提升交互式信息检索中的数据可复用性：来自社区的见解

Tianji Jiang*ti anji008@ucla.edu 加利福尼亚大学洛杉矶分校 洛杉矶，加利福尼亚州，美国

Wenqi Li 北京大学 北京，中国

Jiqun Liu 俄克拉荷马大学 俄克拉荷马城，俄克拉荷马州，美国

摘要

在本研究中，我们对21位交互式信息检索研究人员进行了半结构化访谈，以调查他们的数据再利用实践。本研究旨在通过探索交互式信息检索研究人员在数据再利用方面的信息获取行为，拓展现有研究发现。我们确定了交互式信息检索研究人员在评估数据可再利用性时所需的共享数据特征信息，以及他们通常会查阅哪些来源来获取这些信息。我们认为，这项工作是揭示交互式信息检索研究人员数据再利用实践、并探明该领域群体应采取哪些措施来促进数据再利用的初步尝试。我们希望本研究及未来的研究能够激励更多人投身于持续推进相关标准、基础设施和政策设计的工作，并在该领域培育可持续的数据共享与再利用文化。

CCS 概念

• **通用与参考** → 通用文献；• **信息系统** → **用户与交互式检索**。

关键词

交互式信息检索、科研数据管理、数据可再利用性

ACM 参考格式：

Tianji Jiang、Wenqi Li 和 Jiqun Liu。2025。提升交互式信息检索中的数据可再利用性：来自社区的见解。载于*ACM SIGIR* 人类信息交互与检索会议（*CHIIR*）论文集。ACM，美国纽约州纽约市，5 页。https://doi.org/XXXXXXX。XXXXXXX

1 引言

信息检索（IR）长期以来一直有重复利用共享数据集，并生成供他人再利用的基于系统的实验数据的传统 [11, 13]。随着信息检索研究不断发展，对大规模高质量数据的需求持续增长 [5]。共享和再利用研究数据能够提升用户研究的成本效益，也有助于开展可复现性实验；这两者对于推动信息检索研究和评估技术的发展都至关重要。在信息检索的研究范式根本上立足于比较与评估，使用通用数据集也能确保研究的可靠性、标准化、可比性和可复现性 [5]。这些益处并不局限于以系统为中心的评估。在交互式信息检索中，研究重点是用户与信息检索系统交互过程中的用户行为和任务 [7, 9]，数据共享与再利用同样大有裨益，正如其在更广泛的信息检索领域中所发挥的作用。

IIR研究社区已经认识到共享和再利用研究资源的重要性，尤其是研究数据，同时也包括实验设计和基础设施。一项近期的实证研究发现，尽管仍存在许多实际挑战，但IIR社区大力倡导数据再利用 [?]。对2016年至2022年间发表的CHIIR论文进行的另一项实证分析显示，14.4%的论文再利用了现有数据，这表明数据和资源再利用在IIR研究中已不再罕见 [3]。鼓励和促进此类实践一直是社区积极关注的议题。BIIRRR研讨会系列 [1, 2, 4]，作为最受认可的努力，考察了资源再利用的障碍，并确定了推动再利用所必需的重点领域。在这些讨论的基础上，同一社区将其见解提炼为一份宣言，并于CHIIR 2021发表；该宣言强调了记录和归档研究、支持未来再利用的原则、挑战和要求 [5]。此外，也有一些关于促进或规范资源共享与再利用的实际方法的探索 [6]。例如，Liu [8] 开发了受克兰菲尔德启发的方法，用于评估IIR资源的可重用性并改进报告实践。在我们此前对交互式信息检索研究人员数据再利用实践的研究基础上 [?]，，本研究进一步考察其数据再利用经历中的一个具体方面：他们在决定数据是否可重用时认为研究数据的哪些特征十分重要，以及他们在实践中如何获取有关这些特征的信息。更深入地理解这些要求和策略，将有助于阐明“可重用性”对于IIR研究数据意味着什么，并为知识基础设施的设计提供参考，从而更好地支持此类数据的整理与共享。此外，我们预计，本研究的发现将有助于当前在IIR社区内推动研究资源更顺畅、更有效流通的生态系统的持续建设。

