

depositar
研究資料寄存所

研究資料管理概論

Getting Start to Manage Your Research Data

2021-10-07, Taipei, Taiwan

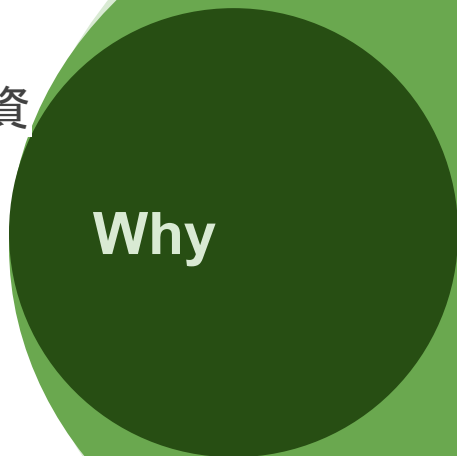
王家薰 Chia-Hsun Wang
allywang@iis.sinica.edu.tw

「研究資料寄存所」是中央研究院資訊所、資創中心、GIS 中心
(人社中心)的協作專案, 部份經費來自台灣科技部計畫



大綱

1. 為什麼需要研究資料管理(RDM)?



2. 該如何進行?
How

3. 有什麼資源?
What

現實面

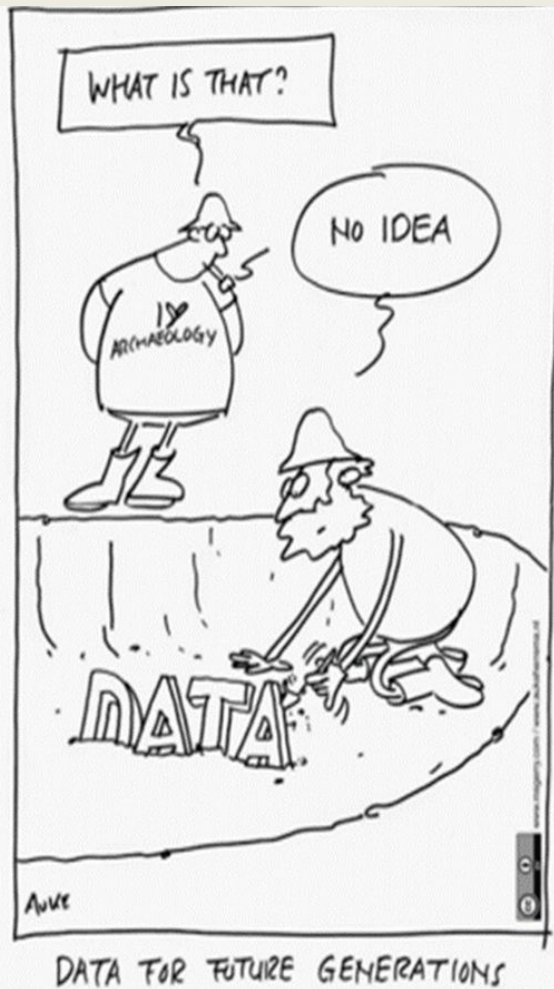
實驗室同仁離職？

人員流動率高

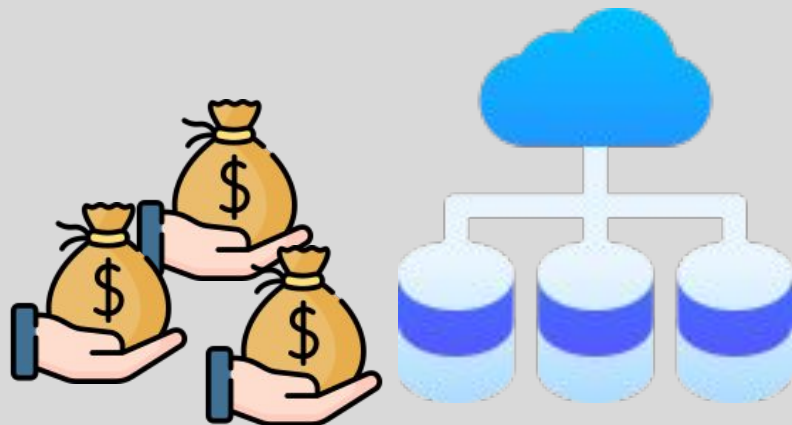
找不到助理？



你需要使用之前的舊資料
被要求復驗你的實驗結果



- 實驗室 NAS 被入侵資料被鎖住？
- 雲端服務開始收費？
- 龐大的資料記錄該何去何從？



做研究已經夠忙了
你是不是想要一個天使教母
幫你一鍵解決所有問題？





研究資料

- 資料是有價值的
- 資料亦是脆弱的
- 跨領域、跨國協同研究的趨勢
- 重視研究質量及效率
- 研究的再現性 (Reproducibility)



Frances Arnold ✓

@francesarnold



It is painful to admit, but important to do so. I apologize to all. I was a bit busy when this was submitted, and did not do my job well.



