arXiv:2603.24147v1 [cs.DL] 25 Mar 2026

将全球科学资助与研究出版物关联起来

Jacob A. Dalsgaard、[1],Filipi N Silva[2],和 Jin Ai[3]

[1]哥本哈根大学社会数据科学中心，丹麦[2]印第安纳大学，美国[3]罗格斯大学纽瓦克分校，美国

摘要

学术出版物中的资助致谢为支持科学研究的组织提供了大规模追踪数据。我们构建了一个数据集，通过系统消歧从出版物元数据中提取的740 万条独特资助致谢字符串，将全球科学资助机构与研究出版物关联起来。我们采用多阶段流程，将资助方名称与标准化组织标识符进行匹配；该流程结合了词汇规范化、基于相似度的聚类、基于规则的匹配、命名实体识别辅助和人工验证。最终的数据集将 190 万条独特资助方字符串关联至规范组织标识符，并记录匹配类型和未解决案例，以支持透明度。技术验证包括跨文献计量学来源的论文级比较，以及对全文致谢部分的人工核验，并报告召回率和精确率指标。该数据集支持对全球科研体系中的资助流向、机构资助组合、地区代表性和集中模式进行分析。

1 背景与概要

自范内瓦·布什战后提出公共资助研究的愿景以来，资助一直构建着现代科学的架构 [1]。它为科学发现和技术创新提供支持，塑造着哪些问题会得到研究、哪些机构和地区能够获得资源，以及研究体系如何随时间演变。然而，全球科学资助体系的结构和稳定性正经历重大变化 [2]。公共研究

资助历来是基础科学和应用科学的主要支持来源，如今却日益受到预算波动、政策优先事项变化，以及来自慈善机构和企业的跨部门参与不断加剧等因素的影响 [3, 2]。与此同时，科学研究的国际合作日益增多，项目、作者署名和资助也越来越跨越国界 [4]。

这些发展共同凸显了建立一体化、透明的数据基础设施的必要性，以便系统地将不同国家和部门的资助信息与科学产出联系起来。缺乏此类基础设施，全球科学资助的实证分析就会持续碎片化，且难以进行比较，从而限制我们追踪资助流向、评估公平性和集中度，或考察资助体系如何塑造科学发现轨迹的能力 [5]。

然而，当前对全球研究资助格局的实证分析，仍受到现有数据来源碎片化、不一致和覆盖不完整的制约 [5]。资助信息的记录方式各异，提取自学术记录的不同部分，并受到不同消歧实践的影响 [6, 7]。这些不一致妨碍了数据库之间的互操作性，并给科学学和政策研究带来隐性测量误差。此外，缺乏对主要资助数据库进行系统的论文级比较，使其覆盖范围、优势和偏差难以明晰，研究人员因而无法明确了解数据库的选择可能如何影响实证结论 [5]。

若干重要数据集提供了关于科研资助的局部视图，但都不够完整。IRIS UMETRICS 提供来自参与其中的美国大学行政记录的高度详细交易数据，便于对科研投入和支出进行细粒度分析。然而，其覆盖范围主要限于美国机构和资助方 [8]。OpenAlex 汇总 Crossref 元数据中的资助信息，并运用基于 Microsoft Academic Graph 和研究机构注册库（ROR）训练的机器学习模型，对机构和资助方进行消歧 [9, 10]。尽管其覆盖范围广泛且开放，资助来源的代表性仍因地区和资助方类型而有显著差异。Dimensions 将结构化资助记录与文本挖掘得到的致谢数据相结合，并侧重于资助项目级关联，但其数据在地域上较为集中，且匹配算法的透明度有限 [11, 12]。高等教育研发调查（HERD）和企业研发调查（BERD）等基于全国调查的数据集提供了有价值的汇总统计数据，但无法

将资助与单篇出版物关联起来，因此在出版物级资助分析中的用途有限 [13, 14]。

这些基础设施共同面临的一项核心挑战，是缺乏统一且具有全球可比性的资助方身份映射。资助致谢通常以自由文本字符串的形式记录，导致命名规范、缩写、翻译方式以及组织粒度层级存在广泛差异。同一资助机构可能以数十种名称变体出现，而相似的字符串也可能指向不同的实体。若缺乏系统性的消歧以及与持久标识符的交叉映射，这些不一致性将降低不同研究之间的可靠性、可比性和可复现性。

近期研究表明，在一定程度上应对这些挑战后，资助致谢数据具有重要的分析价值。例如，Miao 等人 [15] 使用 Web of Science（Web of Science）的资助致谢数据，记录科学资助的集中度和国际相互依存模式。与此同时，他们的研究结果也凸显出，研究结论对基础资助数据的完整性和代表性较为敏感。这表明，有必要采用透明的消歧流程，并开展系统性的跨数据库验证。图 4 进一步印证了这一模式：该图显示，Dimensions 数据库主要收录来自高收入国家的资助方，在许多全球南方地区的代表性有限。相比之下，Web of Science 展现出更广泛的资助机构地理分布，凸显出 Dimensions 中资助方全球覆盖范围较为狭窄。

在本文中，我们通过构建一个大规模、经过消歧的科学资助方名称字符串数据集来应对这些局限。该数据集源自 Web of Science 中的资助致谢，并在可能的情况下系统地关联至 OpenAlex 和 Research Organization Registry（ROR）中的规范组织标识符。我们实施了一个由启发式方法驱动的多阶段消歧流程，整合了词汇规范化、首字母缩略词检测、MinHash局部敏感哈希、基于规则的匹配、命名实体识别辅助、基于相似度的回退方法，以及有针对性的人工验证。重要的是，我们并未强制进行穷尽式整合，而是保留未匹配和多重匹配的资助方字符串，从而透明地保留歧义信息，并支持灵活的下游使用。

除了发布数据集外，我们还对 Web of Science、Ope- nAlex 和 Dimensions 中的资助信息进行技术验证和比较评估。我们通过直接匹配论文层级的数据、依据全文中的资助致谢部分进行人工验证，并比较

