B超越开放性：中国

在 OpenAlex 中的学术数据

张林[1, 2, 3*]，曹哲[1,2]，刘建华[4*]，Nees Jan van Eck[5]

1. 科学、技术与教育评估中心（CSTEA），信息管理学院，武汉大学，中国武汉 2. 信息资源研究中心，信息管理学院，武汉大学，中国武汉 3. 研发监测中心（ECOOM），鲁汶大学 MSI系，比利时鲁汶 4. 北京万方数据股份有限公司，中国北京 5. 科学技术研究中心（CWTS），莱顿大学，荷兰莱顿

***通讯作者：**张林，科学、技术与教育评估中心（CSTEA），信息管理学院，武汉大学，中国湖北省武汉市，电子邮箱：linzhang1117@whu.edu.cn；刘建华，北京万方数据股份有限公司，中国北京，电子邮箱：liujh@wanfangdata.com.cn。

ORCID：
张林：0000-0003-0526-9677 曹哲：
0000-0002-9379-2013 刘建华：
0000-0002-4003-8834 Nees Jan van Eck：
0000-0001-8448-4521

摘要

OpenAlex于2022年推出，是一个完全开放的学术数据来源，承诺比传统专有数据库具有更强的包容性。本研究通过考察OpenAlex对中文期刊及其文章的覆盖率和元数据质量，评估其是否兑现了这一承诺。我们以《中国核心期刊要目总览》2023年版（GCJC）和万方数据为基准，分析三个方面：（1）期刊层面覆盖率；（2）文章层面覆盖；（3）元数据字段的完整性和准确性。结果显示，OpenAlex仅收录了37%的GCJC期刊及其24%的文章，且在学科和时间方面存在显著差异。元数据质量参差不齐：标题、出版年份等基本字段较为完整，但书目信息、作者所属机构和被引参考文献常有缺失或不准确的情况。DOI覆盖率有限，语言信息也常常有误，大多数中文文章被标记为英文。这些发现凸显了在科研评价及相关活动中实现全面包容和可用性所面临的重大挑战。最后，我们提出

改进数据聚合策略、DOI注册实践和元数据标准化方面的建议，以促进本地学术成果融入全球开放基础设施。

1. 引言

《开放研究信息巴塞罗那宣言》（2024年）强调，研究信息不仅应当开放，还应当多样且包容（Babini 等人，2024年）。近年来，新兴的书目数据源（例如 OpenAlex ）为研究信息的开放共享与使用创造了新的可能。OpenAlex 于 2022年推出，是一个完全开放的数据源，可免费访问，且允许不受限制地使用其数据（Priem 等人，2022年）。这些开放数据源日益受到研究机构的关注和支持，也越来越多地被应用于研究与评价实践（Cao 等人，2025年），尤其是作为专有数据源的替代方案；后者传统上一直主导着这一领域。例如，索邦大学秉持其对开放研究的承诺，决定不再续订科学网和 Clarivate 的文献计量工具，并将重点转向探索开放替代方案，其中 OpenAlex 成为一个重要候选（索邦大学，2023年）。同样，莱顿大学科学技术研究中心（CWTS）以 OpenAlex 为基础，推出了全新的 CWTS 莱顿大学排名开放版。这是首个针对全球主要大学、完全透明且可验证的排名，其实现得益于 OpenAlex 数据不受限制的访问与再利用（Van Eck 等人，2024年）。尽管 OpenAlex 为传统专有数据源提供了一种不受限制的替代方案，但问题依然存在：它是否真正兑现了让研究信息更加多样和包容的承诺?

除了免费开放且不限制使用之外，OpenAlex 还因其覆盖范围比许多专有数据源更广而受到认可。它收录了数亿篇论文、期刊、作者和机构的信息，以及数十亿条彼此之间的关联。研究人员从多个角度对其覆盖率进行了实证研究，例如开放获取期刊的覆盖情况（Maddi 等人，2025年）以及对撤稿论文的收录情况（Ortega 等人，2024年）。与此同时，由于数据汇集自元数据质量不一的来源，也出现了一些挑战。Delgado-Quirós 等人（2024年）指出，尽管 OpenAlex 提供大量出版物字段（如持久标识符、书目信息和类型信息），但部分字段的完整性有限。例如，Zhang 等人（2024年）指出，期刊文章经常缺少机构信息。除了完整性，准确性也是一项备受关注的数据质量问题。例如，Céspedes 等人（2025年）发现，尽管 OpenAlex 的语言覆盖比科学网更为均衡，但其语言元数据并非总是准确，导致高估英语研究成果、低估其他语言的作品。

尽管人们对 OpenAlex 的兴趣和使用日益增长，但相对而言，很少有人关注它是否真正促进了多样性与包容性。传统专有数据源长期以来因覆盖范围有限且存在偏差而受到批评，尤其是在收录来自不同地理区域的出版物方面（Asubiaro 等人，2023年）；OpenAlex 这类开放数据源或可帮助解决这一问题。覆盖有限且存在偏差，会影响数据源在研究议程设定、科学评价、国际比较及其他用途中的可靠性和实用性。现有研究已表明，在非洲出版物的覆盖（Alonso-Alvarez 等人，2025年）以及对代表性不足地区或低收入国家开放获取期刊的收录（Maddi 等人，2025年）方面，OpenAlex 明显优于科学网和 Scopus。然而，聚焦亚洲国家的 OpenAlex 包容性与可用性的系统研究仍然鲜见。亚洲地区的非英语学术成果数量可观（Veugelers，2012年）。有证据表明，OpenAlex 对中国等非英语国家出版物的覆盖存在明显断层和系统性偏差（Zheng 等人，2025年），但这一发现主要基于对 OpenAlex 收录的中国机构署名论文的分析，且缺少来自本地数据源的比较基准。

中国是一个尤其具有代表性的案例。中国学者的发表量和引文量已跻身全球前列（Brainard 等，2022年；Xie 等，2014年），使中国的国际科研产出成为全球科学不可或缺的一部分。然而，值得注意的是，中国科研产出中有较大比例发表于国内期刊，而这部分内容长期以来一直被传统的订阅制数据源所忽视。根据中国科协学会服务中心（2024年）的统计，2023年，自然科学领域的中文科技期刊约发表了997,300篇可引用论文，而被科学引文索引收录的期刊中由中国研究人员撰写的文章仅约728,700篇。因此，从地理角度考察中国本地科研产出的覆盖率，是评估 OpenAlex 数据包容性和可用性的重要测试案例。

鉴于上述背景，本研究聚焦于 OpenAlex 对在中国出版的中文期刊（以下简称“中文期刊”）及其文章的收录情况，重点探讨三个问题：（1）在期刊层面，OpenAlex 收录中文期刊的程度如何？（2）在论文层面，OpenAlex 对中国期刊文章的覆盖率如何？（3）OpenAlex 所收录的单篇中国期刊文章的信息有多完整、准确？本研究以定量分析为基础，结合中国本地知识解读实证发现，并为 OpenAlex 等开放研究信息源的未来发展提出有针对性的建议。

