行政院國家科學委員會專題研究計畫 成果報告

智慧財產價值分析系統計畫〔III〕
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 96-2623-7-004-001-
執 行 期 間 ：96 年 08 月 01 日至 97 年 07 月 31 日
執 行 單 位 ：國立政治大學智慧財產研究所

計 畫 主 持 人：劉江彬

計畫參與人員：教授-主持人(含共同主持人)：劉江彬
助理教授-主持人(含共同主持人)：邱仁鈿

處 理 方 式 ：本計畫涉及專利或其他智慧財產權，2 年後可公開查詢

中 華 民 國 97 年 06 月 05 日

關 鍵 字：美國專利轉讓資料、美國專利資料、財務報表、專利家族、專利紅皮書、延伸標記語言

中文摘要：

本研究主要目的是在研究解析 2007 年美國專利轉讓資料、2007 美國專利之 INPADOC 專利家族資料及 2007 新版美國專利磁帶資料，並尋找一種方法將其分解、統整並匯入資料庫。另外針對美國上市上櫃公司財報資料，亦研究一種方法可批次取得。

ABSTRACT

Keyword: USPTO Patent Assignment Data、USPTO Patent Data、Patent Family、Financial Report、Patent Red Book、XML

The goal of this study is to understand the schema of 2007 USPTO Patent Assignment Data、2007 INPADOC patent family data of patent and the 2007 version of USPTO Patent Data. And then find a way to parse, aggregate and import into relational database. For financial report, I will study the best way to batch download the data from SEC Edgar.

目　錄

頁 次

目　錄.....III

第壹章 前言.....1

第貳章 研究目的.....2

第一節 研究動機與目的.....2

第二節 研究架構與研究流程.....2

第參章 文獻探討.....4

第一節 XML（EXtensible Markup Language）.....4

第二節 INPADOC 的資料擷取.....5

第三節 美國證期會財報資料庫.....5

第四節 Redbook 於 2007 年的改變.....6

第肆章 研究方法.....7

第一節　　系統架構.....7

第二節　　實驗流程.....7

第伍章 結論與建議.....9

第一節 研究結論.....9

第壹章 前言

在知識經濟時代來臨以後，企業經營的致勝關鍵已經不再是靠廠房、設備、勞力及資本數量等因素來影響。而是漸漸的被具有高附加價值、可重複再利用的智慧財產取代。智慧財產所包含的範圍很廣，例如品牌、產品、產值、市場、投資組合、技術授權、競爭情報等等商業資訊，除此之外，專利是最能代表一家公司的創新研發能量的智慧財產之一。在近年來政府所推動的兩兆雙星或者綠色矽島等計畫中，都將創新研發視為高科技產業的核心競爭力指標。也因此專利受到政府或者企業經營者的重視。擁有了創新技術的智慧財產權，等於是在創新研發上製造了門檻。不僅能創造競爭優勢，更能透過專利的授權、轉移、佈局、融資、作價投資等策略提高獲利機會。

然而專利資訊本身是高度結構化的資訊。以往這類資料皆為紙本，在資訊的取得上既不方便也沒有效率。在各國專利局的努力之下，近十年的專利電子化已經漸漸普遍。在各國皆以開始公布電子化專利資訊的同時，資訊格式不統一的問題也開始發生。所以先進國家的專利局便開始發展統一的專利描述語言，使用的描述語法為 XML。本研究是針對美國專利局所發佈的新版 Redbook ICE 4.2 格式資料作為研究標的。進行有效率資料格式的拆解與歸納統整。

第貳章 研究目的

第一節 研究動機與目的

本研究主要目的是針對美國專利局所發佈之專利轉讓 XML 資料以及 2007.02.01 版磁帶資料（格式為 Redbook ICE 4.2）進行研究。並將其拆解統整至資料庫。以期後續可於專利指標分析上進行輔助。希望研究成果有助於瞭解完整之專利結構化資料於專利指標分析之研究，並能提出後續研究之方向及改善方法。另外，為了能夠透過美國專利資料分析專利全球佈局情形，本研究會透過 INPADOC 資料庫進行專利家族資訊蒐集。因此，本論文的研究目的可歸納如下：

