XML/DTD理論實務與應用(1)－XML概論

華藝數位 陳嵩榮

大綱

- 標示(Markup)的基本觀念
- SGML、HTML介紹
- XML介紹
- XML相關標準
- XML的應用

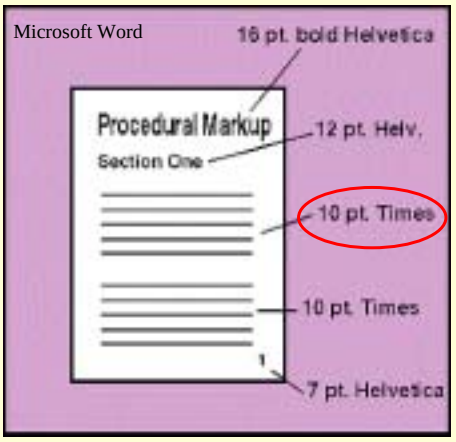
顧名思義

- SGML – Standard Generalized Markup Language (標準通用標示語言)
- XML – Extensible Markup Language （可延伸標示語言）
- HTML – HyperText Markup Language （超文件標示語言）

標示(Markup) 的基本觀念

- 藉著標示傳達某些關於被標示文字的資訊
- 標示的概念在我們生活週遭
- 標示的種類
  - **程序性標示 (Procedural Markup) :**
    - **針對文件的呈現外觀進行標示**
    - **例如：Microsoft Word、PDF**
  - **描述性標示 (Descriptive Markup) :**
    - **針對文件的內容和語義結構進行標示**
    - **例如：SGML、XML**

程序性標示 (Procedural Markup)

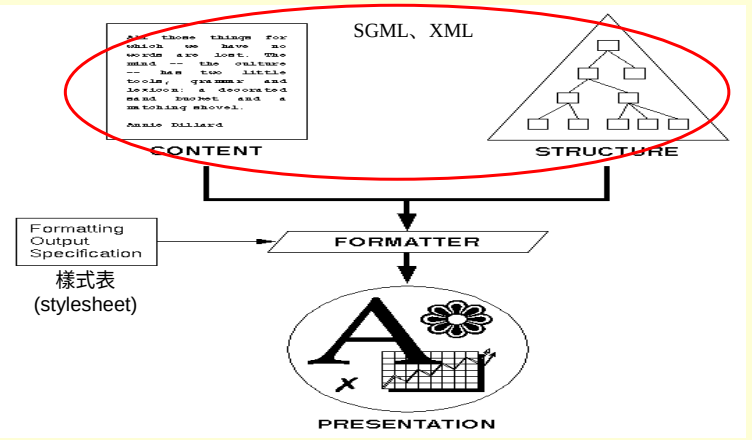


- 針對文件的呈現外觀進行標示
- 所見即所得 (What You See Is What You Get)
- 方便個人使用
- 文件的呈現樣式與內容儲存在同一份檔案
- 標示編碼通常使用專屬於特定平台或系統的控制碼（ex. binary code)

描述性標示 (Descriptive Markup)

- 針對文件的內容和語義結構進行標示
- 文件的內容與呈現樣式分開
- 呈現時須結合樣式表 (stylesheet)
- 標示編碼採用所有電腦平台均能解讀的ASCII code
- 利於大量文件的長期保存、交換與再利用

描述性標示的呈現



描述性標示(SGML、XML)文件與樣式表(stylesheet)分開

- 一份描述性標示文件可以使用不同的樣式表
  - 提高描述性標示文件的再利用性
  - 應用：同一則新聞內容，針對PC, PDA, 手機可有不同的樣式表版本
- 多份描述性標示文件可以使用同一份樣式表
  - 視覺設計維護更為容易
  - 應用：大型網站的網頁設計模式

SGML

(Standard Generalized Markup Language)

- 1986年 ISO 所制定的標準－ISO 8879
- 主要是為了文件交換與電子出版的使用而設計，用來描述文件結構，使得電子文件能在不同軟體系統間交換與傳輸。
- SGML是一套訂定標示語言的元語言（meta-language），藉由DTD（Document Type Definition，文件類型定義）來定義標籤及規範文件結構，並以其所定義的標籤來標示文件的內容
  - meta-language 生 language
  - language 生 document

SGML的優點

- 有彈性 (flexibility) :
  - 使用者可根據需求，制定各種不同的標示語言
  - 能夠描述各種複雜的文件結構
- 非專屬性 (non-proprietary)、平台獨立性 (platform-independence) 與系統獨立性 (system-independence)
  - 標示編碼採用所有電腦平台均能解讀的ASCII code
  - 利於文件的交換與長期保存
- 資訊再利用性 (re-usability)
  - 可利用已標示的結構作內容的加值處理

