國立政治大學資訊科學系

Department of Computer Science

National Chengchi University

碩士論文

Master's Thesis

在語意式雲端環境上資料交換的保護

-以醫療病例為例

Data Exchange Protection in the Semantic Data Cloud-Medical Health Record as an Example

研究生：黃雅玲

指導教授：胡毓忠

中華民國一百零二年三月

March 2013

在語意式雲端環境上資料交換的保護-以醫療病例為例

Data Exchange Protection in the Semantic Data Cloud-Medical Health Record as an Example

研 究 生：黃雅玲　　Student：Ya-Ling Huang

指導教授：胡毓忠　　Advisor：Yuh-Jong Hu

國立政治大學
資訊科學系
碩士論文

A Thesis
submitted to Department of Computer Science
National Chengchi University
in partial fulfillment of the Requirements
for the degree of
Master
in
Computer Science

中華民國一百零二年三月

March 2013

在語意式雲端環境上資料交換的保護

-以醫療病例為例

摘要

近年來，隨著網路資訊的普及和個人隱私意識的提升，個人識別資料的分享和保護已經變成重要網路研究議題之一。資料存放在雲端環境上，因不同資料來源之間結構上的差異，我們將會面臨到如何建立 PII 的分享和保護準則，以確保滿足資料擁有者的隱私偏好。

本研究使用雲端運算做為多個資料源執行資料交換的環境，其好處在於擁有大量的網路存放空間、大幅降低了資料管理成本。舉例來說，我們可在雲端環境上存放大量的醫療資料，當使用者欲查詢不同來源的醫療資料時，可透過資料交換的方式從單一入口取得，不需兩端分別進行查詢，並利用雜湊函數的方式來處理個人資料匿名性的辨識，主要是在不揭露個人資料的狀況下，仍然可以判斷資料是否為同一筆資料。

另外，由於本研究以個人隱私資料做為研究之情境，所以在隱私保護上會以存取控管規範(Access Control Policy，ACP)、資料處理規範(Data Handling Policy，DHP)和資料釋放規範(Data Release Policy，DRP)，三種規範來說明資料保護、資料交換和資料揭露的過程。

最後，本研究主要是使用具有語意化技術本體論和規則的知識表達來解決跨資料源的資料交換，除了理論塑模之外並且利用兩家醫院的情境來加以展示。

Data Exchange Protection in the Semantic Data Cloud-Medical Health Record as an Example

Abstract

Personal Identifiable Information (PII) sharing and protection have become one of the most important research issues for the Internet, especially for cloud computing infrastructure because of its widespread services. The challenge of sharing structured PII data in the cloud is to address the structure differences between data sources. In addition, we face the problem for how to establish the PII sharing and protection principles to ensure that its disclosure criteria are satisfied with the data owners' privacy policies.

In this study, we use cloud computing simulated environment as a multiple data sources exchange platform because of its spacious and cost-effective reasons. For example, we can outsource tremendous amount of electronic health record (EHR) administration services in the cloud without too much cost. Besides, data exchange provides a single point of data access instead of having accessed in a separate entry. We apply hash function of de-identifiable partial PII to enable record linkage services between data sources for data exchange without losing data owners' privacy.

Three types of privacy protection policies are proposed to achieve the data exchange and protection objectives in the multiple sources data cloud. They are Access Control Policy (ACP), Data Handling Policy (DHP), and Data Releasing Policy (DRP). These policies are represented as OWL-based ontologies and enforced as Logic-Program (LP)-based rules. We demonstrate the privacy protection policy concepts for medical record exchange between two hospitals.

致謝

首先感謝胡毓忠教授的指導，在研究的這段期間胡老師總是充滿熱忱的教導，給了我研究方向和建議，總是不厭其繁的與學生討論，讓我學習了許多東西，學生滿懷感謝。此外感謝兩位口試委員，葉慶隆教授與徐國偉教授，經由兩位教授的提點，學生的論文能更趨進於完整。

再來要謝謝實驗室的同學們，不論是已經畢業的峻展學長和協達學長在研究方向都幫了我不少忙，同時也不時提醒我論文的進度；而穩男學長、迪嶸學長和國平則是我互相討論的好朋友，不是論學業上還是生活上，都給予我很多的幫助；還有采衣學妹和容甄學妹適時的鼓勵，讓我倍感窩心；也感謝教學發展中心數位學習組的各位夥伴們，瑋芸、宗瑋、筱慈讓我在研究所期間有個溫暖的工作場所，也謝謝他們陪我度過這些年的喜怒哀樂。

最後，謹以此文獻給我摯愛的雙親。

目錄

圖目錄

表目錄

第一章

導論

1.1 研究動機

近幾年雲端運算流行，網路服務變的更加方便且具有彈性，資料不限於只存本機端，而是放上雲端，並且提出了“Everything-as-a-Service” (XaaS) 的概念。這麼做的好處不但大大降低了資料管理成本與傳統的紙本流通更為快速的分享與流通。然而背後會產生更多的問題。例如：資料放在雲端的安全性、隱私性、私有雲與私有雲之間資料流通的可行性問題等等。語意式雲端的產生，除了原本雲端服務的彈性與方便外，更提供了豐富的表達邏輯、資料建模等功能，進一步的擴展到互聯網及標籤雲的運用上，大大的提升網路效益[1]。

由於個人隱私意識的提升，了解到資料保護的重要性，不論是傳統資料庫服務或是雲端資料庫服務皆需要有妥善的處理機制，來避免資料的散播、流失或被不當的揭露或使用。由於現今企業都慢慢的將資料存放在雲端上[2]，顯然這已成為了一種趨勢，因此，當資料要存放在雲端，又想要得到保護隱私的目的，最常見的想法即是讓雲端廠商看不懂或看不見你的資料，所以將資料做加解密的動作並且將資料分開存放是目前雲端運算保護的方式。基於此，本論文將此項技術進行多重結構化資料源的委外資料交換，當使用者進入雲端中，其查詢的資料來自不同資料來源時，必須先進行身分上的確認才可獲的資料存取的權限，若違反此原則而企圖取得資料則視為違反隱私條例；一旦驗證了使用者身分，則不同資料來源就需要做資料之間的交換，來達到資料共享的效益。

資料交換的保護主要是在雲端資料庫中，當雲端資料庫收集的資料不相同時，因外在條件的因素需要到另一雲端資料庫去查詢資料時，透過資料交換的方式進行資料的傳輸，在此將保護每一個個人資料都是匿名性的比對。本研究利用兩家

醫院情境來加以展示上述說明，讓合法且經授權的使用者可以方便的在雲端內查詢適當的資料。

1.2 研究目的

本研究主要目的是當不同機構或單位收集個人資料時，將資料存放在雲端運算環境上，使用者經由身分認證後能夠透過單一窗口查詢到雲端環境上多個資料源的資料，達到資料分享的目的；但資料分享的過程中，必須謹守個人資料保護法的規定，避免違反個人隱私。本論文主要研究方向如下方所示:

1. 規範框架的整合:本研究假設相同雲端環境上有兩種不同資料來源，而每一個資料來源有各自的存取控管規範、資料處理規範和資料釋放規範，透過本體論整合，將資料做適當的釋放，達到資料分享與保護的目的。
2. 不同來源的資料處理與對齊:各家醫院將資料存放到雲端環境的前提下，當兩家醫院各別將其中的資料源進行資料之間的交換時，為了確保個人資料隱私的保護，資料的接收方必須無法進行個人資料還原的動作，但為了讓雙方資料能夠對齊與整理，本研究加入了匿名性的對齊方式。換句話說，我們在維護個人資料隱私的情況下，整合雙方的資料，可以完成有意義的比對分析目的。

1.3 各章節概述

第本文第一章節為對整篇論文做一個概要性的介紹，包括研究動機、研究目的，以及各章節的概述。第二章是研究的相關背景說明；第三章則是對於相關研究介紹；第四章會完整的描述系統架構作，並且針對存取控管規範、資料處理規範、資料釋放規範與資料交換落實部分詳述。第五章為研究方法的模擬驗證部分說明；第六章則為總結本研究。

第二章

研究背景

2.1 雲端運算

雲端運算的概念是將許多的主機串接在一起視為一個大型主機來做控制中心，是一種新的商業模式。好處在於簡單的獲取了雲中的服務，快速、高效率地完成了工作；而他們獲取的服務類型不盡相同。以下我們將針對雲端運算提供的服務類型和方式，為雲端運算分類[3]。

- 公有雲:表示該組織建立雲端資料中心後，不僅僅讓自己使用也專門規畫給其他人使用，透過收取費用的方式將雲端資料中心的服務與資源分享給其他對象。
- 私有雲:站在雲端資料中心的建立與管理者來講，整個雲端服務從硬體、軟體到管理等，都是該組織自己負責管理維護，等於是組織自己建置雲端硬體、軟體並加以管理使用。
- 混合雲:就是由企業建置雲端運算的系統架構，完成內部私有雲，再視需求和使用量，訂用外部公有雲的服務，打造更具彈性而強大的雲端環境。

本研究主要是利用資料交換的特性，在雲端環境上做資料之間的傳輸，所以將採用私有雲為主要的雲端環境，因為私有雲可能會有特定資料利用資料交換保護的問題產生；若使用公有雲，因其主要目標是分享給其他對象，則不需使用到資料交換保護。

2.2 資料交換 vs.資料整合

資料交換 (Data Exchange) [4] [5] 和資料整合 (Data Integration) [6]皆是資料處理時兩個重要的方式。兩者主要目的皆是將大筆資料做整理、分類，使得使

用者在查詢或使用時可以較快速的得到相符結果。

資料交換可以當作是一種 Global-Local-As-View (GLAV)的形式，可以動態的對應，主要是要對目標資料庫進行查詢，再從另一來源資料庫進行 Schema Mapping，如圖 1。步驟一:先將查詢資料傳送到 Source Schema 進行比對；步驟二:將從 Source Schema 端查詢後的資料回傳到目標資料庫；步驟三:在 Target Schema 內部進行第二次資料比對處理；步驟四:將步驟三得到的結果存放在一個實體資料庫內。

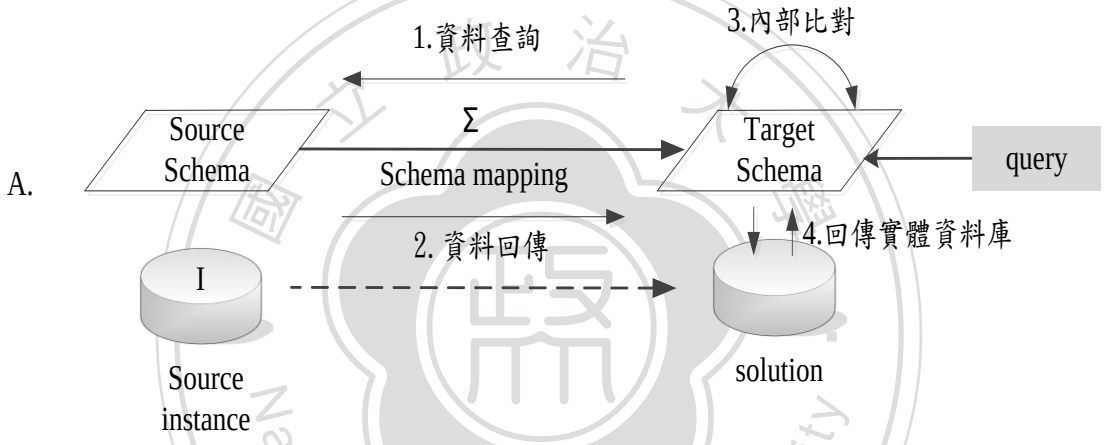

圖 1、資料交換

資料交換特色在於在一方資料查詢時，經特定規則可以到另一方取得相符合的資料，是一次式的查詢，增加查詢的彈性度，資料實體化的特點。資料整合主要是將多方資料來源提取出來做整理，匯集在同一個資料集內，如圖 2。但兩者皆須面臨 Schema 整合的問題。

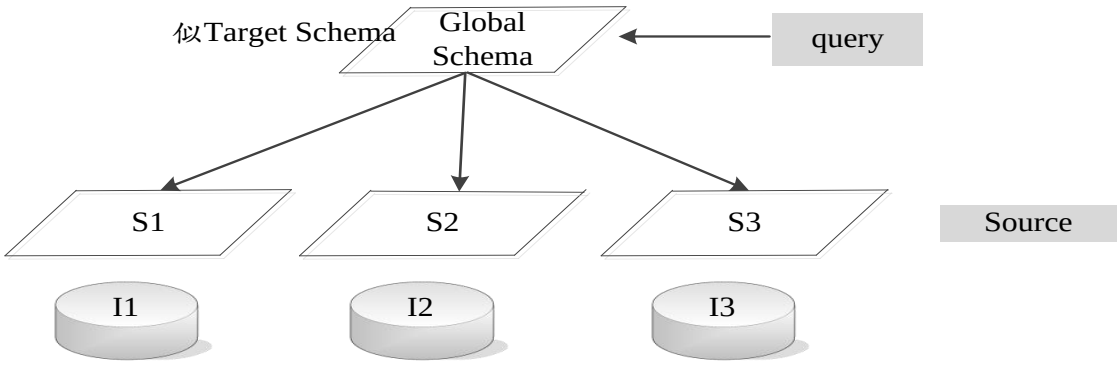

圖 2 資料整合

資料交換和資料整合的第一個差異點在於而資料交換像是做資料傳輸的動作，主要是針對兩個資料源之間的分享；而資料整合像是將資料統整起來，進行分享。第二個差異點在於表示式的不同，資料整合的表示式為 (G,S,M)，其中 G 為 Global Schema，S 為 Source Schema，M 為 G 和 S 之間的關係；資料交換的表示式為 $(\mathbf{S}, \mathbf{T}, \sum \boldsymbol{st}, \sum \boldsymbol{t})$，其中 S 為 Source Schema，T 為 Target Schema，$\sum \boldsymbol{st}$為 S 和 T 之間對 應的關係，$\sum \boldsymbol{t}$為 T 內部對應之關係。在$\sum \boldsymbol{t}$對應時還需要判斷是否為 Weakly Acyclic[4]。Weakly Acyclic 主要是保證在 Polynomial Time 內會有解，不會造導致此次查詢有無法終止的現象發生。

