數位發展部
Ministry of Digital Affairs

建構「臺灣主權 AI 訓練語料庫」

地方政府語料徵集研商會議

115年3月3日

簡報大綱

會議資料

一、建構臺灣主權AI訓練語料庫行動計畫

二、地方政府協力共同推動訓練語料釋出

三、執行期程及地方政府配合事項

建構臺灣主權AI訓練語料庫行動計畫

前言

目前大型語言模型正體中文訓練語料比例偏低

主流模型訓練資料多以英文與簡體中文為主，正體中文語料比例偏低，欠缺臺灣文化觀點與語境脈絡

訓練語料取得涉及著作權議題尚無明確規範

在蒐集AI模型訓練資料階段，訓練資料如受著作權法保護是否構成著作權上的重製、衍生著作或合理使用，實務上仍存在判斷標準模糊的問題

公私部門釋出中文語料主動性仍不足

許多政府機關與民間單位雖掌握大量語料資源，但對中文語料釋出的價值認知不足或缺乏釋出動機，致公私部門在中文語料釋出主動性略顯不足

我們的三大推動策略

① 建立

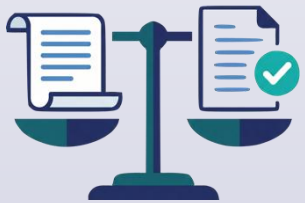
臺灣主權AI語料入口網

支援主權AI模型的訓練需求

基礎建設

② 訂定

(1)促進資料創新利用發展條例
(2)臺灣主權AI訓練語料授權條款

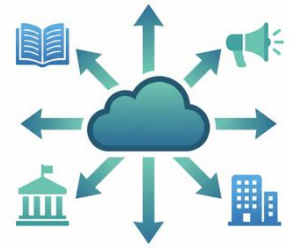


推動語料合法合規共享

法規制度

③ 強化

公私部門語料釋出動能



擴大語料徵集範疇的深度與廣度

輔導推廣

策略一：建立臺灣主權AI訓練語料入口網

https://taic.moda.gov.tw

關鍵字搜尋

快速查找所需資料

開放線上申請

為AI模型訓練者提供合法、便利存取管道

貢獻語料

推動政府機關共同豐富語料庫

徵集高品質、具在地化的正體中文語料

臺灣文化特色與觀點

包含政治、社會、經濟、語言、歷史、地理、民俗、動植物及國家文化記憶等，呈現臺灣文化特色與觀點。

正體中文塊狀資料

具備語意連貫性，內容完整且流暢，非僅由數字、圖表或條列式文字構成。

人工創作並經審核