1. 整理目前國內外相關文獻，研究美國專利轉讓資料XML格式。
2. 整理目前國內外相關文獻，研究新版Redbook ICE 4.2格式。
3. 整理目前國內外相關文獻，研究INPADOC資料取得及解析方式。
4. 整理目前國內外相關文獻，研究美國財報資料格式及擷取方式。
5. 研究目前國內外相關專利資訊提供網站，透過操作瞭解其背後資訊整合的架構。
6. 配合專利本身特性，發展合適且有效率之拆解以及統整方法。
7. 實作一系統，可有效率的針對美國專利局之磁帶資料進行拆解並匯入資料庫。

第二節 研究架構與研究流程

本研究將針對美國專利局專利轉讓 XML 資料及新版 Redbook ICE 4.2 磁帶資料進行研究與分析，主要分成三大部分：

1. 定義與設計：首先確定研究主題與目的
2. 實驗：以參考文獻為基礎，擬訂研究方式與架構，往後則建立實驗系統，

並藉由實驗數據以及資料累積來進行調整。而後不斷進行修正。

3. 結論：藉由實驗數據來進行探討，得到本研究的成果。

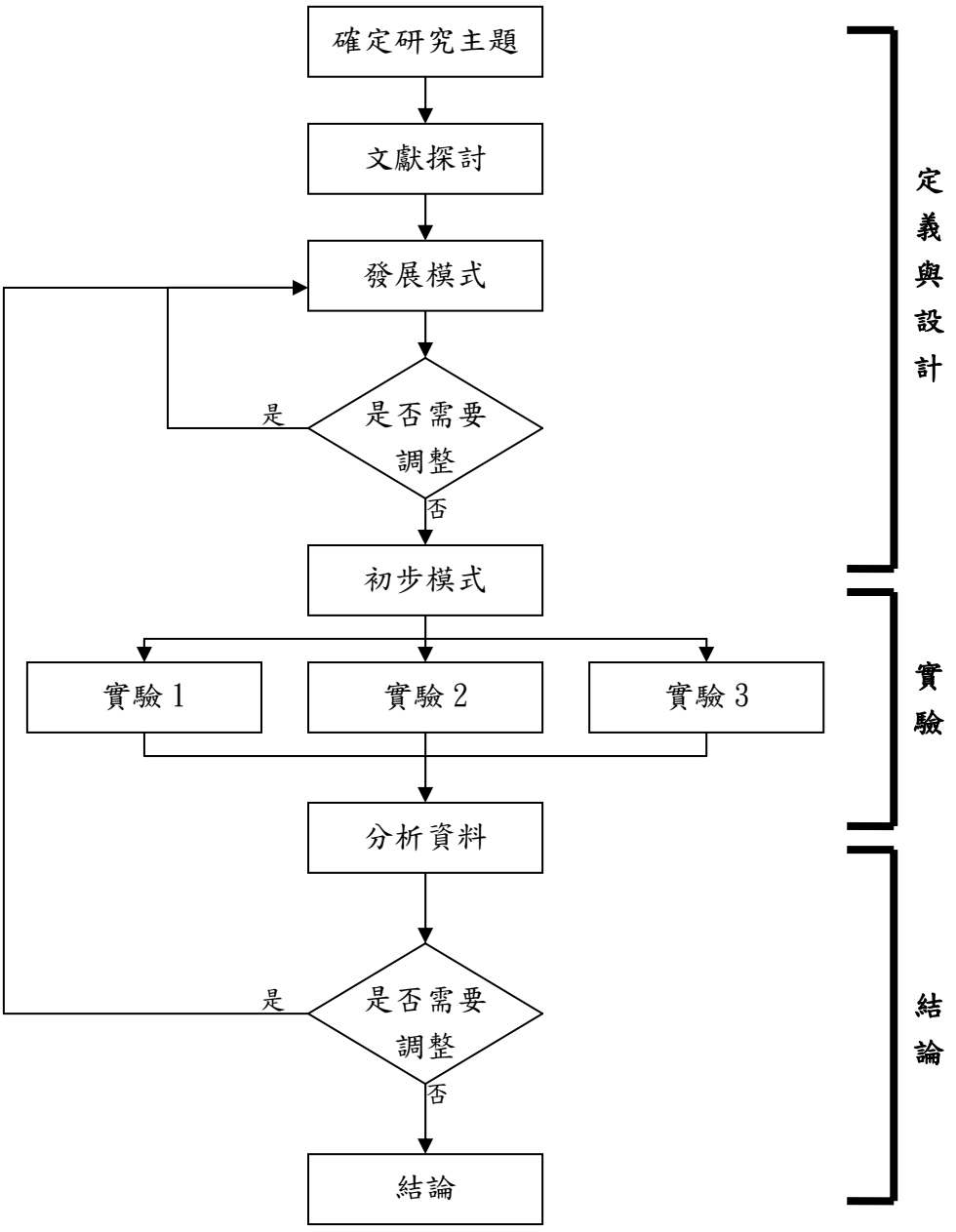

圖 1 研究架構

第參章 文獻探討

第一節 XML（EXtensible Markup Language）

(一) XML 的定義與歷史

XML 全稱為 Extensible Markup Language，中文翻譯為可延伸標記語言、可擴展標記語言或者可延伸標示語言。是一種置標語言。標記指電腦所能理解的訊息符號，通過此種標記，電腦之間可以處理包含各種信息的文章等。如何定義這些標記，既可以選擇國際通用的標記語言，比如 HTML，也可以使用象 XML 這樣由相關人士自由決定的標記語言，這就是語言的可擴展性。XML 是從標準通用置標語言（SGML）中簡化修改出來的。它主要用到的有 XML、XSL、XBRL 和 XPath 等[Wikipedia]。