2.3 個人資料查詢的合理使用

本研究將以個人資料為資料來源進行資料交換，而個人資料是受到個人資料保護法[1]所保護的。目前世界各國皆有相似的法律規範，如歐盟的個人資料保護令中明文規定”企業在蒐集資料以前，必須取得當事人同意，個人可以隨時要求更正、甚至刪除資料，並且個人資料不可傳送到非歐盟國家”[7]。而台灣的個人資料保護法中有明文規定敏感性資料是不可以收集、處理或利用。另外，資料的

[1]台灣全國法規資料庫<<個人資料保護法>>http://law.moj.gov.tw/LawClass/LawAll.aspx?PCode=I0050021

收集也是需要告知資料提供者收集目的、使用時間、地區、方式和對象，本研究會以使用者身分和目的來判斷是否可以查詢部分個人資料，實現給適當的人適當的資料之目的。

一般而言，使用者查詢分成兩類 Subject Based Query (SBQ) 與 Pattern Based Query (PBQ) [8]，主要是在查詢的條件背景與時機的不同時所產生出來的。本研究將充分利用兩者 Query 之間的特性，使得查詢時可以更加明確的知道需要哪些資料，來確保個人隱私無法任意被揭露，在適當的時機給適當的角色適當的資料。

1. SBQ：使用者進行業務上的需求，盡可能的縮小範圍而得到確認的唯一解，像是醫生查詢就診病患的資料。
2. PBQ：使用者是為了統計分析數據時所下的查詢指令，例如:醫生在分析資料的情況下想要得數據資料來做為研究之用途時，運用 PBQ 可以讓查詢結果無法辨識特定人選，可能只會得到概括值，像是幾歲到幾歲之間可能是某些病症的高危險群。

文獻[9]分類了四種資料擁有的資料型態 (DataType) : Identifiers、Quasi-Identifiers、Confidential Attributes 和 Non-Confidential Attributes。

- Identifiers:單一欄位即可完全識別一個人身分 Ex : ID、SSN。
- Quasi-identifiers:多個欄位組合即可提升識別一個人身分的機率 Ex : Gender、ZIP 和 Birthday[10]。
- Confidential attributes:違反隱私的欄位 Ex : Disease 或 Cost。
- Non-confidential attributes:不屬於上述範圍的欄位 Ex:Race。

只有當 Identifiers 和 Confidential Attributes 或者 Quasi-Identifiers 和 Confidential Attributes 同時揭露時才會違反隱私，在其他種組合方式下並不會，例如只揭露 Disease 和 Cost 是不違反隱私的情況。因此本研究在進行資料釋放的時候，也是依照此準則針對 Identifiers 和 Confidential Attributes 或者

Quasi-Identifiers 和 Confidential Attributes 的部分進行保護，預防隱私的侵犯。

2.4 本體論

本體論 (Ontology)最早的概念是從哲學而來的名詞，而根據 W3C (World Wide Web) 對本體論的定義為：「本體論是用來描述與表示各種領域的知識。」簡單來說,就是我們可以利用本體論來架構一個領域知識 (Domain Knowledge)，並進一步分析此領域中各種概念的關係。由知識的概念定義、屬性、實體、及關係的集合體，建構這些元素則需要一套發展程序。

- 概念 (或稱為 Class/Set/Concept)：表示本體論中對某類實體的集合或概念。
- 屬性 (或稱為 Property/Slot/Role/Relation)：表示本體論中實體與實體或概念與概念之間的關係。
- 實體 (或稱為 Individual/Object/Instance)：表示本體論中的個別真實例子。

使用本體論語言來建構本體論，本體論語言允許使用者設計出領域模型 (Domain Model)的明顯與形式的概念化。所要的基本需求是：定義明確 (Well-Defined)的語法：機器處理資訊的必要條件。正規語意 (Formal Semantics)：精確地描述知識的意義且具有語意的方便性、有效推理的支援、充分表達威力。

本研究中本體論使用的時機為使用本體論作為資料的儲存庫，我們透過本體論的知識描述特性，使得詞彙之間的關係有解釋的能力；規則的運用可以推論出隱含的資訊，使得資料能更具有彈性、容易處理與分享，往後只需要透過更改屬性與關係，即可建構出具共享性和再用性的知識本體。

2.5 資料整合-資料庫 vs.本體論

在傳統關聯式資料庫中，資料整合是一門大學問，其主要著重在 Schema 的整合，也就是對不同點的 Local Schema 事先利用查詢語法 SQL 產生一組 View 表示為該 Local Data Source 可提供整合的部分;接著另外產生一個 Global Schema，

作為使用者查詢時，Global Schema 會去對查詢語法在每一個 Local Data Source 進行查詢語法改寫的部份，下到各個點去做查詢，最後將結果回傳給 Mediators[11]，再傳給使用者。Levy 的研究中提到了三種方式[6]：

1. Local-As-View (LAV): Local 的 Relations 或 Concepts 對應到 Global 的 View 或 Queries。
2. Global-As-View (GAV):剛好與 LAV 相反，Global 的 Relations 或 Concepts 對應到 Local 多個 View 或 Queries 組合產生的。
3. Global-Local-As-View (GLAV): Global 的 Views 或 Queries 對應到 Local 多個 Views 或 Queries。

此三種方式主要差別在於 Global 與 Local 的對應角度，像 GAV 是以 Global Schema 為準則，所有 Local Schema 必須要想辦法產生 View 與 Global Schema 能夠對應；反之 LAV 則是以 Local Schema 為準則，所有 Global Schema 的 Relations 必須要想辦法產生 Relations 與 Local Schema 能夠對應；GLAV 則富有最彈性化的設置 Global Schema 和 Local Schema 可以相互對應。基於此，本研究資料交換的技術時採用 GLAV 的方式，使得不同來源的資料可以動態對應且具有彈性。

另外，在本體論的資料整合中，主要注重於 Class 與 Property 彼此之間對應的關係，也就是概念(Concept)的整合，可以表示為特定知識領域中抽象概念的一種階層式框架。對應的方法有 Mapping[12]、Merging 和 Alignment[13]。

- Mapping 主要是因為單一本體論資料有限，所以資料量不大，希望可以透過 Mapping 的方式，對應到另一個本體論上，或者是多個本體論，可以相互對應。這類的方式與 Merging 有異曲同工之妙。
- Merging 是將兩個本體論合併成一個大型的本體論，與最大的差異在於 Mapping 在整合完成後依然會考慮到原本的本體論；而 Merging 是將兩個本體論整合後只考慮整合完成後的唯一本體論，不去理會本來被整合的本體論，其中 PROMPT 方法[14]是由史丹福醫療資訊學系所發展的，其合併兩個知

識本體需要來自相同的領域。PROMPT 合併本體論方法程序與 FCA-Merge 方法一樣為互動性的、人為介入的方式。

- Alignment 的作法上是假設有兩個本體論 A 和 B，以 A 作為主體，而將本體論 B 對應到本體論 A 上。

以上三種方式雖然在概念上不相同，但其實大多的方法都強調在半自動化的整合。如圖 3。

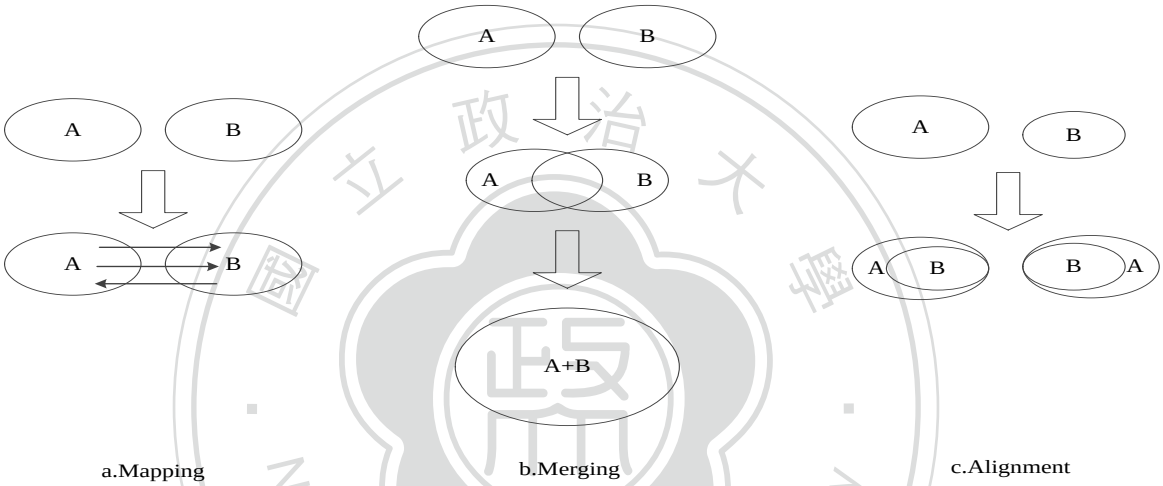

圖 3、三種整合方式

本研究設計三種規範的本體論，每一個資料來源皆有三個本體論產生，當使用者欲查詢不同資料來源時會執行資料交換的步驟，而本研究將會透過本體論 Merging 的對應方式去實現不同資料來源本體論規範之間框架的整合，目的是為了讓兩邊的不同結構的資料在資料交換時有可以對應的 Schema，達到資料交換可以取得不同資料來源的特點。

2.6 雜湊函數

雜湊函數 (Hash Function)常見的有 MD5 (Message-Digest Algorithm 5) [15]、

SHA (Secure Hash Algorithm)、MAC (Message Authentication Code)和 HMAC (Hash-Based Message Authentication Code)，綜合以上，不論是何種Hash Function，都具備下列幾點特性：

- 輸入任意長度的訊息，產生固定長度的雜湊值輸出。
- One-Way Hash 之特性。
- 針對相同訊息進行計算，都會產生出相同結果。
- 雜湊訊息是無法還原成原訊息，因此演算法的設計上必須是不可逆。

本研究的資料為醫療資訊，也因為醫療資料屬於個人隱私的一部份，所以不同資料來源之間可能會有重複性的產生，例如:一個病患去可能到多家醫院就診。在資料交換的過程中，醫院是有權力不提供完整的個人醫療資訊，防止個人資料被還原而辨識為唯一人，在無法辨識的狀況下，查詢出來的資訊可能不夠精確，基於此，本論文將根據雜湊函數的特性，去實現匿名性資料的對齊，將個人資料經由 Hash Function 計算得到固定長的雜湊值，再依據雜湊值的結果，進行資料比對、刪除重複性資料，詳見 4.2.3 章節說明。

第三章

相關研究

3.1 隱私還原保護

在過去的文獻，針對隱私保護提出各種解決方式，其中在[10][16]提出了兩個情境資料，一個可顯示每位投票者的姓名、地址、郵遞區號、性別、生日，這些資料可以和醫療資料中的性別、郵遞區號、生日相互做連結，以至於人們可以利用上述的特性找到特地的個體，基於此，作者提出了資料保護的概念，其主要做法是將這種連結不明確化，就可以阻撓資料還原。另一方面，。K-Anonymity 意即欲將 Table 中的資料化為多個群組，每個群組在敏感屬性上的值皆相同，例如 Birthday, Gender, ZIP 為一個群組，而每個群組中有 K 個 Record，K-Anonymity 目的為將資料庫中資料表達到某一種隱私保護狀態，其主要的用意在於將敏感屬性欄位的 Re-Identification 可能性降到最低，舉例而言，若有使用者惡意的利用 Birthday 和 ZIP 與其他的資料表進行結合比對，進一步地找出某筆紀錄實際上是屬於哪個人，為防止上述的情形發生，K-Anonymity 主要的目的是要能夠消除這種可能性，以達到資料隱私保護的狀態

另外，資料的隱私防護在傳統的作法是對資料進行存取控制，意即只要認為是敏感資訊的欄位便將其值拿掉，這種方式或許是相當直覺且安全，但由於我們除了進行隱私防護之外，還須具備有後續的關聯能力；而當敏感屬性欄位整個移除時，敏感屬性欄位往往又是進行關聯時的關鍵欄位，所以這類的方式對隱私防護是不可行的。

上述方式都是希望能夠將個人隱私資料達到保護的功能，而本研究與之差別在於本研究利用雜湊函數實現個人資料匿名性的對齊，就算把敏感性欄位的值刪除，資料依然具有後續的關聯能力；而我們也可以防止惡意使用者在兩個不同資

料來源分別查詢後，再將資料進行比對，因本研究採用資料交換的特色是一次的查詢可以查詢兩邊來源的資料，當兩邊資料來源可能造成 Re-Identification 的時候，我們會進行資料保護的動作。避免敏感屬性欄位被 Re-Identification。

3.2 雲端委外語意式資料保護

雲端委外語意式資料保護[17]此論文在研究架構上，也是利用存取控管規範、資料處理規範和資料釋放規範這三種規範的合作、分工，將法律上資料隱私的保護與 Data Owner 的隱私偏好做一個結合，最後應用 Microdata 保護技術與 User 的使用情境做比對，實際落實適當的 Microdata 保護技術去保護隱私和滿足 User 查詢資料的目的。

在此篇論文中只針對單一資料源的 Microdata 進行揭露保護的分析，並沒有考量到在多資料源的環境下，Microdata 該如何保護、如何揭露才不會造成個人資料因資料來源的不同，而導致個人資料被揭露的可能性，像是透過多資料源的不同欄位資料查詢，可以辨別特定個人的身分。

本研究提出開放式的環境，在多個資料來源的情況下，要避免資料因外來資料的加入而違反了隱私，例如:在原本的資料上查詢 Gender、ZIP 和 Birthday 的釋放是不會違反隱私，但是若是再經由外來資料的加入，可能導致 Quasi-Identifiers 和 Confidential Attributes 的產生 (2.3 章節)，而造成資料被 Re-Identification。本研究資料的對應與流通是不會產生個人資料被揭露的狀況發生。

3.3 隱私資料的存取控管機制

隨著雲端運算的發展，個人資料被收集到一個集中的企業或者組織手中是不可避免的趨勢，因此資料隱私保護的需求變得相當重要。企業或組織必須擁有彈性且強大的隱私資料保護方式才能讓資料擁有者信賴並且託付資料。文獻[18][19]

