构建可扩展、可重新部署的多层存储库，用于可查找、可访问、可互操作、可重用且安全的 Scientific Data 共享：BIG-MAP 档案库

Valeria Granata, [1,*] Fran.cois Liot, [1,*] Xing Wang, [2] Steen Lysgaard, [3]
Ivano E. Castelli, [3] Tejs Vegge, [3,4,†] Nicola Marzari, [1,2] 和 Giovanni Pizzi[2,‡]

[1] *Theory and Simulation of Materials (THEOS) and National Centre for Computational Design and Discovery of Novel Materials (MARVEL), École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland*
[2] *PSI Center for Scientific Computing, Theory and Data, Paul Scherrer Institute, 5232 Villigen PSI, Switzerland*
[3] *Department of Energy Conversion and Storage, Technical University of Denmark, DK 2800 Kgs. Lyngby, Denmark*
[4] *CAPeX Pioneer Center for Accelerating P2X Materials Discovery, DK 2800 Kgs. Lyngby, Denmark*（日期：2025年12月19日）

arXiv:2512.15815v1 [cs.DB] 17 Dec 2025

大型联盟中的数据共享，例如大型科研合作或产业合作伙伴关系，通常需要同时解决组织和技术问题。共同的数据共享平台对于联盟的成功运作至关重要，有助于促进项目内部的协作，并推动研究成果的交流。开发此类平台面临多项技术挑战，例如设计可扩展、可重新部署的架构，以及创建便于上传和下载数据的用户友好界面；其中，安全和访问控制是首要关注点，必须妥善处理，以防止未经授权的访问或敏感数据遭到滥用。本文介绍 BIG-MAP 档案库，这是一个基于云的、面向特定学科的私有存储库，旨在应对这些挑战。该存储库构建于 InvenioRDM 平台之上，利用该平台的部分功能满足联盟的特定需求，相较于通用数据存储库提供了更适用的解决方案。存储库的访问权限可以限制为特定社区的成员，例如最初为其构建该存储库的 BATTERY 2030+ 倡议：一个由多个项目组成的联盟（其中包括 BIG-MAP），旨在加速先进电池技术的开发与部署。数据和元数据上传至档案库后，访问权限将受到限制（通过细粒度访问控制，仅允许单个 BATTERY 2030+ 项目的成员或该倡议的所有成员访问）。此外，规范化的上传流程可确保数据和元数据已完成格式化，随时可以在需要时发布到开放存储库（例如，在相应科学论文发表时）。本文评述了该存储库的主要功能，着重说明 BIG-MAP 档案库如何通过提供安全、受控的数据存储与访问环境，在保障数据机密性的同时支持灵活的基于权限的访问系统，从而克服大型联盟内部数据共享所面临的挑战；并阐述该档案库如何轻松重新部署，以满足 BATTERY 2030+之外其他联盟的需求，例如欧盟新的 MaterialsCommons4.eu 和 RAISE（欧洲人工智能科学资源）倡议。

一、引言

数字存储库用于存储和保存数据与元数据，便于访问研究信息，使科学家能够分享、重用彼此的研究成果并在此基础上开展工作。它们已成为科学界不可或缺的工具，数量也在持续增加[1–3]，这往往是由资助机构推动的；这些机构要求将与出版物相关的数据发布到适当的公共存储库中[4,5]。

公共数据存储库通常旨在开放共享数据，供任何人访问，这对于涉及敏感或机密信息的协作而言，可能并不理想或合适。在某些情况下，研究人员希望在研究数据产生后立即分享，但在公开之前只与特定同事分享（例如，在研究成果发表于科学期刊时）。例如，这可能发生在由 FINALES[6,7] 等工具驱动的联邦式自动驾驶自主实验室项目中。研究人员也可能需要同事的反馈、正在等待许可或专利的批准，或者所开展的研究属于大型联盟项目，其中涉及与公司和工业界的合作伙伴关系，并且包含机密数据或信息。

这些挑战在大型研究联盟中尤为突出。此类联盟为协调多个机构之间的合作研究提供了共同框架。联盟汇聚了来自不同学科的专家，共同应对宏大的目标，而这些目标通常涉及大量数据。在联盟内部共享数据，对于研究人员获取推进研究所需的信息至关重要。因此，协作型数据存储库是关键工具，充当跨多个项目共享和保存数据的枢纽。当联盟与公司和行业建立合作关系时，协作与安全之间的平衡就显得更加重要。在这种情况下，与公共存储库相比，受限数据存储库能够对资源实施更严格的管控。通过限制数据访问权限，这些存储库使联盟能够维护信息的保密性和完整性，同时继续促进合作伙伴之间的协作。

为满足大型联盟的需求，存储库应实施机制，通过身份验证和授权系统规范访问权限

并通过权限设置提供细粒度的访问控制，使用户能够授予或拒绝对特定数据集的访问权限。此外，在适当情况下，应通过跟踪访问权限、用户操作和数据使用来维护透明度和保密性。

此外，除了存储数据并确保数据得以长期保存外，数据存储库还应让数据易于访问和发现，以便他人能够高效地重复利用。为提高数据的可发现性，存储库应提供或鼓励使用标准化的数据和元数据格式。标准化格式也有助于使数据和元数据与其他数据集、数据处理工具及分析工具实现互操作。如今，在考虑向人工智能（人工智能）系统和机器学习（机器学习）工具提供（可查找、可访问、可互操作、可重用）数据时，这一要求变得更加严格。标准化格式有助于促进不同人工智能或机器学习应用之间的数据共享和重复利用，为数据描述提供通用语言和结构。这样一来，人工智能系统便能有效地从汇总后的数据中学习，从而做出更准确的预测。

