2002 出版與圖書館研討會，台北淡水，頁 14-29，民 91 年 4 月 26 日

基於 XML 之新聞管理與出版系統設計

林信成
淡江大學資訊與圖書館學系

陳勇任
淡江大學教育科技系

楊翔淳
淡江大學資訊與圖書館學系

摘要

本研究以電子新聞的管理與出版為例，藉由自訂的 Metadata 格式，以 XML 語法進行實驗性新聞資料庫之全文標誌，並自行設計一套新聞管理與出版系統，實際在 Web 環境中整合 XML 技術，探討與印證 XML 在電子文件管理與出版方面的優勢。此系統之特色為各個管理與出版模組皆以 XML 為基礎，系統內之所有資料亦採用 XML 格式，相較於傳統的資料處理模式來說，有著更彈性與更易加值處理的特點！再者，藉由 XML 優越的結構化與自我描述性，使得電子文件的「智慧化」程度得以提升，進而增進資訊檢索之精確度。

壹、前言

在資訊充斥的今日社會，人們所感困擾的不是資訊的匱乏，而是資訊的氾濫！如今，電子文件已無所不在，我們不禁要問：當大量電子文件透過網路出版、傳播之後，使用者如何在浩瀚的文件庫中，找到所要的資訊？國家資訊基礎建設所標榜的終極目標，既然是要使任何人能在任何時間、任何地點，皆能透過網路獲得所需的任何資訊或服務[1]，那麼，提供使用者一個有效的檢索機制，便成為電子出版所應考慮的重要課題之一。於是，近幾年來便逐漸形成了兩個蓬勃發展的學派：其一為強調系統智慧化的「資訊檢索」（Information Retrieval，簡稱 IR）；其二為著重文件智慧化的「詮釋資料」（Metadata），這兩者並非互不相容而是相輔相成的，其最終目的無非是為了讓使用者能在浩瀚的電子空間中順利查找到所需資料。

資訊檢索技術歷經數十年的發展，累積了不少經驗與成果，學者 Michael Lesk 將資訊檢索技術的發展歷程，從 1945 年起以每十年為一個年代劃分，每個年代都有重要的突破與進展，是瞭解近代資訊檢索技術發展的重要文獻之一。[2]資訊檢索技術的核心原理是先藉由自動化的文件分析（Document analysis）過程，抽取出足以描述文件的特徵（Feature），再對檢索詞句進行查詢分析（Query analysis），並將其映射（Mapping）至文件空間（Document space）中進行「相似

[1] 林盈達，多媒體網路：趨勢、技術、應用，台北市：松崗，1997。

[2] Michael Lesk, "The Seven Ages of Information Retrieval", 17-June-1995, available at <http://www.ifla.org/VI/5/op/udtop5/udtop5.htm>

度」（Similarity）比對。以此觀之，在文件分析過程中所抽取出的文件特徵，是不是具有足夠的代表性而能充分描述整份文件，對於整個檢索系統的效能有決定性的影響。

一般而言，回現率（Recall Rate）和精確率（Precision Rate）是評估檢索效能的兩大重要指標[3]，兩者經常是無法兼得的。目前，網路上許多採用全文檢索技術的搜尋引擎動輒千百篇的檢索結果，回現率有餘而精確率不足，往往造成使用者資訊需求上的額外負擔；反之，若僅顧及精確率則又往往犧牲回現率。實際上，文件中的結構化資訊，經常是特徵抽取時非常重要的指標，資訊檢索系統在進行文件分析或檢索時，通常可以藉由文件中結構化資訊的輔助，簡化分析過程或提高檢索效能。

因此，加強文件的結構性，增加描述性資料，對於簡化文件分析過程，提昇檢索精確率有極大的幫助。以此觀之，在發展全文資訊檢索技術之外，加強 Metadata[4]的著錄，不失為另一個提昇電子文件檢索精確率的有效方案。Metadata 是個極為普遍的概念，在我們的日常生活中，四處可見 Metadata 的蹤影：例如我們可以用{CPU 型號，記憶體大小，硬碟機容量 …}這一組 Metadata 來描述每部個人電腦的規格，所以我們可以很清楚的知道配備為{Pentium-II 350，64MB RAM，6.4GB HDD…}的 A 電腦，比起配備為{Pentium 133，16MB RAM，1.5GB HDD…}的 B 電腦來得高檔許多；而對於出版品資料則可用{書名，作者，出版社 … }這樣的 Metadata 來加以描述。為了讓 Metadata 發揮更大的功效，於是人們開始制訂各種 Metadata 標準以供遵循，圖書館長期以來所沿用的機讀編目格式 MARC，就是用來描述書目資料的 Metadata 標準。在網路盛行之後，為了因應既多且雜的電子文件，讓使用者都能盡快而且精確的找到所需資料，陸續被制訂出來的 Metadata 標準也就愈來愈多。制訂了 Metadata 之後，尚須付諸實現。以現今的 Web 發展而言，HTML 仍是發行電子文件的標準規格，然而 HTML 標籤著重於文件之版面編排與外觀格式，只有極少關於文件結構之描述者（如 <HEAD>、<META>、<BODY>...等），加以 HTML 並不具備可擴展性，使得雖然可以使用 <META> 標籤在 HTML 文件中著錄 Metadata，但仍不夠理想。

