資料品質及資料淨化技術之探究

－以環保部門資訊整合為例

連志誠　黃素梅　　東吳大學資訊科學系
朱雨其　　行政院環保署監測資訊處

摘　　要

隨著公私企業部門對整合性資訊系統及資料倉儲應用的日益普及，「資料品質」已成為當前資訊系統建構過程及日常操作應用的一項重要課題。由於單一資料來源的資料品質控管程序與多個資料來源的控管程序差異甚大，現行資料倉儲相關技術可否有效提昇資料整合後的資料品質問題，頗值得探討。事實上，資料品質的問題在傳統資料庫與整合性資訊系統（或資料倉儲）有相當差異，在資料整合過程中，來源資料因為使用者輸入錯誤或組織環境隨著時間的推移而改變，這些都會影響所存放資料的品質。本文旨在探討在資料品質的特性及現行提昇資料品質的學理背景及實作方法，特別是資料淨化（data cleaning）技術的相關討論，並以環保部門的資訊整合應用實例輔助說明，以期對資料品質課題有通盤性之瞭解與掌握。

關鍵詞：資料品質、資料淨化、資訊整合系統、資料倉儲

A Study on Applying Data Quality and Data Cleaning Technologies to Government Data Integration

Chin-Cheng Lien, Su-Mei Huang, Dept. of Computer Science, Soochow University
Yu-Chi Chu, Dept. of Information Management, EPA

ABSTRACT

Recently data integration among the departments of government has proposed as an important work to increase the quality of the services provided by government. Data quality is one major factor to make a successful data integration. In this paper, we proposed an approach which advances the current data integration approach of government with data mining technologies. We study the process of data integration in the Environmental Protection Administration Executive Yuan (EPA) of Taiwan. Then we define some metrics of data quality to detect and protect the data produced some unfavorable behaviors such as typing error, unauthorized data, data become out of date, etc. Finally, we design a lot of rules to guard the quality of the data in the related databases in EPA. We got a better data quality with our approach than that of the current approach from EPA databases.

Keyword：Data cleaning , data quality, data warehouses, information integration systems

1. 前言

根據 Data Warehousing Institute 的一份調查報告顯示[8]，美國企業因資料品質的問題，每年損失 6 千億美元。由於網際網路之便捷及資料庫技術之發展，使得資料整合系統之建置日益普遍。在資料整合計畫開始後，大多數企業會採行資料倉儲技術，同時運用資料擷取、轉換並載入 (Eliction, Transform, Load,ETL)作業程序，將各個不同來源的資料整合匯入資料倉儲。在資料擷取的階段中，原有的資料必須經由檢視，找出其中問題或錯誤，並且儘可能解決這些問題。惟以往多以人工方式來清理資料，不僅成本高，也相當耗時，而且效果有限。Data Warehousing Institute 便發現在員工基本資料中，大約會出現 76%的錯誤，進而產生有瑕疵的資料。一個主要的因素是資料來源本身充滿了許多有問題的資料，是以如何有效地處理資料整合系統或是資料倉儲環境所衍生的資料品質問題，已是當前學術界與實務界亟為關注的課題。

一般來說，資料倉儲可視為一種整合性的資料儲存體，其內部所儲存的資料是由多個分散式、自主性及異質性的資訊源中，萃取並整合而來的，由此相對地更突顯出資料倉儲中資料品質問題的重要性。換句話說，決定資料倉儲系統是否能夠開發成功以及有效運作發揮實際功效的一項關鍵因素，取決於資料倉儲系統內部所儲存資料的品質是否足以適用。資料品質的問題在傳統資料庫與資料倉儲中最大的不同在於資料倉儲中的資料主要是做為決策支援，而非操作性的交易應用。因此儲存在資料倉儲中的資料通常是歷史性資料，具有時序上的變化，是以資料品質的問題會因時間延續而加劇並益加複雜。據估計約有六成以上的資料倉儲系統宣告失敗，其主要的原因是沒有充分的時間與努力來解決系統中的資料品質問題[9]。