分析地域覆盖范围和资助方集中度，量化不同文献计量学来源在召回率、精确率和覆盖范围上的差异。我们的结果显示，资助方的呈现情况存在显著差异，尤其是出现频率较低的组织以及位于全球南方的资助方，这凸显了系统性消歧和跨数据库基准测试的重要性。

通过整合主要文献计量数据库中的资助方信息并建立交叉映射，本研究为大规模分析全球科学资助生态系统提供了可复现的数据基础设施。由此生成的数据集可用于研究跨部门和地区的资助流向、机构资助组合、研究领域内资助方的集中度与多样性，以及对现有资助元数据中偏差的系统评估。更广泛而言，该数据集有助于将资助信息透明、可比地整合到科学学和政策研究中。

2 方法

2.1 数据来源

我们使用了三个主要的文献计量数据库，它们均包含从学术出版物中提取的资助信息：Web of Science、OpenAlex 和 Dimensions。这些数据库在覆盖范围、数据结构和资助方消歧方法方面各具优势，也各有局限。

2.1.1 Web of Science

由科睿唯安开发的 Web of Science 是历史最悠久的引文索引数据库之一，收录论文逾 9600 万篇。图1显示，自2005年以来，包含资助信息的新发表研究的覆盖率从5%上升至2021年的50%以上。资助数据通过“资助机构”和“资助编号”字段记录，并经由半自动化解析流程提取。尽管该数据库在完整性和标准化方面存在已知局限，尤其是非英语出版物，但 Web of Science 的历史覆盖范围相对广泛，且索引流程经过筛选，因此被广泛用于全球资助分析 [6, 7]。我们的处理流程使用了 Web of Science CORE-ESCI 2023 快照。

2.1.2 研究机构注册库

研究组织注册库是一个开放、由社区治理的注册库，为从事研究和研究

资助的全球组织提供持久标识符。ROR 标识符旨在支持学术基础设施中的机构消歧，避免歧义，并日益为出版商、资助方和文献计量数据库所采用。截至2024年，ROR 收录了超过 110,000 个组织，包括大学、研究机构、医院、政府机构和资助机构，每个组织都有一个唯一标识符、一个规范名称、多语言别名、地理元数据，以及指向维基数据等外部标识符的连边。我们的处理流程使用了 ROR v1.72 和 v2 模式 [16]。

2.1.3 OpenAlex

OpenAlex 是一个开放、由社区维护的平台，旨在取代 Microsoft Academic Graph (MAG)，整合了 Crossref 的数据、来自 ROR 的机构标识符，以及一个大规模分类模型，以提升作者和机构消歧的效果。2025年以前，OpenAlex 使用 Crossref 的元数据，在可用时推断资助来源。2026年，OpenAlex 将开始挖掘全文 PDF，以将资助方与成果匹配，并直接从资助方导入资助金元数据 [17]。在 OpenAlex 中，资助方被定义为向学术研究提供资金支持的组织。OpenAlex 通过应用程序接口和完整数据库快照向公众开放，因此尤其适用于大规模、可复现科学研究 [9]。我们所用 OpenAlex 快照的最新更新日期为 2025年9月29日。

OpenAlex 中的机构主要根据从出版物中解析出的作者所属机构信息整理而成。OpenAlex 采用双模型分类系统，结合定制的深度学习模型和经过微调的 Distil-BERT 语言模型，将噪声较多的多语言复合机构字符串映射到规范记录。这些模型使用历史 MAG 数据以及根据 ROR 生成的合成隶属关系字符串进行训练，从而实现更广泛、更准确的覆盖 [10]。

2.1.4 Dimensions

由 Digital Science 开发的 Dimensions 是一个商业平台，将出版物元数据与资助项目、专利和政策文件整合在一起。其资助数据来源于多个渠道：与资助方数据库的直接集成、从全文的致谢部分中挖掘信息，以及推断资助项目与成果之间的连边。Dimensions 强调资助项目级关联，包括资助机构、计划和奖励标识符等信息 [12]。然而，由于其专有性质，原始数据和匹配算法并未完全公开，访问权限也可能受到限制

