行政院國家科學委員會補助專題研究計畫

成果報告

建構新聞資料庫芻型：應用「可延伸標記語言」(XML)

發展財經新聞內容標記

計畫類別：■ 個別型計畫　　□ 整合型計畫
計畫編號： NSC 90 – 2412 – H – 004 – 008
執行期間： 90 年 8 月 1 日至 91 年 7 月 31 日

計畫主持人：陳百齡
計畫參與人員： 劉怡麟、謝宛蓉

成果報告類型：■ 精簡報告　　□ 完整報告
報告處理方式：■ 一年後可公開　□ 二年後可公開查詢

執行單位：國立政治大學新聞系

中 華 民 國 92 年 1 月 3 日

建構新聞資料庫芻型：

應用「可延伸標記語言」(XML) 發展財經新聞內容標記

陳百齡

「資料類型定義」表徵數位內容的領域知識，也是建構資料庫的基礎。本報告內容旨在探究財經新聞報導特徵，分析報導內容、並定義資料結構和元素，以作為註記內容的基礎。研究成果不僅用於定義財經新聞資料，以發展 XML 資料庫；本研究之經驗，也將可作為建立各類新聞資料類型定義的借鏡。

關鍵字：可延伸標記語言、資料類型定義、新聞資料庫、財經新聞報導、

Data type definition (DTD) represents domain knowledge, and is a building block of XML-based database. This technical report demonstrates how we a set of data type definition for text-based financial news in Chinese newspapers. The researchers reviewed the literature in news writing, abstracted the critical characteristics from a group of samples of texts, and created the definition. Finally, a formative evaluation was performed to revise the issued DTD.

Keyword：database for news stories, financial news, eXtensible Markup Language (XML), Data type definition (DTD),

壹、前言

新聞產業的內容產製日益倚賴新聞資料庫。然而，新聞資料要能夠做橫跨平台的交換、檢索、以及加值應用，必須具備一套規格標準。在 1998 年問世的「可延伸標誌語言」（eXtensible Markup Language, XML）是一套「後設語言」（meta language）[1]，用於描述資料型態和內容，使數位化的新聞資料更趨於結構化和標準化。隨著網路技術的演進，越來越多的資料庫採用這套標記語言作為處理跨平台新聞資料的機制，以進行產業之間資訊交換、檢索、以及電子商務應用。[2]

[1] 所謂「後設語言」，在資訊科學領域裡通指一套用來描述和分析一般人類語言的符號體系。這套符號體系允許使用者自行發展文件類型定義，以描述資料內容的結構或內容（Simon, 2000）。通常後設語言都利用標誌（markup），也就是在文本前後以標籤加註（tagging）的方式，作為文本傳輸、交換、和儲存的機制。

[2] 產業標準已經應用的標準，包括傳輸協定（protocol）、封裝協定（envelope）、檔頭資料

「標記」（markup）是使用一套後設語言（XML 或 SGML）透過文本類型定義進行資料註記，藉以表達文本的內容、結構、狀態和屬性等資料，俾使電腦可辨識資料內容並加值處理。目前許多新聞產業已經開始建立新聞資料的後設資料（metadata）進行標記，但「內容標記」（content markup；或稱為「語意標記」，semantic markup）規格設計則正在起步階段。目前國外產業所建立新聞內容標記規格（如 NITF），[3]一般均在後設資料的定義中，預留未來內容標記的餘地。[4]

以 XML 標記語言進行任何內容標記，先決條件是建立一套「資料類型定義」（Data Type Definition; DTD）。[5] 新聞領域的資料庫亦不例外。國內學者（如謝瀛春等，1999；黃立夫，2001）發展一般新聞之內容標記之資料定義。本報告延續這個議題，針對財經新聞報導的特定新聞類型建立資料定義。

「資料類型定義」是領域知識的表徵，也是建構 XML 為本的新聞資料庫的第一步。本報告內容旨在建立財經新聞資料類型定義，以作為標示財經新聞內容的基礎。本報告將敘述研究人員如何依據新聞的相關知識，分析財經新聞語料、並定義出財經新聞涉及的若干詞彙和關係，以作為未來資料內容標記的基礎。研究成果不僅可用於定義 XML 資料庫之財經新聞報導資料；本研究的經驗，也作為建立各類新聞資料類型定義的借鏡。

貳、問題分析

首先要解決的問題是：「新聞資料類型定義是什麼？」我們認為，新聞資料的類型定義，便是「新聞知識的表徵」。這個知識可以由一組元素透過其結構與關係而表現出來。所有後設語言（例如 SGML、XML）的「資料類型定義」正

（header）、以及新聞內容（content）的規格等（Dumbill, 2000）。

[3] News Industry Text Format（NITF）由 International Press Telecommunications Council（IPTC）所推出的一套使用 XML 描述新聞內容與結構的標記語言。新聞產業使用使用 NITF 標準描述的新聞資料可透過轉換程式（例如 XSLT）轉換為不同的資料格式、呈現在不同的平台上，例如轉換為 WML 以配合手機或 PDA。採用 NITF 標準編輯而成的新聞資料可就某幾個特定的標籤內容做資料的搜尋以提高準確度，以及透過的描述，將新聞的相關資訊呈現給讀者。

[4] 新聞產業應用 XML 主要在建立「後設資料」（metadata），一般並未納入內容標記之規則。以 NITF 而言，作法是保留「文內註記」（inline markup）項目，NITF 之內容註記項目有兩個特色，一是僅有十項標籤，相當簡略。以人、功能、組織、地點、事件、物件、時間、財貨、數目、引語等十項。