XML 的誕生正好提供了一個可行的解決方案，為 Metadata 的實作提供了一個基礎平台。由於 XML 具有可擴展性、高度結構化和良好的資料組織能力，能夠有效的表達網路上各種知識，為資料的組織、整理與交換提供良好的機制[5]，因此，從 1996 年提出工作草案、1998 年發佈正式建議規格以來，已陸續有許多相關的研究與成果被發表，使其相關文獻逐年大幅增長！我們以圖家圖書館之中華民國期刊論文索引與全國博碩士論文資訊網，搜尋在標題、關鍵字及摘要中有關於 XML 之文獻，從 1996 年至 2001 年間，共得 245 筆；此外，由 Academic Search Elite、Education Resources Information Center、Internet & Personal Computing Abstracts、Library Literature 等涵蓋多元之學術研究領域，包括社會科學、教育、人文、文學、基礎科學之國外資料庫，共得 XML 相關文獻 4914 筆。圖 1是將以

[3] Robert R. Korfhage, Information Storage and Retrieval, Wiley Computer Publishing, New York, 1997, pp. 196-199。

[4] Metadata 的含意是「用來描述資料的資料」（Data describes other data）或「關於資料的資料」（Data about data），其中文譯名有「元資料」、「描述資料」、「詮釋資料」、「後設資料」、「超資料」… 等，並不統一。

[5] 林信成、陳勇任，「基於 XML 之網際網路資料交換雛形系統設計」，教育資料與圖書館學 39 卷第 2 期（民國 90 年 12 月），頁 145-160。

上所查找到的文獻資料以年度區分，自 1996 年至 2001 年為止的文獻成長統計圖。由圖中可以明顯的看出，至 2001 年為止的文獻量，比起 1998 年 XML 正式建議規格定案時整整成長了一倍。

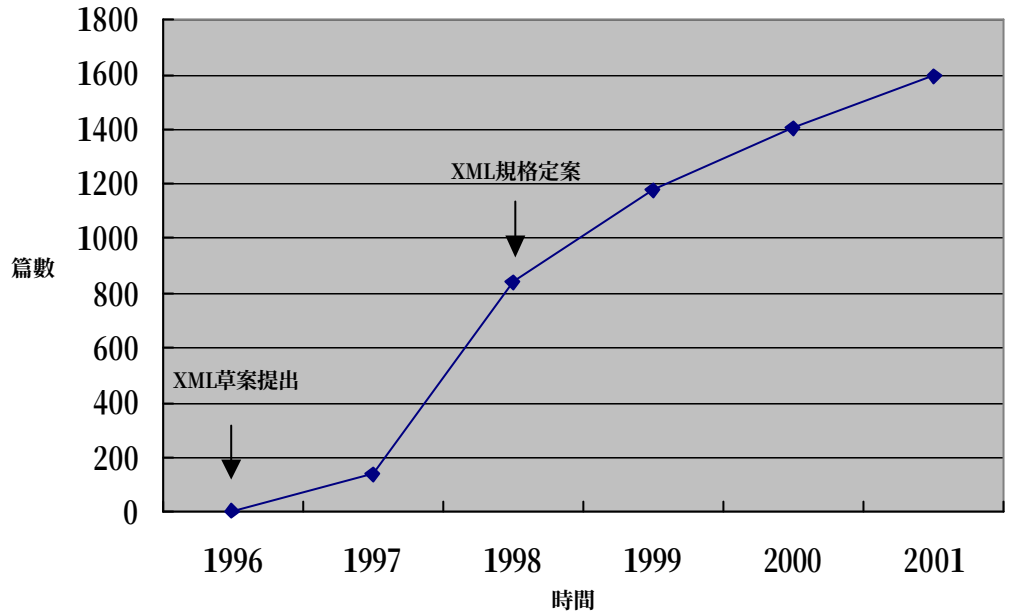

圖1 XML 文獻數量成長趨勢

本研究以電子新聞的管理與出版為例，藉由自訂的 Metadata 格式，以 XML 語法進行實驗性新聞資料庫之全文標誌，並自行設計一套新聞管理與出版系統，實際在 Web 環境中整合 XML 技術，探討與印證 XML 在資料管理與出版的優勢。此系統之特色為各個進行資料處理及交換的模組皆以 XML 為基礎，系統內之所有資料亦採用 XML 格式，相較於傳統的資料處理模式來說，有著更彈性與更易加值處理的特點！全文架構如下：第二節為 XML 與 Metadata 的介紹；第三節探討以 XML 為基礎之新聞 Metadata；第四節為基於 XML 之新聞管理與出版系統之架構規劃、功能說明、系統建置與實驗結果；最後第五節為結論與建議。

貳、Metadata 與 XML

如前所述，資訊檢索系統的回現率和精確率經常無法兼得，因此在文件分析過程中所抽取出的文件特徵，是不是具有足夠的代表性而能充分描述整份文件，對於檢索效能有著決定性的影響。以此觀之，在電子文件的產出過程當中，若能加強文件的結構性，增加 Metadata 的著錄，實不失為提昇檢索精確率的有效方案。而 XML 的適時誕生正好為 Metadata 的實作提供了一個基礎平台，本節將就 XML 技術與 Metadata 之發展作一說明，並闡述兩者之關係。

（一）XML 與 DTD

(A) XML 起源

XML 是 W3C 於 1998 年提出的電子文件規範，使用標誌語言方式對電子文件進行結構化之組織與整理。W3C 全名 World Wide Web Consortium（全球資訊

網協會），成立於 1994 年 10 月[6]，為一國際共同認可的非營利組織。W3C 之成員涵蓋世界各國，目前已擁有超過 400 個不同單位組織之會員，藉由參與會員之努力，W3C 擬定了諸多全球資訊網的公共標準（例如：HTML、XML、CSS 等），因而大幅提昇全球資訊網之互通性（Interoperability），帶動 WWW 世界之迅速發展。[7]W3C 在制定規格時，一般會歷經工作草案（Working Draft）、最後工作草案（Last Call Working Draft）、候選規格（Candidate Recommendation）、待審規格（Proposed Recommendation）建議規格（Recommendation）等階段[8]。有些規格會跳過某些階段，有些則來回審定好幾次，其間的流程與關係如下圖所示。