中提出三種本體論的規範去落實隱私保護。

- Access Control Policy:負責管理資料或服務的存取或釋放。
- Data Handling Policy:讓資料擁有者設定自己的隱私偏好並且在資料流向合作企業時也一併流至合作企業，讓資料的使用滿足資料擁有者的預想。
- Data Releasing Policy:負責管理個人足以辨識身分的資料的揭露時機。

這三種規範的運用，使得企業的存取控管規範和資料擁有者設定的隱私規範都能一起被落實。但是文獻中的 Access Control Policy 並沒有加入隱私保護相關法律的概念，像是查詢的型態 Subject-based Query 和 Pattern-based Query；並在資料揭露的時候也沒有適當資料保護方式可以落實。而本研究在設計三種規範時，除了原先存取控管和隱私規範的概念外，加入了隱私保護相關的概念，並利用資料交換的概念，落實不同資料來源時的保護，避免 Re-Identification 的發生。

第四章

研究方法與架構

4.1 研究情境與架構

本研究以假設情境的方式，來實現不同來源的資料交換與匿名的個人資料保護。以 A 醫院和 B 醫院的醫療資料為例，兩者資料皆受到個人資料保護法所保護且存放在雲端環境資料庫中，研究中所提供 ACP、DHP、DRP 的流程如下:

首先使用者欲查詢兩間醫院來源的醫療病歷資料，會分為兩種查詢情境，第一種為醫生為了診治病患所下的查詢指令，此類的查詢為 SBQ；第二種為研究人員為了統計分析目的所下的查詢指令，此類的查詢為 PBQ。因為本研究是以查詢兩間醫院為主要情境，所以將第一種的查訊型態標記為 $SBQ_a \wedge SBQ_b$，主要可以運用在病患轉院的時候，讓接任的醫生更加了解病患在之前的醫院的情形，這樣不僅對醫生有好處，對病患而言也是一大福音；第二種的查訊型態標記為 $PBQ_a \wedge PBQ_b$，主要的目的是可以得到更多的醫療資料來做分析，讓研究人員可以得到更多的樣本數，增加統計結果的準確性。綜合兩種查詢型態，兩間醫院資料的取的都是透過資料交換的技術去執行。

本研究中 ACP 控管使用者的權限並一併判斷使用者的查詢型態，查詢型態是經由 ACP 的 MoreInfo 欄位判斷，MoreInfo 欄位表示使用者是否需要更多的資訊，若 MoreInfo 欄位若為 1 表示需要多的資訊，此時才可以使用 A 醫院與 B 醫院整合的 DHP 和 DRP；相反的，MoreInfo 欄位若為 0 則直接使用 A 醫院的 DHP 即可。在 DHP 中進行受到統計保護的資料比對與資料交換，利用雜湊函數 (MD5) 計算出個人資料的雜湊值，將兩間醫院分別計算出來的雜湊值做比對和刪除重複性。最後，傳送符合條件的使用情境與符合條件的資料給 DRP，根據文獻[9]中，說明各種情況的統計保護方式，釋放給使用者進行分析作業，整體架構如圖 4。

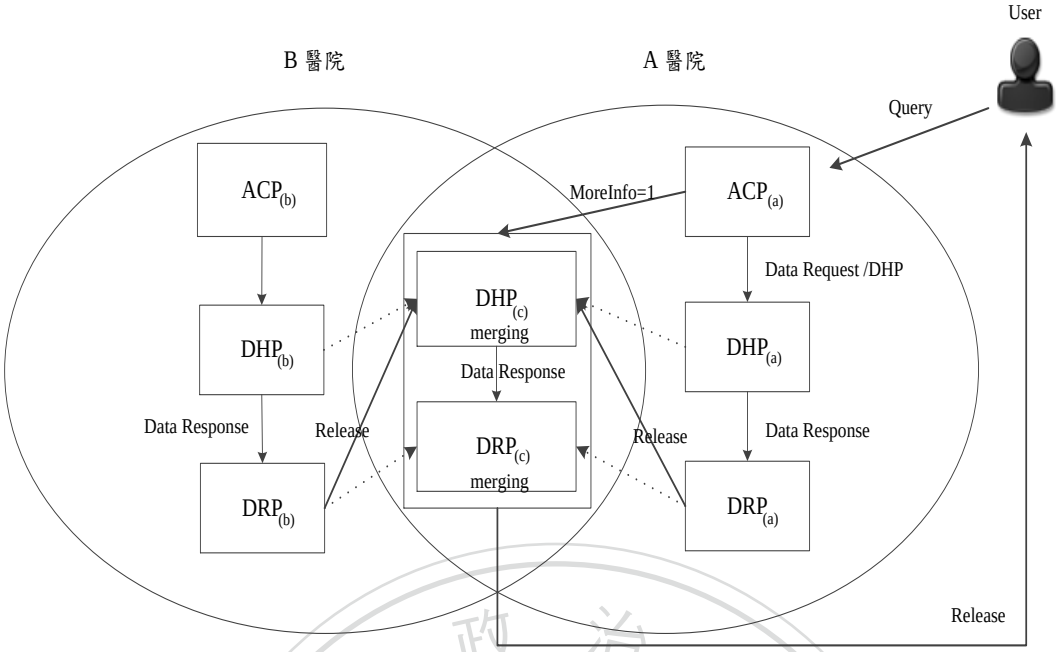

圖 4、研究架構流程圖

本研究架構主要分為兩個部分:不同資料來源之間資料交換的規則和處理隱私資料的對齊。解釋如下所示:

1. 不同資料來源之間資訊的流通是重要的議題，因為網路的發達，資料的取得可能來自不同地方，而本研究採取資料交換的方式處理不同資料來源之間的流通，主要是藉由此技術執行一次式的查詢，得到多方來源的資料，不必分開查詢，另外利用 ACP 的授權判斷後向 A 醫院和 B 醫院進行資料的查詢。

2. 隱私資料的對齊是需要被確立的，若資料有太多的重複性，卻沒有挑選出來，可能會造成統計資料的價值變低，我們主要是要兩邊 Data Sources 各自丟出資料來交換但是不能讓另外一方能夠還原 PII，但是又可以讓他進行資料的對齊與整理，例如整合同一個人的資料，來確保隱私保護的目的，但是卻可以完成有意義的統計分析目的。

本研究可以在不違反個人資料保護法的情況下進行資料交換與流通。使用相同雲端環境上，兩個不同來源的本體論來進行，本體論的建構對象為 A 醫院與 B 醫院資料庫，另外使用 Semantic Web Rule Language (SWRL) [20]作為規則語言，運用本體論加規則語言實現雲端環境上不同資料來源的資料分享和資料保護。

4.2 本體論建構

4.2.1 ACP、DHP 和 DRP 設計

根據 4.1 章節所述，本研究會有 3 種 Policy 來進行資料的存取控管與隱私條件的問題，每間醫院資料都會有內部設定的 ACP、DHP 和 DRP 存在，如圖 4 所示，三者功能如下:

- Access Control Policy (ACP):主要判斷使用者是否有權限能查詢資料並且授權其適當的查詢型態，包含可否去另一間醫院取得更多資訊的條件。假如 Access Control Policy 驗證使用者有權限使用查詢服務，則完成查詢模式的授權後便會啟動 Data Handling Policy。
- Data Handling Policy (DHP):主要落實不同資料來源資料交換的技術，作為資料交換後存放資料的中繼點，利用 Hash Function 來做隱私資料的對齊。
- Data Releasing Policy (DRP):根據查詢結果和授權的查詢方式釋放資料，並且判斷那些資料同時給予會違反了個人的隱私條件。

本研究中 ACP_a 是單一入口處的存取控管，而 ACP_a 不僅僅只判斷是否可以查詢 A 醫院的資料，還可以判斷是否可以一併查詢 B 醫院的資料，若許可將會傳送資料到 DHP_c，DHP_c 則是由 ACP_a 所啟動的，是由 DHP_a (A 醫院)和 DHP_b (B 醫院)整合而成的；DRP_c 則是由 DHP_c 所啟動的，是由 DRP_a (A 醫院)和 DRP_b (B 醫院)整合而成的。

4.2.2 ACP 設計

本研究將使用 A 醫院的 ACP 為主要查詢窗口，除了判斷是否可以查詢 A 醫院的資料，也判斷是否可以一併查詢 B 醫院的資料，如圖 5。MoreInfo 的資料欄位，說明使用者是否可以到其他醫院取得更多資料。

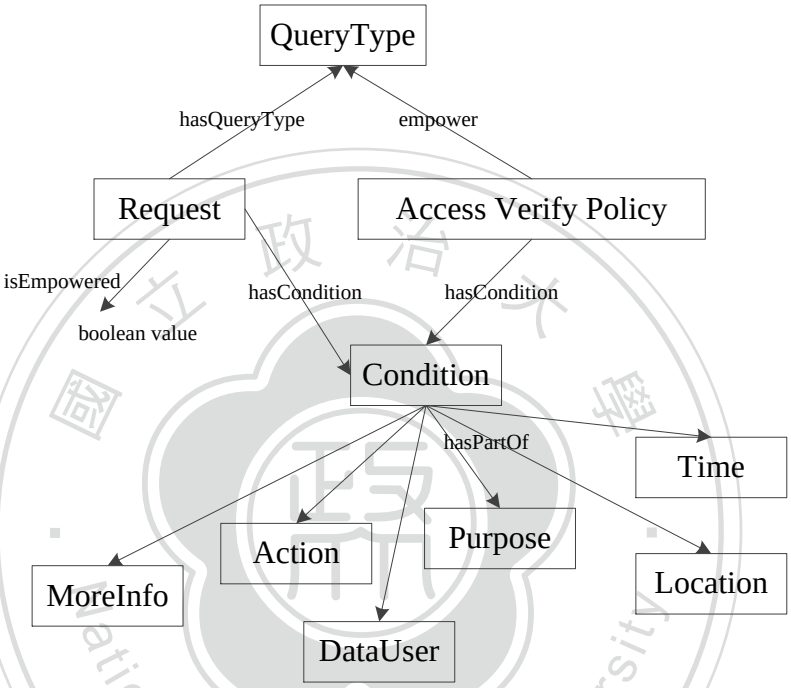

圖 5、ACP 設計

在此，Condition 是一個類別，每一個 Condition 的 Instance 都代表不同的資料使用情境，Request 和 AccessVerifyPolicy 都有各自的 Condition，SWRL 規則如下:

Request(?r) ∧ hasCondition(?r, ?c) ∧ Condition(?c) ∧ AccessVerifyPolicy(?avp)

∧ hasCondition(?avp, ?ac) ∧ Condition(?ac) ∧ sameAs(?ac, ?c)

∧ empower(?avp,?qt)∧QueryType(?qt)→ isEmpowered(?r, 1)∧ hasQueryType(?r,?qt)

---Rule 1

- Rule 1 規則中，主要目的是用來判斷使用者的查詢需求是否與 AccessVerifyPolicy 相同，根據 Request 和 AccessVerifyPolicy 的 Condition，就可以得知使用者是否有權限可以查詢資料以及使用者可以使用的查詢型態。當使用者的資料使用情境符合 AccessVerifyPolicy 中所規範，則會將 Request 的 isEmpowered 屬性所關聯到的 Boolean value 設定為 1，並且授權可以使用的查詢模式?qt。根據 isEmpowered 屬性所關聯到的 Boolean value 的值決定是否進入整合的 Data Handling Policy。如果 isEmpowered 屬性所關聯到的 Boolean value 的值為 0，則會不會進入整合的 DHP。

查詢型態的可能狀況如表格一所示，確認查詢型態後將進入到 DHP。本研究會有 4 種查詢型態的產生，當使用者為醫生在上班期間，想要進行讀取資料的動作，主要目的是為了診病患，地點在醫院裡，就可以讓醫生去選擇是否想要更多醫療資料，如果 MoreInfo 欄位為 Yes，則會給予 $SBQ_a \wedge SBQ_b$ 的查詢型態；若 MoreInfo 欄位為 No，則會給予 SBQ 的查詢型態。當使用者為研究人員為了研究統計分析想要進行讀取資料的動作，在適當的時間與適當的地點的情況下，一樣可以讓研究人員去選擇是否想要更多醫療資料來增加資料準確性，如果 MoreInfo 欄位為 Yes，則會給予 $PBQ_a \wedge PBQ_b$ 的查詢型態；若 MoreInfo 欄位為 No，則會給予 PBQ 的查詢型態。

一、查詢型態

DataUser	Action	Purpose	Location	Time	MoreInfo	QueryType
Doctor	Read	Diagonosis	Hospital	WorkTime	Yes	$SBQ_a \wedge SBQ_b$
Doctor	Read	Diagonosis	Hospital	WorkTime	No	SBQ
Researcher	Read	Analysis	Research Center	WorkTime	Yes	$PBQ_a \wedge PBQ_a$
Researcher	Read	Analysis	Research Center	WorkTime	No	PBQ

4.2.3 DHP 設計

為了使資料交換後，A 醫院能夠得到 B 醫院的資料，包含 A 醫院本身沒有收集的資料，我們必須將兩邊的 DHP 整合，以利資料的存放。本研究使用 PROMPT[14]作為 DHP 的本體論整合方式，PROMPT 在本體論綱要(Schema)的細部整合方法主要是採用字串比對與圖狀架構(Graph)對照的方式將兩者本體論合併，而合併就是將兩個本體論合併成一個大型的本體論，而先前的兩個則不再使用。在比較兩個本體論時，PROMPT 會依照字串比對所設定的參數以及類別在整個本體論中的圖形架構位置來給與使用者合併的建議。

DHP_c 是由 DHP_a (A 醫院)和 DHP_b (B 醫院)整合而成，如圖 6、圖 7、圖 8，分別表示 A 醫院、B 醫院和兩間醫院整合後的圖示，DHP_a (A 醫院)和 DHP_b (B 醫院)主要差別在於 Data 收集不同的欄位的不同。經由 Access Control Policy 驗證授權完後，假如 Request 的 isEmpower 屬性值為 1 代表成立，則啟動 Data Handling Policy，反之則不進行。