除了 JSON（JavaScript 对象表示法）和 XML（可扩展标记语言）等通用数据标准，以及都柏林核心等元数据标准[8] 和 DataCite[9]，之外，基于本体和语义网的更先进技术也正在兴起，帮助科学家更轻松地找到所需的相关数据。本体和语义网技术通过创建描述数据间复杂关系的语言，使计算机能够理解和处理数据[10]，，从而依据含义和逻辑关联检索和管理信息。这意味着计算机搜索内容时，不仅能匹配词语，还能关联相关概念、实体和关系，以提供更准确、更相关的搜索结果。研究人员已探索本体和语义网技术在科学领域的应用。针对特定领域的专门标准正不断涌现，例如电池接口本体（BattINFO）[11]，，这是一个关于电池及其接口的本体，其领域专用标准旨在将有关电池接口的现有知识形式化。

BIG-MAP 档案库（图1）是一个基于 Web、面向特定学科且私有的数据存储库，可存储和共享由 BATTERY 2030+ 联盟[12]生成的研究数据。该档案库依据上述理念开发，旨在通过在联盟内共享数据和信息来促进协作，同时确保保密性。该档案库最初专为 BIG-MAP 项目成员打造，现已向联盟的所有项目开放。它经过专门设计，可满足联盟的需求，为管理敏感数据提供解决方案，同时促进涵盖多种不同模态的项目成员之间的协作[13]。元数据经过结构化处



图 1. BIG-MAP 档案库的主页，这是一个特定领域的私有存储库。只有 BATTERY 2030+ 倡议的成员才能访问该存储库。

理和格式化，以遵循 FAIR 原则[14]，，提升其可查找性和可重用性，使记录能够发布到公共存储库中。该存储库目前托管了 260 多条记录，拥有 450 多名注册用户，上传数据总量约为 1.5 TB。该存储库基于 InvenioRDM [15]，构建，这是一个基于 Invenio Framework、由欧洲核子研究中心开发的研究数据管理（RDM）存储库平台。它利用了 InvenioRDM 的部分功能，例如社区功能和令牌化链接共享，以满足大型联盟的独特需求，为数据共享提供可查找、可访问、可互操作、可重用、安全且灵活的方式。

BIG-MAP 档案库为用户提供量身定制的数据共享解决方案，提供三种开放程度：与特定项目的成员共享、与联盟的所有成员共享，以及可能将数据向全球公众开放（图2），只需轻松、自动地将数据和元数据传输到开放存储库即可。此外，借助令牌化链接，用户可以完全控制共享哪些数据以及与谁共享，从而实现精细的访问控制，并确保只有授权人员才能访问敏感信息。由于该应用采用开源软件编写，且代码可在 GitHub[18]，上自由获取，联盟成员可以检查并验证存储库软件的功能，确保安全机制的实现方式以及数据的存储和处理方式透明。总而言之，BIG-MAP 档案库让联盟能够全面掌控数据共享方式，确保数据所有者监督数据访问权限，从而促进协作中的信任。

II. 存储库的技术与主要特性

BIG-MAP 档案库的代码采用开源许可，使其他联盟能够自由访问权限、修改和再分发

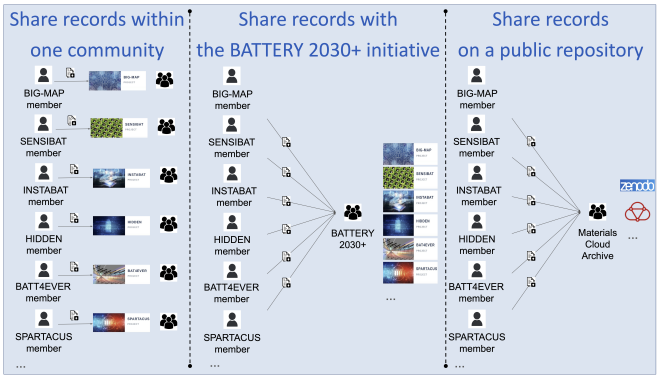

图 2。用户可以选择仅与自己项目的成员（在 InvenioRDM 的术语中称为“社区”）或与整个 BAT- TERY 2030+ 倡议的所有成员共享其记录。在 BIG-MAP 档案库中共享记录要求作者整理数据并准备元数据。这使他们能够更轻松地将研究成果发布到 Zenodo 等公共存储库[16] 和 Materials Cloud Archive[17] 中。

代码[18]。因此，它可以被视为其他希望增强自身数据管理与共享能力的联盟的模板。机构可以在现有工作基础上进一步开发，而不必从头开始；其功能和定制灵活性也可以根据不同研究倡议的独特需求进行调整。

BIG-MAP 档案库基于 InvenioRDM[15], 构建，后者是一个基于 Invenio 框架的研究数据管理存储库平台。InvenioRDM 采用现代化的 Web 架构，依赖 OpenSearch、PostgreSQL、Python 3/Flask 等开源软件和系统，并使用 React 编写前端。

该存储库最初托管在瑞士国家超级计算中心（CSCS）的虚拟机上[19]，，但目前已迁移至丹麦技术大学（DTU）[20] 以及用于加速 P2X 材料发现的长期、大规模 Pioneer 中心（CAPeX）[21]，以确保连续性和长期可持续性。在 DTU，应用程序直接由 DTU 管理的服务器提供服务，所有相关研究数据均存储在 DTU 的存储基础设施中，并通过每日自动备份确保可靠性和长期保存。

InvenioRDM 强大的基础设施和架构[22]旨在实现无缝扩展，使平台能够在不影响性能的情况下应对流量或数据量的增长。这得益于负载均衡、缓存和分布式计算等技术；如有需要，这些技术可使应用程序进行水平扩展（通过增加服务器数量）或垂直扩展（通过提升每台服务器的性能），以满足不断增长的需求。

InvenioRDM 采用混合方式执行读写操作。它将记录存储在关系数据库（PostgreSQL[23]）中，以此作为主要数据源，并将副本索引到一个 NoSQL 系统（OpenSearch[24]）中，以便更快地读取。搜索查询会发送到 OpenSearch 搜索引擎，该引擎基于 JSON，具有很强的可扩展性，并具备强大的搜索和聚合能力。