Frances Arnold ✓ @francesarnold · Jan 3, 2020

For my first work-related tweet of 2020, I am totally bummed to announce that we have retracted last year's paper on enzymatic synthesis of beta-lactams. The work has not been reproducible. science.sciencemag.org/content/364/64...

2:02 AM · Jan 3, 2020 · Twitter Web App

<https://twitter.com/francesarnold/status/1212796266494607360>

FAIR 原則

F

Findable
可被找到

其他人可以找到您的資料 (Metadata、PID等)。

A

Accessible
可被取用

在適當的特定條件或限制下，人和機器應該可以取得的資料。即使資料本身不再可取得，後設資料也仍可取用。

I

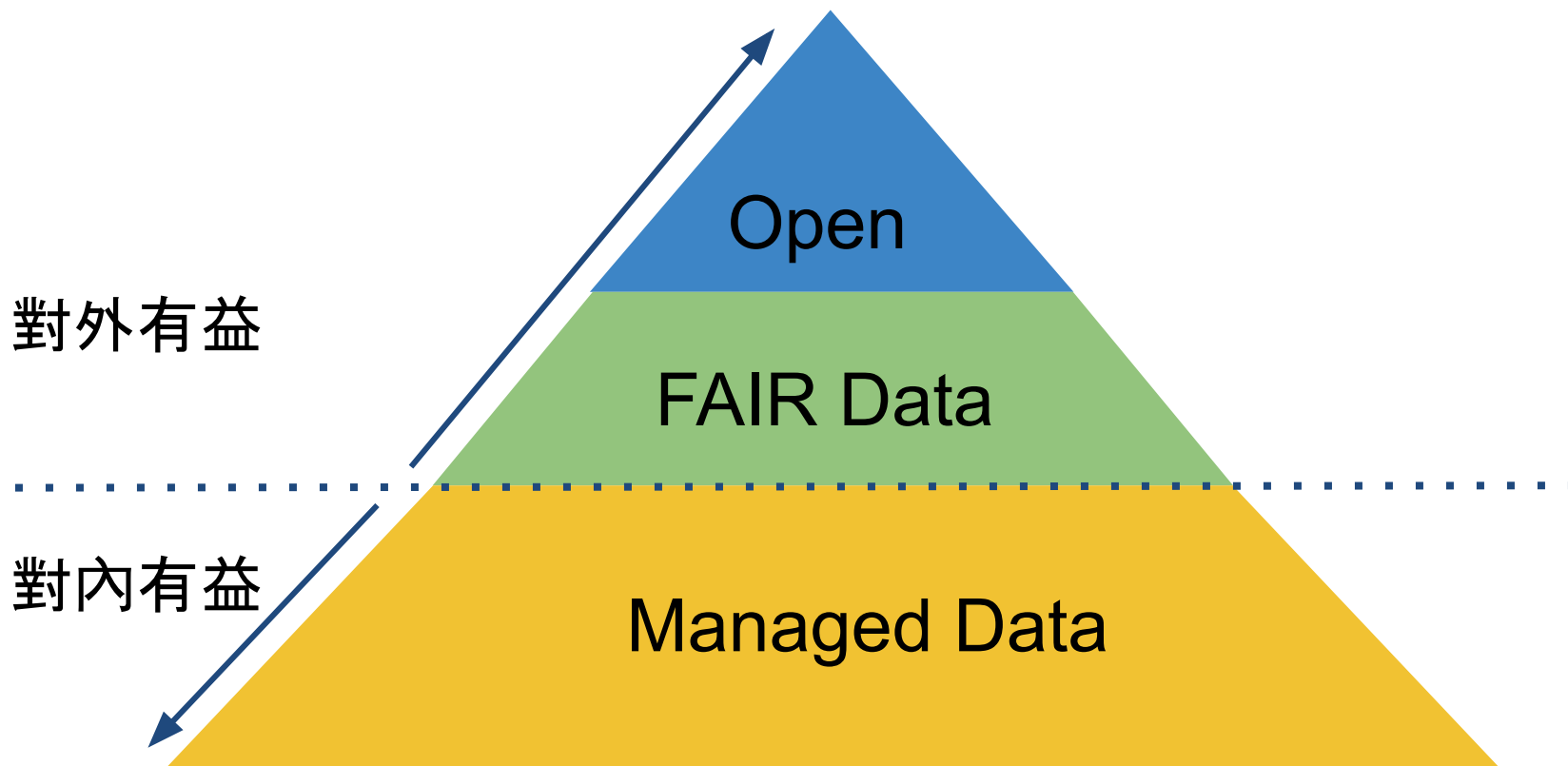
Interoperable
可相互操作

資料使用正規化、可取用、可共享和廣泛適用的語言進行知識再現。(Darwin Core)