2 研究问题与设计

为加深对 IIR 研究中如何评估数据可再利用性的理解，本研究探讨了两个关键问题：

- **研究问题 1：** 在判定研究数据的可再利用性时，需要考量研究数据的哪些特征？
- **研究问题 2：** 研究人员如何收集评估数据可再利用性所需的信息？

特此授权免费复制本作品的全部或部分内容，用于个人或课堂使用，但前提是不得为牟利或商业利益而复制或分发，且复制件首页须附有本声明及完整引文。对于本作品中由作者以外的其他方拥有版权的组成部分，必须尊重其版权。经注明出处后，允许摘录。除上述情况外，如需复制、再版、发布到服务器或重新分发至列表，必须事先获得具体许可和/或支付费用。请发送邮件至 permissions@acm.org 申请权限。*CHIIR*，美国华盛顿州西雅图© 2025 版权所有归所有者/作者所有。出版权授权给 ACM。ACM ISBN 978-1-4503-XXXX-X https://doi.org/XXXXXXX.XXXXXXX

本研究重新使用了我们先前工作中收集的访谈数据 [?], 。先前工作对 21 位交互式信息检索研究人员进行了深入的半结构化访谈，参与者是通过理论抽样选出的。参与者的研究兴趣各不相同，但均至少拥有三年 IIR研究经验，以确保他们充分了解该领域以及严谨研究的构成要素。样本包括来自亚洲、澳大利亚、欧洲和北美洲不同地区机构的八位教授、一位博士后研究人员和十二位高年级博士生。尽管样本规模相对较小，但参与者的研究经历、职业阶段和机构背景涵盖范围广泛。访谈以线上方式进行。为确保访谈期间沟通顺畅，每次访谈均使用参与者首选的语言（英语或中文）进行。

在分析阶段，作者采用了主题编码。他们反复审阅了全部21份访谈文字稿，以找出与研究问题相关的片段。随后，通过持续比较，将概念上相似的片段归为一组，并整理成主题类别。之后，他们进一步完善并综合这些主题，得出直接回应研究问题的关键发现。

3 研究发现

所有参与者都不认为数据可再利用性是数据集本身固有的属性，而是研究人员接触数据后形成的一种评估。他们反复强调，关于数据再利用的决定取决于他们如何看待和解读可获取的信息，这些信息来自数据集本身、随数据附带的文档、相关文献、社区规范以及他们自身的专业知识。参与者并未将可重用性视为固定属性，而是将其描述为一个解读和评估的过程。在这一过程中，他们会根据可获取的信息和自身知识，收集有关影响数据集潜在“再利用适用性”的各种因素的信息（如图 1 所示）。一位参与者这样描述：“一旦我拿到一个（共享的）数据集，就需要根据自己的经验，更深入地了解和熟悉这批数据，而不是只把它看成一串数字。”（*P21*）。

表1：研究发现概要

用于评估可重用性的数据特征 可重用性	数据特征的信息来源 特征
• 数据生产的背景和方法 数据生产 • 数据文档 • 创建者可信度和 社区验证 • 法律与伦理约束	• 相关学术文献 对共享数据 • 个人和职业 网络 • 社区场所和 共享经验

本文回应了两个研究问题，作者将分析重点放在研究参与者评估共享数据可重用性时所寻求的信息，以及他们获取这些信息的来源上。

3.1 交互式信息检索研究人员评估数据可再利用性时考虑的特征

参与者在判断能否以及在多大程度上将其他研究人员收集的数据集用于自己的工作时，会考虑多种因素。这些因素包括数据生产的背景和方法、数据文档、数据创建者的可信度，以及法律与伦理约束。

3.1.1 数据生产的背景和方法。 研究参与者中几乎一致的观点（21 人中有 20 人）是，在评估共享数据集的可重用性时，了解数据收集的背景和方法至关重要。参与者强调，有必要了解这些数据最初是如何、为何以及在何种条件下收集的。此类信息对于评估数据集的可信度是必要的，包括判断数据集是否通过严谨且符合研究目的的研究设计生成、数据收集过程中是否存在可能引入偏差或其他质量问题的因素，以及这些数据能否恰当地支持其新的研究目的。正如 P21 所说：“如果我不知道数据是如何收集和清理的，我就无法信任这些数据。”