元数据按照 JSON 架构存储，这些架构提供了定义其结构和格式的标准化方式。这有助于与其他系统进行数据集成，并能在搜索信息时快速、准确地检索数据。此外，关键词、徽章和筛选器也进一步提升了数据检索效率。这些功能通过缩小搜索结果范围、以视觉方式突出显示相关信息，以及应用特定条件进一步优化搜索，帮助用户快速高效地找到所需信息。

除了搜索功能外，存储库直观且用户友好的界面还提供一系列功能，帮助用户高效管理数据。这些功能包括个人工作区，便于有序访问用户的记录；以及上传表单，用于编辑元数据和上传数据。

上传表单布局简洁，旨在让用户能够直观、轻松地浏览，并通过分步骤的流程引导用户完成上传流程。表单包含用于添加元数据的字段，例如标题、描述和关键词，以便整理数据并使其更易于搜索。此外，为确保作者得到正确引用和署名，上传表单采用标准化标识符。用户可以将记录的作者关联到其 ORCID（开放研究者与贡献者标识符），并将所属机构关联到 ROR（研究组织注册库）ID，从而准确识别作者和研究机构并消除歧义。

上传表单提供多种许可证，帮助用户在决定如何分享和再利用其作品时做出明智选择。可用许可证包括常见的标准化许可证，例如知识共享（CC）、GNU 通用公共许可证或 MIT 许可证，以及存储库的 BATTERY 2030+定制许可证[25]。

用户可以将文件拖放到上传表单的指定框中，或逐个选择文件，轻松完成上传。此外，用户还可以选择同时上传多个文件。进度条会显示上传流程的状态，并同时显示已上传文件的总大小。每条记录的存储限制为 100GB，必要时可以提高。

记录定稿并向档案库的其他用户开放后，用户可以更新其元数据，但不能更改文件。若要为共享记录添加、删除或替换文件，则可以创建新版本，每个版本都与同一记录的其他版本关联，并且可以从其他版本访问。记录元数据可以导出为多种格式，包括 JSON、JSON-LD、DataCite 和都柏林核心。这样便于将记录整合到多种其他系统和工具中，例如其他数据存储库、文献管理软件和研究信息系统。

BIG-MAP 档案库的首个版本是使用 InvenioRDM 开发的，

采用的是 InvenioRDM 9 版本。该版本仅供 BIG-MAP 成员使用，当时尚未引入社区的概念。随后更新至 InvenioRDM 12 版本后，档案库利用并调整了这一新版本提供的社区功能，使其他项目也能访问。如今，BATTERY 2030+ 联盟中九个项目的成员均可访问该档案库。

A. 启用选择性数据共享：社区共享和令牌化链接

BIG-MAP 档案库的一项关键功能是通过权限系统管理记录的访问权限，并精细控制哪些人可以访问数据，从而确保敏感信息仅向其指定对象开放。

该存储库仅限注册用户使用，也就是说，只有拥有账户且已确认电子邮件地址的人才能登录并访问其中存储的信息。账户仅为直接参与一个或多个 BAT- TERY 2030+ 项目的个人创建。每个项目成员名单由项目董事会或其指定的用户管理人员负责管理，他们有责任仅向经授权的个人授予访问权限，并撤销离开项目者的访问权限。

该存储库中创建了多个社区：每个 BATTERY 2030+ 项目各有一个，另有一个 BATTERY 2030+ 社区，涵盖倡议中的所有成员项目。创建用户账户时，系统会将该账户关联至用户所属的所有项目社区，以及 BAT- TERY 2030+ 社区。正如下文所述，任何记录都会在社区内共享，因此用户既可以参与自己所属的项目（促进项目专题讨论、协作和知识共享），也可以参与更广泛的 BATTERY 2030+ 社区（促进跨项目交流）。

记录可设置多个访问权限级别和粒度。这种粒度能够精细控制用户可对记录执行的操作，例如只读、编辑、删除或创建新版本。限制最严格的访问级别是将记录保存为草稿，此时记录仅对所有者（即记录创建者）保密，也就是说，只有所有者才能访问和修改记录。下一级是与用户项目的社区共享记录。用户可以在上传表单中选择要共享记录的社区。共享后，该社区的任何成员都可以访问记录、查看元数据并下载文件，但其他社区的成员无法访问。编辑权限仍仅限于记录所有者。最后，通过与.....共享记录

