開放資料的FAIR原則 與 歐盟資料品質指引

柯皓仁

國立臺灣師範大學圖書資訊學研究所教授

內容大綱

- (科學)開放資料(管理)的FAIR原則
- 歐盟資料品質指引

(科學)開放資料(管理)的FAIR原則

科學資料管理的FAIR原則

- 2016年，《FAIR科學資料管理和管理指導原則(**FAIR Guiding Principles for scientific data management and stewardship**)在《科學資料(Scientific Data)》期刊上發表
- 作者旨在提供指導方針，以提高數位資產的可查找性(Findability)、可近用性(Accessibility)、互通性(Interoperability)和再使用性(Reuse)
- FAIR原則強調機器可操作性(即計算機系統無需或僅需最少人工干預即可查找、訪問、互操作和重用資料的能力)，因為隨著資料量、複雜性和產生速度的增加，人們越來越依賴計算機的支持來處理資料

科學資料管理的FAIR原則 (續)

- 可查找性(Findability)：使用或再使用資料的第一步是找到它們。詮釋資料和資料本身應該很容易被人類和電腦找到。機器可讀的詮釋資料對於自動發現資料集和服務至關重要
  - F1. 詮釋資料和資料本身擁有全域唯一永久識別碼
  - F2. 使用豐富的詮釋資料描述資料
  - F3. 詮釋資料包含它們所描述資料的識別碼
  - F4. 詮釋資料和資料本身在可搜索的儲存庫中註冊或編製索引

科學資料管理的FAIR原則 (續)

- 可近用性(Accessibility)：一旦使用者找到所需的資料，使用者需要知道如何取用這些資料，可能包括身份驗證和授權
  - A1. 詮釋資料和資料本身可藉由其識別碼，使用標準化通信協定進行檢索
    - ✓A1.1 該協定是開放的、免費的、可普遍實施的
    - ✓A1.2 該協定允許在必要時進行身份驗證和授權
  - A2. 即使資料不再可用，也可以取用其詮釋資料

科學資料管理的FAIR原則 (續)

- 互通性(Interoperability)：資料通常需要與其他資料整合。此外，為了分析、儲存與處理，資料需要與應用程式或工作流程進行互操作
  - I1. 詮釋資料和資料本身使用正式的、可取用的、共用的和廣泛適用的語言來表徵知識
  - I2. 詮釋資料和資料本身使用遵循 FAIR 原則的詞彙表
  - I3. 詮釋資料和資料本身包括對其他詮釋資料和資料的限定引用 (qualified reference)。

科學資料管理的FAIR原則

- 再使用性(Reusability)：FAIR 原則的最終目標是優化資料的再使用。為此，應詳細描述詮釋資料和資料本身，以便可以在不同的情境中複製和/或組合它們。
  - R1. 詮釋資料和資料本身應採用多個準確和相關的屬性來詳細描述。
    - R1.1 詮釋資料和資料本身應連同清晰且可取用的「資料使用授權(data usage license)」一同發佈。
    - R1.2. 詮釋資料和資料本身應包含「詳細出處/溯源(detailed provenance)」。
    - R1.3. 詮釋資料和資料本身應符合與領域相關的社群標準。

歐盟資料品質指引

歐盟資料品質指引與FAIR

Findability	Accessiblity	Interoperability	Reusability
Completeness	Accessiblity/availability	Conformity/compliance	Timeliness
Findability		Machine readability/ processability	Consistency
		Openness	Accuracy
			Relevance
			Understandability
			Credibility

可查找性

- 完整性(Completeness)：如果資料包含了表徵一個實體(Entity)所需的所有屬性欄位，則資料是完整的。對於資料本身，完整性常表示資料的屬性欄位盡可能不要有空值。在詮釋資料層面，完整性表示詮釋資料應盡可能完整地描述資源而不要有留空的詮釋資料欄位。
- 可查找性(Findability)：資料集應該可被使用者和電腦代理人發現。資料集的可查找性取決於詮釋資料中的描述，資料描述得越好(如：運用控制控詞彙和關鍵字)，使用者就越容易找到資料。
  - 更名為可發現性(Discoverability)

可近用性

- 可近用性/可得性(Accessibility/availability)：可近用性係指使用者和電腦代理人是否可以在沒有錯誤或近用限制的情況下檢索開放平臺入口網站或資料集的內容。

互通性

- 合規性(Conformity/compliance)：合規性係指資料和詮釋資料遵循公認的標準，例如資料獲取、發布和描述的標準，其中包含資料集和其詮釋資料值合乎標準(如日期格式符合ISO8601)，或是資料集詮釋資料格式符合DCAT。資料或詮釋資料中有效的日期格式也表示符合性。
- 機器可讀性/可處理性(Machine readability/processability)：本指標用以評估資料集和詮釋資料可以被自動化程序理解和處理的程度。
- 開放性(Openness)：資料的開放性對開放資料的概念至關重要。開放性係指資料集乃是以非專屬格式提供，並且可以在開放授權下使用。

