調查資料之隱私保護

李孟諭

調查訪問是研究現代社會現象的一項重要方法，也常用來做為制定政策或解決問題之參考。調查資料除了以研究主題為核心所蒐集的資料外，多少也會涉及受訪者的個人資料[1]，例如：姓名、生日、婚姻狀況、種族、信仰、職業或是收入等。目前我國已有個人資料保護法[2]（簡稱個資法）規範個人資料的「隱私權保護」及「合理利用」。近年來，人文社會科學領域也開始推行 IRB[3]（Institutional Review Boards）審查，主要確保受訪者知情同意，以及研究者不會逾越研究範圍和善盡資料的保密等。因此，隱私的保護與資料的保存已是從事調查研究所應正視的課題。學術調查研究資料庫（Survey Research Data Archive，簡稱 SRDA）[4]為使保管及釋出之調查資料無披露受訪者隱私資料之風險，會針對所保存的調查資料進行資料評估與相對應處理。本文將以 SRDA 經驗來介紹調查資料如何評估以及有哪些處理方式[5]，供研究者或資料管理者善盡資料保密責任之參考。

一、資料評估

首先，先了解調查資料的基本資訊，像是調查的最小單位是個人、家戶或是機構單位，調查是採抽樣調查或是普查，以及成功完訪的樣本數。再檢視調查資料的內容是否包含隱私資料，並考量這些資料是否具有潛在的研究價值。觀察內容包含：

（一）直接可辨識個別單位的資訊

變項能明確指出特定個別單位。例如：姓名、證照及牌照號碼（如身分證號碼、牌照號碼或護照號碼等）、聯絡個人所在地的資訊（如電話號碼、電子郵件位址、IP 位址、通訊地址等），及機構名稱和統一編號等獨一無二的資訊。

1 依個人資料保護法第二條第一項規定，個人資料指自然人之姓名、出生年月日、國民身分證統一編號、護照號碼、特徵、指紋、婚姻、家庭、教育、職業、病歷、醫療、基因、性生活（包括性取向）、健康檢查、犯罪前科、聯絡方式、財務情況、社會活動等及其他得以直接或間接方式識別該個人之資料。

2 立法目的之一，即在保護人格權中的資料隱私權。隱私權是保障個人生活私密領域免於他人侵擾及個人資料之自主控制，只要個人資料非自主地揭露於外界，就代表隱私權被侵犯。

3 目前中央研究院、國立臺灣大學、中國醫藥大學和國立成功大學等都已成立人文社會科學領域的「人類研究倫理審查委員會」，科技部人文司亦要求涉及人類研究者，除符合特定情形之外，應於申請時檢附已送人類研究倫理審查之證明文件，該文件未能於申請時提交者，應於計畫執行前補齊。

4 SRDA 對於資料的保護已累積多年經驗，除了資料整理外，亦導入各項管理機制，如資料分級、限制性資料遠端服務、限制性資料使用室和 ISO27001 資訊安全管理系統。

5 本文改寫自 SRDA 內部訓練文件「資料評估及處理機制」（2012）。

（二）間接可辨識個別單位的資訊

1. 連續變項

數值變項本身不會直接指出特定之個別單位，但因所提供的資訊較細緻，再伴隨其他的變項資訊就可能指出特定的個別單位。例如：身高、體重或收入等。

2. 文字變項

變項的文字內容可能會直接指出特定之個別單位，或是針對受訪對象做過於詳細的描述，再伴隨其他的變項資訊就可能指出特定的個別單位。這些變項如學校系所或行職業等過於詳盡的文字描述。

（三）詳細的地理資訊

地理資訊變項（村里和門號等）與其他受訪資料，如工作性質、教育程度和收入等比對後，辨識出特定之個別單位的風險將增高。例如：地理資訊層級小於（含）村里、文字敘述提及特定位置或機構（如重要地標，像是某捷運站或百貨公司）等。

（四）樣本特性明確

樣本來源為單一學校/機構、或某一個特定、相對小而明確的團體。例如：樣本母體是特定身份，像是職棒球員、原住民族或受刑人等。

（五）可能與外部資料串連

透過外部資料的取得與串連，會增加辨識出特定個人或單位的風險。例如資料檔中雖未提供特定學校名稱，但經由其他變項（如學校規模、學生數或學校位置等）與外部資料（如教育部統計資料）的串連，可能會有識別特定學校之風險。