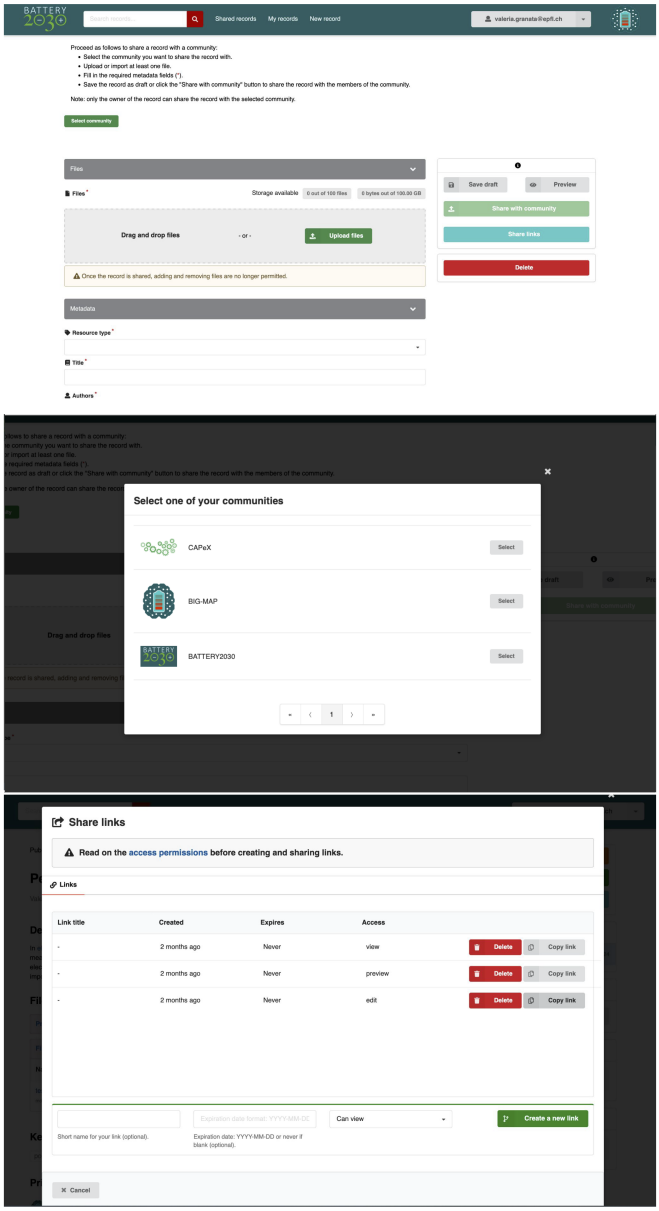

图 3. 除了与社区共享外，用户还可以通过共享链接授予限制更严格的访问权限，向获得链接的指定用户提供特定访问权限（仅查看、编辑等）。

电池 2030+ 社区中，联盟内任何社区的成员都可以访问（即读取）记录。即使记录已与社区共享，记录所有者仍保留对授予该记录的权限的控制权。这确保原始创建者对自己的内容拥有最终决定权，并且可以进行任何必要的更改，而不会干扰他人。

在某些情况下，用户可能需要先向同事开放记录的访问权限，以便他们审阅、提供反馈或批准，然后再与社区共享。此外，记录所有者可能希望向特定同事开放已与某个社区共享的记录的访问权限，而这些同事来自另一个

社区。为应对这些情况，并让用户能够以更精细的方式管理记录的访问权限，而不仅仅是与社区共享，用户可以生成“共享链接”（图3）。记录所有者可以为共享链接设置特定的权限级别（仅查看或编辑），然后像使用 Google 文档、Dropbox 等平台的共享功能一样，轻松地将链接分发给其他用户。具体而言，查看记录的共享链接可以分发给任何社区的成员，甚至可以分发给无权访问该档案库的外部人员。例如，提交一篇介绍档案库中某个数据集的科学论文时，这项功能非常有用：它可以向匿名审稿人提供数据的只读访问权限（在审阅阶段，数据仍可能处于私有状态）。相反，编辑记录的共享链接仅限记录所有者所在社区的成员使用，从而防止协作工作遭到未经授权的篡改。

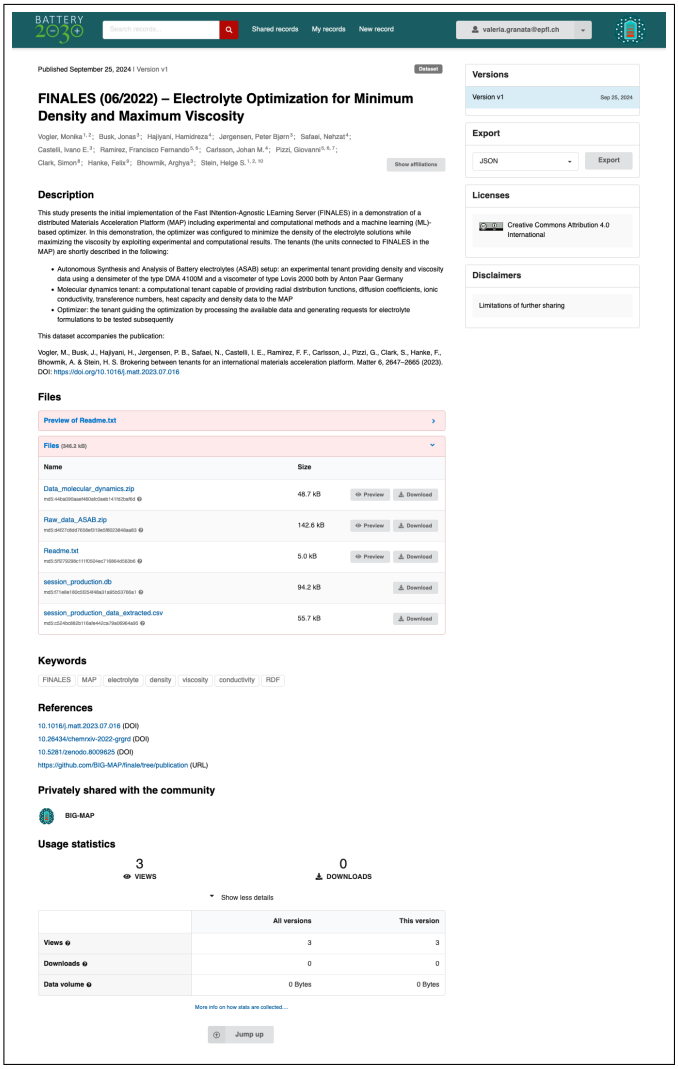

图 4. 记录页面示例。BIG-MAP 档案库会追踪每条记录及其各个版本的使用统计数据，并在每条记录页面上显示单条记录的统计数据和累计统计数据（见图下方）。

B. 安全且匿名的使用统计数据

使用统计数据能让我们深入了解用户与存储库的交互方式。存储库提供诸如浏览次数和文件下载量等指标，揭示记录被访问的频率以及哪些文件最常被下载。这些信息有助于用户了解其成果的可重用性。此外，使用统计数据也是项目经理展示项目成功与影响的重要工具，有助于他们就未来的协作和资源分配作出明智决策。事实上，这些数据为项目成效提供了具体证据，展示研究成果的协作情况和整体可用性，并通过证明以往项目成果的影响，增强资助申请的可信度。此外，这些统计数据也有助于存储库管理员提升内容的可发现性并改善用户体验，方法是找出元数据管理中有待改进的方面。

