构建 FAIR 气候数据的可搜索知识库

Mark Roantree[a]、Branislava Lalic[b]、Stevan Savic[c]、Dragan Milosevic[c]、Michael Scriney[a]

[a]洞察数据分析中心，都柏林城市大学计算学院，爱尔兰[b]塞尔维亚诺维萨德大学农业学院，塞尔维亚。[c]塞尔维亚诺维萨德大学气候学与水文学研究中心，塞尔维亚。

arXiv:2304.05944v1 [cs.DL] 12 Apr 2023

摘要

气候科学数据知识库的开发是一项多学科工作，需要领域专家（气候科学家）、具备知识库设计和构建技能的数据工程师，以及机器学习研究人员共同参与。机器学习研究人员能够为缺口填补等数据准备任务提供专业知识，并就可利用这些数据的不同机器学习模型提出建议。CA20108成本行动的主要目标之一，是开发一个完全符合科学数据管理 FAIR原则的知识门户。第一年，我们开发了一个定制的知识门户，用于采集 FAIR 数据集的元数据。该门户旨在使用世界气象组织标准，为可共享的微气象数据提供详细的元数据描述。系统在本地存储网络、站点和传感器的元数据，同时将实际数据传至 Zenodo，随后接收其返回的 DOI，从而在知识门户与存储平台 Zenodo 之间建立永久链接。用户在知识门户中搜索元数据时，搜索结果会同时提供详细描述以及指向 Zenodo 平台上数据的链接。

关键词： 元数据存储库、微气象数据、FAIR原则、知识门户

1. 引言与背景

作为 21 世纪的一项重大挑战，气候变化正在加速 [6], , 与此同时，与天气相关的自然灾害和极端事件的频率及强度也在增加 [19]。制定应对天气和气候变化的可靠战略计划，有赖于高空间密度且质量优良的站点特定观测与测量数据。为获取所需的高分辨率数值天气预报、气候、农业、环境和城市模拟模型，这一点至关重要。因此，我们需要适当的测量方法，并为建立面向应用的数据集进行战略规划；知识管理可在其中发挥重要作用，并制定新的创新标准。

基于微气象数据获取的环境条件可靠且稳健的时空数据，应为评估和模拟气候变化（CC）的趋势与影响，以及恶劣天气等事件对不同生态系统环境造成的影响提供平台。在欧洲，人们已作出大量努力，集中管理来自地面观测（天气尺度）和卫星测量的数据、天气与气候模拟数据，并创建了哥白尼计划 [2]和 ECMWF [3], 等供公众使用的存储库。尽管这些成熟的数据源已广泛应用于研究、教育和经济领域，但这些存储库仍缺少微气象数据：关于微环境气象条件的数据，而这些数据既应开放，也应供科学研究人员、利益相关方和其他感兴趣的用户群体使用。

微气象数据通常作为科学-不同目的下开发的科研项目和观测网络，却常常沉寂于报告、机构数据存储库，甚至更糟，散落在个人电脑中。为解决这一问题，充分挖掘这一丰富的知识来源，FAIRNESS COST行动 [4] 获资助建立一个微气象知识共享平台。其主要目标包括：建立一个存储库，收录现有且经质量验证的欧洲（及其他地区）微气象原位数据集；管理这些数据，使其符合FAIR原则 [18]；规范数据质量和缺口填补功能；为微气象数据提供丰富的元数据描述；并提供乡村和城市地区的FAIR数据集。

1.1. 成本行动的动机

作为全球气候观测系统（GCOS）的一部分，气象站网络的部署旨在监测天气尺度过程。然而，这些网络的空间分辨率不够精细，无法生成有关微尺度大气状态的数据。当前高质量的微气象数据库主要关注塔式观测以及地表与大气之间的通量交换 [9, 13]，以提供有关低层大气的信息。这些数据库由持续运行的欧洲（欧洲通量数据库集群）或美国（例如 FLUXNET [5]）通量观测网络和倡议组织管理。然而，由于投资和维护成本高昂，其空间覆盖范围有限，而且今后仍将如此。从研究人员的角度来看，FAIRNESS 成本行动的目标群体主要是自动气象站（AWSs）网络。这些自动气象站部署在乡村、城郊和城市地区，作为各类农业气象项目的一部分，例如病虫害预警系统、森林气象学、城市气象学和环境气象学 [7, 12]。