本文將著重討論政府部門在進行資料整合時，所面臨的資料品質問題，例如各來源資料庫對相同物件有不同定義或是不一致的資料內容時應如何處理。我們以行政院環境保護署的環境資料庫為例，該資料庫包含了水質監測、空氣品質監測、毒化物管理等部門所建置的資料來源。本文比較目前詮釋資料作法，嘗試提出以階層性資料品質模式搭配詮釋資料及資料探勘等機制，以自動化偵測資料品質問題，提高資料淨化效果。

本文第二節回顧資料品質相關文獻，第三節描述階層性資料品質模式，第四節以環保部門資訊整合之資料品質管控作為探討實例，第五節為結論。

2. 相關文獻回顧

有關資訊系統中資料品質課題的探究，麻省理工學院的「整體資料品質管理計畫」（total data quality management program）可說是此領域的開路先鋒之一[7]。該計畫每年定期舉辦資訊品質相關議題研討會等活動，對資訊品質研發工作有具體貢獻。Wang et al 在此計畫早期的研究中曾倡議一個基於 ISO9000 標準的資料品質分析架構，這個架構檢視了有關資料品質的一些重要文件，其中有些研究是針對植基屬性之資料品質管理方式進行探討[13]，由於這種管理方式會變動資料庫的原有結構，加入品質相關的訊息，而這些品質

資訊與資料是同存於一個資料表，所以必須在資料更新過程中同步更新品質屬性資料，以免造成錯誤的資訊，如此一來，存取這個資料表的資料庫語言（SQL）就得因應資料綱要的改變而調整，實作上並不十分便利，加以近期資訊整合的來源日益多元，例如 XML 資料及網頁資料等，這種方法有其侷限性。

Jarke et al 倡議一個以擴充性儲存庫為基礎的一種資料倉儲架構，但此架構主要專注在資料倉儲系統設計及建置階段的品質控管，相對的也較少注意到資料倉儲內的資料品質問題。Naumann and Leser 將各個不同面向的資料品質因素，如完整性及時效性等納入「多資料庫查詢語言」中。其作法主要採行「觀點重寫機制」(view-rewriting mechanism)，同時將資訊品質塑模成不同的階層[12]。Dasu and Johnson 將資料探勘（data mining）技術與資料品質的相關技術與學理探討作有系統性歸納整理[3]，顯見資料探勘技術在資料品質的管控方面具有相當正面之功能，我們將在第四節就此課題作說明。

在實作及應用方面， Helfert and Herrmann 以瑞士某家銀行的資料倉儲系統為例，提出一個維持高資料品質的方法（以詮釋資料為基礎之資料品質系統）[4]。Luebbers et al 則以汽車引擎製造商資料倉儲系統為例提出資料品質已被認定為資料倉儲系統成功與否的關鍵要素之一[11]，這份報告中提出以資料探勘為基礎之系統架構來稽核與管控資料品質。Berndt et al 的研究報告則指出，對於一個醫療資料倉儲系統來說，資料品質管理策略更形重要，必須不斷評估、監測並且避免產生錯誤的資料而導致錯誤的醫療決策[2]。醫療資料倉儲系統通常擔負幾個重要使命, 包括醫療政策制定、大量醫療資料搜集維護以及新醫療技術研發的支援，譬如說一份關於突破一些心臟病治療挑戰的最新研究報告公佈後，公共衛生官員就能快速從資料倉儲系統檢索更加詳細的資訊，又如婦女和兒童健康的資訊含括嬰兒死亡率或初生兒體重過輕，公共衛生官員就可以參考資料倉儲系統產生的報表，制定更完善的醫療保健政策。綜上論述，資料品質的優劣不只是資料倉儲系統成敗的關鍵，更攸關國家人民的福祉，但是，在建置資料倉儲的過程中資料品質的維護與提升，通常需要耗費相當大的人力及時間成本，所以，發展自動化、有效率且極具正確性的資料品質管控系統及相關作業流程機制等，益顯重要。