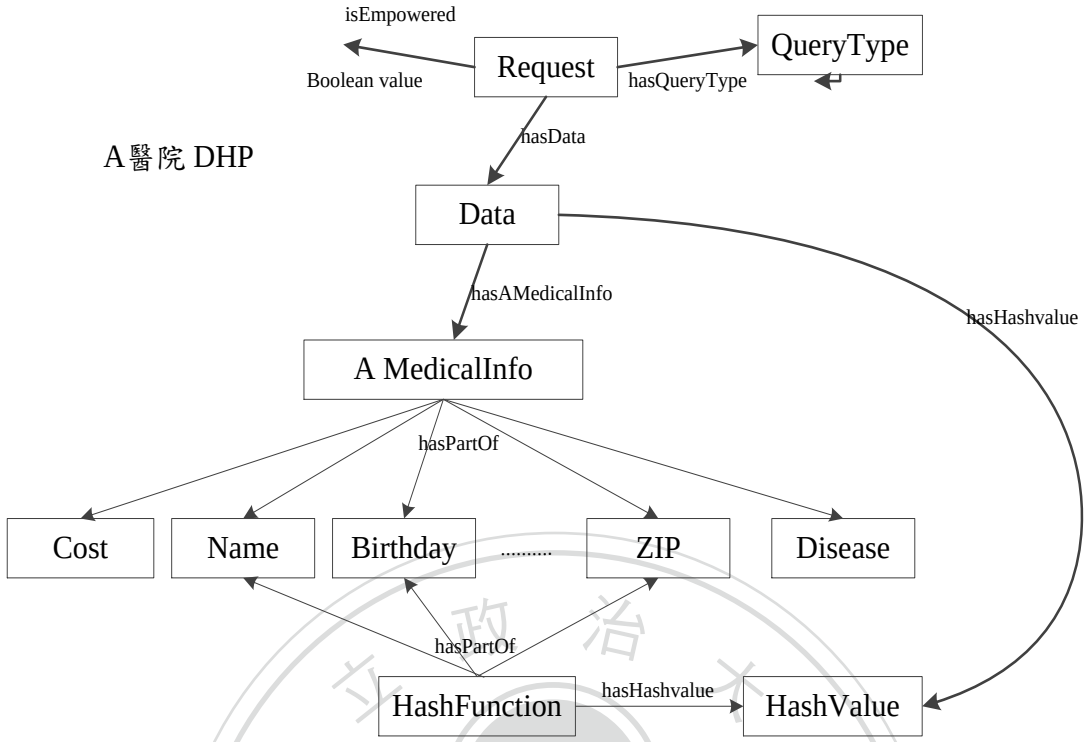

圖 6、A 醫院的 DHP

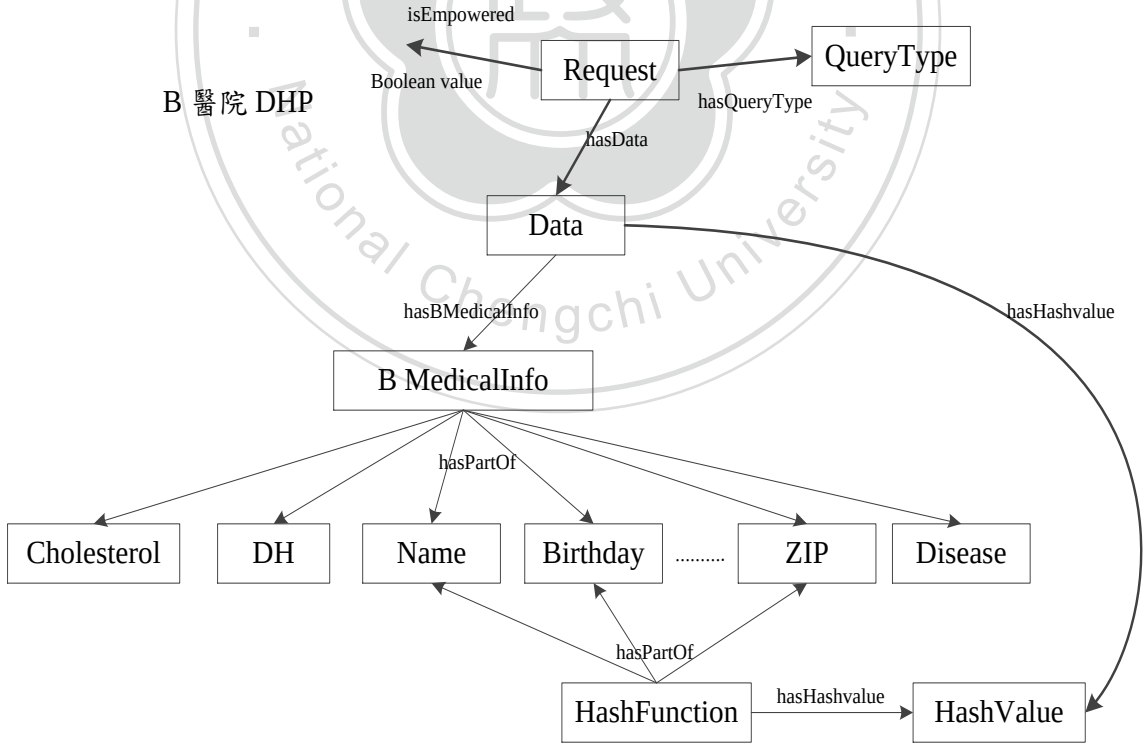

圖 7、B 醫院的 DHP

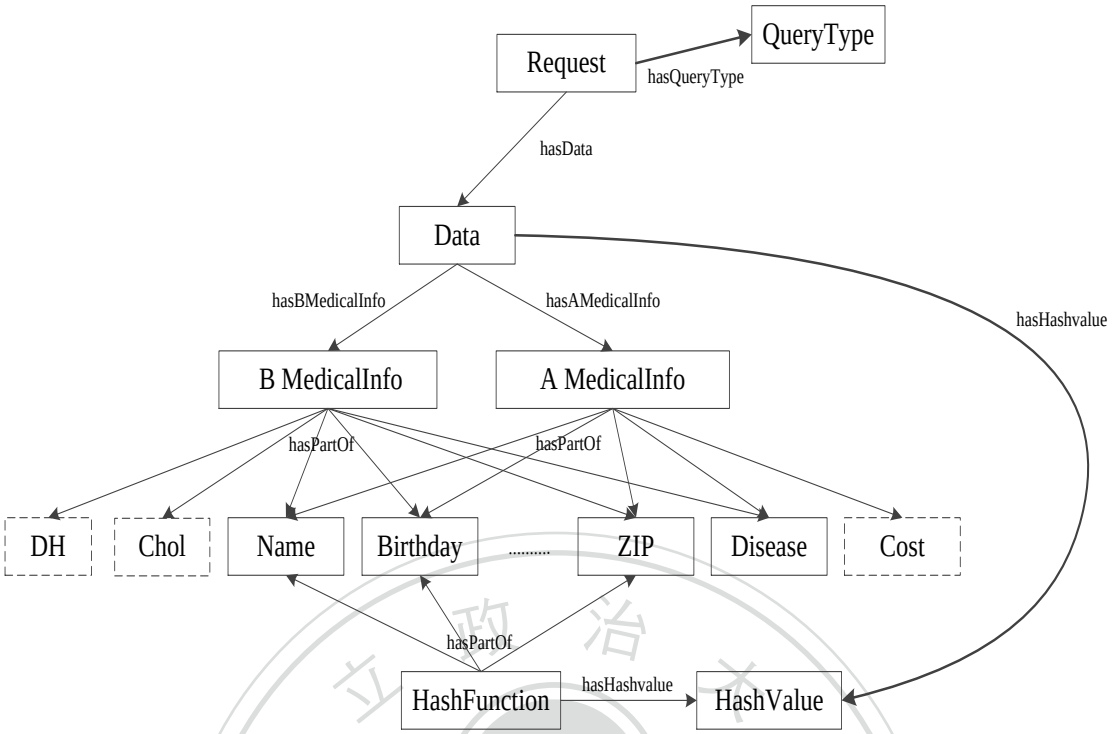

圖 8、兩間醫院 DHP 整合後

DHP 整合主要目的除了將兩邊的 Schema 做整合，落實資料交換，也因為兩邊資料可能收集不同的欄位，所以必須要有存放的相對應欄位。另外是判斷不同資料來源的資料，各別經由相同的雜湊函數處理後所產生的雜湊值來比對兩邊資料是否一致，整合後會根據雜湊函數產生的雜湊值來進行匿名性的個人資料的對應，確認重複性資料的刪除。

4.2.3.1 DHP 中資料交換的表示

根據上章節所說，本研究 DHP 主要功能為使用資料交換的方式執行兩間醫院醫療資料之間資料流通和分享，並加入了雜湊函數的概念進行匿名資料的比對。首先了解資料交換的表示式設置$(S, T, \sum st, \sum t)$，其中 S、T、$\sum st$和 $\sum t$之關係表示如表格二所示。

二、資料交換關係表示

	描述	備註
S	Source Schema	本研究假設為 B 醫院病歷資料庫
T	Target Schema	本研究假設為 A 醫院病歷資料庫
$\sum st$	S 和 T 之間對應的關係	
$\sum t$	T 內部對應之關係	仍需判斷是否為 Weakly acyclic

在進行資料交換之前，我們必須先知道 S 中 Source Schema 的內容和 T 中 Target Schema 之內容，根據表格二表示，S 為 B 醫院病歷資料庫；T 則為 A 醫院病例資料庫；假設兩間醫院的收集內容主要包含 Name、Gender、ID、Disease…等之欄位，差別於 B 醫院有 DH (Day in Hospital)、Blood Pressure…等欄位，而 A 醫院則無；A 醫院有 Cost 和 Doctor 欄位，而 B 醫院則無，B 醫院與 A 醫院收集欄位之差異，如圖 9 所示。

B 醫院	A 醫院
1.ID	1.ID
2.Name	2.Name
3.Birthday	3.Birthday
4.Gender	4.Gender
5.ZIP	5.ZIP
6.Disease	6.Disease
7.Medicine	7.Medicine
8.DH	11.Cost
9.Blood Pressure	12.Doctor
10.Cholesterol	

圖 9、A 醫院與 B 醫院收集欄位與編號

本研究情境的資料交換的步驟是先由 Target Schema A 醫院病歷資料庫下查

詢指令，執行 Mapping Rule ($\sum st$)的對應，找到 Source Schema 中與查詢條件相符的資料欄位，進行資料回傳到 Target Schema 的動作；再根據回傳的資料向 Target Schema 做第二次對應，而 Target Schema 內部可能被分割成許多不同使用者可以使用的資料庫，此時將利用 Mapping Rule ($\sum t$)去連接不同類別間相互的對應關係及限制，以確認是否會產生 Weakly Acyclic [3]。

本研究將資料欄位分成三類，如表格三欄位類別分類表，第一類為 Source Schema 和 Target Schema 內皆有的可相互對應之欄位，本研究有 Name、Gender、Birthday…等；第二類為 Target Schema A 醫院內部進行角色分類後研究人員可顯示的欄位，本研究有 Disease、Cost、Gender 和 Medicine 四種欄位；剩下來的欄位本研究將之統一歸納為第三類，本研究有 DH(Day in Hospital)、Blood Pressure、Cost…等欄位，第三類欄位的產生主要是因為資料來源的不同，使得各自的資料庫收集的資料不同。

三、欄位類別分類表

類別	欄位名稱
第一類	Name、Gender、Birthday…等。
第二類	Disease、Cost、Gender 和 Medicine。
第三類	DH(Day in Hospital)、Blood Pressure、Cost…等。

根據文獻[8]，此文獻說明了資料隱私的保護與資料回溯的可能，多個欄位同時揭露可能會違反隱私，例如文獻提到的 Gender、Birthday 和 ZIP，這些欄位如果個別的揭露其實是無法辨識特定個人，但是若同時把這三項欄位揭露出來可能可以辨別特定個人，也就是多個欄位揭露會違反隱私，我們將這些欄位稱為 Quasi-Identifiers[21]，所以本研究將以這三項欄位為基準，將延伸不同違反隱私的條件型態，做為兩間醫院之間的隱私違反條件。

本研究將利用第一類的欄位兩間醫院的資料對應，因為第一類欄位在 B 醫

院與 A 醫院皆存在，所以第一類欄位將成為資料交換的主要對應欄位；第二類的欄位主要是在做 Target Schema 內部分類後的對應，本研究將 Target Schema 分成兩個類別，一為研究人員(Researcher)可以看到的資料分類，另一為醫院醫療人員 (Medical Employees)可以看到的資料分類，為了簡化欄位需求本研究將不會細分醫生、護士和醫院行政人員可看到的資料等級，在此本研究不是真的將 Target Schema 切割分成兩類，而是採取目的的不同做欄位的遮罩，假設研究人員以數據分析為目的所收到訊息，只會有無法辨識特定個人的欄位，例如:Gender、Medicine 和 Disease。由於本研究流程在通過 ACP 時，即便判斷使用者的身分，所以在 Target Schema 中的欄位分類只是為了要達到$\sum t$的正常運作，確保使用者查詢時可以在有限的時間內完成動作，並且將經由$\sum st$對應後的 B 醫院資料回傳到 A 醫院進行分類。第三類的欄位主要是兩邊欄位皆不出現的其他資料；而本研究是希望取得 Source Schema 與 Target Schema 之間不同欄位資料，使得資料能夠更加的豐富，而有利於分析醫療環境上的關係。

4.2.3.2 重複資料判斷-雜湊函數的應用

在文獻[22]提到不同資料來源時，重複性資料整合的問題，此文獻利用建構第三方平台的概念來收集不同資料來源資料但卻是相同的資料，當存放在第三方平台的資料欲做修改或是使用時，都必須經由各資料來源的許可才可以動作。舉例來說，若兩家醫院同時診療一位病人，某家醫院欲對該資料進行讀取或修改時，則需要另一家醫院的同意，或是兩家醫院必須協商出特定規範，才能在此規範下使用。現在新版的個人資料保護法已經上路，其又更加重視資料的流通與保護，在此建置的第三方平台不一定能夠收集完整的個人資訊，因為無法確定其合法性，所以本研究提出利用雜湊函數的計算達到匿名性個人資料的對齊，來解決不同來源相同資料對應的問題。

