

十分鐘小論文



FastData

快速的人文資料庫撰寫方式

陳鍾誠

2016 年 12 月 2 日

Hi, 大家好！

昨天

- 我和金門大學閩南研究所的《李宗翰》老師一起吃飯，他告訴我正在進行一個《專案》，是有關《台灣歷史人物數位典藏》的計畫！

然後

- 他給我看了一個國外的相關系統
- 該系統是用 MS. Access 寫的！

他希望延伸修改該系統

- 以便納入他們的資料！

於是希望我找學生來工讀

- 建立一個台灣歷史人物的檢索查詢系統！

我看了之後

- 提出了一些問題！

這些問題是

- 為何需要修改系統 ...
- 不能直接沿用國外的那個系統嗎？

李老師面有難色

- 說因為資料欄位不同
- 國外的那個系統主要描述中國官場人物
- 他們現在想做的系統要描述台灣的人物，
包含商人！

所以才需要修改系統

我聽了覺得怪怪的

所以我問

- 不能硬套上去嗎？
 - 官職欄位填入商業公司和職稱...
 - 套不進去的就塞入說明欄位...

李老師還是面有難色

- 我想資料或許差異不小！

這次的討論

- 讓我瞭解到《人文和科技》兩個領域之間果然有個很大的鴻溝！

這個鴻溝

- 不僅僅是因為電腦的使用或程式的撰寫
- 背後有些隱藏的知識，或許是《人文學者》
所迫切需要的！

現在

- 就讓我這個程式人，把這些隱藏的知識揪出來，暫時當一個《科技與人文的中介者》吧！

就我的觀察

- 對於《文史資料》而言
- 像是 MS. Access ， MySQL 這樣的
《關聯式資料庫》，或許是個不太
恰當的資料儲存格式！

為甚麼呢？

因為關聯式資料庫

- 是一種以《表格》透過《欄位》關聯起來的資料體系！
- 表格的欄位數和型態通常必須要固定
- 而且不能任意新增或刪除欄位，否則系統很容易就《掛點》了！

就我的認知

- 關聯式資料庫比較適合用在《工商業紀錄》的儲存與查詢上面
- 但是對於人文領域的那些文史資料，採用關聯式資料庫很容易造成《削足適履》的情況！

因為關聯式資料庫

- 是透過表格的 JOIN 運算進行連接與查詢的
- 這是一種高度格式化的資料庫，而且很沒有彈性！

但問題是

- 文史工作者，常常需要查詢一些
google 這類搜尋引擎做不到的事情
- 例如：請列出 1800-19011 年間所有廣
西布政使的名單！

所以

- 文史工作者很多時候不能只
依靠搜尋引擎！

但是就電腦的角度而言

- 搜尋引擎才是一種萬用的查詢系統，可以檢索任何資料，而且不需要知道資料的欄位格式，完全只要用關鍵字就可以了！

以下是 Google 的搜尋結果

← → ↺ <https://www.google.com.tw/search?q=1800-1900年+廣西布政使&oq=1800-1900年+廣西布政使&>

中正選舉 時空使徒 04 - YouTube 召喚萬歲 - 正文 第八 4399生死狙击 联通二 哩姆播台 - 光華戰記- Facebook

Google 1800-1900年 廣西布政使

全部 圖片 新聞 地圖 影片 更多 ▾ 搜尋工具

約有 30,900 項結果 (搜尋時間：0.46 秒)

广西布政使- 维基百科，自由的百科全书
<https://zh.wikipedia.org/zh-tw/广西布政使> ▾ 轉為繁體網頁
广西承宣布政使、简称**广西布政使**，驻桂林府，明、清两朝广西承宣布政使司行政首长。清朝顺治初 ... 满洲镶黄旗，进士，嘉庆五年二月二十八（1800年3月23日），由湖南按察使升任。 ... 光绪二十六年九月二十六日（1900年11月17日），升任江西巡抚。

廣西布政使- 维基百科，自由的百科全書
<https://zhwiki.nat.moe/zh-tw/广西布政使>
超過 60 筆 - 廣西承宣布政使、簡稱**廣西布政使**，駐桂林府，明、清兩朝廣西承宣布 ...