我國政府目前正積極推動電子化政府相關工作，建置整合性共通平台，希望藉由此平台能提供人民單一服務窗口，因此資料淨化方法在此平台將扮演不可或缺之角色，對提供正確的公務部門資訊有重大影響。目前一般機構僅有基本的資料淨化方法，例如植基屬性方法與詮釋資料方法。

植基屬性方法是以屬性為基礎（attribute-based）的資料品質管理方法，係針對資料的屬性在資料欄位附加上品質因子，這些品質因子就是實際資料所欲達到的目標、特徵以及產生的過程，使用者則依據實際應用需求來評估資料的品質[13]。附有品質因子的資料欄位可聯結至相關的品質資訊，其儲存結構改變了原始關聯資料庫中資料欄位值必須是單一值的限制，使每項有品質因子的資料欄位都以下列序對方式儲存資料。

<Attribute，Quality_Key>

由於使用者對資料品質的觀點與需求被併入資料倉儲的設計中，因此對於不同使用範疇的使用者，可依據其使用權限或需求的資料品質標準，作為篩選資料的標竿，一旦外界環境對資料品質的需求有所改變，也可隨時再重新訂定資料品質的的標準，進行資料篩選與淨化處理，這使得資料品質的驗證評估在資料倉儲發展與應用的任一階段都可進行處理，而非僅限於設計階段或應用階段始可進行[10]。

表 1 為含有品質資訊的資料庫示例,原有的資料表格「水質」<測站名稱, 懸浮固體, 大腸桿菌群, PH 值.>，在屬性欄位附加上品質因子，品質因子中包含有對資料相關的品質訊息，如表 1 中屬性欄位附加了二個品質因子：<輸入日期, 校核者>，分別用以表示資料鍵入日期、資料校核人員。這些擴展後的屬性欄位可以因時因地制宜的增減，如此即可進一步取得相關的品質資料。

表 1：結合資料品質的資料模型

測站名稱	懸浮固體 (mg/L)	大腸桿菌群 (CFU/100ml)	PH 值	輸入日期	校核者
大直橋	38.8	240000	7.4	2005-12-07	Kevin
中山橋	46	10	7.3	2005-12-12	Richard
:	:	:	:	:	:
:	:	:	:	:	:
attributes for data value				attributes for data quality	

但是這種方式會改變原始資料儲存方式，同時還要修改 SQL 的查詢語言結構使其包含對品質資料的處理，因此對品質因子的設定、儲存及擷取必須做進一步的處理，以免造成資料因新增或刪除所形成的異常現象。

詮釋資料方法是以詮釋資料為基礎之資料品質系統[4]，主要專注於資料品質的持續改善，圖 2 顯示了從操作性系統到分析系統的整個資料倉儲系統架構，資料品質在整個資料流的過程中，不斷的被量測以及評估，其中詮釋資料扮演了重要的角色，特別是在資料轉換程序及資料綱要中所用到的詮釋資料都被用來量測資料品質。以詮釋資料為基礎之資料品質系統最重要的一個概念就是整合詮釋資料管理,經由這個詮釋資料管理將所有有關資料品質的重要資訊都納入，包含以下三個部份：

1. 測量及評估資料品質的規則庫(rule base)：透過規則的建立，不但可以設定測量資料品質的條件及標的，還包括了執行的時間程序。
2. 通知規則(notification rules)：當有不符合品質規則的資料或者是偏差的資料產生時，就可以根據通知規則來決定經由何種方式通知相關品質確認工程師，例如利用電子郵件通知品質確認工程師之後，工程師就可以採取適當的處理步驟。
3. 品質聲明(quality statement)：這些聲明包含資料品質測量的結果以及採用何種方式來展現給終端使用者，例如階層式的自動化控制迴圈就可以聚集低階層的品質，再將結果以三種不同的顏色(綠色代表品質良好、黃色代表有部分瑕疵、紅色代表品質低劣)讓使用者很容易辨別以及了解。