由人工撰寫或創作，並經審核確認，來源合法合規，例如政府出版品、研究報告等。

電子資料

電子檔案格式。

策略二：推動語料合法合規共享 （1/2）

《促進資料創新利用發展條例》（草案）

擴大政府資料開放與共享，促進產業及民間資料利他運作，奠定國家數位與科技創新基礎

第二十六條

1. 標準授權

政府機關將政府資料作為開放資料提供利用，應使用標準授權條款。

2. 開放資料

政府機關依標準授權條款提供利用之開放資料，得為人工智慧及其他新興科技研發利用。

第二十七條

1. 非專屬授權

政府機關將政府資料作為共享資料提供利用，應以非專屬授權利用方式為之。

2. 共享資料

授權主管機關訂定政府資料作為共享資料非專屬授權利用之條款範本，且其範本內容應利於人工智慧及其他新興科技研發利用。

資料運用與創新最大化 ＋ **促進高價值資料釋出**

策略二：推動語料合法合規共享 （2/2）

《臺灣主權AI訓練語料授權條款-第1版》

提供一致性授權，在促進AI發展與著作權保護之間取得平衡

授權人

1. 正式授權AI訓練使用

協助被授權人在被明確同意的基礎下，使用該等語料資料於人工智慧的訓練與學習。

2. 合法應用於AI訓練

授予被授權人重製、改作、編輯及其他著作權和著作相關權利上必要之使用權，使其合法用於AI訓練。

3. 免責聲明

同意語料資料可供人工智慧訓練使用，但對其他事項不提供任何擔保。

被授權人

1. 明確規範標示義務

就該語料資料提供之相關識別資訊，如資料集名稱、資料提供者，發布年份等；另AI輸出物應註明「人工智慧生成產出」，落實透明原則。

2. 明示訓練應合法合規

訓練時應維持於合理學習範疇內，其成果不應與原語料實質近似，並不得對原語料之市場或價值造成負面影響。

3. 明定使用政策

嚴禁使用語料資料從事違法或倫理爭議行為。

策略三：強化公私部門語料釋出動能

釐清疑義與即時溝通，手把手輔導機關語料上架

第一階段：語料徵集說明會

溝通階段

舉辦13場研商會議與說明會，
凝聚機關與利害關係人共識

第二階段：語料上架工作坊

