通过 FAIR+ 实施调查工具（FAIRIST）与研究人员互动，提升他们对 FAIR 和开放科学的认识

Kirkpatrick, Christine R.[1,2], Coakley, Kevin[1,3], Christopher, Julianne[1], Dutra, Inês[21] 圣迭戈超级计算机中心，加利福尼亚大学圣迭戈分校，美国加利福尼亚州拉霍亚₂ 计算机科学系，波尔图大学理学院，葡萄牙波尔图 计算机科学系，[illegible]
christine@sdsc.edu，kcoakley@sdsc.edu，jchristopher@sdsc.edu，ines@dcc.fc.up.pt

摘要： 开创性的 FAIR 论文发表六年后，研究人员仍然难以理解如何实施 FAIR。对许多研究人员而言，FAIR 承诺以近期投入换取长期收益，却要求他们具备尚未掌握的技能；同时，它也成为科学家面对的一长串无经费支持的强制任务和繁重要求中的又一项。即使是那些被要求采用 FAIR 研究实践，或确信自己必须为此投入时间的人，也更希望获得及时、适量且与科学制品及研究流程相匹配的建议。由于大多数 FAIR 实施指南过于笼统，研究人员很难根据自身情况调整建议。技术进步，尤其是人工智能（AI）和机器学习（ML）领域的进步，使 FAIR 的采用更加复杂，因为研究人员和数据管理专员都在思考如何让软件、工作流和模型符合 FAIR 原则并具有可重复性。FAIR+ 实施调查工具（FAIRIST）通过系统地整合研究要求与研究提案，缓解了这一问题。FAIRIST 纳入了纳米出版物和笔记本等新型学术成果，以[illegible]

关键词： FAIR、元数据、数据管理计划、调查问卷、可复现性

1. 引言与背景

FAIR原则描述了研究数据管理的15个期望维度（Wilkinson 2016年）。 这些原则为梳理任何给[illegible]

