别再站在巨人的肩膀上构建：找回无法获取的需求质量工件

朱利安 弗拉蒂尼[1], 劳埃德 蒙哥马利[2], 达维德 富奇[1], 扬尼克菲施巴赫[3,4],迈克尔温特 卡尔姆施泰纳[1] 和 丹尼尔门德斯[1,4]

[1]瑞典卡尔斯克鲁纳市瓦尔哈拉韦根*1*号，邮编*371 41*，布莱金厄理工学院[2]德国汉堡市，邮编*20146*，汉堡大学[3]德国慕尼黑市斯特恩大街*5*号，邮编*80538*，*Netlight Consulting GmbH*[4]德国慕尼黑市格里克大街*25*号，邮编*80805*，*fortiss GmbH*

摘要

关于需求质量的文献中有大量出版物介绍各类工件，例如数据集和工具。然而，近期的系统性研究表明，其中超过80%的工件已无法获取，或从未公开，这限制了研究的可复现性和可重用性。在本研究中，我们报告了尝试找回这些工件的工作。为此，我们联系了无法获取的工件的通讯作者，请他们按照开放科学原则找回并公开这些工件。我们的结果基于35位作者中的19份回复（回复率为54%），包括对需求质量工件可获取性的评估，以及作者持续未能提供这些工件的原因分析。总体而言，我们提高了七个数据集和七项实现的可获取性。

关键词

需求质量、开放科学、可用性、制品、数据集

arXiv:2304.04670v1 [cs.SE] 10 Apr 2023

1. 引言

数据集和工具通常被认为是需求质量文献中的重要贡献 [1]。然而，最近的一项次级研究发现，在57项主要研究中，目前公开可用的数据集仅占12%，工具仅占19% [2]。这些制品无法获取会带来两个主要后果。首先，实证结果难以复现，这阻碍了强化科学贡献之实证证据的过程。其次，所呈现的制品难以复用，这使得科学进步不得不一次又一次地重新开始，而不是在现有贡献的基础上不断发展。最终，这些后果阻碍了需求质量研究领域的发展。

Barcelona'23：第六届需求工程自然语言处理研讨会，*2023*年*4*月*17*日，西班牙巴塞罗那
julian.frattini@bth.se (J. Frattini); lloyd.montgomery@uni-hamburg.de (L. Montgomery); davide.fucci@bth.se (D. Fucci); jannik.fischbach@netlight.com (J. Fischbach); michael.unterkalmsteiner@bth.se (M. Unterkalmsteiner); daniel.mendez@bth.se (D. Mendez) https://lloydm.io/ (L. Montgomery); https://dfucci.github.io/ (D. Fucci); https://www.lmsteiner.com/ (M. Unterkalmsteiner); https://www.mendezfe.org/ (D. Mendez) 0000-0003-3995-6125 (J. Frattini); 0000-0002-8249-1418 (L. Montgomery); 0000-0002-0679-4361 (D. Fucci); 0000-0002-4361-6118 (J. Fischbach);0000-0003-4118-0952 (M. Unterkalmsteiner); 0000-0003-0619-6027 (D. Mendez)
© 2023 年本文作者拥有版权。允许[illegible]享署名 4.0 国际许可协议（CC BY 4.0）使用。CEUR 研讨会论文集（CEUR-WS.org）

在软件工程中践行开放科学，有助于提升制品的可获取性并保存未来的贡献 [3]，但事后很难再这样做。因此，过去出版物中制品的可获取性会随时间推移而降低 [1]。由此造成的制品无法获取 [2]，给依赖制品的研究带来了重大挑战，例如需求质量研究，或更广泛的需求工程自然语言处理（NLP4RE）领域。在这项工作中，我们尝试按照开放科学原则请求作者公开制品，以恢复无法获取的需求质量工件。我们的主要贡献是恢复了七个数据集和七个实现。然而，向作者发出的 35 项请求中有 16 项（46%）未获回应，这表明研究界需要进一步强调持续归档研究数据的重要性。

本文其余部分的结构如下：第 2 节介绍开放科学和需求质量研究的背景。第 3 节描述恢复流程，第 4 节呈现恢复结果。我们在第 5节讨论这些结果，并在第 6节总结全文。