二、處理方式

經過資料內容的評估後，對於高低風險的資料可進行相對應的處理，以降低識別特定個別單位之風險。一般常用的方法有：

（一）直接刪除（remove）

資料若具高風險，在最小蒐集原則下，最直接的方式是從資料檔中移除。如圖 1 所示，原始資料中存有受訪者的姓名和電話等隱私資料，若無進一步研究需求，應直接將這些資料刪除。

樣本編號	姓名	電話	題目1	題目2
001	黃一	2787-1234	1	3
002	吳二	2787-5678	1	4
003	張三	2787-9012	2	1

圖 1 刪除隱私資料示意圖

（二）嚴格控管

原始可識別特定個別單位的資料，將其留存有助於日後串連檔案，進行追蹤性的研究之用，應將這些可直接辨識之資料妥善控管，例如：權限控管或密碼管制等，必要時再授權取出串連。

處理上可將原始資料分割成兩個資料集，資料集間僅須保留樣本編號，如圖 2 左下為問卷資料，右下為受訪者隱私資料。前者可用於一般傳輸使用，而後者則是須嚴格控管，必要使用時再透過樣本編號合併兩個資料集。

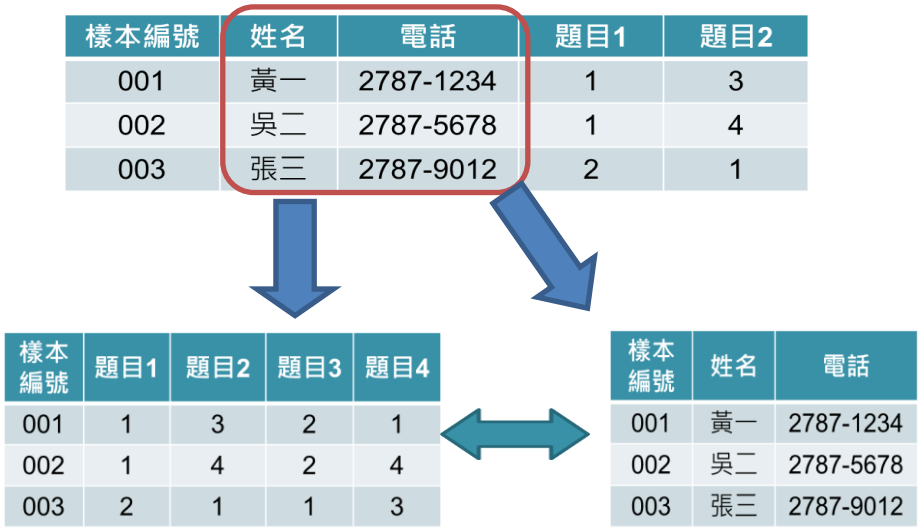


圖 2 分割隱私資料示意圖

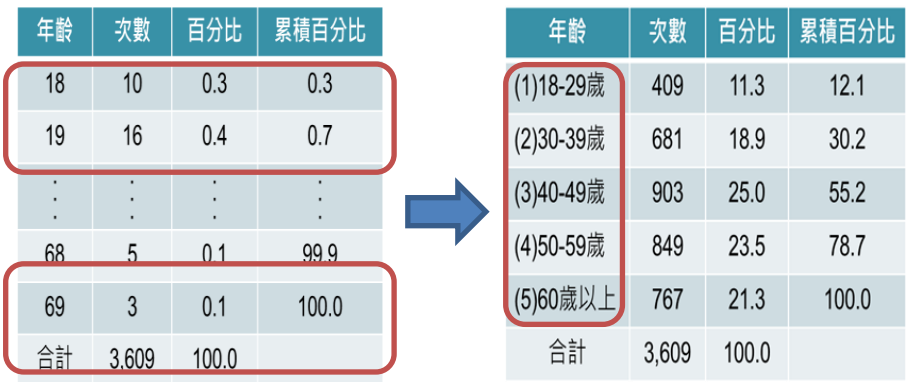
（三）重新分組/編碼（sub-group / top- or bottom-recoding）

1.重新分組

當連續變項的數值過於敏感時，可採用重新分組的方式整理資料。例如：若年齡為一連續變項，可將其分成數個組別，降低它與其他資訊相參照後，特定個人或

單位被辨識的風險；但須注意分組不宜過於粗略，使原始資料所蘊藏的資訊不致因而消失殆盡。

圖 3 左邊為一 18 歲到 69 歲年齡資訊的連續變項，右邊是重新分組（10 歲為一組）後的次數分配結果，最小一組是 18-29 歲（由於 18 和 19 歲樣本數過少，範例是將 18-19 歲組與 20-29 歲組合併為 18-29 歲組），最大一組是 60 歲以上，共分成五組。分組後，各組年齡所占百分比就會升高，可降低年齡資訊的敏感性。