詮釋資料是橫跨整個資料倉儲系統架構的,詮釋資料的儲存與管理可說是資料倉儲系統最重要的一環，不過資料倉儲團隊在作品質確認時通常不會去驗證詮釋資料的正確與否，所以詮釋資料容易被忽視，而這份研究不但建立詮釋資料的管理機制，還加上資料品質的稽核管控,整合了資料倉儲系統的兩個關鍵成功因素,可以提昇資料倉儲系統成功率。

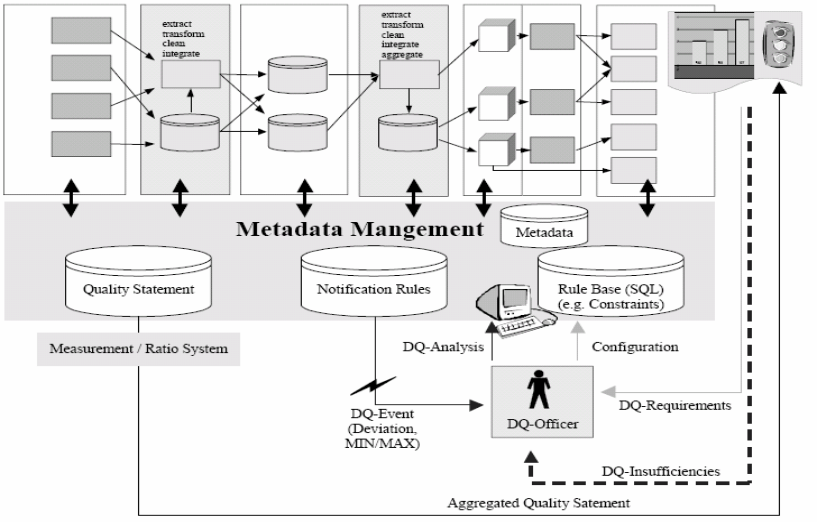

圖 2：以詮釋資料為基底之資料品質系統架構[4]

但若僅以詮釋資料作為基礎之資料品質系統,端賴資料倉儲團隊成員耗費心力持續維護資料品質規則庫，且只能對於已定義的品質問題進行處理，是以有其侷限性，倘能搭配資料探勘的機制，自動發現潛藏的問題，並產生新的調和規則，應能提升資料品質問題的處理速度及適當性。

行政院環境保護署針對環境資料庫建置，已有若干成果，特別是在資料整合的規模與範圍方面，正逐年擴增並有具體效益，惟在資料品質管控方面，尚有下列改善空間：

1. 缺乏完整之資料品質作業準則，以致資料整合時，須仰賴大量人力作業。
2. 資料品質淨化規則與自動化作業工具尚待強化。

以下我們提出階層性資料品質模式搭配詮釋資料及資料探勘等機制，嘗試改善上述情況，以期環境資料庫整合能質量兼備。

3. 階層性資料品質模式

我們參考 EPA 的現況與相關研究建立資料品質的方法。就資料品質本質上的觀點而言，在使用資料庫或資料倉儲中的資料時，使用者最關切的莫過於資料是否適於使用，因此目前不論是學理或實務上，大部份都將資料品質的意義定位在「適於使用」(fit to use)的目標上[5]。由這項定義繼續延伸，必須再明確推衍出使資料適於使用的基本要素，基於這項需求，可將資料品質再細分成四個層面(dimension)來討論分析[13]，每個層面又可再細分為若干個資料品質參數(data quality parameter)，資料品質參數的主要作用是讓使用者評估資料倉儲中的資料品質[14]。品質參數的形成與選擇目前雖沒有一定之規則，但仍