根據 2.6 章節所述，本研究以 MD5 為主要基礎，雖然 SHA 的強度甚於 MD5，

但因本研究目前架構只有在醫療環境上且是用統計的目的上，所以不需要最強大的加密方式，而 MD5 也是目前電腦廣泛使用的雜湊演算法之一。本研究在使用者查詢兩邊資料時會一起將相對應資料的五個欄位(ID、Name、Birthday、Gender 和 ZIP)同時送到雜湊函數(MD5)進行加密動作，如圖 10，然後可以將兩邊計算出來的雜湊值拿來比較判斷資料是否一致，達到匿名性的資料比對。

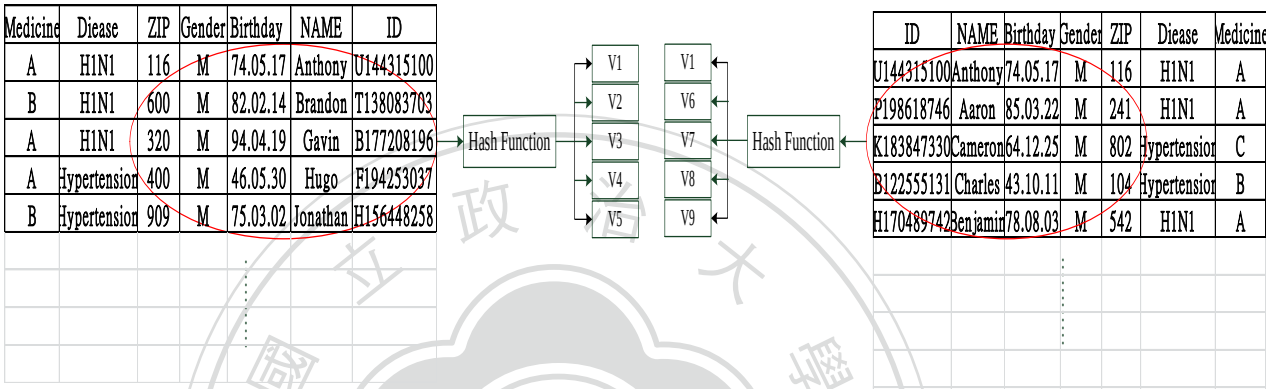

圖 10、個人資料經過雜湊函數處理

舉例來說，一個經過授權的使用者欲同時查詢 A 醫院與 B 醫院的醫療資料時，系統需要有能力去辨識兩間醫院的病患資料是否為同一個人，否則將會有資料的重複計算的問題，例如同一個人在多家醫院看過門診，也是現在社會上的常態。然而在使用者以統計的目的查詢時，可能會因為重複計算的問題導致統計資料的不正確性，例如:同個病患占了太多的樣本數。為此我們必須要有一定的個人資料去執行資料比對、刪除重複性資料，但是若這些比對的個人資料在不經處理的情況下，經由資料交換的方式從 B 醫院流到 A 醫院，這個動作可能是不合法的，基於個人資料保護法的原則，是不允許各單位任意將個人醫療資料恣意透漏給其他單位 (機關)，所以本研究加入了雜湊函數的概念來實現資料交換時，匿名性的資料比對與刪除的問題，如圖 11，左邊為沒有經過比對刪除重複性資

料，所以會有誤差值；右邊為加入了雜湊函數的概念，可以透過雜湊值來比對資料，進一步刪除，會有較小的誤差。

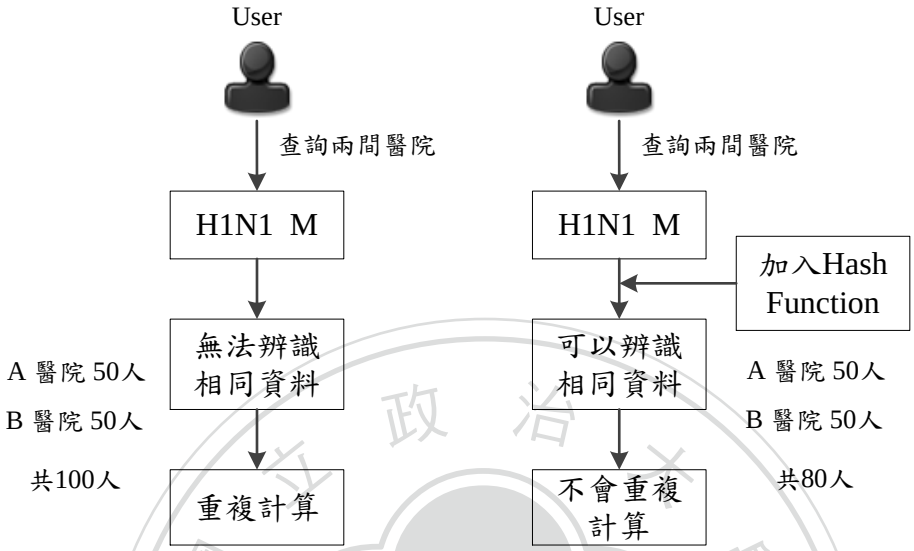

圖 11、不同來源資料對齊範例

4.2.3.3 DHP 中資料交換的落實

首先，本研究的資料是存放在以 TVD 保護的雲端資料庫中，在這個雲端資料庫中會將資料做加密和分割的處理，來增加資料保護的效果；在此本研究將不會探討如何加密、解密、如何分割和分割後的問題；而是討論兩個不同資料來源之間資料流通的問題，如何流通才能不違反隱私，給適當的人適當的資料才是本研究中心主軸。

根據 4.2.1 章節所述，資料交換的落實是在 DHP 階段的時候進行，利用 2.2 章節所述的資料交換方式取得兩方資料。

首先，使用者經由 ACP 的合法的授權與認證後，在 A 醫院下查詢指令，利用$\sum st$來進行兩間醫院之間的醫療資料傳輸的對應，從 B 醫院進行合法的資料取得並回傳，再對回傳後的資料做$\sum t$的處理，$\sum t$主要判斷是否會有 Weakly acyclic 的情況產生。$\sum st$的對應規則如下:

$\sum_{st} = \{B(1,2,3 \ldots 9,10) \rightarrow \exists 8.9.10.\, A(1,2,3 \ldots 9,10,11,12)\}$，詳見圖 12。此規則只是通式，而每一次的對應都會有的不同對應規則產生。本體論架構上的規則如下:

Request (?r) ∧ hasQueryType(?r, PBQ_A∧PBQ_B) ∧ QueryType (PBQ_A∧PBQ_B) ∧ hasData (?r,?rd) ∧ Data (?rd) ∧ hasHashValue (?rd,?hv) ∧ HashValue(?hv) ∧ hasBMedicalInfo (?rd,?brd) ∧ BMedicalInfo (?brd) ∧ hasPartOf (?brd,?bd) ∧ sqwrl:makeSet(?b,?bd) ∧ hasAMedicalInfo (?rd,?ard) ∧ AMedicalInfo (?ard) ∧ hasPartOf (?ard,?ad) ∧ sqwrl:makeSet(?a,?ad) ∧ sqwrl:union(?c,?b,?a)→ sqwrl:select(?c)----Rule 2

- Rule2 規則中，主要是展現資料交換的查詢結果，並顯示資料經由 Hash Function 計算後所得到的 Hash Value，而每一筆資料都會記錄是由 A 醫院或 B 醫院所有，最後會將經由查詢得到的資料聯集起來，達到可以得到外來資料源的資料。然而每一筆資料都會產生一個 Hash Value，主要是用來處理匿名性資料的比對，確認是否為同一筆資料。

我們在使用$\sum st$的時候，會一併的處理個人資料雜湊值的對應，將擁有相同雜湊值的個人資料紀錄成一筆，這樣既有匿名的個人資料保護效果，也避免過多重複性的資料產生。

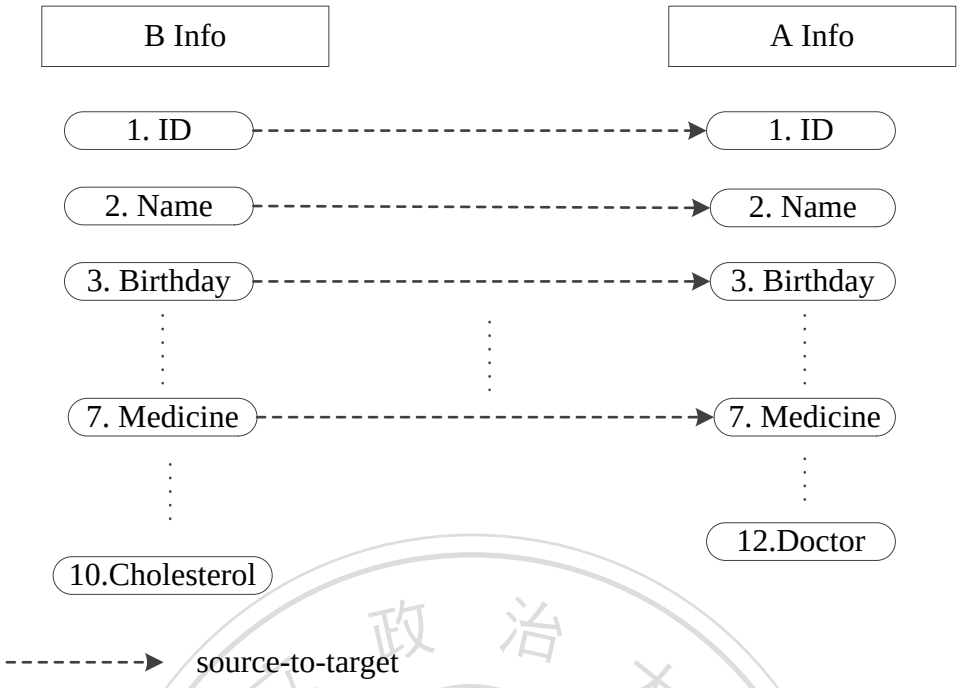

圖 12、Source to Target 對應圖

資料經由$\sum st$對應後，回傳到 A 醫院的資料庫做$\sum t$的處理，而$\sum t$主要判斷是否會有 Weakly acyclic 的情況產生。Weakly Acyclic 的用途主要用來判斷在 Target Schema 內部做$\sum \boldsymbol{t}$時是否會有迴圈的產生，若有 Weakly Acyclic 則表示此次查詢會有終止的時候，不會造成無止境的循環導致無法停止。判斷迴圈是否存在是利用 Weakly Acyclic 中產生 Special Edge 的特性來做決定的，當研究人員類別對應到醫療人員類別時，若有相同欄位則兩者間給予連線，若對應到 Labeled Null (即未匹配或對應到的欄位)則兩者之間的連線即稱為 Special Edge。對應規則如下:

$\sum_t = \{Res(4,6,7,11) \rightarrow \exists 1.2.3.6.8.9.10.12.Med(1,2,\ldots\ldots,12)\}$，詳見圖 13。

在確立 Special Edge 後，才可以判斷是否有由 Special Edge 所產生的迴圈，若有 Special Edge 所產生的迴圈則可能導致此次查詢工作無法停止，因為會在迴圈內一直不停的查詢；若 Special Edge 沒有產生迴圈，則稱此現象為 Weakly Acyclic 表示會有結束的時候。

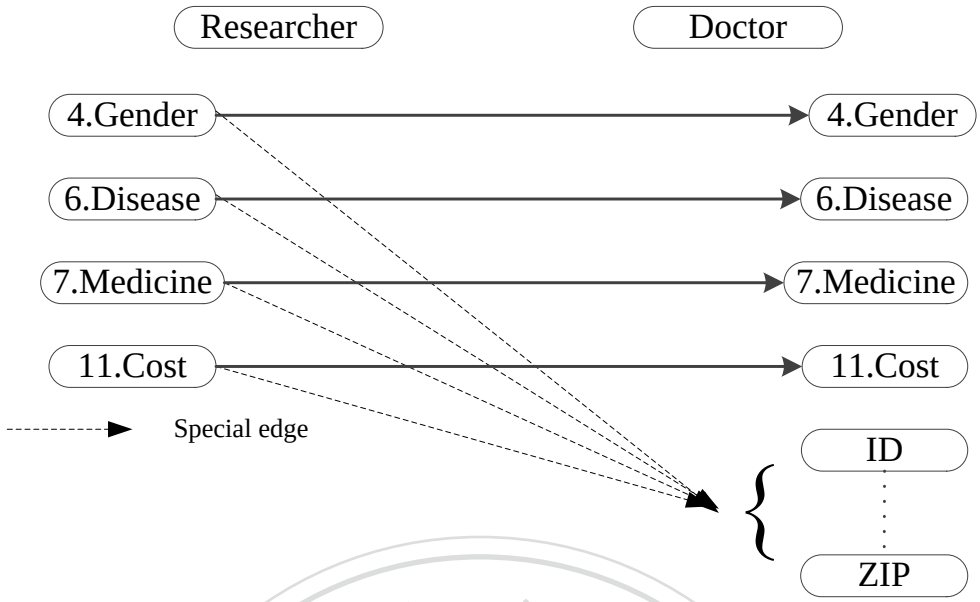

圖 13、在$\sum t$中判斷是否有 Weakly acyclic

本研究將會有 Weakly Acyclic 的產生，所以不會造成無法停止的情況發生，而每一次的查詢都會有一組新的對應情況出現。

4.2.4 DRP 設計

本研究中 DRP 的運用主要依據 ACP 與 DHP 所查詢的結果和授權的查詢方式來釋放資料，避免資料過多的揭露而侵犯隱私。

主要的資料揭露方式有兩種: Microdata 和 Macrodata[9]；當使用者欲將資料做欄位分析時，此時揭露的資料稱為 Microdata；而當使用者欲將資料做為統計分析後的統計數據時，則揭露的資料稱為 Macrodata。依據 Microdata 和 Macrodata 不同的特性，其資料保護的方式也有所不同。目前本研究中只討論 Microdata 的保護，也就是當使用者要求揭露原始欄位資料進行一般分析或者統計分析時，必須落實的保護。Microdata 的保護方式可以分為兩種：Masking 和 Synthetic[23]。Masking 會將資料做修改或隱藏的轉換用來做一般分析，而 Synthetic 會將資料轉換成具有統計特性的資料可以用於統計分析。Masking 又可分為 Non-Prturbative