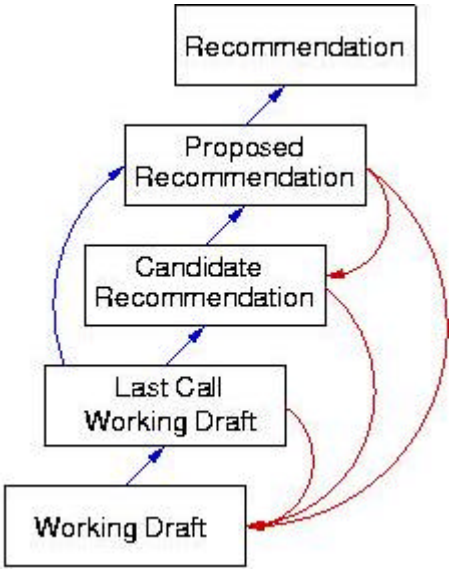

圖2 W3C 制定規格流程[9]

XML 當初在制定時並無最後工作草案及候選規格兩階段，其制定歷程時間如下表所示。

表1 XML 制定歷程

階段	制定時間
工作草案	1996 年 11 月 14 日
待審規格	1997 年 12 月 8 日
建議規格	1998 年 2 月 10 日 2000 年 10 月 2 日（Second Edition）

資料來源：W3C, “Extensible Markup Language (XML),” <http://www.w3.org/XML/>.

(B) DTD 起源

1960 年代，為了便於結構化文件的交換與處理，開始有人倡議標準的文件

[6] "About the World Wide Web Consortium (W3C)", available at <http://www.w3.org/Consortium/#background>.
[7] "XML 小百科：W3C", available at <http://www.xml.org.tw/Function/Fglossary1.asp?key=W3C>.
[8] "Technical Reports", available at <http://web3.w3.org/Consortium/Process-20010719/tr>.
[9] 同註8。

格式。最早是 GCA （Graphic Communication Association ）發展的 GenCode，提供了一套通用的排版碼以利不同的排版資料在不同的廠商間傳送；以及 IBM 所發展的 GML （Generalized Markup Language ），來解決內部大量文件的出版問題，包含了各種類型文件的管理，例如使用手冊、新聞稿、法律合約與專案規格書。1980 年代早期，GenCode 與 GML 這兩個團體的代表加入了美國國家標準局（American National Standards Institute，簡稱 ANSI ）的 Computer Languages for theProcessing of Text 委員會，目標是制定一套標準來定義及使用文件的標示，SGML 主要就是由這個委員會所發展。1986 年，SGML 變成 ISO 的國際標準，編號是 ISO 8879:1986。[10]

SGML 乃針對文件交換和處理而訂，我國也已採納為國家標準「標準通用標示語言」。SGML 有嚴謹的語法，能由機器自動剖析，主要是用來定義 DTD。因此 DTD 也具備嚴謹的語法，提供標籤（tag）、屬性（attribute）及其他語法單元（syntactic unit）來標示文件的結構和內容屬性。[11]

由上述可知，DTD 早在 1986 年就已因 SGML 的制定而存在，但由於 SGML 太過龐大複雜、不易學習及使用，因此並未獲得普及[12]，間接使得 DTD 被人所遺忘。直至 1996 年 W3C 所提的 XML 工作草案[13]中，才又見到 DTD 的身影。

(C) XML 文件類型

據 XML 建議規格[14]的描述，XML 文件依其結構可分為兩類：完構的（Well-Formed）XML 文件和有效的（Valid）XML 文件。

一個完構的 XML 文件必須遵守下列規則：

1. 包含一個以上的元素（elements）。
2. 僅有一個根元素（root）。
3. 所有的元素都有起始標籤與結束標籤。
4. 所有的標籤必需呈巢狀（nest）結構。
5. 空標籤（empty tags）必須導守特殊的 XML 語法。

圖 3 所示即為一個 Well-Formed 的 XML 文件範例，其標籤集可以任意制訂。

```
<?xml version="1.0" encoding="big5" standalone="yes"?>
<淡江大學>
    <資圖所>
        <學號>689070034</學號>
        <姓名>楊翔淳</姓名>
        <級別>二</級別>
    </資圖所>
```

[10] 陳嵩榮，「SGML、XML、RDF 文件標準比較與 Metadata 資料模式設計」(碩士論文，輔仁大學圖書資訊學研究所，民國 88 年)，頁 13。

[11] 朱四明，國內電子公文推展策略研究：SGML 應用的實例，初版（臺北市：松崗，民國 87 年），頁 10。

[12] 林信成，「XML 相關技術與下一代 Web 出版趨勢之研究」，教育資料與圖書館學 37 卷第 2 期（民國 88 年 12 月），頁 184-210。

[13] "Extensible Markup Language (XML)", available at <http://www.w3.org/TR/WD-xml-961114.html >.