2. 中国期刊与数据源的选择

2.1 中国学术期刊

中国的期刊评价实践可追溯至 20 世纪 60 年代，此后历经显著发展，形成了七种用于研究和评价的知名期刊目录（Huang 等，2021）。它们分别是中国科学院国家科学图书馆的中国科学引文数据库（CSCD）和期刊分区表（JPT）；中国社会科学院评价研究院的*AMI*期刊目录；中国科学技术信息研究所的中国科技期刊引证报告（CJCR）；北京大学图书馆的《中国核心期刊要目总览》（GCJC）；南京大学中国社会科学研究评价中心的中文社会科学引文索引（CSSCI）；以及中国知网的世界学术期刊学术影响力指数（WAJCI）。

在七份期刊名录中，GCJC 名录[1] 是最早的名录之一，创立于1992年。该名录涵盖众多学科，在中国广受认可。GCJC 名录是由北京大学图书馆牵头的一项研究项目的成果，另有十余所大学图书馆及相关机构的专家参与。该名录的候选期刊为在中国大陆出版的中文期刊。自2008年起，GCJC 名录每三年更新一次，以反映中国期刊格局的变化。它是中国出版和科研评价的重要参考。因此，本研究聚焦于 GCJC 名录2023年版收录的1,987种学术期刊，以及这些期刊历年发表的文章。

2.2 用于比较的数据源

本研究考察 GCJC 名录中中国期刊及其文章在 OpenAlex 中的覆盖情况，并评估单篇文章信息的完整性和准确性。OpenAlex 是本研究考察的主要数据源，而万方数据则被选作国内代表性数据源，用于评估 OpenAlex 中中国科研产出数据的质量。

万方数据成立于1993年，由中国科学技术部下属的中国科学技术信息研究所（ISTIC）支持，是中国大陆首家以中文信息资源数据库为核心业务的企业。万方数据、清华同方光盘股份有限公司（TTOD）和重庆维普资讯有限公司（VIP）被视为中国三大期刊聚合商（Atwill，2005）。尽管这些聚合商均未实现对入选的1,987种期刊的完整覆盖率，

但万方数据包含了我们分析所需的大部分相关数据。它整合了ISTIC原始文献传递系统以及由中国社会科学院领导的国家哲学社会科学文献中心的数据，从而建立了支撑其信息服务的综合数据库[2]。这种整合确保万方数据充分收录了我们分析所选期刊及其出版物的相关数据。此外，本文作者之一刘建华因其机构职位和资源权限，可以访问万方数据完整的底层数据库，这有助于为本研究有效收集数据。因此，本研究选取万方数据作为与OpenAlex进行比较的基准。

3. OpenAlex 中中国期刊的覆盖率

3.1 期刊匹配与覆盖率

为回答第一个研究问题，本节考察 GCJC 列表中 1,987 种期刊在 OpenAlex 中的覆盖率。我们首先从万方数据获取这些期刊的基本信息，然后使用印刷版国际标准连续出版物号（p-ISSN）将这些期刊数据与 OpenAlex 进行匹配[3]。国际标准连续出版物号（ISSN）是连续出版物的全球唯一标识符，有助于解决区分名称相同的期刊以及识别同一期刊标题的拼写或语言变体等问题，从而实现对期刊的准确识别。在本研究中使用 p-ISSN 是适当的，因为 GCJC 收录的期刊必须有印刷版，而中国的主要报告渠道通常只提供期刊的 p-ISSN。

在本研究纳入的 1,987 种期刊中，只有 6 种期刊在万方数据中缺少 p-ISSN。对于这些期刊，我们进行了人工核查，并确认它们未被 OpenAlex 收录。对于无法通过 p-ISSN 匹配的期刊，我们进一步使用万方提供的中文和英文期刊名称进行检索，但未能找到更多匹配项。这主要是因为 OpenAlex 以不规则的拼音形式记录中文期刊名称。我们曾使用这些期刊名称的各种形式，人工尝试检索其中 20 种期刊，但均未能在 OpenAlex 中匹配到相应期刊。因此，我们认为基于 p-ISSN 的匹配方法适用于本研究。

采用这种方法，我们发现1,987种期刊中有734种（37%）被OpenAlex覆盖。

3.2 期刊覆盖率的学科差异

Th与万方相比，OpenAlex 对中国期刊的覆盖率呈现出显著的

学科差异。表1列出了各学科的期刊数量、被 OpenAlex 收录的期刊数量及相应的期刊覆盖率。可以看出，与人文学科和社会科学相比，生命科学和自然科学领域的中国期刊被 OpenAlex 收录的比例更高。OpenAlex 期刊覆盖率最高的三个学科是生物科学（70%）、医学与健康（64%）和农业科学（59%）。

表1 OpenAlex 按学科划分的中国期刊覆盖率

学科	数量 期刊（收录于万方）	数量 OpenAlex 中的期刊 OpenAlex	期刊覆盖率 在 OpenAlex 中
生物科学	44	31	70%
医学与健康	253	162	64%
农业科学	143	85	59%
天文学与地球科学	101	47	47%
环境科学与安全科学	29	12	41%
普通自然科学	188	76	40%
工业技术	446	152	34%
语言与文字	34	11	32%
经济学	156	47	30%
历史与地理	37	9	24%
文化、科学、教育和体育	158	38	24%
航空航天与航空	25	6	24%
交通运输	30	7	23%
马克思主义、列宁主义、毛泽东思想，邓小平理论	5	1	20%
哲学与宗教	37	7	19%
Arts	32	5	16%
政治与法律	66	10	15%
综合社会科学	166	25	15%
文学	37	3	8%
总计	1,987	734	37%

注：本研究采用中国图书馆分类法区分学科，因为该分类法在中国学术期刊领域具有权威性、标准化和适用性。

4. OpenAlex 中中国期刊文章的覆盖率

4.1 文章覆盖率的学科差异

为回答第二个研究问题，本节针对本研究纳入的每种期刊，考察 OpenAlex 收录的文章数量相对于该期刊发表文章总数的情况。分析以单种期刊为单位进行，而非将所有文章作为一个整体，主要是因为数据质量与学科领域、运营机制和出版规范等期刊特定因素密切相关。在期刊层面评估文章覆盖率，能够获得更详细、更准确的见解。

对于 OpenAlex 和万方数据均收录的 734 种期刊，我们使用印刷版国际标准连续出版物号从各来源获取了完整的文章集合。根据万方数据，这些期刊共发表了 6,453,244 篇文章，我们以此作为评估 OpenAlex 覆盖率和数据质量的基准。OpenAlex 收录了这些期刊的 1,549,503 篇作品。剔除标题缺失的作品和非学术内容（如序言、目录、撤稿通知、社论）后，共保留 1,545,929 篇文章（99.8%）用于分析。

