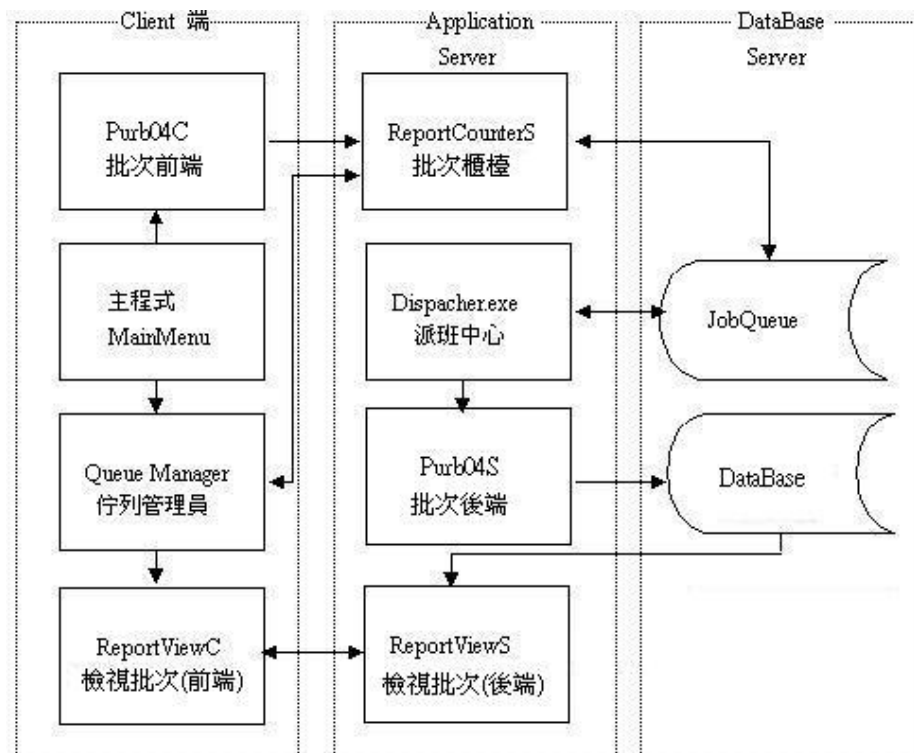


圖 3.31 批次類程式架構圖

圖 3.32 為 ERP 的批次作業，所謂的批次作業主要功能為大量產生資料。下圖為 ERP 的訂單轉請購單複製作業，使用者在畫面上可以挑選要參考的來源訂單以及要產生請購單別及請購日期，按下確定鈕後即由程式依序產生相對應的資料。

在產生資料的過程中，由於我們資料庫資料編碼是採取繁體中文，所以在運算的過程前一律先轉成繁體中文，資料運算後統一以繁體中文的資料型態存進資料庫。

圖 3.32a 批次作業類畫面與資料內容的轉換(繁)

圖 3.32b 批次作業類畫面與資料內容的轉換(簡)

