关于研究数据的长期归档

Cyril Pernet[1*]、Claus Svarer[1], Ross Blair[2], John D. Van Horn[3] 和 Russell A. Poldrack[2]

[1] 神经生物学研究单位，Rigshospitalet，丹麦哥本哈根[2] 心理学系与斯坦福可重复神经科学中心，斯坦福大学，美国加利福尼亚州[3] 心理学系与数据科学学院，弗吉尼亚大学，美国弗吉尼亚州

* 通讯作者：wamcyril@gmail.com

ORCID

Cyril Pernet：0000-0003-4010-4632 Claus Svarer：0000-0001-7811-1825 John D. Van Horn：0000-0003-1537-0816 Russell A. Poldrack：0000-0001-6755-0259

摘要

随时获取研究数据，是 FAIR（可发现、可访问、可互操作、可重用）数据共享旨在大规模实现的目标。然而，我们认为，考虑到维护存储库所需的经济和生态成本，为了快速访问而不断积累并维护所有数据集并不可持续。本文探讨冷数据存储的问题：何时应将数据移至离线存储？如何在此过程中维持 FAIR 原则？谁应负责冷归档和长期保存？

关键词：数据共享、FAIR、成本、长期归档

声明：作者声明不存在利益冲突

数据与代码可用性：原始 OpenNeuro 数据集数量及所用代码均可公开访问，地址为 @https://github.com/CPernet/OpenNeuroCount

引言

数据整理的目标之一，是确保指定用户和再利用者能够日常查找并访问数据。数据的访问频率决定了其温度：‘热’数据是指持续使用的数据，‘温数据’是指需要定期访问的数据，而冷数据则是很少使用的数据。尽管人们已努力建立大规模的温数据存储库，但对于数据随着时间推移变得更冷后该如何处理，却鲜有讨论；这也与关于数据保留政策和长期组织可持续性的一般性建议有关（NSTC，2022）。

数字数据策展包括数据的保存、存储和处置（Higgins，2008）。通常认为，处置意味着这些数据未被

何时处置研究数据?

使用频率是界定数据温度的主要因素。如同分子，活动越剧烈，温度就越高。在此语境下，数据处置涉及确定一个频率阈值，以决定何时处置数据；这一阈值可以理解为温存储的经济和生态成本与数据预期效用之间的权衡。举例来说，我们分析了自2020年1月以来存入 OpenNeuro 平台的167个独立数据集的平均下载次数（下载字节数/数据集大小）（Markiewicz 等人，2021）。下载次数根据存入时间重新对齐。其中~74% 的数据集存入时间至少为24个月，最短为14个月（图1）。从总下载次数中，我们可以观察到异常值占下载量的16% 以上，且对应排名前 ~6% 的数据集。将这些数据随时间绘制出来后，我们可以观察到，对于访问量很高的数据，下载次数并未放缓；而平均或总下载次数较低的数据集则呈现下降趋势。