[11]. 由于访问受限，我们的处理流程中不使用 Dimensions；但在技术验证中，我们使用了于2025年1月5日获取的数据转储。

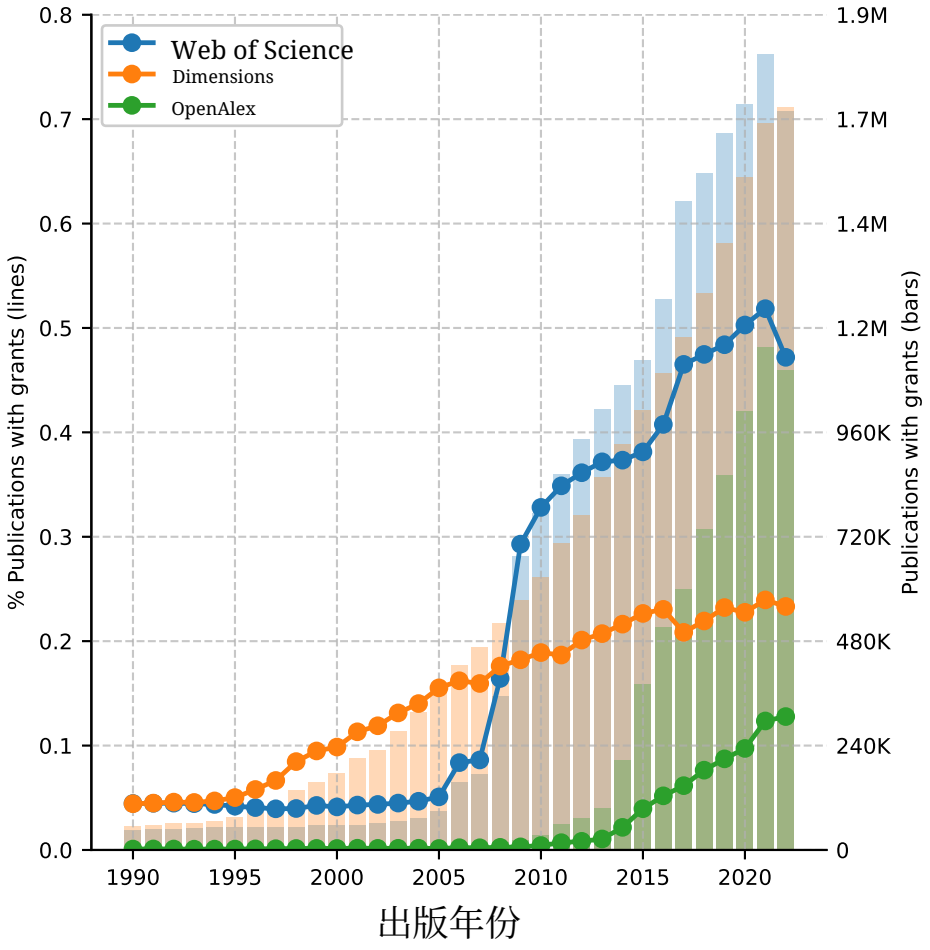


图1：**各文献计量数据库中包含资助信息的出版物覆盖情况**。左侧纵轴显示至少包含一项资助致谢的出版物年度比例（折线图）。右侧纵轴显示至少包含一项资助致谢的出版物年度数量（柱形图）

OpenAlex 和 ROR 的机构参考数据

为了消除 WoS 资助机构字段中所列资助机构的歧义，我们利用 OpenAlex 和 ROR 的数据构建了统一的机构参考索引。OpenAlex 分别提供机构和资助方数据集，但许多组织兼具这两种身份。因此，我们将这些数据集合并为一个机构列表，并使用 ROR 补充缺失的标识符和国家信息。

在合并记录之前，我们统一了 ROR ID 和维基数据标识符的格式，将机构名称标准化为小写，并从主页网址中提取注册域名，将其用作稳定的网络名称。随后，我们分步关联记录：先匹配完全一致的 ROR ID，再匹配完全一致的维基数据标识符，最后匹配相同的主页域名和国家信息。完成初步关联后，我们检查了剩余的重复记录（即多个记录仍指向同一机构的情况）。如果记录具有相同的标准化名称和国家信息，我们就将其视为指向同一机构。发现重复记录时，我们会选取一条规范记录。优先选择来自 OpenAlex 资助方数据集的记录，其次选择具有稳定标识符且活动元数据更完整的记录（例如资助数量或出版物数量）。由于 WoS 使用 ASCII 编码，所有名称均限制为 ASCII 字符。

为确保词汇和检索的完整性，我们将所有已知名称变体合并到一个“其他名称”字段中，并将已识别的首字母缩略词同时存储在该字段内以及单独的首字母缩略词字段中。对于资助数量和出版物数量等数值属性，我们在重复记录中保留最大的非缺失值，以避免重复计数，同时尽可能保留最完整的信息。

这一基于 OpenAlex 构建的综合索引支持与各种实体变体（包括首字母缩略词和译名）进行匹配，从而能够稳健地消歧资助来源。

资助方消歧与匹配流程

我们将 Web of Science 的资助机构字段中列出的资助方名称与统一的组织参考索引进行匹配。该流程如图 2 所示。

首先，我们将所有资助方字符串转换为小写。在可行的情况下，我们还通过提取括号内的大写字母来识别首字母缩略词（例如，从“National Science Foundation (NSF)”中提取“NSF”）（步骤 1）。这有助于识别那些通常以

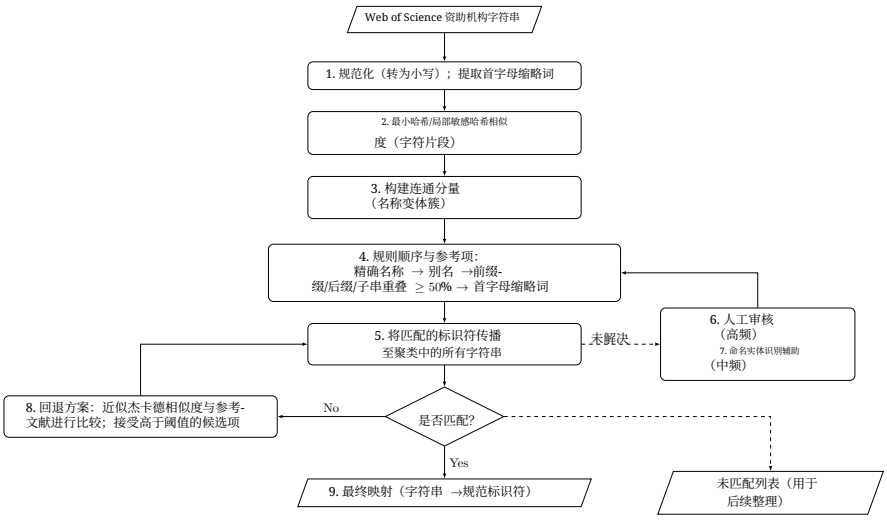


图 2：**资助方名称消歧流程概览**。从 Web of Science 致谢元数据中提取的资助机构名称字符串首先经过规范化处理，并使用 MinHash局部敏感哈希进行聚类，以将相似字符串归为一组。随后，使用一系列确定性规则，将候选聚类与从 OpenAlex 和 ROR 构建的参考索引中的组织进行匹配，这些规则包括精确名称匹配、别名匹配、子串匹配和缩略词匹配。对高频未解决字符串进行人工审查，并使用命名实体识别处理中频案例。其余未匹配字符串通过基于相似度的备用流程进行评估，最终生成资助名称字符串与规范组织标识符之间的最终映射。

缩写形式称呼的组织。

接下来，为了减少格式差异和轻微拼写差异造成的变异，我们将相似的资助方字符串归为一组。我们使用 MinHash局部敏感哈希测量字符串的近似相似度，并将相似度高于 0.95 阈值的字符串视为同一底层名称的变体（步骤 2）。我们构建了一个网络，其中节点代表资助方字符串，连边连接相似度超过该阈值的节点。我们通过识别网络的连通分量对字符串进行分组（步骤 3）。这些组被视为代表单个组织的候选聚类。