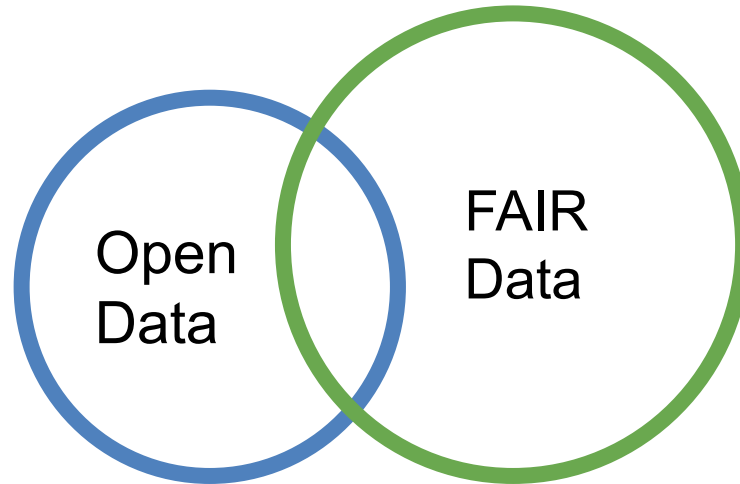
R

Reusable
可再次使用

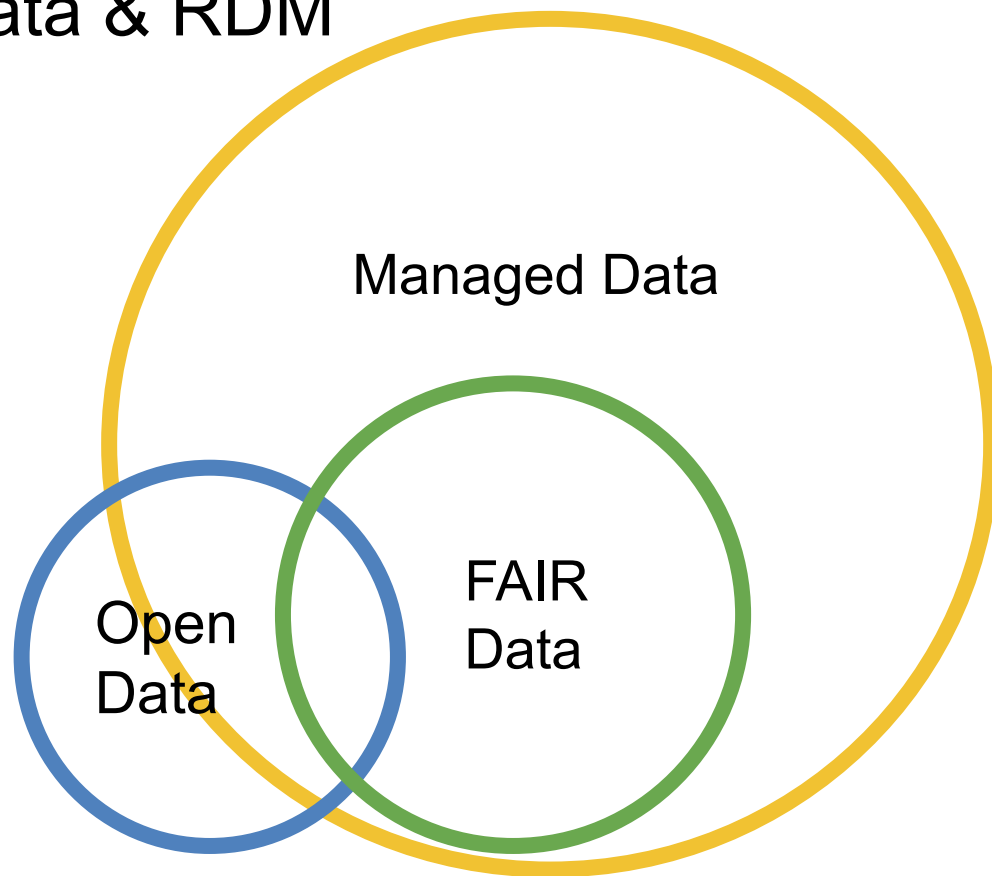
資料借由清晰易讀的資料使用授權進行發佈



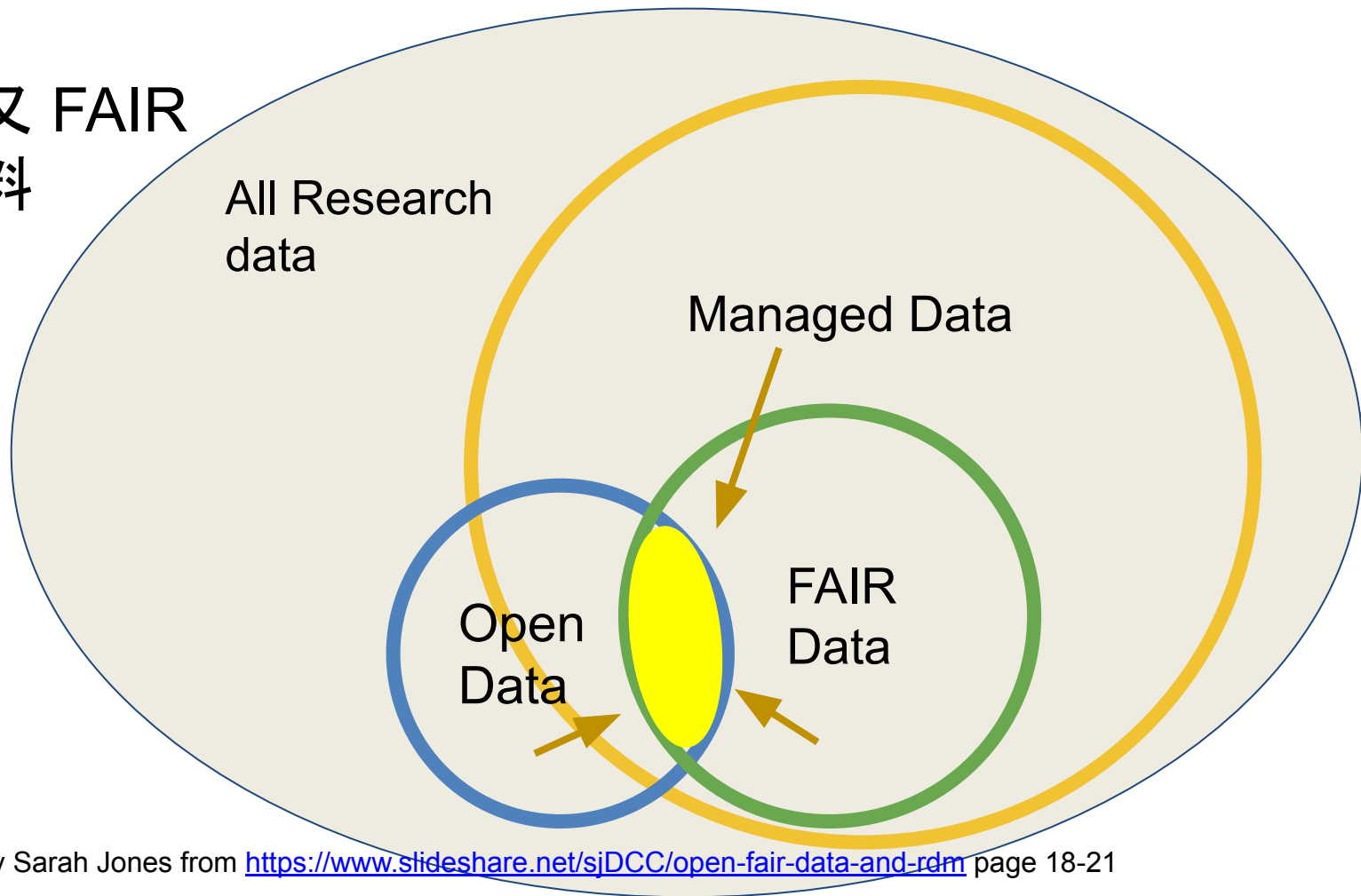
Open Data & FAIR Data



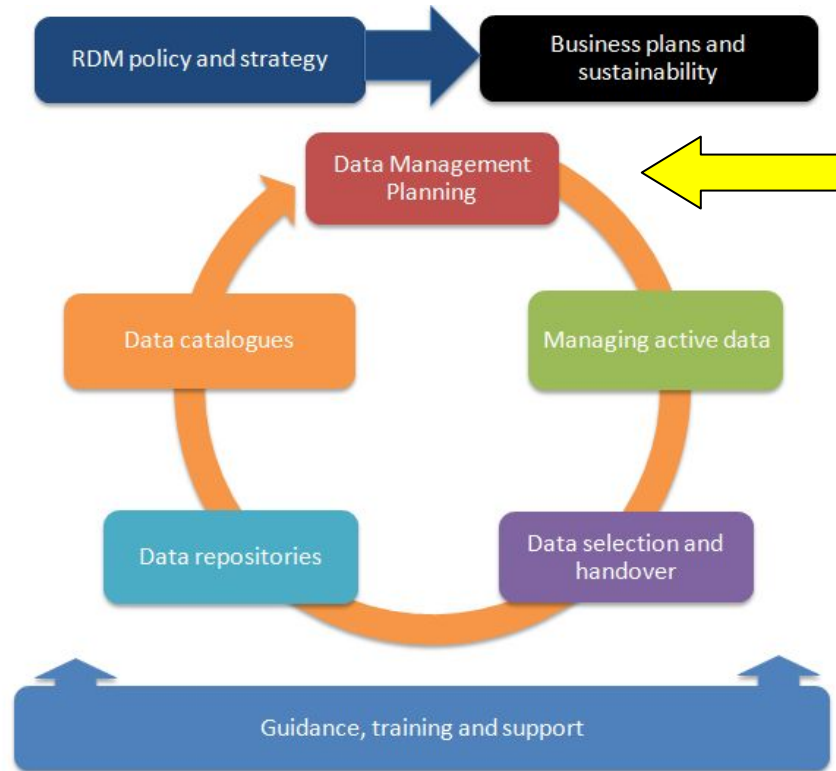
Open Data , FAIR Data & RDM



目標增加 即 Open 又 FAIR 的研究資料



Research Data Management (RDM)



<http://www.dcc.ac.uk/resources/how-guides/how-develop-rdm-services>

資料管理方案的核心需求

CORE REQUIREMENTS FOR DATA MANAGEMENT PLANS



Translations