再使用性

- 及時性(Timeliness)：及時性係指資料集本身和詮釋資料是最新的、且能反映實際和當前情況。易言之，資料集和詮釋資料必須根據現況及時修改。
- 一致性(Consistency)：一致性係指資料和詮釋資料不包含任何矛盾。矛盾的例子包含資料集中具有重複的資料、資料集具有多個且彼此矛盾的授權聲明，或資料修改日期早於創建日期。
- 正確性(Accuracy)：從詮釋資料的觀點而言，正確性係指能以詮釋資料盡可能精確地描述資料集內容，使潛在使用者能夠對資料有實際的認識，並能夠快速評估其與自身情境的相關性。從資料的觀點而言，則是指資料集能正確地描述實體。

再使用性(續)

- 相關性(Relevance)：相關性係指資料集能與潛在使用者相關，資料集應僅包含支持資料產生與開放之目的所需要的資訊。相關性描述了資料有用和適用的程度，以及資料量是否適當。雖然這個指標高度依賴於使用者的感知和擬進行之任務，但資料提供機關可檢視資料產生與開放之目的與所提供資料間的相關性。
- 可理解性(Understandability)：可理解性係指資料和詮釋資料對使用者來說清晰易懂，則它們是可理解的。資料提供機關可以檢視其資料描述、標題和關鍵字、說明文件的品質，以提高可理解性。
- 可信度(Credibility)：可信度係指資料被使用者視為真實可信的程度，如果資料基於可信賴的來源，則被認為是可信的。資料提供機關可以檢視其詮釋資料中是否清楚描述資料發布者、聯繫窗口、資料集擁有者，以提高其資料集的可信度。

FAIR指標驗證(1/3)

113年建立自動化功能，進行FAIR指標驗證

	詮釋資料相關指標		
	檢測項目	**檢測方式**	**檢測建議**
1	**空間範圍是否填寫**	檢查詮釋資料「空間範圍」欄位是否為空	若為空建議填寫
2	**資料收錄時間是否完整**	檢查詮釋資料「開始收錄日期」、「結束收錄日期」欄位是否為空	為空者建議填寫
3	**關鍵字填寫情形**	檢查詮釋資料「關鍵字」欄位是否為空	若為空，提供關鍵字填寫建議
4	**資料集分類是否符合建議**	依據平臺「服務分類建議」結果，不符合建議且未確認調整之分類	提供服務分類修改建議

FAIR指標驗證(2/3)

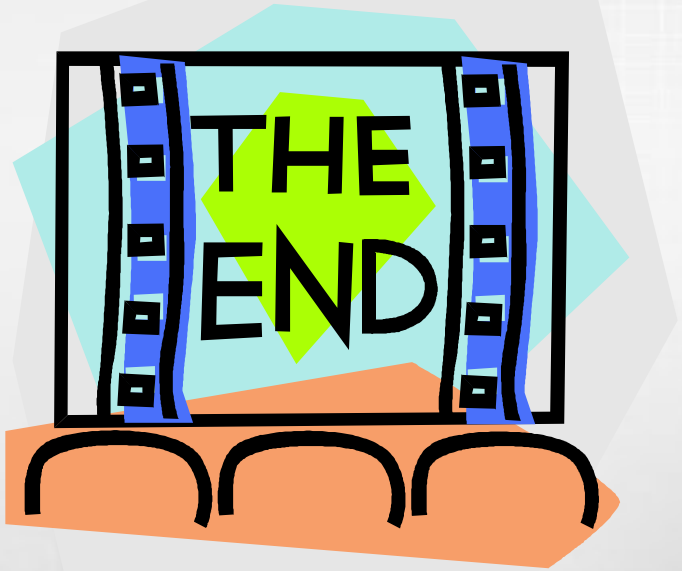
113年建立自動化功能，進行FAIR指標驗證

	詮釋資料相關指標		
	檢測項目	**檢測方式**	**檢測建議**
5	**資料集描述填寫內容是否與標題相近**	計算「資料集標題」與「資料集描述」之間的文本相似度，計算值介於0到1之間，越接近1表示越相似	相似度高於0.8者建議修改
6	**欄位說明是否填寫**	檢查詮釋資料「資料資源欄位」欄位是否為空	若為空，建議機關填寫
7	**資料資源下載網址是否有效**	檢查資料資源下載網址的連線是否有異常	若為異常，建議機關修正
8	**資料欄位是否未引入資料標準**	比對「資料標準」與「主要欄位說明」，是否有可引入資料標準但尚未引入之欄位	提供資料標準引用建議

FAIR指標驗證(3/3)

113年建立自動化功能，進行FAIR指標驗證

	詮釋資料相關指標		
	檢測項目	**檢測方式**	**檢測建議**
9	**結構化資料格式是否有效**	使用工具函式庫(Tool Library)驗證結構化資料(csv/xml/json)格式是否有效(valid)	若回報錯誤則建議調整結構化資料
10	**資料集是否有依更新頻率提供更新資料**	檢查資料集是否有依更新頻率提供更新資料	不符合檢測頻率者，建議提供更新資料
11	**資料數量是否過多或過少**	1. 非API之資料集，提供之資料筆數超過1萬筆 2. 資料筆數未滿6筆者	筆數超過：建議可轉為API方式提供 筆數過少：建議增加資料數
12	**資料欄位是否過少**	資料欄位未滿5欄	若過少建議調整



Thank you for Listening

ChatGPT協助部分翻譯工作