和 Perturbative，Non-Perturbative 並不會修改資料內容但會藉由隱藏資料的方式來保護隱私，如 Local Suppression 或 Top-Coding；而 Perturbative 則會以資料修改的方式來保護隱私，如 Lossy Compression。本研究的統計保護型態只著重在 Non-Perturbative。

利用文獻[17]中 DRP 的設計將 A 醫院與 B 醫院的 DRP 分別表示，如圖 14，黃色部分為 B 醫院本來架構，紅色為 A 醫院本來架構。差在於 Quasi-identifiers 設計的不同，所以經由資料交換後可能會產生不同的結果。圖中的 DataType 則有兩個實體 Continuous 和 Categorical，每一種資料都以 hasDataType 關聯到自己擁有的資料型態。Continuous 是指可以數學運算處理的資料，如加減乘除等，例如 Cholesterol 和 Cost 等，而 Categorical 則是無法進行數學運算，例如 Name 和 Disease 等。

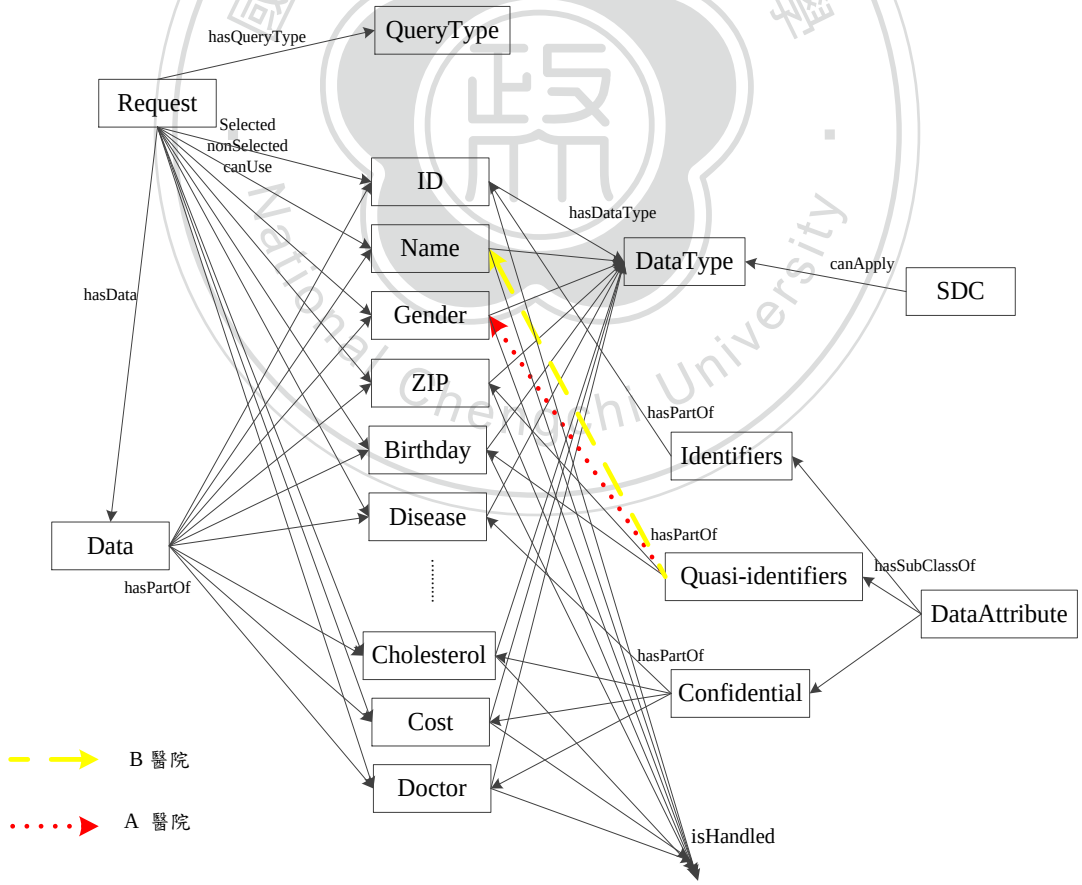

圖 14、A、B 和兩間醫院整合圖

首先，要先判斷使用者查詢型態 (SBQ、PBQ、SBQ_a∧ SBQ_b 或 PBQ_a∧PBQ_b)，假如為 SBQ 或、SBQ_a∧ SBQ_b 則直接將資料揭露給使用者，而 PBQ 或 PBQ_a∧PBQ_b 則判斷是否侵害隱私，本研究只針對 SBQ_a∧ SBQ_b 和 PBQ_a∧PBQ_b 進行討論，規則如下:

Request(?r) ∧ hasData(?r, ?d) ∧ Data(?d) ∧ hasPartOf(?d, ?pod)

∧ hasQueryType(?r, SBQ_a∧ SBQ_b) → canUse(?r, ?pod) ----Rule 3

Request(?r) ∧ hadData(?r, ?d) ∧ Data(?d) ∧ hasPartOf(?d, ?pod)

∧ hasQueryType(?r, PBQ_a∧PBQ_b)∧ sqwrl : makeSet(?rs, ?pod)

∧ sqwrl : groupBy(?rs, ?r)∧ Quasi- identifiers(?qui) ∧hasPartOf(?qui, ?qpod)

∧ sqwrl:makeSet(?qs, ?qpod) ∧ sqwrl : groupBy(?qs, ?qui)

∧ sqwrl : contains(?rs, ?qs)∧ Confidential(?c) ∧ hasPartOf(?c, ?dc)

→sqwrl : selectDistinct(?qui, ?gpod) ----Rule 4

Request(?r) ∧ hadData(?r, ?d) ∧ Data(?d) ∧ hasPartOf(?d, ?pod)

∧ hasQueryType(?r, PBQ_a∧PBQ_b)∧ sqwrl : makeSet(?rs, ?pod)

∧ sqwrl : groupBy(?rs, ?r)∧ Identifiers(?id) ∧hasPartOf(?id, ?qpod)

∧ sqwrl:makeSet(?qs, ?qpod) ∧ sqwrl : groupBy(?qs, ?qui)

∧ sqwrl : contains(?rs, ?qs)∧ Confidential(?c) ∧ hasPartOf(?c, ?dc)

→sqwrl : selectDistinct(?id, ?gpod) ----Rule 5

經由上述規則判斷後，可得知是否有違反隱私，若沒有侵犯隱私則進行揭露，規則如下:

Request(?r) ∧ hasData(?r, ?d) ∧ Data(?d) ∧ hasPartOf(?d, ?pod)

∧ hasQueryType(?r, PBQ_a∧PBQ_b) → canUse(?r, ?pod)--- Rule 6

若發現有侵犯隱私，則會使用 SDC 來處理資料，規則如下:

Request(?r) ∧ hasData(?r, ?d) ∧ Data(?d) ∧ hasPartOf(?d, ?b) ∧ selected(?r, ?b) ∧ hasDataType(?b, ?tp) ∧ DataType(?tp) ∧ SDC(?sdc) ∧ canApply(?sdc, ?tp) → sqwrl:select(?b, ?sdc) ---Rule 7

最後，運用 isHandled 的關係來判斷使用者查詢的欄位是否處理完畢，如果處理完畢，則將資料釋放，規則如下:

Request(?r) ∧ hasData(?r, ?d) ∧ Data(?d) ∧ hasPartOf(?d, ?b) ∧ selected(?r, ?b) ∧ isHandled(?b, 1) ∧ hasPartOf(?d, ?a) ∧ notSelected(?r, ?a) → canUse(?r, ?a) ∧ canUse(?r, ?b) ---Rule 8

- Rule 3 判斷 SBQ_a∧ SBQ_b 的情況，可以直接釋放。
- Rule 4 判斷在 PBQ_a∧PBQ_b 是否違反隱私，準則為是否同時存在 Quasi-Identifiers 加 Confidential，若有會告知違反了那個 Quasi- Identifiers 隱私條件。
- Rule 5 判斷在 PBQ_a∧PBQ_b 是否違反隱私，準則為是否同時存在 Identifiers 加 Confidential，若有會告知違反了那個 Identifiers 隱私條件。
- Rule 6 為 PBQ_a∧PBQ_b 不侵犯隱私情況，可以直接釋放。
- Rule 7 為 Rule 4 或者 Rule 5 兩者其中一條規則顯示出違反隱私，而需要進一步選擇 SDC 的保護方式來處理違反隱私的資料。
- Rule 8 為違反隱私的資料經由 SDC 保護處理完畢後，則可以將資料釋放。

與文獻[24]的不同點在規則判斷的部分，因為兩間醫院可能制定的 Data Attribute 不盡相同，所以規則的判斷哪些條件違反了哪間醫院的隱私項目。在一開始有說明兩邊資料經過資料交換處理後的釋放是由整合的 DRP 所執行，整合原因在於不同資料來源所收集的資料欄位不盡相同，若沒有相對應的欄位可能會

導致資料的不齊全，也失去了資料交換最初的目的，就是要拿取原本 Target 資料庫內所沒有的資料，所以我們將整合兩邊的 DRP。

4.2.5 查詢結果說明

當使用者下查詢指令經由本研究 ACP、DHP 和 DRP 判斷與資料交換後，資料釋放的結果會有侵犯隱私與不侵犯隱私兩種情境發生:

1. 不侵犯隱私:當使用者身分為研究人員，目的是為了研究分析，經過授權後向 A 醫院與 B 醫院進行查詢動作，若使用者欲查詢 Gender 和 Disease 的資料，因為兩邊都有相符合的欄位設計，所以會先將兩邊符合 Gender 和 Disease 的個人資料經由雜湊函數計算，計算完成後一併與 Gender 和 Disease 資料傳送到整合的 DHP 進行資料比對是否有重複的資料，若有則記錄為一筆，不再重複計算，最後經由 DRP 揭露資料，交換對應規則如下:

$$\sum_{st} = \{B(gender, disease, hashvalue) \rightarrow$$
$$\exists A(gender, disease, hashvalue)\}$$
$$\sum_{st} = \{B('F,'H1N1',' hashvalue') \rightarrow A('F,'H1N1','hashvalue')\}$$

其中，規則內的 hashvalue 會在系統執行時會自動產生，然後進行比對與刪除。

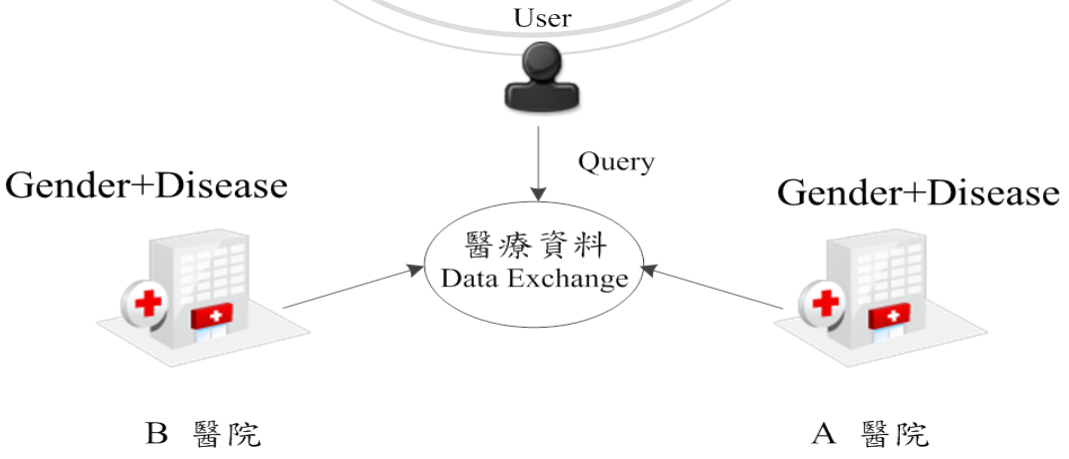

圖 15、不侵犯隱私範例

2. 侵犯隱私:當使用者身分為研究人員，目的是為了研究分析，經過授權後向

A 醫院與 B 醫院進行查詢動作，若使用者欲查詢 Birthday、Gender、ZIP 和 DH 的資料，在此條件狀況下因為在 A 醫院的收集欄位中沒有 DH 的資料，所以必須經由 B 醫院的資料所取得，而原本 A 醫院中若只查詢 Birthday、Gender 和 ZIP 是不會違反隱私條例的，但因外來的資料 (DH)導致 Quasi-Identifiers 和 Confidential Attribute 同時出現而違反隱私，於是整合的 DRP 將會啟動 Microdata 的保護方式來揭露資料，交換對應規則如下:

$$\sum_{st} = \{B(bir, gen, zip, dh, hashvalue) \rightarrow$$

$$\exists dh.\, A(bir, gen, zip, dh, hashvalue)\}$$

$$\sum_{st} = \{\mathrm{B}('500710', 'm', '247', '20, 'hashvalue') \rightarrow$$

$$\exists dh.\, A('500710', 'm', '247', '20, 'hashvalue')\}$$

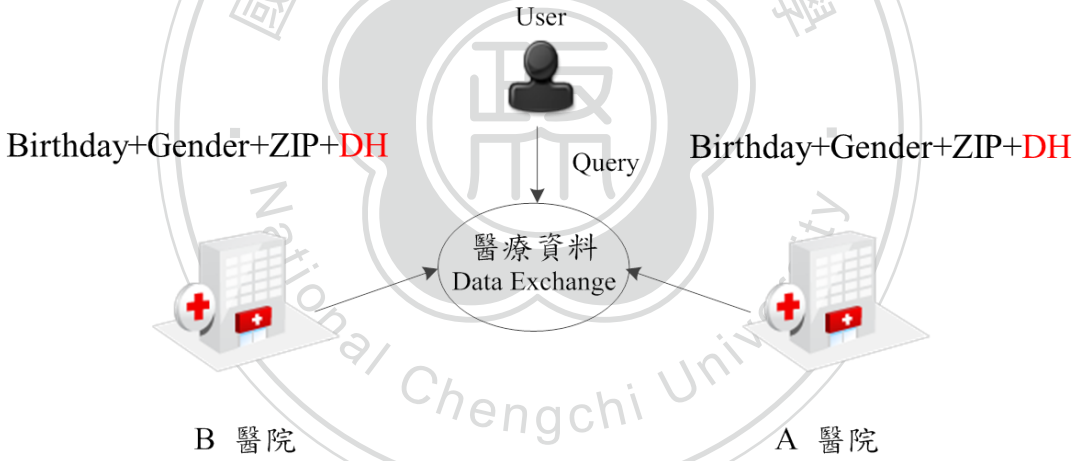

圖 16、違反隱私規則

4.3 不同資料來源的分析與優勢

4.3.1 SBQ 分析與優勢

在 $SBQ_a \wedge SBQ_b$ 的查詢條件下，本研究可以根據資料交換特點，得到 Target Schema 內所沒有的資料，例如:病患欲從 B 醫院轉院到 A 醫院，在傳統的做法上通常是人到病歷到，如果我們可以在病患運送的途中，將這些資料透過資料交換

