研究人员如何以及为何引用数据？学术文章中数据引用的修辞特征与功能研究

Sara Lafia (slafia@umich.edu)*[1], Andrea Thomer[3], Elizabeth Moss[1], David Bleckley[1], Libby Hemphill[1,2]

1. ICPSR，密歇根大学，密歇根州安娜堡，美国 2. 信息学院，密歇根大学，密歇根州安娜堡，美国 3. 信息学院，亚利桑那大学，图森，亚利桑那州，美国 * 通讯作者

作者贡献

概念化：A.T.、S.L.、E.M.、D.B. 和 L.H.；形式分析：S.L.、D.B. 和 E.M.；经费获取：L.H. 和 A.T.；方法论：S.L.、E.M.、D.B.、A.T. 和 L.H.；撰写初稿：S.L.、L.H.、A.T.、D.B. 和 E.M.

摘要

数据再利用是社会科学中的一种常见实践。尽管已发表数据在社会科学研究的产出过程中发挥着至关重要的作用，但它们并未得到一致的引用，这使得评估其完整的学术影响力并认可原始数据生产者的贡献变得困难。此外，理解研究人员引用数据的动机也可能颇具挑战。与对学术文献的参考一样，数据引用具有多种修辞功能，例如致敬、表明分歧或进行比较。本文研究研究人员如何以及为何在学术写作中引用社会科学数据。我们构建了一套类型学，用于模拟锚定数据引用的实体之间的关系，以及这些实体的特征 （访问、行动、位置、风格、类型）和功能 （批评、描述、说明、互动、使合法化）。我们通过对引用存档于密歇根大学校际政治与社会研究联合会（ICPSR）的社会科学数据的多学科研究文章（n=30）进行编码，说明该类型学的使用方法。我们展示了该类型学如何捕捉研究人员与数据的互动以及他们引用数据的目的。我们的类型学为记录和分析研究人员关于数据使用的叙述提供了一种系统方法，拓展了我们认可支持研究的数据贡献的能力。

关键词

引文分析、数据再利用、研究评估、学术交流 n

1. 引言

随着数据集进入科学记录，引文将已发表数据与更广泛的研究网络联系起来（Hey et al., 2009）。数据引文确立了研究结果的优先权，提供证据以表明研究发现的质量与重要性，并使人们能够研究研究人员如何使用现有数据。引文分析依赖引用标准化来评估学术交流模式，例如思想在不同科学学科间的传播范围或可见度（如通过论文引文网络）。学术论文的引文索引，如科学引文索引（Garfield, 1964年），使研究人员能够了解哪些人被大量引用、哪些已发表作品被大量引用，以及哪些出版物渠道具有影响力。近期对文献计量网络的审查还揭示了不平等现象，表明引文并非客观的（Kwon, 2022年）。例如，引文率因研究主题以及作者的种族与性别而异，这表明社会因素在研究人员了解已发表作品并决定是否引用这些作品方面发挥着重要作用（Kozlowski et al., 2022年）。

尽管用于研究科研出版物引文趋势的基础设施已经相当完善，但三项主要挑战限制了对数据引文进行大规模分析。第一项挑战是明确识别数据引用。尽管已有成熟的系统用于引用他人的工作（Chernin，1988），许多作者仍未能引用数据。尽管数据存储库已就数据引文的最佳实践提供指导（Fenner 等，2019），且资助方和出版商也施压要求“让数据发挥作用”（Cousijn 等，2019），许多作者仍在其写作中以非正式方式提及数据。当将数据与出版物关联起来的负担主要落在作者身上时，研究论文中对数据集的引用往往不完整或不一致（Boland 等，2012）。非正式的数据引文实践使读者难以了解作者访问了哪些数据，以及他们是对数据进行了分析，还是仅仅对其进行了描述（Moss 和 Lyle，2018）。

第二项挑战涉及理解数据引用的意图。文献计量分析通常将引文视为一种可以标准化的对象，并普遍将其解释为赋予已发表作品合法性（Cronin，1981）。然而，与学术文献的引文一样，研究人员出于不同目的引用数据。现有的引文类型学涵盖了研究人员引用材料的各种原因（例如说服、批评和对比）。许多研究人员通过实证研究传达其研究发现，在这类研究中，他们提出的论断与其他学术成果（包括已发表数据）相关联。作者的主张在具体程度上从明确到隐含不等，且通常得到数据的支持（Blake，2010）。

第三个相关挑战是推断引文的质量。用于定量影响评估的文献计量指标（如 h指数）能够反映来源的受欢迎程度或可见度（Egghe, 2010年），但很少说明围绕该来源的互动性质。关于学术文献引文的先前研究区分了表层引用与深入利用来源材料的引用（Cronin, 1984; Leydesdorff, 1998; Spiegel-Rosing, 1977; White 和 Wang, 1997），并有助于确定引文的目的或极性（Abu-Jbara 等人, 2013; Cohan 等人, 2019年; Hernández-Alvarez 和 Gomez, 2016; Teufel 等人, 2006）。例如，致敬性质的引文（如致敬自己的导师或其他有影响力的研究人员）也会造成累积优势，使最知名的研究人员获得远多于他人的工作认可（Merton, 1968）。因此，某一来源收到的引文数量并不能表明引文的目的，也未必能作为研究质量的可靠替代指标（Garfield, 1979年）。

鉴于分析数据引用面临的挑战，本研究采用定性方法，旨在确定科学研究中社会科学数据再利用的类型、原因及相关互动。关于数据引用的先前研究主要关注正式的书目引文（Belter, 2014年; Jiao 和 Darch, 2020年; Mooney 和 Newton, 2012; Park 等人, 2018）。相比之下，我们还仔细分析论文中间接的数据提及——即提到某个数据集或其部分内容、但未正式引用的句子。我们发现，数据引用承担的功能有限，并通过一种类型学加以界定，以概括并描述研究人员引用数据的各种方式。随后，我们运用这一类型学分析社会科学出版物中研究数据的使用情况。

2. 背景

2.1 数据的定义与引用

“数据”和“数据集”这两个术语的含义因语境而异。通常，科学家在互动并记录观测值时所作的选择，决定了什么会成为“数据”。因此，“数据”是解释的副产品，实际上可以理解为“仅指经过分析的内容”（Coombs，1964年）。定义“数据”的难点之一，在于其身份“难以驾驭”且“边界模糊”，这使其难以作为可供轻松引用和检索的数字对象发挥作用（Wynholds，2011年）。技术和科学文献中对“数据集”一词的用法也存在分歧，这给数据共享和保存带来了挑战（Renear等人，2010年）。“数据”是表达内容的符号抽象组合；“数据集”由多个数据承载实体构成，并可能包含有关数据的其他上下文信息，例如收集方法（Furner，2016年；Wickett等人，2012年）。在我们的分析中，我们聚焦于已归档的社会科学数据集，这些数据集包含上下文元数据和文档，由他人出于研究目的制作并共享。

捕捉数据集与其他学术成果之间的关系，对于赋予数据集应有的认可至关重要。数据的“参考”“提及”和“引文”彰显其重要性，并通过署名赋予认可（Altman M, Borgman C, Crosas M, Matone M, 2015）。虽然“参考”和“引文”常被交替使用，但“参考”通常表示某项成果列于出版物的参考文献部分（Gilbert and Woolgar, 1974）。“引文”一词则意味着使用持久标识符（PID）；相较于“参考”或“提及”，在文献计量学中，它带有更正式的意味（Ball and Duke, 2015）。带有 PID 的引文将已发表作品与其使用情境关联起来，使人们能够验证并再利用现有的科学分析（Buneman et al., 2022; Shotton, 2010）。虽然作者开始使用数字对象标识符（DOI）来引用数据，但关于何时以及为何应这样做的最佳实践尚未得到广泛遵循（Mayo et al., 2016）。例如，近期一项对使用 DOI 引用数据集的文献综述发现，许多作者引用了他们曾提及或描述过（例如数据收集方法）、但并未重新分析或以其他方式使用的数据（Banaeefar等人, 2022）。在本研究中，我们使用“数据引用”这一通用术语，涵盖非正式的数据提及（即仅使用数据集名称）、正式引用（即纳入数据集 DOI 的 APA格式引文）以及对数据使用情况的描述。

2.2 学术交流中的引文

对科学学科内引文的分析揭示了信息流，并将彼此分散的信息脉络汇集起来，从而构建科学领域内各主题的“共识模型”（Garvey and Griffith, 1972）。引文反映了对作者的影响，而引文模式则追踪活跃研究网络中的交流（Edge, 1979）。引文在微观层面和宏观层面发挥着不同作用。在微观层面，引文体现职业关系，并发挥奖励作用；在宏观层面，引文群体则作为概念符号，在层级社会网络中对知识进行编码（Leydesdorff, 1998）。论文被引用时，可以作为其文本所表达思想的象征（Small, 1978）。引文具有强大的影响力，因为它们具有持久性，并且会获得独立于参与其形成的人们之外的身份。它们是“言语行为”，即在文献中留存并可随时间推移接受审视的简短陈述（Smith, 2014）。被引用的来源常用于佐证陈述或假设、指向更多信息、承认同一领域的先前研究，以及进行批判性比较以彰显研究质量（Spiegel-Rosing, 1977）。

研究基础设施研究依赖引文作为追踪署名和表明数据集或软件等学术成果影响力的指标（Mayernik 等，2017）。期刊和数据出版商等机构通过出版指南和格式手册，规范学科与文化层面的写作样式和引文方式。使用特定标识符的数据引文能够让读者识别、获取研究数据并给予认可。尽管已有相关建议和最佳实践，正式数据引文在学术写作中仍不普遍（Mooney 和 Newton，2012）。不完整、非正式或格式不当的引文，给追踪数据使用带来了障碍（Zhao 等，2018）。例如，模糊或隐含的参考会使不熟悉变量或其他数据特征的读者难以理解作者使用了哪些数据以及如何使用这些数据（Moss 和 Lyle，2018）。因此，如果只关注正式引文实践（例如使用 DOI），就会忽视许多潜在的数据引用。我们希望更全面地理解作者在论文中以正式和非正式方式提及数据时所采用的修辞方式。

2.3 引文的意义与动机

引文赋予科学研究以荣誉与认可（Cronin，1984年）。然而，作者的引文实践与利用引文评估表现、衡量研究影响力之间，可能存在脱节。换言之，引文表明了引用了“什么”以及引用频率，却没有解释引用作品的“原因”。作者引用出版物的动机可分为科学性动机和策略性动机。科学性引文用于提供背景、指出研究空白，并为比较奠定基础。策略性引文则体现对主观规范的认可，并宣传已发表的作品（Lyu 等，2021年）。尽管人们往往认为，引文代表着高质量的作品，且这些作品对作者的研究产生了影响，但一项调查发现，作者决定引用哪些作品时，更多是受战略因素驱动，而非基于对相关研究的熟悉程度或对其质量的判断（Teplitskiy 等，2018）。

对作者开展的行为调查和访谈揭示了他们对所引用内容及其引用原因的判断，而这些判断并未反映在学术记录中（刘，1993）。在一项此类研究中，作者在决定是否引用某些出版物时，会考虑出版物的新近程度、主题针对性和易用性