須符合能表示出原始資料特徵的先決條件。

圖 1 是構成資料品質定義之階層圖，資料品質的四個層面與品質參數說明如下：

1. 存取性(accessibility)：對使用者而言，具有良好資料品質特性的資料倉儲應具有輕易取得所需資料以便進一步分析操作的功能。其次是安全性(security)的考量，對於機密性的資料為確保其安全性與隱密性，必須有效限制使用者的存取。乍看之下，資料的存取性與安全性考量在某些情形下是相衝突的，但若完全無法取得這些保密性資料，則分析人員將無法研究解決資料不適用的問題，資料的需求管理者也無法作出相關的決策。所以發展一套合理的機制，有效限制機密資料的存取是必要的。例如環保機關對於未經確認的環境監測數據，必須適當規範存取的限制，以確保資料不被誤用。
2. 詮釋性(interpretability)：詮釋性層面的品質參數主要用來描述資料本身的定義，資料的定義明確清晰將有助於使用者或分析人員提高對資料的了解度，了解的層面應包含資料的格式、內容與主要用途。因此在詮釋性層面中包含了資料格式(syntax)與資料語意(semantics)等兩個品質因子，用來定義資料項的屬性。例如水質檢測資料必須定義其數據所用的單位、檢測地點名稱與座標位置，以確保資料被正確解讀。
3. 議題關聯性(contextual)：資料是否切於議題，在於資料的適用性。資料是否適於使用則可由資料量的多寡、資料之相關性(relevancy)及合時性(timeliness)等三方面來探討，其中合時性品質參數又將資料的時間性質區分為非揮發性(non-volatile：資料的使用無時效限制)與具時效性(current：資料項被存入資料庫的時間及其有效期限)。匯入資料倉儲的資料量多寡應視應用為基準，而非將所有資訊源中的資料完全整合匯入系統中。例如空氣或水質的監測資料就必須分成即時性資料與歷史性資料二類，二者所服務的對象及其對品質的要求有相當程度的差異。
4. 可信度(believable)：資料倉儲的資料除了要與議題具有相關性外，還要能取得使用者的信任；假設民眾不信任環保機關所發布的監測數據，則這樣的資料是否具有利用價值，不無疑問。資料的可信度，一般可由完整性(completeness)、一致性(consistence)、正確性(accurate)及可靠性(credible)等四個品質參數來衡量[1,6]。

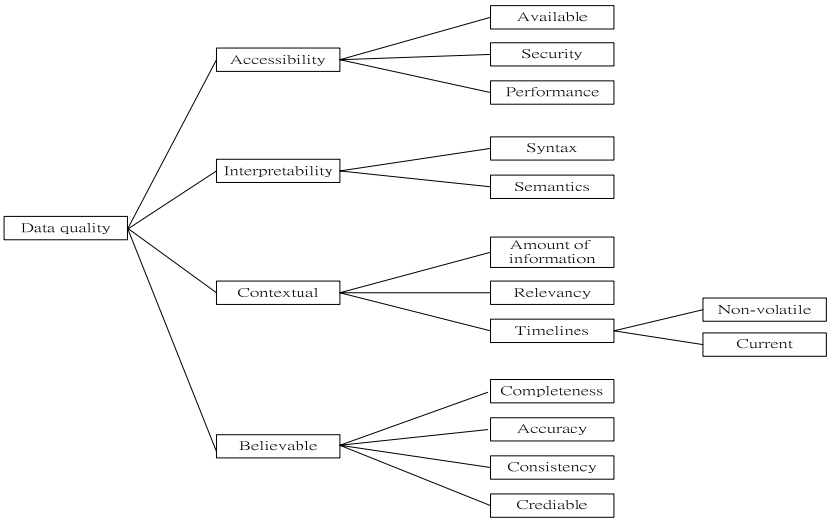

圖 1：資料品質定義之階層圖

一般而言，資料品質的概念是屬於多維度的，而品質參數的訂定也最好要符合各種不同的資料型態，來自各個不同的應用領域中，都有著共同的品質參數用以評估其資料品質的需求。這些品質參數除了可用來標示資料項之外，還可以標註於品質因子上，以確保品質因子的品質。