的技術事先讓 A 醫院的醫生知道病患在 B 醫院的治療情況，像是住院天數、血壓...等，這樣對醫生和病患都是有益的，可以讓醫生在診治時作為參考的依據。

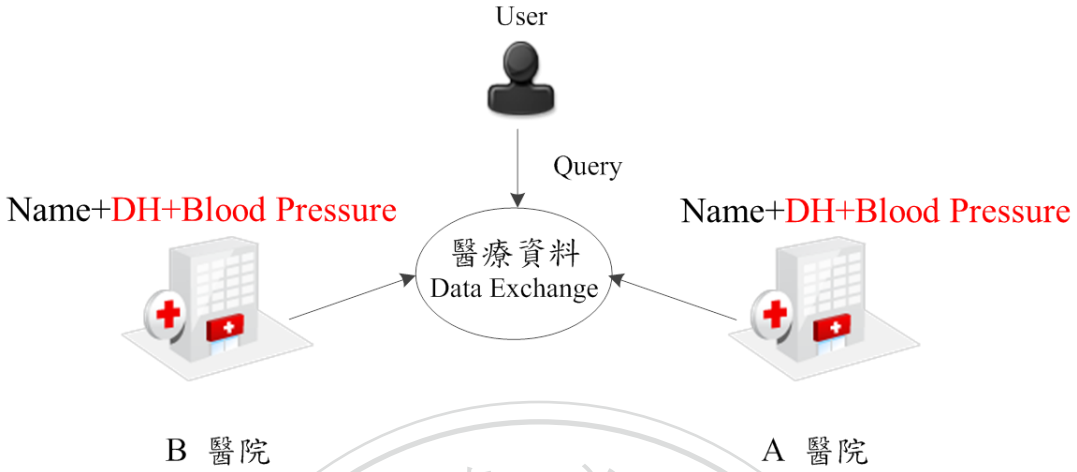

圖 17、找到單一資料源沒有的資料

4.3.2 PBQ 分析與優勢

本研究主要是分析相同 Domain 的情況下，資料來自不同資料來源時，對於使用者在查詢使用上或結果分析上，以及資料保護的安全性，與資料來自於相同來源之間的差異。

在 $PBQ_a \wedge PBQ_b$ 的查詢條件下，以分析結果來說，不單僅僅增加了樣本數，提高了分析的可信度，更可以得到更豐富的結果加以運用。這裡提出一個情境為例，假設 A 醫院與 B 醫院皆有相同種類的癌症病患，當使用者欲分析藥物上的治療對於相同癌症的藥效差異性，而針對 A 醫院與 B 醫院進行資料查詢；假設 A 醫院可能使用藥物 A 來進行此種癌症病患的治療；B 醫院可能使用藥物 B 來進行相同種類癌症病患的治療，經由資料交換的查詢，我們不僅僅可以分析出哪些治療藥物相對來說比較有效，更可以經由匿名性雜湊函數的概念進行資料的交叉比對得知同時使用藥物 A 與藥物 B 的病患有多少人，這是單一資料來源各別查詢無法達成的。

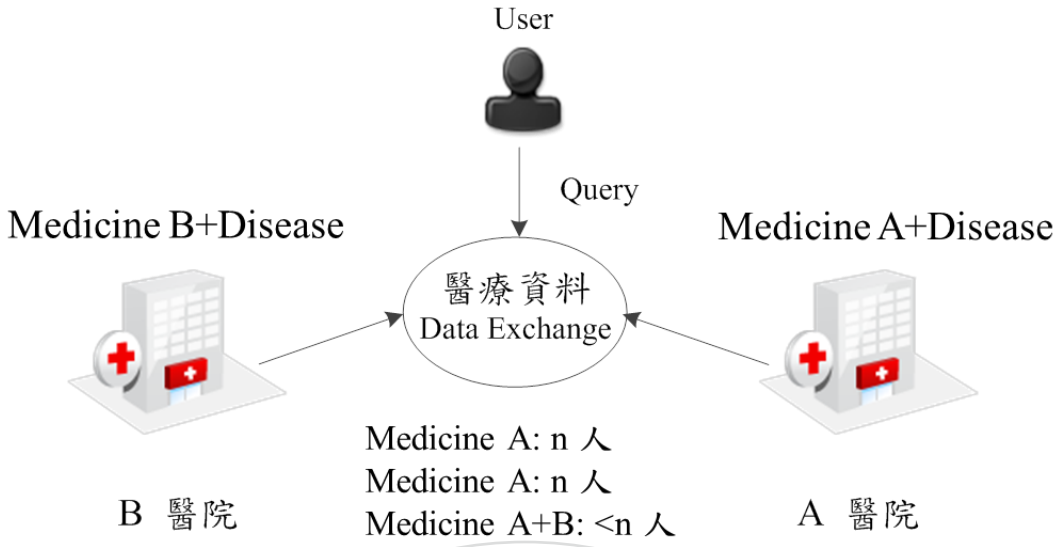

圖 18、可以交叉比對

另外，在 PBQ 的查詢條件下，以資料保護而言，與單一資料來源不同的地方在於，使用者經過授權後可以到 A 醫院進行資料的查詢，假設使用者查詢 Birthday、Gender 和 Disease，這三項要素不需要有 Microdata 隱私保護即可揭露；另一位使用者經過授權後可以到 B 醫院進行資料的查詢，假設使用者查詢 ZIP、Gender 和 Disease，相同的，這三項要素也不需要有 Microdata 隱私保護即可揭露；但是若將 A 醫院與 B 醫院倆便個別查詢出來的資料經過比對，我們同時會發現有四項要素存在 Birthday、Gender、ZIP 和 Disease，其中 Birthday、Gender、ZIP 為 Quasi-identifiers，再加上 Disease 為 Data Attribute 這些資訊是可以識別唯一人的，這樣的狀況是我們不希望發生的，如圖 19。所以就資料保護而言，本研究資料交換的好處在於可以避免上述各別查詢然後在人工比對的狀況發生，當使用者欲查詢 Birthday、Gender、ZIP 和 Disease 這四項要素時，會知道已違反了 Quasi-identifiers 與 Data Attribute 共同存在的問題發生，所以系統會經由 SDC 選擇適合的保護方式來揭露資料。

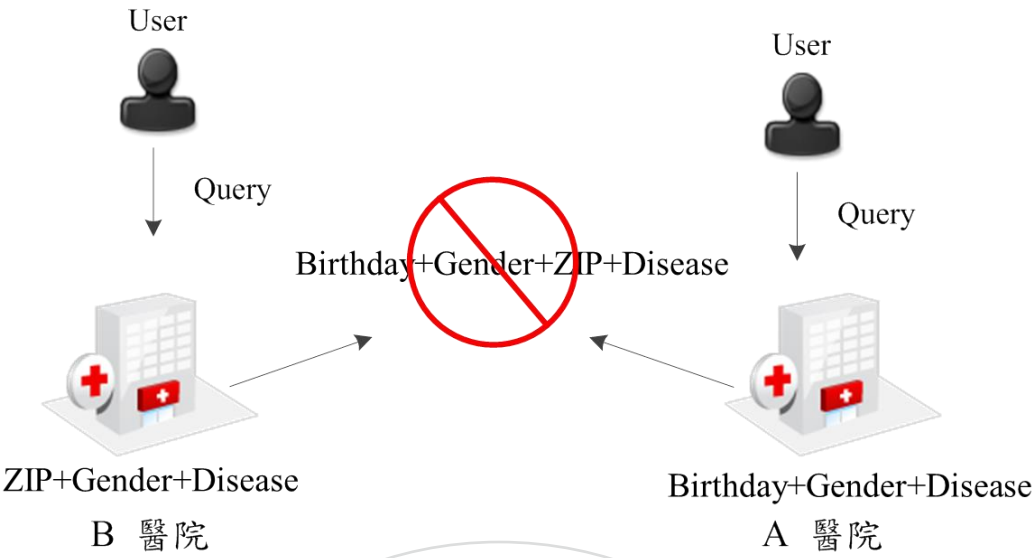

圖 19、各別查詢違反隱私

第五章

模擬驗證

5.1 模擬架構

本研究目標是在雲端環境上以醫療資料為主要情境,在個人資料不違反個人資料保護法的狀況下，可以相互對應、流通，達到給適當的人適當的資料。透過資料交換與語意網的技術來實行，落實資料提供者所期待的隱私保護，幫助使用者以不違反隱私期待的前提下尋找適當的資料。

5.2 模擬驗證之環境需求

為了驗證我們本研究假設的真實性,我們使用了模擬驗證的方式來加強我們假設的可靠度。在模擬驗證上使用工具 Protégé[24]做為編輯本體論的工具，使用 PROMPT 來進行 ACP 與 DHP 的整合,SWRL 落實推論的式子,另外再透過 Jess 推論引擎來推理 SWRL 的規則。使用的工具都可在 Protégé 工具完成，其元件說明如下：

- Protégé OWL Plugin：Protégé 是一種本體論建構與開發的開放式環境平台，擁有大量的 plugin 來支援本體論的推論與分析。
- Jess:一套以 java 語言為基礎的規則推論引擎,使用 Rete 演算法來處理規則；JessTab 是 Protégé 的 plug-in，其允許一同使用 Jess 及 Protégé。它擴展 Jess 額外的功能,來允許對映 Protégé 知識庫至 Jess facts(classes 及 individuals)。
- SWRL Tab Plug-in:可以掛載在 Protégé 平台上來進行規則的撰寫，亦可以當作 Referenced Libraries 導入至 Eclipse 平台上進行程式的開發。SWRL 規則語言可以來編寫規則和推論,其中所有的詞彙皆來自於我們所建構的本體論中，此外可以採用 Jess 推論器來推論轉寫的規則。

- PROMPT: 可以掛載在 Protégé 平台上將兩個本體論合併起來。PROMPT 外掛有四種模式，分別是比較 (Compare)、移動 (Move)、合併 (Merge)及萃取 (Extract)。在此我們只介紹實作將會使用合併 (Merge)模，合併模式則是將兩個本體論合併出一個新的本體論。在合併模式下選擇要合併的兩個本體論之後，執行合併模式，接著便會進入合併模式的選單。在選單中 PROMPT 演算法會自動判斷兩個本體論之間哪些類別是相似或完全相同，並提出建議，使用者可以依據 PROMPT 提出的建議來合併類別、屬性與實例，完成兩個本體論的合併。

5.3 Protégé 實作本體論與規則

本研究採用 Protégé 作為本體論的開發工具，版本為 Protégé3.4.4，建立模擬驗證所需要的情境資料與各種 Policies(ACP/DHP/DRP)，圖 20。