### 3.8.5 資料查詢作業

查詢類作業與建檔類作業的架構類似，其運用範圍、繼承架構及程式架構如下：

#### 1) 運用範圍

提供線上快速查詢資料內容，類型包括：基本資料查詢、彙總資料查詢。

#### 2) 繼承架構

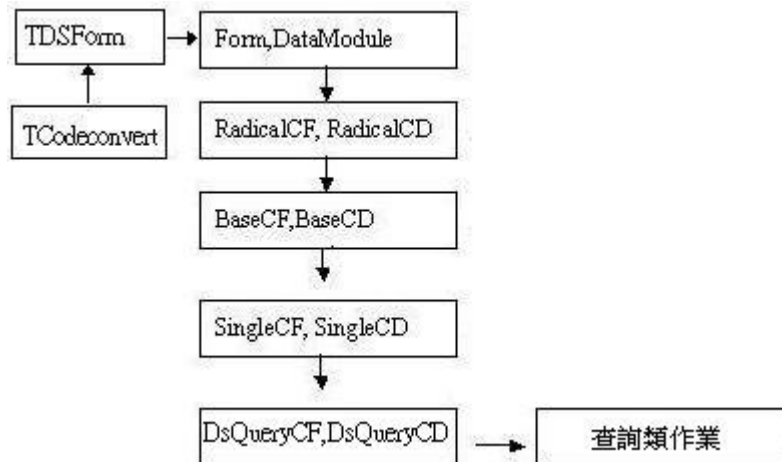


圖 3.33 查詢類繼承架構圖

#### 3) 程式架構

查詢類作業架構圖，如圖 3.34 所示，其中 CopQ01C.DLL 即是該支查詢作業的名稱。

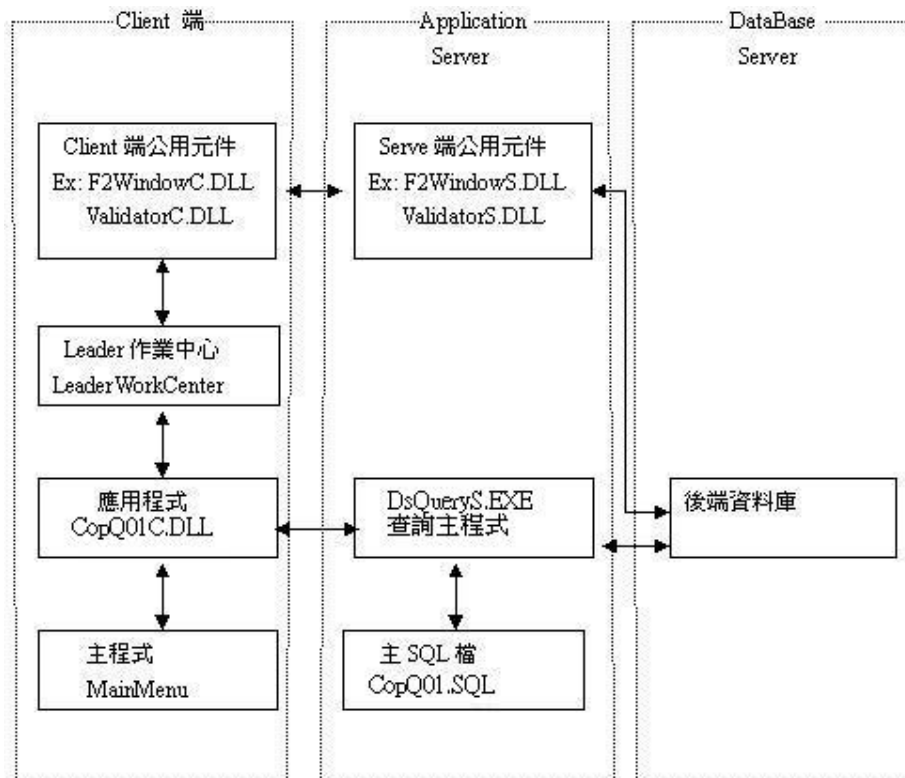


圖 3.34 查詢類程式架構圖

圖 3.35 為 ERP 系統的查詢作業(客戶資訊查詢作業)，此類作業特性如同報表類的程式，皆是經由搜尋資料庫的資料再依照特定的顯示規則，最後整理出使用者想要看的資料及格式呈現出來。

與報表不同的是此類作業不需要做輸出的動作，純粹只做畫面上的顯示，所以在繁簡中文內容的轉換上同建檔類作業，只做操作介面及資料內容轉換。

客戶代號	客戶簡稱	業務人員代號	業務員名稱	部門代號	部門名稱	交易幣別	折扣率	銷售評等
1101	大同家具	0558	朱南安	5400	出貨業務組	NT\$	0.00%	A
1102	和大家俱行	0558	朱南安	3500	營業國內部	NT\$	95.00%	A
1103	林森家俱	0558	朱南安	5400	出貨業務組	NT\$	0.00%	A
1105	東興傢具	0558	朱南安	3500	營業國內部	NT\$	0.00%	A
1900	零售商			3500	營業國內部	NT\$	0.00%	

