

Wiederverwendung von Mikrobiom-Daten

Fair teilen, besser forschen

ALEXANDER J. PROBST¹, ROLAND HATZENPICHLER²¹ UMWELT-METAGENOMICS, UNIVERSITÄT DUISBURG-ESSEN² DEPARTMENT OF CHEMISTRY AND BIOCHEMISTRY, MONTANA STATE UNIVERSITY, BOZEMAN, USA

DOI: 10.1007/s12268-026-2710-2

© The Author(s) 2026

■ Forschungsdaten müssen öffentlich verfügbar und wiederverwendbar sein; das ist das Grundprinzip von *Open Data*. Nur das Einhalten dieses Grundprinzips für Sequenzdaten von Mikrobiomen erlaubt es uns, große Metastudien anzufertigen, globale Zusammenhänge zu verstehen, aber auch phylogenetische Beziehungen fein aufzulösen. Die *Bermuda Principles* von 1996 fordern eine öffentliche Freigabe von Sequenzdaten innerhalb von 24 Stunden nach Fertigstellung bzw. Qualitätssicherung – ein Prinzip, das später durch das *Fort Lauderdale Agreement* 2003 nochmal verstärkt wurde [1]. Betrachtet man die Entwicklung der Menge an Sequenzdaten, die seither jährlich produziert wurde, und die Entwicklung der Geschwindigkeit von Wissenschaften in diesem Bereich im Allgemeinen, ist es klar, dass eine Regelung, die zur Zeit der Sequenzierung des Humangenoms erdacht wurde, nicht mehr aktuell sein kann und überarbeitet werden muss.

Ausgangslage: veraltete Regeln, aktueller Tumult

Dass sich heutzutage kaum ein:e Wissenschaftler:in an die *Bermuda Principles* hält und die Daten 24 Stunden nach Generierung öffentlich zur Verfügung stellt, ist offensichtlich. Alle gängigen Zeitschriften sind sich zumindest einig, dass zum Zeitpunkt der Publikation eines Artikels alle Sequenzdaten öffentlich zur Verfügung stehen müssen, und im Idealfall sind die Daten zumindest auch für den Begutachtungsprozess eines Artikels vorhanden. Manche Zeitschriften verlangen die Veröffentlichung der Daten bereits bei der Einreichung des Artikels. Darüber hinaus gibt es auch internationale Wissenschaftsinstitute, die Sequenzdaten nach dem Erstellen direkt online stellen, auch wenn sie diese für Kooperationen mit anderen Instituten generieren. Diese können dann andere Wissenschaftler:innen verwenden, sodass Konkurrenzsituationen entstehen könnten. Genau ein derartiger Fall hat 2022 viel Aufsehen in Social Media generiert: Ein Wissenschaftler an einem amerikanischen Institut hatte wertvolle Proben, die in mehreren aufwändigen Expeditionen gesammelt wurden, an einer öffentlichen Einrichtung sequenzieren lassen. Eine Zusammenfassung des bewilligten Antrags zur Sequenzierung der Proben war ebenfalls online erschienen. Diese Proben wurden direkt von anderen Wissenschaftler:innen in einer Veröffentlichung verwendet – zum Schaden des Wissenschaftlers aus den USA und vor allen Dingen der Doktorand:innen, die auf dem Projekt arbeiteten. Leider war dies kein Einzelfall, Gleichzeitig fühlten sich jene, die die Proben nahmen und auf Expedition waren, Metadaten erhoben, also die „Datenersteller“, nicht gewürdigt. Es war Zeit, die *Bermuda Principles* und das *Fort Lauderdale Agreement* zu überdenken.

Die Würdigung von Datenersteller:innen (*Data Creators*) bei der Verwendung von

öffentlichen Daten ist ein Thema, das schon sehr viele Jahre in der Wissenschaft existierte. Es wurden einige Vorschläge gemacht, z. B. Datenzitationen [2, 3], jedoch wurde keiner in der Wissenschaftsgemeinschaft angenommen, sodass es nun an der Zeit ist, eine neue Lösung zu finden, die eher auf Kooperation als auf Konkurrenz basierte.

Vorgeschlagene Roadmap

In einem kürzlich erschienenen Artikel [4] schlagen wir zusammen mit 226 anderen Wissenschaftler:innen eine Roadmap vor. Ziel ist eine *Data Reuse Information* (DRI) für jeden Datensatz einzuführen, wenn dies die Datenersteller:innen wünschen. Dieser Eintrag besteht aus einer oder mehreren ORCIDs (*Open Researcher and Contributor IDs*), womit die Datenersteller:innen kontaktiert werden sollen, wenn die Daten wiederverwendet werden. Ziel ist es, dass die Nutzenden der Daten mit den Ersteller:innen in Kontakt treten, zusätzliche Metadaten und Informationen fließen und eine Kooperation entsteht, die einerseits die Wissenschaft verbessert aber andererseits auch die Arbeit der Datenersteller:innen würdigt. Darüber hinaus könnten damit auch Konflikte von Projekten, auf denen Promovierende arbeiten, möglicherweise verhindert werden, sodass diese in einer gesunden, für die Wissenschaft nachhaltigen Geschwindigkeit an neu erhobenen Daten arbeiten können ohne Angst, dass sie von anderen Arbeitsgruppen durch Datenwiederverwendung genutzt werden.