在 BIG-MAP 档案库中，我们使用 InvenioRDM 的使用情况跟踪功能来收集统计数据。使用情况跟踪完全匿名，并在服务器端进行。系统会针对每条记录及其每个版本收集使用情况（图4）。针对浏览和下载事件所跟踪的信息包括匿名访客 ID、请求来源国家（根据用户 IP 地址确定）以及引荐域名。匿名访客 ID 用于统计独立浏览和下载次数。其生成方式是将用户个人标识符（例如存储库中的用户 ID、用户会话 ID 或用户 IP 地址）与盐值（随机文本值）结合，再应用单向密码哈希函数对数据进行混淆。盐值每 24 小时重新生成一次。

使用情况跟踪符合 COUNTER 实践准则[26] 和研究数据实践准则使用指标[27]，使存储库符合 COUNTER 标准。

C. 用于数据交换和与外部应用程序集成的 REST API

RESTful 应用程序接口提供了一种灵活且可扩展的方法，便于软件应用程序之间进行通信，从而能够以结构化、标准化的方式交换数据。

InvenioRDM 的 REST API 是与外部应用程序交换数据的可靠高效方式，可与存储库无缝交互。要执行创建、更新记录、与社区分享记录或创建记录新版本等操作，用户可以绕过图形界面，通过脚本直接与存储库的后端交互——

API 脚本可以手动运行，也可以使用 cron 作业或任务调度器等工具安排定期运行。这对于系统性地检索和更新数据尤其有用，有助于实现自动化工作流以及跨系统的数据同步。

虽然 REST API 为数据交换提供了强大的工具，但用户可能至少需要掌握脚本语言的基础知识，才能有效地使用它们，例如 Python、JavaScript 或 Bash 等。这是因为应用程序接口需要发送 HTTP 请求，并以结构化格式处理数据。为使 REST API 面向所有用户开放，并免去编程专业知识的要求，我们开发了一个命令行 API 客户端。BIG-MAP Archive API 客户端[28] 提供了简单的命令，让用户能够轻松地与存储库交互并执行 API 调用，而无需编写代码。该客户端以 Python 包的形式发布，可从 PyPI[29]下载并安装。要验证存储库访问权限，用户只需提供其身份验证 Bearer 令牌。这是一段简短的字符串，用户可以在网页界面上生成一次（通过账户设置页面）。

作为一个应用示例，BIG-MAP Archive API 客户端已成功用于与 FINALES 中介软件系统进行数据交换。FINALES（Fast INtention-Agnostic LEarning Server）[6,7] 是一款用于编排多个租户的研究工作的软件，其中包括优化器、实验、模拟和数据库。API 客户端的一条命令可以将 FINALES 服务器的 SQLite 数据库备份到 BIG-MAP 档案库，并立即在存储库中共享备份数据。在自主研究活动期间，当产生新的测量数据或模拟结果时，它们会被整合到 FINALES 平台中。可以安排某个特定租户定期触发 API 客户端，并自动将数据库的最新版本上传到存储库，从而使所有数据立即对其他项目用户开放。电池优化活动中演示了这一做法，详见参考文献 7。遵循 FAIR 原则，两次自主 FINALES 研究活动中生成的数据[6,7] 在论文发表时已通过 Materials Cloud Archive[17] 公开发布，参见参考文献30 和31。

III. 未来目标与展望

随着档案库不断发展，预计将进行三项主要更新：使用 Keycloak 建立身份验证服务；通过简化记录在公共存储库中的发布流程，自动化开放性的第三个也是最后一个级别；以及通过 BattINFO 本体集成新的数据标准。为确保这些开发工作及长期可维护性所需的资源，档案库目前由 DTU 和 Pioneer Center CAPeX 管理并托管，详情如下文所述。

A. 使用 Keycloak 进行身份验证和用户管理

BIG-MAP 档案库用户账户的创建和管理目前由存储库管理员负责，他们遵循项目经理的指导，更新项目成员名单。为便于管理存储库中的用户账户并实现自动化，同时实现职责（功能）的分离，我们将实施一项基于 Keycloak[32] 的集中式服务。Keycloak 是一款开源的身份与访问管理（IAM）解决方案，提供稳健且可扩展的身份验证和授权系统。它基于标准协议，支持 OpenID Connect、OAuth 2.0 和 SAML。Keycloak 的授权框架建立在角色、权限、作用域和策略等概念之上，这些概念共同实现对用户访问权限的精细控制。InvenioRDM 支持与外部身份验证方法集成，包括 OAuth，并提供开箱即用的 Keycloak 集成功能。

设置 Keycloak 服务器并定义预期的角色和权限后，用户凭证及其他个人数据将安全地存储在 Keycloak 数据库中。Keycloak 使用端到端加密和安全协议来保护用户数据，并确保身份验证和授权事务的完整性。此外，Keycloak 还具备单点登录（SSO）功能。如果除 BIG-MAP 档案库之外的其他应用程序也配置为访问同一个 Keycloak 提供程序，则 SSO 功能允许用户使用一组登录凭证访问所有这些应用程序，而无需分别登录每个应用程序。

这种方式使 BIG-MAP 档案库应用程序能够专注于其核心业务功能，同时将所有用户身份验证和授权流程委托给 Keycloak。这也实现了项目中角色与任务的分离：存储库管理员专注于管理 BIG-MAP 档案库；用户管理员则可以直接在 Keycloak 中维护更新用户列表。

B. 在公共存储库中共享数据

BIG-MAP 档案库目前为用户提供两种共享记录的方式：与项目成员共享，或与整个 BATTERY 2030+倡议共享。第三种方式将用于将记录发布到公共数据存储库 Materials Cloud Archive[17,33]（图 5），这是一个经审核的开放存储库，收录与计算材料科学相关的研究数据。该存储库为每条已发布记录提供数字对象标识符（DOI），确保其易于查找和引用。