品號	品名	規格	單位	幣別
111001	辦公椅SL	500	PCS	NT\$
111001	辦公椅SL	500	PCS	NT\$
111001	辦公椅SL	500	PCS	NT\$
25001	輪子	一般規格	SET	NT\$
25001	輪子	一般規格	SET	NT\$
25001	輪子	一般規格	SET	NT\$
25001	輪子	一般規格	SET	NT\$

圖 3.35a 查詢作業類畫面與資料內容的轉換(繁)

客戶代號	客戶簡稱	業務人員代號	業務員名稱	部門代號	部門名稱	交易幣別	折扣率	銷售評等
1101	大同家具	0558	朱南安	5400	出貨業務組	NT\$	0.00%	A
1102	和大家俱行	0558	朱南安	3500	營業國內部	NT\$	95.00%	A
1103	林森家俱	0558	朱南安	5400	出貨業務組	NT\$	0.00%	A
1105	东兴家具	0558	朱南安	3500	營業國內部	NT\$	0.00%	A
1900	零售商			3500	營業國內部	NT\$	0.00%	

品号	品名	规格	单位	币别
111001	办公椅SL	500	PCS	NT\$
111001	办公椅SL	500	PCS	NT\$
111001	办公椅SL	500	PCS	NT\$
25001	轮子	一般规格	SET	NT\$
25001	轮子	一般规格	SET	NT\$
25001	轮子	一般规格	SET	NT\$
25001	轮子	一般规格	SET	NT\$

圖 3.35b 查詢作業類畫面與資料內容的轉換(簡)



