

上好味研究工作論文系列

欄位變了，詮釋資料跟上了嗎？臺灣111-114年傷亡道路交通事故開放資料的年度schema drift與機器可判讀性稽核

Did the Metadata Keep Up with the Fields? A Reproducible Audit of Annual Schema Drift and Machine Interpretability in Taiwan's 2022-2025 Road-Crash Open Data

周秉漢／CHOU, BING-HAN

ORCID：0009-0000-7018-5096

SHWRP-2026-036 | 版本1.1 | 2026-08-07

公開工作論文 | 未經外部同儕審查

DOI：10.5281/zenodo.21830987；本研究為公開工作論文，版本1.1，未經外部同儕審查。

CC BY 4.0僅適用於作者有權授權的原創分析、文字、研究程式及原創圖表；第三方名稱、商標、網址、資料、圖片、文字摘錄及其他內容仍依原權利人或來源條款使用，不因收錄於本研究而改以CC BY 4.0授權。

中文摘要

開放資料的年度檔案若新增欄位，資料目錄的機器可讀詮釋資料是否同步反映？本研究在2026年8月7日匿名讀取政府資料開放平臺JSON API，稽核內政部警政署111至114年（西元2022至2025年）傷亡道路交通事故資料。研究只保存白名單欄位、集合差異與原始回應SHA-256，不下載ZIP，也不保存逐案事故紀錄、地點、地址、聯絡資訊、精確座標或具名平台值。

四個年度資料集的description分別列出51、51、51與52個領域欄位；111至113年的欄位集合相同，114年相較113年新增1個欄位，集合Jaccard相似度為51/52，即98.08%。但四筆distribution.resourceField都只列5個通用名稱：序號、年度、name、schema與description。這個「52對5」不是資料品質分數，也不能證明ZIP內部錯誤；一個可行的替代解釋是該ZIP為清單或manifest，真正欄位由內含schema描述。作為反例，交通部112年道路交通事故30日死亡清單的CSV詮釋資料在resourceField明列9個領域欄位，顯示平臺結構可承載較具體欄位名稱。

本研究的結論限於「目錄層的兩種欄位表述存在不對稱，且年度新增欄位只在自由文字描述中直接可見」。它不評估事故資料正確性、道路安全成效、機關績效、縣市風險或政策因果效果。

關鍵詞：臺灣開放資料、道路交通事故、詮釋資料、schema drift、機器可判讀性、資料治理、可重現研究。

Abstract

When an annual open-data file gains a field, does its machine-readable catalog metadata expose that change in a structured form? This study anonymously retrieved JSON metadata from Taiwan's Government Data Open Platform on 7 August 2026 and audited the National Police Agency's annual road-crash datasets for ROC years 111–114 (2022–2025). Only allow-listed metadata derivatives, set differences, and SHA-256 hashes of the complete responses are retained. The release does not download the ZIP distributions or retain case-level crash records, locations, addresses, contact details, precise coordinates, or named platform values.

The free-text description enumerates 51, 51, 51, and 52 domain fields across the four annual records. The 2022–2024 field sets are identical; the 2025 set adds one field, yielding a Jaccard similarity of 51/52, or 98.08%, against 2024. In contrast, each annual record exposes only five generic names in distribution.resourceField. This 52-to-5 contrast is not a quality score and does not establish an error in the ZIP. A plausible alternative is that the ZIP is a manifest whose embedded schema describes the underlying files. As a counterexample, the Ministry of Transportation and Communications' 2023 30-day road-fatality CSV lists nine domain-specific names in resourceField.

The supported finding is therefore narrow: two catalog-level representations are asymmetric, and the annual addition is directly visible only in free text. The study does not assess crash-data accuracy, road-safety performance, institutional performance, geographic risk, or causal policy effects. Version 1.1 is a public working paper, has not undergone external peer review, and is identified by DOI 10.5281/zenodo.21830987.