Recommendations from the Practical Guide are being taken up all around the world. The guide has also been translated by several organisations to support their researchers in their national language. The following versions are currently available:

- **Chinese (traditional)**, on Depositar
- **French**, on Ouvrir la Science

<https://scienceeurope.org/our-priorities/research-data/research-data-management/>

https://data.depositar.io/en/dataset/se_rdm_guides



資料管理方案 (DMPs) 的核心需求

1. 資料的描述、資料的蒐集、以及既有資料的再次使用
2. 文件與資料的品質
3. 研究過程中的資料儲存及備份
4. 遵循法律和研究倫理以及行為守則
5. 資料共享和長期保存
6. 資料管理的責任和資源

資料管理方案的核心需求

1. 資料的描述、資料的蒐集、以及既有資料的再次使用：

a. 如何蒐集或產生新的資料？如何再次使用既有資料？

- 請說明收集或產生新資料的方法或軟體。
- 陳述既有資料在再次使用上的任何限制或使用條款(如果有的話)。
- 闡述如何記錄資料溯源 (data provenance)。
- 若有考慮過重用任何既有資料來源但最終決定不採用, 請簡述原因。

b. 要蒐集或產生什麼(種類、格式和數量的)資料？

- 提供資料類型的詳細資訊: 數值、文本、圖像、音訊、影片、混合媒體等。
- 提供資料格式的詳細資訊: 編碼方式、副檔名等。
- 決定使用某資料格式的理由為何？
- 請優先選用開放、標準的檔案格式。
- 提供資料大小及數量的詳細資訊。