[5] 一般而言，XML 的文件結構可分為兩個部分，包括「邏輯結構與實體結構（Simon, 2000）。物理結構主要功能在於定義標誌之間的結構組織關係，由「文件類型定義」（Data Type Definition; DTD）和 XML 標籤集所構成（tag sets）。物理結構則是資料內容的部份，由 XML 文件中的元素和資料實體（entities）所構成。數位內容產製人員依照原始文件的領域知識，制訂文件類型定義（DTD），以作為進一步標誌的基礎，DTD 不但揭示文件的結構、同時也定義結構裡各個單元的內涵。DTD 產生標籤集（tag sets），以規範 XML 文件。

是要透過其語法，把特定內容（例如財經新聞）的知識結構元素以及關係，轉換成為電腦可以辨識的語言規則，從而建立電腦判讀這類資料的基礎。換言之，在建立資料類型定義之前，我們必須從領域知識中萃取出若干規則，作為未來電腦判讀資料的基礎。因此，本研究有必要說明「新聞」與「財經新聞」（特定的新聞內容類型）的表徵。

一、新聞報導

「新聞報導」（news stories）和一般敘事內容同樣含有結構、方向、重點及觀點；新聞報導作為一種公共論述，目的在將新聞工作者對於新聞事件的心象表徵傳遞給特定讀者（臧國仁／蔡琰，1998；臧國仁／施祖琪，1999）。本文則認為新聞報導是當代描述事實的一種資訊內容，在媒體與社會文化的情境下發展出特定的文體與內容。

首先，新聞報導必須以「事實」（facts）為基礎。新聞報導與一般故事不同的一個特點，在於新聞必須報導事實。當下存在各種敘事文本目的各有不同，倘若以新聞形貌出現時，則其目標在讓讀者將信文本內容為可信的事實，也正因此，新聞報導論述內容必須遵守若干語言常規（臧國仁／施祖琪，1999）。例如，新聞報導中的敘事主體較少使用第一、二人稱（例如避免「我」、「你們」，而使用「記者」、「讀者」）、引述消息來源（「某某指出」或「某某表示」）、使用中性詞語、同時也強調「擬真」（以一問一答保留採訪現場原貌）等，均是新聞作為事實論述的表現。

其次，新聞報導的內容反映新聞組織文化及其常規。新聞組織常規及語言常規分別表現在新聞報導文本類型和語言層次。新聞機構為因應每日變化的事件，故設定各種機制，使得新聞處理例行化（routinization）。新聞報導通常在稿頭中交代新聞組織、報導時間和報導者（中央社記者某某廿日台北電）、區分路線版面（立院新聞通常在要聞版）、批次處理（以截稿時間為單位，包裝處理一批新聞）、標準化新聞處理程序及修辭用語（例如，媒體採用編採手冊以規範人名、職稱、及機構等文字體例）。

第三、新聞報導是累積發展的敘事文本。新聞報導內容必須建立在先前新聞內容的基礎上。誠如 Bell（1994)）將新聞報導喻為「連續劇」，新聞報導內容固然是各自獨立的內容，但是若干新聞報導之間，又彼此有所關連。

第四、新聞報導具有特定的文本類型（genre）。新聞文體具體表現在倒金字塔式寫作格式和新聞元素。新聞報導的倒金字塔式寫作，也就是「把越是重要的話題依序排在前面」。首先，新聞報導文本結構區分為「摘要—導言—軀幹」（van Dijk, 1988）。導言摘要自內文，而標題則摘要自導言；換言之，越在新聞報導文本前段位置，則文字越為精鍊。例如標題一般不超過 20 字長度，而導言不超過

百字。其次，則是新聞元素的重複出現；新聞報導標題、導言和內文均指向同一核心事實，三者之間具有連貫性，遣詞用字雖不盡相同，但若干新聞元素（人事時地物等）卻重複出現。第三、新聞報導時間結構亦於一般敘事文本。新聞報導敘事排序，係按照事件重要程度而非時間順序，因此未必反映事件發展之時間順序。第四、缺乏結尾（ending），新聞報導雖然提供事件發生的重要情節，但是往往沒有結局（臧國仁／蔡琰，1998）。

最後、新聞報導出現的詞類頻率不一。新聞報導由若干元素組成，新聞寫作是選擇／組織／呈現新聞資訊的過程，而新聞工作者辨識／選擇／組織／呈現新聞事實的對象，新聞學教科書歸納為六項元素（即 5W1H）：「何人／何事／何時／何地／何物／何事／為何」。首先，這些新聞元素重複出現在標題、導言、內文之中，而呈現若干特徵。其次，詞類出現的頻率不同。人、事、時、地、物多以名詞呈現，但是人、時、地、物可能較容易由單一詞組表達，「何事」與「為何」則較複雜，要靠多類詞組之組合（動詞、或動詞加名詞）才能表達，因此新聞元素之辨識有難易之別，辨識「何事」與「為何」的難度高於人、時、地、物。其次，新聞報導以呈現事實為主，必須使用中性詞語寫作，在相當程度上限制代名詞、形容詞和副詞的使用，因此名詞和動詞出現頻率應高於其它詞類。

二、財經新聞報導

一般而言，指涉財經新聞相關詞彙如「財經新聞」[6]、「經濟新聞」[7]、「商情新聞」[8]（business news, financial news, etc）等。概略地來說，「商情新聞」較偏向較重視產業的動態，例如中央社商情新聞。「經濟新聞」或「財經新聞」是國內學術或一般著作較常使用的分類名稱，其中「經濟新聞」又常與過去強調經濟發展的社會脈落結合，例如：王洪鈞（1959）、楊士仁（1989）等。目前，一般以「財經新聞」為最常用的詞彙，坊間相關書籍也多採用。[9] 承上述定義，「財