。研究人员只需在资助方要求的范围内处理这些议题。美国国立卫生研究院（NIH）要求提交数

本文的组织结构如下。第2节介绍定义和术语，并综述相关研究。第3节介绍 FAIRIST 的动机和

2. 文献综述、定义与术语

本文涉及信息科学和计算机科学中的概念。

FAIR 数据是符合四类原则的数据：可发现性、可访问性、互操作性和可重用性。***FAIR***原则是使

***FAIR*数字对象**： 尽管最初发表的 FAIR 原则指出，所有类型的研究制品都需要符合 FAIR 要求，

FAIR+：为简洁地指代 FAIR 和可复现性，本文使用 FAIR+ 一词。

开放科学一词被不同的利益相关者用于关注开放性的不同方面，涵盖从技术架构、知识创造的

可复现性：本文采用 Gundersen 的可复现性定义，核心在于科学研究应当能够复现；复现并非

相关工作

FAIR原则及相关文献为如何处理研究项目中的数据提供了很好的建议和指导。但研究人员通常使

FAIR类型和数据管理类型的工具，以及即将推出的用于发布和重复使用数据管理计划的工具。

Argos： 这是 OpenAIRE 与 EUDAT 共同开发的平台，可用于创建和管理数据管理计划（

记录研究项目及其成果，主要是数据集。Argos 用户界面经过精简，似乎采用了现代的用户界面

DMPTool：该工具由加州数字图书馆（CDL）提供，数年前创建，旨在帮助研究人员编制随项

FAIR 实施档案（***FIP***）向导：该向导由 GO FAIR 国际办公室和 GO FAIR 基金会的成员创建，

FAIR Connect：这项由 GO FAIR 基金会和 iOS Press 发起的新计划，旨在为数据管理专员和研

3. 动机与利益相关者

研究人员希望获得有关如何落实 FAIR原则 的实用建议，但数据管理所需的陡峭学习曲线和背景

采用罗列所有内容、让研究人员自行挑选相关原则的做法。此类工具的设计应只涉及与研究人员

FAIRIST 最初是一套模板化回复，用于协助同事撰写符合 FAIR 且具备人工智能意识的数据管理

FAIR维度	
可发现	• 数据将被分配持久标识符 <如何？ > 并将在 Renaissance Simulation 网站 https://rensimlab.github.io/ • 目录条目将添加到 SDSC 的 FAIR数据点。 • RSL 网站上将提供元数据以及相关本体的链接。 • 如有标签，将采用 schema.org 描述符。
可访问	• 可通过 SDSC 云（对象存储）获取，无需专用软件即可访问。其中包括原始数据和晕目录。 • 代理模型和其他机器学习基准测试将提供给 MLCommons，供科学工作组使用，并纳入未来的基准测试分发版本。 • 任何应用程序接口都将进行版本管理并加以描述，相关链接将发布在 RSL 网站上。
可互操作	• 代码存储在 GitHub 上（并链接自 RSL 网站） • 使用 RSL 中采用 xyz 标准的图书馆<或标准 Python 库等>。 • 使用天体的标准参考资料<更多内容见此处>。 • 使用<某种以地理空间特定方式描述形状的标准方法>。 • 输入数据和输出数据均采用 HDF5 格式。

可复用	• 机器学习模型和数据将存储在 OpenML.org 上 • 笔记本将展示如何组装模型和样本训练数据集。 每个笔记本成果都会被分配一个数字对象标识符 <，使用 Zenodo 还是加州大学圣迭戈分校图书馆？ > • RSL 笔记本界面位于 GitHub 上。 • 仿真的创建溯源信息将作为元数据的一部分提供。 • 我们将发布公告，声明所有数据均依据以下许可协议授权：知识共享署名 4.0 国际许可协议。