对于主要使用源自实验室研究的数据集的研究参与者而言，这种对方法学清晰度的需求尤为强烈；这些实验室研究考察真实用户如何在给定主题上执行特定任务类型。对这些研究人员来说，行为数据是在特定背景下构建的，而数据生产的背景与数据本身不可分割。这包括实验设计、研究前后与参与者的互动、研究期间要求用户完成的任务，以及研究人员如何指导他们执行这些任务等因素。这些要素并非无关紧要的细节；它们界定了所研究的现象。另一位参与者（P20）解释道：“用户行为取决于特定的界面。如果不了解背景，数据就会失去意义...这些数据是为他们的研究目的收集的。如果任务不一致， 它就无法直接服务于我的研究目的。”

3.1.2 数据文档。几乎所有参与者都强调，完善的文档对于评估共享数据的可重用性至关重要。即使数据已公开，许多人也指出，由于缺乏对数据集实际包含内容及其旨在说明什么的清晰描述，他们仍无法重用这些数据。当无法直接从数据集中轻易推断出这类信息时（大规模数据集或包含许多变量的数据集往往如此），参与者希望能从随附的

文档。参与者将文档描述为任何形式的附带材料，包括 README 文件、元数据模式、数据描述符或方法附录，这些材料提供解释性指引。有趣的是，参与者对他们期望在文档中看到的内容各有侧重，其中一些内容见表 2。

参与者对于数据文档应包含哪些内容的期望，可归纳为实现三个相互关联的目的。首先是认识论透明度，指的是

揭示数据是如何生成的，以及生成时所处的条件。第二，实用可用性，即提供足够的信息，让他人了解如何使用该数据集，以便重现分析或有效地重复使用数据

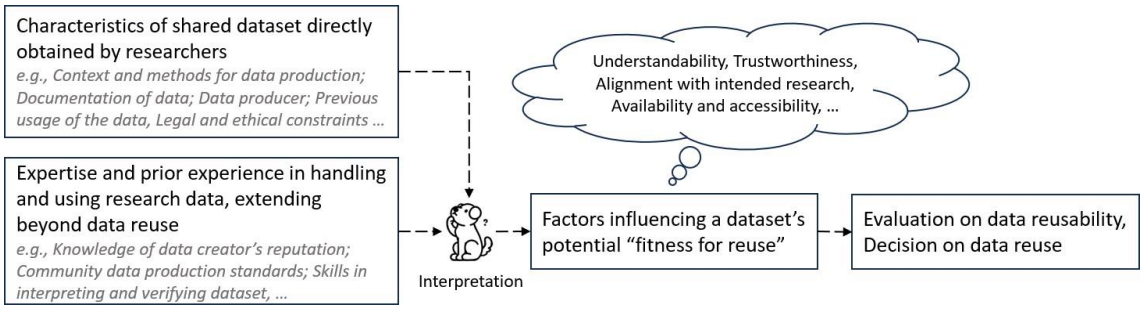


图 1：评估共享数据可重用性的流程

表 2：研究参与者提及的数据文档所需的描述

参与者	对数据文档的描述
P1	至少对我而言，解释数据结构的文档是非常重要。在研究中，你需要了解这些数据结构提供了哪些信息、没有提供哪些信息，以及如何恰当地使用它。这一点极其重要。
P4	在文档中，通常会说明如何使用这些数据。了解这一点后，你也就能对数据集形成整体认识从统计视角来看……如果你的研究更多关注简短的日常对话，这类信息可以帮助你判断该数据集是否符合你的需求。所有这些印象都可以大致从其文档中获得。
P7	在数据文档中，我们希望准确了解该数据集是如何开发的、提供了哪些说明，以及有哪些评估人员所使用的标准……在重新使用之前，这些细节非常重要重新使用它。
P21	理想情况下，我们会有 README 文件。我很喜欢直接把数据放进 CSV 文件……并附上一份文档，其中包含论文链接，说明人们应如何引用它，以及对每一列的说明。