现有项目的主要弱点之一，是数据缺乏可查找性、可访问性、互操作性和可重用性：它们在管理数据和元数据时未能满足 FAIR指导原则 [18]。

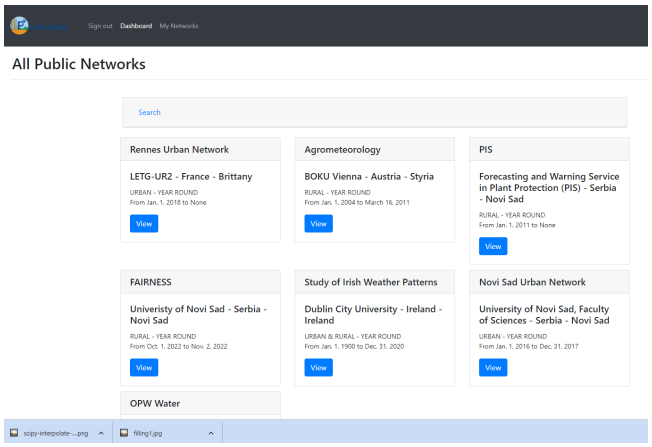


图 1：KSP 首页

本文介绍了CA20108成本行动 [4] 中的研究人员所做的工作：设计并构建一个知识门户 [10, 11]，用于存储微气象元数据，以便使大量现有数据符合 FAIR 标准，并尽可能开放获取。图 1 展示了当前版本的 FAIR知识门户首页，其中显示了已发布的网络元数据。成功部署该知识门户的潜在益处包括：更好地了解农村和城市地区小尺度气候变化及恶劣天气影响的现状；更好地规划田间作物作业（灌溉）和植物保护措施（喷洒），并改进造林成效；更好地规划城市能源消耗；改进应对危险病毒传播媒介的适应方案；更有效地组织对入侵物种（如白纹伊蚊）的监测。

论文结构。在§2中，我们简要概述 FAIR 原则，并引用一些将 FAIR指标 应用于现有项目的初步工作；在§3中，我们讨论如何使用知识门户进行搜索；在§4中，我们简要概述系统分析；最后，在§5中，本文以结论收尾。