將資料品質塑模成階層式的架構有助於系統管理者及使用者明確地掌握資料品質相關因素；同時，不同領域、不同需求的使用者可依其需求特性，將該階層架構作彈性調整，以契合實際需求。是以，在資料倉儲建構時，宜將資料品質課題依階層架構分階段納入建構要項，例如在資料倉儲設計規劃時，明確定位各資料品質之需求及其投資成本效益，而在資料彙入時，則是資料品質把關的最重要時機點，在資料品質階層中，比較基層的因素應能充分評量及掌握，如此在後續的維護管理階段，才能具體的維持資料品質。

4. 環保部門資訊整合之資料品質管控

以下介紹我們所建議的環保部門資料品質管控程序，以往政府部門大都依個別機構的權責劃分資訊處理作業，也就是從「政府觀點」發展，不論各政府機關間，或是機關內部各單位間，係以獨立方式進行業務電腦化流程，於是形成許多「煙囪式」（stove-piped）系統。這些系統彼此不盡相容，同時重複建置資料。民眾要擷取資料時，需分別從各業務主管機關進行查詢檢索。例如，有民眾欲申辦「工廠設立許可」，他可能要查詢縣市政府建設局、環保局、工務局（建築管理）等單位，甚或是中央政府經濟部（商業司、工業局）環保署等單位的網站才能取得完整資料。政府部門應該以「民眾觀點」發展，提供主題導向（subject-oriented）的整合與服務，才能提昇為民服務的水準，然欲達成資訊基礎面的整合，尚有相當多課題待克服，但綜觀現行實務應用上，資料品質的課題是其中之一。

由於組織文化、資料特性、使用者作業習慣及成本效益考量等因素，在公部門實施資訊系統的整合及資料品質的管控作業與一般企業有相當程度的差異性。例如前面所提的詮釋資料方法在一般企業可以獲致不錯的成果，但是在公部門由於使用者配合程度，可能無法完全掌握各個資料來源的資料品質特性，以致不易彰顯整合效果。我們認為在公部門的資料品質提昇工作必須要通盤考量資料來源特性、人員參與及技術工具等各個面向，同時要搭配適當的行政措施，才能克竟事功。

圖 3 說明一個公部門資訊整合加上資料品質管理平台的架構及流程。我們可以分為三個面向來討論：

1. 人員方面：主要參與人員應含括業務分析人員（business analysts）及資訊技術人員。業務分析人員除了應來自各相關業務部門外，也應同時顧及不同層級，包括決策人員及基層業務人員等，他們在界定資料品質的需求及各項定義方面是非常重要的。例如對水質監測資料各個不同測值單位的對應關係，水質監測項目的名詞定義及語法格式等。而資訊技術人員則根據業務分析人員所歸納出的業務知識與品質需求，發展相對應的程式模組及系統工具，用來發覺、偵測或是過濾可能發生的資料品質問題。由於公部門的組織型態及文化與一般企業有所差異，是以參與人員之表現良窳可能影響資料品質管理平台的執行效率及正確性。