[6] 方怡文／周慶祥（1999：16）認為，凡涉及財政、經濟與各項商業行為有關的活動，均稱為財經新聞，如經濟部財政部與中央銀行所發布的金融、財政政策，財經學術機構研究調查或民間的商業交易行為，均為財經新聞。徐筠惠（2002）則認為，凡是涉及財政、經濟與各項工業、商業行為有關的活動，均稱為財經新聞，例如如經濟部、財政部與中央銀行所發布的金融、財政政策，財經學術機構研究調查或民間工業生產、商業交易行為，均稱為財經新聞。此外，Lanson & Fought (2001)也指出，財經新聞涵蓋的範圍從生產線及於辦公室；無論是網路或紙本紀錄，從公司年度報告到政府檔案，都是財經資訊的來源。企業經營盈虧、產品的生產與銷售、利率匯率的升降變化，也都是財經新聞報導的重點。

[7] 王洪鈞（1959：255）認為一般所指之經濟新聞，包括物價行情、金融、市場、進出貿易、工業建設、勞工、農業、商業及其他有關之經濟活動。陳淑美（1992：7）指出，經濟新聞的內容包括：財稅、金融、經濟政策、證券、貿易、工業、農礦業、市場及工商報導等。方怡文／周慶祥（1999：16）則區隔經濟新聞與財經新聞，所有的經濟活動都個人息息相關，舉凡人民的納稅、銀行的存款、股票的投資、進出機場的商品查驗等，都是屬於經濟新聞報導範疇。

[8] 「商情新聞」或「商業新聞」，常見於國外的新聞分類（business news / business reporting）（Itule & Anderson, 1997；Mencher, 1997）。

[9] 例如《解讀財經新聞》（培生）、《股票財經新聞解讀入門》（寶川）、《財經資訊與媒體》（商周）等均冠以財經新聞。

經新聞」包括財政、經濟、商業等各活動，應屬當中涵蓋面最廣。

本報告採用「財經新聞」一詞，範圍包括財政金融、經濟、工商活動等消息，在意義上也涵蓋了「經濟新聞」、「商情新聞」所指涉的範疇。初步分析，財經新聞報導的內容可以歸納為以下三個特徵：財經術語、數字與數量、以及比較趨勢。

首先，財經新聞報導使用大量術語。[10] 術語是「用來表示特殊意義的專門用語」，例如通貨膨脹、躉售物價指數、消費者物價指數、所得稅等。財經新聞報導裡所涉及的術語有幾個特徵，第一、這些術語指向經濟活動相關的人、事、時、地、物。第二、術語通常言簡意賅，以較為簡短的字詞表現特定情境裡的若干事實，例如「財測」意指「企業之財務預測」。第三、術語通常以特定詞彙指向固定的意涵，例如「台積電」是指「台灣積體電路公司」；「逾放」則指「金融機構放款逾期未能回收」；比起一般新聞用語，財經新聞報導使用的術語較不易產生一詞多義。

其次，財經新聞報導使用大量數字相關的詞彙。[11] 數字用語包括數目字（如「一百」）和量詞（如「美元」），數目字本身非詞彙必須和量詞結合（如「一百美元」）方有意義。其次，由於數目字出現頻率高，因此財經新聞報導使用種類繁多的量詞。第三、財經新聞報導中的數字用語，可再細分為兩類；一種數字用語是「額度」，指涉固定的數量，例如金錢或商品的數量（例如，每股「100 美元」、每張「8.8 元」）

第三、財經新聞報導透過「比較」與「趨勢」彰顯新聞之核心意義。[12] 經濟活動的相關事實常以數量表達，但是單一數字並不足以說明，因此必須透過共同期間或類目，與另一組數目字相比較，才能表現出新聞意涵。股市漲跌變化、利率與匯率升降、原料與商品價格波動等新聞報導，均包含比較或趨勢。使用數字時，數量比較必須包含兩個特定情境（例如，「同年」或「同月份」）、並表現數目字之改變幅度（漲／跌，升／降）。

如上所述，一般新聞和財經新聞各可以萃取出若干特徵。新聞報導以事實為基礎、受到語言常規、敘事文本累積發展、展現特定文本類型，以及詞類出現頻

[10] 例如方怡文／周慶祥（1999：250-7）探討財經新聞的採訪與寫作，羅列了財經新聞相關術語與解釋，並認為這些術語是財經新聞的內容要素。此外，徐筠惠（2002）也指出，財經專業詞彙是財經新聞用語中最大的特色。

[11] 王洪鈞（1959：263）指出：「經濟新聞多半由數字砌成，每個數字皆代表一定數量。」Mencher(1997)也表示，商情記者必須具有處力數字的能力。Hough (1995) 談論新聞寫作中的數字，也表示商情新聞中含有大量數字。

[12] 王洪鈞（1959：261-2）指出：「經濟新聞比須著重比較方法，許多財政消息、生產消息、年度報告之類若沒有比較就會變得毫無意義…。比較方法多用再以線再數字和過去同一時期的統計數字比較，或以一地區和另一地區數字的比較。」徐筠惠（2002）亦認為「漲與跌近義詞」是財經新聞用語中的一類。