在本文余下部分中，FAIRNESS 一词指 FAIRNESS CA20108成本行动，FKP 指 FAIRNESS知识门户。

2. FAIR指标的应用

FAIRNESS 的主要目标和宗旨完全聚焦于开放科学这一理念，它是未来科学战略的主要支柱之一，并与多个使命领域相一致。因此，理解和检验 FAIR指标很有意义。

我们先简要回顾 FAIR原则。

- **可查找**。 自动发现数据集和服务需要机器可读元数据。使用该

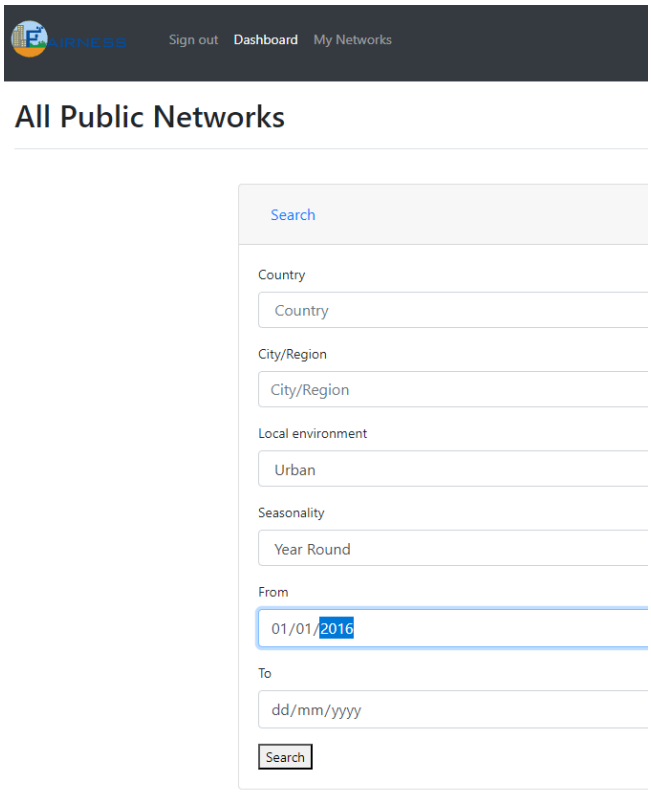


图2：可查找的元数据：启动搜索

KPG，对于系统中共享的所有微气象数据，数据所有者都会提供元数据描述，随后系统通过关键词或网络、站点和传感器搜索词驱动搜索引擎。在图 2 中，当前版本 FKP 中检索元数据的方法以一系列搜索字段呈现：国家、地区、局地环境、季节性和日期。目前，该方法正在扩展，以便在更广泛的元数据描述中进行关键词搜索。

- **可访问**。 识别出符合搜索条件的数据集后（见图3），结果页面会提供网络详细信息。假设数据可自由访问，则会提供 Zenodo 数字对象标识符和链接，以便直接访问数据。

- **可互操作**。 数据互操作性是指共享和整合来自不同用户和来源的数据的能力[16]。只有采用标准（元）数据模型来描述数据，才能实现这一点。这是一个重要概念，通常需要具备数据工程技能才能落实。在 FKP 中，世界气象组织指南提供了元数据的设计和结构。

- **可重复使用**。 若要真正实现可重复使用，元数据应尽可能详细地表述。这样，便可根据不同的科学需求复制和整合数据。虽然 FKP 支持非常详细的元数据描述，但并非所有元数据都必须提供。

由于已承认，在某些情况下，提供这些信息的开销可能非常高昂，因此仅作建议。

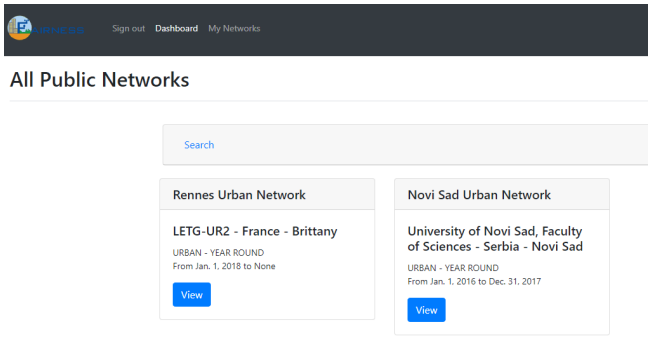


图3：搜索结果

需要指出的是，最初针对数据制定的 FAIR 规范较为抽象，可能导致不同的解读。为进一步阐释 FAIR原则，甚至提出加强原始描述的方法，人们已做出一些努力 [1]。这些努力通过考察 EOSC 等现有开放存储库，并深入探讨如何评估存储库在多大程度上符合 FAIR 来实现。包括采用 XML标准构建的可互操作数据仓库在内的解决方案，例如 [14], , 为开发利用通用数据模型的 FAIR 存储库基础设施提供了指导。这应成为任何知识门户的核心，否则这些门户将难以满足 FAIR 的可互操作原则。在开发 FKP 期间，人们努力更详细地界定 FAIR 的内涵，并首先聚焦于用于记录元数据描述的标准数据模型（WMO GAMP [20]）。这一设计决策确保了底层对 FAIR 的支持，因为它使所有 FKP 元数据和数据都符合可互操作原则。

在 [8], 作者介绍了一项案例研究，详细说明了如何对微气象数据开展 FAIR测试。针对 PIS 微气象数据的这项全面自我评估分为四个阶段：使用自我评估工具开展初步自我评估，以识别基础功能方面的不足；开展定性自我评估，以检验现有解决方案的质量并找出改进空间；确定落实前两个步骤所得结果所需的工具、方法和算法；执行并测试塞尔维亚植物保护预报与报告服务的一个组成部分。