我们计算了所有可用出版年份中每种期刊在 OpenAlex 收录的文章数量与在万方数据收录的文章数量之比，发现 734 种期刊中有 91% 的文章覆盖率低于 50%。这表明，OpenAlex 对中国期刊文献的覆盖不仅限于收录的中国国内期刊数量，也体现在对这些期刊文章的覆盖程度上。

与期刊覆盖率类似，文章的覆盖率在不同学科之间也存在显著差异。如表 2 所示，与人文学科和社会科学领域的期刊相比，生命科学和自然科学领域的期刊中，收录于 OpenAlex 的文章比例往往更高。

表 2 OpenAlex 中中国期刊文章按学科划分的覆盖率

学科	文章数量 万方中的	文章数量 OpenAlex 中的	OpenAlex 中的文章覆盖率
航空航天与航空	29,321	9,857	36%
普通自然科学	609,156	249,286	35%
哲学与宗教	27,941	9,906	34%
生物科学	218,508	72,604	31%
天文学与地球科学	274,975	76,172	29%
农业科学	735,289	169,364	28%
工业技术	1,458,769	399,563	27%

交通运输	41,943	9,639	25%
环境科学与安全科学	99,771	20,830	24%
经济学	266,955	50,457	21%
医学与健康	2,097,189	390,121	19%
文化 ，《科学》教育和体育	242,721	43,689	19%
政治与法律	47,767	9,301	18%
综合社会科学	144,431	19,968	16%
马克思主义，列宁主义，毛泽东思想，邓小平理论	6,118	918	15%
语言与文字	41,453	4,687	11%
文学	13,694	1,206	10%
Arts	38,132	3,977	9%
历史与地理	59,111	4,384	9%
总计	6,453,244	1,545,929	24%

注：本研究采用中国图书馆分类法区分学科，因为该分类法在中国学术期刊领域具有权威性、标准化和相关性。

为探究中国期刊文章是否确实未被 OpenAlex 收录，我们考察了这样一种可能性：部分文章实际上已被索引，只是无法通过其印刷版国际标准连续出版物号检索到。此次分析纳入了本研究所涉及的 1,987 种期刊在万方数据中记录有 DOI 的全部文章，共计 7,182,093 个唯一 DOI。其中，192,108 篇文章无法通过印刷版国际标准连续出版物号在 OpenAlex 中检索到，但可以通过 DOI 匹配。在这些文章中，13% 在 OpenAlex 中没有来源信息，79% 已关联到某个来源但缺少国际标准连续出版物号，8% 已关联到某个来源，但其国际标准连续出版物号似乎有误。这些发现证实，通过印刷版国际标准连续出版物号检索文章的策略总体上是可行的。不过，这些发现也揭示了 OpenAlex 的一个重要弱点：大量文章缺少来源信息或包含不准确的来源元数据，这限制了期刊层面分析的可靠性。

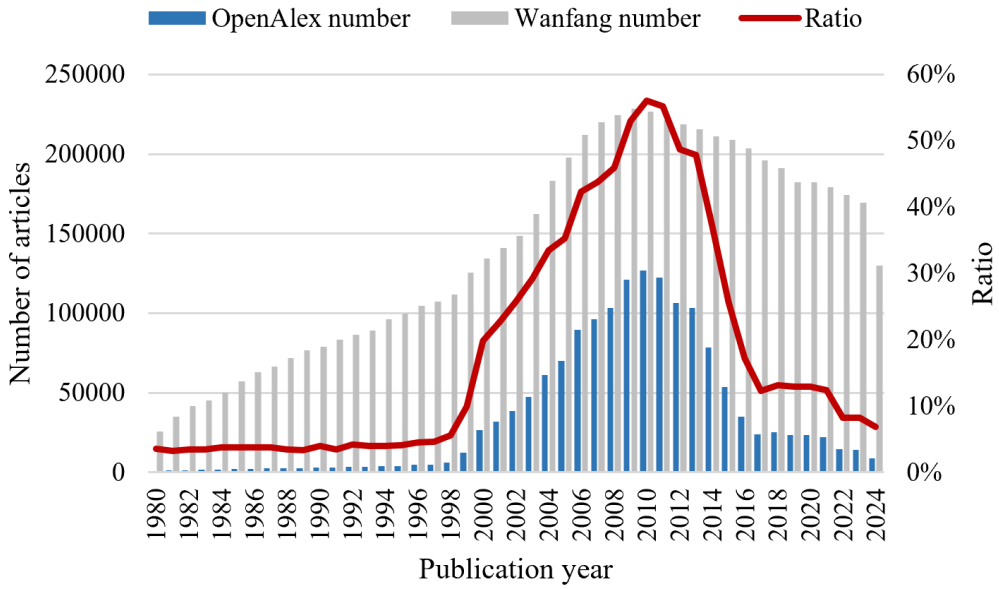
4.2 文章覆盖的时间趋势

为进一步探究文章层面覆盖，我们考察了 OpenAlex 对中国期刊文章的收录时间分布。图1展示了1980年至2024年间，734种中国期刊中 OpenAlex 覆盖的文章数量与万方数据覆盖的文章数量之比。