XML 是從 1996 年開始有其雛形，並向 W3C（全球資訊網聯盟）提案，而在 1998 二月發佈為 W3C 的標準（XML1.0）。XML 的前身是 SGML（The Standard Generalized Markup Language），是自 IBM 從 60 年代就開始發展的 GML（Generalized Markup Language）標準化後的名稱。而 GML 的重要概念為有兩個，一為文件中能夠明確的將標示與內容區隔，二為有文件的標籤使用方法均一致[Wikipedia]。

同時 W3C 意識到 HTML 的原罪[Wikipedia]：

不能解決所有解釋資料的問題 - 像是影音檔或化學公式、音樂符號等其他型態的內容。

效能問題 - 需要下載整份文件，才能開始對文件做搜尋的動作。

擴充性、彈性、易讀性均不佳。

為了解決以上問題，專家們使用 SGML 精簡製作，並依照 HTML 的發展經驗，產生出一套使用上規則嚴謹，但是簡單的描述資料語言，正是所謂的 XML[Wikipedia]。

第二節 INPADOC 的資料擷取

(一) Open Patent Service

INPADOC 提供 Open Patent Service 這樣的 Web Service 服務，可以透過該服務取得特定專利之專利家族資訊。服務的 WSDL 在 http://ops.espacenet.com/OpenPatentServices/webService/getPatentData?wsdl 可以瀏覽。

第三節 美國證期會財報資料庫

（一）財報資料庫檢索模式

財報資料的擷取與專利不大相同。專利由於數量過於龐大，因此多為透過檢索的方式至 USPTO 查詢符合條件的專利，再行下載；而財報則是透過公司名稱作為區隔，可以查到每個公司各年度發布的各種資訊。

因此於本系統中，建立的財報檢索擷取介面，也應從公司別、年度、報告類型作為出發點進行擷取。美國財報的擷取，使用者可以選擇輸入公司名稱或 CIK 代碼、報表類型、以及資料時間範圍，系統便會將該篇財報從美國證期會擷取回系統。同樣的，如果該篇資料已存在於系統之中，則不會重複擷取。