[14] "Extensible Markup Language (XML) 1.0 (Second Edition)", available at <http://www.w3.org/TR/2000/REC-xml-20001006#sec-documents>.

```
</淡江大學>
```

圖3 Well-Formed 的 XML 文件

而所謂的 Valid XML 是指其不但符合 Well-Formed 標準，並且在 XML 文件中的元素（Element）、屬性（Attribute）以及其他單元（如 Entity、Notation … 等），皆需符合相對應的 DTD 規範。例如，以圖 3 的文件為例，我們可以制定其 DTD 規範如圖 4所示，那麼所有引用此份 DTD 的 XML 文件，其標籤出現的順序與次數便被侷限於此份 DTD 規範中。

```
<?xml version="1.0" encoding="Big5"?>
<!ELEMENT 淡江大學 (資圖所+)>
<!ELEMENT 資圖所 (學號, 姓名?, 級別, 專長*)>
<!ELEMENT 學號 (#PCDATA)>
<!ELEMENT 姓名 (#PCDATA)>
<!ELEMENT 級別 (#PCDATA)>
<!ELEMENT 專長 (#PCDATA)>
```

圖4 DTD 範例

在此例中，淡江大學為 XML 文件的根元素，而之後的資圖所、學號、姓名、級別等，就是其下層的元素名稱；此外各元素後的+、?、* 則代表該元素出現的次數，其含意如下：

符號	出現次數
+	大於等於 1 次。
?	0 或 1 次。
*	大於等於 0 次。
沒有符號	1 次。

表2 元素出現次數及其代表符號

在 XML 文件中，引用特定 DTD 並符合其規範，即是 Valid XML 文件。如圖 5 即是將原來圖 3 的 Well-Formed XML 文件，加入了圖 4的 DTD 引用之後，成為一份 Valid XML 文件的例子。

```
<?xml version="1.0" encoding="Big5"?>
<!DOCTYPE 淡江大學 SYSTEM "student.dtd">
<淡江大學>
    <資圖所>
        <學號>689070034</學號>
        <姓名>楊翔淳</姓名>
        <級別>二</級別>
    </資圖所>
</淡江大學>
```

圖5 Valid XML 文件

（二）Metadata 的發展

Metadata 在資訊組織界最普遍的解釋是"data about data"，意指有關資料的資

料，即資料之描述性資訊，如圖書館的 MARC 記錄，即為一種 Metadata，[15] 而 XML 與生俱來的結構化及自我描述特性遂成為實作 Metadata 的跨平台語言。

(A) 國內發展

有關 Metadata 的研究在台灣起步甚早，相關研究單位及計畫包括各圖書資訊學系、中央研究院、國科會各項數位博物館計畫、台灣省公共圖書館的地方文獻數位計畫、國家檔案局等[16]，茲將其整理如下表。

表3 國內 Metadata 發展現況

研究單位與計畫	研究成果
數位博物館專案故宮文物之美計畫 Metadata 小組	1. MICI－DC for NPM 2. 人名權威檔及其 DTD 3. 主題權威檔及其 DTD 4. 地名權威檔及其 DTD 5. 時代權威檔及其 DTD 6. 文獻 MICI－DC、DTD 及著錄範例 7. 書畫 MICI－DC、DTD 及著錄範例 8. 展覽 MICI－DC、DTD 及著錄範例 9. 器物 MICI－DC、DTD 及著錄範例 10. 參考書目 MICI－DC、DTD 及著錄範例
中央研究院史語所博斯年圖書館	1. 傅斯年圖書館拓片詮釋資料及著錄範例（草稿） 2. 傅斯年圖書館善本書詮釋資料及著錄範例（草稿） 3. 傅斯年圖書館善本書明人文集詮釋資料及著錄範例（草稿）
機關檔案管理系統研究小組	1. 檔案詮釋資料格式 2. 以都柏林核心集為基礎的檔案詮釋資料格式
國立自然科學博物館、暨南大學、數位博物館計畫資訊組織與檢索規範研究小組	1. 蝴蝶生態面面觀 2. 蘭嶼生物／文化多樣性數位博物館之詮釋資料
數位博物館計畫資訊組織與檢索規範研究小組	淡水河溯源
中央研究院數位博物館專案平埔文化計畫設計	平埔文化
中央研究院數位博物館專案人文與自然資源地圖計畫設計	人文與自然資源地圖詮釋資料
內政部資訊中心設計	國土資訊系統相關數值資訊詮釋資料規格及範例

[15] 陳昭珍，電子圖書館整合檢索之理論與實作，初版，（臺北市：文華，民國 89 年），頁 83。
[16] 國家圖書館 Metadata 研究小組。中文詮釋資料(Metadata)格式彙編。臺北市：國家圖書館，民國 89 年。

公共圖書館自動化及網路諮詢員會地方文獻數位化小組及宜蘭文化局	戲曲唱片詮釋資料及 DTD
公共圖書館自動化及網路諮詢員會地方文獻數位化小組及高雄縣文化局	皮偶劇本詮釋資料及 DTD
公共圖書館自動化及網路諮詢員會地方文獻數位化小組及台中縣文化局	地方發展古照片詮釋資料及 DTD
公共圖書館自動化及網路諮詢員會地方文獻數位化小組及新竹縣文化局	客家文物詮釋資料及 DTD
國家圖書館 Metadata 研究小組	古籍善本詮釋資料及著錄範例
交通大學圖書館	楊英風數位美術館詮釋資料

資料來源：國家圖書館 Metadata 研究小組。中文詮釋資料(Metadata)格式彙編。臺北市：國家圖書館，民國 89 年。

由上表可知目前國內所提出的 Metadata 格式，絕大多數皆採用 XML 的 DTD 語法[17]，並搭配 XML 使用，其範圍大部分為文史及圖書館相關領域。

(B) 國外發展

OASIS 為國際上一個提供資料和資訊內容交換技術的重要組織，為一國際性非營利公協組織。在角色扮演上，OASIS 與國際相關標準組織(如 W3C,NIST,UNCEFACT 等)的功能具有彼此互補相乘的效果。換言之，OASIS 專注於結構化資訊標準的推廣和導入，透過各會員彼此之間的技術交流及資訊分享，收集思廣益之效，藉以確實反應市場運作機制於結構化資訊標準應用面的真實需求。[18]

為了促進資訊系統結構發展，OASIS 在 1999 年六月成立一非營利性組織 www.XML.org 其目的是使 XML 標準重疊的部份減到最少，並提供公開的 XML 資訊。[19]在 XML.org 中，收錄了各行各業所用或制定中的 Metadata 資料，至本文截稿為止共 49 類合計有 941 種之多，茲將其整理如下。[20]

表4 OASIS 收錄之國外 Metadata 發展現況

分類	數量	分類	數量
會計（Accounting）	5	地理（Geography）	2
廣告（Advertising）	6	健康（Healthcare）	15
航空（Aerospace）	24	人力資源（Human Resources）	23
農業（Agriculture）	4	保險（Insurance）	6
藝術（Arts/Entertainment）	26	網際網路（Internet/Web）	21
天文（Astronomy）	15	法律（Legal）	10
汽車（Automotive）	12	文學（Literature）	14
銀行（Banking）	10	製造業（Manufacturing）	3

[17] "Extensible Markup Language (XML) 1.0", available at <http://www.w3.org/TR/1998/REC-xml-19980210>.