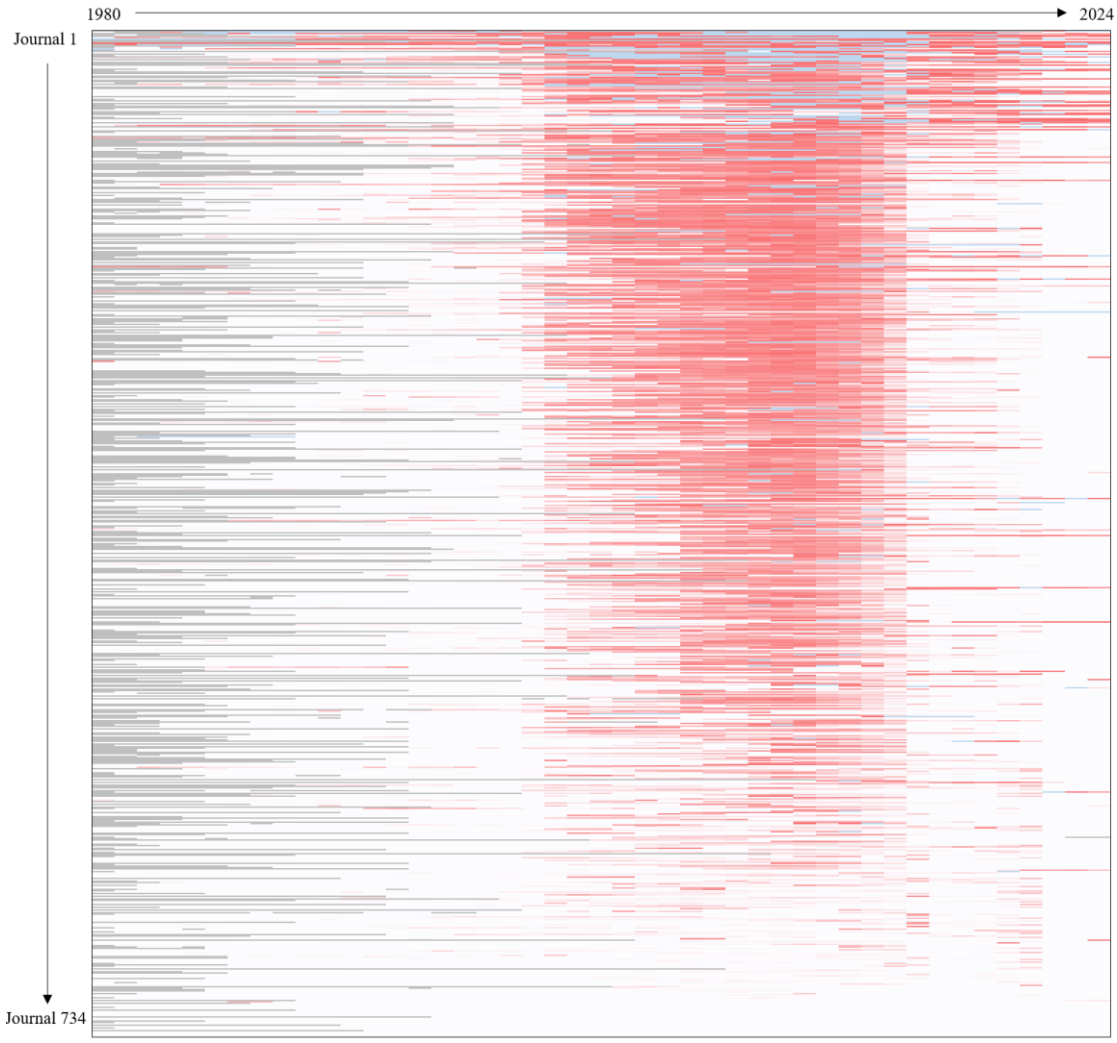
结果显示，2000年至2015年间的比率相对较高。相比之下，较早年份和最近年份的覆盖率明显较低。这一模式表明，OpenAlex 对中国期刊文献的收录在时间上并不均衡，并在2010年前后达到峰值。这与郑等人（2025）的研究结果一致。

我们还考察了所涵盖的文章是否在 OpenAlex 中包含 *mag_id*，这表明该文章此前已被纳入微软学术知识图谱（MAG），即 OpenAlex 的前身。这一点很重要，因为它可以显示 OpenAlex 是从 MAG 继承这些记录，而不是独立添加这些记录的。我们发现，超过 94% 的文章具有 *mag_id*，这一趋势与整体文章集相似。这表明，OpenAlex 索引的大多数中国期刊文章都源自 MAG 运行期间。自 MAG 关闭以及 OpenAlex 于 2021 年上线以来，中国期刊文章的收录量并未显著增加。

（a）总体趋势



（b）期刊特定趋势



注：红色越深，比例越大。比例大于 1 的单元格以蓝色表示。根据万方数据，在相应年份没有文章的单位所对应的单元格以深灰色阴影表示。

图1 1980年至2024年OpenAlex收录的734种中国期刊文章与万方数据的文章数量之比

然而，值得注意的是，图1中的比率并不能准确反映实际覆盖率。通过对期刊卷册进行全面的人工检查，并开展基于DOI的分析，我们发现OpenAlex存在若干数据质量问题，包括元数据缺失、字段标引错误和重复作品。因此，OpenAlex映射到各期刊的文章可能与相应的本地数据并不完全一致。

更具体地说，我们观察到，OpenAlex 中收录的一些文章缺少期刊信息，因此无法通过 p-ISSN 检索。此观察结果与我们的发现一致：万方数据中有 192,108 篇具有有效DOI的文章无法通过 p-ISSN

检索，但可以通过 DOI 匹配。其中一个案例是题为“岩石磁组构能否揭示应变?来自华南地块早三叠世石灰岩的案例研究”的论文[4]，该论文发表于 2009 年《地球物理学报》的第 10th 期。在 OpenAlex 中，该文章的来源字段为空，这导致覆盖率被低估。

相反，OpenAlex 有时会将名称相似的期刊混为一谈，从而导致高估。例如，《地质学报》（国际标准连续出版物号 0001-5717）和《地质学报（英文版）》（国际标准连续出版物号 1000-9515）是两种不同的期刊，但 OpenAlex 错误地将其中一种期刊的文章归到了另一种期刊名下。题为“西藏活动构造带及其对地震的控制作用”的文章便是一个例子：该文发表于后者，却在 OpenAlex 中被归到了前者名下。这一现象是导致上述比率被高估的原因之一，可能造成图1中出现蓝色单元格。

对于使用相同国际标准连续出版物号、但刊载内容不同的期刊，也存在类似问题。例如，OpenAlex 将与 https://journal.xidian.edu.cn/xdxb/CN/1001-2400/home.shtml 和 https://xadzkjdx.cn/ 这两个网站相关联的期刊视为同一个期刊实体，因此难以区分它们各自刊载的出版物。尽管这两种期刊标示使用相同的国际标准连续出版物号，但实际上彼此无关，刊载的文章也各不相同。这表明，海外注册期刊可能会使用中国国内期刊的国际标准连续出版物号，引发对其滥用甚至合法性的担忧。具体而言，这凸显了 OpenAlex 在有效区分合法期刊与欺诈性期刊方面的局限性。

此外，通过分析 OpenAlex[5]，中作品的位置信息，我们发现，除期刊网站外，OpenAlex 中的数据还来自聚合平台，例如中国的中国知网和日本的 J-GLOBAL。这些平台的数据质量直接影响 OpenAlex 检索到的数据质量。例如，J-GLOBAL 可能会分别以英文和日文收录同一篇中国期刊文章，从而导致 OpenAlex 中出现重复作品。这类不一致凸显了核验和协调开放研究信息源网络抓取数据所面临的挑战。

上述示例很可能反映了更广泛的规律，而非仅仅是我们样本中的孤立案例；它们也表明 OpenAlex 存在系统性问题，并对整个数据库产生更广泛的影响。在下一节中，我们将考察单篇文章元数据字段的完整性和准确性，以进一步评估这些问题。

5. OpenAlex 中中国期刊文章的元数据质量

To 为回答第三个研究问题，本节评估

OpenAlex 收录的中国期刊文章中，所选元数据字段的完整性和准确性。完整性是指 OpenAlex 中包含特定字段信息的文章所占比例，也就是说，该字段是否存在且已填充。这一概念不同于准确性；准确性是指 OpenAlex 中该字段内容与万方数据本地来源中可用信息相符的文章所占比例。