随后，我们尝试将每个聚类与参考索引中的一个组织关联起来，

采用结构化的匹配规则序列（步骤 4）。我们首先检查是否与已知名称精确匹配，并记录别名。如果未找到精确匹配，我们会查找一个名称是否包含在另一个名称中的情况。我们先检查名称是否出现在前缀或后缀位置，然后再考虑子字符串匹配。无论哪种情况，只有当子字符串至少覆盖字符串中一半的字符时，我们才将其视为匹配。如果仍未解决，我们会比较首字母缩略词。匹配完成后，便为聚类中的所有名称分配规范组织标识符（步骤 5）。

仅依靠自动化规则，在一次遍历步骤 1-5 的过程中，有些字符串无法得到可靠匹配。我们对其中最常见的情况进行人工审核，将出现次数超过 1,000 次的资助方字符串定义为此类情况（步骤 6）。在人工标注过程中，我们首先搜索 OpenAlex API，以查找候选匹配项。随后，我们通过验证候选项或进行网络搜索，手动分配 OpenAlex 资助方 ID。完成人工标注后，我们回溯至步骤 4，在再次运行步骤 4 和 5 之前应用人工标注。接下来，我们改进了对未匹配的中频情况（即出现次数在 100 至 1,000 次之间的资助方字符串）的匹配（步骤 7）。我们使用 GLiNER [18] 进行零样本命名实体识别，以识别字符串中的机构名称，然后再次尝试匹配（步骤 4-5）。最后，对于仍未解决的字符串，我们使用另一个由剩余字符串构建的 LSH 索引进行额外的相似性搜索，并查询所有可能的参考名称。我们接受相似度高于 0.9 这一保守相似度阈值的匹配。有关所用阈值、参数和模型的详情，请参阅补充信息和代码。

这种方法能够高效地对多样且多语言的资助方字符串进行消歧。如果匹配过程的任何环节产生了多个合理的候选项，我们都会将其保留，并将最终的机构消歧推迟到单篇论文层面，以便利用上下文元数据（如合著者所属机构、国家、学科领域）进行更准确的分类（见第 5.1 节）。具体而言，我们的方法优先考虑字符串之间的词汇和结构相似性，这一优先级也体现在字符串的匹配顺序中。使用大语言模型进行的初步实验表明，在要求其判断我们流程识别出的匹配是否正确时，它们往往会混淆实体，或给出信心十足但错误的判断（见补充信息）。消歧流程的所有步骤，包括预处理、聚类、基于规则的匹配以及人工标注整合，均通过随本研究提供的开源软件包实现。代码仓库包含必要的脚本，可根据原始 Web of Science 资助

名称字符串和 OpenAlex/ROR 参考索引重现该数据集，但须遵守底层文献计量数据库的许可限制。

3 数据记录

该数据集提供了一份消歧后的全球科学资助机构列表，其构建方法是将出版物致谢中的资助方字符串与标准化的机构标识符进行匹配。该数据集以单个 Parquet 文件发布，文件名为 *wos funders.parquet*，可从 Zenodo [19] 获取。

表 1 描述了与该数据集相关的数据字段。数据集中的每一行代表一个资助机构字符串—ID 对。我们统计每个资助方字符串的出现频率，并尽可能将每个字符串关联到两个标准组织标识符：OpenAlex ID 和 ROR ID。

列	Type	# 行数 with non-空值	描述
资助机构	字符串	11,086,346	从出版物致谢中提取的原始资助方名称字符串。
id	整数	5,639,378	资助方条目的 OpenAlex ID。
数量	整数	11,086,346	出版物数量，其中识别出此资助机构字符串。
ids:ror	字符串	5,639,271	消歧后 ROR（研究组织注册库）标识符，与资助方关联。
来源	字符串	5,639,271	该标识符是否为 OpenAlex 资助方 ID 或机构 ID。
显示名称	字符串	5,639,271	规范名称或首选名称来自 OpenAlex。
匹配类型	字符串	5,639,378	消歧的具体部分负责此项工作的处理流程匹配。

表 1：消歧后的科学资助者数据集的字段描述。

4 技术验证

在最初的 740 万个不同资助方字符串中，我们为其中 190 万个找到了至少一个可能的匹配项，覆盖了 Web of Science 中论文提及资助方的所有实例的 72%。表2按资助方字符串的出现频率分组，汇总了不同资助方字符串的总数和匹配数。通过关注结构相似性和词汇相似性，该数据集对高频和中频资助方字符串实现了良好覆盖，但仍有大量多样且低频的资助方字符串未能匹配。表 3 描述了消歧流程的各个部分分别负责找到多少个匹配项。

频率范围	字符串总数	匹配的字符串	未匹配率
≥ 1000	2901	2898	0.001
100–1000	24001	19069	0.205
10–100	263449	110836	0.579
1–10	7108735	1881783	0.735

表2：根据是否存在 OpenAlex ID，按频率范围统计未匹配的资助方字符串比例。

匹配类型	数量
未匹配（NaN）	5,384,500
前缀或后缀匹配	1,049,038
首字母缩略词匹配	522,745
子串匹配	231,125
文档聚类	91,996
名称（精确）	43,021
别名（精确）	42,707
Jaccard 回退	33,307
人工标注	647

表 3：按消歧方法分类的资助机构字符串匹配数量。

4.1 跨数据集论文的直接匹配

为评估 Web of Science、OpenAlex 和 Dimensions 之间资助信息的召回率与互补性，我们首先找出同时被 Web of Science 和 OpenAlex 收录的 200 万篇文章，以及同时被 Web of Science 和 Dimensions 收录的 400 万篇文章——

这些文章均使用持久标识符 DOI 进行匹配，并且仅纳入资助信息完整的记录。对于每篇论文，我们计算 Web of Science 与 OpenAlex 中资助方标识符的交集。Dimensions 使用不同的机构标识符（GRID），因此我们借助 ROR 元数据在不同标识符之间进行转换。交集非空即定义为命中。如果交集包含任一数据库中的全部资助方，则定义为完全命中。我们发现，88% 至 89% 的论文中，OpenAlex 和 Dimensions 报告的至少一位资助方也出现在 Web of Science 中；而在 OpenAlex 和 Dimensions 收录的论文中，分别有 80% 和 67% 的论文所列出的全部资助方也都列于 Web of Science 中。反过来，在 OpenAlex 和 Dimensions 中包含资助信息的论文里，分别有 66% 和 64% 列出了 Web of Science 论文中归属的全部资助方（表 4）。