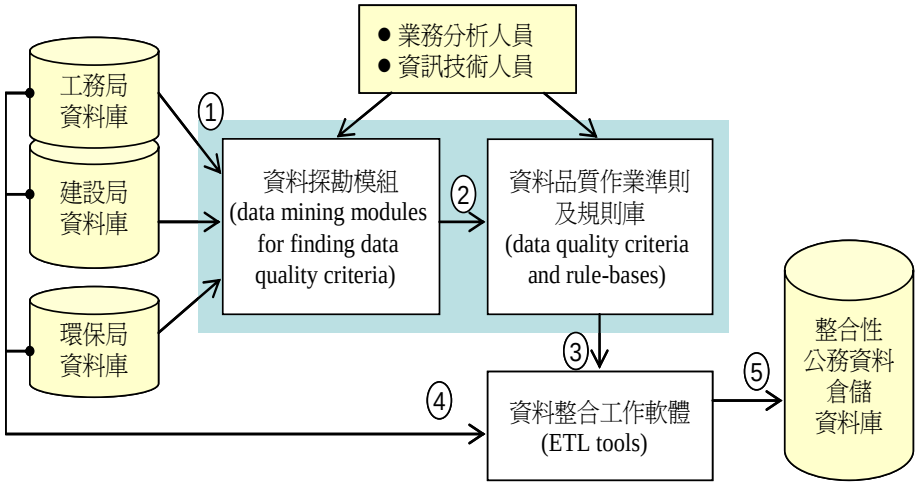

圖 3：結合資料品質管控的公部門資訊整合架構

2. 技術及工具方面：我們倡議運用資料探勘的方法，先針對現有的資料來源群體中，發掘出潛藏的資料品質問題，例如某工廠其所生產的產品中具有某些毒性化學物質，但工廠廠址卻坐落在住宅區，則其來源資料庫，也就是環保局或工務局(都市計畫部門)的資料可能有問題，這種資料一致性及完整性的品質課題，可逐步藉由資料探勘的方式發掘。但有時看似不合理的資料，卻是現實世界中真實的現象，例如當受到沙塵暴影響時，空氣中懸浮微粒值變得非常高，此時若依一般資料品質準則，容易將其誤判為異常值，此時可搭配詮釋資料方法加以註記，以確保真實呈現原始資料，這些作業準則可以構成資料品質規則庫，提供後續資料整合工作軟體（ETL tools）參用。各個資料來源的資料量非常龐大，如果單靠人工方式來檢核，所要付出的代價是相當昂貴的，因此採用自動化機制進行檢核，可以加速效率。其中，資料探勘就是一種能夠增進資料檢核偵錯率的方法，不過，選擇一個好的資料探勘演算法卻是這種方法最困難的一部分。我們建議程序如下：
   (1). 首先，由領域專家（業務分析人員）來辨識真實資料庫中的屬性、特徵與結構。
   (2). 辨識的結果將作為產生測試資料的參數，並將產生的測試資料存檔。
   (3). 由資料探勘和資料檢核專家（資訊技術人員）根據結構化歸納或資料偏差偵測結果，持續調整資料探勘演算法，直到令人滿意的評估結果產生。
   (4). 最後，由業務分析人員及資訊技術人員共同使用這種客製化的資料檢核工具來辨識整合性資料庫中的錯誤並提出正確資料值。
3. 作業流程方面：參照圖 3 所示，首先將資料來源群體作為資料探勘的輸入材料，而輸出的結果構成資料品質作業準則及規則庫，接著 ETL 工具軟體將上述步驟之規則庫作為資料品質的主要參用依據，當匯入資料時必須檢核其資料品質是否符合規則庫所定規則，而後再整合為資料倉儲。依此作業流程進行，我們認為可以改善以往投入大量人力檢核資料品質反致績效不佳之情況。以環保機關對環境監測資料品質管控為例，經過上述程序，我們可以獲得類似下列表現方式的資料品質推論規則，

用以輔助提昇資料品質參數：

1. 存取性的規則：例如討論驗證與授權的關係，可以有下列關係：

   若資料沒有經過驗證與授權，則須嚴格限制其存取的對象

   可以表示成下列規則：

   IF Certificated Data (A , No) and Authorized (A, No)

   THEN Accessibility (A) = 0.9

   其中，0.9 表示存取限制嚴格。若 0.0 則表示不必限制。

2. 詮釋性的規則：例如討論測站名稱，可以有下列關係：

   若相同座標位置之測站資料，測站名稱不同時，則其詮釋性較低

   可以表示成下列規則：

   IF XYDim (A, 00.00) and XYDim (B, 00.00)

   THEN Interpretability (A, B) = 0.2

   其中，0.2 表示詮釋性非常低。若 0.9 表示詮釋性非常高。

3. 議題關聯性的規則：例如討論空氣品質即時監測資料時，可以有下列關係：

   若監測時間與目前時間相差一小時以上時，其「即時性」之關聯就有誤用之虞，則其議題關聯性較低

   可以表示成下列規則：

   IF TimeConstraint (Airquality, Realtime) and

   CurrentTime – MonitorTime(Airquality) < 1 hr

   THEN Contextualability = 0.9

   其中，0.9 表示議題關聯性非常高。若 0.1 表示詮釋性非常低。

4. 可信度的規則：例如討論工廠之屬性彼此間之資料一致性，可以有下列關係：

   若毒性化學工廠的土地使用分區座落在「工業區」，則工廠資料可信度高

   可以表示成下列規則：

   IF Attribute (Factor A, Baneful) and Region(Factor A, Industry)