5.1 元数据完整性

表3展示了在 OpenAlex 中发现的中国期刊文章关键元数据字段的完整性。样本集中的所有文章均包含出版年份信息。由于本研究使用 ISSN 检索期刊文章，并排除了缺少标题的文章，因此标题和国际标准连续出版物号字段的完整性也达到100%。大多数文章还包含作者信息。然而，卷、期号和页码等书目信息往往缺失。

表3 OpenAlex 中中国期刊文章的元数据完整性

元数据字段	OpenAlex 中包含该字段的文章数量 其中元数据字段已填充	元数据完整性 在 OpenAlex 中
DOI	418,811	27%
标题	1,545,929	100%
ISSN	1,545,929	100%
Year	1,545,929	100%
卷	554,819	36%
期号	560,639	36%
起始页	544,270	35%
末页	507,780	33%
作者	1,536,991	99%
机构字符串	1,046,425	68%
机构	682,471	44%
资助项目	624	0%
参考文献	104,708	7%

样本集中文章的数字对象标识符（DOI）可用性极低[6]。一个可能的影响因素是，中国的DOI注册机构并未提供可与Crossref所提供的服务相媲美的标准化开放元数据服务（Wang等，2018）。这一局限使OpenAlex难以实现DOI信息的完整性。DOI完整性较低还意味着，在论文层面对OpenAlex和万方数据进行匹配时，不能仅依赖DOI。为评估元数据字段的准确性（第5.3节），需要采用更先进、更灵活的匹配策略（第5.2节）。

An另一个需要注意的重要问题是，OpenAlex 提供两种类型的作者

机构信息：直接从上游数据源或文章落地页和 PDF 中获取的原始机构字符串，以及 OpenAlex 根据这些原始机构字符串关联的标准化机构列表。如表3所示，OpenAlex 中包含机构信息的中国期刊文章比例相对较低。只有 68% 的文章包含原始机构字符串，只有 44% 关联了机构。这凸显了 OpenAlex 在检索、处理和统一中国期刊文章的机构信息方面存在局限性。在后续分析中，我们主要关注关联的机构，因为这是对用户而言最便捷、可直接使用的机构信息。

5.2 文章匹配

为评估 OpenAlex 中中国期刊文章相较于万方数据的元数据准确性，我们采用基于规则的方法，将从 OpenAlex 检索到的文章与万方数据中的相应记录进行匹配。该方法包含五条匹配规则，按严格程度从高到低依次应用：

(1) 按数字对象标识符匹配：此规则使用数字对象标识符，它能够唯一标识学术作品。按数字对象标识符匹配可确保最高的精确率，因此优先应用。

(2) 按标题 + 国际标准连续出版物号 + 年份 + 期号匹配：在某一期刊的特定期号中，两篇文章标题相同的情况并不常见。因此，结合标题、国际标准连续出版物号、年份和期号可实现可靠匹配。

(3) 按标题 + 国际标准连续出版物号 + 年份匹配：由于许多 OpenAlex 记录经常缺少期号信息，此规则省略期号字段，同时保留合理的精确率。只要排除征稿启事和目录等非研究类一般作品，该规则仍能确保较高质量的匹配。

(4) 按标题 + 国际标准连续出版物号匹配：此匹配规则移除了年份限制，以应对 OpenAlex 中年份元数据可能存在的错误。该规则假设期刊很少发表标题完全相同的多篇文章。

(5) 仅按标题匹配：这是限制最宽松的规则，仅根据文章标题进行匹配。当期刊曾更改国际标准连续出版物号，或 OpenAlex 错误地将文章关联到错误来源时，此规则尤其有用。不过，解读此匹配规则的结果时需要谨慎，因为它可能导致匹配结果含糊不清，尤其是对于常见或笼统的标题。

值得注意的是，随着学术出版国际化趋势日益明显，一些发表于中国期刊的文章会附有英文书目信息。因此，我们在基于标题的匹配过程中同时使用了中文和英文标题信息[7]。在此过程中，我们发现

OpenAlex 的标题字段存在格式问题，可能导致匹配失败。为解决这一问题，我们对标题进行了一系列预处理步骤。具体而言，我们进行了编码操作（例如，将“&”转换为“&”），并采用模糊匹配技术，在匹配前删除标题中的标点符号（包括空格）。

根据上述匹配规则，共成功匹配了1,394,154篇文章（90%）。根据第一条匹配规则，共有282,105篇文章（18%）通过DOI匹配。该比例相对较低，部分原因是DOI字段的覆盖率有限（27%）。另一个原因是DOI元数据中存在错误。我们对OpenAlex中有DOI但未能匹配的文章进行了随机抽样，人工核查了50篇，发现其中一些记录在OpenAlex中包含错误的DOI[8]。其余大多数未匹配的情况，是因为万方数据中的DOI由ISTIC注册，而OpenAlex通常收录的是其他注册机构提供的DOI。因此，两个来源中的DOI不兼容，导致OpenAlex中的文章无法与万方数据中的对应记录匹配。根据第二条匹配规则，另有196,519篇文章（13%）成功匹配。该比例较低，可归因于OpenAlex中期号信息有限。根据第三条匹配规则，我们又成功匹配了859,362篇文章（56%），因此这是主要的匹配方法。根据第四条和第五条匹配规则，分别匹配了9,029篇（1%）和47,139篇文章（3%）。

其余 151,775 篇未匹配文章（10%）可归因于多种因素。其中一个因素是，OpenAlex 从托管于不同国家的众多聚合平台获取数据。例如，根据位置链接，我们的样本中有超过 18,000 条文献（1%）是从日本的 J-GLOBAL 平台检索到的。因此，我们数据集中的一些文章标题既不是中文也不是英文，而是日文，且没有数字对象标识符（例如：https://openalex.org/W2742465672），这使得它们难以与万方数据匹配。

5.3 已匹配文章的元数据完整性和准确性

表4展示了使用不同规则匹配的文章中所选元数据字段的完整性和准确性。不同匹配组中特定字段的完整性和准确性各不相同。具体而言：

- 出版年份字段完整且准确，但仍存在少量标注错误。值得注意的是，规则4的准确性为0%是符合预期的，因为年份信息正确的文章本应已通过规则3完成匹配。对年份信息不正确的记录进行人工检查后发现，部分差异较小，通常是由于在线发布与正式出版之间存在延迟，或

出版与在某些平台上建立索引之间存在延迟。例如，文章 https://doi.org/10.16511/j.cnki.qhdxxb.2015.24.011 的出版年份被记录为2016年，而实际出版时间为2015年。OpenAlex以2016年对该论文建立了索引。在其他情况下，也观察到较大的差异，但难以轻易解释。例如，根据该期刊官方网站，文章“Effect of genistein on cell cycle and apoptosis in PC12 cells transfected with APP695MT gene”于2012年出版，但OpenAlex将其索引的出版年份记为2026年。

- 相比之下，期号和页码信息通常缺失，这可能是因为 OpenAlex 使用的主要中文数据源（例如中国知网英文版）不提供这些字段。不过，在这些字段存在时，其准确性通常是可以接受的。期号错误也可能源于中国国内期刊不规则的出版方式。例如，《中国科学通报》每月出版三期，偶尔还会合并两个期号。万方有时会为此类合并期号分配新的期号（例如 Z1），从而导致其与 OpenAlex 记录的期号不一致。