2. 背景

2.1. 软件工程中的开放科学

科学工作需要具备可复现性 [3]，以增强其对某一研究领域所贡献的证据 [4]。开放科学 是一项旨在确保研究制品 [3] 公开可用，并因此促进可复现性的倡议。在开放科学中，与软件工程最相关的是出版物的开放获取 、数据集的开放数据以及源代码的开源 [5], ，其中开放科学的每个方面都涉及不同的技术和最佳实践，以公开其各自类型的研究制品。包括欧盟在内的多个政府研究资助机构都规定必须开放获取科学研究成果（包括数据、工具等）[1]。

在文献中，Minocher 等人提出了四个属性数据可恢复性、数据可用性、分析清晰度和结果一致性，并明确强调了这些属性之间的顺序依赖关系——例如，如果数据不可恢复，那么数据的分析清晰度便毫无意义。

近期，为激励学者遵循开放科学原则，人们推出了开放科学徽章 [6] 和注册报告 [7]。然而，软件工程研究社区仍在适应开放科学原则 [3], ，制品无法获取的情况也很常见 [1]。制品无法获取的主要原因包括数据敏感，以及通讯作者变更所属机构后无法再访问其制品 [8]。尽管某些制品无法获取的原因（例如公司所有数据的敏感性）可能确实需要付出巨大努力才能解决，但其他原因（例如制品丢失、疏于管理）则可以通过遵循建议的指南 [3] 并使用现代制品共享工具轻松避免。

[1] https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science/ 开放获取_en

2.2. 需求质量文献

近期进展已经确立，需求工程（RE）中产生的制品会对下游软件开发活动产生重大影响 [9],，甚至可能导致项目失败 [10]。因此，需求制品值得进行质量保证 [1]。需求质量文献致力于提供理解需求质量以及衡量和改进需求质量的支持 [1]。其中一种广受欢迎的方法是提出质量因素。需求质量出版物通常会提出一个或多个质量因素——例如，使用协调歧义会导致不同的解释 [11]——在数据集中标注该质量因素的实例，最后提出一种实现（即一种算法或功能完备的工具），以自动检测这些实例。

这些制品——包括数据集和实现——都是重要的贡献，有助于推动实证研究和技术转移。虽然（经标注的）数据集是开发新的、改进现有质量因子检测实现的主要驱动力，但实现则是部署到业界、用于实际集成并改进软件工程流程的工具。NLP4RE 研究领域将自然语言处理（NLP）技术应用于 RE [12]，并构成需求质量文献 [1], 中很大一部分贡献，尤其关注工具的交付与改进。除了这些由 NLP 驱动的工具依赖训练数据的可用性和可靠性之外，这也使 NLP4RE 研究领域处于开放科学挑战 [12]的前沿。因此，NLP4RE 社区尤其清楚自身对制品可用性的依赖 [13]。

然而，近期的系统性研究显示，其中相当一部分制品已无法获取[2]，或者从未被提供过 [12, 1, 2]。表1 报告了从我们此前发表的、关于需求质量因素的文献综述所涉及的57项主要研究中提取的57个数据集（D）和36个实现（I）的可获取状态 [2]。

3. 恢复流程

人们认识到，需求质量工件的可获取性不足 [1, 2]，这促使我们设定目标：通过确保数据的可恢复性，改善需求质量文献中的开放科学现状；数据可恢复性是科学工作可复现的必要前提 [14]。

本节将按照所采取的步骤，记录制品恢复过程。在第 3.1 节中，我们介绍主要研究样本的选择。我们在第3.2 节详细说明联系通讯作者的方法，并在第 3.3节记录与其保持通信的过程。最后，我们在第 3.4节记录恢复过程及其成功情况的评估。所有生成的数据、脚本和文档均在我们的复现包[4]中公开。

[2]其中，“可获取”表示状态为应要求提供（见表1）或更高。[3]在此语境下，我们使用“开源”一词时，采用其通常的理解，而不是开放科学框架中所使用的含义 [5],，后者意味着遵循开放数据下列出的属性。[4]可访问于https://doi.org/10.5281/zenodo.7708571