[18] "XML 小百科：OASIS", available at <http://www.xml.org.tw/Function/Fglossary1.asp?key=Oasis>.

[19] "About XML.org", available at <http://www.xml.org/xml/aboutxml.shtml>.

[20] "Vertical Industry Directory", available at <http://www.xml.org/xml/industry_industrysectors.jsp>.

生物（Biology）	4	數學（Math/Data）	12
商業服務（Business Services）	3	多媒體（Multimedia）	24
化學（Chemistry）	2	新聞（News）	10
電腦（Computer）	5	其他行業（Other Industry）	4
建築（Construction）	9	公益服務（Public Service）	3
顧客資訊（Customer Relation）	6	出版業（Publishing/Print）	26
關稅（Customs）	2	不動產（Real Estate）	15
資料庫（Databases）	7	宗教（Religion）	2
電子商務（E-Commerce）	55	人工智慧（Robotics/AI）	5
電子資料交換（EDI）	19	科學（Science）	63
企業資源規劃（ERP）	3	電腦軟體（Software）	73
經濟（Economics）	2	供應鏈（Supply Chain）	24
教育（Education）	44	電信（Telecommunications）	24
公共事業（Energy/Utilities）	22	運輸（Transportation）	7
環境（Environmental）	1	旅行（Travel）	4
金融服務（Financial Service）	56	XML 技術（XML Technologies）	206
食品服務（Food Services）	3		

資料來源：XML.org,"Vertical Industry Directory," <http://www.xml.org/xml/industry_industrysectors.jsp>.

參、基於 XML 之新聞資料庫 Metadata 探討

為了驗證經由 XML 註錄 Metadata 之電子文件，能有效提升檢索系統之精確度，本研究以一實驗性質之新聞資料庫為例，除對每篇新聞內容以 XML 註錄相關之 Metadata 外，並在其全文中以 XML 加註描述性標籤，此外，藉由自行設計的新聞管理與出版系統，測試加註了這些描述性資料之後的檢索結果。但新聞 Metadata 的標準化並非本文之重點，本研究之核心在於設計一套以 XML 為基礎之新聞管理與出版系統，因此，為了使系統開發及實驗能順利進行，本研究為該系統制訂了一個簡易的新聞資料庫 DTD，作為新聞資料之組織、檢索、出版之基礎。

首先，本節先就新聞資料庫之 Metadata 進行探討，下一節再針對整體系統架構及實驗結果詳加說明。在新聞資料庫的 Metadata 方面，國內目前只有政大新聞系謝瀛春教授發表過有關科學新聞的內容標誌[21]，國外則是以 NITF（News Industry Text Format）[22]與 XMLNews[23]為兩大主流。

（一）科學新聞內容標誌

此內容標誌依據新聞學及新聞寫作的相關原理，運用 XML 標記新聞內容，並以純淨新聞體裁之科學新聞為樣本，進行標誌工作。[24]在新聞事件（event）的陳述上，以人（who）、事（what）、時（when）、地（where）、如何（how）、為何（why）六大方向來描述，將新聞內容以此為表達的重點，讓使用者清楚的明

[21] 謝瀛春、黃學碩、維習安、雷約翰、謝清俊，「新聞內容的標誌-XML 之應用」，海峽兩岸資料庫/數據庫與資訊/信息服務交流與合作論文集（民國 90 年 1 月），頁 205-212。

[22] "NITF News Industry Text Format", available at <http://www.nitf.org/site/index.html>.

[23] "XMLNews.org", available at<http://www.xmlnews.org/>.

[24] 同註21。

自發生的新聞事件始末。[25]

（二）XMLNews

XMLNews 分成兩部份：XMLNews-story 定義新聞的內容，XMLNews-meta 則敘述新聞稿件。XMLNews-story 是借用另一個 XML 規格 NITF 而來的，事實上是 NITF 的 Subset。[26] NITF 是在 1998 由許多新聞機構及美國報業協會，共同研製的一個 XML 標準規格。NITF 的訂定是要取代以列印為主的 ANPA 1312 老舊規格[27]，但由於 NITF 太龐大，許多 Tag 大多數人都不會用到，並且不夠靈活，要加一個 Tag 都不容易，所以才會有 XMLNews 的出現。

XMLNews-meta，是依據 W3C 標準的規則 RDF（Resource Description Framework）[28]制訂的，RDF 是以物件導向的理念，用 XML 作規範，制訂出來一套描述資料的規則。XMLNews-meta 是描述新聞稿件的規格，其 DTD[29]中的 Element 有新聞機構代碼與名稱、版權、使用權、作者、新聞類別、產業名稱、語言名稱與代碼、人物名稱、發稿時間等多項。[30]