方向	完全命中率	命中率
OpenAlex ⊆Web of Science	0.8	0.89
Dimensions ⊆ Web of Science	0.67	0.88
Web of Science ⊆ OpenAlex	0.66	0.89
Web of Science ⊆ Dimensions	0.64	0.88

表 4：数据库间的方向性交集与完全命中覆盖率

4.2 人工评估

为进一步评估我们的消歧方法的准确性，我们从 Web of Science 中随机抽取 250 篇文章进行人工验证，将识别出的资助机构与已发表论文的实际内容进行比对。对于样本中的每篇论文：

1. 使用论文的数字对象标识符（DOI）获取全文版本。
2. 手动查找并审阅资助致谢部分。
3. 我们将论文中的资助方提及与数据集中的消歧后的名称集合进行比较。

对于每篇论文，我们记录以下内容：

- **资助方总数**：我们统计全文资助和致谢部分中报告的不同资助机构总数。

- **正确识别数量**：我们统计正确识别的资助方数量。

- **错误数量**：我们统计错误识别的资助方数量

通过这种标注，我们可以计算每篇论文的召回率和精确率，并通过对所有论文取平均值进行汇总，同时将所有论文的数值相加，得出总体统计数据。此外，我们还报告错误率，其定义为错误数量除以一篇论文的资助方总数。最后，我们报告命中率和全命中率，分别定义为至少有一个资助方被正确识别的论文比例，以及所有资助方均被识别的论文比例。

表 5 报告了 Web of Science、OpenAlex 和 Dimensions 的各项指标。OpenAlex 和 Dimensions 的指标分别基于样本的两个子集（N=100）计算，并要求每篇文章必须同时被 Web of Science 和 Dimensions、或 Web of Science 和 OpenAlex 收录，且至少关联一个资助方。Web of Science 中经过消歧的资助信息总体表现更好。单篇论文的标注可在随附的代码仓库中查阅，其中还提供了 DOI 和标题。

	平均召回率	平均精确率	平均错误率
Web of Science	**0.78**	**0.96**	**0.04**
OpenAlex	0.68	0.94	0.05
Dimensions	0.68	0.88	0.13

	召回率	精确率	错误率	命中率	全部命中
Web of Science	**0.73**	**0.96**	**0.03**	**0.98**	**0.54**
OpenAlex	0.64	0.95	0.04	0.95	0.42
Dimensions	0.6	0.86	0.1	0.96	0.48

表 5：资助方识别性能

4.3 不同来源数据库的覆盖范围比较

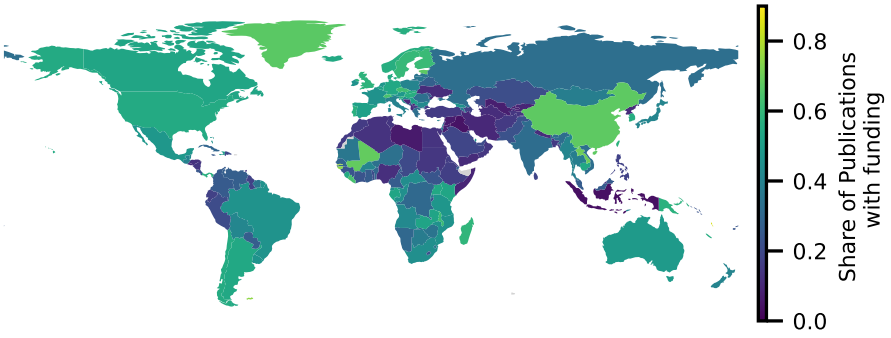
由于发布的数据集是基于 Web of Science 的资助致谢构建，并以 OpenAlex 和 Dimensions 为参照进行评估，因此

有必要说明不同来源的资助元数据覆盖率有何差异。因此，我们比较了自2010年以来发表的与资助方相关的出版物在国家层面的归属情况，以及资助记录在各组织之间的集中度。这些比较旨在对依赖来源的覆盖模式进行技术验证，并记录潜在偏差，供用户在决定使用哪种资助元数据时参考。

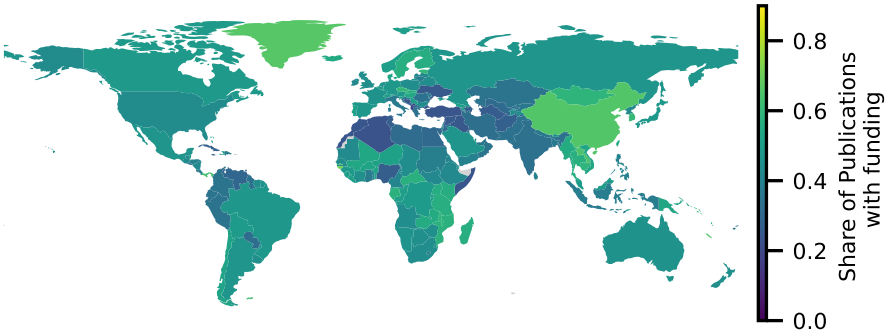
图3按照 [15] 的方法，根据作者隶属机构将论文归属到各个国家，并比较其中记录了资助信息的出版物在各国出版物中的占比。在这项比较中，Web of Science 和 Dimensions 在北美洲、中国和欧洲记录的资助覆盖率较高，而 OpenAlex 的整体覆盖率较低。Dimensions 在全球南方许多国家记录的资助覆盖率也较低。这些差异体现了不同来源在采集资助元数据方面的差异。