（White 和 Wang，1997）。作者还认为，引文体现了文献的突出性或新颖性，是一种“概念标记”，而引用这一“概念标记”将通过与现有研究保持一致或对其进行批评，增强自身工作的权威性（Case 和 Higgins，2000）。Silvello 识别出跨科学领域共有的六种数据引用主要动机：数据归属（问责和贡献）、数据关联（与出版物中的论断相关联）、数据发现（识别和检索）、数据共享（声誉）、数据影响（评估影响范围）以及可复现性（验证和程序）（Silvello，2018）。然而，仅凭这些动机并不能解释作者的数据引用行为及其为何会因出版场所和背景而异。

2.4 分析引文内容和背景

许多引文分析的计算方法是在既有关于作者引用动机的认识基础上提出的。多项内容分析研究所识别出的常见引文类别包括背景信息、理论框架、先前的实证或实验性证据、负面区分以及方法论解释（Ding 等人，2014年）。引文出现的出版物章节等特征，可以与引文的语义内容结合使用，以预测引文意图（Nakov 等人，2004年）。研究者提出了各种分类方案，用于标注作者引用已发表研究时的意图（Hernández-Alvarez 和 Gomez，2016年）。其中一种方案通过区分并加权负面、中性和正面引文，来说明引文目的（即作者意图）和极性（即作者情感）（Abu-Jbara 等人，2013年）。更细粒度的基于规则的编码方案则区分弱点陈述、与其他研究的对比或比较、赞同、与其他研究的兼容性以及中性引文（Teufel 等人，2006年）。相较之下，粒度较低的方案通过区分背景信息、方法和比较类引文，支持一般性的引文意图分类（Cohan 等人，2019年）。这类方案还可以帮助区分引文框定（例如使用、动机、未来、扩展、比较或对比、背景）（Jurgens 等人，2018）。

定性方法通过聚焦参考文献的背景和特征，回应了预测作者意图的困难。对学术文献中数据引文的综述发现，这些引文在两个主要维度上存在差异：被引实体和风格（Fear，2013年）。数据生产者（创建数据的研究人员）和数据提供者（提供数据的个人或机构）通常会在数据引文中被标明。另一项近期研究发现，研究人员倾向于使用他人创建的数据进行比较（例如，真实值验证、校准和确定基线测量）以及整合（例如提出新问题和开展新研究）（Pasquetto 等，2019年）。一项大规模调查发现，现有数据经常被用作新研究的基础、为新项目做准备、产生新想法、开发新方法、通过分析和意义建构验证其数据，以及用于教学（Gregory 等人，2020年）。鉴于数据再利用可以满足多种需求，研究人员引用数据的目的也各不相同。我们以有关数据再利用实践的既有认识为基础，展示研究人员如何认可数据的贡献。

3. 材料与方法

3.1 抽样框

我们分析了一组出版物样本，这些出版物引用了一个或多个可通过密歇根大学校际政治与社会研究联合会（ICPSR）获取的数据集；ICPSR 是密歇根大学下属的大型社会科学数据档案库。我们以现有的学术出版物引文方案为基础构建类型学，并通过迭代编码对其进行扩展和细化。该类型学涵盖了作者在参考研究数据时采用的结构特征和修辞功能。先前研究主要关注特定的出版物风格，例如数据论文（Jiao 和 Darch，2020年；Li 和 Jiao，2022年），或特定的出版平台，例如 PLoS One（Zhao 等人，2018）。与之不同的是，我们选取了引用 ICPSR 已归档社会科学数据的多学科出版物。这种方法使我们能够涵盖更广泛的数据引用背景。我们考察了研究出版物完整篇幅背景中出现的数据引用。此外，我们唯一的选择要求是每份出版物都提及一个或多个已归档的社会科学数据集。

我们分析了作为 ICPSR 扩充《ICPSR数据相关文献书目》工作的一部分而检索到的论文。该书目收录了超过 100,000 篇使用可通过 ICPSR 获取的现有社会科学数据的出版物。书目的审查过程包括在书目数据库中搜索有关已发表 ICPSR 研究所提供数据的参考文献。工作人员会人工审查出版物检索结果的元数据和全文，以查找数据使用的证据。ICPSR 执行严格的收录标准，以确保书目中的出版物确实反映了数据使用情况。若出版物明确提及一项或多项可通过 ICPSR 获取的研究，且能够确定作者访问并分析了相关数据，则会予以收录。如果出版物未能证明对 ICPSR 数据进行了实质性使用，或者无法确定作者分析所使用的具体研究或序列，则会被拒绝收入书目。

为了构建一个用于检验我们类型学的抽样框，我们首先从当前的 ICPSR书目中确定了五篇代表 ICPSR 数据多学科使用情况的出版物。我们仔细阅读这些出版物，以识别数据引用并制定初步类型学。随后，我们检索了由 Dimensions 文献计量数据库提供的外部出版物全文索引（Hook 等人，2018），以查找截至 2022年2月可获取的 ICPSR 11,639 个研究DOI中的任何引用。在 ICPSR书目工作人员的支持下，我们评估并将 2,546 条检索结果分为六个类别，以表明这些出版物是否符合《ICPSR书目》的收录标准。这些类别由 ICPSR 工作人员提出（Banaeefar等人，2022年），定义见**附录 A**

（补充文件 1）。随后，我们从每个类别中随机选取出版物纳入分析，最终共纳入三十篇出版物。我们为每篇出版物收集了额外的元数据，例如来自 Dimensions 的研究领域类别，以确定样本的学科覆盖范围。我们在**附录 A（补充文件 1）**中报告了出版物抽样框和筛选标准。

3.2 定性编码

我们的团队开展了两个阶段的编码。第一阶段旨在制定编码手册，以描述数据引用的多样性。为制定和完善编码，我们团队的三名标注员多次阅读**附录 A（补充文件 1）**中列出的每篇标记为“阶段 I”的出版物全文。标注员分别提出修订意见，并共同完善编码手册的共享版本。为实现定性可靠性，标注员参考了大量示例，并在每周会议上对其进行讨论（Bauer，2000年）。标注员每周会面，在交互式编码会议中审查并纳入所提出的变更。不断涌现的观点和相互冲突的意见促成了对话，并据此形成了编码手册。除编码外，标注员还讨论了数据引用的范围，以及编码手册的重点、目的和定义。每位团队成员都独立应用编码手册每次迭代后的更新编码，以识别并标注全文中的所有数据引用。团队重复这一过程，直至达到饱和且不再提出新的编码（Charmaz，2006年）。第一阶段编码形成了一份稳定的编码手册，我们在**附录 B（补充文件 2）**中报告了该手册。

我们还描述了标注员之间编码的一致程度。鉴于标注员从出版物的非结构化全文中选取并编码文本片段，我们使用 Holsti指数（Holsti，1969）作为赞同度测量指标。最终的 Holsti指数为 67.9%，表明赞同度相对较高；这是因为标注员选取了不同的文本片段，并使用多个编码对其进行编码。团队的一名成员独立地将该类型学应用于抽样框中标记为“第二阶段”的 25 篇出版物留出集。第二阶段展示了该类型学的实际应用，并呈现了**附录 A（补充文件 1）**中所报告的研究发现，这些发现将在**第4节**中介绍。

4. 结果

数据引用类型学由四个父级编码组成（数据实体、数据引用、特征和功能），其内容在**表1**中进行了汇总和定义。数据实体锚定数据引用，其定义基于 Renear等人（Renear等人，2010）提出的实用性区分，用于界定科学文献中数据集的组成部分。Renear等人区分了数据内容（包括文件和观测值）、分组（它们所属的集合或研究）以及用途（用于解释数据的元数据）。同样，我们将数据实体定义为“表示已记录观测值的一个或多个词”，并将其记录为“文件”“元数据”“研究”或“变量”。我们排除了数据实体不具体或未在论文正文中讨论的情况。例如，如果某个数据集的名称出现在出版物标题中，但数据未在正文中描述，我们不会将其视为数据实体。鉴于某些数据分析讨论较为宽泛（例如统计检验的结果），我们仅考虑引用了特定数据实体的陈述。

一条数据引用是一个或多个数据实体出现的上下文窗口。我们尝试了各种上下文窗口，并确定段落能够充分呈现数据实体前后的细节。我们重点关注三种数据引用类型，这些类型在**第2.1节**中介绍：引文、提及和使用。数据引文包含明确指向已发表数据源的指示，但不提及数据集，也不表明作者使用了

数据。数据提及会提及某个数据集以对其进行描述，但不表明使用了数据。数据使用会提及某个数据集，并描述作者与数据之间的互动。我们为每条数据引用应用特征和功能编码。完整的编码手册以及定义、规则和示例，均见**附录 B（补充文件 2）**。

表1。父级编码、子代码和定义概览

父级代码	子代码		定义
	第1级	**第2级**	
数据实体	File		表示.....的一个或多个词记录的观测值
	元数据		
	研究		
	变量		
数据引用			其中一个或更多数据实体被提及
特征	访问	提供、接收	结构 , form, and 数据的呈现方式 参考
	行动	引用，提及 , Uses	
	位置	摘要、致谢、附录、图注.....	
	样式	首字母缩略词、通用项、名称、括号内的	
	Type	派生数据、主要、二次使用	
功能	批评	比较、局限性	数据的用途 参考
	描述	构成、来源	
	说明	背景 ，展望	
	互动	解释、处理	
	使合法化	论证、透明度	

4.1 数据引用的特征

这些特征描述了数据引用的结构、形式和外观。我们识别出的数据引用共有二十四项特征，归纳为五个子代码（访问、行动、位置、样式和类型）。访问 代码用于说明作者是否在引用中描述了数据共享或检索。行动代码用于捕捉作者与数据之间的距离，并以一个连续谱表示，包括“引用”（即仅以括号形式提及数据，不提供其他背景信息）、“提及”（即描述或暗指数据）和“使用”（即描述实际动手处理数据的工作）。行动代码以 Pasquetto 等人提出的比较性数据再利用与整合性数据再利用之分为基础（Pasquetto 等，2019）。位置 代码标注数据引用在出版物中出现的章节，例如摘要、致谢、图注、图、表、脚注，或扩展 IMRAD 结构中所述的方法部分（ Sollaci 和 Pereira，2004）。样式代码记录作者如何

通过使用首字母缩略词（如“美国国家选举研究”）、通用名词（如“数据”或“研究”）、正式名称（如“美国国家选举研究，2016”），或采用括号内的引文标注作者及出版年份，来指明数据实体。类型代码用于判断数据实体是源自现有数据、代表作者创建的一手来源，还是供其他研究人员使用而出版的二手来源。

4.2 数据引用的功能

功能反映了每个数据引用的目的。我们识别出的数据引用的十种修辞功能归纳为五个子代码（批评、描述、说明、互动和使合法化）。各代码及子代码的定义和示例见

附录 B（补充文件 2）。

批评代码包括 (1) 比较，即将作者的研究与使用该数据或其他来源研究发现的其他研究进行对比。作者进行比较，以对照自己的研究与先前研究发现，或总结先前使用数据得出的结论。批评代码还包括 (2) 局限性，表明作者意识到使用数据时需要谨慎。此代码包括承认质量问题，例如可能影响数据使用范围的潜在错误或抽样偏差。