### 3.8.6 憑證類

憑證類作業主要是提供交易行為過程中的憑單，其運用範圍、繼承架構及程式架構如下：

#### 1) 運用範圍

類型包括：交易單據類的憑證、標籤類的憑證。

#### 3) 繼承架構

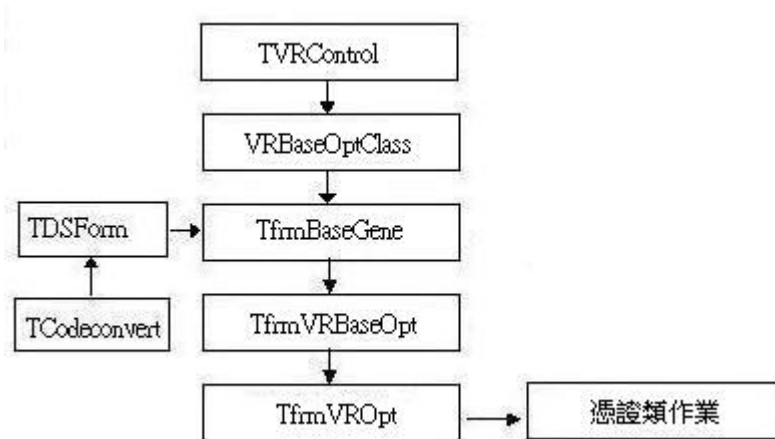


圖 3.36 憑證類繼承架構圖

#### 3) 程式架構

憑證類作業架構圖，如圖 3.37 所示，其中 Acpr01C.DLL 即是該支憑證作業的名稱。

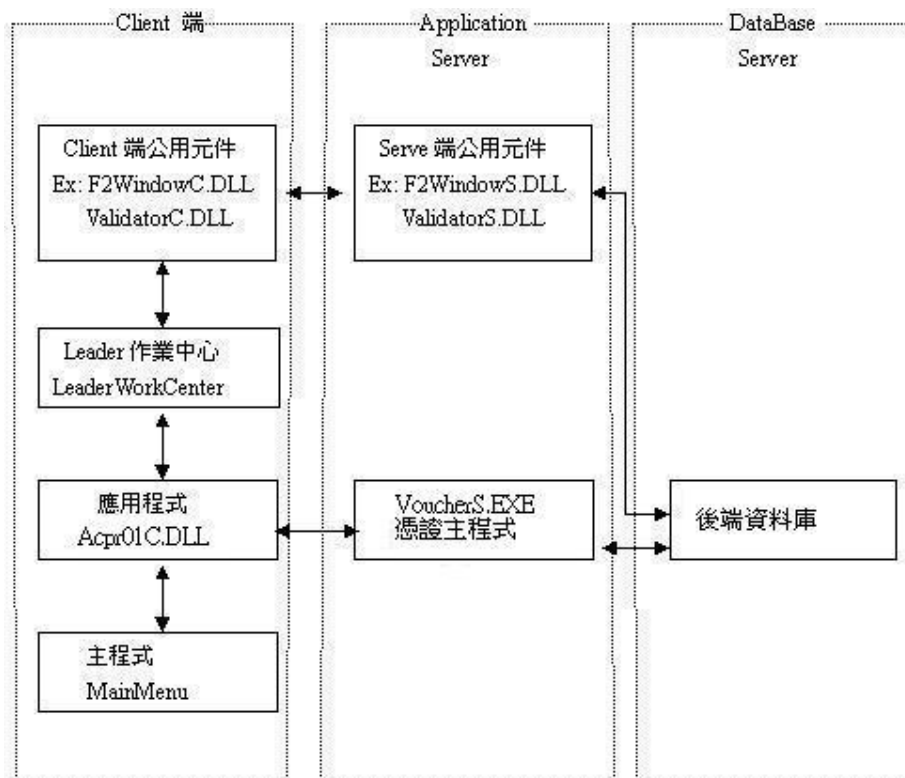


圖 3.37 憑證類程式架構圖

圖 3.38 為 ERP 系統的一支憑證(客戶訂單憑證)，憑證類的作業主要在於產生 ERP 系統在交易行為上的證明文件。

以本研究的 ERP 系統為例，其憑證畫面構成的要素為一張底圖與底圖上面的動態資料，所以在文字的轉換上分為圖形文字轉換與資料內容的轉換。圖形文字的轉換同前面流程圖的方式，採取兩份底圖的做法，在程式運作過程中動態載進繁體或簡體的底圖；文字的轉換時機點抓在資料呈現出來之前。

### 測試用公司 國內訂單

公司電話: 公司傳真: 頁次: 1/1  
製表日期:94/11/10

訂單單別:220 國內訂單 訂單單號:88062300001 單據日期:88/06/23 客戶代號:1101 大同家具 電話: 傳真: 送貨地址:	客戶單號: 統一編號: 交易幣別:NT\$ 匯率:1 課稅別:應稅外加 營業稅率:5%	出貨廠別:001 部門:5400 業務人員:0558 價格條件: 付款條件:	台北廠 出貨業務組 朱南安 材積單位: CUFT 備註: 確認碼:Y
---	--	--	---

序號	品名 規格 客戶品號	訂單數量 已交數量 贈品量 贈品已交量	單位 小單位 折扣率 包裝方式	訂單單價 訂單金額 毛重(Kg) 材積	預交日 交貨庫別 前置單別 前置單號	專案代號 備註 結案碼
0001	111001 辦公椅SL 500	300 300 0 1	PCS 100%	480 144000 0 0	88/06/24 1A 0 0	自動結案
0002	111002 辦公椅BL 500KL	200 200 0 0	SET 100%	540 108000 0 0	88/06/24 1A 0 0	自動結案
	以下空白//					

圖 3.38a 憑證作業類畫面與資料內容的轉換(繁)