（三）自訂的新聞 Metadata

Metadata 與資料間存在互動的關係，因為資料實體是藉由 Metadata 描述而成，由於 Metadata 主要目的在於描述資源的屬性、特徵。就本研究架構而言，是依據 DTD 來描述系統之綱要（schema），再藉由系統綱要控制與管理 Metadata。[31]而在 Metadata 的部分則是以 XML 格式來做描述，並遵循自定之系統 DTD，在符合 Valid 的文件類型前提之下，完整呈現資料內容。

在參考國內外的新聞 Metadata，並考慮本系統之實際需求乃在於提升內文檢索的精確度之後，我們提出一個自行制訂的簡易版新聞 Metadata DTD，其中，每則新聞除了包含諸如新聞編號、索引、日期、標題、作者 … 等基本資料外，另將新聞內容分以人、事、時、地、物等加以標誌，以便進行內文語意搜尋之用，最後，另含一個排版用的標籤作為不同樣式之套版。

肆、基於 XML 之新聞管理與出版系統設計

（一）系統架構與功能說明

本研究所規劃並擬完成之系統，其介面共分為前端使用者介面與後端管理者介面，如圖 6 所示，若依功能劃分則可區分為四大功能模組，分別是：

1. 資料出版模組（Data Publishing Module）
2. 資料檢索模組（Data Searching Module）
3. 資料編輯模組（Data Editing Module）
4. 資料管理模組（Data Management Module）

---

[25] 同註21。

[26] "XMLNews.org XMLNews Specifications", available at<http://www.xmlnews.org/XMLNews/>.

[27] "nitf introduction", available at <http://www.nitf.org/site/intro.html>.

[28] "Resource Description Framework (RDF) - W3C Semantic Web Activity", available at<http://www.w3.org/RDF/>.

[29] "XMLNews-Meta Documentation", available at <http://www.xmlnews.org/docs/xmlnews-meta.html>.

[30] "XML 傳遞新聞稿", available at<http://www.brainnew.com.tw/Article/na1999/F_042099.htm>

[31] 余顯強，「以 XML 框架設計之 Metadata 系統」，書藝第 37 期（民國 90 年 6 月），頁 41。

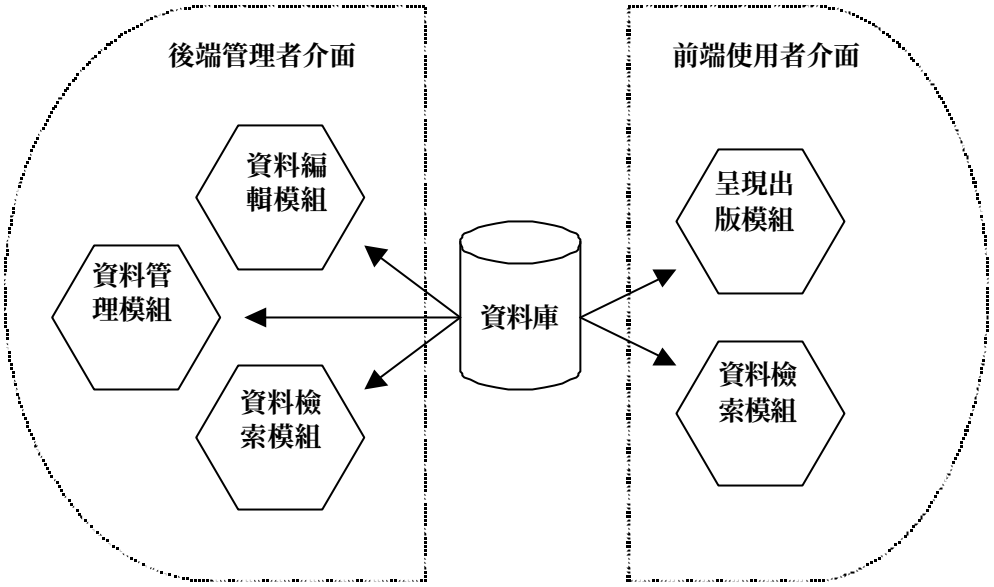

圖6 基於 XML 之新聞出版系統功能模組

上述四大模組彼此分工合作，其主要功能說明如下：

(A) 資料出版模組

本模組負責將資料內容呈現給使用者，並且具備解讀 XML 文件的功能。此模組透過 DSO[32]以及 DOM[33]來解讀 XML 文件，並結合該資訊所需之相關功能及超連結，加以包裝、排版，只要是符合系統之 DTD 規範的 XML 文件，皆可透過此模組呈現內容。

(B) 資料檢索模組

本模組提供使用者檢索所需新聞資料。一般檢索模組僅提供欄位的檢索功能，並未提供針對內容某些特定目標加以檢索，如人名、地名等，本模組除提供新聞類別、關鍵字詞檢索功能之外，藉由 XML 將文件結構化的特性，可針對新聞內容的人、事、時、地、物加以檢索，提高檢索結果的查準率。

(C) 資料編輯模組

本模組提供管理者編輯新聞文件內容。對於 XML 並不了解或不熟悉者，皆可透過此模組，輕易的將所需之新聞內容編輯成符合系統之 DTD 規範的 XML 文件，並針對新聞內容給予人、事、時、地、物不同的標記，提供檢索模組使用，另外經由 XML 文件內容與呈現資料分離的特點，同一份文件可選擇不同樣式做為出版的選擇。

(D) 資料管理模組

本模組提供管理者異動／修改資料。透過網路可遠端開啟管理模組，對所需的新聞資料做新增、修改與刪除等動作，另外，對於異動過的資料，藉由資料管理模組即可做查詢的動作，檢視其 XML 內容是否可正確呈現，無需回到一般使

---

[32] DSO(Data Source Object，資料來源物件)可將 XML 文件視為一份文件資料庫進行資料存取的動作。

[33] DOM(Document Object Module，文件物件模型)為一 W3C 所制定的介面標準。DOM 並非針對 XML 量身訂做的，而是一套普遍適用於 HTML、XML 等文件的應用程式介面(Application Programming Interface, API)。