表1需求质量工件的可用性状态 [2] 包括数据集（D）和实现（I）

状态	说明	(D)	(I)
开放数据	该制品托管在满足以下标准的服务中：(1) 不可变 URL（作者或其他人无法修改）其他），（2）永久（托管组织有一项使命，旨在在可预见的未来维护制品）、（3）可访问（那里是一个指向真实数据源 URL 的数字对象标识符），以及（4）开源许可证（该制品拥有授予访问和再利用权限的许可证）	1	0
开源[3]	[仅用于实现] 该实现对所有人可用使用 [illegible]，且代码库已被披露	-	5
论文中提供	[仅适用于数据集] 该数据集足够小，作者在手稿中披露完整的数据集	5	-
可访问链接	制品现在可以访问了，但缺少部分开放数据方面（见上文）	1	1
应要求	作者声称该制品可应要求提供	0	1
链接失效	论文中包含指向制品的链接，但该链接无法打开。	10	1
无链接	提供了一个制品，但未说明如何访问它。	15	27
私有	提供作者指出，某个制品存在，但对某些人而言是私有的原因（例如涉及私有数据的产业合作等）。	24	0
专有	该制品是专有的，付费后即可获得访问权限	1	1
总计		57	36

3.1. 研究样本选择与准备

我们采用了便利抽样，因为作为我们研究结果基础的主要研究是根据便利性选取的 [15]。具体而言，我们从一组用于构建需求质量因素本体的主要研究中获取了制品 [2]。为了开发该本体，我们从另一项次级研究 [1]所报告的原始出版物集中收集了报告质量因素的手稿。从此类出版物中提取数据集和实现，揭示了制品可用性的令人遗憾的现状。

我们用以下信息补充了此前研究 [2] 中有关数据集和实现的数据。

- 通讯作者：每个制品都关联了一位通讯作者。
- 提及：每个制品都与其在手稿中的原文提及相关联。

此外，我们更正了有关一个数据集和三个在先前研究中持续存在的实现的信息 [2]。

我们使用电子表格收集了有关的数据。

1. **作者**(n=35)，列出每位作者的姓名和电子邮箱地址；2. **数据集** (n=57)，列出每个数据集所属的出版物、原文提及、通讯作者及其当前可用性；3. **实现** (n=36)，列出每项实现所属的出版物、原文提及、通讯作者及其当前可用性。

3.2. 联系作者

我们编写了一个 Python 脚本，自动为每位通讯作者生成一封电子邮件。邮件包含以下内容：

1. 标题：解释我们的工作，并请求为开放科学作出贡献（或者说明为什么这是不可能的）。2. 制品列表：列出作者出版物中包含的、未采用开放获取的制品。3. 说明：简要如何根据开放科学原则正确披露制品，以及提出在此过程中为他们提供帮助。4. 联系方式：与我们取得联系的方式。

我们于2022 年11月 30^{th} 日首次通过邮件联系作者，随后于12 月 13^{th} 日发送提醒，并于 2023年 1 月 11^{th} 日发送最后一次提醒。对于截至最后一次提醒仍未回应我们请求的作者，我们还联系了他们的合著者，以提高获得回应的可能性。我们于 2023 年 2 月 8 日完成了恢复流程，整个过程历时 70 天。

3.3. 对应关系

我们通过在工作日的24小时内回复，与联系的作者保持密切联系。在此过程中，我们澄清疑虑并提供帮助。我们在一个电子表格文件中处理并记录作者答复中包含的信息。我们在额外的列中跟踪回复状态，将请求标记为无法投递、未答复、已答复或已完成。对于某个恢复请求，当通讯作者针对其所有制品提高了可获取性，或解释了无法恢复或披露这些制品的原因后，我们便将该恢复请求标记为已完成。

此外，我们记录了首次发送电子邮件的日期、收到首次回复的日期以及完成请求的日期，同时记录了作者发送的电子邮件数量、制品的更新后的可用性状态，以及最终作者未采取建议行动的解释。两名作者分别独立地对这些解释进行编码，并就未能恢复的原因类型达成了完全一致。当通讯作者的电子邮箱地址不再使用时，我们通过个人联系人或 Twitter 和 LinkedIn 等社交网络与其取得联系。