- 由于 OpenAlex 和万方在作者姓名索引的语言和格式标准方面存在差异，准确匹配单篇文章的作者姓名颇具挑战。因此，本研究使用作者人数作为评估准确性的替代指标。尽管 OpenAlex 为大多数中国期刊文章提供了作者列表，但作者人数的准确性相对较低。例如，根据规则3，准确率仅为15%。人工核查发现，许多依据规则3匹配的文章来自中国知网英文版平台，而且其中大多数文章只列出了第一作者。这种差异可能是由自动化数据提取所使用的解析算法存在局限性造成的。此外，即使作者人数正确，OpenAlex 中的作者顺序也经常有误，并且常常将第一作者错误地列为最后一位作者。

- 对于机构信息，我们采用与作者信息类似的方法，重点关注机构数量。机构信息的完整性低于作者信息的完整性。这与此前关于 OpenAlex 中机构元数据缺失的研究结果一致（Zhang 等人，2024年）。机构数量的准确性介于 40% 至 80% 之间，表明其可靠性相对较低。

- 最后，OpenAlex 中被引参考文献信息的完整性和准确性较低，这与此前的评估结果一致（Gusenbauer，2024）。只有少数文章包含参考文献信息，而且参考文献计数的准确性有限。这表明，OpenAlex 在中国期刊文章所引参考文献方面存在严重的数据质量问题。

表4 OpenAlex 中匹配文章的元数据完整性和准确性

	规则1：按 数字对象标识符 (282,105, 18%) 规则2：按标题+国际标准连续出版物号+年份 +期号 (196,519, 13%) 规则3：按标题+国际标准连续出版物号+年份 (859,362, 56%) 规则4：按标题+国际标准连续出版物号 (9,029, 1%) 规则5：按标题 (47,139, 3%) 总计 (1,394,154, 90%)											
	C	A	C	A	C	A	C	A	C	A	C	A
Year	100%	95%	100%	100%	100%	100%	100%	0%	100%	97%	100%	98%
期号	97%	99%	100%	100%	1%	76%	94%	97%	17%	93%	35%	99%
起始页	94%	82%	98%	93%	1%	92%	85%	85%	16%	90%	34%	87%
末页	91%	68%	93%	76%	1%	48%	80%	81%	16%	88%	35%	72%
作者计数	99%	83%	99%	89%	100%	15%	99%	74%	100%	40%	100%	40%
机构计数	33%	47%	22%	53%	56%	62%	13%	55%	52%	75%	46%	59%
参考文献计数	18%	5%	12%	6%	0%	11%	39%	8%	1%	6%	6%	5%

注：“C”表示完整性，“A”表示准确性。

除上述字段外，我们还发现 OpenAlex 中的语言字段存在显著的不可靠性。样本中的所有文章均来自以中文发表文章的期刊。然而，根据对样本中在 OpenAlex 里包含语言信息的 1,537,378 篇文章进行的分析，只有 5% 被标记为中文，92% 被标记为英语，另有少数被标记为韩语、越南语或丹麦语等语言。此前的研究已证实，OpenAlex 的语言元数据不准确，且低估了俄语和中文等语言的占比（Céspedes 等，2025）；OpenAlex（2025）也已正式承认这一点。此问题后来已得到改善（Demes，2025），但它引发了人们对开展语言多样性及类似主题分析的担忧。

5.4 匹配文章的元数据完整性和准确性方面的学科差异

我们进一步考察了不同学科在元数据完整性和准确性方面的差异。如图2所示，生命科学和自然科学领域的期刊在期号、页码范围等元数据字段方面表现出更高的完整性。这一模式反映出这些领域的出版实践具有更高的标准化和国际化程度。相比之下，人文学科和社会科学领域的期刊往往表现出更高的作者和机构计数准确性。

这可能是因为这些学科的合作作者团队规模较小（Lewis 等，2012；Melin, 2000），从而降低了作者和隶属机构数据出错的可能性。

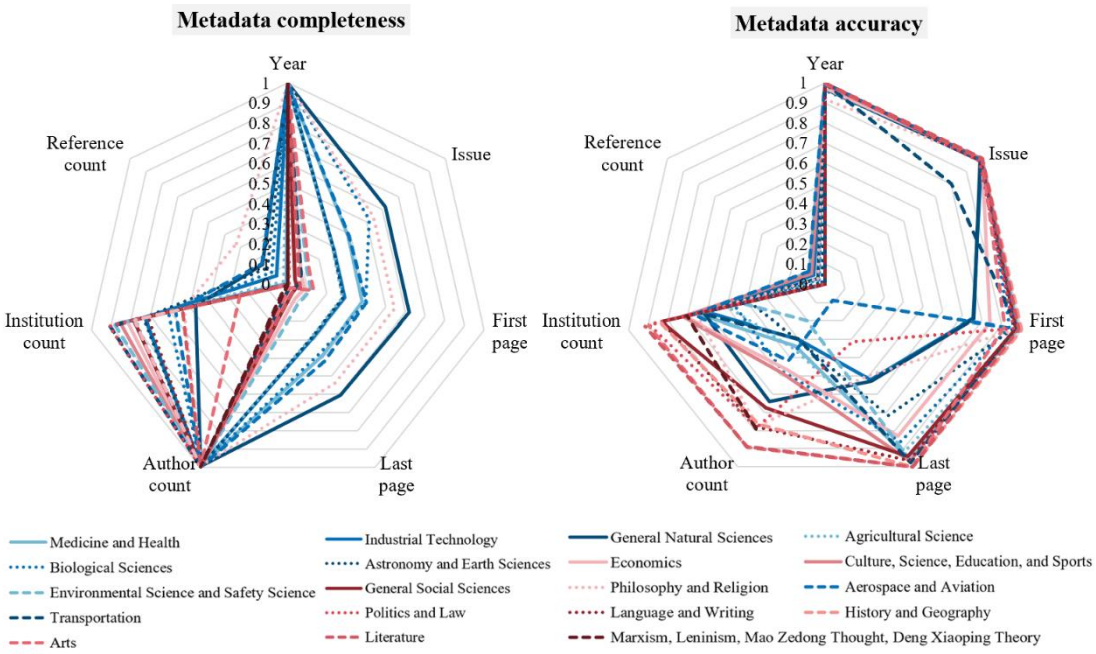


图2 OpenAlex 中匹配文章的元数据完整性和准确性（按学科划分）　　e

6. 结论与讨论

通过比较开放数据源 OpenAlex 与本地数据提供方万方数据所收录的期刊和文章元数据，本研究发现，OpenAlex 目前仅收录略多于三分之一的中国核心期刊，来自这些期刊的文章占比则更低，且不同学科和出版年份之间存在显著差异。在单篇文章层面，我们的分析进一步揭示了普遍存在的元数据质量问题，例如书目信息缺失、作者机构信息和引文数据不完整。此外，页码、作者、机构和引文等细节的准确性也较低。尽管与其他专有数据源相比，OpenAlex 的覆盖范围更广、包容性更强，但本研究结果表明，在提高中国期刊文章的覆盖率和相关元数据的质量方面，仍有很大的改进空间。这些发现凸显了 OpenAlex 在全面呈现全球学术交流、以及用于科研评价和相关活动时确保各国公平性方面的现有局限。同时，中国学术出版体系的独特特点也使实现这些改进颇具挑战。尽管如此，这些局限与挑战也带来了

