個人資料去識別化過程驗證案例報告

日期： 105.2.23

財政部財政資訊中心
Fiscal Information Agency, Ministry of Finance

簡報大綱

- 緣起及依據
- 導入及驗證過程
- 驗證標準涵蓋內容及驗證範圍
- 實作案例
  - 風險評鑑
  - 個人資料去識別化
- 結論與建議
- 誌謝

緣起及依據

- 104.8.18 行政院研商「個人資料去識別化」驗證標準規範（草案）會議紀錄：請財政部財政資訊中心率先運用前開標準進行「個人資料去識別化」驗證，於本 (104) 年 11 月底前完成驗證程序。

- 104.9.17 隨前揭會議紀錄函送由經濟部標準檢驗局研訂之「個人資料去識別化過程驗證要求及控制措施」。

導入及驗證過程

8.18 行政院研商會議

8.20 中心召開研商導入會議

9.1 第 1 次輔導團隊教育訓練

9.17 標檢局驗證要求及控制措施定稿

10.2 文件導讀

10.7 第 1 次文件檢視及討論會議

10.23 提出驗證申請

11.5 第 7 次文件檢視及討論會議

11.10 ETC 現場訪談

11.13 內部稽核

11.16 管理審查會報

11.18 及 19 ETC 現場驗證

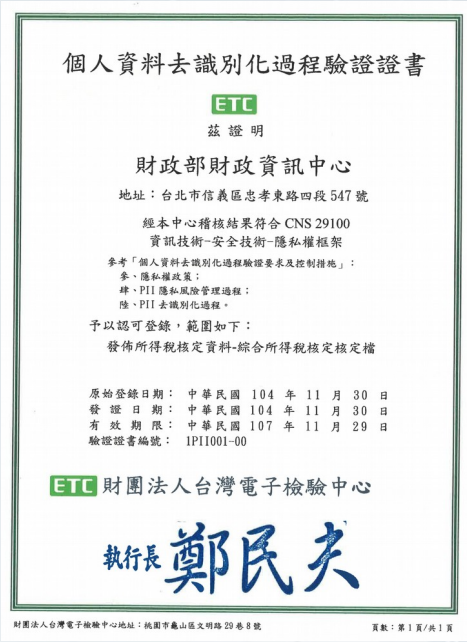
11.25 矯正計畫

11.30 取得證書

12.14 至行政院進行案例報告

共完成

一階文件	1 份
二階文件	11 份
三階文件	1 份
四階文件	15 份



個人資料去識別化過程驗證證書

ETC

茲 證 明

財政部財政資訊中心

地址：台北市信義區忠孝東路四段 547 號

經本中心稽核結果符合 CNS 29100
資訊技術-安全技術-隱私權框架

參考「個人資料去識別化過程驗證要求及控制措施」：
柒、隱私權政策；
肆、PII 隱私風險管理過程；
陸、PII 去識別化過程。

予以認可登錄，範圍如下：

發佈所得稅核定資料-綜合所得稅核定核定檔

原始登錄日期： 中華民國 104 年 11 月 30 日
發 證 日 期： 中華民國 104 年 11 月 30 日
有 效 期 限： 中華民國 107 年 11 月 29 日
驗證證書編號： 1PII001-00

ETC 財團法人台灣電子檢驗中心

執行長 鄭民夫

財團法人台灣電子檢驗中心地址：桃園市龜山區文明路 29 巷 8 號　　頁數：第 1 頁/共 1 頁

各階文件列表

- 一階文件 (1 份)
  - 隱私權政策
- 二階文件 (11 份)
  - 組織管理要點
  - 文件暨紀錄管理要點
  - 人力資源安全管理要點
  - **隱私風險評鑑管理要點**
  - 隱私風險處理管理要點
  - 個人可識別資訊蒐集、處理及利用管理要點
  - 當事人權利行使管理要點
  - 資訊安全與責任管理要點
  - 遵循性管理要點
  - 稽核作業管理要點
  - 非預期資料揭露管理要點
- 三階文件 (1 份)
  - **個人資料去識別化過程作業規範**
- 四階文件 (15 份)
  - 紀錄表單

驗證標準涵蓋內容及驗證範圍

▶驗證標準涵蓋內容

個人資料去識別化過程驗證要求及控制措施

- ▶參、隱私權政策： 1 項要求、 10 項控制措施。
- ▶肆、 PII 隱私風險管理過程： 1 項要求、 1 項控制措施。
- ▶陸、 PII 去識別化過程： 5 項要求、 24 項控制措施。