本文通过针对全部 16 项FAIR指标进行测试（4 项原则中的每项各有 4 项测试），并为每项测试设置“是”“否”“部分”这几类基本类别，报告了预报与报告数据符合 FAIR 要求的程度。**可发现**指标包括：必须为数据分配持久标识符（F1）；有描述数据的丰富元数据（F2）；元数据在线存储于可搜索资源中（F3）；元数据记录注明持久标识符。对于这些指标，测试数据集获得了 2 个是 和2 个部分 的回答。**可访问**测试指标包括：根据持久标识符进行访问必须指向数据或相关元数据（A1）；数据可通过以下协议进行访问检索结果遵循公认的标准（A2）；访问过程必须包括身份验证和授权步骤（A3）；并且在可能的情况下，即使数据不是开放的或不符合FAIR，元数据也应可访问（A4）。对于可访问性指标，测试数据集获得了 2 个是 和2 个部分 的响应。**可互操作性** 指标包括：数据以易于理解且最好是开放格式的形式提供（I1）；元数据遵循相关标准（I2）；在可能的情况下使用受控词表、关键词、叙词表或本体（I3）；并为所有相关数据提供限定引用和链接（I4）。对于可互操作性指标，测试数据集仅在 I2 上获得了一个是 ，以及3 个部分 的响应。**可复用性** 指标包括：数据准确，并通过许多相关属性得到充分描述（R1）；数据具有清晰且可访问的数据使用许可（R2）；明确说明数据为何以及由谁创建和处理（R3）；并且数据和元数据符合相关领域标准（R4）。在最后一项测试中，又获得了 2 个是和 2 个部分 的响应。

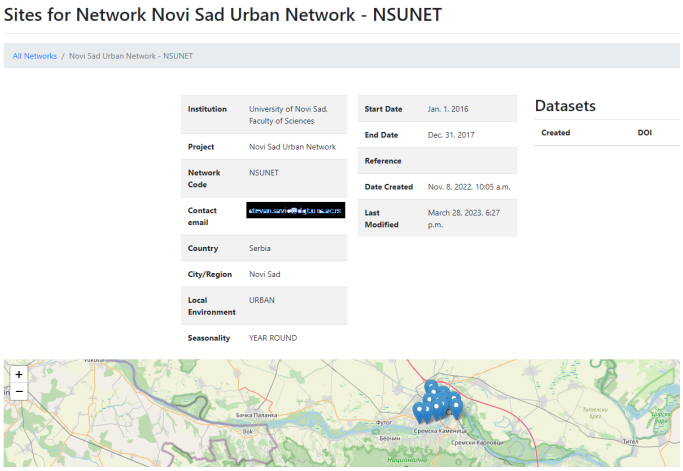


图4：检查搜索结果

该研究在 FKP 开发之前开展，为如何将 FAIR测试纳入 FKP 提供了有益的指导，并促使人们重点关注互操作性指标，因为这是最难完全达标的指标。

3. 微型气候知识门户

在本节中，我们使用一个中等规模的微气象网络来展示当前版本 FKP 的功能。生成该数据集的研究项目 [17] 促成了诺维萨德城市网络（NSUNET）的构建。该网络由 28 个远程站点和 2 台服务器组成，完全基于开源技术构建。其目标是监测气候事件；采集诺维萨德城市区域的长期气象数据；以及向城市应急服务部门发送当前城市天气状况的预警通知。这类部署的主要优势之一，是能够以较低的互联网服务费用运行，并确保低预算硬件上的高可靠性和高性能 [17]。在这里，我们可以

Name	Latitude	Longitude	Altitude (M)	Time Zone	Macroscale Environment	
s2-2	45.249166	19.837222	79.00000	Europe/Belgrade	Urban Street Canyon	View
s2-3	45.261388	19.848888	78.00000	Europe/Belgrade	Residential Area (Multi-Story Buildings)	View
s3-2	45.233333	19.809722	79.00000	Europe/Belgrade	Residential Area (Houses)	View
s5-2	45.25	19.816111	75.00000	Europe/Belgrade	Boulevard	View
s5-3	45.2625	19.826388	78.00000	Europe/Belgrade	Residential Area (Multi-Story Buildings)	View
s5-4	45.238055	19.832777	81.00000	Europe/Belgrade	Residential Area (Multi-Story Buildings)	View
s5-5	45.253055	19.8475	80.00000	Europe/Belgrade	Residential Area (Multi-Story Buildings)	View
s5-6	45.2425	19.847222	78.00000	Europe/Belgrade	Residential Area (Multi-Story Buildings)	View
s6-4	45.233611	19.791944	76.00000	Europe/Belgrade	Residential Area (Houses)	View
s6-8	45.251388	19.875555	76.00000	Europe/Belgrade	Residential Area (Houses)	View
s6-9	45.240555	19.881111	92.00000	Europe/Belgrade	Residential Area (Houses)	View