率不一。另一方面，財經新聞報導則大量使用術語、數字用語和重視趨勢比較。這些特徵，正是我們可以用於建立內容標記。以下我們將把財經新聞報導的知識應用到內容標記設計。

三、從財經新聞到內容標記

「內容標記」註記新聞報導內容，使電腦得以辨識並處理新聞資訊內容，進而可以跨越平台使用。每一則新聞報導內容的標記過程可區分以下幾個階段：(1) 數化（digitization）：將新聞報導轉換為數位資料；(2) 辨識（recognition）：確認加上標籤的資料內容；(3) 選擇（selection）：依據「資料類型定義」，找出詞組應註記的特定標籤；(4) 註記（tagging）：為詞組加上標籤：以及 (5) 驗證（parsing）：透過編碼軟體、確認內容標記「有效並且妥善」（valid and well-formed）。

理想的內容標記過程，在為新聞報導的所有詞組找到妥善的註記標籤，但是由於語言的多義，而使得內容標記未必周全。新聞內容與標記之間要產生「一對一」的對應，相當困難。這些考量，涉及標籤之定義是否清楚、以及資料類型定義之層級是否精簡，也關係著後續之內容標記工作是否繁複，成本高低，以及未來檢索是否容易。原因來自新聞報導的語意模糊、一詞多標籤、詳細度、以及關連性（Allen & Mohr, 1998），依序說明如下。

首先，標籤註記會面臨「語意模糊」的問題。這是新聞報導裡的特定詞組，在意義上無法只有一種意涵。換言之，標籤雖可「一對一」對應於詞組，但是其意義仍模糊而無法窮盡。無法明確對應之原因，有可能來自於同一詞組的分歧定義，例如「台北」雖是一個「地名」，若指「一個行政區域」，則應標記為「城市」；若指「以台北市及附近區域」，則應標記為「區域」。又例如「包青天」可能指向「包拯，一位歷史人物」，可能指向「某傳統戲劇裡的一個角色」。在標記時雖都可以標記為「人名」，但是意義卻不盡相同。因此在標籤裡必須進一步定義其屬性，以區隔其意涵。

其次，就算新聞內容相當清晰，可能也會發生「一詞指向多個標籤」(multiple categorization)。換言之，新聞內容可能對應到數個標籤或屬性。詞組可以指向多個標籤可能由於情境轉換而產生語義變化之故。變化可能來自於新聞文本裡的轉喻，例如一則新聞中「中南海」若指「北京市裡的一個區域」，新聞內容標記無法充分對應則應標為「地名」，但是在另一則新聞中可能指向「中國政府領導人」，應標記為「國家名稱」。此種混淆也可能來自於新聞報導用字的精簡，例如以「財政部說」代替「財政部發言人說」，此處「財政部」一詞其實是「財政部」和「發言人」兩個詞組的精簡，因此完整的註記應該包括「政府部門」和「職稱」兩個標籤註記。

第三、新聞內容標記也涉及「詳細程度」(specificity)，也就是標籤集要多詳盡的問題。以「總統」一詞為例，可以只標記處理「總統」這個標籤，但也可以將不同狀態的「總統」(除「現任總統」之外，還包括「前總統」、「先總統」、「故總統」等）納入標籤、或設定為屬性。另一方面，標籤集所涵蓋的範圍也涉及資料層級。以「2008 年北京奧運會」為例，共有三個詞組（2008 年、北京、奧運會）可同時一併用「事件」這個標籤註記，但也可以將事件內容再區分層級、或者再區分屬性（在「事件」標籤下加入「時間」、「地點」標籤或屬性)。一般而言，詞類區分如果越詳盡，則越能指出特定概念，但必須處理的標記越多，而成本也越高。因此標籤集設計者必須考量效益問題，在內容標記之目標和成本之間取得平衡。

第四、新聞內容標記也涉及「關連性」(serial relations)，也就是標籤與標籤之間在語意上關連的問題，有些詞組概念定義可以分割，有些則否。新聞報導的數個詞組之間可能產生語意關連，例如「經濟部國營事業管理委員會」指向「一個政府部門下的次級機構」，包含兩個互為關連的詞組，可以分別註記為不相屬的兩個概念定義（例如標記「經濟部」為「政府組織名稱」，而「國營事業管理委員會」為「部門名稱」)。但有些詞組的語意關連，則無法分割為個別的，例如「每股一百美元」則涉及數量與定量詞之間的關係，即使要切割，也必須在一個標籤下的第二層標籤或歸為屬性。因此，有些詞組可能必須結合成為一個標籤下的次層標籤或屬性，意義才會完整；然而標籤集層級越越多，將使資料辨識判斷過程越趨複雜，而成本也越高。

參、研究方法

本研究旨在建立財經新聞資料庫之內容標記文本類型定義。研究過程大致可分為兩個階段。第一個階段的任務是製作新聞內容標記之標籤集及標示說明，研究人員以前述歸納的新聞報導相關規律和現有產業標準為基礎，綜合制定一套標籤集，並擬定相關的標示說明。第二個階段主要任務是進行形成性評估(formative evaluation)，研究人員將標籤集及若干新聞樣本交付編碼員進行測試，然後最後依據測試結果修正

一、製作標籤集

標籤集製作共分為三個步驟，研究人員首先參考新聞產業之相關規格，接著進行財經新聞報導之語料分析，最後將歸納分類的詞類轉換成為標籤集和標記說明。以下依序描述這些步驟：

（1）參考媒體產業之內容標記規格：如前所述，新聞產業應用 XML 主要在建立「後設資料」(metadata)，一般並未納入內容標記之規則。就以產業主流規

的 NITF 而言，作法是保留「文內註記」（inline markup）項目。[13]NITF 公佈之文內註記項目有兩個特色，一是標記項目簡略，僅設定人物、功能、組織、地點、事件、物件、時間、財貨、數目、引語等十項，類似新聞學教科書中之 5W1H。另一個特色則是類型定義層級扁平，結構僅有兩層。研究人員參考此一結構，應用於第二階段財經新聞「語料分析」之類目編纂。