▶驗證範圍：所得稅核定資料 - 綜合所得稅核定檔

- ▶以 102 年度綜合所得稅核定資料為實作案例。
- ▶資料檔共 7,200,807 筆紀錄。

隱私風險評鑑

定義可接受風險值為 1.2

衝擊構面評分 8 ~ 40 分

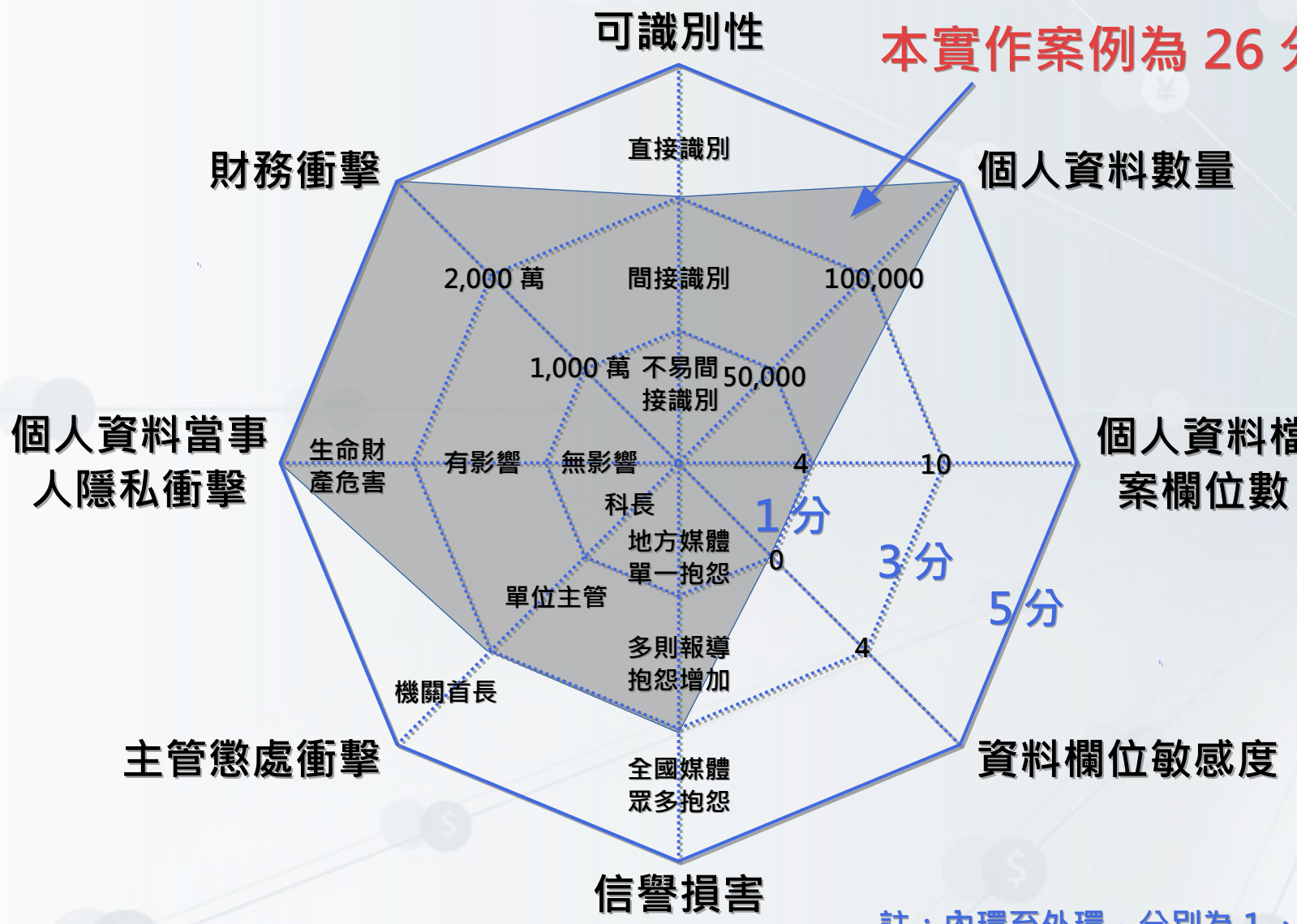
重現性、資源可用性及區別性三者擇其最高者

風險值 = 衝擊值 x 重新識別可能性

本實作案例：

1.18 = 26 x 1/22

衝擊值：衝擊構面評分



註：內環至外環，分別為 1 、 3 、 5 分。

重新識別可能性：重現性

- 不變動型資料集：1
- 部分變動型資料集：0.75
- 變動型資料集：0.5

$$\text{重現性} = \text{權重} \times \frac{\text{鄧巴數}}{\text{資料集中總個體數}}$$

本實作案例為 2.08×10^{-5}

- 資料接收者團隊恰巧認識資料集中個體的數值。
- 某個人維持緊密人際關係的人數上限。
- 通常設定為 150 。

本實作案例為 7,200,807

重新識別可能性：資源可用性

$$\text{資源可用性} = \frac{\text{可重新識別資料筆數}}{\text{重新識別測試資料筆數}}$$

本實作案例為 0

本實作案例為 0

本實作案例為 375

- 母體為 7,200,807 筆，合格群組數為 63,730 。
- 基於 95% 信心水準，誤差為正負 5% ，共抽樣 375 筆進行重新識別測試。

重新識別可能性：區別性

威脅模型

T1: 評估內部控管及外部動機與能力。

T2: 資料遺失，攻擊機率 =1/(365xN)
其中 N 表近 N 年資料未遺失。

T3: 資料開放，攻擊機率 =1

區別性 = 攻擊機率 x 猜中的最高機率

本實作案例為 1/22

本實作案例為 1/22
以 K 匿名法實作，
猜中的最高機率為 1/K

個人資料去識別化

- 本實作案例係以K匿名法進行去識別化處理。
- K匿名法 (K-Anonymity)
  - 原理
    - 當一個資料集中，對於一個或多個屬性值結合起來的組合 (如地址、年齡、性別等) ，若是可以找到 K 筆資料具有相同的組合，則此資料集就符合 K 匿名。
  - 處理方式
    - 概化 (generalization) ：將屬性值以區間值代替。
    - 抑制 (suppression) ：以其他符號代替、隱藏或刪除屬性值。
  - 判定方式
    - 將同一屬性值結合起來的組合進行分群，計算各群組之個體數量，以判定K值。

K匿名法 (K-Anonymity) 範例

姓名	年齡	性別	住所	宗教
Ramsha	29	F	Tamil Nadu	Hindu
Yadu	24	F	Kerala	Hindu