姓名	籍貫	任職
清安泰	滿洲鑲黃旗	嘉慶五年二月二十八（1800年3月23日），由湖南按察使升任
齊布森	滿洲鑲紅旗	嘉慶十五年十一月十四（1810年12月10日），同貴州布政使互調

「布政使」到底是一個多大的官？ - 每日頭條
<https://kknews.cc/zh-tw/history/ryvren.html>
2016年8月22日 - 所以說現在知道電視里的那些官們為什麼爭奪**布政使**這一職了吧。 ... 官也難當。1900年發生庚子之亂，之後到辛亥革命滿清滅亡共11年。11 ... 福建省閩侯人。1882年清光緒八年舉人，曾歷任**廣西**邊防大臣，安徽廣東按察使，湖南布政 ...

缺少字詞： 4800

但是文史學者希望一查出來就長這樣

姓名	籍貫	出身	任職	離職
清安泰	滿洲鑲黃旗	進士	嘉慶五年二月二十八（1800年3月23日），由湖南按察使升任。	嘉慶七年十一月二十五（1802年12月19日），調任浙江布政使。
恩長	滿洲鑲藍旗		嘉慶七年十一月二十五，由安徽按察使升任。	嘉慶十一年八月十六（1806年9月27日），升任廣西巡撫。
李鉉	山東壽光縣	進士	嘉慶十一年八月十六，由山西按察使升任。	嘉慶十五年四月二十六（1810年5月28日），年老召京以三品京堂用。
齊布森	滿洲鑲紅旗		嘉慶十五年四月二十六，由戶部主事升任。	嘉慶十五年十一月十四（1810年12月10日），同貴州布政使互調。
陳預	順天宛平	庶吉士	嘉慶十五年十一月十四，由貴州布政使調任。	嘉慶十六年九月二十三（1811年11月8日），同江寧布政使互調。
史積容	順天宛平	進士	嘉慶十六年九月二十三，由江寧布政使調任。	嘉慶十九年六月初五（1814年7月21日），革職。
葉紹桂	浙江歸安	庶吉士	嘉慶十九年六月初五，由大理寺少卿調任。	嘉慶二十二年九月十二（1817年10月22日），升任廣西巡撫。
富綸			嘉慶二十二年九月十二，由安徽按察使升任。	嘉慶二十三年十月二十（1818年11月18日），解任，同年十月二十四日（11.22日），革職（以偏執誣訐，出入罪名日）。