未来改进的机会，因此我们针对中国出版体系中的关键参与者以及 OpenAlex 等开放研究信息源提出了若干建议。

首先，OpenAlex 等具有全球视野的开放数据源应加强其从多样化上游数据源聚合和集成数据的策略。我们的分析显示，OpenAlex 中约三分之二的中文期刊和文章数据来自中国知网等本地数据聚合平台，其余则来源于包括 Crossref、期刊网站和外国数据聚合平台在内的其他渠道。尽管这种多样性能够提高覆盖率和包容性，但也会带来格式不一致、重复记录等风险。为提升数据的可用性和可信度，建议 OpenAlex 加强对上游数据源可靠性的评估，并监测中国知网等平台是否已被完整索引，还是仅进行了部分集成，尤其是考虑到近期中国知网收录文章数量有所下降。此外，还建议制定清晰、透明的数据集成策略。这包括明确处理上游数据源提供的冲突信息的协议，以及在上游数据源未提供某些字段时的处理方式。如果某个字段为空，OpenAlex 能够说明这是因为无法获取相关数据，还是因为原始作品本身不包含此类信息，也将十分有益。例如，对于作者所属机构和被引参考文献，用户目前无法判断某项作品缺少这些信息，还是 OpenAlex 无法获取这些信息。另一项建议是，不仅应使原始数据的来源透明可见，还应在 OpenAlex 创建的清理统一版本旁提供对原始数据本身的访问权限。纳入来源信息将使用户能够更容易地追溯问题，从而推动社区更快地进行修正，也有助于用户更好地评估数据是否适合其用途。

其次，中国的 DOI 注册实践仍需进一步发展。目前，全球共有 12 家 DOI 注册机构，其中 ISTIC&Wanfang 和中国知网是为中文期刊提供 DOI 服务的两家机构[9]。起初，ISTIC&Wanfang 提供免费注册服务。后来，中国知网也参与其中，接手了部分注册工作，但其期刊文章和学位论文的注册数量仍显著低于 ISTIC&Wanfang[10]。每篇文章都可以通过这两家机构之一获取 DOI。然而，中国的 DOI 注册体系尚未完全成熟，也未充分发挥 DOI 作为唯一标识符的作用。一方面，中国的 DOI 注册流程缺乏全面的自动化规范。许多中国期刊在发表文章时并未启用 DOI，而是采用先预分配、后注册并在后续阶段启用的机制。另一方面，在中国注册 DOI 所需的元数据仅限于国家标准规定的一些必要字段。中国 DOI 注册机构填写的元数据字段远不如 Crossref[11], 丰富，例如关键词、摘要和被引参考文献等详细信息均未纳入

此外，中国的 DOI 代理机构目前尚未像 Crossref 和 DataCite 那样，通过提供数据快照或应用程序接口来实现便捷、开放的数据访问。这些限制目前给将中国本地数据整合到 OpenAlex 等国际数据源带来了重大挑战，需要多个利益相关者协同努力。

第三，万方数据等本地数据聚合商，尤其是提供标识符注册服务的聚合商，应成为原始数据生产者与下游用户及平台之间可靠的中介。应鼓励国内提供方加强协调，避免出现不一致的情况，例如期刊同时在 ISTIC 和中国知网注册 DOI，这会损害标识符的唯一性，并使下游数据集成更加复杂。能够兼顾区域和学科差异的国际标准，有助于本地数据聚合商向 OpenAlex 等下游平台提供质量更高、更一致且更易于集成的数据。

最后，作为出版元数据的原始来源，期刊和出版商决定了数据质量的上限。除了传播高质量研究成果，他们还必须维护学术出版的诚信与公信力。然而，我们的分析发现了一些不良做法，例如滥用 ISSN，以及为同一篇文章添加多个 DOI，这表明期刊管理实践存在不足，最终影响元数据质量。这一挑战在中国尤为突出，因为大多数期刊由大量机构管理和出版。据统计，中国的 5,211 种科学期刊归 3,217 家机构所有，其中 76.87% 的机构仅拥有一种期刊；这些期刊由 4,470 家机构出版，其中 96.04% 的机构仅出版一种期刊（中国科协学会服务中心，2024年）。因此，加强中文出版流程的标准化和国际互操作性，对于提升元数据质量、增强本地中文期刊在全球学术生态系统中的可见度也至关重要。

本研究仅聚焦于《中国核心期刊要目总览》（*2023*年）》所列期刊，以及这些期刊及其文章在 OpenAlex 中的覆盖情况。因此，研究结果并不代表中国所有类型文学的覆盖情况，也不能推广至其他国家或地区。尽管如此，鉴于所选期刊目录规模庞大且具有权威性，同时中国在全球出版活动中发挥着重要作用，本研究提供了有价值的见解，并提出了一种可拓展至其他国家和地区的方法。此外，必须认识到，OpenAlex 等数据源在不断演变。包括本研究在内的元数据质量调查，虽然能提供分析时点上的宝贵见解，但随着新数据的加入和现有记录的完善，其相关性可能会减弱。因此，读者应结合这些数据源的动态特性及其持续发展来解读研究结果。

致谢

作者谨此感谢国家自然科学基金委员会（项目编号：72374160、L2424104）以及武汉大学国家图书情报科学实验室中心的支持。

资助信息

国家自然科学基金委员会，项目编号：72374160、L2424104。

尾注

[1] http://hxqk.lib.pku.edu.cn/

[2] 根据官方网站统计，TTOD的中国知网（CNKI）收录了1,979种GCJC期刊，维普数据收录了其中的1,968种。

[3] 本研究所用的 OpenAlex 数据于 2024年11月通过应用程序接口获取。

[4] 尽管我们的样本由中国期刊及其发表的文章组成，但由于出版业的国际化，部分期刊会提供英文书目信息。

[5] OpenAlex 为每篇论文提供落地_页_网址和*PDF*_网址字段，反映了论文可通过哪个来源访问。

[6] 我们在此报告一项用于比较的基准数据：根据通过我们的匹配流程识别出的万方记录（详情见第 5.2 节），72% 的文章关联有数字对象标识符。

[7] 在本研究中成功匹配的文章里，根据万方数据，近 90% 提供了英文标题。然而，OpenAlex 根据其来源平台提供的信息，每篇文章仅提供一种语言的标题。为此，我们使用中文和英文文章标题，将 OpenAlex 的文章标题与万方数据进行匹配。如果一篇文章在任一语言下成功匹配，我们便视该匹配有效。