描述代码包括 (3) 构成，即解释或讨论有关数据或元数据的知识。构成类数据引用描述数据包含的内容（例如抽样总体）或研究背景（例如数据收集方法）。描述代码还包括 (4) 来源，作者在其中描述数据的来源信息。来源类参考文献会注明相关数据的来源地以及数据生产者或提供者。

标注为说明 代码的参考文献具有说服力。该代码包括（5）背景，作者在其中提供背景信息、研究发现或源自引用数据的统计数据。在背景 参考中，数据具有转喻性，代表作者提出的观点或主张。说明代码还包括（6）展望，作者在其中推测其研究中未进行或审查的数据的潜在应用。展望参考基于数据的属性，主张其具有潜在效用。

互动 代码描述了对数据进行的实际操作。互动 包括一个子代码（7）解释，作者在其中提出源自对引用数据进行分析的经验性主张。解释参考通常出现在作者描述其分析之后。互动代码还包括一个子代码（8）处理，作者在其中描述处理数据时执行的步骤。处理涉及选择变量、为分析准备或转换数据，以及抽样、相关分析、整合和验证分析等具体的数据准备技术。

最后，使合法化 代码用于通过对数据所作的价值陈述来说服读者的参考。使合法化代码包括（9）论证，即强调数据的某项特征，该特征能够提升作者选择的可信度或权威性。论证 参考的示例包括作者说明选择数据的原因、讨论数据的可信度或代表性，以及描述以往数据使用情况并据此限定其选择。使合法化代码还包括（10）透明度子代码，作者在其中解释为何或如何应用某项分析程序，并

说明分析中考虑的信号质量或其他因素。透明度参考的示例包括通过将分析纳入补充材料，使方法公开或可复现。

4.2.1 数据引用为读者提供数据访问途径

数据引用提供了多种访问数据的方式。尽管有些期刊要求作者附上数据提供声明，说明作者会向读者提供其分析所用的数据，但在我们审阅的出版物中，这类声明并不常见。我们遇到的示例包括作者提供其分析派生数据的访问途径，并明确表示目的是实现复现。或者，作者也可能提供数据访问途径，以此招募未来的合作伙伴。以下是一则关于数据提供 的说明：

> “作者已公开本文所呈现分析所依据的数据（见 Styck、Beaujean 和 Watkins，2019年），从而使合格的研究人员能够复现并进一步拓展这项工作。后续使用者有义务将数据创建者纳入其发表计划，若数据创建者有此意愿。”（Styck 等，2019年）

关于作者从提供者处访问或接收数据的陈述，通常伴有正式的括号内数据引用。我们将数据接收定义为对现有数据实体的参考，以及对如何访问该数据的说明。例如，如果某项参考以括号内形式出现，那么参考文献列表中的对应条目必须提供一种访问机制，例如统一资源定位符，以便他人访问该来源。在以下示例中，作者通过括号内引文正式注明数据创建者，并列出所进行分析的详细信息，以及促使其选择该数据集的历史背景：

> “为了检验核试验产生的放射性沉降物是否对农业部门产生了持久影响，我根据1940年至1997年的《美国农业史》资料，构建了包含可比变量的面板数据 ~~Haines等人（~~2015）。~~这些普查数据来自美国~~最全面的农业调查，相关调查可追溯至1840年。从1920年开始，农业普查每十年开展一次。我使用这些数据，探究放射性沉降物沉积对国家层面长期结果和农业发展的影响。”（Meyers，2019）

4.2.2 数据引用表明作者与数据的互动

数据引用涵盖了作者与数据之间互动的三个层面。首先，我们找到了表层数据引文的示例：作者以引用学术文章的相同方式引用已发表的数据集。在这些情况下，作者没有在文中指出具体的数据集，而是通过脚注或括号内的引文，在参考文献列表中正式注明数据集。大多数数据引文出现在引言部分，且具有背景性，旨在提供背景、研究发现或统计数据，作者借此佐证某个观点。然而，作者引用的统计数据或图究竟如何与源数据相关，或如何从源数据派生出来，往往并不清楚。在下例中，数据引用提供了研究发现，却没有进行直接分析。文中没有使用任何动词来描述对数据进行的行动

；读者只能假定作者此前有分析这些数据的经验，或认为所引图表与该数据集已发表的汇总统计数据相关。在下例中，作者提供了统计数据，并附有相应的脚注；该脚注指向文章参考文献列表中关于印度人类发展调查数据的正式引用：

> “贫民窟的住房、水、卫生设施及其他服务质量较差，这会导致多种后果，其中包括更高的疾病发生率和死亡率。[17] [18] 另一方面，富裕家庭通常居住在有管道供水的地区；遇到缺水时，他们可以建造储水设施、抽取地下水井中的水，也可以付费购买送上门的水。印度城市人口中最贫困的五分之一家庭，只有38%能在室内使用管道供水；相比之下，最富裕的五分之一家庭中，这一比例为62%。”（Frumkin 等人，2020年）

作者提及数据而非引用数据时，会描述数据集的构成或来源。与引文不同，提及会在正文中直接点明数据的名称。我们发现，数据提及主要出现在文章的方法、引言、讨论和脚注中。许多数据提及介绍了数据产品的构成细节，并传达了有关研究依据、收集方法或总体的信息。提及数据既能提供背景信息，介绍作者随后在分析中使用的数据，也能表明作者知晓某些数据，但评估后决定不予使用。在下面的示例中，作者描述了一项调查各波次之间抽样总体的变化，以说明其样本选择方法、表明其了解数据质量，并为所采用的方法提供依据：

> “在教育、健康、关系状况和就业状况方面，第1波中未能留在分析样本内的受访者，相较于留在样本内的受访者处于劣势。因此，如果更容易出现抑郁症状的人留在分析样本内的可能性较低，那么第1波与第2波之间的样本流失可能会导致我们保守地评估经历经济衰退的背景如何影响当地居民的抑郁症状。”（Settels，2021年）

4.2.3 数据引用是实证论证的构成要素

在对数据进行批评的示例中，作者会描述其他人先前的工作或研究发现，并将其与自己的方法进行对比。在某些情况下，作者会说明自己如何以不同方式使用同一数据，或基于所提供的理由决定不使用该数据。下面给出了一个数据比较的示例。作者介绍了几项研究相似总体的纵向研究，并解释了研究发现可能存在差异的原因，指出这些差异源于研究构成不同。通过这种方式，作者表明自己已尽到应有的审慎责任；他们了解相关研究，也能描述其局限性：

> “全国律师资格考试通过情况纵向研究针对比 After the JD 更早的队列开展研究，发现少数族裔的长期律师资格考试通过率明显低于白人。[5] 因此，对所有法律学位持有者（包括未通过律师资格考试者）进行研究，

> 可能会发现更大的收入种族差距。人口普查调查（例如本论文使用的调查）不包含律师资格考试通过情况，因此与 After the JD 相比，其中可能包含比例更高的低收入人群。[6]" (McIntyre 和 Simkovic，2018)

在文章的方法和讨论部分，以及图注、图和表中，都发现了更多表明数据使用的引用。根据作者使用的动词和人称代词，对提及和数据使用声明进行了区分。大多数使用声明描述了行动，具体而言，是数据处理（例如处理数据时执行的步骤）和解释（例如提出源自数据分析的经验性主张）。描述使用的数据引用也出现在附录和补充材料中，而非文章中指定的部分，例如致谢或数据可用性声明。数据处理的示例包括从引用数据中选取变量，以及对引用数据进行准备、转换、修改、抽样、子集选取、比较或相关分析。数据解释包括构建理论、进行比较，以及解释图中的实证证据。以下示例展示了作者如何在详细说明其分析方法时，引用某项研究的两个波次及相关变量：

> "要评估遗传和环境因素随时间保持稳定的程度，需要将经典双生子设计扩展至涵盖重复测量。在此，我们采用了双变量乔列斯基分解法（见图1）：对于 n 个测量变量中的每一个，乔列斯基分解都会指定 n 个潜在的 A、C 和 E 因子。若将其视为一幅图，并把潜在因子排列在测量变量上方，则每个因~~子都与其~~下方的测量（显变量）相连，并与右侧的所有变量相连。这样，每个潜在因子所连接的变量数都比前一个因子少一个。这一设计有助于回答当前问题，因为它既可以估计第1波的 A、C 和 E 效应，也可以估计这些效应在多大程度上能够解释第2波的方差，以及第2波出现的新方差。"（Lewis 和 Bates，2017）

5. 讨论

我们的类型学将数据使用的概念扩展到重新分析之外。例如，一些研究人员可能会访问并重新分析已发表的调查数据，而更多人可能会将该调查的问卷或抽样设计作为黄金标准加以再利用。用户也可能通过指出调查数据在探讨特定主题时的局限性来对其提出批评。我们的方法广泛搜集此类数据引用，从而深入了解社会科学数据如何支持研究。我们的类型学有助于为研究人员提供建议情境，说明他们何时、为何以及如何引用已发表数据。此外，它也为制定新颖的数据再利用指标奠定了基础，这些指标反映了人们使用数据的多种形式，从再利用调查设计到重新分析调查数据。

我们的类型学还揭示了数据引用与传统书目引文之间存在差异和相似之处。首先，被引用实体的规模可能有所不同；我们发现了对单个文件、元数据记录、整体研究以及单个变量的引用。书目引文的规模同样可能有所不同（例如，引用论文中的特定短语或章节，而非整篇论文

）。数据实体的组成部分则具有不同且可能更为广泛的范围。通过审阅其他领域的论文，进一步拓展和完善这一类型学，或许能够发现更多子实体（例如，考古学或古生物学研究可能会引用特定的人工制品，以及由这些人工制品派生的数据）。还需要进一步研究，以理解这些不同引用规模的影响；不同规模（例如变量级与完整数据集级别）的参考是否与不同类型的使用和论证相关？不同规模的数据是否更有可能或更不可能促成对数据集的正式引用？数据实体还可能具有多个可供引用的版本（尽管我们在样本中未发现这种情况）；这会如何使我们追踪学术影响的流动变得更加复杂？

其次，我们发现，数据引用可以像书目参考文献一样，发挥“概念符号”（Leydesdorff，1998）的作用。通过首字母缩略词或名称非正式地提及数据集（且不附正式引用），表明作者熟悉这些数据集，正如人们熟悉学术研究中的经典著作一样。换言之，生物学家提及达尔文或经济学家提及洛克时所表现出的熟悉程度，同样也体现在对数据集的引用中。未来的研究若能识别出这些基础性或经典数据集，或许有助于揭示作为概念符号的数据集与书目参考文献有何不同。数据集或许有其独特之处：它们还可以具有鲜明的转喻功能，即对整个数据集的引用可以代指对数据集某个具体部分或特征的引用（正如我们的“背景”代码所揭示的那样）。

第三，数据引用与数据实体之间存在一些通常不会出现在书目实体中的互动，即数据的提供和存档。数据集既是可供使用的资源，也是可供引用或向他人提供的学术成果。在我们标注的出版物中，作者很少直接提供其基于现有数据得出的研究发现；更常见的是，作者仅描述他们所访问的数据源或数据提供者，以建立可信度和信任。关于研究人员对数据共享与再利用态度的先前研究表明，研究人员不愿提供数据访问权限，因为他们认为数据对他人没有价值，或者因为囤积数据有助于吸引未来的合作（Cragin 等，2010；Pasquetto 等，2019）。尽管我们的样本并不具代表性，但我们发现了一些与这项先前研究相符的初步迹象。