用者介面。

（二）系統建置與實驗結果

本系統建構於 Microsoft Windows 2000 Advance Server（NT 技術平台）之上：Web 伺服器為 Microsoft IIS，中介軟體及各大模組使用 ASP（Active Server Pages）語言開發，至於後台資料庫系統則採用 Microsoft SQL Server。使用者端則可以使用支援 XML 之瀏覽器（如 Microsoft Internet Explorer 5.0 以上版本之瀏覽器），進行連線。[34]

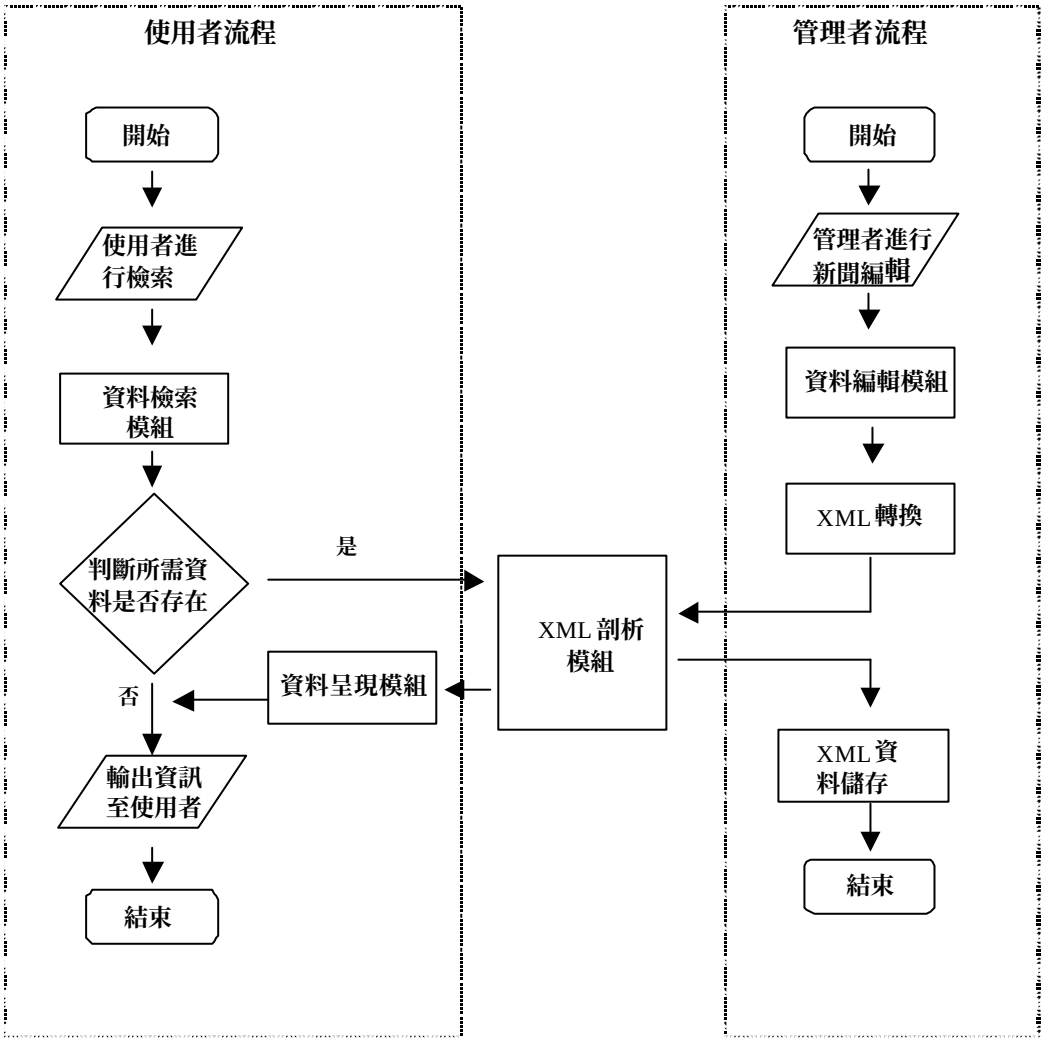

圖7 系統運作流程圖

此系統的運作流程如圖 7 所示，分為使用者與管理者兩種流程。使用者可經由資料檢索模組，輸入欲查詢之條件，如圖 8 所示，接著資料檢索模組至資料庫中搜尋所需資料。檢索模組的工作完成後，由資料庫所取出之資料將交由資料出版模組，呈現出版模組將依照系統所訂定之 DTD 來確認 XML 資料的合法性，接著驗證後的 XML 資料將呈現給使用者。呈現的 XML 資料透過 DSO 以及 DOM 的解讀，可在不更改原始新聞內容之下，給予各種排版樣式，如圖 9所示。

[34] 本系統目前架設於http://163.13.176.7/xp。

圖8 資料檢索模組

圖9 各種排版樣式

由於本系統之新聞資料已經事先加值處理，著錄了 XML 標誌，故使用者在檢索本系統時，除可依據標題、作者、時間 … 等簡要資料進行檢索外，也可針對新聞內容，依據人、事、時、地、物等限制條件進行更精確的全文檢索，例如圖 10 乃是以「大學」作為關鍵字詞，選擇以不限標誌的方式進行內文檢索之結果，共找到八篇文章；對於同樣的檢索詞，如果將檢索條件限制於「地」，表示使用者想要查詢的只是與大學有關的地方、地名或地點，而非所有與「大學」概念相關的文章，結果如圖 11 所示，找到七篇文章；再者，若檢索詞仍為「大學」，