年齡	次數	百分比	累積百分比
18	10	0.3	0.3
19	16	0.4	0.7
⋮	⋮	⋮	⋮
68	5	0.1	99.9
69	3	0.1	100.0
合計	3,609	100.0	

年齡	次數	百分比	累積百分比
(1)18-29歲	409	11.3	12.1
(2)30-39歲	681	18.9	30.2
(3)40-49歲	903	25.0	55.2
(4)50-59歲	849	23.5	78.7
(5)60歲以上	767	21.3	100.0
合計	3,609	100.0	

圖 3 重新分組資料示意圖

2.重新編碼

若連續變項的敏感性不高，但其極大或極小值占整筆樣本的百分比過小，容易有洩漏受訪者身份的風險，因此可將這類連續變項重新編碼，歸類至鄰近的數值。極端值的認定，可由次數分配結果觀察，重新編碼後的極大/極小值比例參考美國人口普查局的規則，建議至少需占總樣本數的 0.5%6。

例如圖 4 左邊為受訪者體重資訊的連續變項，可得知 20 到 29 公斤的樣本數不多，所占總樣本數的百分比也都在 0.5%以下，編碼上可將 20 到 29 公斤的樣本與 30 公斤的樣本結合成一組「30 公斤（含）以下」的組別，從圖 4 右邊可看出，重新編碼後的百分比可提高到 2.16%。

6 Checklist on Disclosure Potential of Proposed Data Releases (July 1999) http://fcsm.sites.usa.gov/files/2014/04/checklist_799.doc。

體重	次數	百分比	累積百分比
20	2	0.01	0.01
21	1	0.01	0.02
22	2	0.01	0.03
23	6	0.04	0.06
24	2	0.01	0.08
25	7	0.04	0.12
26	10	0.06	0.18
27	23	0.14	0.31
28	43	0.25	0.57
29	52	0.31	0.87
30	218	1.28	2.16
31	162	0395	3.11
⋮	⋮	⋮	⋮

體重	次數	百分比	累積百分比
30公斤(含)以下	366	2.16	2.16
31	162	0.95	3.11
32	283	1.67	4.78
33	177	1.04	5.82
⋮	⋮	⋮	⋮

圖 4 重新編碼資料示意圖

（四）移除明顯文字敘述

若資料中的文字資訊對於研究無特殊價值，但由於記載過於詳盡且具敏感性，應移除名稱等文字敘述。如圖 5 所示，原始資料的文字敘述過於特定或敏感，對於研究分析並無特殊意義，應將特定工作單位的名稱移除，僅保留在什麼行業上班即可。

樣本編號	在哪裡工作
001	在台灣銀行工作
002	中國信託商業銀行
003	在台北富邦銀行

樣本編號	在哪裡工作
001	在銀行工作
002	商業銀行
003	在銀行

圖 5 移除明顯文字敘述示意圖

（五）合併變項（combining variable）

過細的資料可能會透露過於詳盡資訊，為了降低敏感性，可將兩個或以上之細項變項合併為一個加總變項（summary variable）。例如圖 6 將家戶中某類支出的各細項金額，加總成為此類支出總金額後，再行釋出。

樣本編號	細項 1	細項 2	細項3
001	500	350	600
002	800	780	700
003	600	610	800

樣本編號	細項加總
001	1,450
002	2,280
003	2,010

圖 6 合併變項示意圖

（六）轉換變項

某些文字型變項雖然不適合釋出原始資訊，但依照特性或屬性經過分類後，仍具有研究的價值與意義，則可針對這類變項進行轉換處理。例如圖 7 將學校名稱轉換為新增變項後取代，新增變項保留其特定資訊，如公立/私立、普通/技職等；或是將系所資訊轉換成領域、學門。

樣本編號	大學
001	臺灣大學
002	政治大學
003	世新大學

樣本編號	公私立
001	(1)公立
002	(1)公立
003	(2)私立

圖 7 轉換變項示意圖

（七）分類編碼（coding）

針對行業或職業的文字型變項，可採用行職業標準分類代碼（如主計總處版 7），來進行分類編碼的工作（不釋出文字變項），既可保留分析價值，又不至於暴露風險。

如圖 8 的範例將原始文字變項中的中華電信、郵政和大學等資訊，依行職業標

7 行政院主計總處行業標準分類
http://www.dgbas.gov.tw/lp.asp?CtNode=5479&CtUnit=566&BaseDSD=7&mp=1