最后，我们还发现，这与描述作者引用文献动机的既有分类方案相一致。我们所识别的修辞功能，通过帮助读者发现作者分析过的数据，体现了作者研究发现的质量、可验证性或可复现性（Silvello，2018）。总体而言，我们考察的数据引用要么提供了有关数据集构成的详细信息，要么描述了数据处理。在我们识别出的示例中，作者补充了其所进行分析的背景信息，从而将数据源与其用途联系起来。此外，当作者提供数据的具体访问信息时，读者便能够获取相同的数据集。

5.1 局限与未来工作

本研究提出了一种类型学，用以建模作者如何参考研究数据。我们通过细读ICPSR书目中的论文，并不断添加新类别直至达到饱和，来构建这一类型学。本分析无意提供

关于特定引文趋势的定量证据。我们需要在更大规模上开展标注，并采取额外的测量指标，以验证标注员之间的一致性。此外，我们的抽样框是通过选取先由专家（即ICPSR书目工作人员）审阅并分类的论文构建的；样本在**附录 A（补充文件 1）**所列类别之间保持平衡。虽然这种抽样策略有助于开发和分析ICPSR书目，但未来将该类型学用于其他目的时，可能需要不同的筛选标准。

我们设想将这一类型学用于研究社会科学学科间数据引用的差异（例如，将数据引用策略的序列或共现情况作为科学学科或分析方法的标志）。近期一项关于 ICPSR 数据引文实践的研究发现，已发表文献中存在意料之外的数据集 DOI 使用情况，而这些引用并未表明数据被实际使用（Banaeefar等人，2022年）。我们的类型学可用于研究研究人员何时以及为何使用数据集 DOI，并区分描述数据的引用与暗示进行了数据分析的引用。

6. 结论

尽管研究数据在现代科学分析中日益重要，但从历史上看，它们并未被视为主要研究成果。研究数据的出版、长期保存和传播，以及描述性元数据，使其他人能够发现、使用并引用其他研究人员为其他目的收集的观测值。我们提出了一种数据引用类型学，用以刻画数据在科学出版物中发挥的功能：批判性、描述性、说明性、交互性和合法化。该类型学涵盖研究人员与数据的互动（例如，利用数据开展的工作或分析）以及他们对数据的判断（例如，根据已知的数据情况对其适用性作出的论断）。理解作者为何引用研究数据，对于给予数据生产者和提供者应得的学术研究认可至关重要，因为他们为科学研究工作的开展提供了便利。

作者说明

ICPSR数据相关文献书目可在
https://www.icpsr.umich.edu/web/pages/ICPSR/citations/ 查阅。

致谢

我们感谢密歇根大学信息学院的 Elizabeth Yakel、Morgan Wofford 和 Lizhou Fan 对早期草稿提出意见。

资金支持信息

本研究得到美国国家科学基金会资助（项目编号 1930645）。本项目部分得益于博物馆与图书馆服务研究所的资助（LG-37-19-0134-19）。

附录 A（补充文件 1）：抽样框

阶段	标题	年份	作者	DOI	页码	和引用	筛选标准	标准定义		出版物研究领域	ICPSR研究已引用	ICPSR研究作者
I	美国奴隶制的政治遗产美国奴隶制	2016	Avidit Acharya，Matthew Blackwell，Maya Sen	10.1086/686631	20	187	已在参考文献	示范用途ICPSR数据的引用数字对象标识符。标题已收录于ICPSR参考文献（重复记录）。	16 人类社会研究人类社会；1606 政治学《科学》	研究	三代合并，1965—1997年；青年—父母社会化面板研究，1965—1997年：四波合并；历史、人口、经济和社会数据：美国，1790—2002；美国国家选举研究1984年–1998年序列研究	Elliot, Patrick；Jennings, M. Kent，Markus, Gregory B., Niemi, Richard G.，Stoker, Laura；Haines, Michael R.；Warren E. Miller E.；密歇根大学密歇根大学。研究所社会研究研究。美国全国选举研究
I	《妈妈有了避孕药》：Anthony Comstock 如何以及 Griswold 诉康涅狄格州塑造了美国生育	2010年	Martha J. Bailey	10.1257/aer.100.1.98	31	85	已收录于参考文献	示范用途的 ICPSR数据引用数字对象标识符。标题已经在ICPSR参考文献（重复记录）。	14 经济学；15 商业，管理，旅游和服务	研究	出生率详细档案，1968年至1980年	美国部卫生与公众服务部。国家健康统计中心统计数据
I	农村隔离、小城镇与亲密伴侣暴力亲密伴侣暴力	2022年	Kathryn O. DuBois	10.1177/0886260520943721	23	1	已在参考文献	示范用途ICPSR数据引用的数字对象标识符。标题已在ICPSR书目（重复记录）。	16 项研究发表于人类社会；17 心理学与认知科学；1602 犯罪学；1607 社会工作；1701 心理学	研究	国家犯罪受害调查，合并文件，1992年至2015年	美国美国部司法部。司法项目办公室司法项目办公室。司法统计局统计数据
I	父母与同伴的社会规范与青少年的中学后教育态度：潜在类别分析	2018	Kremer, Kristen P.，Vaughn, Michael G., Loux, Travis M.	10.1016/j.childyouth.2018.08.026	13	5	已收录于参考文献	示范用途ICPSR数据引用数字对象标识符。标题已在ICPSR参考文献（重复记录）。	14 经济学；16 人类社会社会；1402 应用经济学；1607 社会工作	研究	高中纵向研究，2009-2013 [美国州]	美国部教育。教育科学研究所教育科学。国家中心教育统计数据

阶段	标题	年份	和作者	DOI	页码	和引用	筛选标准	标准定义		出版物领域 研究	ICPSR研究 已引用	ICPSR研究 作者
I	指数验证 功能上的 重要的呼吸 ……的症状 全国代表性 成年人 人口评估 烟草与健康 研究，2014—2016年	2021	Halenar, Michael J., Sargent, James D., Edwards, Kathryn C., Woloshin, Steven, Schwartz, Lisa, Emond, Jennifer, Tanski, Susanne, Pierce, John P., Taylor, Kristie A., Lauten, Kristin, Goniewicz, Maciej L., Niaura，Raymond，Anic，Gabriella，Chen，Yanling，Callahan-Lyon，Priscilla，Gardner，Lisa D., Thekkudan, Theresa, Borek, Nicolette, Kimmel, Heather L., Cummings, K.M., Hyland, Andrew, Brunette, Mary F.	10.3390/ijerph18189688	6	4	已在 参考文献	示范用途 的 ICPSR数据 引用数字对象标识符。标题 已收录于 ICPSR 参考文献 （重复记录）。	11 医学与 健康科学； 1102 心肺 医学与 血液学；1117 公共卫生与 卫生服务	总体 评估 烟草与 健康（PATH） 研究 [美国 州] 受限使用 文件	美国 部 健康与人类 服务。国家 卫生研究院 健康。国家 药物研究所 滥用；美国 部 卫生与 公众服务部。 食品药品 监督管理局。 烟草制品 中心
II	在……的阴影之下 蘑菇云的阴影下》： 核试验、 放射性尘埃，以及 对美国 农业的损害，1945年至 1970	2019年	Keith Meyers	10.1017/s002205071800075x	38	3	已收录于 参考文献	示范用途 ICPSR数据的 引用的数字对象标识符。标题 已存在于 ICPSR 书目 （重复记录）。	14 经济学； 1402 应用 经济学	美国 农业数据， 1840—2012； 历史， 人口统计、 经济和 社会数据： 美国， 1790—2002	Haines, Michael; Fishback, Price; Rhode, Paul; Haines, Michael R.; 跨大学 联盟 政治与社会 研究
II	治疗效应估计的推广 来自 样本到总体：一项 关于……的案例研究 寻找的困难 充分数据	2016	Elizabeth A. Stuart，Anna Rhodes	10.1177/0193841x16660663	32	33	已在 参考文献	示范用途 的ICPSR数据 引用数字对象标识符。标题 已经在 ICPSR 参考文献 （重复记录）。	16 人类社会研究 人类社会；13 教育	启蒙计划影响 研究（HSIS）， 2002-2006 [美国 州]	美国 部 卫生与公众 服务。 儿童事务管理局 儿童事务管理局 家庭事务局。规划、研究与评估办公室 规划、 研究与评估办公室 评估

阶段	标题	年份	和作者	DOI	页码	和引用	筛选标准	标准定义		出版物研究领域 研究	ICPSR研究 已引用	ICPSR研究 作者
II	患病率和 预测因素 患者报告结局 长期心理与 身体健康 捐献后的 成人间 活体供肝 移植队列 研究	2018	Mary Amanda Dew，Zeeshan Butt，Qian Liu，Mary Ann Simpson，Jarcy Zee, Daniela P. Ladner, Susan Holtzman, Abigail R. Smith, Elizabeth A. Pomfret, Robert M. Merion, Brenda W. Gillespie, Averell H. Sherker, Robert A. Fisher, Kim M. Olthoff, James R. Burton 小, Norah A. Terrault, Alyson N. Fox, Andrea F. DiMartini	10.1097/tp.0000000000001942	14	25	已在 参考文献	示范用途 ICPSR数据 引用数字对象标识符。标题 已经在 ICPSR 参考文献 （重复记录）。	1117 公共卫生 与健康 服务；11 医学与健康 科学	协作 精神病学 流行病学 调查（CPES）， 2001-2003 [美国 州]	Margarita Alegria；Jackson, James S. (James Sidney)；Ronald C. Kessler C.；David Takeuchi，David
II	复合 劣势 经济衰退 城市和 社区层面 老年美国人的 抑郁症状	2021年	Jason Settels	10.1177/1535684120980992	29	0	已在 参考文献	示范用途 ICPSR数据 引用数字对象标识符。标题 已在 ICPSR 参考文献 （重复记录）。	12 建成 环境与 设计；16 人类社会研究 在人类社会； 1604 人文 地理学；1605 政策与 行政管理′； 1205 城市与 区域规划	国家社会 生活、健康与 老龄化项目 （NSHAP）：轮次 1, [美国], 2005-2006； 国家社会 生活、健康与 老龄化项目 （NSHAP）：第二轮 和伴侣数据 收集，[美国 州], 2010-2011年	Waite，Linda J.；劳曼，爱德华 O.；莱文森，温迪·S.；林道，斯泰西·苔丝勒；奥穆尔切尔泰，科尔姆·A.；韦特，Linda J.; Cagney，Kathleen A.; Dale，William；Huang，Elbert；Laumann，Edward O.；McClintock，Martha K.；O'Muircheartaigh，Colm A.；Schumm，L. Phillip；Cornwell，Benjamin
II	纵向稳定， 基于脑的预测 模型解释了 儿童期智力与 儿童期智力 社会人口学、 心理和 遗传因素	2021	Narun Pornpattananangkula, Richard Anneyb, Lucy Riglinb, Anita Thaparb	10.1101/2021年.02.21.432130	37	2	已收集	示范用途 ICPSR数据 引用数字对象标识符。 出版物标题 未收录于 ICPSR 书目（新 记录）。	08 信息与 计算 科学；17 心理学与 认知 科学；0801 人工 智能与 图像处理； 1701 心理学	统一犯罪 报告计划 数据： 县级 详细逮捕 与犯罪数据， 美国， 2010	美国 部 司法项目办公室 司法项目办公室。 联邦调查局 联邦调查局