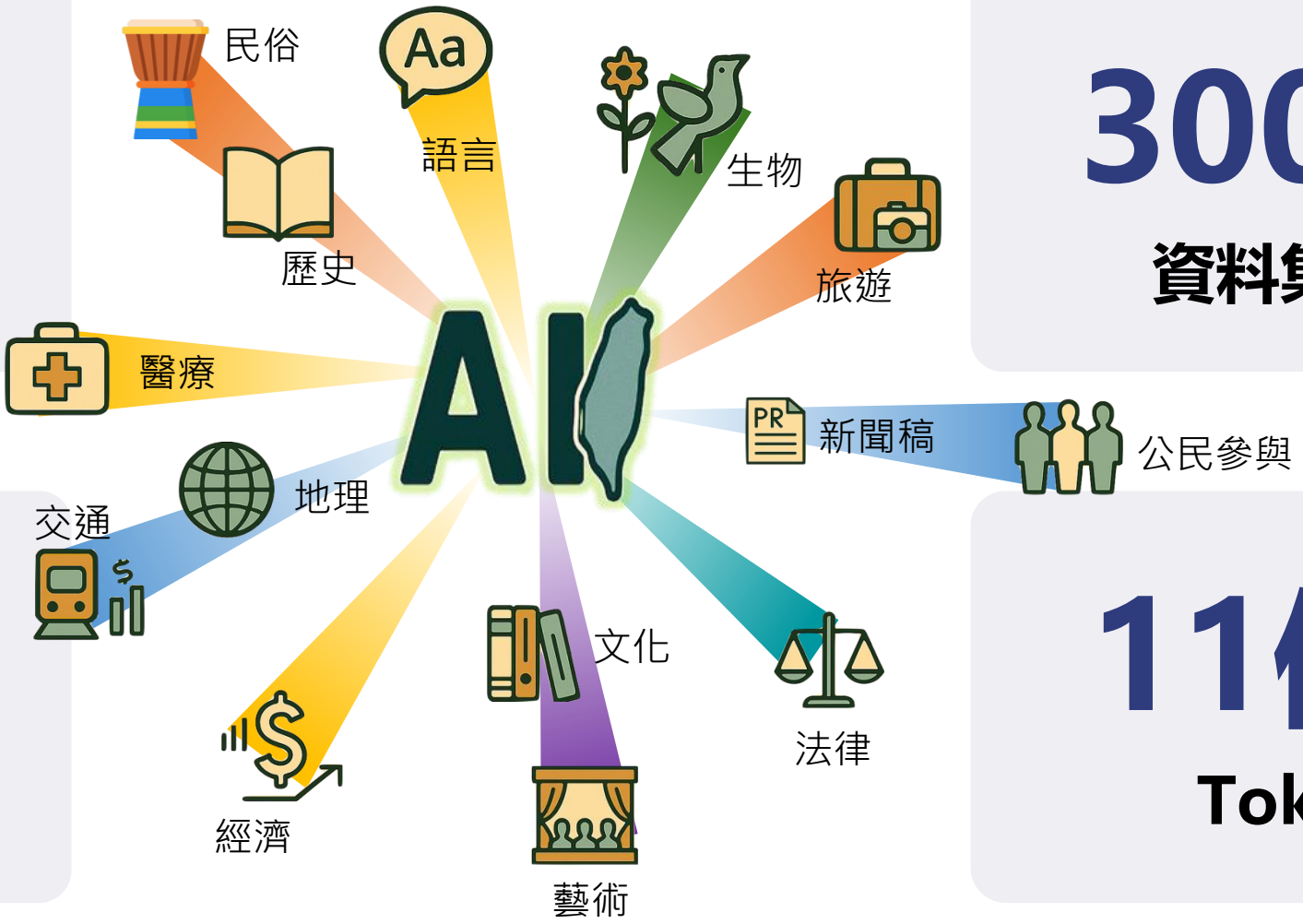
實作階段

辦理5場實作工作坊，
加速政府機關語料上架作業

攜手中央機關，打造豐富多元語料庫

200+
政府機關參與

80+GB
資料量



3000+
資料集數量

11億+
Tokens

地方政府協力
共同推動訓練語料釋出

語料徵集範疇

依下列**十大文本資料類型**（詳見附件1），系統化徵集高品質、具臺灣語言文化價值語料

政府計畫/報告
- 縣市綜合發展計畫
- 年度施政計畫

委託 / 自行研究報告
- 文化資產修復調查
- 地方產業振興研究
- 環境影響評估報告

政策與施政文件
- 政府公報
- 議會施政報告
- 各局處重要政策白皮書

活動與宣傳文本
- 活動導覽手冊
- 各類節慶活動文宣

互動式溝通語料
- 市民專線常見QA
- 客服信箱回覆訊息（須去識別化）

政府出版品
- 縣市文化期刊
- 政府定期出版各類專業手冊（如：防災手冊、長照資源地圖）
- 縣市年鑑

台灣在地文化
- 地方志
- 文化館藏資料
- 旅遊指南
- 在地動植物生態調查報告

語言資源 / 詞彙
- 本土語言推廣詞彙表
- 地方特色產業術語表
- 地名淵源檢索資料

教育與學習資源
- 訓練教材
- 科普教材
- 教育教材
- 語言學習教材

其他
- 非屬上述類型但可作為 AI 訓練語料的文本資料

臺灣主權AI訓練語料庫

語料釋出四步驟 —— 分階段推動各機關訓練語料釋出

1 盤點現有電子資料：

依十大文本資料類型，全面檢視與盤整機關內部各項資料來源。
如涉及政府資訊公開法第18條所列限制公開或不予公開之語料，若可透過技術手段予以排除，即可釋出部分資料開放或供AI訓練使用。

2 檢視資料權利完整性：

2.1 開放資料優先
2.2 僅限AI訓練使用

3 規劃語料提報時程：

完成第一階段語料上架（4至5月）：
各機關擁有完整著作財產權之政府出版品、委託研究報告等優先
完成第二階段語料上架（6至7月）：
擴大盤點範圍，經評估確認可提供AI訓練使用

