# 國立陽明交通大學 AI 模型及共享資料集管理規定

# 資料集管理作業

資料集之詮釋資料規範

為最大化資料集之共享，促使資料集於使用者之流通，資料集之內容應至少包含以下欄位，並採用JSON/XML/CSV等機器可讀檔案類型為詮釋資料之格式。

(1) 主題群(Group)

(2) 資料集名稱(name)。

(3) 資料集內容描述(description)。

(4) 資料集上架時間(createtime)。

(5) 資料集共享範圍及提供使用方式，例如無需申請、需提出申請、可下載、僅供VPN使用等。

(6) 資料集標籤(Tags)。

(7) 資料集來源(creator)。

(8) 其他：例如資料是否涵蓋個資與隱私、資料檔案大小與類型、標記資料的來源與類型等。

# 資料集上架之FAIR原則規範

資料集上架應符合 FAIR 原則 (Findable 、Accessible、Interoperable、Reusable)，以利資料再利用，其原則說明如下：

(1) 可搜尋(Findable)：新增或編輯資料集時須建立標籤 (tag)、額外索引資訊 ，本平台據以建立資料庫索引，方便使用者搜尋。

(2) 可近用(Accessible)：本平台之資料集可透過API 方式搜尋及下載，機敏資料下載需要使用授權令牌(token)方得存取。

(3) 可互操作(Interoperable)：資料與其詮釋資料必須使用標準的語彙及格式，建議使用本體論 (ONTOLOGY) 來描述資料，例如臨床醫學詞彙可參考 <https://bioportal.bioontology.org/ontologies/SNOMEDCT>，資料集須提供 pdf 文件說明資料欄位格式及利用方式

(4) 可再利用(Reusable) ：資料集必須充分描述資料及使用授權，並可追溯資料來源，資料與其詮釋資料符合相關領域本體論詞彙。

# 模型管理作業

模型之詮釋資料規範

模型上架本平台，應提供詮釋資料(metadata)，內容應至少包含以下欄位，並建議採用JSON/XML/CSV等機器可讀檔案類型為詮釋資料之格式，以利模型之交換與共享。

(1) 類別(Group)：模型(AI model)。

(2) 模型名稱(name)。

(3) 模型內容描述(description)。

(4) 模型上架時間(createtime)。

(5) 模型共享範圍及授權方式(license)，例如：一般商業授權、自由開源授權等。

(6) 模型關鍵字(Tags)。

(7) 模型來源(creator)：自主開發、修改或引用他人模型等。

(8) 模型支援之框架(model framework)與版本，如Python Function、R Function、Keras、PyTorch、Spark MLlib、TensorFlow等。

(9) 訓練資料集名稱、來源、資料量、訓練環境說明、訓練結果說明等資訊。

(10) 測試資料集名稱、來源、資料量、測試環境說明、測試結果說明等資訊，若資料集已上架於本平台，則應提供相關連結。

# 模型測試

為確保模型之可用性，模型上架於本平台前，應對模型進行相關測試，並將測試結果利用 CSV 檔案方式上傳本平台，內容應包含以下相關資訊：

(1) 所使用之測試資料來源：需提供測試資料的資料來源，包含以下內容，若測試資料集已上架於本平台，則應提供相關連結。

A. 取得方式(如：網址、檔案等)。

B. 資料是否包含個資與隱私、資料授權方式等。

C. 檔案大小(包含：單一圖片長寬、結構化資料筆數、總量大小)。

D. 檔案類型(結構化資料：文字、數字；非結構化資料：影像、視訊)。

E. 標記資料來源(網址、檔案)、類型(pixel-wise、 bounding box、classes)。

(2) 運算資源評估：需提供資源使用情況。包含：

A. GPU或CPU運算與所使用核心數。

B. 記憶體使用量。

C. 模型之使用空間。

(3) 使用方法

A. 評估測試結果之程式碼。

B. 做推論需要之程式碼。

(4) 預期結果：需提供前述條件下的預期結果，包含:

A. 輸出類型(如：bounding box、classes等)。

B. 結果類型(如：正確率、recall、precision、f1、miss rate)。

C. 測試結果(如：accuracy=95%、F1 score=80)。

# 模型效能度量

除模型測試相關資訊外，建議模型上架前，亦可針對模型進行效能度量，並將結果利用CSV檔上傳本平台，唯效能量測所使用的訓練資料應盡可能與測試資料互斥或各自獨立，訓練資料與測試資料之劃分也應盡可能保持資料分佈一致性，避免因資料劃分過程引發偏差。效能量測結果應包含以下相關資訊：

(1) 所使用之訓練資料來源的背景說明：應提供訓練資料的資料來源，包含以下內容，若訓練資料集已上架於本平台，則應提供相關連結。

A. 取得方式(如：網址、檔案等)。

B. 資料是否包含個資與隱私，資料授權方式。

C. 檔案大小(包含：單一圖片長寬、結構化資料筆數、總量大小)。

D. 檔案類型(結構化資料：文字、數字；非結構化資料：影像、視訊)。

E. 標記資料來源(網址、檔案)、類型(pixel-wise、bounding box、classes)。

(2) 所使用之測試資料來源，可供使用者測試：應提供測試資料的資料來源，包含以下內容，若測試資料集已上架於本平台，則應提供相關連結：

A. 取得方式(如：網址、檔案等)。

B. 資料是否包含個資與隱私，資料授權方式

C. 檔案大小(包含：單一圖片長寬、結構化資料筆數、總量大小)。

D. 檔案類型(結構化資料：文字、數字；非結構化資料: 影像、視訊)。

E. 標記資料來源(網址、檔案)、類型(pixel-wise、bounding box、classes)。

F. 測試檔案輸入格式。

(3) 模型使用之運算資源評估：需提供資源使用情況。包含：

A. GPU或CPU運算與所使用核心數。

B. 記憶體使用量。

C. 模型之使用空間。

(4) 使用者將得到的預期測試結果：需提供效能度量下的各項結果。

A. 用途(如：分類、物件偵測、對話機器人、語意分析等)。

B. 執行時間。

C. 結果類型(如：正確率、recall、precision、f1、miss rate等)。

D. 測試結果(如：accuracy=95%、F1 score=80等)。

E. 輸出類型(如：bounding box、classes等)。