SGML的缺點

- SGML標準規格過於複雜
  - 開發相關應用軟體的成本太高
  - 即使是SGML大廠，在成本效益考量下也沒100%支援SGML標準
- SGML文件不易在Web上應用
  - SGML的起源比Web來得早
  - 目前Web主流瀏覽器(IE, Netscape)不支援SGML

HTML的限制

- HTML的標籤集與每個標籤的意義是固定的(pre-defined)，使用者不能自行定義標籤
- HTML大部分的標籤是用來控制呈現樣式，無法描述較複雜的文件結構
  - 和內容結構有關的標籤：<head>,<meta>,<title>,<body>,<p>
  - 無法支援較精確的查詢(通常採用全文檢索)
  - 資料庫中的資訊轉成HTML後，常會造成資訊遺失
  - 不同廠商所發展的HTML Extension不相容

XML的發展背景

- 隨著Web發展，人們希望在Web上除了展示資訊，網路內容可做更多自動化的應用，Web需要比HTML功能更強大的標示語言，像SGML那樣可以處理各種結構化資訊
  - SGML需要改造，才能適應Web的發展需求
  - 取SGML精華，補HTML不足— XML誕生

XML
（Extensible Markup Language)

- W3C在1996年底提出的標準
  - W3C : World Wide Web Consortium, 全球資訊網協會，宗旨是透過制定相關標準，提升網路的互通性，帶動Web發展
- 從SGML衍生出來的簡化格式，也是一種元語言(meta-language)
- 1998/2，XML1.0 Recommendation (W3C建議規格)
- Microsoft, IBM, Sun Microsystems, Oracle, Adobe, ArborText,...等軟體大廠支持，趨勢已成

W3C制定XML的目標

- XML為網路應用而生
  - XML能直接在Web上使用(SGML不行)
- 兼顧過去SGML的投資
  - XML能與SGML相容
- 易於開發應用程式才容易普及
  - 處理XML文件的程式能很容易被開發
  - XML的選項功能盡量保持最少，最好是零 (增加語法限制讓XML應用程式更容易開發)
  - 語法省略，例如省略結尾標籤，對於XML標示來說並不重要(增加語法限制讓XML應用程式更容易開發)
- ….

XML文件實例

- XML文件片段：

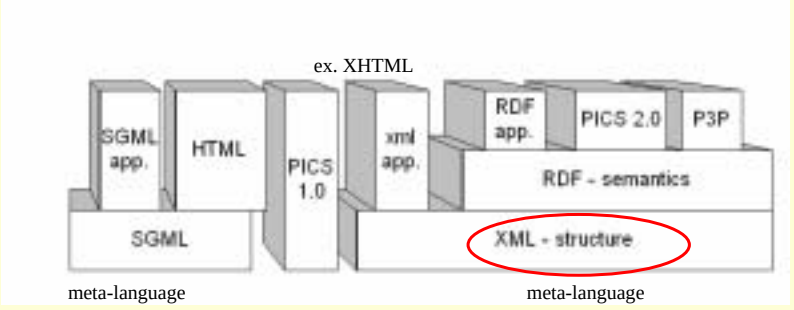
```
<customer-details id="AcPharm39156">
<name>Acme Pharmaceuticals Co.</name>
<address country="US"> <street>7301 Smokey
Boulevard</street> <city>Smallville</city>
<state>Indiana</state> <postal>94571</postal>
</address> </customer-details>
```

- 標籤的表示法如同HTML

SGML、HTML、XML的關係

- SGML(1986)
- HTML是SGML的兒子
  - 1989 invented by Tim Berners-Lee
  - 1996, HTML 3.2
- XML是SGML的弟弟(1998, XML 1.0)
- 所以XML是比HTML晚出生的叔叔
  - SGML、XML是meta-language，HTML是language，輩分不同
  - XML是SGML的簡化，而非HTML的延伸

W3C 的資料格式



XML與SGML的相同點

- 均是meta-language，可依據需求制定各種不同的標示語言
- 能夠描述各種複雜的文件結構，並可藉由（Document Type Definition，文件類型定義）來驗證文件結構的完整性與正確性
- 都須結合樣式表來設定文件內容的呈現格式
- 都具備跨平台、利於長期保存與再利用、能被人直接閱讀等特性。