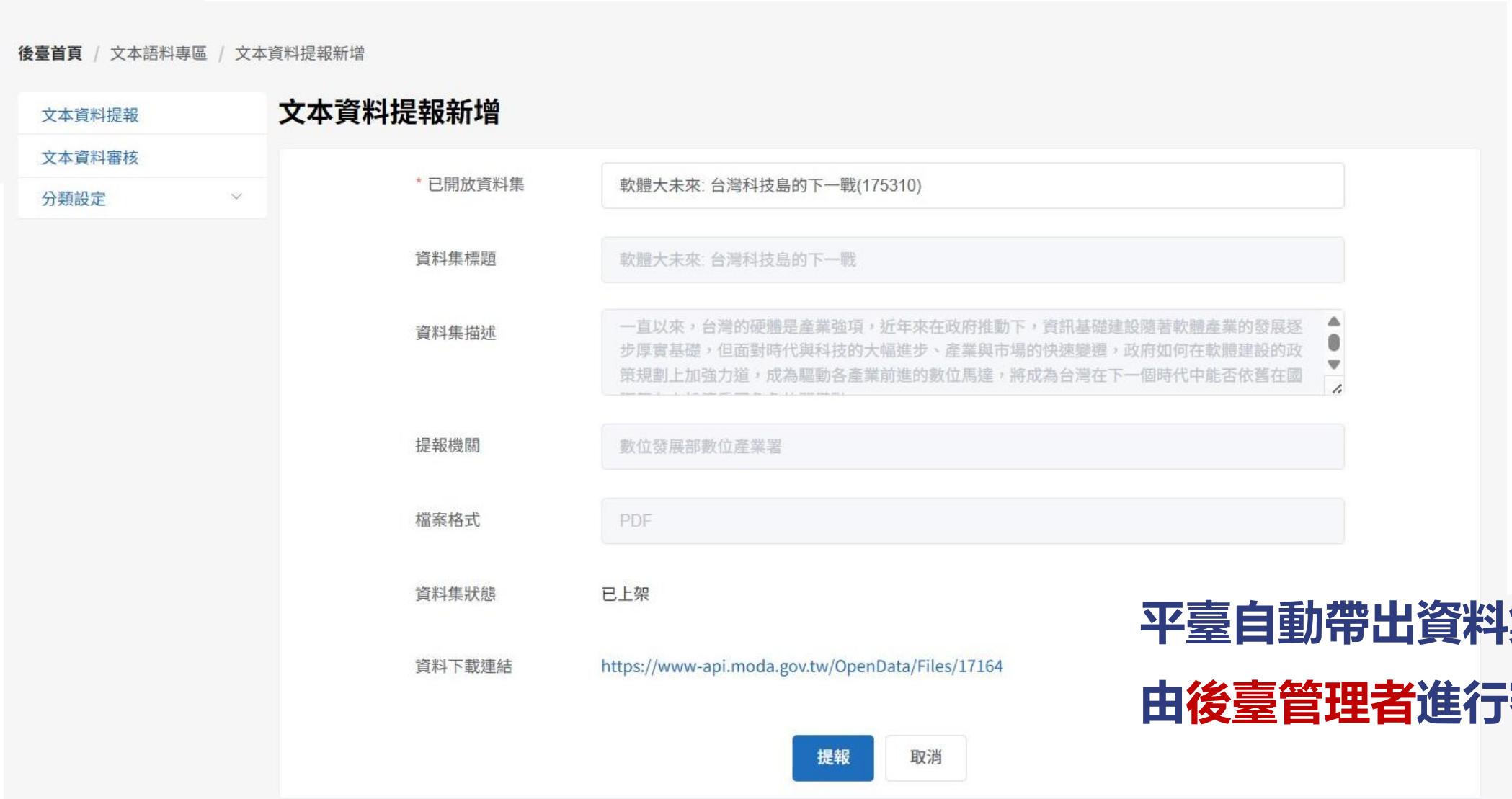
4 完備詮釋資料發布資料集：

完備詮釋資料內容：
4.1 基礎欄位：資料提供機關、資料集名稱、資料集描述、授權資訊、時間等關鍵資訊
4.2 進階欄位：檔案大小、格式等

Step1. 盤點現有資料-已上架OD平臺文本語料(1/3)

可透過後臺進行文本語料提報

Step1. 盤點現有資料-已上架OD平臺文本語料(2/3)



平臺自動帶出資料集內容，

由後臺管理者進行審核

Step1. 盤點現有資料-已上架OD平臺文本語料(3/3)

審核狀態　已上架

提報歷程

動作	執行人員	時間
新增提報單	[redacted]	2025-11-20 16:03:04
自動解析中		2025-11-20 16:05:02
自動解析完成，待審核上架		2025-11-20 16:05:03
審核通過	[redacted]	2025-11-20 16:47:09

若符合標準即審核通過，該資料集將標註為文本資料。

Step1. 盤點現有資料-未開放或釋出(1/2)

智慧財產局

機關可直接釋出無須另外授權：

1.非著作權標的之資料

依著作權法第9條規定，以下資料不受著作權法保護，機關可直接釋出，而無須另行授權。

1. 憲法、法律、命令或公文（包括公務員於職務上草擬之文告、講稿、新聞稿及其他文書）。
2. 中央或地方機關就前款著作作成之翻譯物或編輯物。
3. 單純為傳達事實之新聞報導所作成之語文著作。
4. 依法令舉行之各類考試試題及其備用試題等。

2.逾越保護期間之著作

機關管理之資料如屬已逾著作保護期間之公共財，可直接釋出而無須另行授權。

各類著作之存續期間規定於著作權法第30條至第34條，主要類型如下：

1. 一般著作：著作人生存期間及著作人死亡後50年（§30）。
2. 共同著作：若為二人以上共同完成之著作，不能分離利用者，其著作財產權存續至最後死亡之著作人死亡後50年（§31）。
3. 法人著作、攝影、視聽、錄音及表演著作：著作公開發表後50年（§33、§34）。