3.4. 评估

为了评估制品恢复流程，我们根据表格中的文档生成了以下数据的统计信息。

1. 通信（作者响应时间和频率），用于评估恢复过程的工作量。2. 恢复请求成功率（工件可用性的变化），用于评估恢复过程的成功程度。

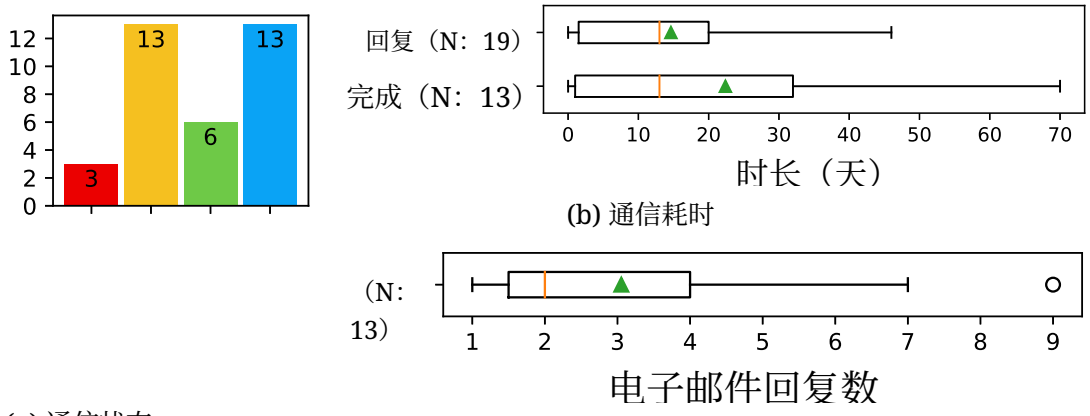
3. 未能恢复的原因（作者为未恢复所作的辩解），以评估阻碍开放获取的原因。

我们根据文档生成描述性统计数据，对数据进行了评估。

4. 结果

4.1. 通信

在联系的 35 位通讯作者中，19 位回复了恢复请求，其中 13 位完成了请求。尽管我们搜索了有效的联系方式，仍有三位作者无法联系上。通信状态的分布如图 1a 所示。通讯作者平均用 14.6 天回复我们的请求，并额外用 22.4 天完成请求。平均而言，一项请求通过与通讯作者往来 3 封电子邮件得以解决。这些统计数据的分布分别如图 1b 和图 1c 所示。



(a) 通信状态

(b) 通信耗时

(c) 通信频率

Fi图1： 与制品恢复 r

equest

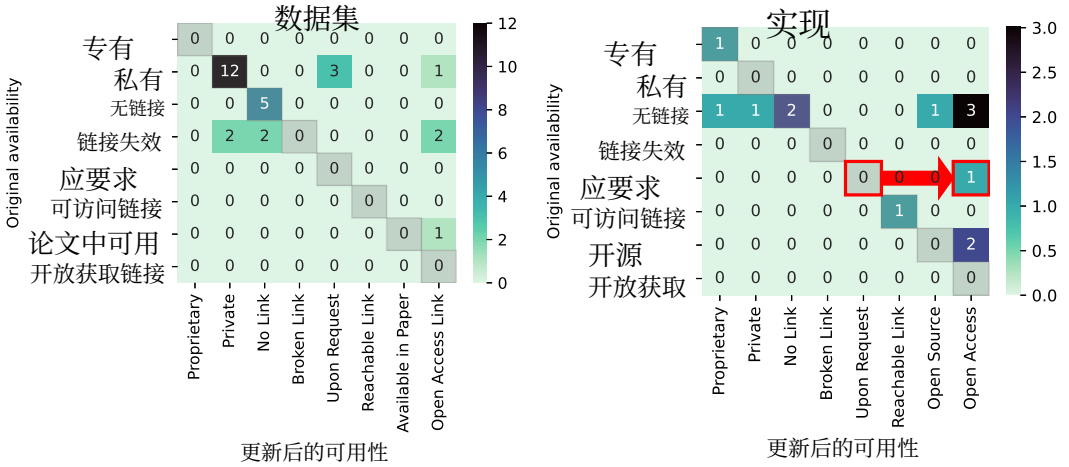
4.2. 制品恢复成功

通讯作者提高了七个数据集（其中四个遵循开放获取原则）和七个实现（其中六个遵循开放获取原则）的可获取性。这使数据集的可获取性从12.3%（7/57，1个开放获取）提高到22.8%（13/57，5个开放获取），并使实现的可获取性从19.4%（7/36，0个开放获取）提高到30.6%（11/36，6个开放获取）。作者还进一步确认了21个数据集和六个实现不可获取，并说明了无法恢复或披露这些数据集和实现的原因。