图5：检查站点详细信息

使用该网络展示部分搜索功能；在通过图2发起的搜索中，它是定位到的 2 个网络之一，并在结果中的图3中显示为诺维萨德城市网络。

存储在 FKP 上的数据集是原始实验数据的一个子集，并且可公开获取 [15]。元数据描述了在 2016 年和 2017 年这两年期间，从诺维萨德的 12 个城市站点每小时采集的气温数据。该集合中有 2 个数据集：一个数据集提供有关安放温度传感器的 12 个站点的详细信息，另一个文件则包含这 12 个地点的气温数据。第二个数据集总共包含 17,544 条气温数据记录。气温数据经过清理并补齐缺失值，因此每个站点每天都有 24 条测量记录。从结果页面选择目标网络后，图4会显示网络元数据以及一张交互式地图，其中标出了 12 个站点的位置。FKP 门户支持逐层查看站点，如图5所示。

随后，用户可以进一步深入查看任意站点的元数据。图6展示了诺维萨德网络中某个站点所采集的元数据。

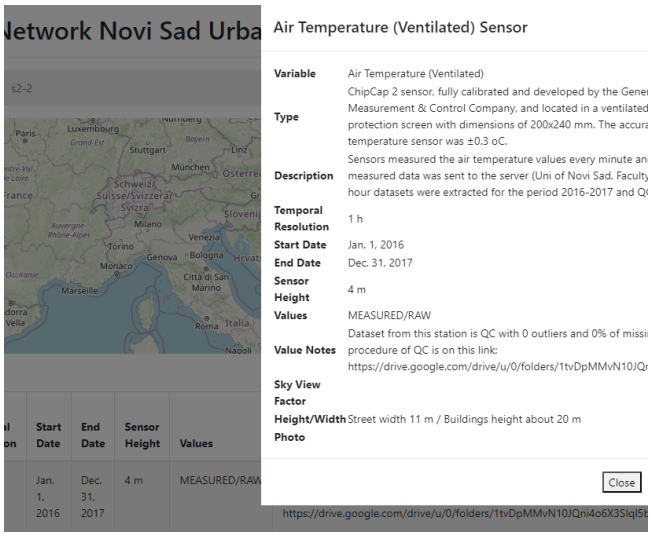


图6：特定站点详情

当前指标包括：网络数量；网络平均规模（站点数量）；各网络/站点的数据集日期和规模；各站点的传感器数量和类型。

4. KSP分析

系统已实现维度分析，用于监测网络的数量和规模，并在网络类型、日期和地理位置方面提供一定程度的覆盖。维度分析是指可以按网络、地理位置、日期以及城市/农村值分析 FKP 元数据。随后可通过下钻进行二维或三维分析，将维度指标组合起来。例如，可以统计或定位特定日期范围内、特定国家的城市网络数量。

5. 结论

本文介绍了 FAIRNESS 知识门户。该门户作为 CA20108成本行动 的一部分开发而成，并通过分步指南展示如何搜索和浏览网络中的微气象元数据。FKP 的当前版本（v2.0）仅向成本行动成员开放，因为随着新功能的持续开发，我们目前仍处于测试阶段。预计这一状态将在成本行动结束时改变，届时气候科学界的更广泛成员都将能够访问。目前的计划包括新增