不同的目的。第三，社会信任信号，即展现数据创建者的专业性、能力和问责意识，以及最初收集这些数据的研究过程。这最后一个方面构成了评估数据可再利用性时另一项至关重要的考量，正如下文将讨论的那样。总体而言，研究参与者认为，能够实现上述一个或多个目的的高质量数据文档，是数据集具有可再利用性的明确标志。

3.1.3 创建者的可信度与社区验证。 研究参与者普遍指出，数据创建者的可信度，以及数据集在多大程度上经社区使用得到验证，是影响数据可再利用性的另一项反复出现的因素。研究人员很少会“照单全收”某个数据集。相反，他们会对数据集作出判断，主要关注信任问题，依据的是原始数据或随附文档中未包含的信息，例如谁制作了该数据集、谁曾再利用过该数据集、该数据集在哪里共享，以及它在其领域内的传播范围有多广。

研究发现，对共享数据的信任往往依托于声誉。最初由知名研究人员、信誉良好的机构或研究声誉卓著的个人（尤其是熟识的人）收集的数据，被认为比来自知名度较低来源的数据更具可复用性。相比之下，数据集来源不清或未被识别的数据往往会受到怀疑，即使其文档或相关文献在技术层面上似乎无可挑剔。正如 P14 所说：“权威性与原始数据收集者的声誉自然会让人觉得，他们的数据质量更高。声誉更好的团队可能会被认为收集的数据更加准确或可靠。”有评估导向的信息检索经验的研究参与者

研究还将可信度描述为一种机构担保。由文本检索会议或大型企业研究团队等长期开展的项目发布的数据，被认为天然更加可靠，当然也更容易重新使用，因为它们经过了多轮审核、标注和再利用。正如 P19 所指出的，文本检索会议数据集的优势不仅在于其设计，也在于反复验证：“它们经历了多轮迭代，也有许多人审阅过……我们有相关性评判员，也有查询——这些都是参考依据。”另一个值得注意的发现是，主要使用源自实验室研究的数据集的参与者，会将可信度与对创建者的熟悉程度联系起来，而非机构担保。

对他们而言，对数据的信任建立在人际关系之上。由于这些研究人员在数据收集时往往遵循不同的规程，而且社区内没有统一的标准化方法来开展此类研究，他们只能凭借对数据创建者的个人了解，判断其方法学能力和伦理可靠性。一位研究参与者（P11）指出：“你得了解开展这项研究的团队。如果我认识他们，就能信任他们的数据；否则，我宁愿自己收集。”

3.1.4 法律与伦理约束。 几位研究参与者还提出，法律与伦理约束是共享数据的另一项令他们关注的关键特征，有时甚至决定了数据是否完全可能被重新使用。对于许多参与者而言，尤其是那些有处理人类受试者或用户行为数据经验的参与者，知情同意、隐私和数据保护问题优先于对分析潜力的任何考量。在审查共享数据集时，他们首先查找并密切关注的信息之一，就是围绕其使用的法律与伦理约束。正如一位参与者（P2）简洁地说道：“我首先会问：我是否能够在符合伦理的前提下访问这个 数据集，以及如何恰当地使用它？这是在开展其他任何工作 之前必须跨过的门槛。”

此外，研究参与者还提到，践行伦理责任本身就是科研专业性的标志，这

反过来也有助于提升共享数据的可重用性认知。透明地处理知情同意、访问权和允许用途，不仅能保护最初提供数据的个人，也体现了数据创建者的诚信，从而增强后续使用者的信任与可信度。正如研究参与者所说：

> “只要数据集清楚说明了适用哪些权限和许可协议，我就有信心使用它。”（*P17*）
>
> “完善的伦理文档，例如说明用户是否给予了知情同意，以及如何合法使用数据，表明创建者承担了责任……也说明创建者应该是一位优秀的研究人员。”（*P11*）

3.2 交互式信息检索研究人员获取信息以评估数据可再利用性的来源

由于缺少专门用于存放行为数据或用户研究数据的存储库，交互式信息检索研究人员必须从各种来源拼凑有关共享数据集的信息，才能做出明智的判断。正如P8所说：“并没有一个地方可以浏览用户研究数据集及其相关信息……”除了检查数据集本身及其随附文档外，参与者还提到了其他几种信息来源，包括与共享数据相关的学术文献、个人和职业网络，以及社区场所和共享经验。