資料管理方案的核心需求

2. 文件與資料的品質：

a. 有哪些後設資料和文件(如資料蒐集方法和組織資料方式的描述)會伴同資料？

- 將提供哪些後設資料來協助他人識別和尋得資料。
- 將使用哪些後設資料標準(例如 DDI, TEI)
- 說明如何組織資料以及相關之專業標準、版本控制、檔案夾結構。
- 需附哪些其他文件記錄才能使資料可被再用？
- 如何擷取以上提到的資訊以及將其記錄在何處？

b. 將採取哪些措施以維持資料品質？

- 闡述收集回來的資料是一致、優良，及如何記錄這一切。這過程可能包括：校準，重複收集樣本或測量，標準化的資料擷取，資料驗證與勘誤，資料的同儕評審，使用的控制詞彙表述等。

資料管理方案的核心需求

3. 研究過程中的資料儲存及備份：

a. 研究過程中，如何儲存資料和後設資料，並建立備份？

- 哪裡存儲和備份資料，以及每隔多久做備份。
- 優先選用具自動備份功能、運作穩固而且有管理人員掌管的託管存儲空間，例如由研究人員原屬機構/單位的資訊支援服務提供的備份空間。

b. 研究執行期間，如何確保資料安全並保護敏感資料？

- 說明一旦發生意外時該如何復原資料。
- 如何管理使用者角色權限設置。
- 保護資料免受安全威脅。評估及描述主要風險以及如何管控。
- 說明機構內有哪些現行的資料保護政策。

資料管理方案的核心需求

4. 遵循法律和研究倫理以及行為守則：

a. 如何確保涉及個人資料的處理過程能符合相關法規並保障資料安全？

- 確保在處理個人資料時遵守資料保護法規
 - 獲得個人資料的知情同意。
 - 建議個人資料匿名化 (anonymisation) 以進行保存和共享。
 - 考慮將個人資料擬匿名化 (pseudonymisation)。
 - 考慮將個人資料加密。
 - 說明是否有為具存取個人資料權限的使用者訂立管理資料取用的程序。

b. 如何處理智慧財產權與所有權等各類可能衍生的法律問題？有哪些適用的法規？

- 誰是資料的擁有者。
- 是否會影響到智慧財產權。
- 闡明對第三方資料的再使用是否有任何限制。

c. 如何考慮可能的倫理問題，並遵守行為準則？

- 在顧及倫理問題，是否也影響到資料的存儲和傳輸方式、限制了檢視或使用人群以及可保存的期限？
- 遵守國內和國際對資料處理及保護的行為守則和機構倫理準則。

資料管理方案的核心需求

5. 資料共享和長期保存：

a. 資料該在何時、以何種方式共享？在資料共享上是否有特定限制？若延後開放的期限，其理由為何？

- 說明如何讓資料更容易被尋獲及以何種方式共享。
- 規劃長期保存資料的方案。
- 說明資料可供開放使用的預期發佈時間。是否限定某些資料只供某使用者取用，如果需要，請闡述原因以及時限。
- 指明誰有使用資料的權限。若有必要訂立資料共享協議，請說明方式和原因。說明將採取什麼措施來克服限制或將限制減至最小。

b. 如何甄選需要長期保存的資料？在何處(如資料寄存庫或典藏庫)長期存放？

- 列明出於契約、法律或法規目的而必須保留或銷毀哪些資料。
- 列明如何決定要保留哪些資料及需要長期保存的資料。
- 評估資料的潛在研究用途和使用者。
- 說明將資料存放在何處。如果不打算建立或將資料放在資料寄存庫，請在資料管理方案中證明，怎樣可以在計畫經費動支期過後仍有效地管理資料。建議資料寄存庫的策略和過程(包括任何後設資料標準和涉及的成本)已被審核。