图 2 展示了恢复请求的成功情况。该热图考虑了所有制品（图 2a 中的数据集，以及图 2b中的实现），其中通讯作者完成了恢复请求。单元格中的数字表示制品的数量，这些制品的原始可用性（纵轴）已更新为新的可用性（横轴）。可用性保持不变的制品数量（例如，因为某位作者

已确认无法提高制品的可用性）报告在对角线上（灰色阴影部分）。制品可用性的提高对应对角线右侧的单元格，可用性的下降则对应左侧的单元格。

例如，之前有一项实现可通过申请获取，[16]。如今，作者按照开放获取原则公开了该实现[5],，对应条目向右移动了三个单元格（见图2b）。



（a）数据集可用性的变化　　（b）实现可用性的变化

图2：制品可用性的变化

无法找回或披露制品的情况报告如下：在 21 个无法找回的数据集中，15 个已丢失（即作者再也找不到这些数据，或作者认为掌握数据的联系人无法联系），另有 6 个因内容敏感而无法披露。在 6 个无法找回的实现中，3 个变为专有，另有 3 个已丢失。

5. 讨论

在这一流程的 70 天时间范围内，需求质量出版物的作者找回了若干数据集和实现，这些资料现已可用于科学结果复现，也可在未来项目中重复使用。我们在需求质量因素本体 [2][6] 以及复现包中引用了找回的制品，以便人们访问。

此外，作者还确认另有若干制品无法获取。虽然这并未直接提高用于复现的制品可用性，但澄清了若干数据集和实现的模糊状态。总体而言，作者在回复恢复请求时，要么找回了自己的数据，要么说明无法找回，并提供了有帮助的

[5]现已公开发布于 https://doi.org/107484023 ₆5281./zenodo.参见内容位于 http://reqfactoront.com/

解释。恢复请求失败的原因包括：（1）没有回应；（2）制品遗失；或（3）制品包含敏感信息。我们没有遇到恢复请求失败的其他原因，这印证了所抽样的需求质量社区致力于开放科学的善意。这与其他制品恢复尝试的经历形成对比，在那些尝试中，研究人员遇到过诸如要求报销费用，或认为恢复工作对个人没有任何收益等原因 [8]。

由于本研究存在局限性，我们无法声称这些观察结果对软件（需求）工程领域普遍适用。首先，主要研究的集合是通过便利抽样从一项先前研究中获取的 [2]。这一样本存在已知局限，因为有若干与需求质量文献相关的主要研究缺失。因此，关于恢复成功率和回应情况的结果并未涵盖完整的需求质量文献和研究群体。此外，我们关于回应状态的结论，尤其是未获答复和已获答复请求的情况，受到我们联系通讯作者方式的限制。以电子邮件作为沟通方式会妨碍回复率，因为人们更换所属机构后往往会弃用原有邮箱 [17]。回应成功率有限，是本研究所用时间范围和沟通渠道造成的结果，并不代表研究群体对开放科学的态度。

6. 结论

由于数据集和实现不可用，需求质量文献中先前出版物的可信度和可复用性均受到阻碍。我们请求57篇出版物的通讯作者依据开放科学原则公开其制品。我们提高了七个数据集和七个实现的可用性，其中若干现在遵循开放科学原则。

通过这项研究，我们希望提高人们对恢复早期出版物相关制品之重要性的认识。尽管得益于全面的指南（见 [3]）或社区举措（例如会议上的制品评估环节），人们近期对遵循开放科学原则的重视程度有所提高，但这些做法很少追溯应用于以往的出版物。此外，我们希望所创建的材料能够帮助 NLP4RE 等高度依赖制品的研究领域的研究人员恢复更多制品。