用于评估数据质量的工具和服务，包括数据缺口的程度；在某些情况下，还将提供机器学习工具，为符合特定标准的数据集填补缺口。

参考文献

[1] D'Aquin 等。《FAIREST：评估研究数据仓库的框架》，Data Intelligence 5（2023）。DOI：10.1162/dint a 00159。-[2] 哥白尼计划：欧盟地球观测计划，2019。网址：copernicus.eu/sites/default/files/Brochure Copernicus 2019%20updated 0.pdf [3] 欧洲中期天气预报中心。《用户指南》，2011。网址：https://confluence.ecmwf.int/display/FUG/Forecast+User+Guide[4] 微气象测量 FAIR 网络，欧盟 COST 行动，2021。网址：https://www.fairness-ca20108.eu/ [5] https://fluxnet.org/data/。最后更新：2020。[6] 政府间气候变化专门委员会。《全球升温 1.5°C》，2019。可查阅：https://www.ipcc.ch/site/assets/uploads/sites/2/2022/06/ SR15 Full Report HR.pdf [7] Lalic 等。《塞尔维亚的景观物候建模与决策支持》。载于《景观建模与决策支持》，Mirschel 等编，施普林格，2020。DOI 10.1007/978-3-030-37421-1 29。[8] Lalic B.、Koci I. 和 Roantree M.《微气象数据的 FAIR 性与负责任的研究与创新：气候研究的开放框架》。第二届 AGROECOINFO 会议论文集，希腊，2022。可查阅：https://doras.dcu.ie/28012/1/Lalic.Koci.Roantree-Open.Framework.for.Climate.Research.pdf [9] Lappalainen 等。《泛欧亚实验（PEEX）计划的丝绸之路议程》，《大地球数据》，2:1，8-35，2018。DOI： 10 1080 20964471 2018 1437704 。/ . . Loebbecke, [10] Claudia 和 Crowston, Kevin，《知识门户：组成部分、功能与部署挑战》（2012）。ICIS 2012 论文集 6。可查阅：https://aisel.aisnet.org/icis 2012 6 / [11] proceedings/KnowledgeManagement/ McCarren 等。《农业数据仓库构建中的异常检测》。澳大利亚计算机科学周（ACWS）论文集，ACM 出版社，2017。可查阅：https://dl.acm.org/doi/pdf/ 10 1145 3014812 3014829 。/ . Muller 等。《全球大气陆地温度的十年变化》。《地球物理研究杂志：大气》，第118卷，5280–5286，2013–。[13] Pastorello 等。《一套用于更精确监测陆地与大气交换的新数据集》。《EOS 地球与空间科学新闻》。98，2017。DOI：10.1029/2017EO071597。[14] Roantree M. 和 Liu J.《一种选择待实体化视图的启发式方法》。《软件：实践与经验》44(10)，第1157–1179页，Wiley，2014。[15] Savic S.、Secerov I.、Dunjic J. 和 Milosevic D.《诺维萨德市逐小时气温数据集——NSUNET 系统》，2023。DOI：10.5281/zenodo.7738093。[16] Scriney M. 等。《从半结构化数据源自动构建数据集市》。《计算机杂志》，第62卷，第3期，第 394 413, 2019 10 1093 064 –页。[17] https://doi.org/ . /comjnl/bxy ˇSecerov, I.B., Savic, S.M., Miloˇsevic, D.D. 等。借助高密度监测网络系统推进城市气候研究。《环境监测与评估》191, 89, 2019 10 1007 10661 019 7210 0 。DOI：. /s - - [18] - Wilkinson, M., Dumontier, M., Aalbersberg, I. 等。《科学数据管理与管理的 FAIR 指导原则》。《科学数据》3，160018，2016，10 1038 2016 18（）。https://doi.org/ . /sdata. . . [19] 世界经济论坛。《全球风险报告》（第14版），2019。可查阅：https://www3.weforum.org/docs/WEF Global Risks Report 2019.pdf。[20] 世界气象组织。《农业气象实践指南》（GAMP），WMO-No.134，2010。

致谢

本研究获得以下经费支持：欧洲科学技术合作组织（COST）资助项目 CA20108、爱尔兰科学基金会资助项目 SFI/12/RC/2289 P2，以及塞尔维亚共和国教育、科学和技术发展部资助项目 451-03-47/2023-01/200117。

如何引用：

Roantree, M.、Lali´c, B.、Savi´c, S.、Miloˇsevi´c, D. 和 Scriney, M.，《构建 FAIR 气候数据的可搜索知识库》，2023 年 EGU 大会，奥地利维也纳，2023 年 4 月 24–28 日，EGU23-7786，https://doi.org/10.5194/egusphere-egu23-7786，2023 年。