3.公眾領域貢獻宣告（CC0）之資料

原著作權人已拋棄著作財產權，機關即可直接釋出。

1. 「公眾領域貢獻宣告」（即CC0）或標示此圖樣：Ⓒ0 之著作。
2. 選擇其著作「不保留權利」，而將作品貢獻至公眾領域。
3. 任何人皆可以任何方式、為任何目的使用（包含商業目的）。

Step1. 盤點現有資料-未開放或釋出(2/2)

智慧財產局

請機關評估權利完整性：

4. 機關享有完整著作財產權之資料

機關可本於著作財產權人之地位，決定採用授權條款將著作提供AI訓練與學習。

1. 機關委外採購就委外廠商履約完成之著作，依照雙方所簽訂的採購契約，若約定「機關為著作人」、「機關取得著作財產權」或「（全部）著作財產權讓與機關」等。
2. 履約完成之著作若內容涉及第三人著作，機關仍需向該第三人（即著作財產權人）洽取授權及再授權之權利，始得另行授權供AI訓練與學習。
3. 公務員職務上完成之著作（如非屬著作權法第9條的研究報告等），依著作權法第11條規定，在沒有特別約定的情形下，其著作財產權歸機關享有者，機關可決定採用授權條款將該著作提供AI訓練與學習。

5. 機關無著作財產權或未享有完整著作財產權之資料

機關應向該著作財產權人洽取授權及再授權之權利，始得另行授權供AI訓練與學習。

1. 僅是其他著作人授權給機關在被授權的範圍內利用，機關並無著作財產權。
2. 著作人僅讓與機關部分的著作財產權。

Step2. 依資料可授權範圍評估適用授權類型

不限目的自由應用

政府資料開放授權條款 - 第1版 (OGDL-1.0)

1. 授權使用者**不限目的**、時間及地域、非專屬、不可撤回、免授權金進行利用，包括重製、改作、散布、公開播送等。
2. 資料提供機關擁有完整著作財產權，或經授權得再轉授權第三人利用之資料，並可依本條款授權使用者再轉授權他人利用。
3. 使用者應明確標示原資料提供機關之相關聲明。
4. 授權條款詳細內容請參閱政府資料開放平臺（https: / / data.gov.tw / license ）