在计划的升级中，用户将能够自动将其在 BIG-MAP 档案库中创建并共享的记录传输到 Materials Cloud Archive，在那里

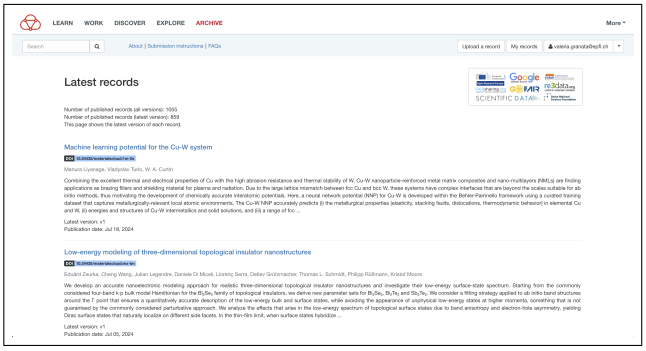

图 5. Materials Cloud Archive 是一个面向计算材料科学的公共审核型存储库。用户只需单击一下，即可将 BIG-MAP 档案库中共享的记录直接发布到 Materials Cloud Archive。

这些记录随后将由版主审阅，并在获批后发布。为自动将记录从 BIG-MAP 档案库上传至 Materials Cloud Archive，最初将在 BIG-MAP Archive API 客户端中实现一条特定的命令行界面命令。此命令会将 BIG-MAP 档案库中已共享记录的元数据和文件复制到 Materials Cloud Archive 上的一条草稿记录中。创建草稿后，用户可以根据需要对其进行修改，然后提交审阅。为确保身份验证安全，用户还必须在 Materials Cloud Archive 上拥有账户，并在运行命令行界面命令时提供凭据或令牌。不过，我们指出，如果两个档案库都实现 Keycloak 单点登录功能并使用同一身份验证服务器，这个额外步骤将不再必需。在这种情况下，可以重复使用同一身份验证信息，无需分别登录每个档案库。

在第二阶段，两个档案库之间的数据传输或可直接通过 BIG-MAP 档案库的用户界面实现；用户只需单击一下按钮即可启动传输，从而进一步简化流程并减少手动命令的使用。

C. 为自主平台启用遵循 FAIR 设计原则的数据

BIG-MAP 档案库的主要目标之一，是为其他系统重新使用和重新处理数据做好准备。为了有效集成到其他系统中，数据需要使用通用标准化词汇表对概念进行结构化和标注。应提供映射，将数据格式化为语义数据模型，使数据字段与其相应含义相匹配，从而确保数据具备机器可读性和互操作性。此类映射的一个示例是 BattINFO 本体[11]，，旨在支持 BIG-MAP 项目及更广泛范围内计算工具的开发和互操作性数据的部署。BattINFO 本体专用于电池领域，并描述了数千个与电化学和电池相关的概念，以及这些概念之间的关系。通过赋予数据含义，它将数据与其他数据片段建立链接，并根据这些含义实现机器推理。

可生成符合 BattINFO 规范的元数据的系统之一，是前文介绍的 FINALES 自主框架[6,7] 。与 FINALES 的所有通信都必须通过标准化数据模式进行，这些模式还可以补充本体注释。如前所述，当数据和元数据存储在 BIG-MAP 档案库中时，也可以将语义注释以 JSON-LD 等标准格式上传到档案库。随后，可以对档案库进行调整，以突出显示条目中语义注释的可用性，并以可视化方式呈现这些注释。值得注意的是，这既便于理解数据，也便于查找特定信息；借助定制的本体感知搜索平台，用户便能通过语义搜索轻松找到相关条目。

通过将数据及其语义注释上传至档案库，项目的每个阶段都将纳入可查找、可访问、可互操作、可重用（FAIR）原则，确保从项目一开始，数据就以提升其可用性和价值的方式被创建、共享和维护。因此，这种方法通过将数据管理的 FAIR 原则融入 FINALES 等项目的设计与开发，采用了“设计即 FAIR”（FAIR-by-Design）的方法论。

D. 未来方向与规模扩展

如前所述，为确保服务的长期可持续性和连续性，BIG-MAP 档案库已迁移，目前由 DTU 和 CAPeX 中心维护。档案库的原始网址（https://archive.big-map.eu/）现在会自动重定向至新的托管位置 https://archive-capex.energy.dtu.dk。档案库基于稳健的 InvenioRDM 技术构建，具备充分的可扩展性，可纳入 BATTERY 2030+之外及其内部的其他项目，例如 Salamander、UltraBat 和 BatCAT，并且还可进一步扩展至电池以外的领域，涵盖欧盟委员会将于 2026年 启动的 Materials Commons 倡议（例如参见 https://materialscommons4.eu）。此外，档案库还可通过其应用程序接口与其他数据存储设施实现互操作，例如参与开发绿色氢能技术的地平线欧洲计划项目 DECODE[34]。

该档案库还旨在成为新的泛欧洲 RAISE 计划的核心组成部分[35] 。该计划由欧盟委员会于2025年11月4日在哥本哈根启动，作为人工智能科学峰会[36]的一部分。在峰会上，人工智能材料科学界特别发起了一项行动倡议，提出建立 RAISE-AM：“泛欧洲高级材料 RAISE 实验室”，并将其作为人工智能

材料科学研讨会的成果。

Materials Commons 和 RAISE-AM 等倡议将高度依赖 BIG-MAP 档案库等存储库，以支持访问大规模、多源且可互操作的数据集，例如用于训练欧洲本土的基础模型，如面向材料化学的 MACE-MP-0 基础模型[37]。为此，计划建立一个 Mat4EU 中央枢纽，以连接联邦式档案库和先进材料数据空间及相关倡议。

IV. 讨论与结论

BIG-MAP 档案库应对大型研究联盟内数据共享所面临的独特挑战，尤其是涉及敏感信息的情况。该存储库提供了一个安全且易于使用的平台，专为 BATTERY 2030+ 倡议的需求量身打造，在促进协作的同时，通过受控访问确保信息保密。通过实施令牌化共享链接和精细化权限设置等高级功能，研究人员能够有效管理数据，并促进不同项目成员之间的协作。