1. 研究問題與判定邊界

本研究回答三個可由公開詮釋資料重現的問題：

- 1 警政署111至114年傷亡道路交通事故年度資料集在description中各列出多少個領域欄位？
- 2 相鄰年度欄位集合是否改變；若改變，新增與移除欄位是什麼？
- 3 description中的領域欄位表述，與distribution.resourceField的結構化欄位表述是否一致？

「機器可判讀性」在本文只指匿名程式能否從目錄API直接辨認欄位集合、版本差異與資源屬性。它不等於檔案可下載、內容正確、統計可用、隱私安全或資料適合任何特定決策。本文也不將resourceAmount=0解讀為零件事故；該值只能被記錄為目錄詮釋值，不能取代ZIP內實際列數。

研究單位是五筆目錄詮釋資料：四筆警政署年度傷亡事故資料，以及一筆交通部112年30日死亡清單反例。這不是臺灣交通開放資料普查，沒有對發布者、縣市、事故原因、平台、企業或產品排名。

2. 資料、來源與權利

核心量化輸入來自政府資料開放平臺匿名GET端點：

- 警政署111年傷亡道路交通事故資料：<https://data.gov.tw/api/v2/rest/dataset/161199>
- 警政署112年傷亡道路交通事故資料：<https://data.gov.tw/api/v2/rest/dataset/167905>
- 警政署113年傷亡道路交通事故資料：<https://data.gov.tw/api/v2/rest/dataset/172969>
- 警政署114年傷亡道路交通事故資料：<https://data.gov.tw/api/v2/rest/dataset/177136>
- 交通部112年道路交通事故30日死亡清單：<https://data.gov.tw/api/v2/rest/dataset/172799>

詮釋資料欄位的制度背景參照數位發展部「資料集詮釋資料標準指引」與跨平臺GET介面說明（<https://data.gov.tw/about/doc?chapter=11&doc=4>；<https://data.gov.tw/about/doc?chapter=27&doc=8>）。國際對照使用W3C於2024年8月22日發布的DCAT 3

Recommendation（<https://www.w3.org/TR/vocab-dcat-3/>），僅用於區分dataset、distribution、catalog record與版本描述的概念。

道路安全的公共議題背景參照交通部路政及道安司2026年3月12日發布的114年交通死亡統計新聞稿（https://www.motc.gov.tw/ch/app/news_list/view?id=14&module=news&serno=b9b8dd21-c943-495f-a206-be4a47a7d087）。該新聞稿只界定議題脈絡，不進入本文欄位計數。

政府資料依來源標示的政府資料開放授權條款第1版使用；W3C文件依其文件授權條款。版本1.1公開資料包不重製完整第三方JSON、HTML、ZIP或文件全文，只保存研究必要的白名單衍生欄位與雜湊。CC BY 4.0只涵蓋作者有權授權的原創分析、文字、程式與原創圖表；第三方權利不因收錄而改變。

3. 方法

3.1 擷取與白名單化

fetch_sources.py以標準TLS驗證匿名讀取五個JSON端點。程式先對完整回應計算SHA-256，再只保留資料集ID、標題、發布與修改日期、資料品質代碼、更新頻率狀態，以及第一個distribution的格式、編碼、品質檢查時間、資料量、OAS狀態與結構化欄位名稱。聯絡人姓名、電話、電子郵件、下載URL與其他非必要欄位不落盤。

年度資料的領域欄位從description最外層括號內，以中文頓號U+3001切分。前後空白被移除，空字串捨棄；未進行同義詞合併。distribution.resourceField[].name另作結構化欄位集合。兩者不直接宣稱同一層級，只比較目錄頁可見表述。

3.2 年度差異

相鄰年度新增欄位定義為後一年集合減前一年集合；移除欄位反向計算。集合相似度採Jaccard：

$$J(A,B) = |A \cap B| / |A \cup B|$$

所有百分比由整數計數計算，四捨五入至小數點後兩位。欄位順序另以完整串列SHA-256保存，但本文不把順序差異解釋為語意差異。

3.3 反例與可否證條件

交通部112年30死亡清單不是同一資料產品，因此不拿來評比發布者；它只檢查resourceField是否能容納領域欄位。若該反例也只有通用wrapper名稱，則「平臺可承載具體欄位」的反例不成立。若ZIP內schema日後證明已完整、穩定且可直接連結至目錄欄位，本文關於「目錄層需要更清楚連結」的建議應相應縮小。

