

分享您的数据 – 数据存档和引用的检查清单

欢迎来到“开放科学之旅”！

如何规划、管理和共享您的数据？
可使用本检查清单来助您改进数据管理和共享实践。

目标受众

主要群体: 团队或项目研究员

次要群体: 团队或项目负责人

本篇为通用清查清单，需要根据具体机构、实验室、研究团队和/或资助方的要求进行调整。

A. 在研究项目期间规划和管理数据

1. 就数据文件的格式做好提前规划，并统一管理文件：

- a. 使用 [非专有格式](#)（例如纯文本表格）和不会丢失信息的压缩文件格式（例如 TIFF），以确保文件长期可供访问。加密文件格式可能会给未来访问埋下隐患。分析中使用的专业软件的开源替代方案也是优选方案。
- b. 选择人类和机器都能阅读的 **描述性强和有用的文件名**。改进文件名的具体做法可以包括：1）使用诸如“-”之类的分隔符将各术语连接在一起，使用“_”来分隔不同的信息；2）避免使用特殊字符（如\$）、空格和标点符号；3）创建有意义的名称和元数据、以此对文件内容做出解释；4）使用便于默认排序的方法，例如 YYYY-MM-DD 日期格式（国际标准）。参见 Jenny Bryan 的 [如何命名文件](#)。
- c. 使用能够产生高质量、干净数据的数据组织技巧。有关数据组织的通用信息，请参见 EDI 存储库的 [数据清洗指南](#)。
- d. 确定项目的分层文件夹和文件结构。保持一致，一个项目使用一个目录，按功能（如数据、代码）和输出（如图表、结果）分类。确保原始数据是独立的、不可编辑的，并生成原始数据的副本以供编辑。参考 Turing Way 开发的 [“文件结构”简编](#)。
- e. 创建一个或多个自述文件（README files），为有关数据提供解释信息。提前规划在收集信息时如何捕获自述文件中所用的信息，以确保日后对数据做出正确解释。那些能帮助您创建数据集中元数据的基本元素也应包括进来。
 - 根据您的文件结构和复杂性而定，您可能需要为每个数据集都创建一份自述文件以及对整个文件结构作出综合描述的一份自述文件（也称为“文件目录”）。

在您的自述文件中包括以下元素：标题、研究者联系信息、收集日期、地理信息、描述性术语、语言信息、资金来源、共享/访问信息、数据和文件概述、方法论信息和数据特定信息。参见康奈尔大学的 [编写“自述”风格元数据的指引](#)，其中包括 [可下载的自述文件的模版](#)。此外，请参见 Smithsonian 图书馆的 [描述数据：数据字典](#)。有关元数据的一般信息，请参见 EDI 存储库的 [创建元数据](#)。

请保持自述文件处于最新更新状态。

2. 使用版本控制、工作流程和笔记本工具更高效地管理数据：

- a. 使用版本控制系统来跟踪文件更改或项目历史（例如，获取的各个步骤、重新格式化）。诸如 [GitHub](#), [GitLab](#), 或者 [开放科学框架 \(OSF\)](#) 等平台可以提供额外的协作和在线备份功能。
- b. 采用工作流程管理工具生成可重建和可扩展的数据分析。如 [通用工作流程语言 \(CWL\)](#) 等工作流语言提供了一种描述命令的工具，能将工具连接在一起并形成工作流。类似的可选工具包括 [Snakemake](#), [Nextflow](#) 和 [Galaxy](#)。
- c. 对于实验室而言，可以考虑使用电子实验室笔记本来记录研究、实验和程序。诸如 RSpace 等实验室笔记本，可以提供链接到研究工作流的功能，并更有效地记录和共享研究信息。
- d. 知晓并遵守您或其他人先前创建/收集的数据使用许可和/或数据请求协议。

3. 存储和备份数据文件：

- a. 确保文件由一套可信的系统进行自动备份、保障文件安全。如果可能，应充分利用你所在机构提供的备份服务。备份系统最好位于不同的地理位置（例如云存储服务）。

B. 公开分享数据

当你在研究中完成一个里程碑并且在发表之前，可使用以下步骤。

1. 选择数据存储库

- a. 为您自己创建的数据提前规划存放地点，这有助于为您更高效地搭建和管理数据，方便日后记录和共享数据。这份 [存储库指引](#) 大有帮助。

2. 决定要保存哪些数据

- a. 项目中您收集或创建的数据（如原始数据）以及处理后的数据应得到妥善保存和公之于众，以允许其他读者分析评估您的结论并在您的研究基础上继续开展工作。
- b. 对于模型或模拟数据，通常应保留模型和具体配置信息。EarthCube 建模社区开发出了一个框架、就应该保留的内容给出相关指引，该框架可根据具体情况进行调整。
- c. 对于非常大的数据量（大于 1tb 或 TB）来说，数据保存工作将颇具挑战，因为这通常需要费用和额外资源。可以考虑机构能否提供数据保存和合规的解决办法。其他方案请参见 [这篇博客文章](#)。[[互联网档案链接](#)].

3. 在存放之前做好数据记录工作，包括自述文件、元数据和许可授权等。

a. 准备元数据。（推荐）

许多存储库为理解数据所需的文档类型提供指导（例如 [EDI 存储库指引](#)）。如果无法获得，请遵循自述文件指南（如上）。

b. 为数据使用设置许可权限（必需）

根据您对自己数据的预期使用类型来选择许可权限。为了鼓励他人再次使用，尽可能公开您的数据。首选许可是 CC0 或 CC-BY 4.0。参见 [知识共享](#)。

4. 將數據存放在選定的存儲庫中

a. 遵循選定存儲庫的指引。

包括自述文件和/或元數據文件

確定許可授權形式。這可能由存儲庫管理，但如果不是这样的话，則應說明您確定的授權方法。

b. 確保存儲庫清晰無誤地展示您的首選引文。

C. 為所有數據提供引用來源

1. 在論文的參考文獻內，列出您使用的原始數據和處理後數據的數據來源（這包括您創建的數據以及其他創建的任何數據）。這樣做可以確保為數據提供合理的可信度。
2. 在論文加入**數據可用性聲明**，描述您的數據處於何處以及如何可用，包括訪問數據的在線方式。在向期刊提交論文之前請檢查好各項鏈接和文件，以確保同行能找到數據、開展同行評審。
3. 使用**可用性聲明和引用檢查清單**，用以确保您的數據引用和可用性聲明完整無缺。參見**數據和軟件共享方面的作者須知**了解更多詳情。

相關清單的快速鏈接

- [您的數字存在](#)
- [軟件存檔和引用檢查清單](#)

如推薦更新信息，請發送電子郵件至 datahelp@agu.org。在電子郵件中請注明清單的名稱和 DOI。

如要引用本清單：Song, Dada, Stall, Shelley, Specht, Alison, O'Brien, Margaret, Machicao, Jeaneth, Corrêa, Pedro Luiz Pizzigatti, David, Romain, Miyairi, Nobuko, Murayama, Yasuhiro, Santos, Solange, Wyborn, Lesley, Vellenich, Danton Ferreira, & Mabile, Laurence. (2022). 《數據存檔和引用檢查清單》。Zenodo <https://doi.org/10.5281/zenodo.13897204>