### 测试用公司 国内订单

公司电话: 公司传真: 页次: 1/1  
制表日期:94/11/10

订单单别:220 国内订单 订单单号:88062300001 单据日期:88/06/23 客户代号:1101 大同家具 电话: 传真: 送货地址:	客户单号: 统一编号: 交易币别:NT\$ 汇率:1 课税别:应税外加 营业税率:5%	出货厂别:001 部门:5400 业务人员:0558 价格条件: 付款条件:	台北厂 出货业务组 朱南安 材积单位: CUFT 备注: 确认码:Y
---	--	--	---

序号	品名 规格 客户品号	订单数量 已交数量 赠品量 赠品已交量	单位 小单位 折扣率 包装方式	订单单价 订单金额 毛重(Kg) 材积	预交日 交货库别 前置单别 前置单号	项目代号 备注 结案码
0001	111001 办公椅SL 500	300 300 0 1	PCS 100%	480 144000 0 0	88/06/24 1A 0 0	自动结案
0002	111002 办公椅BL 500KL	200 200 0 0	SET 100%	540 108000 0 0	88/06/24 1A 0 0	自动结案
	以下空白//					

圖 3.38b 憑證作業類畫面與資料內容的轉換(簡)

## 第四章效能分析

許多支援顯示介面繁簡切換的軟體或網頁，常常採取的做法是同時維護兩套不同語言的原始碼，再依照使用者選擇的語言別(繁體或簡體)分別提供不同的操作介面。此方法的好處是執行速度不受切換繁簡動作的影響，缺點是對於一些動態的資料無能力即時做轉換且系統維護尚須付出雙倍的成本。本研究所提出的方法可有效降低系統維護的成本及支援線上即時轉換的功能。在整個效能表現上也有令人滿意的結果，其主要的原因在於(1)預先把字庫載入記憶體(2)字碼轉換採取直接由原字碼運算得出其對應字碼的陣列位置，一次搜尋即可找到要轉換的字碼。

### 4.1 字對字轉換

字對字的轉換單純是以一個中文字對一個中文字的轉換，不考慮句子前後的關係，這種方式的轉換在效能上表現為最好。針對字碼轉換部份，效能測試數據圖如圖 4.1。其中測試的硬體環境為：

CPU：Pentium 1.3GHZ

記憶體：512MB

作業系統：Windows XP SP1

硬碟空間：80GB

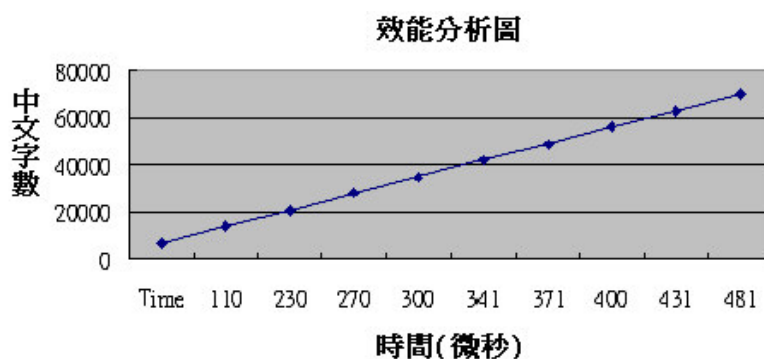


圖 4.1 轉字效能測試數據圖

## 4.2 詞對詞轉換

中文字轉詞的精確度通常與執行效能是成反比的，也就是說要達到百分之百的準確度，相對要建立的詞庫就大，程式運算邏輯也複雜很多。針對詞庫瘦身的部份，本研究採取詞庫分割的概念，即是詞庫之下再依照不同類型來分類，如給 ERP 應用系統就用 ERP 詞庫、電腦操作手冊就使用電腦專有名詞詞庫等等。本研究除了運用詞庫分割來增進效能，另外在詞庫的搜尋及轉換上，採取(1)預先把詞庫載入記憶體(2)詞彙轉換採取直接由原詞碼運算得出其對應詞碼的陣列位置，一次搜尋即可找到要轉換的詞句群組(3)詞句排列依照長度由大排到小，長詞優先的斷詞方法。