阶段	标题	年份	作者	DOI	页码	引用	筛选标准	标准定义		出版物 研究领域	ICPSR研究 已参考	ICPSR研究 作者
II	感知到的种族 歧视、种族 怨恨和支持 平权行动 和优先招聘 和晋升：一项 多种族分析	2021	Maruice Mangum， Ray Block Jr.	10.1080/21565503.2021.1892781	24	1	已收集	示范用途 ICPSR数据 引用数字对象标识符。 出版物标题 不在 ICPSR 参考文献（新记录）。		21 历史与 考古学；2103 历史研究	国家政治 研究，2008年	Jackson, James S.（James Sidney）； Hutchings, Vincent L.; Wong, Cara; Brown, Ronald
II	转介样本中 认知能力 分测验分数的剖面信度 样本	2019	Kara M. Styck, A. Alexander Beaujean, Marley W. Watkins	10.1037/arc0000064	10	5	已收集	示范用途 ICPSR数据 引用数字对象标识符。 出版物标题 不在 ICPSR 参考文献（新记录）。		17 心理学与 认知 科学；1701 心理学	剖面信度 认知能力 分测验分数的 转介样本； 国家调查：关于 药物使用与 健康，2014年	Styck, Kara M.; Beaujean, A. Alexander; Watkins, Marley W.; 美国 卫生与公众服务部 卫生与公众服务部 服务。国家 美国国立卫生研究院 健康。尤妮丝 肯尼迪·施莱佛 国家研究所 儿童健康与 人类 发展
II	时间稳定性 内群体偏袒 主要归因于 遗传因素	2017	Gary J. Lewis， Timothy C. Bates	10.1177/1948550617699250	22	6	篇已收录	示范用途 的 ICPSR数据 引用数字对象标识符。 出版物标题 不在 ICPSR 参考文献（新增记录）。		17 心理学与 认知 科学；1701 心理学	美国中年期研究 美国 (MIDUS 2)， 2004—2006年	Ryff, Carol D.； Almeida, David M.； Ayanian, John Z.； Carr, Deborah S.； Cleary, Paul D.； Coe, Christopher； Davidson, Richard J.; Krueger, Robert F.; Lachman， Marge E.; Marks， Nadine F.； Mroczek，Daniel K.; Seeman， Teresa E.; Seltzer， Marsha Mailick； Singer, Burton H.； Sloan, Richard P.； Tun, Patricia Ann； Weinstein, Maxine； Williams, David R.

阶段	标题	年份	和作者	DOI	页码	和引用	筛选标准	标准定义		出版物研究领域的研究	ICPSR研究已引用	ICPSR研究作者
II	障碍和促进因素 成功过渡的 从长期 住院式药物滥用 治疗	2016	Jennifer I. Manuel， Yeqing Yuan， Daniel B. Herman， Dace S. Svikis， Obie Nichols，Erin Palmer，Sherry Deren	10.1016/j.jsat.2016.12.001	7	33	数据提及	已致谢 a 简短说明（例如，某项 统计数据），附有数字对象标识符， 但数据未在 出版物中进行 分析。		17 心理学与 认知 科学；1701 心理学	全国药物使用与 健康调查 健康，2011年	美国 美国部 健康与公众 服务。 药物滥用 与心理健康 服务 管理局。 行为健康统计 与质量中心 统计数据与 质量
II	考察风险 频繁使用可卡因： 聚焦非裔 美国治疗人群 总体	2016	Tamika Chere Barkley Zapolski， Patrick Baldwin & Carl W. Lejuez	10.3109/10826084.2016.1155618	11	6	数据提及	已致谢 a 简要要点（例如， 统计数据），带有数字对象标识符， 但数据并未 在该 出版物。		11 医学与 健康科学； 1117 公共卫生 与健康 服务	国家药物使用调查 药物使用与 健康，2013年	美国 美国部 健康与人类 服务。 药物滥用 与心理健康 服务 管理局。 中心 行为健康 统计数据与 质量
II	Anzansi家庭项目： 一项针对……的研究方案 组合干预 应对 发育和 健康结果 面临风险的青春期女孩 无陪伴 迁移	2020	Ozge Sensoy Bahar，Fred M. Ssewamala， Abdallah Ibrahim， Alice Boateng， Proscovia Nabunya，Torsten B. Neilands， Emmanuel Asampong，Mary M. McKay	10.1186/s40814-020-00737-4	12	2	数据提及	致谢 a 简要信息（例如， 带有数字对象标识符的统计数据）， 但数据并未 在该 出版物。		17 心理学与 认知 科学；11 医学与健康 科学；1701 心理学；1117 公共卫生与 卫生服务	青少年 发展 研究，1988-2011 [圣保罗， 明尼苏达州]	Jeylan T. Mortimer
II	保护青少年 免受低收入和 中等收入 国家的 人际暴力侵害 （PRO YOUTH 试验）： 一项研究的研究方案 整群随机 对照试验，针对 强化家庭 计划 10-14 （“强大家庭”）在 巴拿马	2018	Anilena Mejia， Richard Emsley， Eleonora Fichera， Wadih Maalouf， Jeremy Segrott， Rachel Calam	10.1186/s13063-018-2698-0	13	4	工具 提及	已致谢 a 研究工具 或来自……的问题 带有……的工具 一个数字 对象标识 符。		11 医学与 健康科学； 1117 公共卫生 与健康 服务	全国青年 调查 [美国 州]：第七波， 1987	Elliott, Delbert S.

阶段	标题	年份	和作者	DOI	页码	和引用	筛选标准	标准定义		出版物研究领域的研究	ICPSR研究已参考	ICPSR研究作者
II	考察队列 效应在发展性 物质使用轨迹中的 use	2017	Alison Reimuller Burns, Andrea M. Hussong, Jessica M. Solis, Patrick J. Curran, James S. McGinley, Daniel J. Bauer, Laurie Chassin, Robert A. Zucker	10.1177/0165025416651734	11	7	工具 提及	注明了一项 研究工具 或来自某个工具的问题 带有 一个数字 对象标识 符。	17 心理学与 认知 科学；1701 心理学；1702 认知科学	监测 未来：A 青年生活方式与价值观持续研究 的生活方式 与价值观 青少年，1978年	杰拉尔德·G·巴赫曼 ；约翰斯顿， 劳埃德，D.； Patrick M.
II	日常人际交往 与 幸福感 ：作用 情绪调节的 策略	2020	Kira S. Birditt， Courtney A. Polenick，Gloria Luong，Susan T. Charles, Karen L. Fingerman	10.1037/pag0000416	30	15	工具 提及	已致谢 a 研究工具 或来自……的一个问题 一件带有……的工具 数字对象标识符。	17 心理学与 认知 科学；1701 心理学	家庭交流 研究第1波， 费城， 宾夕法尼亚州， 2008	Karen Fingerman L.
II	学校不平等与 城市福利：利用 数据科学超越 社会经济地位 数据科学	2019	Renato P. dos Santos, Şahin Bülbül, Isadora L. Lemes	10.17648/acta.scientiae.5494	26	1	方法论 提及	使用数据 DOI 来 致谢其 研究的 方法论或 设计。	16 人类社会研究 人类社会； 1608 社会学	平等的 教育 机会 （COLEMAN）研究 （EEOS），1966	Coleman, James S.
II	……之间的关系 失业 与意大利的生育率：一 时间序列分析	2016	Alberto Cazzola， Lucia Pasquini， Aurora Angeli	10.4054/demres.2016.34.1	40	43	方法论 提及	使用数据 DOI 来 致谢一项 研究的 方法论或 设计。	16 人类社会研究中的 人类社会； 1603 人口学	解决 生态 推断问题： 重构 个体行为 从总体 Data	King, Gary
II	社会生态 功能与 脆弱性分析框架 用于分析森林政策 改革	2012	Fanny Rives， Martine Antona， Sigrid Aubert	10.5751/es-05182-170421	18	8	方法论 提及	使用数据 DOI 来 致谢一项 研究的 方法论或 设计。	41 环境 科学；4101 气候变化 影响与 适应	议程， 替代方案，以及 公共政策， 1976年、1977年、1978年， 1979 [美国 州]	John W. Kingdon
II	干旱 城市的健康保护：相关考量 政策制定者	2020	Howard Frumkin， Maitreyi Bordia Das, Maya Negev, Briony C Rogers, Roberto Bertollini, Carlos Dora, Sonalde Desai	10.1136/bmj.m2936	8	5	来源提及	关联至某个 数据来源 来自另一项研究 或关联至某个 被提及但 建议但不热在线精品 分析的数据源。	11 医学与 健康科学； 1117 公共卫生 和健康 服务；1103 临床科学	印度人类 发展 调查小组 （IHDS、IHDS-II）， 2005年、2011—2012年	Sonalde Desai； Reeve Vanneman； 国家经济应用委员会 应用经济学 研究，新 德里
II	当代视觉艺术中的激光物理学 当代视觉 艺术：研究 方案	2016	Diaa Ahmed Mohamed Ahmedien	10.3897/rio.2.e11150	11	1	来源提及	数字对象标识符关联至一个 数据来源 来自另一项研究 或与某个数据源关联的数字对象标识符 该数据源 被提及但未被 分析。	19 创意艺术与写作研究 创意艺术与 写作；1902 电影、 电视与 数字媒体	报道艺术 II [2003]	Szanto, Andras