資料管理方案的核心需求

5. 資料共享和長期保存：

c. 需要透過哪種方法、軟體或工具才能取用資料嗎？

- 說明潛在用戶是否需要特定工具來存取、使用或重用資料？請考量用來存取資料的軟體的可持續性如何——是否穩定和長期可用。
- 說明資料是否將利用資料寄存庫達到共享，或按照請求批核資料使用權，還是使用其他機制？

d. 如何確保編配具唯一性及永久性的識別碼（如數位物件識別碼 DOI）給每一個資料集？

- 解釋如何在其他情況下重複使用資料。每項資料應該編配一個固定的編號 / 識別碼，這可讓查找和引用資料更可靠及有效率，還有助於跟踪資料被引用和重用的記錄。
- 闡明是否納用具唯一性及永久性的識別碼（如數位物件識別碼 DOI）。通常，信賴度高、可以長期固定地提供服務的資料寄存庫都會提供每個資料一個固定的識別碼。

資料管理方案的核心需求

6. 資料管理的責任和資源：

a. 何人(如角色、職位及機關)負責管理資料(如資料託管員)?

- 概述資料管理 (management) 和託管 (stewardship) 的角色和職責, 例如: 何方負責資料擷取、產出後設資料、把關資料品質、維護存儲和備份、處理資料歸檔以及資料共享。盡量列明負責人的名稱。
- 至於多方協作的計畫, 請說明合作夥伴之間如何分配及協調資料管理的職責。
- 列明誰負責實施資料管理方案、確保方案有被評核並在必要時做修訂。
- 建議定期更新資料管理方案。

b. 哪些資源(如經費及時間)將專用於管理資料及確保資料的 FAIR 特質, 亦即確保資料是可被搜尋 (Findable)、可被取得 (Accessible)、可互操作 (Interoperable), 並且可再使用 (Re-usable)。

- 將資料準備為可發佈共享及在資料度用 (data curation) 上的必要成本(如人力、金錢、時間), 需要估算並闡釋估算的方法。
- 將資料準備為資料寄存庫可收受的這項工作, 可需要投放額外的資源? 來自資料寄存庫的費用如何滿足? 若需要, 請說明金額以及如何支應這項開支。



資料管理方案

依研究歷程成長

一開始的選擇很重要

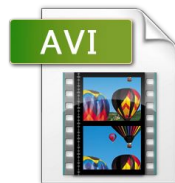
需要專人專務進行規劃

撰寫研究資料管理方案：工具與樣本

- DMPTool (University of California Curation Center)
 - 線上工具
<https://dmptool.org/>
 - 公開的樣本
https://dmptool.org/public_plans
- DMPonline (Digital Curation Centre, UK)
 - 線上工具
<https://dmponline.dcc.ac.uk/>
 - 公開的樣本
https://dmponline.dcc.ac.uk/public_plans

資料格式建議

- Open Format eg. csv
- 可互操作 eg. json 、 xml
- 非專有或商用軟體才可開啟的格式



...

Type

Text documents

• Preferred format(s)

- PDF/A (.pdf)
- ODT (.odt)

• Non-preferred format(s)

- Microsoft Word (.doc)
- Office Open XML (.docx)
- Rich Text File (.rtf)
- PDF other than PDF/A (.pdf)

Plain text

- Unicode text (.txt)

- Non-Unicode text (.txt)

Markup language

- XML (.xml)
- HTML (.html)
- Related files: .css, .xslt, .js, .es

- SGML (.sgml)
- Markdown (.md)

<https://dans.knaw.nl/en/about/services/easy/information-about-depositing-data/before-depositing/file-formats>



OpenRefine

專門對付雜亂資料的強大工具；
清理、轉換資料的格式與內容；
支援 Windows 、 Mac 、 Linux

A free, open source, powerful tool for working with messy data
<https://openrefine.org/>

可使用的研究資料寄存庫

- Zenodo
 - developed under the EU OpenAIRE program and operated by CERN
 - <https://zenodo.org/>
- DANS (Data Archiving and Networked Services)
 - an institute of the Dutch Academy KNAW and funding organisation NWO
 - <https://dataverse.nl/> (short-term data management)
 - <https://easy.dans.knaw.nl/> (long-term archiving)
- 研究資料寄存所 (depositor)
 - developed at Academia Sinica and funded in part by MOST
 - <https://data.depositor.io/>
- CoreTrustSeal Certified Repositories
 - <https://www.coretrustseal.org/why-certification/certified-repositories/>

Research Data Management (RDM)



<http://www.dcc.ac.uk/resources/how-guides/how-develop-rdm-services>



<https://rdm.depositar.io/>

[illegible]

Thank You!

<https://data.depositar.io/>

<https://rdm.depositar.io>

data.contact@depositar.io

rdm.contact@depositar.io