图4根据匹配到的资助机构所在地，将出版物归属到各个国家，并比较其中记录了资助信息的出版物的国家分布。在这项比较中，Dimensions 记录的与资助方相关的出版物中，高收入国家所占比例更高；而 Web of Science 显示，匹配到的资助方地理分布更广，其中与非洲、南美洲和东南亚国家相关的出版物更多。OpenAlex 总体上呈现出与 Web of Science 相似的模式，但在中非和西非部分地区以及中亚的覆盖明显更为稀疏。这些比较表明，不同数据库所记录的资助方地理分布差异显著，在解读国家层面分析时应将其纳入考量。

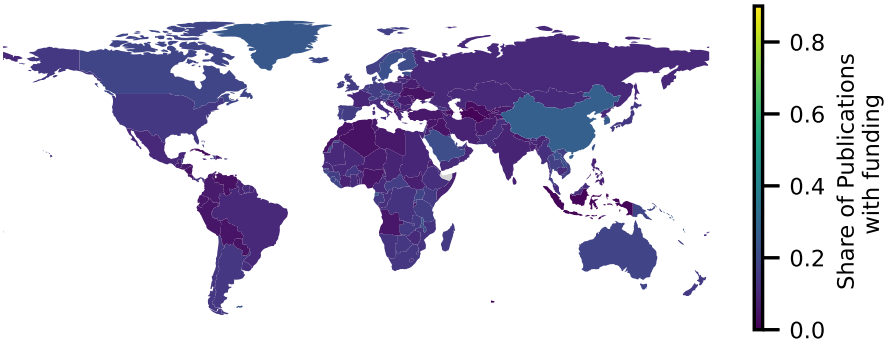
最后，图5展示了出版物的秩频分布。我们发现，1311个资助方为全部出版物中的80%提供资助。尽管这种集中度表明，聚焦排名靠前的资助方即可涵盖大部分资助活动，但这也凸显了排除排名较低的资助方可能造成的遗漏：20%的资助方覆盖了全部出版物的97.5%。我们还展示了不同部门的集中度，结果表明，与企业和慈善资助相比，政府资助集中在少数机构。许多规模较小或具有地区特色的资助机构，为小众研究领域、职业生涯早期资助以及多样化的机构支持作出了贡献。Dimensions 这样的数据集只包含不到700个不同的资助方，因此可能忽略这些贡献，并可能使分析偏向历史悠久、知名度高的资助机构。



(a) Dimensions

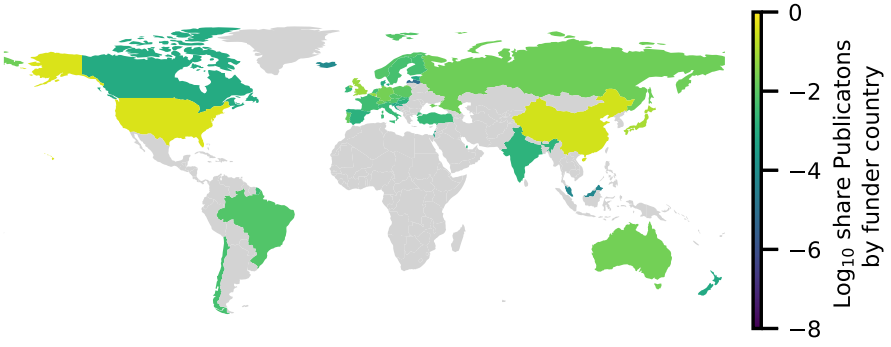


(b) Web of Science

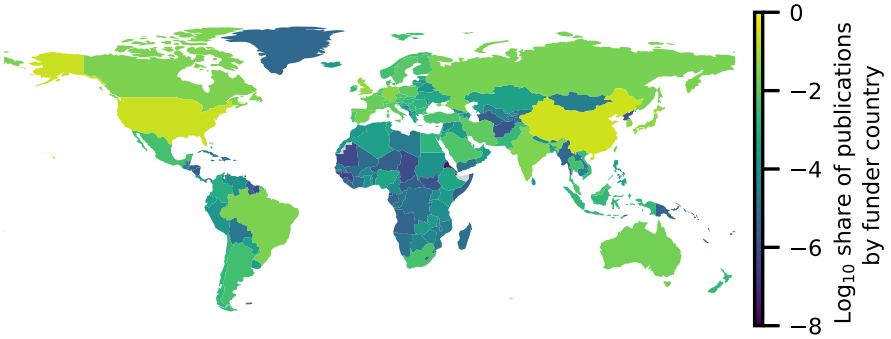


(c) OpenAlex

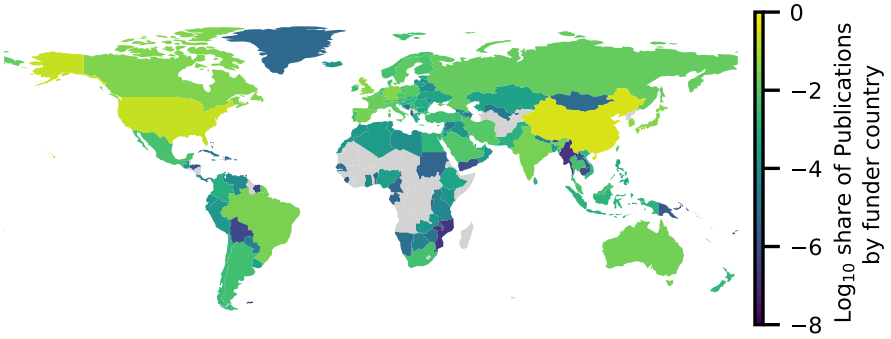
图3：**按作者所属机构划分的国家层面资助归属覆盖率**。地图显示了在 (a) Dimensions、(b) Web of Science 和 (c) OpenAlex 中，按作者所属机构所在国家划分的、记录了资助信息的出版物比例。



(a) Dimensions



(b) Web of Science



(c) OpenAlex

图4：**各文献计量数据库中资助机构的地理分布**。地图展示了各国资助方所资助的出版物中带有资助致谢的比例。各面板对比了 (a) Dimensions、(b) Web of Science 和 (c) OpenAlex 中与资助方相关的出版物的地理分布