[8] 例如，与链接“https://openalex.org/W4205460699”对应的文章，其数字对象标识符应为“10.37188/CO.2019-0255”。

然而，OpenAlex 提供的数字对象标识符是“10.37188/co.2020-0255”。

[9] https://www-old.doi.org/registration_agencies.html

[10] 截至2025年4月3日，ISTIC&Wanfang 收录的期刊文章 DOI 总数为36,329,523；截至2025年4月10日，中国知网收录的 DOI 总数为19,087,061。

[11] ISTIC 提供的元数据包括期刊名称、卷、期号、国际标准连续出版物号、出版年份、起始页、末页、DOI、作者、机构和文章标题。相比之下，Crossref 提供了更多信息，包括摘要、参考文献、出版商、资助方等（https://www.crossref.org/documentation/）。

参考文献

Alonso-Alvarez, P., & Van Eck, N. J. (2025). 非洲研究出版物在 OpenAlex 中的覆盖率、元数据完整性和准确性：比较分析。定量科学研究。https://doi.org/10.1162/QSS.a.396

Asubiaro, T. V., & Onaolapo, S. (2023). 非洲期刊在科学网、Scopus 和 Crossref 中的覆盖范围比较研究。《信息科学与技术协会会刊》，*74*(7), 745-758. https://doi.org/10.1002/asi.24758 Atwill, Y. Y. (2005). 中国的电子期刊：技术与馆藏问题。学术图书馆学杂志, *31*(6), 598-604. https://doi.org/10.1016/j.acalib.2005.09.001 Babini, D., Garcia, A. B., Costas, R., et al. (2024). 不仅开放，而且多元与包容：迈向去中心化与联邦式研究信息源。*Leiden Madtrics*. https://doi.org/10.59350/gmrzb-e2p83 开放研究信息巴塞罗那宣言，Kramer, B., Neylon, C., & Waltman, L. (2024). 开放研究信息巴塞罗那宣言。*Zenodo*. https://doi.org/10.5281/zenodo.10958522 Brainard, J., & Normile, D. (2022). 中国在被引次数最多的论文方面跃居首位。《科学》, *377*(6608), 799. https://doi.org/10.1126/science.ade4585 Cao, Z., Zhang, L., Huang, Y. et al. (2025). 学术界如何引用开放研究信息数据源？基于 OpenAlex 和 Microsoft Academic系列的综述研究。《科学计量学》. https://doi.org/10.1007/s11192-025- 05347-6 Céspedes, L., Kozlowski, D., Pradier, C., et al. (2025). 评估语言

覆盖率：对 OpenAlex 元数据准确性和完整性的评估。《信息科学与技术协会会刊》, *76*(6), 884- 895. https://doi.org/10.1002/asi.24979 Delgado-Quiros, L., & Ortega, J. L. (2024). 八个免费开放学术数据库中出版元数据的完整程度。定量科学研究,*5*(1), 31-49. https://doi.org/10.1162/qss_a_00286Demes, K. (2025). 更好地检测 OpenAlex 中的语言，也能更好地开展合作。*OpenAlex* 博客. https://blog.openalex.org/a-better-way-to-detect-language-in-openalex-and-a-better-way-to-collaborate/ Gusenbauer, M. (2024). 超越谷歌学术、Scopus 和科学网：对 59 个数据库引文索引的后向和前向引文覆盖率进行评估。研究综合方法, *15*(5): 802-817. https://doi.org/10.1002/jrsm.1729 Huang, Y., Li, R., Zhang, L., et al. (2021). 中国期刊评价体系的全面分析。定量科学研究，*2*(1), 300-326. https://doi.org/10.1162/qss_a_00103Lewis, J. M., Ross, S., & Holden, T. (2012). 学术合作的方式与原因：学科差异及政策影响。高等教育,*64*, 693-708. https://doi.org/10.1007/s10734-012-9521-8 Maddi, A., Maisonobe, M., & Boukacem-Zeghmouri, C. (2025). 开放获取期刊的地理与学科覆盖范围：OpenAlex、Scopus 和 WoS。*PLOS One*, *20*(4): e0320347. https://doi.org/10.1371/journal.pone.0320347 Melin, G. (2000). 实用主义与自组织：个体层面的科研合作。研究政策, *29*(1), 31-40. https://doi.org/10.1016/S0048- 7333(99)00031-1 OpenAlex (2025). *OpenAlex* 技术文档：*Work* 对象 - 语言. 检索于 2025-02-28，网址：https://docs.openalex.org/api-entities/works/work- object#language Ortega, J. L., & Delgado-Quiros, L. (2024). 七个主要学术数据库中撤稿文献的收录情况：Dimensions、OpenAlex、PubMed、Scilit、Scopus、The Lens 和科学网的覆盖范围比较。《科学计量学》,*129*(7), 3769-3785. https://doi.org/10.1007/s11192-024-05034-y Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex：全面开放的学术作品、作者、出版载体、机构和概念索引。*arXiv*. https://doi.org/10.48550/arXiv.2205.01833 中国科协学会服务中心。（2024）。中国科技期刊发展蓝皮书。北京：科学出版社。

索邦大学。（2023年）。索邦大学取消订阅《科学》网络版 *f*
《科学》。检索于2025-02-05，网址：https://www.sorbonne-universite.fr/en/news/sorbonne-university-unsubscribes-web-science Van Eck, N. J..、Waltman, L. 和 Neijssel, M.（2024年）。CWTS 莱顿大学排名开放版2024年上线。*Leiden Madtrics*。https://doi.org/10.59350/r512t-r8h93 Veugelers, R.（2012年）。亚洲在科学领域的崛起。*Breugel*。https://www.bruegel.org/blog-post/rise-asia-science Wang, W.、Deng, L.、You, B. 等（2018年）。数字对象标识符及其在中国核心学术期刊中的应用：中国视角。《学术出版》，*31*(2)，149-154。https://doi.org/10.1002/leap.1137 Xie, Y.、Zhang, C. 和 Lai, Q.（2014年）。中国崛起为科学技术领域的主要贡献者。《美国国家科学院院刊》，*111*(26)，9437-9442。https://doi.org/10.1073/pnas.1407709111 Zhang, L.、Cao, Z.、Shang, Y. 等（2024年）。OpenAlex中缺失的机构：可能的原因、影响及解决方案。《科学计量学》，*129*(10)，5869-5891。https://doi.org/10.1007/s11192-023-04923-y Zheng, M.、Miao, L.、Bu, Y. 等（2025年）。了解OpenAlex覆盖率方面的差异：中国案例。《信息科学与技术协会会刊》，*76*(11)，1591-1601。https://doi.org/10.1002/asi.70013