基于 InvenioRDM 平台构建存储库所使用的技术，增强了其可扩展性和适应性，确保它能够满足用户不断变化的需求。对 FAIR 原则的践行进一步确保数据不仅得到保存，还能被发现和重用，从而支持与其他应用程序和工具集成。

展望未来，计划中的改进包括集成 Keycloak 以改善用户管理、支持将记录发布到 Materials Cloud Archive 等公共存储库，以及集成 BattINFO 本体。这些改进将显著增强存储库的功能，旨在打造更强大的科学协作系统，促进数据的无缝交换。

BIG-MAP 档案库为其他联盟和研究倡议树立了宝贵的典范。事实证明，它是 BATTERY 2030+倡议的重要工具，在数据管理和共享方面发挥着关键作用[38]。此外，其开源代码也为其他希望开发自有数据存储库的联盟提供了模板。

V. 数据可用性声明

本文未生成或分析任何新数据。不过，我们强调，BIG-MAP 档案库本身就是一个用于托管数据的存储库。与记录共享的社群成员获准访问相应记录。BIG-MAP 档案库的源代码已在 GitHub 上公开提供，地址为 https://github.com/team-capex/big-map-archive，并已存档于 Zenodo，DOI: 10.5281/zenodo.17939528。

VI. 致谢

本项目获得欧盟地平线2020研究与创新计划资助，资助协议编号为957189（BIG-MAP）。作者感谢BATTERY2030PLUS、UltraBat和BatCAT的支持；这些项目由欧盟地平线2020研究与创新计划资助，资助协议编号分别为957213、101103873和101137725；同时感谢用于加速P2X材料发现的先锋中心（CAPeX），丹麦国家研究基金会资助编号为P3。VG、XW、NM和GP感谢国家研究能力中心NCCR MARVEL提供的经费支持；该中心由瑞士国家科学基金会资助（资助编号205602）。我们感谢ETH理事会开放研究数据计划的支持（项目“PREMISE”：开放且可复现的材料科学研究；以及项目“API-03 IntER”：ETH领域存储库之间的互操作性）。

* VB 和 FL 贡献相同。 † teve @dtu.dk
‡ giovanni.pizzi@psi.ch
[1] K. Stracke 和 J. Evans, Commun Chem **7** (2024), https://doi.org/10.1038/s42004-024-01143-0。[2] H. Pampel、P. Vierkant、F. Scholze、R. Bertelmann、M. Kindling、J. Klump、H.-J. Goebelbecker、J. Gundlach、P. Schirmbacher 和 U. Dierolf, PLoS ONE **8**, e78080 (2013)。[3] H. Pampel、N. L. Weisweiler、D. Strecker、M. Witt、P. Vierkant、K. Elger、R. Bertelmann、M. Buys、L. M. Ferguson、M. Kindling、R. Kotarski 和 V. Petras, Scientific Data **10**, 571 (2023)。[4]“美国国立卫生研究院：数据管理与共享政策”，（访问日期：2025年11月30日），https://grants.nih.gov/policy-and-compliance/policy-topics/sharing-policies/dms。[5]“瑞士国家科学基金会：开放研究数据指南与要求”，（访问日期：2025年11月30日），https://www.snf.ch/en/dMILj9t4LNk8NwyR/topic/open-research-data。[6] M. Vogler、J. Busk、H. Hajiyani、P. B. Jørgensen、N. Safaei、I. E. Castelli、F. F. Ramirez、J. Carlsson、G. Pizzi、S. Clark、F. Hanke、A. Bhowmik 和 H. S. Stein, Matter **6**, 2647 (2023)。[7] M. Vogler、S. K. Steensen、F. F. Ram´ırez、L. Merker、J. Busk、J. M. Carlsson、L. H. Rieger、B. Zhang、F. Liot、G. Pizzi、F. Hanke、E. Flores、H. Hajiyani、S. Fuchs、A. Sanin、M. Gaber˘s˘cek、I. E. Castelli、S. Clark、T. Vegge、A. Bhowmik 和 H. S. Stein，Advanced Energy Materials

14, 2403263 (2024).[8]“都柏林核心元数据元素集”，（访问日期：2025年11月30日），https://www.dublincore.org/specifications/dublin-core/dces/.[9] “DataCite 元数据架构”，（访问日期：2025年11月30日），https://schema.datacite.org. [10] S. Clark, C. Battaglia, I. E. Castelli, E. Flores, L. Gold, C. Punckt, S. Stier, and P. Veit, ChemSusChem 18, e202500458 (2025). [11] S. Clark, F. L. Bleken, S. Stier, E. Flores, C. W. Andersen, M. Marcinek, A. Szczesna-Chrzan, M. Gaberscek, M. R. Palacin, M. Uhrin, and J. Friis, Advanced Energy Materials 12, 2102702 (2022).[12]“Battery2030+ roadmap”，（访问日期：2025年11月30日），https://battery2030.eu/wp-content/uploads/2022/07/BATTERY-2030-Roadmap_Revision_FINAL.pdf. [13] T. Vegge, J. Tarascon, and K. Edstr¨om, Advanced Energy Materials 11, 2100362 (2021). [14] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Ap- pleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. ’t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Pers- son, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. San- sone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolsten- croft, J. Zhao, and B. Mons, Scientific Data 3, 160018 (2016).[15]“InvenioRDM”，（访问日期：2025年11月30日），https:// inveniosoftware.org/products/rdm/.[16] “Zenodo 存储库”，（访问日期：2025年11月30日），https://zenodo.org. [17] L. Talirz, S. Kumbhar, E. Passaro, A. V. Yakutovich, V. Granata, F. Gargiulo, M. Borelli, M. Uhrin, S. P. Hu- ber, S. Zoupanos, C. S. Adorf, C. W. Andersen, O. Sch¨utt, C. A. Pignedoli, D. Passerone, J. VandeVondele, T. C. Schulthess, B. Smit, G. Pizzi, and N. Marzari, Scientific Data 7, 299 (2020).[18]“BIG-MAP Archive GitHub 存储库”，（访问日期：2025年11月30日），https://github.com/team-capex/big-map-archive.[19] “瑞士超级计算中心（CSCS）”，（访问日期：2025年11月30日），https://www.cscs.ch/.[20] “丹麦技术大学（DTU）”，（访问日期：2025年11月30日），https://www.dtu.dk/.[21] “CAPeX 先锋中心”，（访问日期：2025年11月30日），https://capex.dtu.dk.[22] “InvenioRDM 架构”，（访问日期：2025年11月30日），https://inveniordm.docs.cern.ch/maintenance/ architecture/.[23] “PostreSQL”，（访问日期：2025年11月30日），https://www.postgresql.org/.[24]“OpenSearch”，（访问日期：2025年11月30日），https:// opensearch.org/.[25]“BATTERY2030+ 非独占分发许可”，（访问日期：2025年11月30日），https://archive-capex.energy.dtu.dk/licenses/nonexclusive-distrib/1.1.[26] “Counter Metrics”，（访问日期：2025年11月30日），https: // www.countermetrics.org.