3.2.1 与共享数据相关的学术文献。 大多数参与者（n=20）表示，学术出版物是他们了解现有数据集的主要信息来源，往往也是唯一的信息来源。对于任何审查他们自己复用这些数据的研究的人来说，这一来源同样易于获取，因此参与者对自己工作的透明度更有信心，因为他们可以依据审查者能够获取的相同信息来评估数据。参与者还指出，原始数据创建者通常会在学术论文中最详细地说明其数据生产过程，但其目的未必是促进复用，而是确保研究设计的报告透明。通过查阅方法部分、附录和补充材料，参与者推断出数据是如何收集的、参与者执行了哪些任务，以及研究设计是否符合自己的研究目的。正如P11 所述，“主要是通过阅读论文……没有其他渠道……我会了解他们如何设计任务、招募了哪类参与者，然后判断是否适合。”

3.2.2 个人和职业网络。 个人网络和合作在了解数据特征、评估其可重用性方面发挥着至关重要的作用。研究人员常常依靠同事、导师或会议人脉来了解共享数据集，或确认其可信度。P15 描述了她获取共享数据集信息的经历：“这往往要通过人来了解。如果我认识实验室或作者，就会放心使用他们的数据……如果有些事情不确定，我可以联系数据创建者。”

3.2.3 社区场所与共享经验。CHIIR 和 SIGIR 等社区场所也充当了间接机制，帮助人们发现并获取有关他人共享数据的信息。尤其是，这些会议设立了资源共享专题分会，鼓励研究人员不仅分享自己的数据集和经验，也了解在哪里可以找到共享数据及相关见解。通过会议演讲和非正式讨论，研究人员常常了解到他人制作的数据集，并对其内容和背景获得第一手了解。正如P20 所描述的：“通常在会议上，有人会介绍一个数据集。如果它很有意思，我可能会之后联系他们，看看是否能获取该数据集，并提出我仍有的（进一步）问题，了解相关情况。”

这些来源——社区场所和共享经验、已发表文献以及专业网络——表明，在 IIR 领域，了解和评估共享数据集与社区互动密切相关，而非依赖正式存储库。关于数据集的知识通过会议演讲、非正式交谈和学术出版物传播；研究人员交流的不仅是数据，也包括解读数据所需的情境理解。由此，社区本身发挥着一个活态存储库的作用，它由人员、论文和实践构成网络，数据相关知识经由这一网络不断产生、验证和传递。认识到这一模式，凸显了设计知识基础设施的必要性：应立足于现有的社区互动模式，让这些非正式交流更加可见、持久且易于获取，从而支持 IIR 领域负责任的数据共享与再利用。

4 讨论与结论

与所有科学探索一样，IIR研究的进展取决于能否在以往的思想、方法和资源基础上继续推进[1, 10]。人们已开展实践并进行讨论，旨在建立共享的学科研究工具和数据存储库，总结和分析当前的数据再利用实践，并促进 IIR 领域的数据共享与再利用。研究界已发现 IIR 研究资源再利用面临若干障碍，包括社区资源组织方式零散、缺乏对这些资源存在情况的认知，以及现有资源的文档和组织不充分[2]。然而，当前关于 IIR 研究中数据再利用方法与障碍的讨论主要用于提出想法，而该领域需要更多实地探索，以提供证据和进一步见解，支持这些讨论。本研究旨在通过探索 IIR 研究人员为支持数据再利用而寻找数据和资源的行为，拓展当前研究成果。我们发现了 IIR 研究人员在评估数据可再利用性时会考量的共享数据的某些特征信息，以及研究参与者通常获取此类信息的来源。本研究为持续开展的工作作出贡献，帮助人们了解 IIR 社区中围绕数据共享与再利用的实践、挑战和决策过程。此外，研究结果揭示了发展知识基础设施的机会，这些基础设施可建立在现有社区实践（例如文本检索会议及类似活动[e.g. 12]）之上，让与数据相关的知识更易被发现、彼此关联并便于获取，从而支持该领域更有效且负责任的数据共享与再利用。