在需求质量因素本体[7], 的范畴内，我们的计划包括建立一个中央存储库，提供相关制品可用性及其存放位置的最新信息。我们诚邀研究人员为此贡献力量，充实本领域的证据。

致谢

KKS 基金会通过布莱金厄理工学院的S.E.R.T.ResearchProfile项目支持了这项工作。此外，我们还要感谢审稿人提出的宝贵意见，正是基于这些意见，本文得以改进。

[7] http://reqfactoront.com

参考文献

[1] L.Montgomery、D. Fucci、A. Bouraffa、L.Scholz、W. Maalej，需求质量的实证研究：一项系统映射研究，《需求工程》（2022）1–27。[2] J. Frattini、L. Montgomery、J.Fischbach、M. Unterkalmsteiner、D.Mendez、D. Fucci，文本需求质量因素的实时可扩展本体，载于：2022年IEEE第30届国际需求工程会议（RE），IEEE，2022年，第274–280页。[3] D.Mendez、D. Graziotin、S. Wagner、H.Seibold，软件工程中的开放科学，载于：《软件工程中的当代实证方法》，Springer，2020年，第477–501页。[4] B.C. Anda、D.I.Sjøberg、A.Mockus，软件工程中的可变性与可复现性：对开发同一系统的四家公司的研究，《TSE》35（2008）407–429。[5] J. Tennant、J.Beamer、J. Bosman、B.Brembs、N. C. Chung、G. Clement、T. Crick、J. Dugan、A.Dunning等，开放学术战略制定的基础（2019）。[6] M. C. Kidwell、L. B. Lazarević、E. Baranski、T.E. Hardwicke、S. Piechowski、L.-S.Falkenberg、C. Kennett、A. Slowik等，认可开放实践的徽章：一种提高透明度的简单、低成本且有效的方法，《PLoS Biology》14（2016）e1002456。[7] B. A. Nosek、C. R. Ebersole、A. C. DeHaven、D. T.Mellor，预注册革命，《美国国家科学院院刊》115（2018）2600–2606。[8] M. Gabelica、R. Bojčić、L. Puljak，许多研究人员未遵守其发表的数据共享声明：一项混合方法研究，《临床流行病学杂志》（2022）。[9] S. Wagner、D. M.Fernández、M.Felderer、A.Vetrò、M.Kalinowski、R.Wieringa、D. Pfahl、T. Conte、M.-T. Christiansson、D. Greer等，需求工程现状：一个理论与一组全球调查，《TOSEM》28（2019）1–48。[10] D. Mendez、S.Wagner，探寻需求工程中的痛点：全球调查体系的设计及德国的初步结果，载于：第17届国际软件工程评估与测量会议论文集，2013年，第183–194页。[11] S.Ezzini、S. Abualhaija、C. Arora、M.Sabetzadeh、L. C.Briand，使用领域专用语料库改进需求歧义处理，载于：2021年IEEE/ACM第43届国际软件工程会议（ICSE），IEEE，2021年，第1485–1497页。[12] L. Zhao、W. Alhoshan、A. Ferrari、K. J. Letsholo、M. A. Ajagbe、E.-V. Chioasca、R.T. Batista-Navarro，需求工程的自然语言处理：一项系统映射研究，《ACM计算调查》（CSUR）54（2021）1–41。[13] F. Dalpiaz、A. Ferrari、X.Franch、C. Palomares，需求工程的自然语言处理：最好的时代尚未来临，《IEEE软件》35（2018）115–119。[14] R. Minocher、S.Atmaca、C. Bavero、R. McElreath、B. Beheim，63年的社会学习研究中，可复现性呈指数级提升（2020）。[15] S.Baltes、P. Ralph，软件工程研究中的抽样：批判性评述与指南，《实证软件工程》27（2022）1–31。[16] F.-L.Li、J. Horkoff、L. Liu、A. Borgida、G. Guizzardi、J.Mylopoulos，使用desiree进行需求工程：一项实证评估，载于：国际高级信息系统工程会议，Springer，2016年，第221–238页。[17] J.D. Wren、J. E. Grissom、T. Conway，Medline中通讯作者的电子邮件衰减率：通过电子邮件地址过期与作者沟通并向其索取材料的能力正在逐渐消失，《EMBO报告》7（2006）122–127。