（2）財經新聞語料分析：本研究之語料分析是以人工方式進行。研究人員從中央社全文檢索資料庫財經類別之新聞條目中隨機抽取廿則新聞報導，作為語料分析的對象。[14] 接著將這些新聞中需要進行註記的詞組篩選出來，接著以 NITF 結構類目區別詞類，再下來則檢討這些詞類的結構關係，藉以決定標籤集內容和屬性。標籤集初步決定之後，研究人員進行前側，隨機選出新聞數則試行註記，並使用 XML 編碼軟體驗證（parsing），以確認標籤集是否「有效且妥適」（valid and well-formed）。本研究所建立的「財經新聞內容標記標籤集」如附錄一及附錄二。

二、形成性評估

本報告進行評估有兩重目的。一個目的是透過評估而修正資料類型定義，另一個目的則在觀察內容標記之編碼問題。本研究以人類編碼員註記，並由研究人員比對標記，並對標記結果進行分析。

本研究以兩位編碼員進行標記，編碼員具新聞科系背景，並稍微熟悉 XML 標記概念。標記過程採用編碼軟體 XML Spy 進行標示。編碼員需依照研究人員所制定之標籤集及標記說明進行標記工作。在編碼員進行編碼之前，研究人員事先說明標記方式及操作規則，並先以 10 則樣本進行演練。待編碼員熟悉軟體操作以及標記規則之後，開始標記 30 則樣本。編碼員進行內容標記工作時，均可參考研究人員提供的使用手冊。

編碼員標記內容的工作完成之後，研究人員以人工比對兩個編碼員的標記結果，並將編碼員針對相同樣本的兩份標記文件，進行兩兩比對。藉以發現編碼員之間的同意程度。凡標記不同之處均予紀錄、事後並加以分類。研究人員經歸納後所發現的標記歧異及其種類，作為分析的資料。

---

[13] 以新聞資料而言，「文內註記」就是將文章中有特殊意義的詞或句子以有意義的標籤名為其做標記的動作，讓讀者看出新聞中所要強調的東西，或是透過標籤的超連結功能得到更多資訊。

[14] 本研究以財經新聞為研究範疇，所建立之標籤集希望適用於一般所有財經新聞。本研究以中央社新聞資料庫中的新聞為抽樣對象，因為中央社資料庫收錄 1991 年 1 月 1 日以來至今的新聞，且財經新聞完整而數量龐大，具有代表性。抽樣方法採分層抽樣法，由於中央社新聞資料庫中財經相關的分類有三：國內財經、國外財經、大陸財經，本研究便依此三類的新聞數量比例來分層抽選樣本數，如此可以模擬三類新聞出現的比例。分層之下，再採用隨機抽選的方式，確定選取的樣本。

肆、結果與討論

研究人員發現，人類編碼員即使經過相當訓練，仍出現若干標記不一致之處，包括：對應於同一詞組的標籤不一致、同一標籤適用詞組範圍不一致、註記詳細程度不一致、以及詞組之間關係不一致。這些結果回應 Allen & Mohr（1998）的說法，在標籤集和使用手冊方面應予補強。資料同時也顯示，人類編碼員在辨識和標記受到相當限制，因此有必要（Kando, 1996）。

研究人員對原先制定的 DTD 作下列修訂： (1) 標籤名稱的變更：對於易造成錯誤理解、混淆的標籤，更改名稱，例如<稱謂>改為<職稱>、<數字說明>改為<數字名稱>、<地域>改為<區域>等。

(2) 標籤增減：對於無法標記的重要內容新增元素或屬性以供標記，例如，<數字>增加子元素<專有名詞>；<數字>增加子元素<幣制>，標記數字描述中的新台幣、美金等貨幣描述；<數值>增加屬性、參數，取消 "not"、上限"ul"、下限"ll"，等預設值。此外，分析顯示，定義不明造成標記內容歧異，研究人員也增加規則與範例中的文字說明部分，使得標籤集使用人員減少錯誤。

本研究也指出：(1) 財經新聞資料元素辨識和標記難易程度不一，人名、時間、地點、物件等易於辨識，但事件和原因的標示則不容易；(2) 形成性評估以編碼員標記歧異為分析基礎，固可提供初步資料，若採用量化資料，則應可獲得更精確的資訊；(3) 作為萃取新聞元素樣本數較少，未來研究分析樣本應再增加；(4) 未來可應用中文分詞程式（例如 CKIP）[15]做更精密的語料分析，並建立專業語料庫，作為自動化判讀基礎。