準分類代碼編碼成(610)電信業、(540)郵政及快遞業和(850)教育服務業。

樣本編號	在哪裡工作
001	中華電信
002	中華郵政
003	中華科技大學

樣本編號	行業代碼
001	(610)電信業
002	(540)郵政及快遞業
003	(850)教育服務業

圖 8 分類編碼示意圖

（八）虛擬或匿名化

若村里資料有保留的必要性，可利用虛擬編號來降低辨識的風險。如圖 9 將原始料的村里代碼，透過自訂的公式重新編碼（recode），將代號編碼成另一組虛擬村里代號，如此可用於分類和分析但卻無法得知是那一村里。

另外，特別提醒若樣本編號中包含鄉鎮市區的資訊，而樣本編號又有保留的必要性，也可利用虛擬編號建立新的樣本編號。例如目前 SRDA 釋出的「人力資源調查年資料」，鄉鎮市區代號即經過虛擬處理，實際的地理資訊僅提供到縣市等級。

樣本編號	村里代號
001	(103)八張里
002	(201)礁溪里
003	(303)頂城里

樣本編號	虛擬村里代號
001	506
002	802
003	605

圖 9 虛擬編號示意圖

三、其他處理機制

除了運用上述的處理方式來降低資料被辨識的風險外，國內外政府機構在整理或釋出特別敏感的資料時，另有數種處理方式。不過，由於這些方式會影響樣本數或變更原資料內容，所以需要進行一些統計上的確認，以免影響分析結果。一般用在樣本數很大或是普查資料。

（一）取樣（sampling）

當樣本人數占母群體總人數的比例很高（甚至普查）時，樣本被辨識的風險就會提高。若資料的樣本數非常大，可從原始資料中抽取足以提供合理推論的樣本大小，代替釋出所有的原始資料。例如：SRDA 釋出的「台灣教育長期追蹤資料庫」（Taiwan Education Panel Survey，簡稱 TEPS）公共版，是從實際受訪樣本中隨機抽取 70%的樣本資料作為釋出檔案，只要具網路會員身份即可直接下載。

（二）擾亂（disturbing）

當變項的原始數值可能會透露敏感資訊時，可考慮增加變項的隨機變異或隨機誤差，降低資料被識別的風險。如同一群體的樣本有相同的權數，可將權數加上微小的隨機數值，可解決同一群體樣本因權數一樣導致被辨識風險，且不至於影響分析。例如：TEPS 的學生資料因同一班級有相同的權數，為使學生的班級身份不被辨識，已將學生權數加上微小的隨機數值，降低同一班學生權數都一樣的被辨識風險。

（三）置換（swapping）

當某些樣本的特定變項數值，容易洩漏樣本的個人資訊，則可針對那些可被間接識別出的樣本，置換其重要變項之數值或將其換成特定範圍的數值。一種方式是紀錄置換（record swapping）8，就是將敏感變項之數值調換（如兩個家戶樣本位在不同地理區，都只住 1 人，除了收入之外，其他個人背景資料都一樣，可將兩個家戶資料置換）；另一種方式為等級置換（rank swapping），也就是將某些級距內的資料，其一小部分換成特定範圍的數值（如級距之平均值，適用於連續變項）。由於這兩種方式都會變更原資料內容，實務上 SRDA 並未使用過。

隨著個資法修法通過後，個人資料保護的議題不斷被提及，調查資料的個人隱私保護已是研究者所必須肩負的責任，一旦處理不當，可能會有民事、刑事和行政責任。希望透過本文的介紹，能協助研究者妥善管理自己所掌管的調查資料，以保障受訪者的權利。當然，若研究者礙於人力、物力或財力無法對調查資料有效管理，亦可選擇捐贈給 SRDA，一方面資料得以被安全妥善、長久地保存下來，提高研究成果的能見度，另一方面當資料再次被利用，其他資料使用者引述資料來源時，亦可增加原研究者的學術聲譽。

8 Census Statistical Disclosure Control - The use of record-swapping to protect data confidentiality
http://www.nrscotland.gov.uk/files//statistics/seminars/Scotlands-People-28-November/12-pams-conf-nov2013-census-disc-control.pdf