表*1*：启发 FAIRIST 创建的模板

手动填写这些模板几次之后，我们很快意识到，可以通过自助式调查问卷简化这一流程。尽管每

FAIRIST 会为研究人员提供定制文本，供其纳入数据管理计划或提案中。许多联邦资助机构要求

开，并采用允许免

像 FAIRIST 这样的工具，其利益相关者包括来自各个领域和行业的研究人员，包括学术界和工业

科研经费申请课程（NSF HSI 国家 STEM 资源

设的数据管理计划培训课程。

4. FAIRIST 与所用技术

FAIRIST 会调查项目的各个方面，然后将其映射到可能的选项和建议。这是一种以用户为中心的

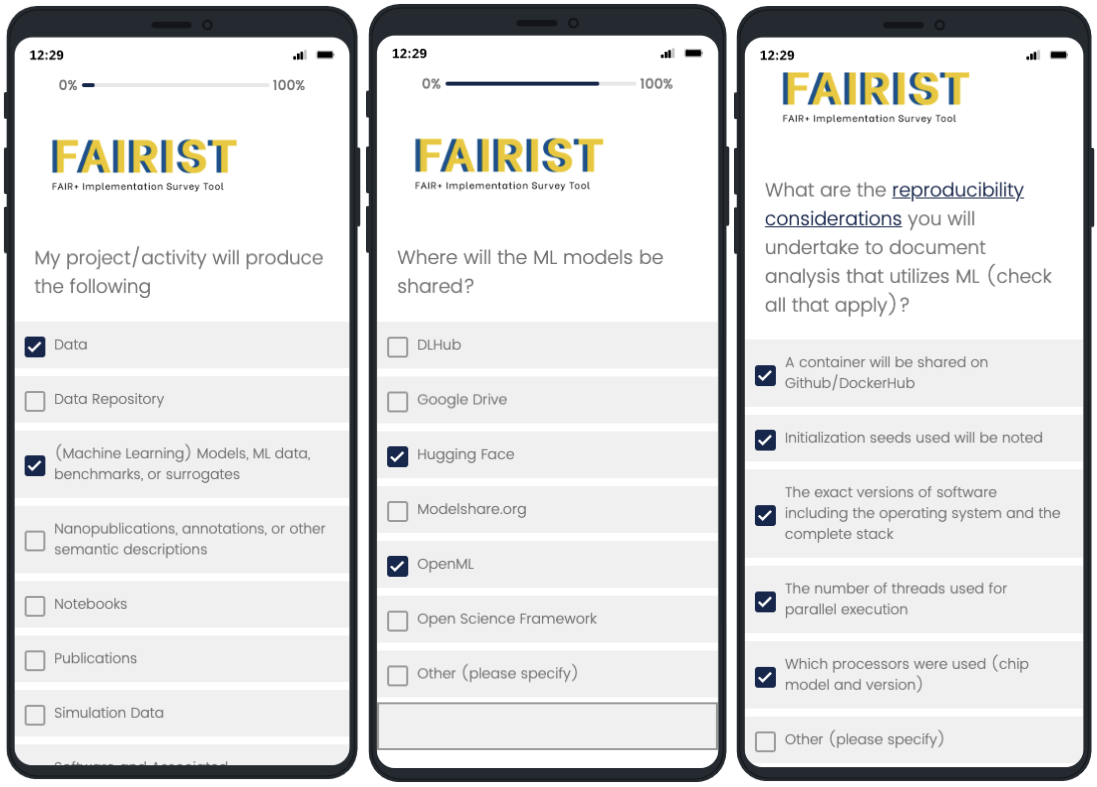

图 *1*：FAIRIST 中嵌入的逻辑会扩展调查问题，使其适用于所描述的项目。

可复现性相关考量选项，提炼自一篇篇幅更长、内容更复杂的计算机科学论文，该论文探讨了导

。这一方法可用于迅速将其他研究数据管理领域的学术成果付诸实践。

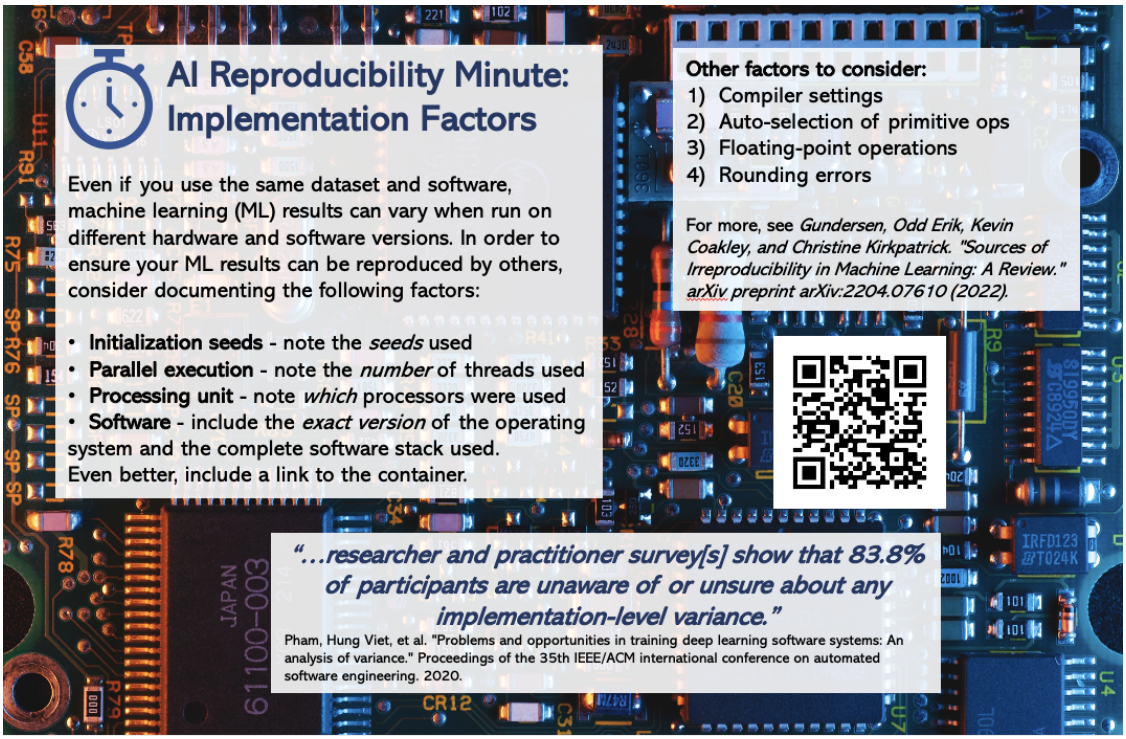

图2：推广与意识提升材料，可与 FAIRIST 配合使用，供图书馆、科研计算协调员及其他研究人

FAIRIST 是使用 Qualtrics 构建的。Qualtrics 是一款商业调查工具，具有功能齐全的用户界面

Qualtrics API 功能有限且速度较慢；输出可能需要 1-5 分钟才能生成。FAIRIST 输出准备好供

FAIRIST 建议

根据您的回答，以下建议供您考虑和/或纳入项目的数据管理计划。

数据类型

R与该项目相关的研究对象可分为以下几类 :

- ● 数据
- ●（机器学习）模型

计划采用的数据管理实践

表1展示了项目期间将开展的具体数据管理行动，以及这些行动与FAIR高层目标的关联。

FAIR 维度	计划实施的研究数据管理实践
可发现	• 研究成果将发布在项目网站上。 • 数据将按照社区最佳实践分配唯一标识符，并将在项目网站上引用。 • 元数据及相关本体的链接将在项目网站上提供。 • 如有标签，将使用 schema.org 描述符。