[15] 中央研究院資訊科學研究所中文詞知識庫小組之陳克健／黃居仁（1993）等人曾發展分詞程式，建立新聞語料字頻統計表。相同技術可作為本研究下一步之基礎。

參考書目

方怡文／周慶祥(1999)。《新聞採訪理論與實務》。台北：正中。

王洪鈞(2000)。《新聞報導學》。台北：正中。

徐筠惠（2002）：《財經新聞教材規劃初探》。國立台灣師範大學華語文研究所碩士論文

陳百齡（2002）：〈讓電腦也能夠認識新聞：如何標示新聞資料的內容？〉。中華傳播學會 2002 年會暨研討會，台北：深坑世新會館。

陳克健／黃居仁（1993）：新聞語料字頻統計表，技術報告 CKIP 93-02，台北：中央研究院資訊科學研究所中文詞知識庫小組。

-- 《新聞常用動詞詞頻與分類》，技術報告 CKIP 93-03，台北：中央研究院資訊科學研究所中文詞知識庫小組。

-- 《新聞常用名詞詞頻與分類》，技術報告 CKIP 93-04，台北：中央研究院資訊科學研究所中文詞知識庫小組。

陳淑美(1992)。《財經新聞自動分類之研究》。台灣大學圖書館學研究所碩士論文。

黃立文（2000）個人化網路新聞系統：雛形設計，新竹：國立交通大學碩士論文

黃罡慶、張鈺華(1997)。《股票財經新聞解讀入門》。台中市：寶川。

蔡琰／臧國仁（1999）：〈新聞敘事結構：再現故事的理論分析〉。《新聞學研究》58：1-28。

楊士仁、鄭優、趙政岷(1991)。《財經資訊與媒體 Q & A》。台北：商周。

臧國仁／施祖琪（1999）：〈新聞編採手冊與媒介組織特色：風格與新聞風格〉。《新聞學研究》60：1-37。

Allen, D. and Wiebke Möhr (1998). *Considerations for the Semantic Markup with the NITF*, available online.

Bosak, Jon & Tim Bray (1999, May). XML and the Second Generation Web, *Scientific American*, URL http://www.sciam.com/1999/0599issue/0599bosak.html.

Bray, Tim (1998). *News Wire Services Heading for XML*, available online, URL: http://www.xml.com/print/98/08/nitf.html.

Dumbill, E. (2000). *XML in News Syndication*. available online URL: http://www.xml.com/print/2000/07/17/sydication/newsindustry.htm.

Hall, Richard (2000, May). *Why XML is Important for Printing and e-publishing?* Online Technology, URL http://www.newsandtech.com/issues/2000/05-00/ot/05-00 _hall.htm

Hsieh, Ying-chun, Shyue-shuo Huang, Christian Wittern, Rick Jelliffe and Ching-chun Hsieh (2000). *Chinese Newspaper Metadata: Presenting Content of Science News Electronic* Cultural Atlas Initiative Conference, London, (June 26-28 2000). available online URL: http://www.som.uaf.edu/ffjal/papers/nitf.html

Mencher, Melvin(1997). *News reporting and writing* (7th ed). Madison, Wis. : Brown & Benchmark Publishers.

Mikula, Norbert (2000). *The Future of Internet Publishing*, Intranet Design Magazine. also available URL: http://idm.intranet.com/text/features/datachannel/xml.shtml.

Simon, Hank (2000). *XML: Strategic Analysis of XML for Web Application Development*. Charleston,

SC: Computer Technology Research Corp.

Berners-Lee, T., James Hendler & Ora Lassila (2001, May). The Semantic Web, *Scientific American*, URL http://www.sciam.com/2001/0501issue/0501berners-lee.html

van Dijk, Teun (1988) The Analysis of News As Discourse, *News Analysis: Case Studies of International and National News in the Press*, 8-30.

附錄一　財經新聞標示結構

新聞標示結構

絕對日期與絕對時刻兩屬性的時間描述法請參考 ISO 8601，語言屬性的描述請參考 ISO 639，貨幣代碼參考 ISO 4217。

(1) metadata 部分：

<metadata>
- ├ <新聞分類>
- ├ <關鍵字>
- ├ <識別符號>
- └ <發布時間>　屬性：絕對日期、絕對時刻

(2) inline 部分：

<標題>
- ├ #PCDATA(字串)
- ├ <引語>
- ├ <註記>
- └ %inline

<稿頭>
- ├ #PCDATA(字串)
- ├ <報導機構>
- ├ <撰文者>
- ├ <報導地點>
- ├ <外電來源>　屬性：縮寫
- ├ <報導時間>　屬性：絕對日期
- └ <文體>

<content>(內文)
- ├ #PCDATA(字串)
- ├ <導言>
- └ %inline

<導言>
- ├ #PCDATA(字串)
- └ %inline

/********** %inline **********/

<人>
- #PCDATA(字串)
- <姓名>　屬性：語言(預設值=zh)
- <職稱>
- <關係>　屬性：描述(value=姓名/身分)
  - <人>
  - <關係描述>　屬性：關係詞彙

<組織>　屬性：類型、全稱、簡稱、語言(預設值=zh)
- #PCDATA(字串)

<地方>
- #PCDATA(字串)
- <國家>
- <地區>
- <城市>
- <地點>

<時間>　屬性：絕對日期、絕對時刻
- #PCDATA(字串)

<期間>　屬性：絕對起始日期、絕對結束日期、絕對起始時刻、絕對結束時刻
- #PCDATA(字串)
- <起始時間>　屬性：絕對日期、絕對時刻
- <結束時間>　屬性：絕對日期、絕對時刻

<數字>　屬性：絕對日期、絕對時刻、類別
- #PCDATA(字串)
- <數字名稱>
  - #PCDATA(字串)
  - <專有名詞>
- <專有名詞>
- <幣制>
- <狀態>　屬性：比較符號(預設值=gt)
- <數值>　屬性：絕對數值、十進位字元、千進位字元、區間(非區間"not",上限 "ul",

下限"ll"，預設值=not)

├ <分數>

├ <下標>

└ <上標>

└ <單位> 屬性：關係、單位描述、貨幣代碼

<引言>

└ #PCDATA(字串) 屬性：發言者(以 ID 表示)

<專有名詞>

└ #PCDATA(字串) 屬性：類型(以 ID 表示)

<事件>

└ #PCDATA(字串) 屬性：類型(以 ID 表示)