[27] M. Fenner、D. Lowenberg、M. Jones、P. Needham、D. Vieglais、S. Abrams、P. Cruse 和 J. Chodacki，《PeerJ 预印本》6，e26505v1（2018）。[28]“BIG-MAP API 客户端源代码”，（访问日期：2025 年 11 月 30 日），https://github.com/team-capex/big-map-archive-api-client。[29]“BIG-MAP Archive API 客户端的 PyPI 项目”，（访问日期：2025 年 11 月 30 日），https://pypi.org/project/ big-map-archive-api-client。[30] M. Vogler、J. Busk、H. Hajiyani、P. B. Jørgensen、N. Safaei、I. E. Castelli、F. F. Ramirez、J. M. Carlsson、G. Pizzi、S. Clark、F. Hanke、A. Bhowmik 和 H. S. Stein，Materials Cloud Archive，materialscloud:2024.142（2024），https: // doi.org/10.24435/materialscloud:ph-jb。[31] S. K. Steensen、M. Vogler、F. F. Ramirez、L. Merker、J. Busk、J. M. Carlsson、L. H. Rieger、B. Zhang、F. Liot、G. Pizzi、F. Hanke、E. Flores、H. Hajiyani、S. Fuchs、A. Sanin、M. Gaberˇsˇcek、I. E. Castelli、S. Clark、T. Vegge、A. Bhowmik 和 H. S. Stein，Materials Cloud Archive，materialscloud:2024.73（2024），https://doi.org/10.24435/materialscloud:qt-1s。[32] “Keycloak”，（访问日期：2025 年 11 月 30 日），https://www. keycloak.org。[33] “Materials Cloud Archive”，（访问日期：2025 年 11 月 30 日），https://archive.materialscloud.org。[34]“DECODE 项目”，（访问日期：2025 年 11 月 30 日），https://decode-energy.eu/about/。[35]“《科学人工智能欧洲战略》”，（访问日期：2025 年 11 月 30 日），https://eur-lex. europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX: 52025DC0724。[36] “2025 年科学人工智能峰会（AIS25）”，（访问日期：2025 年 11 月 30 日），http://www.ais25.eu。[37] I. Batatia、P. Benner、Y. Chiang、A. M. Elena、D. P. Kov´acs、J. Riebesell、X. R. Advincula、M. Asta、M. Avay-lon、W. J. Baldwin、F. Berger、N. Bernstein、A. Bhowmik、F. Bigi、S. M. Blau、V. Cˇarare、M. Ceriotti、S. Chong、J. P. Darby、S. De、F. Della Pia、V. L. Deringer、R. Elijoˇsius、Z. El-Machachi、E. Fako、F. Falcioni、A. C. Ferrari、J. L. A. Gardner、M. J. Gawkowski、A. Genreith-Schriever、J. George、R. E. A. Goodall、J. Grandel、C. P. Grey、P. Grigorev、S. Han、W. Handley、H. H. Heenen、K. Hermansson、C. H. Ho、S. Hofmann、C. Holm、J. Jaafar、K. S. Jakob、H. Jung、V. Kapil、A. D. Kaplan、N. Karimitari、J. R. Kermode、P. Kourtis、N. Kroupa、J. Kullgren、M. C. Kuner、D. Kuryla、G. Liepuoniute、C. Lin、J. T. Margraf、I.-B. Magdˇau、A. Michaelides、J. H. Moore、A. A. Naik、S. P. Niblett、S. W. Norwood、N. O’Neill、C. Ortner、K. A. Persson、K. Reuter、A. S. Rosen、L. A. M. Rosset、L. L. Schaaf、C. Schran、B. X. Shi、E. Sivonxay、T. K. Stenczel、C. Sutton、V. Svahn、T. D. Swinburne、J. Tilly、C. van der Oord、S. Vargas、E. Varga-Umbrich、T. Vegge、M. Vondr´ak、Y. Wang、W. C. Witt、T. Wolf、F. Zills 和 G. Cs´anyi，《化学物理学杂志》163，184110（2025）。[38] I. E. Castelli、D. J. Arismendi-Arrieta、A. Bhowmik、I. Cekic-Laskovic、S. Clark、R. Dominko、E. Flores、J. Flowers、K. U. Frederiksen、J. Friis、A. Grimaud、K. V. Hansen、L. J. Hardwick、K. Hermansson、L. K¨oniger、H. Lauritzen、F. L. Cras、H. Li、S. Lyonnard、H. Lorrmann、N. Marzari、L. Niedzicki、G. Pizzi、F. Rahmanian、H. Stein、M. Uhrin、W. Wenzel、M. Winter、C. W¨olke 和 T. Vegge，《电池与超级电容器》4，1803（2021）。