（二）財報資料庫瀏覽

財報的格式多半以 HTML、PDF、Word 文件呈現，其中

HTML 的部分系統會自動進行適當的項目拆解。唯至於 PDF 以及 Word 文件，由於其格式限定，並沒有辦法輕易的做到這種拆解或是 highlight 的功能。

第四節 Redbook 於 2007 年的改變

（一）官方的改變說明

美國專利局針對 RedBook 的資料格式，在 2007 年於前置處理語言上有做了很多的改進。透過這些改進將有助於解析引擎的產生。本案前期的成果報告中已經有針對 2007 初所修正的版本進行研究。本期不再說明。

第肆章 研究方法

本研究綜合第參章相關文獻的探討結果，進行研究之設計，研究設計中包含了系統架構、設計、實驗流程以及驗證方法，分述如下。

第一節 系統架構

系統包含解析引擎產生器、專利資料驗證器、解析引擎評估及使用者介面四大部分。每個部分的說明如下：

1. 解析引擎產生器：負責從專利轉讓資料 DTD 以及新版Redbook ICE 4.2 DTD產生專利資料解析引擎，並將專利欄位拆解，取得專利書目欄位。
2. 專利資料驗證器：透過準備好的驗證資料，驗證解析引擎的正確性，以及資料的完整性。
3. 分類引擎評估：透過物件分析、資料掃瞄、時間判讀以及空間應用等方面，評估解析引擎的整體效能。
4. 使用者介面：使用者可以透過使用者介面操作系統的相關功能。