<物件>

└ #PCDATA(字串) 屬性：類型(以 ID 表示)

/**/

附錄二　新聞標籤集說明

新聞標籤集說明文件

■ 屬性說明：

絕對日期：以 ISO 8601 標準的日期表示法(YYYY-MM-DD)來描述。Y 表示西元、M 表示月份(01-12)、D 表示日(01-31)。

絕對時刻：以 ISO 8601 標準的時間表示法(hh:mm:ss)來描述。h 表示小時(00-24)、m 表示分(00-59)、s 表示秒(00-60)。

語言：以 ISO 639 標準的語言代碼為屬性的內容，預設為中文(zh)。

貨幣代碼：以 ISO 4217 標準的貨幣代碼為屬性的內容，新台幣為 TWD。

類型：利用分類碼或分類用的關鍵字，目的在對所標記的文字內容做一分類。

■ 新聞標籤集說明：

◆ **新聞**：新聞標記的起頭，其下子元素依序為 metadata，標題，稿頭，內文。

◆ **metadata**：新聞本文前的資訊。此部分的資料可提供新聞數位化後，各項應用所需的資訊，例如，與新聞檢索有關的資料(例：關鍵字)，或是新聞分類代碼等。

子元素：**新聞分類**—新聞的分類代碼。

關鍵字—與新聞相關的關鍵字，可用作新聞資料庫的查詢。

發布時間—新聞發布的日期，有兩個屬性(絕對日期、絕對時刻)，用以記錄新聞的標準日期與時間。

識別符號—新聞的識別碼。

範例：

標記前：

EF5F1010.CAP

08/14/01　08:28:43

標記後：

```
<metadata>
  <新聞分類/>
  <關鍵字/>
  <識別符號>EF5F1010.CAP</識別符號>
  <發布時間 絕對日期="2001-08-14" 絕對時刻="08:28:43">08/14/01　08:28:43
  </發布時間>
</metadata>
```

◆ **標題**：標記新聞的標題。

子元素：**引語**—標題中被括號標示的導引語詞。

註記—標示新聞重要性的關鍵字。

範例：

標記前：

中共黨報宣稱抑制股市投機是長期政策『ＥＭ』

標記後：

<標題>中共黨報宣稱抑制股市投機是長期政策<註記>『ＥＭ』</註記></標題>

◆ **稿頭**：用以標記報導機構、撰文者、外電來源、報導地點、報導時間、文體。

子元素：**報導機構**—報導新聞的組織，例如：中央社。

撰文者—採訪記者的姓名或撰寫此篇新聞的撰文者姓名。

外電來源—外電新聞的報導機構，包含一個屬性(縮寫)，記錄外電來源的縮寫。

報導地點—新聞的報導地點，例如：紐約、台北等。

報導時間—新聞的時間，包含一個屬性(絕對日期)，以記錄此新聞的標準時間。

文體—新聞文章的文體，例如：特稿、電。

範例 1：

標記前：

中央社記者楊允達日內瓦特稿

標記後：

<稿頭>
<報導機構>中央社</報導機構>
記者<撰文者>楊允達</撰文者>
<報導地點>日內瓦</報導地點>
<文體>特稿</文體>
</稿頭>

範例 2：

標記前：

中央社華盛頓十三日法新電

標記後：

<稿頭>
<報導機構>中央社</報導機構>
<報導地點>華盛頓</報導地點>
<報導時間 絕對日期="2001-08-13">十三日</報導時間>
<外電來源>法新</外電來源>
<文體>電</文體>
</稿頭>

◆ **內文**：新聞的所有文字內容，第一段為導言。文章內所有的重要資訊皆可使用以下所介紹的 inline 標籤集加以標記。

子元素：**導言**—新聞的第一段。

範例：

標記前：

納稅人申報綜合所得稅時列報大陸地區扶養親屬，所檢附的親屬關係證明文件，已經稽徵機關查對核認在案者，以後年度申報同一親屬，可以原經海峽交流基金會的生存公證書正本核認，免每年辦理公證。

北區國稅局表示，自兩岸關係條例實施以來，有愈來愈多的納稅人於申報綜合所得稅時列報大陸地區扶養親屬，但納稅人申報所檢附的親屬關係證明文件，大陸地區少數縣市公證處只肯核發一次，不再重複，為避免納稅人無法再提示前述證明文件，造成稽徵實務上徵納雙方困擾，特別放寬規定。

標記後：

<內文>

<導言>納稅人申報綜合所得稅時列報大陸地區扶養親屬，所檢附的親屬關係證明文件，已經稽徵機關查對核認在案者，以後年度申報同一親屬，可以原經海峽交流基金會的生存公證書正本核認，免每年辦理公證。</導言>

北區國稅局表示，自兩岸關係條例實施以來，有愈來愈多的納稅人於申報綜合所得稅時列報大陸地區扶養親屬，但納稅人申報所檢附的親屬關係證明文件，大陸地區少數縣市公證處只肯核發一次，不再重複，為避免納稅人無法再提示前述證明文件，造成稽徵實務上徵納雙方困擾，特別放寬規定。