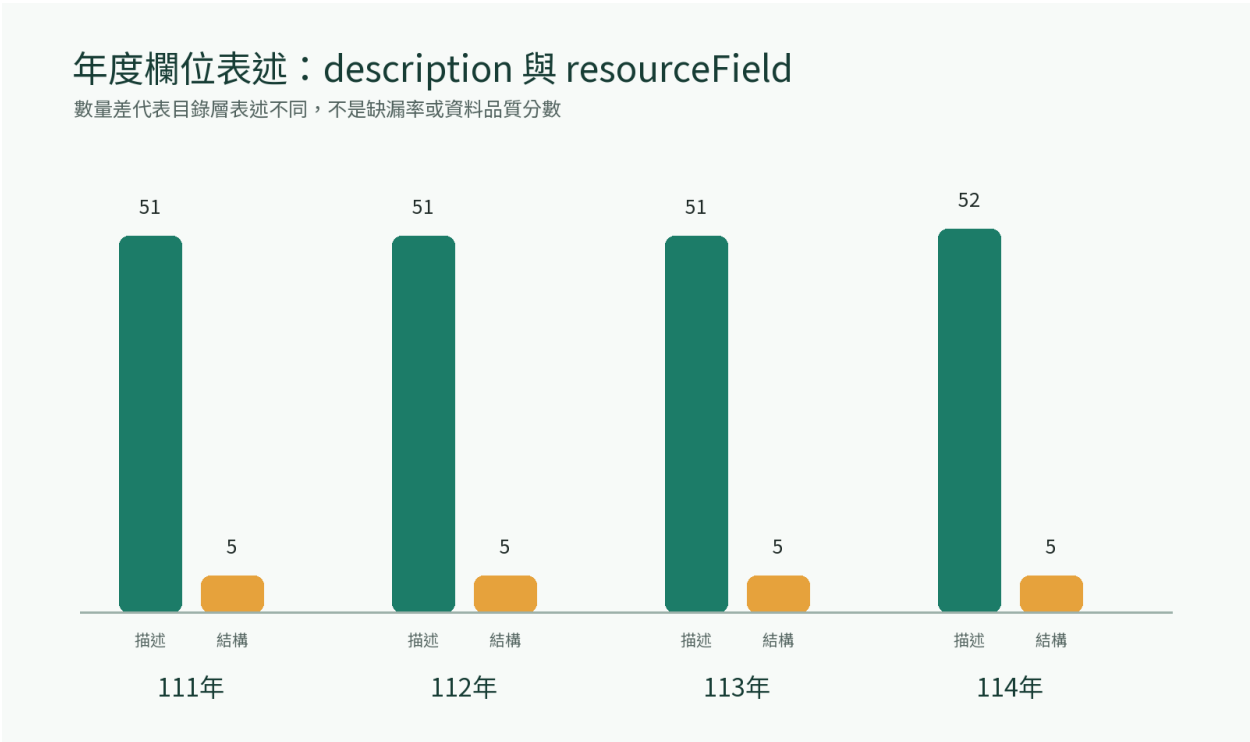


圖1：年度description領域欄位與resourceField名稱數量；作者原創圖表，不是發布者排名。

4. 結果

4.1 四個年度的description欄位數

111、112、113年各有51個領域欄位；114年有52個。111至113年的欄位集合完全相同。114年相較113年新增1個與共享經濟或外送平台名稱有關的欄位，沒有移除欄位。本文只報告欄位名稱的存在，不讀取也不發布任何平台值。

113與114年的交集為51、聯集為52，因此Jaccard相似度為0.980769，四捨五入為98.08%。高相似度與一個新增欄位可以同時成立；它表示年度schema大致延續，不表示新增欄位不重要。

年度資料	西元資料年	description欄位數	resourceField名稱數	相鄰年度新增	相鄰年度移除
111年	2022	51	5	不適用	不適用
112年	2023	51	5	0	0
113年	2024	51	5	0	0
114年	2025	52	5	1	0

4.2 結構化欄位表述與資源屬性

四個年度的distribution.resourceField均列出5個名稱：序號、年度、name、schema、description。四筆資源同時都宣告ZIP、UTF-8與品質檢查時間；resourceAmount均為0，更新頻率均標為不定期，且檔案型資源未提供OAS。OAS空白對檔案資源不是缺失判定；resourceAmount=0也不是事故列數。



