

GIJC25工作坊

清資料 (Data Cleaning) :

「提升新聞準確性的關鍵步驟」

在數位時代，數據是我們挖掘新聞故事的基石，但數據的品質直接決定了故事的準確性。我們將聚焦在一個對所有記者來說都至關重要，卻常常被忽視的環節--資料清理。

資料清理或稱為資料整理，核心任務是讓數據變得可靠且可用。它涉及到識別和糾正錯誤、填補缺失值、移除重複，並解決數據集中的各種不一致性。

為什麼清資料很重要？

- **現狀挑戰：**現今 80% 的數據資料（如電子郵件、PDF、社群貼文）皆屬於非結構化格式，難以直接分析。
- **潛在風險：**各國資料格式不一致、缺失值或人為輸入錯誤，若未經處理直接引用，容易導致報導失真。

假設你搜集到一個數據報表，城市那欄出現了「台北市」和「北市」兩種寫法，會導致分析結果出錯。另外像是日期格式混亂、缺失值，都會導致誤導性的分析結果。

根據統計，全球大約有 **80%** 的數位資料，像是 PDF 文件、電子郵件或社群媒體貼文，都是非結構化的數據，無法直接用於分析。

清資料的目的是「準備數據」，確保每一步都經過仔細調整和紀錄，建立一個可持續的工作流程，才能確保新聞的真實性與準確性。

清資料的標準流程概覽

資料清理並不是一個單一的動作，而是一個標準化的三步驟流程：

1. **理解與備份 (Understanding and Backup)**
 - 在開始調整任何數據之前，先花時間觀察數據結構、欄位意義和潛在問題。
 - 最重要的一步：務必先備份原始檔案，無論後續操作出任何問題，都能夠隨時回到起點。
2. **執行清理 (Execution)**

- 這是實際操作的環節，我們在這裡會執行移除重複、統一日期格式、處理缺失值以及合併相似類別等具體動作。
3. 驗證與紀錄 (Validation and Documentation)
- 清理完畢後，必須檢查數據的邏輯性。例如，如果某人的出生年份是 2060 年，這顯然是個輸入錯誤。
 - 同時，詳細記錄您做的每一個清理步驟，這是確保透明度，並讓您的報導經得起檢驗的關鍵。

實務操作：具體的清理技巧

我們將清理動作分成兩大類：數據一致性與數據完整性。

數據一致性 (Consistency) 確保數據的表達形式標準化：

- 統一格式：將日期、貨幣或數字格式標準化（例如，全部使用 YYYY-MM-DD）。
- 合併類別：這是調查記者最常遇到的問題。例如將 “F”、“FEMALE”、“female” 這三種輸入，統一合併為單一的 “Female”。
- 修正大小寫：在處理英文數據時，利用工具函數統一英文字首大寫，避免系統將大小寫不同的單字視為不同類別。

數據完整性 (Integrity) 確保數據的準確與無誤：

- 移除重複項：使用試算表軟體的「移除重複」功能，刪除完全相同的資料列。
- 處理缺失值：識別所有空白儲存格。需要判斷是直接刪除含有空白的列（如果數據不重要），還是依據前後文或邏輯進行填補（例如複製上方的城市名稱）。
- 邏輯檢查：進行異常值檢查，確保數據在邏輯上合理。

簡單實用的清資料工具

 Google Sheets / Excel 最熟悉的選擇。適合基礎清理，利用內建公式（如 TRIM, UNIQUE）快速去除重複項與修正格式。	 OpenRefine 強大的開源工具。擅長處理大量雜亂數據，其「分群 (Clustering)」功能可快速標準化相似的文字。	 Tabula PDF 救星。專為從 PDF 文件中提取表格而設計，能將鎖在文件中的數據轉換為可編輯的試算表。
---	--	--

對於不是程式設計師背景的人來說，有許多工具能幫忙完成這些清理工作。

1. Google Sheets / Excel：

- 這是我們最熟悉的工具。它適合處理數據量較小、結構較為簡單的數據。
- 利用 TRIM（去除多餘空格）和 UNIQUE（找出不重複值）等內建公式，就能完成大部分基礎清理工作。

2. **OpenRefine:**

- 這是一個強大的開源桌面工具，專門處理大量且雜亂的數據。
- 它最著名的功能是「分群 (Clustering)」，能夠自動找出寫法相似但並不完全相同的文字 (例如「Apple Inc.」和「Apple INC」)，讓您一鍵標準化。