</內文>

■ inline 標籤集：可在新聞中標記重要資訊，例如，人、事、時、地、物等。

◆ **人**：標記具有姓名的人物，標記的範圍包含此人物的職稱或人際關係的形容。

子元素：**姓名**—人物的全名，包括姓與名。

稱謂—此人物在工作上的頭銜。

關係—當此人物在文中是以人際關係或親屬關係形容時使用。其下包含兩個子元素來描述此人物的關係。

子元素：**人**—標記與此主角相關聯的人物。

關係描述—標記關係名詞。包含一個屬性(關係詞彙)，用以統一關係的描述。

範例 1：

標記前：

中共「外貿部長」吳儀

標記後：

```
<人>
  <稱謂>中共「外貿部長」</稱謂>
  <姓名>吳儀</姓名>
</人>
```

範例 2：

標記前：

陳總統的夫人吳淑珍

標記後：

```
<人>
  <關係>
    <人>陳總統</人>
    的<關係描述 關係詞彙="太太">夫人</關係描述>
  </關係>
  <姓名>吳淑珍</姓名>
</人>
```

◆ **組織**：描述特定的國際組織、政府組織、財團法人等多人形成的團體。其中不包含籠統的集合名詞，例如，中小企業、上市公司等。標籤中包含四個屬性(類型、全稱、簡稱，語言)，分別記錄了與此組織相關的資訊。

範例：

標記前：

世界貿易組織

標記後：

```
<組織 類型="經濟" 簡稱="WTO" 語言="zh">世界貿易組織</組織>
```

◆ **地方**：標記新聞中出現的地方名詞，包括國家、地域、城市、地點。

子元素：**國家**—明確的國家名詞。

地域—對一個地區的泛稱。

城市—大至城市，小至鄉鎮。

地點—標記不屬於上述範圍的地理資訊，例如，公園、大樓等。

範例 1：

標記前：

台灣台北市

標記後：

<地方>
<國家>台灣</國家>
<城市>台北市</城市>
</地方>

範例 2：

標記前：

台北市二二八和平公園

標記後：

<地方>
<城市>台北市</城市>
<地點>二二八和平公園</地點>
</地方>

◆ **時間**：描述文章中所出現的日期或時刻。包含兩個屬性(絕對日期、絕對時刻)，以方便記錄絕對時間。

範例 1：

標記前：

財政部預定本月十九日標售

標記後：

財政部預定<時間 絕對日期="1997-09-19">本月十九日</時間>標售

範例 2：

標記前：

股市改為下午三點收盤

標記後：

股市改為<時間 絕對時刻="15:00:00">下午三點</時間>收盤

◆ **期間**：用以標記一段時間的文字敘述。包含四個屬性(絕對起始日期、絕對結束日期、絕對起始時刻、絕對結束時刻)，記錄絕對的日期或時間值。

子元素：**起始時間**—標記文字中所描述的起始時間。

結束時間—標記文字中所描述的結束時間。

(上述兩子元素皆包含絕對日期與絕對時刻兩屬性，以記錄絕對的時間值)

範例 1：

標記前：

創下九個月以來的單季新高

標記後：

創下<期間 絕對起始日期="1993-07" 絕對結束日期="1994-04">九個月以來</期間>的單季新高

範例 2：

標記前：

此會議將自二十五日舉行，二十八日結束

標記後：

此會議將自<期間><起始時間 絕對日期="2001-09-25">二十五日</起始時間>舉行，<結束時間 絕對日期="2001-09-28">二十八日</結束時間></期間>結束

◆ 數字：標記文章中出現的數字。具有絕對日期與絕對時刻兩屬性，以記錄與此數字相關的時間；類別屬性可記錄此數字的分類，例如，股市。

子元素：**數字說明**—標記描述此數字的說明文字。

狀態—標記此數字的漲跌或正負等狀態說明的文字，包含一個屬性(比較符

號)，以"gt"表示大於、"lt"表示小於、"eq 表示相等"。

數值—標記純數字的部分。可利用以下三個子元素來標記特殊的數字。

子元素：**分數**—標記以分數表示的數字。

上標—標記以上標表示的數字

下標—標記以下標表示的數字。

單位—標記數字的單位。具有三個屬性(關係、單位描述、貨幣代碼)，用以描述單位的相關資訊。

範例 1：

標記前：

今年大陸春夏播種糧食作物近十二億畝，比去年減少兩千多萬畝

標記後：

<數字 絕對日期="1993"><數字說明>今年大陸春夏播種糧食作物近</數字說明><數值>十二億</數值><單位/>畝</數字>，<數字 絕對日期="1992"><狀態 比較符號="lt">比去年減少</狀態><數值>兩千多萬</數值><單位>畝</單位></數字>

範例 2：

標記前：

成交金額四億一百五十萬美元

標記後：

<數字><數字說明>成交金額</數字說明><數值>四億一百五十萬</數值><單位>美元</單位></數字>

◆ 引言：標記新聞中人物或組織的發言語句，包含一個屬性(發言者)以記錄發言者的 ID。

範例：

標記前：

陳水扁總統說：海峽兩岸，一邊一國

標記後：(此例中， AD75 為陳水扁的 ID)

陳水扁總統說：<引言 發言者="AD75">海峽兩岸，一邊一國</引言>

◆ 專有名詞：標記財經新聞中的專有名詞。可利用一個類型屬性來表示此專有名詞的類別。

範例：

標記前：

美國那斯達克指數連跌三天

標記後：

<專有名詞 類型="股市指數">美國那斯達克指數</專有名詞>連跌三天

◆ 事件：標記會議等具有特定名稱的事件。同樣具有類型屬性為事件做一分類。

範例：

標記前：

二ＯＯ一年地球科學聯合學術研討會

標記後：

<事件 類型="會議">二ＯＯ一年地球科學聯合學術研討會</事件>

◆ 物件：標記產品等看得見或是具體的物件。具有類型屬性為物件做一分類。

範例：

標記前：

電玩遊戲軟體軒轅劍四上市兩個月

標記後：

電玩遊戲軟體<物件類型="電玩">軒轅劍四</物件>上市兩個月