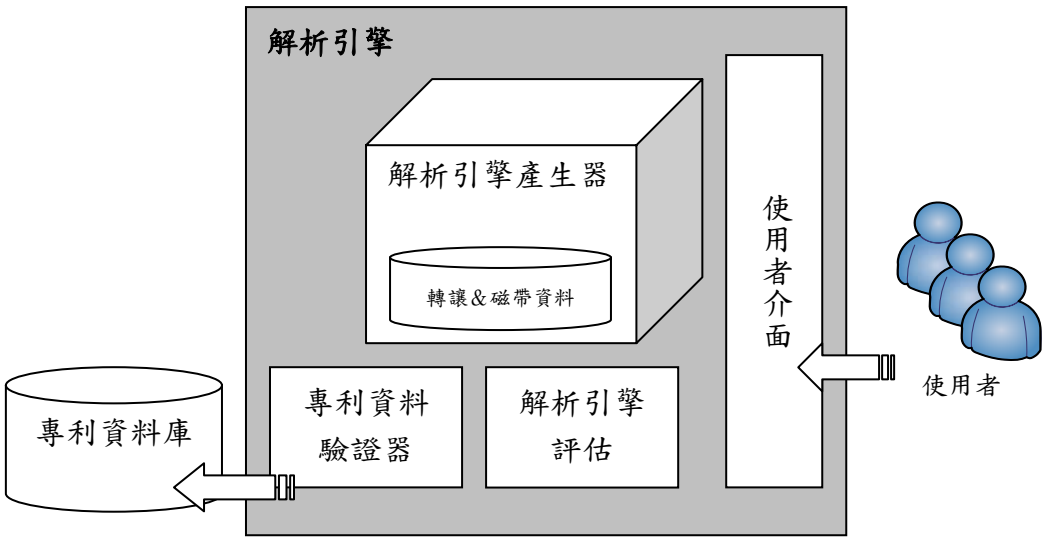

圖 2 系統架構圖

第二節 實驗流程

本研究透過事先設計好之實驗流程進行研究。透過依系統化的流程可清楚得知各步驟的產出以及預期成果。並可重複套用到其他欲研究之資料格式上。

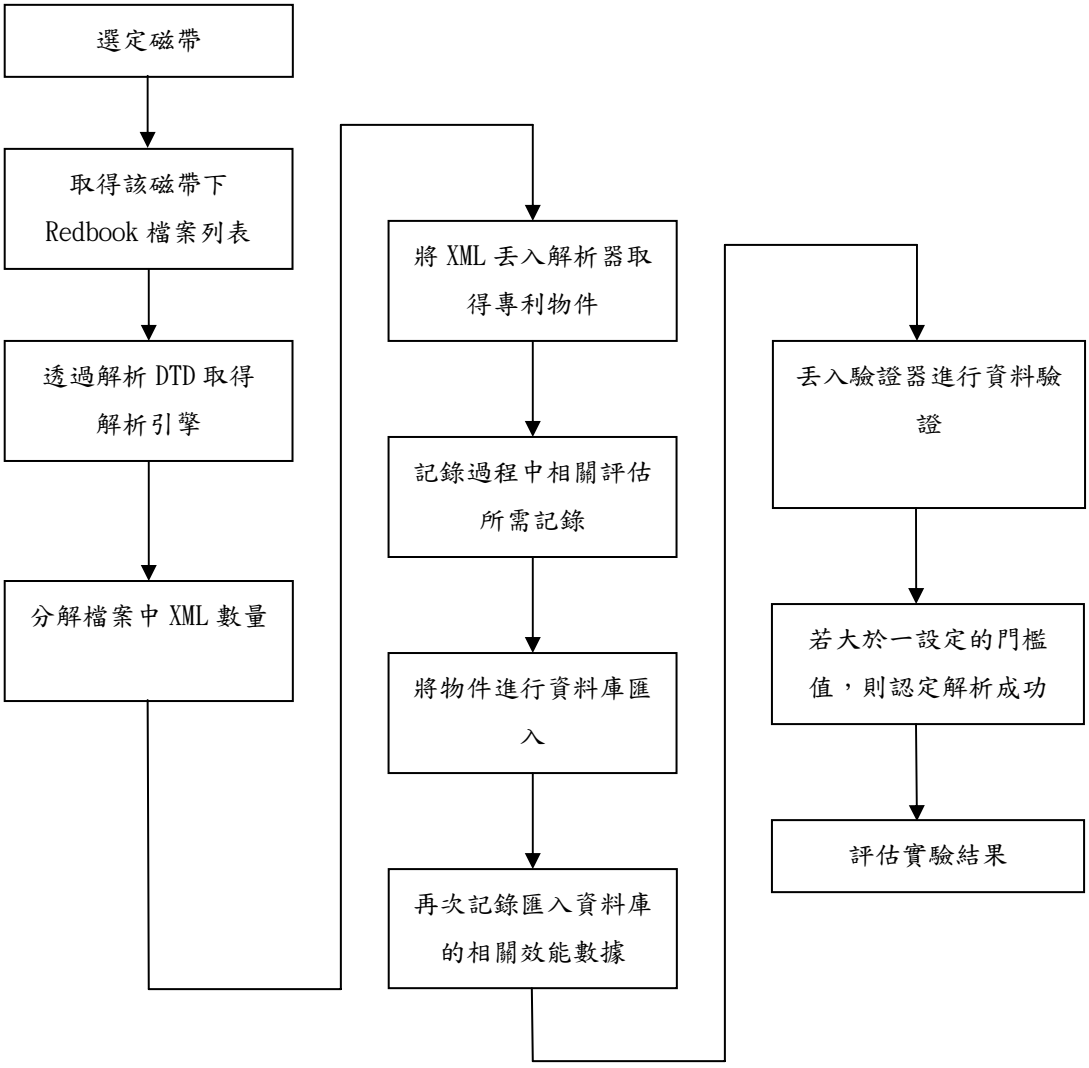

圖 3 實驗流程

第伍章 結論與建議

第一節 研究結論

根據去年的研究結論中，我們已經產出新版的 XML 格式 parser。讓程式可以自動判斷所解析的 XML 的 Redbook 版本。再載入對應的 parser 進行資料解析。不過專利發明人、專利申請人一樣必須透過同樣的統一程序來處理。將採用同樣的方式，在處理完所有的專利資料後再一次性校正。針對專利家族及財報資料庫的部分，可以透過網際網路直接擷取資料。並且將資料儲存於資料庫當中。可以簡化資料更新的步驟。