圖2：五筆目錄詮釋資料的機器欄位矩陣；作者原創圖表，不是發布者排名。

交通部112年30日死亡清單反例為CSV、UTF-8，並在resourceField列出9個領域欄位。它支持一個有限主張：相同目錄資料模型可以提供領域欄位名稱。它不證明年度ZIP的5個名稱必然錯誤，因為ZIP可能包裝manifest，且兩個資料集的發布流程與粒度不同。

4.3 變動位置

114年的新增欄位直接出現在自由文字description的末端；四筆年度資料的結構化resourceField仍維持相同5個通用名稱。因此，僅依賴resourceField的目錄用戶端無法從這五個名稱辨認本次領域欄位新增。能解析中文description的程式可以發現變動，但必須依賴標點與文字格式。

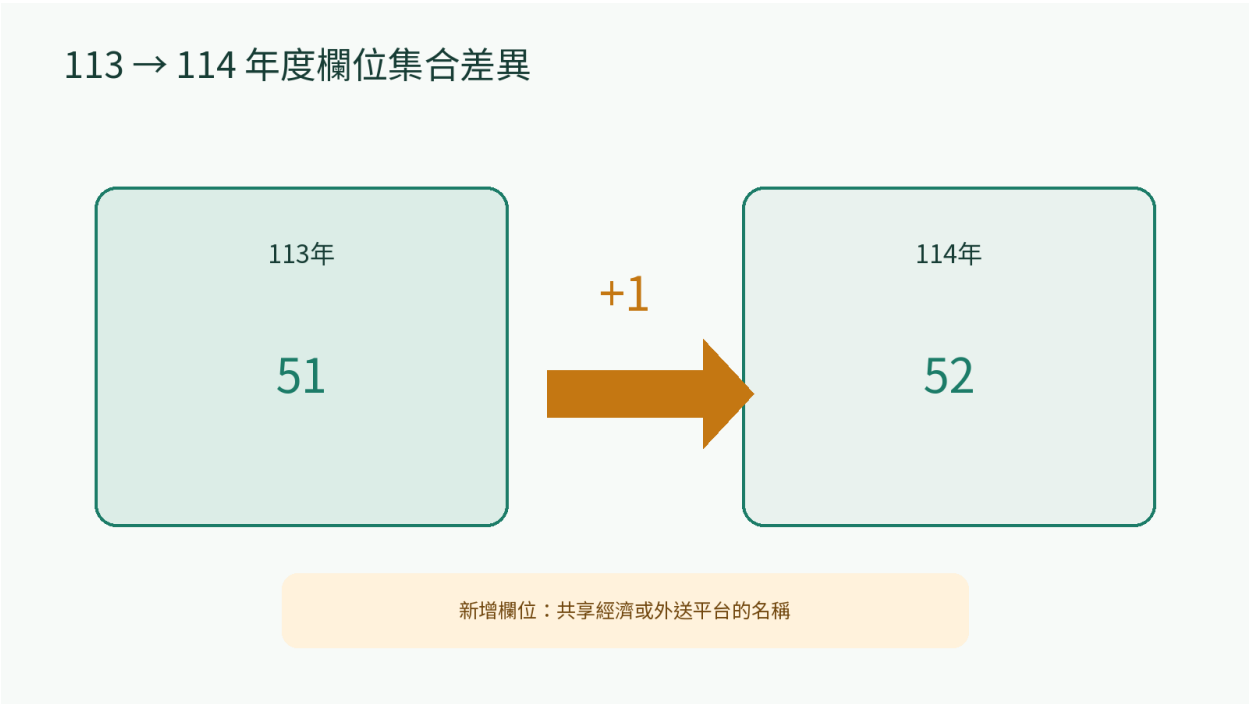


圖3：113至114年度欄位集合差異；作者原創圖表，不是發布者排名。

5. 反例、不確定性與替代解釋

第一個反例是30日死亡清單的9個具體欄位，說明結構化表述不是技術上不可能。第二個反例是四筆年度資料都具有格式、編碼與品質檢查時間；因此不能把本文簡化成「詮釋資料全部不完整」。第三個反例是111至113年集合完全相同，顯示自由文字解析在三個年度得到穩定結果，並非每年都漂移。