XML與SGML的相異點

- 每一份SGML文件，均有相應的DTD；對XML文件而言，DTD可有可無
- XML排除SGML某些複雜不常用的語法：
  - SGML有包含 (Inclusion)、除外(Exclusion) 兩種語法規則來指定內容模式(content model) 的例外處理(Exceptions)；XML則不支援例外處理
  - XML不支援AND(&)內容模式、SDATA內部實體（internal entities）…等語法
  - …

XML與HTML之比較

	HTML	XML
擴展性 (extensibility)	HTML 的標籤集與每個標籤的意義是固定的 (pre-defined)，使用者不能自行定義標籤	使用者可根據需求自行定義特定的標示語言(包含標籤與屬性)
結構性 (structure)	大部分的標籤是用來控制呈現樣式，無法描述較複雜的文件結構	能夠描述各種複雜的文件結構
驗證性 (validation)	沒有驗證文件結構完整性與正確性的機制	可藉由 DTD 來驗證文件結構的完整性與正確性

XML文件

- 所有的XML文件都須符合well-formed XML 規則
- XML文件的種類：
  - well-formed XML文件：
    - 沒有對應DTD或XML Schema
    - 只需符合well-formed XML 規則
  - valid XML文件：
    - 有對應DTD或XML Schema
    - 除符合well-formed XML規則外，文件內容結構須符合DTD或XML Schema之規範

well-formed XML 規則

- 包含一個以上的元素。
- 恰有一個根元素。
- 不能省略起始標籤或結尾標籤。(如 .....)
- 所有的 標籤必須呈現適當的巢狀 (nest) 結構。
  (如 <B><I>bold and italic</B>italic</I> 是不允許的)
- 空標籤必須遵守特殊的XML語法。
  (如)
- 所有的屬性值必須括上單引號或雙引號。
  (如 <font size="2">)

→XHTML

XHTML

- W3C Recommendation 31-May-2001
- HTML, XHTML都是製作超文件網頁 (hypertext) 的標示語言
  - HTML是SGML的應用，HTML文件不符合well-formed XML規則
  - XHTML是XML的應用，XHTML文件符合well-formed XML規則

XHTML vs. HTML的不同(1/6)

- HTML不同元素的標籤容許互相交錯
  - 例：

    bold and italicitalic
- XHTML不同元素的標籤不容許互相交錯(必須是適當的巢狀結構)
  - 例：

    bold and italicitalic 或
    bold and italicitalic

XHTML與HTML的不同(2/6)

- HTML元素與屬性名稱大小寫均可
  - 例：

    <LI>….
- XHTML元素與屬性名稱必須是小寫
  - 例：

    <li>….

XHTML與HTML的不同(3/6)

- HTML某些元素可以省略結尾標籤
  - 例：

    ** here is a paragraph.**

    ** here is another paragraph.**
- XHTML除了空元素外，都必須有結尾標籤
  - 例：

    ** here is a paragraph. **

    ** here is another paragraph. **

XHTML與HTML的不同(4/6)

- HTML屬性值加不加引號均可
  - 例：

    <table rows=3>
- XHTML屬性值必須加引號(單引號或雙引號均可)
  - 例：

    <table rows="3">

XHTML與HTML的不同(5/6)

- HTML某些屬性可以採用簡化表示法
  - 例：

    <dl compact>
- XHTML屬性均不能採用簡化表示法
  - 例：

    <dl compact="compact">

XHTML與HTML的不同(6/6)

- HTML空元素表示法如起始標籤
  - 例：

    **
<hr>**
- XHTML空元素有特殊的表示法
  - 例：

    **
<hr/>**

撰寫well-formed XML文件

1. 使用純文字編輯器(如Windows的記事本)
2. 撰寫XML宣告
3. 撰寫根元素
4. 撰寫其他XML標籤與內容
5. 存檔，副檔名設為 .xml
6. 以瀏覽器IE檢視XML文件(瀏覽器IE內含XML Parser，可剖析XML文件是否well-formed)

撰寫 XML宣告

- 幾種XML宣告：
  - <?xml version=”1.0”?>
  - <?xml version=”1.0” encoding=”UTF-8” ?>
  - <?xml version=”1.0” encoding=”big5” ?>
  - <?xml version=”1.0” standalone=“yes” ?>
  - <?xml version=”1.0” standalone=“no” ?>
- 語法說明：
  - <? 是XML文件宣告的起始符號
  - xml 表示這是一份XML文件
  - version 用來指定XML標準的版本，目前是”1.0”
  - encoding 用來指定XML文件所使用的字集編碼，預設是”UTF-8”
  - standalone 用來指定XML文件是否獨立存在，預設是 “yes”
  - ?> 是XML文件宣告的結束符號