本研究存在一些潜在局限。首先，研究结果基于参与者对

数据再利用的自我报告想法和行为。我们认识到，这些报告可能与参与者在研究中的实际做法有所不同。其次，由于本研究采用半结构化方法，与参与者的访谈可能受到访谈者关注点的影响，从而引入偏差并影响研究结果。未来的研究中，我们计划对 IIR 研究人员的数据发现和复用行为进行现场观察，以更好地了解他们的预期、顾虑和复用实践。

参考文献

[1] Toine Bogers、Samuel Dodson、Luanne Freund、Maria Gäde、Mark Hall、Marijn Koolen、Vivien Petras、Nils Pharo 和 Mette Skov。2019年。CHIIR 2019交互式信息检索资源再利用障碍研讨会（BIIRRR 2019）概述。*CEUR*研讨会论文集（2019年3月）。https://oda.oslomet.no/oda-xmlui/ handle/10642/7005 R19。[2] Toine Bogers、Maria Gäde、Luanne Freund、Mark Hall、Marijn Koolen、Vivien Petras 和 Mette Skov。2018年。交互式信息检索资源再利用障碍研讨会。载于*2018*年人类信息交互与检索会议论文集（*CHIIR'18*）。计算机协会，美国纽约州纽约市，382–385。doi:10.1145/3176349.3176901 R24。[3] Toine Bogers、Maria Gäde、Mark Michael Hall、Marijn Koolen、Vivien Petras 和 Birger Larsen。2023年。我们在CHIIR中的工作、共享与再利用。载于*2023*年人类信息交互与检索会议论文集（美国得克萨斯州奥斯汀，2023 03 19 351 356 10 1145 3576840 3578305 - - ）。ACM，–。[4]
doi: 。/。Toine Bogers、Maria Gäde、Mark Michael Hall、Marijn Koolen、Vivien Petras 和 Paul Thomas。2022年。第三届构建信息交互与检索资源再利用研讨会（BIIRRR 2022）。载于*2022*年人类信息交互与检索会议论文集（*CHIIR'22*）。计算机协会，美国纽约州纽约市，374–376。doi:10.1145 3498366 3505838 /。Maria Gäde、Marijn Koolen、Mark Hall、Toine [5]
Bogers 和 Vivien Petras。2021年。关于交互式信息检索资源再利用的宣言。载于*2021*年人类信息交互与检索会议论文集。ACM，澳大利亚首都领地堪培拉，141–149。doi:10.1145/3406522.3446056 R11。[6] Diane Kelly、Jaime Arguello 和 Robert Capra。2013年。美国国家科学基金会关于基于任务的信息搜索系统的研讨会。载于*ACM SIGIR*论坛，第47卷。ACM，美国纽约州纽约市，116–127。[7] Jürgen Koenemann 和 Nicholas J. Belkin。1996年。交互的价值：一项关于交互式信息检索行为及其有效性的研究。载于*SIGCHI*人类因素与计算系统会议论文集：共同点——*CHI'96*（加拿大不列颠哥伦比亚省温哥华，1996年）。ACM出版社，205–212。doi: 10 1145 238386 238487 。/。Jiqun Liu。。面向受克兰菲尔[8] 2022
德方法启发的交互式信息检索评估可重用性评估。59，5（2022年），103007。doi:10.1016/j.ipm.2022. 103007[9] Jiqun Liu。2021年。解构交互式信息检索中的搜索任务：任务维度与预测因子的系统综述。《信息处理与管理》58，3（2021年），102522。[10] Jiqun Liu 和 Chirag Shah。2019年。交互式信息检索用户研究的设计、评估与报告。施普林格国际出版，瑞士查姆。https://link.springer.com/10.1007/978-3-031-02319-4 R12。[11] Mark Sanderson 和 W. Bruce Croft。2012年。信息检索研究史。《*IEEE*会刊》100，百年纪念特刊（2012年5月），1444–1451。doi:10. 1109/JPROC.2012.2189916 3。[12] Ellen M Voorhees。2007年。TREC：延续信息检索的实验传统。《*ACM*通讯》50，11（2007年），51–54。[13] Ellen M Voorhees、Donna K Harman 等。2005年。*TREC*：信息检索中的实验与评估。第63卷。麻省理工学院出版社，剑桥。

收稿日期：2025年10月29日