可访问	• 可通过开放的网络可访问文件夹获取。 • 所有数据均开放。
可互操作	• 代码存储在 GitHub 上（并从项目网站链接）。 • 使用代码中包含的图书馆。 • 输入数据和输出数据均为 HDF5 格式。
可复用	• 机器学习模型和数据将存放在 OpenML.org。 • 发布的声明将指定研究对象采用知识共享署名许可协议。

表1：按 FAIR维度划分的计划数据管理实践

表2：FAIRIST显示建议的示例输出

虽然这并非长期解决方案，但使用 Qualtrics 可以根据用户测试结果以及其他专家的意见逐步改

从敏捷软件开发社区借鉴的一条经验是，未来的工具开发者应当避免过分沉迷于开发网络工具或

可通过以下网址使用 FAIRIST：http://fairist.sdsc.edu/

结论与未来工作

尽管 FAIR实施取决于各研究项目和领域的具体因素，也会随着技术的发展而变化，但仍有可能

未来工作应包括由信息科学和科研计算领域的专家，对问卷选项和输出进行更广泛的审查。如果

还有许多其他来源可用于挖掘潜在问题和实施建议。本研究聚焦于计算机科学和人工智能，这是

，但未来版本的 FAIRIST 可以实现这一点。例如，对于将创建新领域存储库且属于地球科学领域

致谢
感谢 Melissa Cragin（莱斯大学）对 FAIRIST 调查问题提供的早期反馈，并感谢 Odd Erik Gundersen

资金支持
本研究部分由美国国家科学基金会（NSF）资助，项目编号为 #2226453、#1916481；并获得圣

参考文献

AAAI会议：可重复性检查清单，2022年2月22日至3月1日。网址：https://aaai.org/Conferences/AAAI-22/reproducibility-checklist/ [最后访问于12月9日 2022]。

Argos：规划并跟踪您的数据。网址：https://argos.openaire.eu/ [最后访问于 2022]年12月9日。

Carroll, S R 等。2020。《原住民数据治理CARE原则》。《数据科学杂志》，19(1)：43。数字对象标识符：https://doi.org/10.5334/dsj-2020-043