撰寫根元素

<?xml version=”1.0” encoding=“big5” ?>

<record-list>

……

</record-list>

撰寫其他XML標籤與內容

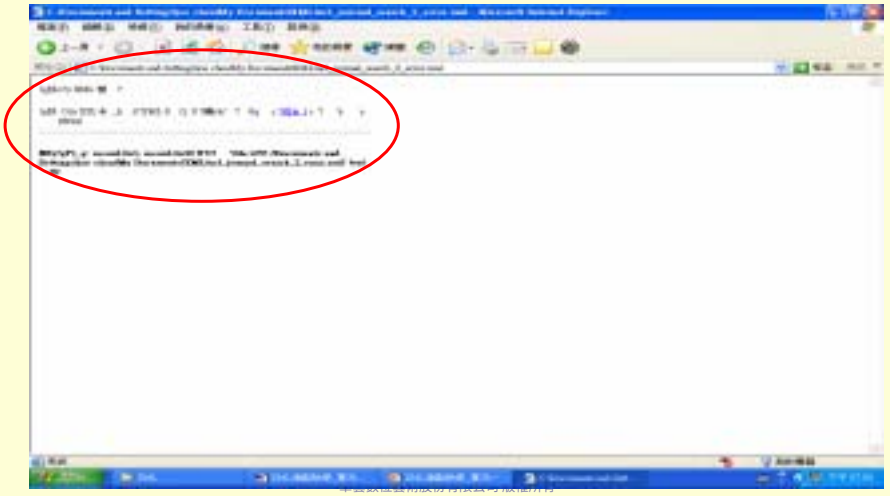
```
<?xml version="1.0" encoding="big5"?>
<record-list>
….
<record>
<seq>1</seq>
<title>XML及RDF技術介紹</title>
<creator>梁高榮</creator>
<journal>機械工業</journal>
<vol>220</vol>
<date>90.07</date>
…..
<url>http://www2.read.com.tw/…..</url>
…..
</record>
…..
</record-list>
```

注意事項：

- 標籤大小寫意義不同，如<title>與<Title>為不同標籤
- 標籤內容若有XML保留字元，須用替代表示法，如
  - < 以 < 表示
  - > 以 > 表示
  - & 以 & 表示
  - “ 以 " 表示
  - ‘ 以 ' 表示
- 存檔時，副檔名須設為 .xml

一份正確的Well-formed XML文件在Microsoft Internet Explorer 6.0 的瀏覽結果

一份有語法錯誤的XML文件在Microsoft Internet Explorer 6.0 的瀏覽結果



XML的樣式表(Style Sheet)

- CSS (Cascading Style Sheet)
  - W3C Recommendation
  - 為HTML制定的樣式表標準，XML也可使用
- XSL (Extensible Stylesheet Language)
  - W3C Recommendation, 2001/10/15
  - 專為XML制定的樣式表標準

XSL與CSS之比較

	XSL	CSS
能否使用在 HTML？	no	yes
能否使用在 XML？	yes	yes
使用語法？	XML	CSS

連結XSL樣式表的XML文件

```
<?xml version="1.0" encoding="big5"?>
<?xml-stylesheet href="ncl_journal_search_2.xsl" type="text/xsl" ?>
<record-list>
….
<record>
<seq>1</seq>
<title>XML及RDF技術介紹</title>
<creator>梁高榮</creator>
<journal>機械工業</journal>
…..
</record>
…..
</record-list>
```


一份結合XSL的XML文件在Microsoft Internet Explorer 6.0 的瀏覽結果

XML相關標準制定現況

- 主要標準：
  - XML 1.0 : W3C Recommendation 10-Feb-1998
- 解決元素與屬性命名衝突問題：
  - Namespaces in XML : W3C Recommendation 14-Jan-1999
- 結構定義：
  - XML Schema : W3C Recommendation 2-May-2001
- 呈現樣式 :
  - XSLT : W3C Recommendation 16-Nov-1999
  - XSL : W3C Recommendation 15-Oct-2001
- 更強大的超連結 :
  - XLink : W3C Recommendation 27-June-2001
- 超文件網頁標示語言：
  - XHTML 1.1 : W3C Recommendation 31-May-2001