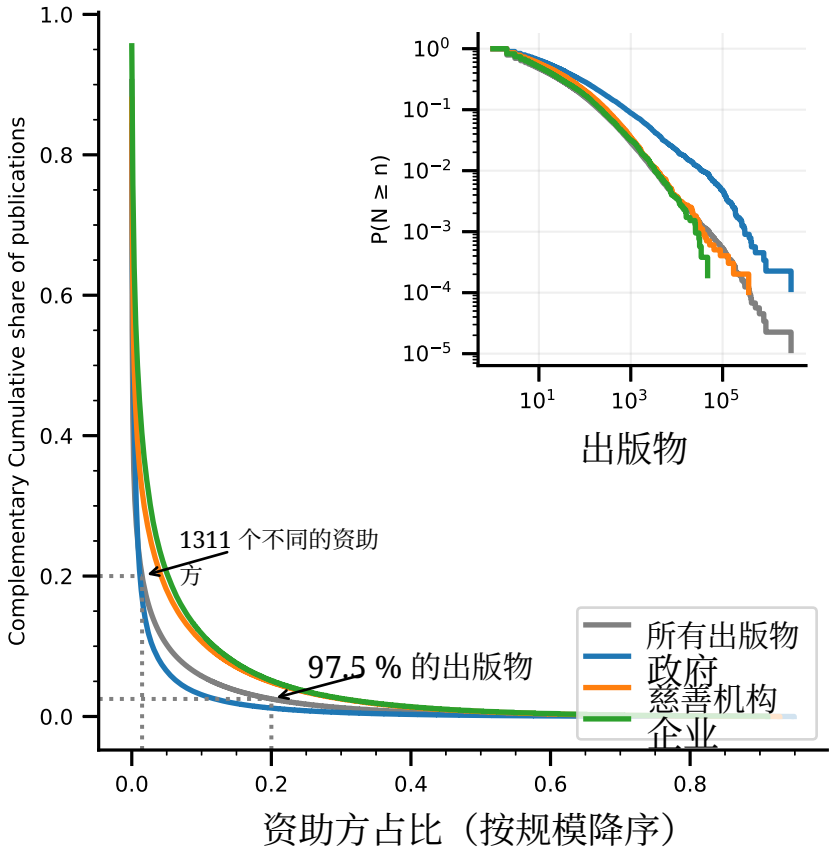


图5：**归属于资助机构的出版物频率排名分布**。根据 Web of Science 数据集中注明获得资助支持的出版物数量对资助机构进行排名。该图展示了各资助方所关联出版物的累积分布，说明资助归属集中在数量相对较少的组织中。图中还展示了按部门划分的分布（政府、企业和慈善机构），以突出不同类型组织的资助活动集中度差异。

5 使用说明

该数据集支持科学政策、文献计量学和创新研究交叉领域的广泛分析。用户可以开展多种研究，包括：

• **追踪跨部门和地区的资助路径**。通过将资助方标识符与机构元数据关联起来，用户可以研究政府、慈善机构和企业资助方流向科研机构的资源，进而分析跨部门合作和公私资助动态。

•**考察机构专业化与战略协同**。该数据集使用户能够绘制哪些机构从哪些类型的组织（如国家机构、使命导向型资助方和国际组织）获得资助，从而开展有关机构资助组合、专业化和战略合作伙伴关系的研究。

• **分析特定研究领域中资助方的集中度与多样性**。通过将该数据集与主题分类体系（如研究领域、期刊类别）相结合，可以考察某些领域的研究是由多元化的资助方群体资助，还是仅由少数资助方资助，从而为有关独立性、竞争和创新风险的讨论提供依据。

• 地理空间资助流向制图</style】：按资助方来源国家或地区汇总资助致谢信息（利用 OpenAlex 地理位置元数据），从而以空间可视化方式呈现科学投资模式。

5.1 解决论文—资助方字符串对中的模糊匹配问题

发布的数据集为每个 Web of Science 资助机构字符串提供一个或多个候选 OpenAlex/ROR 标识符。在大多数情况下，一个资助机构字符串可以映射到单个规范组织。然而，有些字符串仍存在歧义，因此保留了多个合理的候选项。重要的是，这种歧义出现在论文—资助方字符串对这一层级：同一篇论文可能会致谢多个不同的资助方；当存在多个候选匹配项时，我们的目标是确定每个被致谢的资助方字符串对应哪个规范组织。

为支持需要为每个论文—资助方字符串对指定单一标识符的下游分析（例如，按国家、领域或机构汇总资助情况），我们实施了一个可选的论文层级消歧步骤，为每个存在歧义的对选择一个候选机构。我们在验证中使用了此流程，也可以借助随附的软件包复现。

对于存在多个候选匹配项的任意（论文、Web of Science 资助方名称）对，我们首先从 Web of Science 元数据中提取作者所属机构所在地。我们按作者的位置顺序处理：第一作者、最后一位作者、第二作者，然后是第三作者；并尝试识别每位作者所属机构所在的国家。对于每个作者位置，我们检查是否有候选资助机构在同一国家注册。如果发现某个作者位置对应的国家匹配项，我们就选择匹配的候选机构并结束处理。

如果没有候选匹配项与任何可用的作者所属国家相匹配，或者有多个候选资助机构符合国家匹配标准，我们就依据资助方在 OpenAlex 中的普遍程度来消除歧义。具体而言，我们选择在 OpenAlex 中总体普遍程度最高的候选项（例如，作为资助机构出现的频率最高）。当地域信息缺失或不足以唯一确定某个机构时，这一基于知名度的标准可作为保守的平局裁决依据。

对于欧盟资助方，有一种特殊处理方式。我们手动整理了一组欧盟级资助机构（例如欧盟委员会、欧洲研究委员会、地平线计划）。对于这些实体，我们将欧盟机构候选项视为与作者所属国家为任何有资格获得欧盟资助的论文相匹配，其中包括英国和关联国家。这一例外考虑了欧盟资助的跨国性质，并避免仅依据作者所属国家而系统性地错误判定欧盟资助方。