阶段	标题	年份	作者	DOI	页码	引用	筛选标准	标准定义		出版物研究领域的研究	ICPSR研究已参考	ICPSR研究作者
II	法律学位对少数族裔有价值吗？	2017	Frank McIntyre，Michael Simkovic	10.1016/j.irle.2017.09.004	57	3	来源提及	数字对象标识符关联到一个数据来源来自另一项研究或与某个数据源关联的数字对象标识符该数据源虽被建议使用，但未进行分析。		18 法律与法学研究；1801 法律	After the JD，第3波：职业转型纵向研究，2012—2013年，美国	Robert Nelson; Ronit Dinovitzer; Gabriele Plickert; Joyce Sterling; Bryant G. Garth
II	距离、规模和动荡：南北地中海互动	2014	J. Patrick Rhamey，William R. Thompson，Thomas J. Volgy	10.4000/cdlm.7840	21	3	已收集	示范用途ICPSR数据引用数字对象标识符。	出版物标题不在ICPSR参考文献（新记录）。	16 项研究人类社会；1606 政治学《科学》	冲突与和平数据库(COPDAB)，1948—1978	Edward E. Azar
II	福岛的影响在德国？变化情况媒体报道与公众对……的舆论核能	2015	Dorothee Arlt, Jens Wolling	10.1177/0963662515589276	16	37	数据提及	致谢 a 简短要点（例如，带有数字对象标识符的统计数据），但数据未被分析于该出版物中。		13 教育；创意艺术与写作研究艺术与写作；22 哲学与宗教研究；1302 课程与教学论；2202 特定领域的历史与哲学特定领域；1903年 新闻业与专业写作	全球速访民意调查关于日本海啸及其影响对核能看法的调查，2011	盖洛普国际有限公司
II	评估短期影响媒体知情家长项目一项网络项目面向家长，旨在青少年性健康促进	2019	Tracy M. Scull，Christina V. Malik，Elyse M. Keefe，Alexander Schoemann	10.1007/s10964-019-01077-0	21	17	工具提及	已承认研究工具或一个问题来自一项工具数字对象标识符。		17 心理学与认知科学；1701 心理学	爱荷华州青年与家庭项目，1989-1992	Rand Conger；Lasley, Paul；Lorenz, Frederick O.; Simons, Ronald; Whitbeck, Les B.; Elder Jr., Glen H.; Norem, Rosalie
II	是否为自我报告能够测量指标的将 Individuals 定义为身体活动还是不活动？	2015	约斯坦 Steene-Johannessen, Sigmund A. Anderssen, Hidde P. van der Ploeg, Ingrid J.M. Hendriksen, Alan E. Donnelly, Søren Brage6，以及 Ulf Ekelund	10.1249/mss.0000000000000760	35	154	方法论提及	使用数据 DOI 来致谢研究的方法论或设计。		11 医学与健康科学；1117 公共卫生与健康服务	欧洲晴雨表64.3：外国语言，生物技术，有组织犯罪，和健康用品，11月至12月月 2005年	Papacostas, Antonis

阶段	标题	年份	和作者	DOI	页码和	引用	筛选标准	标准定义		出版物研究领域的研究	ICPSR研究已引用	ICPSR研究作者
II	时变效应 收入对海马的体积轨迹的影响 青春期女孩	2017	Monica E. Ellwood-Lowea ，Kathryn L. Humphreysa，Sarah J. Ordaza，M.Catalina Camachoa, Matthew D. Sacchetb, Ian H. Gotliba	10.1016/j.dcn.2017.12.005	10	34	来源提及	与一个 数据来源 来自另一项研究 或与一个 数据源关联的数字对象标识符 建议但未分析。		17 心理学与认知科学；1701 心理学	教育机会 教育 平等 （COLEMAN）研究（EEOS），1966	Coleman, James S.

附录 B（补充文件 2）：编码手册

父级代码	第一级子代码	第二级子代码	定义	注释	示例
数据实体			表示记录的一个或多个词 观测值	首先识别并编码所有相关的数据实体 在论文中。若抽样框中提供了研究名称和研究 作者姓名，也请一并查找。一个 或多个作为数据专有名词的词语 产品（例如，调查的名称）或关键词 用作代称（例如，“数据”、“文件”）。	
	File		对观测值进行编码的项目	搜索术语“文件”及其同义词：“数据”。 排除“数据收集”或“摘要”等短语 关于数据的讨论。	我们的分析使用了成人 *W2* 和 *W3* 限制使用文件（*https://doi.* *org/10.3886/ICPSR36231.v21* （访问日期：*2021*年*2*月*17*日）。
	元数据		描述一组观测值的项目	搜索术语“元数据”及其同义词： “编码手册”、“文档”、“用户指南”。	*PATH*研究设计与方法 *[6,16,17]*，访谈程序、 问卷、抽样、加权、 以及响应率均载于*PATH* 研究限制使用文件用户指南。
	研究		指包含一组观测值或 代替一组观测值的括号内陈述 （作者，年份）	搜索术语“study”及其同义词： “数据集”、“序列”、“调查”、“波次”。排除 诸如“本研究”或“案例研究”之类的参考。	烟草与健康人口评估 （*PATH*）研究是一项持续进行的， 具有全国代表性的纵向 队列研究，可获取详细的 美国烟草使用信息 总体。
	变量		表示一个维度的项目 观测值	搜索术语“变量”及其同义词： “指标”和“权重”。排除提及 抽象意义上的变量的短语，例如“预测变量、 调节变量和结果变量”。	三种喘鸣变量（曾有 喘鸣、过去*12*个月（*P12M*） 喘鸣，以及*P12M*喘鸣发作 频率）合并后构成 一个变量（*0* = 从未喘鸣；*1* = 曾有喘鸣，但过去*12*个月内未出现喘鸣 或过去*12*个月内出现喘鸣，但没有喘鸣发作 发作；*2* = 过去*12*个月内出现喘鸣，且有*1–3*次 发作；*3* = 过去*12*个月内出现喘鸣，且有*4–12*次 发作；*4* = *P12M* 喘鸣，且更多 发作）。
数据引用			提及一个或多个数据 实体的上下文窗口	对包含一个或多个数据的段落进行编码 作为数据引用的实体。每个数据引用 包含一个或多个数据实体。	
特征			数据的结构、形式和呈现方式 文本中的引用	如适用，请分配一个或多个第二级子代码 数据引用。	
	访问		说明数据可以从何处 访问	作者必须提供统一资源定位符/数字对象标识符或联系人 以便访问数据。组织 仅有信息是不够的。	
		提供	将数据分享回研究 共同体	作者创建了一个数据实体，并说明了其所在位置 ，以便他人访问	数据和支持材料 可用于复现数值 论文中结果所必需的内容可在 *JOP Dataverse*（*https://dataverse.* *harvard.edu /dataverse/jop*）中获取。

父级代码	第一级子代码	第二级子代码	定义	注释	示例
		接收	访问来自研究共同体的数据	作者提及现有数据实体，并说明 可在何处访问。如果该参考是 括号内的，则参考文献列表中的相应条目 需要提供访问机制。仅说明某个 个人或机构并不足够，如果访问 未提供机制。	*Elliot, Patrick. 2007*年。“三代合并 ，*1965–1997*年。”校际大学 政治与社会研究联盟 （*ICPSR*）。*http://doi.org/10. 3886/ICPSR04532.v1*。
	行动		作者与数据之间的互动	确定数据引用的接近程度 表明作者与数据产品的关系 产品	
		引用	表面上指向数据，但未命名 或提供背景。给出了研究发现，但未进行 分析。	括号内的引文提及数据，但该 陈述未提及数据实体。没有使用动词 来描述对数据采取的行动。	据估计，有*2150*万人患有 可诊断的物质使用障碍 （物质使用障碍），占美国 总体的*9%*，另有约*4*万人 从事被认为对医学有害的滥用行为（ *McLellan* 与 *Woodworth*，*2014*年；物质滥用 与精神卫生服务 管理局，*2015*年）。
		提及	提及、描述或暗指数据	对数据实体的参考具有描述性	在许多年中，美国国家选举研究都包含 “情感温度计”问题，要求 受访者评估自己对 政治人物和群体的感受 （包括种族或族裔群体），并按照 *0* 至 *100* 的量表进行评分。
		Uses	表明积极、亲自参与的数据处理工作	对数据采取的行动显而易见。使用动词以及 与数据实体相关的“我们”表述。	对于*2010*年合作国会选举研究（*CCES*），当两个 问题都被提出时，我们对两个问题均进行了重新量表化处理， 并将它们取平均值以创建 一个测量指标。
	位置		数据引用出现的位置	确定数据引用所在论文的 IMRAD结构位置 数据引用出现的位置	
		摘要	在出版物摘要中	在简要概述 论文	……的逻辑回归估计 半年期亲密伴侣暴力患病率采用 广义估计……进行建模 方程和稳健标准误来 弥补重复测量和 *NCVS* 的复杂抽样设计 。
		致谢	在致谢部分	在致谢声明或致谢部分中， 其中可能包含有关 资金支持、首席研	我们也感谢 *Michael Haines*， *Eitan Hersh* 和 *Heather O’Connell* 分享 数据给我们。
		附录	在附录部分或补充材料中	究员的信息 在论文作者称为 附录、补充材料等的部分	*1955*年美国家庭增长研究 。该研究以*18*至*39*岁的人群为样本， 岁的白人女性，她们 目前已婚。关于 受访者是否曾使用过 *1955*年美国家庭增长研究中的避孕方式， 相关信息取自两个 问题。

父级代码	第一级子代码	第二级子代码	定义	注释	示例
		图注、图、表	在表或图的图注中	在表、图或其图注中。优先于其他 章节信息（例如，代码脚注而非 脚注和结论）。为提及数据实体的整个表或 图编码。	注：此处的不平等通过以下指标测量 白人和黑人 县内收入中位数之比的对数 *2014*年。数据来自美国社区 调查，*2009—2014*年。
		结论	在结论部分	在论文末尾提供摘要或结论的部分 论文末尾	然而，如果这种区分仅基于 通常用于划分 城市、郊区和农村 聚居地的 *NCVS* 分析， 那么，这项研究的结果表明， 所谓的城乡差异 不过是一个迷思，我们更应该 把精力放在理解郊区与 非郊区在风险方面的差异上。 *IPV.*
		数据可用性	在数据可用性声明中	在文章的独立字段中。可能由 期刊要求，并包含访问信息。	数据可用性声明：以下数据来自 *PATH*研究第*1*波至第*3*波， 可作为受限数据下载 使用文件（*https://www.icpsr.umich.edu/icpsrweb/NAHDAP/studies/36231*）。 使用申请指南和条件 可在上述网站获取。
		讨论	在结果或讨论部分	在描述或解读研究发现的部分	这些局限性通过以下方式得到处理 将*PATH*研究的数据与 美国国家健康与营养调查的数据进行比较，并使用已有的问题 这些问题已验证可用于捕捉其他人群中的呼吸系统 疾病。
		脚注	在脚注中	在脚注或尾注中。覆盖其他章节 信息（例如，代码脚注而非脚注 和结论）。	请参阅 *IPUMS* 文档：*https //usa.ipums.org/* 以获取完整的 这些人口普查指标的说明。
		资金支持	在资金支持信息中	作者提供了有关研究资助方和/或 资助详情	作者声明已收到 以下资金支持，用于 本文章的研究、作者工作和/或出版 ：部分经费由以下项目提供： 国家科学基金奖项 *DRL- 1335843* 美国国家科学基金会的 *DRL-1335843* 号资助（共同首席研究员 *Stuart* 和 *Olsen*）以及 *R305D150003* 号资助 来自教育科学研究所 、美国教育部 教育（共同首席研究员 *Stuart* 和 *Olsen*）。