   THEN Belief = 0.95

   其中，0.95 表示可信度非常高。若 0.1 表示可信度非常低。

上述方法將使環境資料庫中之資料品質得以量化性描述，有助於系統管理者及使用者掌握資料與真實環境之差距。其次，透過這種規則性的描述，可以讓各來源資料庫資料匯入環境資料庫前之篩選、過濾效果得以有效提昇。本文所提的方法可以比目前的方式得到更嚴謹有效的資料淨化結果。

5. 結論

隨著「電子化政府」工作之推展，未來整合型之公部門資訊服務勢將成為政府便民服務的業務主流，同時也是政府部門未來重要的施政項目，因此如何確保整合性資訊的資料品質，以輔助決策分析，將是一項重要工作。多個資料來源整合後，在資料品質的控管程序上，更須嚴謹面對時序特性，以避免因長時間延續而累積出無法控管的資料品質。同時，

整合後之資料品質要能「適於使用」於公部門不同單位，而且也能符合民眾的需求。本文所述方法可以使環境資料庫所儲存之資料品質作定量性之表現，而非傳統之定性描述，對民眾掌握環境資訊極有助益。

在未來研究方面，我們將著手針對以資料探勘演算法發掘資料品質特性及規則方面進行實證性的實作與驗證，並且將就實際的公務資料整合系統進行測試。其次，由於網頁及 XML 資料愈來愈普及，有淩駕傳統資料庫的趨勢，我們也將探討這類資料整合時，其資料品質與傳統資料型態的不同，並尋求擴大本文倡議方法的適用範圍。

參考文獻

1. Ballou, D.P. and H.L. Pazer, "Cost/Quality Tradeoffs for Control Procedures in Information Systems," *International Journal of Management Science*, 15(6), pp.509-521, 1987.
2. Berndt, D.J. et al., "Healthcare Data Warehousing and Quality Assurance," *Computer Society* IEEE, 2001.
3. Dasu, T. and T. Johnsou, *Exploratory data mining and data cleaning*, Wiley, 2004.
4. Helfert, M. and C. Herrmann, "Proactive Data Quality Management for Data Warehouse Systems - A Metadata based Data Quality System-,"
5. Hufford, D., "Quality and the Data Warehouse," http://www.datawarehouse.com/resources/articles/
6. Huh, Y.U. et al., "Data Quality," *Information and Software Technology,* Vol. 32, No. 8, pp.559-565, 1990.
7. http://web.mit.edu/tdqm, 2005
8. http://www.cpro.com.tw/channel/news/
9. IBM White Paper, *The IBM Information Warehouse Solution – A Data Warehouse Plus!*, IBM Corporation, 1996.
10. Kesh, S., "Evaluating the quality of entity relationship models," *Information and Software Technology,* Vol. 37, No. 12, pp.681-689, 1995.
11. Luebbers, D. et al., "Systematic Development of Data Mining-Based Data Quality Tools," *Proceedings of the 29th VLDB Conference,* Berlin, Germany, 2003.
12. Naumann, F. and U. Lsesr, "Quality-driven integration of heterogeneous information systems," *Proceedings of the 25th VLDB Conference*, Edinburgh, Scotland(1999).
13. Wang, R.Y. et al., "Toward Quality Data：an Attribute-based Approach," *Decision Support Systems* Vol.13, pp.349-372, 1995.
14. 朱雨其，楊珊珊，楊鍵樵， "資料倉儲環境中資料品質的評估機制，" 中華民國八十八年全國計算機會議， pp. A284-A291.