整個效能測試數據圖如圖 4.2，詞庫大小約 18000 筆資料。由圖得知，在轉換一篇一萬個中文字的文章大概花了 0.3 秒的時間，整個呈現出來的效能還算令人滿意。其中測試的硬體環境為：

CPU：Pentium 1.7GHZ

記憶體：2GB

作業系統：Windows XP SP1

硬碟空間：80GB

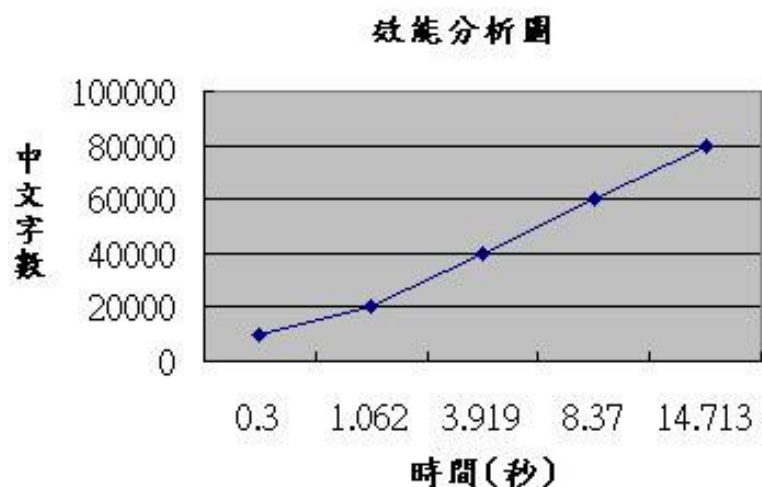


圖 4.2 轉詞效能測試數據圖

### 4.3 資料內容轉換

本研究轉換的對象不侷限於畫面上靜態的中文字，對於由資料庫存取出來的動態中文資料也進行做轉字與轉詞的動作。每次轉換的資料量大約 2000 個中文字數，其呈現出來的效能相當於轉字時間加上轉詞的時間，以一萬個中文字內容來測試，轉換的時間約耗費 0.41 秒的時間。

## 第五章 結論

中文字是中華五千年文化的精華，其組成概念來自於象形、會意、形聲等等，是我們老祖宗的智慧結晶。目前雖因政治環境的因素導致有繁體中文與簡體中文的區別，但是其本質是不變的，藉由本研究所提出的方法，希望能更拉近雙方彼此間的關係，熱絡彼此經貿以及文化上經驗的交流。

隨著兩岸政策的開放，ERP 系統也必須同時支援兩種不同文化背景及語言的使用者來操作。本研究利用中文字的內碼特性，配合建立好的字庫及詞庫，可以提供一種即簡單又快速的繁簡中文轉換方法。研究中所提出的方法並不侷限應用於 ERP 系統，諸如 BBS、Web Mail、搜尋引擎、繁簡網頁、線上遊戲、兩岸資料的傳遞皆可適用。能夠幫助兩岸資料交流順暢，不管是由簡入繁還是由繁返簡，皆能透過一些簡單的轉換技巧，讓使用者能無閱讀資料的障礙是本研究最原始的動機。另外把該方法實作在企業的應用軟體上，驗證該方法的確是可行方案，幫助企業花最少的成本，快速建置具有彈性且合用的繁簡 ERP 系統是本研究最終的目標。