这一消歧步骤对每个论文—资助方字符串对独立执行；因此，经过处理后，一篇论文仍可能对应多个资助方，每个不同的致谢资助机构字符串对应一个资助方。

5.2 使用更新后的数据源

消歧流程取决于所使用的文献计量数据和参考数据的具体版本。由于资助致谢字符串和注册记录会随时间变化，将该流程应用于更新后的数据版本，可能会产生不同的字符串分布和匹配结果。尤其是人工审核步骤（图 2 中的步骤 6）侧重处理高频未匹配字符串，因此容易受到特定数据集版本中观察到的频率和词汇模式影响。将该流程应用于其他数据快照的用户应重新评估此步骤，并记录任何额外修改，以确保过程透明且结果可复现。

数据可用性

本文介绍的消歧后科学资助者数据集已在 Zenodo [19] 上公开发布。该数据集以单个 Parquet 文件的形式发布，包含从出版物致谢中提取的资助方名称字符串及其匹配的标识符。每条记录都将一个资助机构字符串与标准化的组织标识符关联起来，包括 OpenAlex ID，以及在可用时提供的研究组织注册库 ID。数据集还包含 Web of Science 中每个资助机构字符串的频次、OpenAlex 提供的规范机构名称，以及描述消歧流程的元数据。

OpenAlex 和研究组织注册库中的数据均采用开放许可证（CC0）发布，并根据这些许可证将其标识符纳入数据集。与本文描述一同发布的数据集包含从 Web of Science 索引的出版物致谢字段中提取并经过处理和汇总的信息。底层 Web of Science 记录为专有内容，受许可限制，无法再分发。共享的数据集仅包含资助方名称字符串、汇总数量以及匹配到的组织标识符，不包含任何可能用于重建原始数据库的书目元数据。

代码可用性

用于对资助方字符串进行消歧以及进行论文级归因的代码可在 https://github.com/jacdals97/linking-funding-landscapes/ 获取。该代码库还包含用于人工评估的表格。

参考项

[1] 范内瓦·布什。《科学：无尽的前沿：关于战后科学研究计划的总统报告》，1945年。美国国家科学基金会重印版，2023年。[2]经济合作与发展组织。《经合组织*2025*年科学、技术与创新展望：在不断变化的格局中推动变革》。经合组织出版社，巴黎，2025年。[3] 经合组织科学、技术与创新司。《经合组织主要科学与技术指标：2024年3月刊中的研发及相关要点》，2024年3月。

[4] 美国国家科学委员会。《发现：研发活动与研究发表》，2025年7月。[5]Peter Kokol 和 Helena Blaˇzun Voˇsner。《Scopus、Web of Science 和 PubMed 收录医学期刊文章资助信息的差异》。《医学图书馆协会期刊：*JMLA*》，106(1):81–86，2018年1月。[6] Ad`ele Paul-Hus、Nadine Desrochers 和 Rodrigo Costas。《Web of Science 中资助致谢数据的特征、描述及使用考量》。《科学计量学》，108(1):167–182，2016年7月。[7] Nicola Grassano、Daniele Rotolo、Joshua Hutton、Fr´ed´erique Lang 和 Michael M. Hopkins。《出版物致谢中的资助数据：覆盖范围、用途与局限》。《信息科学与技术协会期刊》，68(4):999–1017，2017年。[8]创新与科学研究所（IRIS）。IRIS UMETRICS 2024版发布数据集，2024年。[9] Jason Priem、Heather Piwowar 和 Richard Orr。《OpenAlex：一个完全开放的学术作品、作者、发表平台、机构和概念索引》，2022年6月。[10]OurResearch。OpenAlex 机构解析。GitHub
https://github.com/ourresearch/openalex-institution-parsing (2026)。[11] Daniel W. Hook、Simon J. Porter 和 Christian Herzog。《Dimensions：构建检索与评估的背景信息》。《研究计量与分析前沿》，3:23，2018年8月。[12] Kidambi, M.。用于强化研究未来的拨款数据。Dimensions博客
h
t
t
p
s://
w
w
w.
dimensions.ai/blog/grant-data-for-strengthening-research-futures/ (2024)。[13] 美国国家科学与工程统计中心。2024年高等教育研发（HERD）调查，2025年2月。[14] 美国人口普查局。企业研发与创新调查，2026年1月。[15] Lili Miao、Vincent Larivi`ere、Feifei Wang、Yong-Yeol Ahn 和 Cassidy R. Sugimoto。《全球科学资助中的合作与相互依存》，2024年2月。

[16] 研究组织注册库。ROR 数据（版本 1.72）。Zenodo。https://doi.org/10.5281/zenodo.17280449（2025年）。[17]Kyle Demes。OpenAlex 中的资助元数据。https://blog.openalex.org/funding-metadata-in-openalex/，2026年1月。[18] Urchade Zaratiana、Nadi Tomeh、Pierre Holat 和 Thierry Charnois。Gliner：使用双向 Transformer 进行命名实体识别的通用模型。载于*2024*年北美计算语言学协会分会年会论文集：人类语言技术（第 *1* 卷：长篇论文），第 5364–5376 页，2024年。[19]Dalsgaard, J. A., Silva, F. N. 和 Ai, J. 将全球科学资助与研究出版物数据集相关联。Zenodo。https://doi.org/10.5281/zenodo.18895449（2025年）。

作者贡献

JAI 和 FNS 提出了初步构想。JAD 和 FNS 负责数据收集。JAD 开发了核心消歧流程和可视化，并撰写了论文初稿。所有作者均参与了论文撰写。作者共同参与了人工评估。

利益冲突

作者声明不存在利益冲突

致谢

数据来源于 Dimensions，这是由 Digital Science 提供的互联研究信息系统 https://www.dimensions.ai

资助

本研究得到了 VILLUM FONDEN 提供的研究资助（37394）的支持。本研究使用了通过分配获得的印第安纳大学 Jetstream2 CPU

CIS230183 来自先进网络基础设施协调生态系统：服务与支持（ACCESS）计划，该计划由美国国家科学基金会的资助项目 #2138259、#2138286、#2138307、#2137603 和 #2138296 提供支持。本研究还得到了美国国家科学基金会 APTO 计划的支持，资助项目编号为 #2404109。