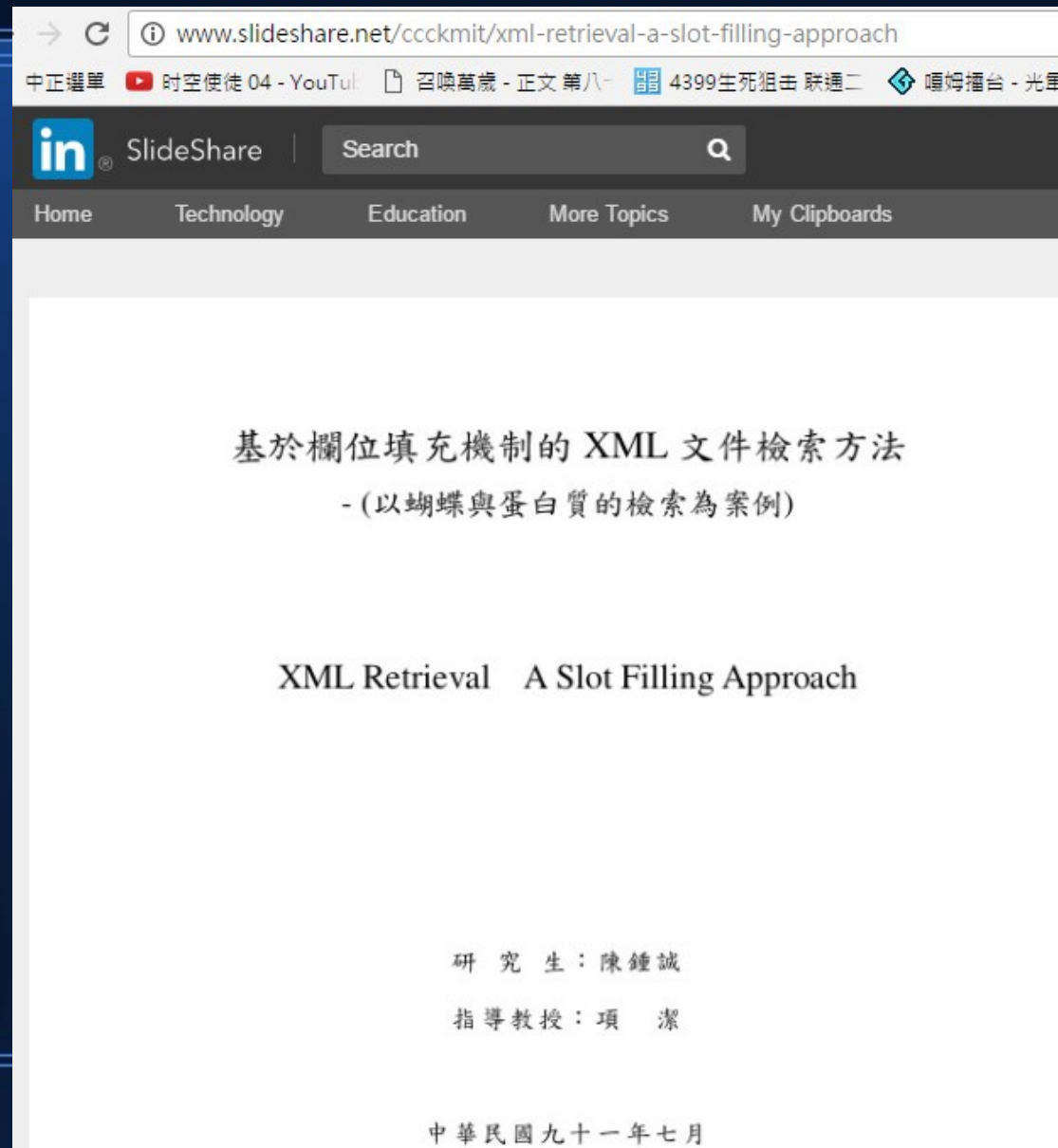
所以

- 對於文史工作者而言
- 需要一個介於《關聯式資料庫》
和《搜尋引擎》之間的系統！

這就讓我又回想起

- 自己當年的博士論文
- 就是提出《半結構化 XML 文件的檢索方法與自動建構工具》

那篇論文我最近才放上 Slideshare



The image is a screenshot of a web browser displaying a Slideshare presentation. The browser's address bar shows the URL www.slideshare.net/ccckmit/xml-retrieval-a-slot-filling-approach. The browser's taskbar at the top includes icons for a menu, YouTube, a document, and other applications. The Slideshare website interface features a dark header with the LinkedIn logo, the text "SlideShare", a search bar, and navigation links for "Home", "Technology", "Education", "More Topics", and "My Clipboards". The main content area is white and contains the following text:

基於欄位填充機制的 XML 文件檢索方法
- (以蝴蝶與蛋白質的檢索為案例)

XML Retrieval A Slot Filling Approach

研 究 生：陳鍾誠
指 導 教 授：項 潔

中 華 民 國 九 十 一 年 七 月

但是當年的 XML

- 其實是種非常囉嗦的結構！

直接拿來撰寫文史資料

- 會一直打很多

《標記》，

囉嗦的要死！

```
<?xml version="1.0"?>
<quiz>
  <qanda seq="1">
    <question>
      Who was the forty-second
      president of the U.S.A.?
    </question>
    <answer>
      William Jefferson Clinton
    </answer>
  </qanda>
  <!-- Note: We need to add
  more questions later.-->
</quiz>
```