Salima	28	F	Tamil Nadu	Muslim
Sunny	27	M	Karnataka	Parsi
Joan	24	F	Kerala	Christian
Bahuksana	23	M	Karnataka	Buddhist
Rambha	19	M	Kerala	Hindu
Kishor	29	M	Karnataka	Hindu
Johnson	17	M	Kerala	Christian
John	19	M	Kerala	Christian

抑制　概化　抑制

姓名	年齡	性別	住所	宗教
*	20<Age30	F	Tamil Nadu	*
*	20<Age30	F	Kerala	*
*	20<Age30	F	Tamil Nadu	*
*	20<Age30	M	Karnataka	*
*	20<Age30	F	Kerala	*
*	20<Age30	M	Karnataka	*
*	Age20	M	Kerala	*
*	20<Age30	M	Karnataka	*
*	Age20	M	Kerala	*
*	Age20	M	Kerala	*

2-Anonymity
相同屬性之紀錄皆至少有 2 筆

個人資料去識別化程序

- 判定直接識別欄位。有重新識別要求，採擬匿名化處理；若無，則刪除。
- 依需要選擇最少必要欄位，並判定間接識別欄位。
- 進行風險評鑑，決定 K 門檻值。
- 調整參數，進行間接識別欄位去識別化處理。
- 驗證是否符可接受風險值。若無，重覆前一步驟。
- 處理離群值 (以前 5% 個體回應全體 50% 支配模式估算) 。
- 檢查，進行重新識別測試，並重新確認風險評鑑。
- 輸出，並保留相關紀錄。

實作案例

- 狀況 1
  - 確認戶籍地址概化處理方式。
- 狀況 2
  - 確認間接識別欄位及低度間接識別欄位皆納入 K 匿名法之分群計算。
- 狀況 3
  - 確認最終方案僅納入戶籍地址、所得總額兩項間接識別欄位。

實作案例：狀況 1

- 威脅模型：資料開放 / 重新識別要求：無
- 間接識別欄位
  - 戶籍地址、所得總額
- 低度間接識別欄位：無
- 概化處理方式
  - 戶籍地址：最小統計區、一級發布區、二級發布區
  - 所得總額： 4 等分位組
- K 門檻值： 22

實作結果：狀況 1

- 限於最小統計區轉換速度瓶頸，故以臺北市 (資料共 939,640 筆) 為例，進行觀察。
- 以戶籍地址及所得總額 (4 等分位組) 進行分群。

戶籍地址概化處理方式	總群組數量	小於 K=22 之群組數量	小於 K=22 之群組所佔資料比例
最小統計區	42,570	22,321 (52%)	**31%**
一級發布區	29,993	6,926 (23%)	**11%**
二級發布區	**3,583**	**15 (0.4%)**	**0.004%**