僅授權AI訓練使用

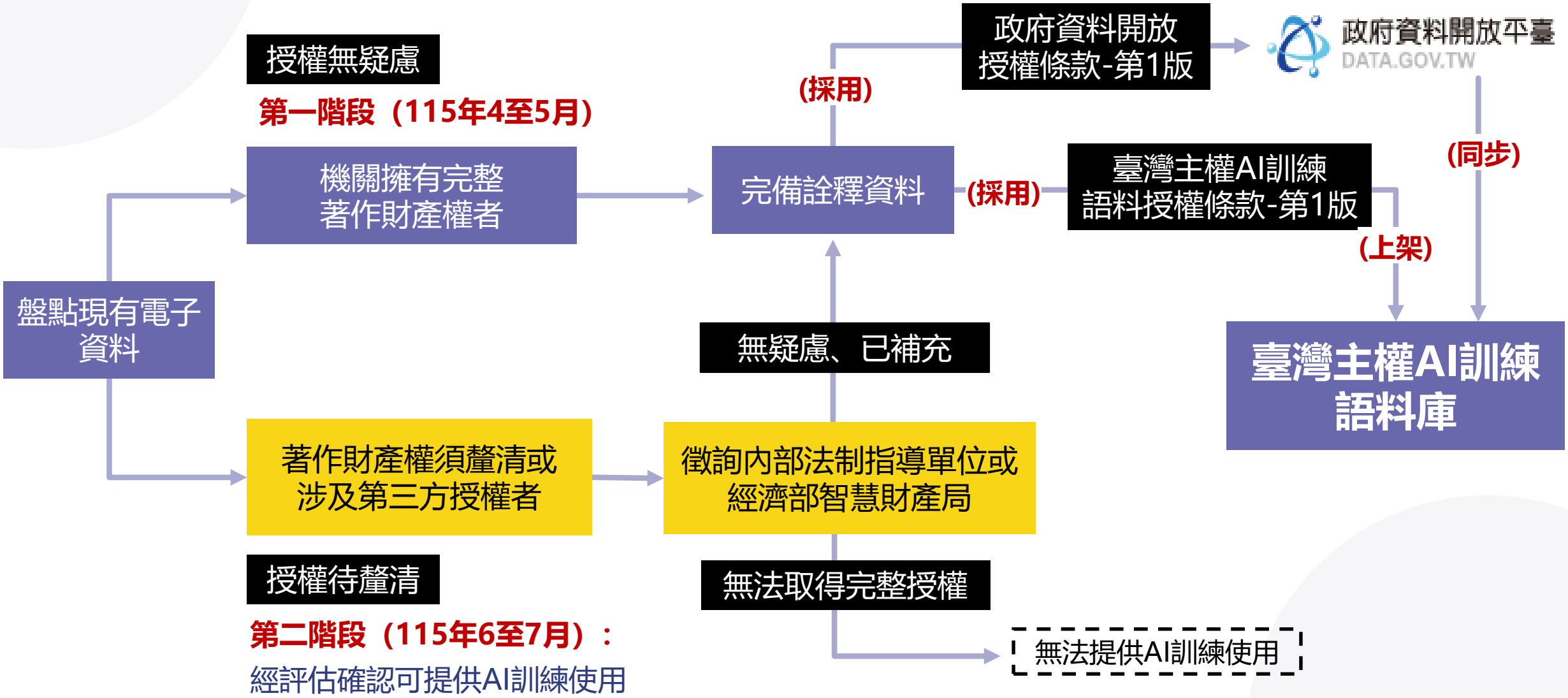
臺灣主權AI訓練語料授權條款 - 第1版 (SAITDL-1.0)

1. 明確授權語料資料**僅限AI訓練使用**，訓練成果與原作授權脫鉤，允許後續成果自由運用。
2. 資料提供機關擁有完整著作財產權，或經授權得再轉授權第三人利用之資料。
3. 被授權人應明確標示語料資料出處資訊，包含資料提供者、資料集名稱、年份等，毋須就原語料資料內含個別素材逐項標示。
4. 授權條款詳細內容請參閱臺灣主權AI訓練語料平臺(https://taic.moda.gov.tw/content/license)