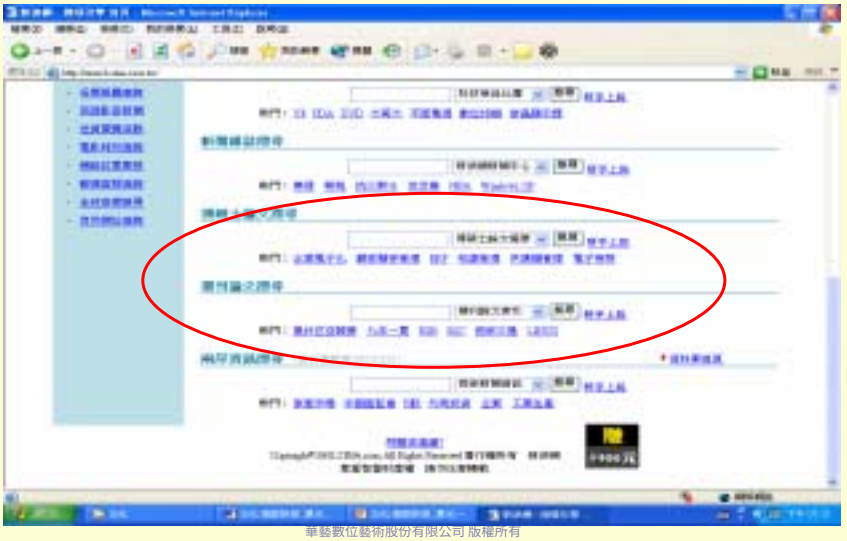
XML的應用(1/2)

- 定義各種不同用途的標示語言
  - XHTML, MathML (Mathematical Markup Language), SMIL (Synchronized Multimedia Integration Language), RDF (Resource Description Framework), WML (Wireless Markup Language)…
- 詮釋資料(Metadata)交換
  - XML是比HTML, SGML更適合的Metadata交換語法
  - 更開放的資料交換標準，如MARC XML DTD vs. ISO 2709
  - 簡易的異質資料庫整合模式，如國家圖書館與入口網站在期刊論文搜尋服務的合作
- 數位典藏
  - 以XML作為文字資料數位典藏的資料格式，可紀錄內容結構並確保長期保存

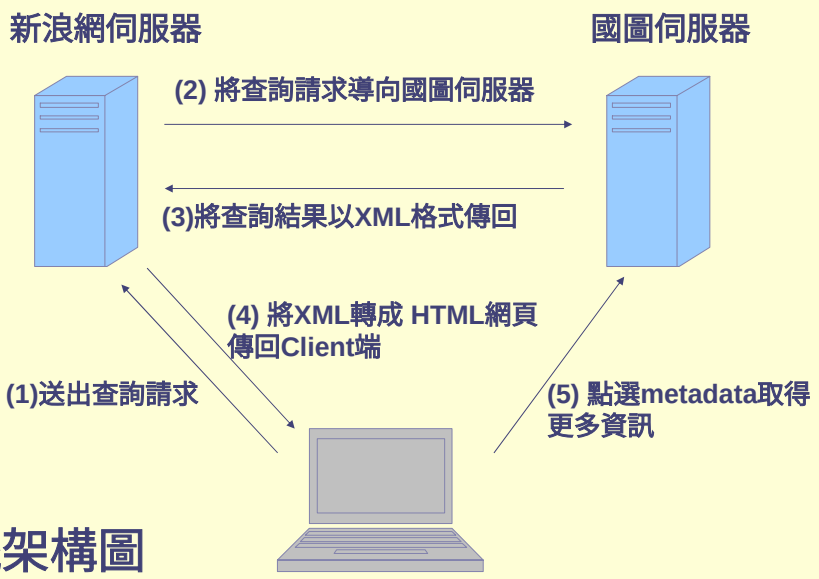
XML的應用(2/2)

- 同一份內容，多種呈現
  - 同一則網路新聞，可用PC, PDA, 手機介面瀏覽
- 電子書
  - 內容順序從固定、線性到多元、可自由跳躍
  - 針對不同的族群、需求、情境可設定不同的內容組合、順序與呈現方式
- 電子商務
  - 更有效率的B2B線上交易
- 資訊過濾、比價搜尋…
  - 如果新聞內容、產品銷售資訊是用XML紀錄
- 更有彈性的資訊再利用
  - 可試需要整批抽取XML文件部份內容

國家圖書館與新浪網在期刊論文搜尋服務的合作



國家圖書館與新浪網在期刊論文搜尋服務的合作



系統架構圖