3. **Tabula:**

- 如果您經常處理 PDF 文件中的表格，Tabula 是您的救星。
- 它專門用於從 PDF 中提取表格數據，並將其轉換為可編輯的試算表格式，解救那些被鎖在文件中的寶貴數據。

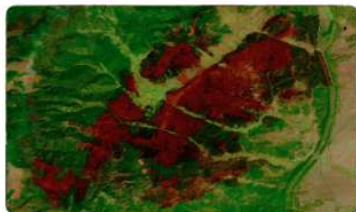
衛星影像：

「即使是靜態地圖，只要能有效傳達資訊，就能像複雜的互動地圖一樣強大。」

衛星影像的價值，不在於它的解析度有多高，而在於如何利用它來簡潔有力地傳遞核心訊息。

遙測 (Remote Sensing): 非接觸式的資料收集。不只像普通相機一樣拍下肉眼可見的「真實色彩」，還能捕捉到許多人眼看不見的光譜波段，像是紅外線。這些看不見的數據，能告訴我們地表的溫度、植被的健康程度，甚至土壤的含水量，這是傳統攝影無法做到的。

應用案例：用影像說故事



火災偵測

分析火災的燃燒痕跡 (Burn Scars)，
評估災害影響範圍。



環境監測

追蹤冰川萎縮、森林砍伐或沙漠化，
展示長期的環境變遷。



人權調查

外媒利用 Google Earth Pro 揭露集中營
位置，提供視覺證據。

1. 火災偵測：
 - 如圖所示，我們可以利用 Sentinel-2 影像來分析一場火災留下的燃燒痕跡 (Burn Scars)，精確地評估受影響的面積，並追蹤火勢的移動方向。
2. 環境監測：
 - 可以將不同年份的衛星影像進行對比，追蹤冰川的萎縮、森林的砍伐率或沙土的入侵，將長期的環境趨勢具象化。
3. 人權調查：
 - 這是衛星影像在媒體中影響力最大的應用之一。著名的案例包括 BuzzFeed News 利用 Google Earth Pro 影像，來定位和揭露一些偏遠地區的設施位置，提供了難以否認的視覺證據。

免費衛星資料來源



Copernicus Browser

歐盟太空總署託管，提供 Sentinel-2 影像。約每 5 天經過同一區域，是監測短期變化的最佳工具。



NASA Earth Observatory

提供免費、高解析度或經校正的影像，專注於重大災害或科學上有趣的特定主題。



USGS GloVis

Landsat 8/9 衛星影像的下載平臺。採集頻率約 8 天一次，是 Copernicus 的可靠備案。

1. **Copernicus Browser**: 這是歐盟太空總署 (ESA) 的專案。它提供來自 **Sentinel-2** 衛星的影像，它的優勢是採集頻率高，約每 5 天就能經過同一個區域一次，非常適合用來監測近期發生的變化，例如一場森林大火의 蔓延。
2. **NASA Earth Observatory**: 這是美國太空總署的平臺。他們提供許多經過科學家校正、易於閱讀的影像，通常專注於展示重大的環境變化或災害。
3. **USGS GloVis**: 這是 Landsat 衛星影像的主要下載平臺。Landsat 的歷史悠久，採集頻率約 8 天一次，是 Sentinel-2 的一個很好的備案，尤其當你想尋找更早期的歷史影像時，Landsat 往往能提供更長的時間跨度。

衛星資料的限制

1. 解析度與細節: Sentinel-2 影像的解析度在 10 公尺左右。這足以讓你看到一座大橋、一條燃燒痕跡或一片煙霧。
 - 但如果你的目標是看清一輛汽車的品牌或者觀察一個人，那麼免費影像就不足夠了，你需要使用解析度更高的商業級影像。
2. 採集頻率:
 - 雖然 Sentinel-2 約 5 天一次，但這不是保證。如果你需要在特定的小時甚至分鐘內觀察某個事件，免費衛星可能無法滿足需求。
3. 雲層覆蓋:
 - 衛星從太空拍攝地球，如果地面被雲層遮蔽，我們就無法獲得清晰的數據。當你在查詢資料時，可利用平臺上的「雲層覆蓋滑桿」篩選掉雲量過高的影像。