Coakley, Kevin，2022。FAIR矩阵调查API。网址：https://github.com/kevincoakley/fair-matrix-survey-api [最后访问于 2022]年12月15日。

CODATA-RDA-DataScienceSchools/Materials。网址：https://github.com/CODATA-RDA-DataScienceSchools/Materials/blob/master/docs/DataAtlanta2022/index.md [最后访问于 2022]年12月9日。

CODATA-RDA研究数据科学学校。网址：https://www.datascienceschools.org/ [最后访问于 2022]年12月9日。

数据设施委员会：地球科学研究。可在 https://www.earthcube.org/council-of-data-facilities 查阅。[最后访问于 2022] 年 12 月 9 日。

De Smedt, K., Koureas, D. 和 Wittenburg, P., 2020. 面向科学的 FAIR 数字对象：从数据片段到可操作知识

FAIR Connect。网址：https://fairconnect.pro/ [最后访问于 12 月 9 日 2022]。

Fecher, B. 和 Friesike, S., 2014。《开放科学：一个术语，五种思想流派》。载于《开放科学》（第

FIP Wizard 文档。访问地址：https://fip-wizard.readthedocs.io/en/latest/ [最后访问日期：12 月 9 日 2022]。

FIP Wizard。访问地址：https://fip-wizard.ds-wizard.org/ [最后访问日期：12 月 9 日 2022]。

Gundersen, O E 和 Kjensmo, S，2018。《现状：人工智能的可复现性》。第三十二届 *AAAI* 人工智能大会，32(1)。数字对象标识符：https://doi.org/10.1609/aaai.v32i1.11503

Gundersen, O.E.、Coakley, K. 和 Kirkpatrick, C., 2022。《机器学习不可重复性的来源：综述》。

Law, Asha, 2022。《WorldFAIR 中的 FAIR 实施配置文件（FIPs）：我们学到了什么？》可访问：https://worldfair-project.eu/2022/09/30/fips-in-worldfair-what-have-we-learnt-public-workshop-tue-25-october/ [最后访问日期：12 月 9 日 2022]。

LibGuides：研究数据管理：FAIR 数据。可访问：https://ufs.libguides.com/rdm/fair [最后访问日期：12 月 9 日 2022]。

磁学信息联盟（MagIC）。可访问：https://www2.earthref.org/MagIC/about [最后访问日期：12 月 9 日 2022]。

美国国立卫生研究院。2020。《NIH 数据管理与共享最终政策》，2020 年 10 月 29 日。可访问：https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html [最后访问日期：12 月 9 日 2022]。

NSF HSI 国家 STEM 资源中心：撰写资助申请——培训与资源。访问地址：https://hsistemhub.org/grantsmanship/ [最后访问于12月9日 2022]。

Praetzellis, M 和 Riley, B。2019年，加州数字图书馆：DMPTool，2019年11月8日。https://cdlib.org/services/uc3/dmptool/ [最后访问于12月9日 2022]。

Qualtrics XM——体验管理软件。访问地址：https://www.qualtrics.com/ [最后访问于12月9日 2022]。

Wilkinson, M D 等。2016年。科学数据管理与管理维护的 FAIR指导原则。*Scientific Data*，*3*(1)：160018。数字对象标识符：https://doi.org/10.1038/sdata.2016.18

WorldFAIR 项目等。网址：https://worldfair-project.eu/ [最后访问于2022年12月9日].

作者信息

Christine R. Kirkpatrick是加州大学圣迭戈分校圣迭戈超级计算机中心研究数据服务部主任。她

Kevin Coakley是加州大学圣迭戈分校圣迭戈超级计算机中心的高级云集成专家。Kevin 是挪威

Julie Christopher是加州大学圣迭戈分校圣迭戈超级计算机中心的信息技术项目经理。

Inês Dutra是葡萄牙波尔图大学理学院计算机科学系的助理教授。她的主要研究兴趣包括并行性、