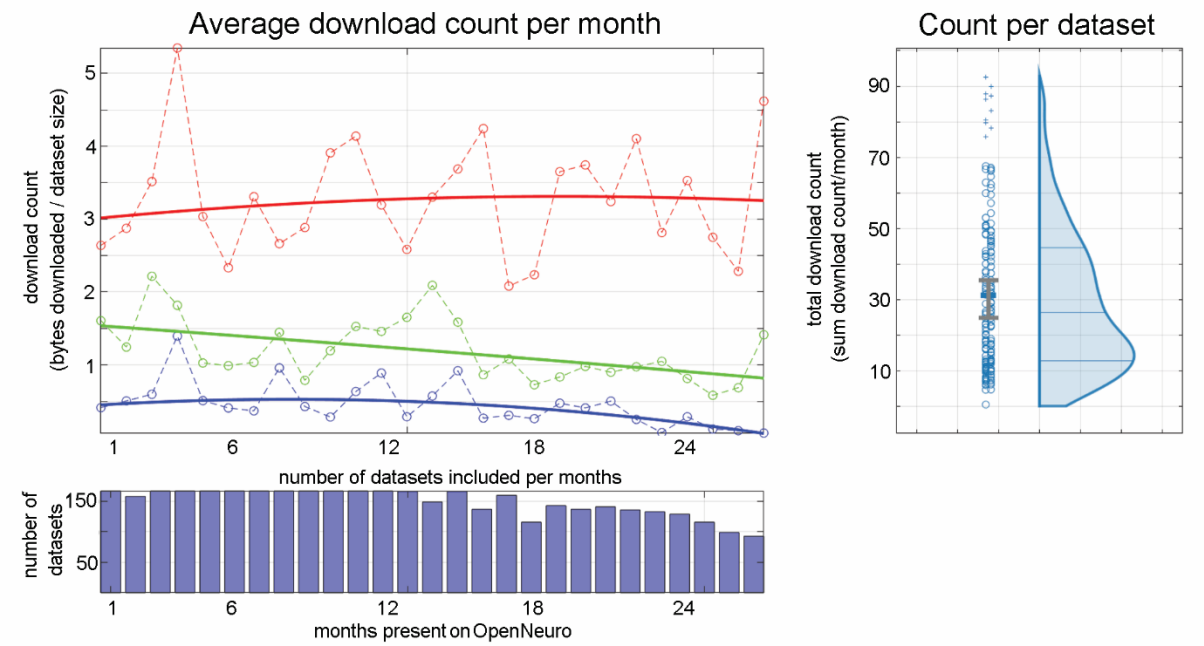


图1. 自2020年1月以来，每个数据集的平均下载次数随时间的变化，数据来源：*https://openneuro.org/*。左侧显示高访问量数据集的下载次数（*>75* 总下载次数，以红色表示）、平均访问量数据集（总下载次数介于24至35之间，以绿色表示）以及低访问量数据集的下载次数（*<18* 总下载次数，以蓝色表示）；粗线表示二阶多项式拟合。底部的条形图显示每月纳入的数据集总数。右侧显示总下载次数（圆圈和叉号代表各个数据集；带叉号的数据集是根据中位数绝对偏差确定的异常值；灰色条表示均值31及其95%最高密度置信区间）*[24 35]*, ，蓝色区域表示总数据计数的非参数核密度估计。

在决定保留哪些数据时，最重要的是数据的效用。数据经常被下载，可能是因为其质量、丰富度和复杂性使其能够得到有效的进一步分析，或被重新用于解决新的科学问题。一个很好的例子是 INDI数据集（1000功能连接组项目、ADHD-200、自闭症脑成像数据交换（ABIDE）和可靠性与可重复性联盟（CoRR）数据），这些数据集已被多次重新分析，并由此产出了高质量的出版物（截至2017年3月共913篇），估计节省了超过一百万美元（Milham等人，2018年）。数据集也可能因其简单易用而被下载，例如用于方法开发或教学；但目前缺乏针对这类情况的衡量指标。我们认为这些情况很可能存在，因为OpenNeuro上下载量最高的数据集并非只有规模最大的，其中一些也是小型数据集（如图1所示，下载量最高的数据集平均为~54GB [分钟，最小5.6Mb，最大847Gb]；而下载量最低的数据集平均为~34Gb [分钟，最小311Mb，最大171Gb]；单侧t检验 t(26.35)=0.9 p=0.18）。

然而，使用频率只是效用的一种替代指标。显而易见，某些数据在非常特定的情境中可能极具价值，却很少被访问（例如，关于一种鲜为人知的病毒序列的数[illegible]
[illegible]用这一概念则有助于决定何时处置数据，并将数据销毁与冷归档区分开来；仅凭使用频率无法估算效用。由于数据效用高度依赖领域和具体情境，我们无法就何种数据有用提出普遍建议。鉴于此处讨论的数据已经过整理，且数据收集与整理均需付出成本，我们可以避免销毁数据，并建议默认采用冷归档。回到OpenNeuro数据集的例子，我们可以考虑在24至36个月后，将数据集中的较低端部分（~38%）转入冷归档。接下来的问题是，冷数据是否仍能保持可发现、可访问、可互操作和可重用（Wilkinson等人，2016年）。