但將檢索條件限制於「事」，表示使用者想要查詢的是有關大學的事件而非地點，則如圖 12所示，則更精確的找到兩篇含有大學相關事件的文章。

圖10 不限檢索標誌之檢索結果

圖11 限定檢索標誌為「地」之檢索結果

圖12 限定檢索標誌為「事」之檢索結果

在管理者方面，可透過資料編輯模組進行新聞資料的新增，其內容經由 XML 轉換及剖析之後，儲存至資料庫，並可透過資料管理模組，對新聞內容進行修改、刪改的動作。此外，本系統藉由 XML 將新聞內容結構化，將其人、事、時、地、物等字詞分析並標記，如圖 13所示。

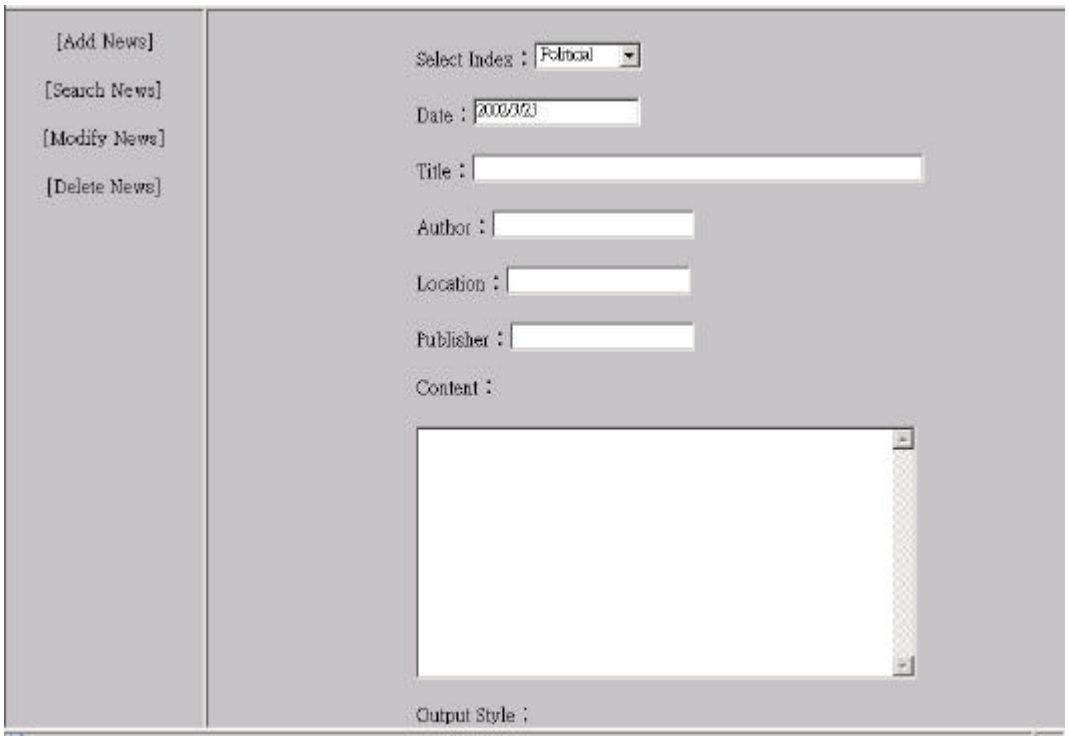

圖13 資料管理模組

伍、結論與建議

為因應網路普及後資料的快速成長，新一代資訊檢索系統，特別是全文式查詢系統，必須更有效率與精準的提供服務。XML 支援語言中立（language neutral）的定義和平台中立（platform neutral），並且能提供定義在 Web 環境上結構化文件交換的格式。XML 允許使用者自行定義所需的標示語言。因此，可以將資料內容以清楚的標籤表現其意義，並可廣泛地應用在各種領域。[35]就資訊檢索方面而言，由於 XML 資料具有自我描述性質，因此可以提供語意層次的搜尋，避免全面性的盲目搜索，進而提昇檢索結果的精確度，這在網路文件氾濫成災的今日尤其重要。對於系統的開發而言，由於 XML 具備可擴展性、資料與樣式分離等特色，各個系統可根據自身的需求，對 XML 資料進行其他加值處理，這使得 Web 應用程式（Web Application）的發展更具彈性。[36]

本研究透過新聞管理與出版系統的實作，將各資料模組以 XML 為基礎，系統內所有資料亦採用 XML 格式，以 XML 將新聞作結構化的處理，並自訂 Metadata 來描述其內容，搭配資料檢索模組，可確實針對新聞內容作精確檢索，其精確度優於傳統的全文檢索結果。此外，由於 XML 資料與樣式分離的特性，使得新聞呈現的樣式非常有彈性，可因不同的使用者需求做更改，而不用更動原

[35] 同註31，頁 40。
[36] 同註12，頁 195。

始的新聞資料內容。以奇摩新聞為例，其合作的對象有聯合新聞網、中央社、中時電子報、台灣日報及路透社等國內外新聞媒體[37]，發佈其即時與每日新聞，在這樣的一個合作關係之下，每家媒體有著不同的新聞格式與其排版方式，一但需要發佈在同一網站之上，必定另外制定合作的標準，才能解決彼此間不相容的問題；若運用本新聞管理與出版系統，只需遵守 DTD 的規範，同一份新聞內容，就可依照不同的需求，迅速方便的在網站上出版，而無需另外做內容上的更改。

網路出版是必然的發展趨勢，如何運用新的技術讓網路環境更加的有效率是大家所追求的目標，XML 的出現帶來不同的解決方式，藉由系統實作的驗證，更加深應用 XML 的信心，期望在不久的將來，XML 發展更成熟之際，能全面的應用於網路之上！

[37] "Yahoo!奇摩新聞", available at <http://tw.news.yahoo.com/>.