最重要的替代解釋是ZIP wrapper：name、schema與description可能描述壓縮檔內的檔案清單，而不是逐案CSV的資料欄。若如此，5與52屬於不同抽象層級。本文沒有下載ZIP，無法判定wrapper解釋是否正確，也不把兩數之比當成缺漏率。

其他不確定性包括：API回應是2026年8月7日截面，發布者可在之後修訂；description的頓號解析假定欄位名稱本身不含頓號；欄位名稱相同不保證型態、定義、代碼表或缺值規則相同；欄位新增不表示資料已完整填值；metadata修改日不等於底層事故資料修改日。這些限制都阻止因果、績效與內容正確性推論。

6. 可重現性與主張追溯

data/metadata-snapshot.json保存五筆白名單快照與完整回應SHA-256。analysis.py從快照重建年度欄位集合、相鄰年度差異、Jaccard值、反例表與三張圖。build_artifacts.py固定字型、頁面尺寸與PDF invariant模式產生可搜尋PDF。verify_reproducibility.py在臨時目錄離線重跑分析與PDF，逐檔比較SHA-256。

每個精確量化主張在claim-provenance.json與outputs/claim-results.csv連結至來源ID、JSON欄位、清理規則、程式步驟與輸出表。版本化ZIP排除自身、checksum與外部package-validation-report.json，避免ZIP雜湊自我引用；最終ZIP雜湊由外層收據驗證。

7. 結論與窄幅建議

本研究可支持的結論只有三點：四個年度description欄位數為51、51、51、52；114年相較113年新增1個欄位，集合相似度98.08%；四筆年度resourceField均維持5個通用名稱，而一筆不同產品的CSV反例提供9個領域名稱。

對目錄治理的窄幅建議是：若年度ZIP確為wrapper，可在目錄層提供可機器連結的內含schema或版本說明；若description仍是唯一完整欄位表，則可同步提供結構化欄位清單與年度delta。這是互通性建議，不是對機關、資料品質或道路安全政策的負面判定。

8. 限制與更正方式

本研究未下載或驗證事故ZIP內容，未計算事故件數、死傷人數、縣市率、平台分布或任何因果效果。來源錯誤、欄位更名或解析錯誤可透過本研究頁的公開來源網址重查；更正時應保留舊版、說明變更、重跑全部衍生表與PDF，並在新修訂公開前完成機器驗證。

版本1.1已透過Zenodo

new-version流程公開，並取代版本1.0作為最新版本；版本1.0仍由Zenodo保留。本稿未經外部同儕審查。

建議引用：周秉漢（2026）。〈欄位變了，詮釋資料跟上了嗎？臺灣111-114年傷亡道路交通事故開放資料的年度schema drift與機器可判讀性稽核〉。《上好味研究工作論文系列》，SHWRP-2026-036，版本1.1。https://doi.org/10.5281/zenodo.21830987

參考來源

- 1 內政部警政署。111至114年傷亡道路交通事故資料集詮釋資料。政府資料開放平臺，資料集161199、167905、172969、177136。

- 2 數位發展部。資料集詮釋資料標準指引與跨平臺介接指引。https://data.gov.tw/about/doc?chapter=11&doc;=4
- 3 交通部。112年道路交通事故30日死亡清單詮釋資料。https://data.gov.tw/api/v2/rest/dataset/172799
- 4 World Wide Web Consortium. Data Catalog Vocabulary (DCAT) — Version 3. W3C Recommendation, 2024-08-22. https://www.w3.org/TR/vocab-dcat-3/
- 5 交通部路政及道安司。114年交通死亡人數持續下降。2026-03-12。https://www.motc.gov.tw/ch/app/news_list/view?id=14&module;=news&serno;=b9b8dd21-c943-495f-a206-be4a47a7d087

權利聲明

© 2026 周秉漢。CC BY 4.0僅適用於作者有權授權的原創分析、文字、研究程式及原創圖表；第三方名稱、商標、網址、資料、圖片、文字摘錄及其他內容仍依原權利人或來源條款使用，不因收錄於本研究而改以CC BY 4.0授權。