實作結論：狀況 1

- 僅考慮戶籍地址及所得總額進行分群，戶籍地址概化處理需擴大至二級發布區，滿足 K 門檻值 22 之群組所佔資料筆數方大於 95% 。
- 經徵詢潛在使用者意見
  - 二級發布區地理空間大小類似於村里 (全國共 9,458 個二級發布區，村里數則為 7,851 個) 。
  - 一般民眾對村里名較為熟悉。
- 考量前列因素，改以傳統村里界進行戶籍地址概化處理。

實作案例：狀況 2

- 威脅模型：資料開放 / 重新識別要求：無
- 間接識別欄位
  - 戶籍地址、所得總額、所得淨額、應納稅額
- 低度間接識別欄位
  - 扶養人數：70 歲以上扶養人數、70 歲以下扶養人數
  - 扣除額人數：身心障礙特別扣除額人數、幼兒學前扣除額人數
- 概化處理方式
  - 戶籍地址：村里
  - 所得總額採 10 等分位組，所得淨額及應納稅額因高度相關於所得總額 (相關係數分別為 0.99 及 0.97) ，採 10 等分位組概化，但不納入分群計算。
  - 低度間接識別欄位：僅概化處理，不納入分群計算。
- K 門檻值： 22

實作結果：狀況 2

- 以全國資料進行計算。
- 以戶籍地址及所得總額 (10 等分位組) 進行分群。
- 為提交現場驗證之實作案例。

戶籍地址概化處理方式	總群組數量	小於 K=22 之群組數量	小於 K=22 之群組所佔資料比例
村里	78,549	14,819 (18.8%)	2.87%

實作結論：狀況 2

- 經現場驗證時與稽核團隊討論：
  - 以下欄位與其它間接識別欄位組合時，仍有可能重新識別單一個體，且不符 K 匿名法之定義，故皆應併同納入分群計算，包含：
    - 間接識別欄位：所得淨額、應納稅額
    - 低度間接識別欄位： 70 歲以上扶養人數、 70 歲以下扶養人數、身心障礙特別扣除額人數、幼兒學前扣除額人數
- 矯正計畫：
  - 所有間接識別欄位及低度間接識別欄位皆納入 K 匿名法之分群計算。

實作案例：狀況 3

- 威脅模型：資料開放 / 重新識別要求：無
- 間接識別欄位
  - 戶籍地址、所得總額、所得淨額、應納稅額
- 低度間接識別欄位：無
- 概化處理方式
  - 戶籍地址：村里
  - 所得總額、所得淨額及應納稅額： 10 等分位組
- K 門檻值： 22

實作結果：狀況 3

- 以全國資料進行計算。
- 以戶籍地址、所得總額、所得淨額及應納稅額 (皆 10 等分位組) 進行分群。

戶籍地址概化處理方式	總群組數量	小於 K=22 之群組數量	小於 K=22 之群組所佔資料比例
村里	291,319	199,587 (68%)	18%

實作結論：狀況 3

- K 匿名化之間接識別欄位愈多，則分群愈細，每群之個體數量愈少，達一定 K 門檻值之資料顆粒度將愈加粗糙。
- 以狀況 3 為例，若需達 K 門檻值 22 ，則戶籍地址需以鄉鎮以上等級進行概化處理。
- 考量資料顆粒度不宜過粗，故回歸以狀況 2 之方式處理，惟不提供所得淨額、應納稅額、扶養人數及扣除額人數等欄位。

實作結論：總結

- 威脅模型：資料開放
- 重新識別要求：無
- K門檻值： 22
- 間接識別欄位
  - 戶籍地址、所得總額
- 低度間接識別欄位：無
- 概化處理方式
  - 戶籍地址：村里
  - 所得總額： 10 等分位組
- 資料損失
  - 以群組計： 18.8%
  - 以筆數計： 2.87%

- 資料損失範圍

結論與建議

- K 匿名法不適用於連續型資料，可能會產生資料損失，且隨資料欄項增加，資料顆粒度將愈粗糙。

- 隱私保護技術除 K 匿名法外，尚有不同實作方式，建議持續引入新技術以滿足不同需求。

- 統計資料及去識別化後之原始資料各有不同應用標的，宜先確認應用情境，再選擇合適之方法。

誌謝

- 內政部
- 財團法人資訊工業策進會科技法律研究所
- 財團法人工業技術研究院巨量資訊科技中心
- 財團法人台灣電子檢驗中心

簡報完畢
敬請指教

財政部財政資訊中心

Fiscal Information Agency, Ministry of Finance