依授權類型上架至對應平臺

- 非屬著作權標的者
- 政府機關享有著作財產權者
- 已取得第三方完整授權者

Step3. 規劃語料釋出時程



Step4. 完備詮釋資料、發布資料集

資料集樣態	樣態1	樣態2	樣態3	樣態4	樣態5
資料集提供機關	範例：外交部	範例：數位發展部數位產業署	範例：金融監督管理委員會	範例：文化部	範例：考試院
資料聯絡人	OOO	OOO	OOO	OOO	OOO
聯絡人電話	oo-oooooooo	oo-oooooooo	oo-oooooooo	oo-oooooooo	oo-oooooooo
資料類型	**政府出版品**	**政府出版品**	**活動與宣傳文本**	**委託研究報告**	**政府出版品**
資料集名稱	中華民國外交年鑑	112 Taiwan數位內容產業報告	金管會新聞稿	因應高齡化少子化之文化政策及資源調整策略委託研究報告	考銓季刊
資料集描述	彙整外交部及各駐外館（處）歷年各項外交業務及活動，章節分為本文、外交活動圖片及附錄三部分。	2022年全球數位內容產業市場規模約為25,121億美元，較2021年同比成長約7.28%，至2026年預計增加至29,322億美元。2022年疫情衝擊降低，除了傳統電視與家庭影音產業呈現負成長外（受OTT影音產業替代效應影響），其餘數位內容次產業皆呈現成長趨勢；就次產業的表現與趨勢來檢視，主要有五個軸向的觀察，提供讀者掌握各行各業數位內容產業的發展動向。	本會暨四局涉及金融機構相關政策、財務資訊、消費者保護等重要事項，提供對外新聞稿。	為瞭解我國文化領域面臨高齡化少子化等人口結構轉變下的衝擊，例如影視音及出版產業如何針對不同人口進行節目製作或書籍出版、文化創意產業因應人口變遷會面臨之衝擊與解決方式等問題本研究以文化政策、文化創意產業以及「黃金人口參與村落文化發展計劃」等三項作為研究主軸，研究未來我國文化政策的執行策略，期望能透過政策的調整滿足全民的文化需求。	彙整考試院《考銓季刊》2004-2008年發表期刊，內容涵蓋我國人事制度、考選與銓敘政策等研究與實務探討。
關鍵字（以半形逗號隔開）	外交施政方針與計畫,對外關係,領事事務,外交行政,外交大事日誌,外交活動圖片,總統之外交言論摘錄	政府出版品,科學,社會,財政經濟,國家發展及科技	新聞稿	文化政策,文化產業,高齡社會,文創,社區營造	考銓領域,考選,銓敘,期刊
授權方式	□ **政府資料開放授權條款** ■ **臺灣主權AI訓練語料授權條款** □ **授權需確認**	■ **政府資料開放授權條款** □ **臺灣主權AI訓練語料授權條款** □ **授權需確認**	■ **政府資料開放授權條款** □ **臺灣主權AI訓練語料授權條款** □ **授權需確認**	□ **政府資料開放授權條款** ■ **臺灣主權AI訓練語料授權條款** □ **授權需確認**	□ **政府資料開放授權條款** □ **臺灣主權AI訓練語料授權條款** ■ **授權需確認**
發佈年份	1988-2023	2025	2013-2025	2016	2004-2008
檔案格式	PDF	PDF	JSON	PDF	PDF
資料集檔案大小	2.61 GB	61.2 MB	2.17 MB	6.33 MB	50MB
規劃釋出時程	第一階段	第一階段	第一階段	第一階段	第二階段
備註					

4月10日前填妥語料盤點清單之詮釋資料內容，以利後續語料上架作業。

執行期程及地方政府配合事項

115年執行期程規劃

地方政府

- 盤點現有電子資料
- 檢視資料權利完整性
- 規劃語料釋出時程

3至4月

地方政府

完成**第一階段**語料上架

4至5月

地方政府

完成**第二階段**語料上架

6至7月

持續推動

- 推動政府機關語料釋出（**數發部、政府機關**）
- 調整既有授權與新合約調適作為（**政府機關**）
- 研訂語料貢獻獎勵機制（**數發部**）
- 精進語料庫相關功能與機制（**數發部**）
- 建立民間語料貢獻機制（**數發部**）
- 推動民間語料釋出（**數發部、民間**）

1至3月

- 辦理**2**場語料徵集說明會(1、2月)
- 辦理地方政府研商會議(3月)
- 推動地方高品質訓練語料釋出

數發部、地方政府

4月

- 辦理**2**場語料上架工作坊
- 手把手帶領機關語料上架

數發部、地方政府

地方政府語料盤點及釋出三階段

1

提交語料盤點清單

115年4月10日前

- 提交語料盤點清單（附件2），寄送至 tsaitc@moda.gov.tw
- **推派1位語料庫窗口**

2

第一階段語料上架

115年4-5月

- 完成第一階段語料（授權無疑慮之電子資料）上架

3

第二階段語料上架

115年6-7月

- 完成第二階段語料（確認授權狀態無疑慮之電子資料）上架，並持續推動所屬機關及法人單位

歡迎地方政府踴躍上架語料

共同助力主權AI發展與應用

- **語料庫後臺：**
  https://cms.taic.moda.gov.tw
- **聯絡電話：0800-023-300**
- **聯絡信箱：tsaitc@moda.gov.tw**

擴大高品質文本語料徵集，拓展語料來源，提升語料多元性。

涵蓋臺灣語言文化價值及
教育等多元範疇

- **歷史與文化記憶**
- **族群及地方文化**
- **宗教與信仰**
- **藝術與文學**
- **語言與詞彙教材**
- **教育與學習資源**
- **文化資產**
- **生物多樣性**
- **其他……**

提供具臺灣觀點、文化特色高品質語料，支援主權AI模型的訓練需求

Q & A

意見交流