冷 FAIR 数据

在基于网络的存储库中，可发现性高度依赖全局唯一标识符（GUID），也称为唯一持久标识符（PID）。数据集通常会分配此类标识符，常见形式为数字对象标识符（DOI）。将数据转入冷归档时，必须保留元数据信息，并注明数据当前的温度。由于 FAIR 也关注机器可操作性，而冷归档往往由人工负责，因此必须建立明确的机制，以便在需要时申请获取此类数据，并说明申请方式和预计延迟时间。对于云服务存储库，只需将数据云分层设为冷层即可；如果需要访问，检索时间会更长、费用也会更高，但由于这些数据很少被访问，总体生态成本和财务成本会更低。

考虑长期数据保存时，我们必须牢记，一般而言，数字媒介和机构都无法确保长期可靠，因此有必要规划数据长期保存，尤其要考虑存储库停止运营时如何保存数据。fMRI数据中心项目（Van Horn & Gazzaniga，2013）就是一个例子：该项目整理并归档了 2000年至[illegible]

[illegible]

互操作性和可重用性主要取决于所使用的数据格式和数据完整性。对于数据格式，无论是热数据还是温数据，我们都只能建议使用开放且有完善文档记录的格式，以便任何人随时检索和读取数据。可重用性还取决于数据完整性，这是任何数据存储系统的重要方面。已有研究提出并实际应用了四种风险来源：存储硬件、物理环境、[illegible]

[illegible]数据完整性通常通过同时使用多个数据副本来解决，例如定期将备份副本存放在物理位置分隔的服务器上，而这些服务器自身采用独立磁盘冗余阵列，从而保护数据免受磁盘故障影响。然而，冷归档往往依赖单一副本

因此，位级信息完整性更值得关注。冷存储的典型建议是使用磁带，尤其是线性磁带开放标准（LTO），

[illegible]，以便在需要时独立存储和检索数据。

结论

研究人员乐于承担数据创建的责任，并且越来越多地承担起数据共享这一艰巨但必要的任务。随着数据

致谢

C.R.P. 获诺和诺德基金会 NNF20OC0063277 资助。

参考文献

Altman, M. & Landau, R.（2020）。选择高效可靠的保存策略：利用大规模分层离散事件仿真对长期[illegible]。《国际数字策展期刊》，15。http://dx.doi.org/10.2218/ijdc.v15i1.727

Currie, A & Kilbride, W（2021）《永远符合 FAIR 吗？研究数据保存中的问责与责任》。《国际数字策展期刊》，16。).2218/ijdc.v16i1.768

Gorgolewski, K., Auer, T., Calhoun, V. 等（2016）《脑影像数据结构：用于组织和描述神经影像实验输出的格式》。 Scientific Data，3，160044。https://doi.org/10.1038/sdata.2016.44

Higins, S.（2008）。《DCC策展生命周期模型》。《国际数字策展期刊》，3。https://doi.org/10.2218/ijdc.v3i1.48

Markiewicz, C. J., Gorgolewski, K. J., Feingold, F., Blair, R., Halchenko, Y. O., Miller, E., Hardcastle, N., Wexler, J.,

美国国家科学技术委员会，《联邦资助研究数据存储库的理想特征》（2022）https://doi.org/10.5479/10088/113528

Van Horn, J.D. & Gazzaniga, M.S.（2013）。为何共享数据？从 fMRIDC 中汲取的经验。*Neuroimage*，82。

Wilkinson, M., Dumontier, M., Aalbersberg, I. 等。《科学数据管理的 FAIR 指导原则》与管理。*Scientific Data* 3, 160018（2016）。https://doi.org/10.1038/sdata.2016.18