XML

所以文史工作者

- 其實需要一套更彈性的
《快速撰寫方式》！

這就是我們今天

- 十分鐘小論文要探討的重點了！

在程式領域

- 目前我們寫程式的時候，也常常需要寫文件！

但是程式人

- 總是很懶得寫文件
- 他們需要快速的撰寫工具！

在 2004 年

- John Gruber 和 Aaron Swartz 創造出了一種稱為 markdown 的快速語言！

結果

- 現在很多程式人都用 markdown 來撰寫說明文件！
- 像是 github, stackoverflow 等等都採用 markdown 的格式撰寫！

以下是一個 markdown 範例 與轉換為 HTML 後的呈現結果

Text using Markdown syntax	Corresponding HTML produced by a Markdown processor	Text viewed in a browser
<pre># Heading ## Sub-heading ### Another deeper heading Paragraphs are separated by a blank line. Two spaces at the end of a line leave a line break. Text attributes <i>italic</i>, *italic*, bold, **bold**, `monospace`. Horizontal rule: --- Bullet list: * apples * oranges * pears Numbered list:</pre>	<pre><h1>Heading</h1> <h2>Sub-heading</h2> <h3>Another deeper heading</h3> <p>Paragraphs are separated by a blank line.</p> <p>Two spaces at the end of a line leave a
 line break.</p> <p>Text attributes italic, italic, bold, bold, <code>monospace</code>.</p> Horizontal rule: <hr /> <p>Bullet list:</p> apples oranges pears </pre>	<p>Heading</p> <hr/> <p>Sub-heading</p> <p>Another deeper heading</p> <p>Paragraphs are separated by a blank line.</p> <p>Two spaces at the end of a line leave a line break.</p> <p>Text attributes <i>italic</i>, <i>italic</i>, bold, bold, <code>monospace</code>.</p> <hr/> <p>Horizontal rule:</p> <p>Bullet list:</p> <ul style="list-style-type: none">• apples• oranges• pears <p>Numbered list:</p> <ol style="list-style-type: none">1. apples2. oranges3. pears

但是文史工作者

- 需要融合《表格、文件、彈性描述》等等結構於一體的資料建構格式！

於是我決定

創造一個格式給文史工作者用

這個格式稱為 FastData

以下是 FastData 的範例

FastData 融合了 CSV, Markdown, JSON 與 XML 格式

```
:::format=csv
```

```
#People
```

name,	bornDate,	dieDate,	domain,	detail
Steve Jobs,	1955/2/24,	2011/10/5,	Computer Entrepreneur,	#SteveJobs
Bill Gates,	1955/10/28,		, Computer Entrepreneur,	#BillGates

```
#SteveJobsTimeTable
```

time,	event
1976/4/1,	#Apple Inc. Created
1977/4/16,	#AppleII computer introduced at West Coast Computer Faire.
2001/10/23,	#iPod released
2007/6/29,	#iPhone released
2010/4/30,	#iPad released

```
#AboutSteveJobs
```

Steven Paul "Steve" Jobs (/ˈdʒɒbz/; February 24, 1955 - October 5, 2011) was an American businessman, inventor, and industrial designer. He was the co-founder, chairman, and chief executive officer (CEO) of Apple Inc.; CEO and majority shareholder of Pixar;^[2] a member of The Walt Disney Company's board of directors following its acquisition of Pixar; and

希望對文史工作者會有幫助

```
:::format=JSON

#SteveJobs
{
  company:"Apple|Pixier|Nextstep",
  product:"Apple II|iMac|iPod|iPhone|iPad",
  timetable:#SteveJobsTimeTable
  about:#AboutSteveJobs
}

#BillGates
{
  company:"Microsoft",
  product:"DOS|Windows",
  timetable:#BillGatesTimeTable
  about:#AboutBillGates
}
```

這就是我們今天的

- 十分鐘小論文！

希望您會喜歡

我們下回見！

Bye Bye!