父级代码	第一级子代码	第二级子代码	定义	注释	示例
		引言	在引言或背景部分	在提供引言或 论文开头的背景介绍中	在本文中，我将考察来自 全国犯罪受害调查 （*NCVS*）的数据，以估计风险的变化 不同聚落类型中的亲密伴侣暴力，使用一种 能够区分孤立 农村聚落与人口更多的 非大都市聚落 尽管这些聚落通常被编码为农村 尽管其居住人口相对较多 密度。
		关键词	在关键词中	将整个关键词块编码，包括“数据” 实体，作为数据引用	关键词：亲密伴侣暴力， 患病率、*NCVS*、聚落类型
		方法	在材料与方法或数据部分	在描述材料与方法的部分	本研究所使用的数据来自 n = *578,471* 名年龄在 *18* 岁及以上的女性 共访谈了 n = *1,672,999* 次，时间跨度为*1994*年到*2015*年期间的 *NCVS* （美国司法统计局，*2016*年）。
	样式		说明作者如何命名 数据引用中的实体		
		首字母缩略词	使用首字母缩略词指代数据实体	仅以首字母缩略词给出数据实体	本节详细介绍 *REDI* 的使用情况 以及 *HSIS* 数据，以评估 *REDI* 项目的可推广性 全国符合启蒙计划资格的儿童， 由 *HSIS* 代表。
		通用项	使用通用名词指代数据实体， 例如“数据”“研究”“波次”“变量”	数据实体使用小写形式非正式地指定 词语且不含版本信息	数据由以下机构提供： 政治与社会研究校际联盟 和社会研究，密歇根州安娜堡， 美国。原始数据的收集者 和联盟均不承担任何 对分析或 此处所呈现的解释负责。
		Name	使用数据实体的正式或非正式名称	数据实体使用明确无歧义的名称，可能 也可能不包含具体的版本信息（例如 使用了哪个波次）	研究了早于 *After the* *JD* 的队列后，全国律师资格考试通过情况纵向 研究发现，长期律师资格考试 通过率明显较低 少数族裔低于白人。
		括号内的	使用作者和年份链接到参考 文献列表，以指代数据实体；也可以 采用尾注、脚注式指示符或 APA格式 引用	数据实体未在正文中提及，而是通过 括号标注，并列于引用文献中	为了检验放射性沉降物是否 对农业部门产生了持续 对农业部门的影响，我 根据可比变量构建了面板数据 取自《美国农业史》 中*1940*年至*1997*年的人口普查 *Haines*等人（*2015*年）。
	Type		作者与数据的关系	根据背景判断作者是否 创建了数据，或使用了他人创建的数据	

父级代码	第一级子代码	第二级子代码	定义	注释	示例
		派生数据	从现有数据创建新数据	通过分析或组合现有数据派生的新数据 作者可能会分享这些数据。合并或 应用函数会创建派生数据，但 子集选取不会。合并数据的嵌套层级 文件（例如：从个人层级到家庭层级）或多个年份 纵向数据的推导不会被视为 推导	我们将在下一节讨论我们的数据 并呈现将以下内容联系起来的核心结果： *1860*年奴隶制的患病率与 当代态度，具体内容如下 部分，并附有额外的稳健性 检验，详见附录， 可在线获取。
		主要	体现数据的创建	作者自行收集的数据。这 涵盖作者正在解读但 可能尚未发表或未被明确命名的原始数据	作者已提供 支撑分析的数据 本文所呈现的内容（见 *Styck*， *Beaujean* 和 *Watkins*，*2019*年），因此 以便合格人员进行复现，并可能 拓展这项工作。 研究人员。后续使用者有义务 让数据创建者参与其中，融入他们的 发表计划中征求数据创建者的参与，前提是创建者 有此意愿。
		二次使用	使用他人创建的数据或已提供 给他人使用的数据（例如，存档数据）	作者归属于另一位创建者的数据 或称为存档数据的情况。这涵盖了以下情形： 作者重复使用自己的数据，但 将其称为已发表成果	对于所有这些变量，我们将 *1860*年数据映射到现代县 界，使用了以下程序 附录*A*节中描述的程序。
功能			数据引用的用途	为每个第二级子代码选择一个或多个实例 应用于每个数据引用。一个数据引用 可以具有多个修辞功能	
	批评		批评所引用的数据	目的是通过批评来劝说读者 数据产品	
		比较	对比作者的研究 与其他基于这些数据开展的研究，或呈现 不同数据源之间的比较	与先前研究发现相反，得出的结论 来自先前对引用数据的使用可能包括 “不同于”“但是”等术语	*20*世纪*50*年代和 *20*世纪*60*年代的效应量在大小上大致相似， 但方向与其相反， 奴隶制对奥巴马白人 在*2008*年白人选票中的影响，根据*CCES*估算 白人受访者。
		局限性	表明作者意识到并采取了谨慎措施 （例如，限定解释的范围） 与数据	承认潜在错误、缺陷， 质量问题、适用性、隐私——背景 引用数据的权衡取舍、抽样问题。可能 包括“弱点”“缺陷”等词语	首先，*1860*年奴隶数据是历史数据 且可能存在测量误差。
	描述		中性地描述引用数据	目的是向读者提供详细信息，以便 帮助他们理解	
		构成	说明或讨论有关……的知识 数据产品，例如其元数据	描述数据的内容（研究的基础 、其收集方法、总体等）。可能会 使用“包括”“代表”等词语	将样本范围限制为南部地区后 白人，我们有一个美国国家选举研究样本，包含 来自南部*64*个县的*3,123*名个体， 南部。
		来源	数据来源州，从而建立一种 与数据来源地的关联	注明数据生产者／提供者的认可，并描述 其可访问性及来源信息。可能使用 “来自”“由……提供”等表述	我们使用以下数据构建了这些测量指标： 联合国粮食及农业 组织（*FAO*）的数据。
	说明		说明由以下内容支持的观点 引用数据	目的是用实证证据加以说明	

父级代码	第一级子代码	第二级子代码	定义	注释	示例
		背景	背景、研究发现或统计数据取自 引用数据，其中数据代表 作者的观点（即具有转喻性）	数据的参考方式与研究 文章相同（引用整体实体以阐明 观点）	例如，人们可能会关注 *REDI*干预对 所有*4*岁儿童的影响，包括一个更加 社会经济背景多样化的儿童总体， 其社会经济背景比符合启蒙计划 资格的儿童更加多样化；在这种情况下，一项全国性的 具有代表性的儿童数据集，涵盖 美国，例如 *ECLS-B*， 将是更合适的总体 数据集（*http://nces.ed.gov/ecls/birth.asp*）。
		展望	关于数据潜在应用的评论 这些应用并未在 本研究	指出数据“可能有用”，或具有 对某项应用有用的属性	请注意，*HSIS* 使用的是较早版本，但 该测量指标的可比版本， 可用于未来的研究， 进一步诊断 基于研究、 符合发展规律的结果对 *HSIS* 的 可推广 性。
	互动		描述对 引用数据	目的是告诉读者对数据产品做了什么，或 从数据产品派生出的内容	
		解读	基于……提出经验性主张 数据分析	构建理论、进行比较、解读 引用数据中的实证证据，遵循 分析描述，和/或伴随 图。可能包括“显示”“发现”等词语	相反，如图*4*所示，数据 显示区域 差异在重建时期结束后出现。 重建时期。
		处理	描述了处理以下内容时所执行的步骤 data	选择变量、文件、波次和工具 引用数据；准备、转换、 修改、抽样、比较、相关分析、 验证、整合、扩展、更新对引用数据的分析 可能包括“子集”、 “样本”等术语	我们将这些数据筛选为前者 邦联州，以及密苏里州和 肯塔基州，这两个州都曾有相当多的 人支持美利坚联盟国，*5* 以及自我认定为白人的人，因此我们 拥有超过*4*万名受访者 覆盖南部 *1,435* 个县中的 *1,329* 个 县。
	使合法化		通过……赋予该研究合法性 引用数据相关联，赋予该研究合法性	目的是通过价值说服读者 关于数据产品的陈述	
		论证	强调数据的某项特征，该特征赋予 作者所作选择以可信度或权威性	通过推理和选择数据来建立信任， 数据的可信度；通过……传递规范性 典型用途、先前研究使用数据； 证实或提供依据、论证、限定 数据选择可能包含“尽管”等词， “具有代表性”	由于县界发生了变化 自*1860*年以来，我们采用面积加权 将数据从*1860*年映射到 人口普查对应到县界 *2000*年，从而使我们能够估算 *1860*年现今的县中被奴役人口比例 现今的县
		透明度	解释为何或如何应用某项分析程序 以表明质量或 分析中考虑的因素	使方法公开，并可能具有可复现性，以便 遵守期刊或资助方的数据共享要求 包括指向附录中数据分析的引用，或 补充材料可能包含术语 "文档"、"提供"	此外，我们还提供了额外的 证据，见附录 *D*，其中我们 分析了当代流动性数据 来自 *2000*年美国人口普查的数据。

参考文献

Abu-Jbara A、Ezra J 和 Radev D（2013年）引文的目的与极性：迈向基于自然语言处理的文献计量学。载于：*2013*年北美计算语言学协会分会会议论文集：人类语言技术，美国佐治亚州亚特兰大，2013年6月，第596–606页。计算语言学协会。可访问：https://aclanthology.org/N13-1067。

Altman M、Borgman C、Crosas M、Matone M（2015年）数据引用联合原则简介。《美国信息科学学会通报 / *ASIS* 41(3)。Wiley：43–45。数字对象标识符：10.1002/bult.2015.1720410313。

Ball A 和 Duke M（2015年）如何引用数据集并关联出版物。数字对象标识符：10.1007/1-4020-5340-1。

Banaeefar H、Burchart S、Moss E 等（2022年）最佳实践可能还不够：使用 DOI 进行数据引用的差异。在国际社会科学信息服务与技术协会年会上展示的海报，2022年6月9日。数字对象标识符：10.7302/4809。

Bauer MW（2000年）经典内容分析：综述。载于：使用文本、图像和声音进行质性研究：实用手册。Sage，第131–151页。数字对象标识符：10.4135/9781849209731。

Belter CW（2014年）衡量研究数据的价值：海洋学数据集的引文分析。《*PLoS One*》9(3)：e92590。数字对象标识符：10.1371/journal.pone.0092590。

Blake C（2010年）超越基因、蛋白质和摘要：从生物医学全文文章中识别科学论断。生物医学信息学杂志43(2)：173–189。数字对象标识符：10.1016/j.jbi.2009.11.001。

Boland K、Ritze D、Eckert K 等人（2012）在出版物中识别数据集参考文献。载于：数字图书馆理论与实践，2012年，第150–161页。施普林格柏林海德堡。数字对象标识符：10.1007/978-3-642-33290-6 17._

Buneman P、Dosso D、Lissandrini M 等人（2022年）数据引用与引文图。定量科学研究2(4)。麻省理工学院出版社期刊：1399–1422。数字对象标识符：10.1162/qss_a_00166。

Case DO 和 Higgins GM（2000）我们如何研究引文行为？一项关于传播领域文献引用原因的研究。《美国信息科学学会杂志》。Wiley在线图书馆。数字对象标识符：10.1002/(SICI)1097-4571(2000)51:7<635::AID-ASI6>3.0.CO;2-H。Charmaz K（2006）《建构扎根理论：定性分析实用指南》。SAGE。Chernin E（1988）“哈佛体系”：谜团已解。《英国医学杂志》297(6655)。BMJ Publishing Group：1062。网址：https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1834803/（访问于2023年2月15日）。