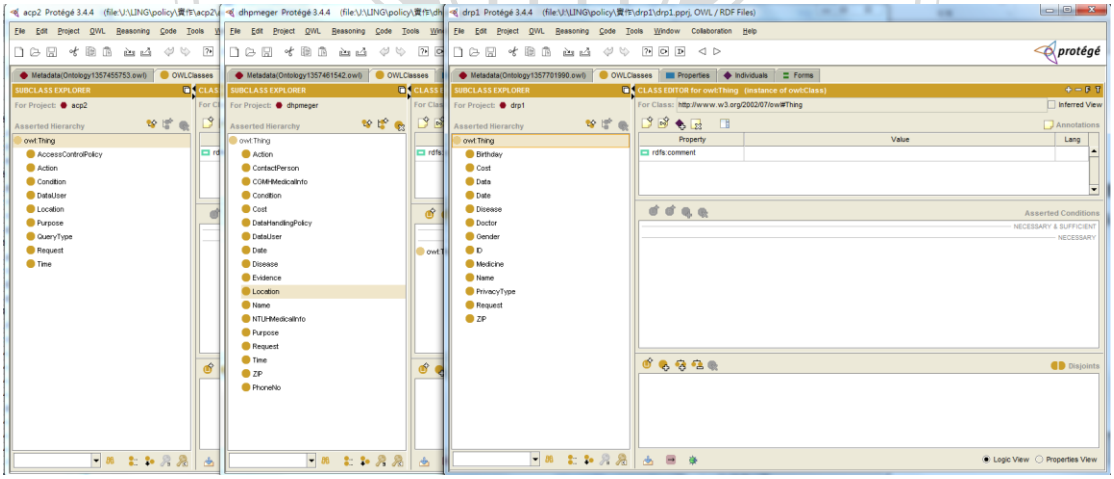

圖 20、使用 Protégé 3.4.8 模擬 ACP、DHP、DRP

根據章節四步驟說明，先利用 ACP 做身分和查詢型態判斷，如圖 21。可以知道每一個需求是哪一種查詢型態。

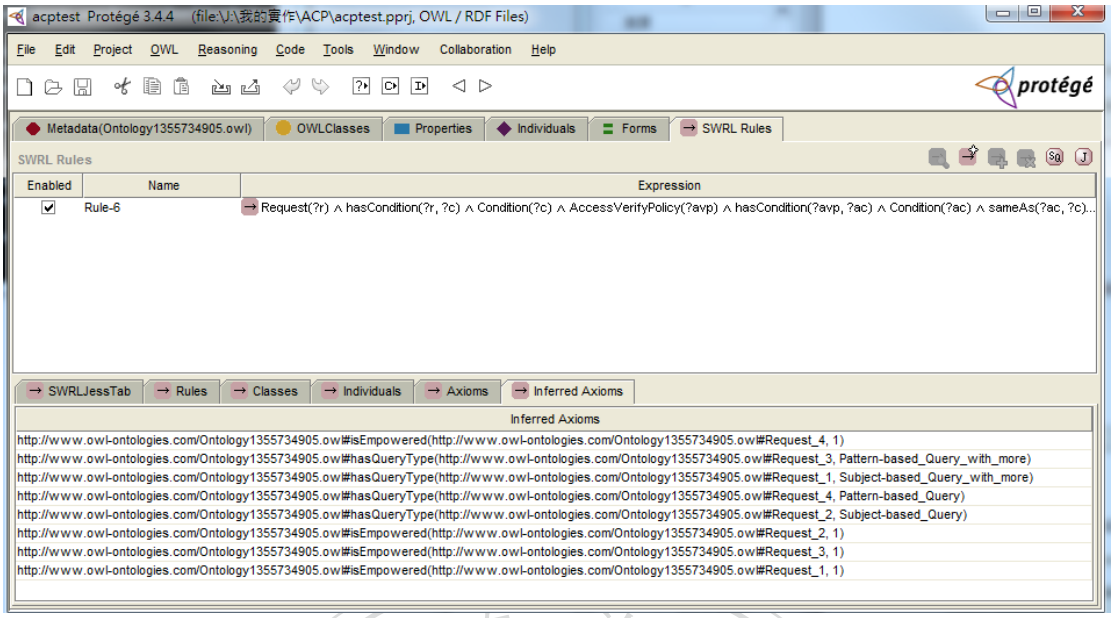

圖 21、ACP 推論圖

在使用進行 PROMPT 將兩邊的 DHP 做整合，如圖 22。

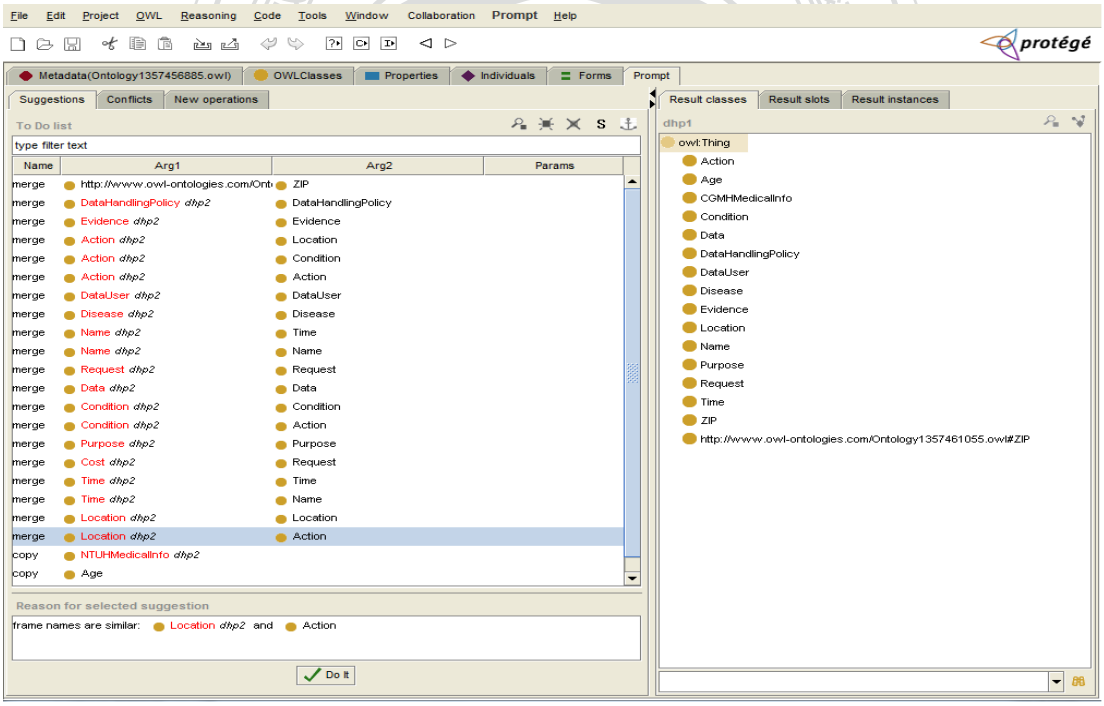

圖 22、DHP 整合圖

整合完成後，將透過 SWRL 的推論，可以得到兩間醫院的醫療資訊，如圖 23。

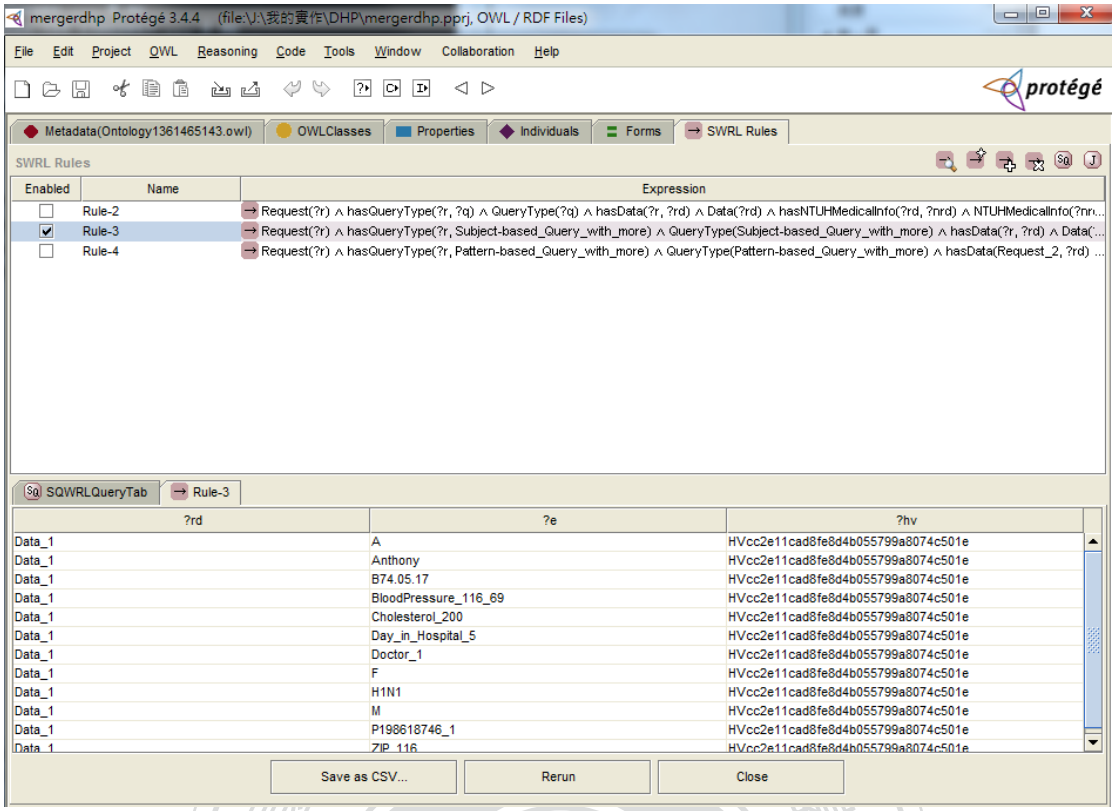

圖 23、Protégé SWRL Tab 推論 Data Handling Policy 的規則畫面

最後，可經由 DRP 判斷是否會違反隱私，若有則會顯示違反哪一個隱私規則。

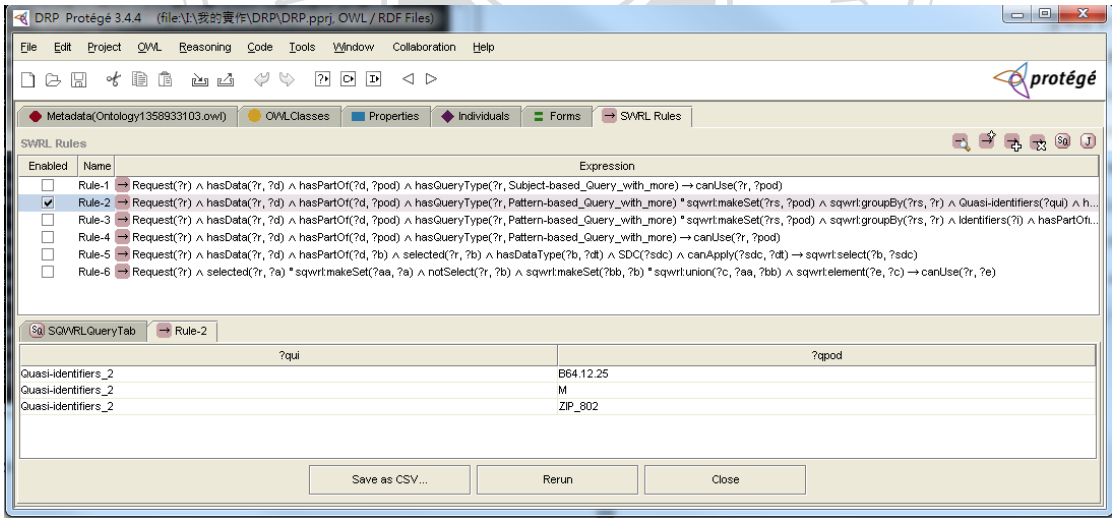

圖 24、Protégé SWRL Tab 推論 Data Release Policy 的規則畫面

第六章

結論與未來展望

本研究以語意化的技術結合了資料隱私存取控管的規範，並運用資料交換的技術，避免進行不同資料來源查詢時，因外來資料的加入而造成個人資料被 Re-Identification 發生，也運用雜湊函數的特性實現個人匿名性資料的對應問題。本研究利用醫療資訊來說明此概念，但其實只要是將個人資料存放在於雲端資料庫上的領域皆可以使用此架構達成上述目的，例如校園資訊系統、金融資訊系統等等都足以適用，只要依照不同的情境類別分類與動態的加入在領域規範的描述即可。

而在未來的研究中，可以使用本研究之方法，擴充在不同雲端環境上，不同資料來源的運用，以不違反法律的條件下，讓此方式可以運用在不同領域環境上，探討不同領域環境上的可行性，同時對隱私資訊的釋放也需要一併考量。

參考資料:

[1] Eberhart, A. *et al*., "Semantic Technologies and Cloud Computing." *In Foundations for the Web of Information and Services*, Fensel, D., Ed.; Springer, 2011, pp. 239–251.

[2] 新北市政府資訊中心,新北市打造雲端檔案櫃省紙減碳節省公帑千萬,2012 http://www.imc.ntpc.gov.tw/web/News?command=showDetail&postId=262250

[3] Bill Claybrook. "Differences Explained: Private vs. Public vs. Hybrid Cloud Computing." Sponsored by: HP & INTEL, 2011.

[4] R. Fagin , *et al*.," Data Exchange: Semantics and Query Answering", *Lecture Notes in Computer Science*, vol.2572, pp.207-224, 2003.

[5] A. Hernich, *et al.*," Logic and Data Exchange: Which Solutions Are "Good" Solutions?", *Lecture Notes in Computer Science*, vol.6006,pp.61-85, 2010.

[6] A. Y. Levy, *et al.*, "Querying Heterogeneous Information Sources Using Source Descriptions," *Presented at the Proceedings of the 22th International Conference on Very Large Data Bases*, 1996.

[7] R. Herold, "European Union (EU) Data Protection Directive of 1995 Frequently Asked Questions " *Computer Security Institute,2002.*

[8] R. Popp, *et al*.," Countering Terrorism Through Information and Privacy Protection Technologies", *IEEE Security and Privac*, vol.4,pp.18-27, 2006.

[9] V. Ciriani, S. Capitani di Vimercati, *et al.*, "Microdata Protection," in *Secure Data Management in Decentralized Systems*. vol. 33, 2007, pp. 291-321.

[10] L. Sweeney, *et al*.," k-Anonymity: A Model for Protecting Privacy", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*,vol.10,pp.557-570,2002.

[11] D. Calvanese and G. D. Giacomo, "Data Integration: A Logic-Based Perspective," *AI Magazine,* vol. 26, pp. 59-70, 2005.

[12] Y. Kalfoglou and M. Schorlemmer, "Ontology Mapping: The State of The Art", *The Knowledge Engineering Review*, vol. 18, pp. 1-31, 2003.

[13] J. Euzenat and P. Valtchev, "Similarity-Based Ontology Alignment in OWL-Lite", *ECAI* , 2004.

[14] N. F. Noy and M. A. Musen, "The PROMPT Suite: Interactive Tools for Ontology Merging and Mapping," *International Journal of Human-Computer Studies*, vol. 59, pp. 983-1024, 2003.

[15] R.L.Rivest.," The MD5 message digest algorithm ", *RFC 1321*, 1992.

[16] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression." *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 2002.

[17] 鄭國平,"雲端委外語意式資料保護," 碩士, 資訊科學學系,國立政治大學,2013.

[18] C.A. Ardagna, *et al.*, "A Privacy-Aware Access Control System*," *J. Comput. Secur.*, vol. 16, pp. 369-397, 2008.

[19] C. A. Ardagna, J. Camenisch, *et al.*, "Exploiting cryptography for privacy-enhanced access control: A result of the PRIME Project," *J. Comput. Secur.*, vol. 18, pp. 123-160, 2010.

[20] Ian Horrocks, Peter F. Patel-Schneider, Harold Boley, Said Tabet , Benjamin Grosof , Mike Mike Mike Dean(2004). ”SWRL: A Semantic Web Rule Language Combining OWL and RuleML”, http://www.w3.org/Submission/SWRL/

[21] OECD定義Quasi-identifiers http://stats.oecd.org/glossary/detail.asp?ID=6961

[22] 楊竣展,"整合資料在雲端環境上的分享與隱私保護-以電子病歷資料為例," 碩士, 資訊科學學系,國立政治大學,2011.

[23] J. Mateo-Sanz, A. Martínez-Ballesté, et al., "Fast Generation of Accurate Synthetic Microdata," *in Privacy in Statistical Databases*. vol. 3050, 2004, pp. 298-306.

[24] Knublauch, H., M. A. Musen, and A. L. Rector(2004). “Editing description logics ontologies with the Protégé OWL plugin”, *International Workshop on Description Logics*., Vol.104.