Diese Roadmap kann aber nicht blind eingesetzt werden, ohne dass sie die Wissenschaft einschränkt. Eine Kontaktaufnahme und Kooperation solle nur zustande kommen, wenn die wiederverwendeten Daten auch in einer Abbildung der neuen Studie erscheinen oder in einem fundamentalen Zusammenhang mit jener stehen. Reine Sequenzvergleiche, um neue Sequenzen zu annotieren, Metaanalysen oder KI-basierte Reanalysen von großen Datensätzen, oder gar Nutzung von sekundären Datenbanken ohne DRI sind in der Roadmap ausgenommen. Natur-



lich ist es wünschenswert, dass auch hier Kontakt und Kooperation entsteht, jedoch kann dies aus Zeitgründen von Datenerstellenden nicht immer verlangt werden. Der allgemeinen Nutzung der DRIs bei Einreichung eines Datensatzes und bei der Wiederverwendung von Daten mit DRI wurde in einer Umfrage in der oben genannten Studie mit großer Mehrheit (94 % bzw. 96 %) zugestimmt.

Schwachstellen und Zukunft der Verwirklichung

Neben der Nutzung des DRIs durch die Wissenschaftsgemeinschaft steht und fällt der Erfolg der Roadmap mit seiner Implementierung in Datenbanken. Hier ist der nächste Schritt essenziell; das *European Nucleotide Archive* hat im vergangenen Jahr bereits einen Eintrag für ORCIDs von Datensätzen eingeführt, sodass die Frage nach der Möglichkeit der Implementierung beantwortet ist. Es fehlt jedoch der spezifische Eintrag des DRI und der damit verbundenen Roadmap. Neben der Implementierung muss auch geklärt werden, wie sich der DRI über Datenbanken hinweg fortpflanzt. Zum Beispiel kann dies sehr leicht für *Sequencing Read*

Archive(SRA)-Daten eingeführt werden, jedoch haben diese Datensätze auch assoziierte Genome sowie vorhergesagte Gen- und Proteinsequenzen. Wie sich der DRI über Datensätze fortpflanzt, ist noch unklar und müsste je nach Möglichkeiten erschlossen werden.

Um den DRI vernünftig und ohne großen Aufwand global verwenden zu können, wird direkt nach seiner Implementierung in Datenbanken Software notwendig sein, die entsprechende DRIs in Datensätzen identifiziert und direkt die entsprechenden Personen per E-Mail automatisiert zu kontaktieren. Die Deutsche Forschungsgemeinschaft fördert im Zuge des INF-Projekts im SFB RESIST (1439) die Erschließung der Etablierung der Implementierung und der Entwicklung einer entsprechenden Software.

Danksagung

Wir danken allen Wissenschaftler:innen, die sich an der Erstellung der Roadmap beteiligt haben, vor allem Laura A. Hug, Cristina I. Moraru, André R. Soares und Folker Meyer für jahrelange Planungen und Diskussionen. AJP bedankt sich für die Finanzierung des INF-Projekts im SFB 1439/2 RESIST (Projekt-

nummer 42654780) der Deutschen Forschungsgemeinschaft (DFG).

Literatur

- [1] Undefined (2013) Report of the International Strategy Meeting on Human Genome Sequencing held at the Princess Hotel, Southampton, Bermuda, on 25th–28th February 1996. <http://hdl.handle.net/10161/7715>
- [2] Borgman CL (2016) Data Citation as a Bibliometric Oxymoron. In: Sugimoto CR (ed.) *Theories of Informetrics and Scholarly Communication*. Berlin, De Gruyter, 93–116
- [3] Cousijn H, Feeney P, Lowenberg et al. (2019) Bringing citations and usage metrics together make data count. *Data Sci J* 18: 9
- [4] Hug LA, Hatzepichler R, Moraru C et al. (2025) A roadmap for equitable reuse of public microbiome data. *Nat Microbiol* 10: 2384–2395

Funding: Open Access funding enabled and organized by Projekt DEAL.
Open Access: This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Korrespondenzadresse:

Prof. Dr. Alexander Probst
 Arbeitsgruppe Environmental Metagenomics
 Research Center One Health Ruhr
 Fakultät für Chemie
 Universität Duisburg-Essen
 Universitätsstraße 5
 D-45141 Essen
alexander.probst@uni-due.de