## 參考文獻

- [1] E-Works 中國製造業信息化網站，網址：  
<http://www.e-works.net.cn/xxhjj/jj8-1.htm>
- [2] 行政院主計處電子處理資料中心，「中文碼簡介與定義」，  
CNS11643 中文標準交換碼全字庫，網址：  
<http://www.cns11643.gov.tw/web/word.jsp#1>
- [3] 行政院主計處電子處理資料中心，「BIG-5碼介紹」，CNS11643  
中文標準交換碼全字庫，網址：  
<http://www.cns11643.gov.tw/web/word.jsp>
- [4] 邱俊穎(民90)，「中文網站繁簡互訪協助系統」，中央大學，資  
訊工程研究所碩士論文，第6-9頁
- [5] 行政院主計處電子處理資料中心，「中文碼簡介與定義」，  
CNS11643中文標準交換碼全字庫，「BIG-5碼主要版本」，網址  
：<http://www.cns11643.gov.tw/web/big5/index.html>
- [6] 維文編碼標準化信息，「GB2312-80介紹」，網址：  
<http://www.uighurlinux.org/gbcode.htm>
- [7] The Unicode Consortium, 網址：<http://www.unicode.org/>
- [8] 陳克健、陳正佳、林隆基，「中文語句分析的研究-斷詞與構詞  
」，中央研究院資訊所技術報告，TR86-004，1986
- [9] Kai-Ying Liu, and Jia-Heng Zheng, "Research of automatic Chinese  
word segmentation", Machine Learning and Cybernetics, Vol.2,  
pp.805 - 809,. Nov. 2002

- [10] Keh-Hwa Shyu, Chun-Jen Lee, and Mu-King Tsay, "An OCR based translation system between simplified and complex Chinese characters", Pattern Recognition, Vol.2, pp.368 - 370, Oct. 1994
- [11] 張如瑩(民90), 「多語系平行關鍵頁搜尋引擎之設計與建構」, 元智大學, 資訊管理研究所碩士論文, 第14-18頁
- [12] 中央研究院資訊科學研究所, 「中文詞知識庫」, 網址：  
<http://ckipsvr.iis.sinica.edu.tw/>
- [13] 兩岸文件通, 網址：<http://www.winpreton.com/>
- [14] 南極星, 網址：<http://www.njstar.com/njstarc.php>
- [15] 信使網路繁簡通, 網址：<http://www.infoscape.com.cn/>
- [16] 簡繁轉換網, 網址：<http://www.web-change.com/>
- [17] 簡體網頁轉成繁體網頁, 網址：<http://chinagate.yam.com/>
- [18] 鼎新電腦股份有限公司, 網址：  
<http://www.dsc.com.tw/dsc/dsc1.asp>
- [19] 黃景祥, "三層式架構－應用系統概論", 網址：  
<http://www.ares.com.tw/communicate/magazine/資通通訊雜誌第22期/三層式架構應用系統概論.htm>
- [20] Jack Halpern, and Jouni Kerman, "The Pitfalls and Complexities of Chinese to Chinese Conversion", 網址：<http://www.basistech.com/>
- [21] 資訊與通訊標準推廣網站－中文碼之應用與發展, 網址：  
<http://www.tips.org.tw/nii/storehouse/chinese/87-1.htm>
- [22] 有關內碼的小知識, 網址：<http://peopledaily.com.cn/>
- [23] 華通之家, 學中文用電腦, 「中文的內碼特別多」, 網址：  
<http://home.netteens.net/~bookcity/yiyan9821.htm>



- [24] Xiaohong Huang, Zhensheng Luo, and Jian Tang, "A quick method for Chinese word segmentation," Vol.2, pp. 1773 – 1776, Oct. 1997
- [25] 台灣寶蘭，網址：<http://www.borland.com/tw/>
- [26] Microsoft SQL Server Home, 網址：  
<http://www.microsoft.com/sql/>
- [27] DreamDesigner, Dream Company, 網址：  
<http://www.dream-com.com/>
- [28] QuickReport, QuSoft, 網址：<http://www.qusoft.com/>
- [29] InfoPower, woll2woll, 網址：<http://www.woll2woll.com/>
- [30] TeeChart, Steema Software SL compan, 網址：  
<http://www.steema.com/>