Cohan A、Ammar W、van Zuylen M 等（2019）科学出版物中用于引文意图分类的结构框架。载于：*2019*年北方会议论文集，美国宾夕法尼亚州斯特劳兹堡，2019年。计算语言学协会。数字对象标识符：10.18653/v1/n19-1361。Coombs CH（1964年）数据理论。英格兰牛津：Wiley《数据理论》。网址：https://psycnet.apa.org/fulltext/1965-00053-000.pdf。Cousijn H、Feeney P、Lowenberg D 等（2019年）整合引文与使用指标，使数据发挥价值。《数据科学期刊》18。Ubiquity Press, Ltd. 数字对象标识符：10.5334/dsj-2019-009。Cragin MH、Palmer CL、Carlson JR 等（2010年）数据共享、小科学与机构知识库。《皇家学会哲学汇刊*A*辑：数学、物理与工程科学》368(1926)。英国皇家学会出版部：4023–4038。数字对象标识符：10.1098/rsta.2010.0165。Cronin B（1981年）引文理论的必要性。《文献学期刊：专业知识的记录、组织与传播》37(1)。Emerald：16–24。数字对象标识符：10.1108/eb026703。Cronin B（1984年）《引文过程：引文在科学传播中的作用与意义》。Taylor Graham。Ding Y、Zhang G、Chambers T 等（2014年）基于内容的引文分析：新一代引文分析。《信息科学与技术协会期刊》65(9)。Wiley：1820–1833。数字对象标识符：10.1002/asi.23256。Edge D（1979年）科学传播的定量指标：批判性综述。《科学史：文献、研究与教学年度评论》17(36 Pt 2)：102–134。数字对象标识符：10.1177/007327537901700202。Egghe L（2010年）赫希指数及相关影响力指标。《信息科学与技术年度评论》44(1)：65–114。数字对象标识符：10.1002/aris.2010.1440440109。Fear KM（2013年）《测量并预测数据再利用的影响》。网址：https://deepblue.lib.umich.edu/handle/2027.42/102481。Fenner M、Crosas M、Grethe J S 等（2019年）学术数据知识库的数据引文路线图。《科学数据》6(1)：28。数字对象标识符：10.1038/s41597-019-0031-8。Frumkin H、Das MB、Negev M 等（2020年）干旱城市的健康保护：政策制定者的考量。《英国医学杂志》371：m2936。数字对象标识符：10.1136/bmj.m2936。Furner J（2016年）“数据”：数据。载于：Kelly M 和 Bielby J（编）《数字时代的信息文化：献给拉斐尔·卡普罗的纪念文集》。威斯巴登：Springer Fachmedien Wiesbaden，第287–306页。数字对象标识符：10.1007/978-3-658-14681-8_17。Garfield E（1964年）“科学引文索引”——索引编制的新维度。《科学》144(3619)。美国科学促进会（AAAS）：649–654。数字对象标识符：10.1126/science.144.3619.649。Garfield E（1979

) 引文分析是　y否　g合法的评价工具？《科学计量学》1　(4). Springer

施普林格科学与商业媒体有限责任公司：359–375。数字对象标识符：10.1007/bf02019306。Garvey WD 和 Griffith BC（1972）科学学科内部的交流与信息处理：心理学的实证研究发现。《信息存储与检索》8(3)：123–136。数字对象标识符：10.1016/0020-0271(72)90041-1。Gilbert GN 和 Woolgar S（1974）书评：科学的定量研究：文献考察。科学研究4(3)。SAGE 出版社：279–294。数字对象标识符：10.1177/030631277400400305。Gregory K、Groth P、Scharnhorst A 等（2020）失而复得？发现研究所需的数据。《哈佛数据科学评论》。数字对象标识符：10.1162/99608f92.e38165eb。Hernández-Alvarez M 和 Gomez JM（2016）引文语境分析调查：任务、技术与资源。《自然语言工程》22(3)：327–349。数字对象标识符：10.1017/S1351324915000388。Hey T、Tansley S 和 Tolle K（2009）第四范式：数据密集型科学发现。华盛顿州雷德蒙德：微软研究院。Holsti OR（1969）社会科学与人文学科内容分析。马萨诸塞州雷丁：艾迪生-韦斯利出版社。Hook DW、Porter SJ 和 Herzog C（2018）Dimensions：构建搜索与评价的背景。《研究计量与分析前沿》3。数字对象标识符：10.3389/frma.2018.00023。Jiao C 和 Darch PT（2020）数据论文在学术交流中的作用。《信息科学与技术协会会议录》57(1)。Wiley。数字对象标识符：10.1002/pra2.316。Jurgens D、Kumar S、Hoover R 等（2018）通过引文框架衡量科学领域的演变。《计算语言学协会汇刊》6。麻省理工学院出版社：391–406。数字对象标识符：10.1162/tacl a 00028/43437。_ _ Kozlowski D、Larivière V、Sugimoto CR 等（2022）科学中的交叉性不平等。《美国国家科学院院刊》119(2)。数字对象标识符：10.1073/pnas.2113067119。Kwon D（2022）引文正义的兴起：学者如何让参考文献更加公平。《自然》603(7902)：568–571。数字对象标识符：10.1038/d41586-022-00793-1。Lewis GJ 和 Bates TC（2017）内群体偏袒的时间稳定性主要归因于遗传因素。《社会心理与人格科学》8(8)：897–903。数字对象标识符：10.1177/1948550617699250。Leydesdorff L（1998）引文理论？《科学计量学》43(1)。施普林格科学与商业媒体有限责任公司：5–25。数字对象标识符：10.1007/bf02458391。Li K 和 Jiao C（2022）作为社会语言学认识论对象的数据论文：对数据论文摘要中所用修辞策略的内容分析。《信息科学与技术协会期刊》73(6)。Wiley：834–846。数字对象标识符：10.1002/asi.24585。Liu M

研究。《文献学杂志》49(4)。MCB UP Ltd：370–408。数字对象标识符：10.1108/eb026920。Lyu D、Ruan X、Xie J 等（2021）引用动机分类：一项元综合。《科学计量学》126(4)：3243–3264。数字对象标识符：10.1007/s11192-021-03908-z。Mayernik MS、Hart DL、Maull KE 等（2017）评估和追踪科研基础设施的成果与影响。《信息科学与技术协会期刊》68(6)。Wiley：1341–1359。数字对象标识符：10.1002/asi.23721。Mayo C、Vision TJ 和 Hull EA（2016）引文的位置：出版物引用 Dryad数字知识库中原始数据的实践变化。11(1)：150–155。数字对象标识符：10.2218/ijdc.v11i1.400。McIntyre F 和 Simkovic M（2018）法律学位对少数族裔而言是否同样有价值？《国际法律与经济学评论》53：23–37。数字对象标识符：10.1016/j.irle.2017.09.004。Merton RK（1968）科学中的马太效应。探讨科学的奖励与交流体系。《科学》159(3810)：56–63。可访问：https://www.ncbi.nlm.nih.gov/pubmed/5634379。

Meyers K（2019）《蘑菇云的阴影下：核试验、放射性尘埃与对美国农业的损害，1945年至1970年》。《经济史杂志》79(1)。剑桥大学出版社：244–274。数字对象标识符：10.1017/S002205071800075X。

Mooney H 和 Newton MP（2012）《数据引用剖析：发现、再利用与认可》。《图书馆学与信息科学杂志》。数字对象标识符：10.7710/2162-3309.1035。

Moss E 和 Lyle J（2018）《不透明数据引用：实际引文实践及其对追踪数据使用的影响》。获取地址：https://deepblue.lib.umich.edu/handle/2027.42/142393。

Nakov PI、Schwartz AS、Hearst M 等（2004）《引文句：用于生物科学文本语义分析的引文句》。载于：《SIGIR会议论文集》，2004年，第81–88页。Citeseer。数字对象标识符：10.1.1.59.2666。

Park H、You S 和 Wolfram D（2018）在生物医学领域，为数据共享和再利用而进行的非正式数据引用比正式数据引文更常见。《信息科学与技术协会期刊》69(11)。Wiley：1346–1354。数字对象标识符：10.1002/asi.24049。

Pasquetto IV、Borgman CL 和 Wofford MF（2019）科学数据的使用与再利用：数据 a
创作者的优势。*1.2*1(2)。麻省理工学院出版社 - 期刊。数字对象标识符：10.1162/99608f92.fc14bf2d。
Renear AH、Sacchi S 和 Wickett KM（2010）科学与技术领域中数据集的定义
al
文献。美国信息科学与技术学会会议录47(1)。Wiley：1–4。数字对象标识符：10.1002/meet.14504701240。Settels J（2021）城市和社区层面的经济衰退与美国老年人抑郁症状之间的复合劣势。《城市与社区》20(3)。SAGE 出版社：260–288。数字对象标识符：10.1177/1535684120980992。Shotton D（2010）CiTO，即引文类型本体。《生物医学语义学杂志》1 Suppl 1：S6。数字对象标识符：10.1186/2041-1480-1-S1-S6。Silvello G（2018）数据引用的理论与实践。《美国信息科学与技术学会杂志

n

《科学》。Wiley在线图书馆。数字对象标识符：10.1002/asi.23917。

Small HG（1978年）作为概念符号的被引文献。《科学社会研究》8(3)。SAG E
出版社有限公司：327–340。数字对象标识符：10.1177/030631277800800305。
Smith B（2014年）文献行为。《制度、情感与群体施动者》：19–31。数字对象标识符：10.1007/978-94-007-6934-2 2._Sollaci LB 和 Pereira MG（2004年）引言、方法、结果与讨论（IMRAD
）
结构：五十年调查。《医学图书馆协会杂志：JMLA》92(3)：364–367。可访问：https://www.ncbi.nlm.nih.gov/pubmed/15243643。Spiegel-Rosing I（1977年）科学研究：文献计量与内容分析。《科学社会研究》7(1)。SAGE出版有限公司：97–113。数字对象标识符：10.1177/030631277700700111。

Styck KM、Beaujean AA 和 Watkins MW（2019）转介样本中认知能力分测验分数的剖面信度。科学心理学档案7(1)：119–128。数字对象标识符：10.1037/arc0000064。

Teplitskiy M、Duede E、Menietti M 等（2018）为什么我们所知的关于引文的几乎一切都是错误的：来自作者的证据。*STI 2018*。可访问：https://scholarlypublications.universiteitleiden.nl/handle/1887/65227。

Teufel S、Siddharthan A 和 Tidhar D（2006）引文功能的自动分类。载于：*2006* 年自然语言处理实证方法会议论文集，澳大利亚悉尼，2006 年 7 月，第 103–110 页。计算语言学协会。可访问：https://aclanthology.org/W06-1613。

White MD 和 Wang P（1997）《引用行为的定性研究：贡献、标准与元层次文献记录问题》。《图书馆季刊》67(2)。芝加哥大学出版社：122–154。数字对象标识符：10.1086/629929。

Wickett KM、Sacchi S、Dubin D 等（2012）《科学数据中的内容与表示层级识别》。《style id='1'>美国信息科学与技术学会会议录49(1)。Wiley：1–10。数字对象标识符：10.1002/meet.14504901199。

Wynholds L（2011）《链接科学数据：无序且边界模糊的数字对象的身份问题》。《style id='1'>《国际数字保存杂志》6(1)。爱丁堡大学图书馆：214–225。数字对象标识符：10.2218/ijdc.v6i1.183。

Zhao M、Yan E 和 Li K（2018）《数据集提及与引文：对全文出版物的内容分析》。《style id='1'>《信息科学与技术协会期刊》69(1)。Wiley：32–46。数字对象标识符：10.1002/asi.23919。