NETZWERK Digital GreenTech

VERÖFFENTLICHUNG: JANUAR 2023

Publikation von Forschungsdaten und Forschungsmanagement am Beispiel der Bioökonomie

Von Prof. Dr. Ulrike Lucke, Institut für Informatik und Computational Science, Ulrike.lucke@uni-potsdam.de
Prof. Dr. Anette Prochnow, Leibniz-Institut für Agrartechnik und Bioökonomie, aprochnow@atb-potsdam.de

Dieses Papier ist im Projekt Netzwerk Digital GreenTech (NetDGT) entstanden. NetDGT leistet wissenschaftliche Querschnittsarbeit zur Fördermaßnahme Digital GreenTech des Bundesministeriums für Bildung und Forschung (BMBF). Es vernetzt und informiert zu Themen an der Schnittstelle von Umwelttechnik, Digitalisierung und Nachhaltigkeit. Die Veröffentlichungs-Reihe bereitet Wissen, das in und um die Maßnahme entsteht auf und macht sie einer interessierten Öffentlichkeit zugänglich.

1. Reproduzierbarkeit als Herausforderung und Chance

Aus Schwächen in statistischen Analysen (Ioannidis, 2005), aufgedeckten Fehlern in Software (Eklund et al., 2016) und der mangelnden Reproduzierbarkeit von Studienergebnissen insgesamt (Open Science Collaboration, 2015) resultierte in den 2010er Jahren eine Bewegung, die in der Psychologie beginnend mittlerweile alle Wissenschaftsdisziplinen erfasst hat. Gefordert werden insbesondere eine bessere Nachvollziehbarkeit der in der Forschung genutzten Daten und Methoden sowie eine Reproduzierbarkeit der Ergebnisse.

Im Ergebnis erfuhr die schon Jahre zuvor gestartete Open-Access-Bewegung einen deutlichen Aufwind. FAIRness (Wilkinson et al., 2016) entstand als neues Ziel im Umgang mit (digitalen) Forschungsobjekten, die auffindbar (findable), zugänglich (accessible), interoperabel (interoperable) und wiederverwendbar (reusable) sein sollen. Damit verbunden waren auch begriffliche und methodische Schärfungen (Plesser, 2018), Analysen etablierten Haltungen und Praktiken (Baker, 2016) sowie die Etablierung von Normen und Qualitätssicherungsmechanismen in einzelnen Communities, z.B. die die Vorabregistrierung von Experiment-Designs (Nosek et al., 2018),

die Prüfung von Daten und Software hinter Publikationen (ACM, 2022) oder allgemeine Richtlinien zum Umgang mit Forschungsdaten (DFG, 2022).
Reproduzierbarkeitsstudien in verschiedenen Fächern und Publikationsplattformen wiesen jedoch auch Jahre später nur geringe Reproduzierbarkeitsraten der veröffentlichten Forschungsergebnisse nach (Collberg et al., 2016; Nüst et al., 2018; Stagge et al., 2019; Riedel et al., 2022). Als Ursachen hierfür wurde neben einer mangelnden Verfügbarkeit bzw. Interpretierbarkeit der genutzten Datensätze auch wiederholt die enge Kopplung an die für die Erhebung, Verwaltung oder Verarbeitung genutzte Software identifiziert. Selbst wenn diese gemeinsam mit den Daten veröffentlicht wurde, war sie häufig nicht lauffähig. Es verbleibt also in allen Disziplinen weiterhin großer Handlungsbedarf insbesondere bei Forschungssoftware (Lucke, 2022).

2. Forschungsdaten und –softwarte in der Bioökonomie

Die Bioökonomie steht vor der Aufgabe, die gesunde Ernährung einer wachsenden Weltbevölkerung zu sichern, die Wirtschaft von fossilen auf erneuerbare Rohstoffe und Energieträger umzustellen und dabei gleichzeitig die Umweltbelastungen zu reduzieren, Treibhausgasemissionen zu mindern und sich an den Klimawandel anzupassen. Die Bioökonomie bezieht sich auf die „Erzeugung, Erschließung und Nutzung biologischer Ressourcen, Prozesse und Systeme, um Produkte, Verfahren und Dienstleistungen in allen wirtschaftlichen Sektoren im Rahmen eines zukunftsfähigen Wirtschaftssystems bereitzustellen“ [BMBF und BMEL 2020]. Sie umfasst die Bereitstellung von Biomasse, ihre Nutzung als Lebensmittel, Biomaterialien und Energieträger und das integrierte Management biogener Reststoffe. Die Bioökonomie beruht auf Biomasse, biologischen Prinzipien und Prozessen und biologischem Wissen.
Damit verbunden sind eine hohe Komplexität und Diversität der Systeme. Die Diversität bioökonomischer Produktionssysteme drückt sich in ihrer Gebundenheit an sehr unterschiedliche Standortbedingungen, in einer hohen zeitlichen und räumlichen Heterogenität und in einer hohen Variabilität und Individualität aus. Dementsprechend umfangreich und vielfältig sind Forschungsdaten in der Bioökonomie.

Die Forschungsdaten in der Bioökonomie sind von einer großen Diversität gekennzeichnet. Sie werden auf unterschiedliche Weise generiert, vor allem durch:

- Messungen (mit Sensoren, Instrumenten, ggf. auch händisch)
- Experimente (z.B. im Feld, in Technika, Pilotanlagen, Praxisanlagen, Reallaboren)
- Beobachtungen (z.B. Tierverhalten, Arbeitsprozesse)
- Erhebungen (z.B. bei Erzeuger:innen und Verbraucher:innen)
- Befragungen
- Modell-Simulationen

Die Forschungsdaten liegen anschließend in unterschiedlicher Form vor und sind dementsprechend zu behandeln, z.B. als:

- numerische Werte, ggf. mit Maßeinheiten
- Bilddaten (z.B. Fotografien, Luftbilder, Satellitenbilder, Karten, Grafiken, Zeichnungen, Videos)
- Texte (z.B. genetischer Code, Volltexte, Fragebögen)
- Modelle

Bislang existieren hierfür noch keine durchgängigen Infrastrukturen, und es fehlt damit verbunden an etablierten Standards für Daten und Metadaten, an Instrumenten für die Qualitätssicherung, an rechtlichen und ethischen Richtlinien u.v.m. (NFDI4Agri, 2019).

Die im Prozess der Forschung genutzten Daten lassen sich auch in der Bioökonomie nur selten von der für ihre Produktion und Verarbeitung eingesetzten Software trennen. Daher muss auch Software FAIR behandelt werden (Chue Hong et al., 2022). Selbst bei offenen Quellformaten wie CSV-Dateien können nachgelagerte Verarbeitungsschritte je nach eingesetzter Software und Systemarchitektur variieren, bspw. in der Zahl verwendeter Nachkommastellen oder in der Rundung von Ergebnissen. Software kann in der Forschung verschiedene Rollen einnehmen. Sie ist manchmal ein eingesetztes Werkzeug, das von Dritten stammt und in unveränderter Form Verwendung findet; hier finden sich neben freier Software häufig auch kommerzielle Produkte. Oft ist Forschungs-Software auch ein eigenständiges Produkt der Forschung, die als ein weiteres Artefakt aus dem Forschungsprozess heraus entsteht. Und manchmal – etwa in der Informatik – ist Software auch der unmittelbare Gegenstand von Forschung. Manche Forschungs-Software ist generisch, wird in verschiedenen Fächern eingesetzt. Andere ist auf eine bestimmte Disziplin und deren Bedarfe zugeschnitten. Software in der Forschung ist demnach so vielfältig wie die Forschungslandschaft selbst (Lucke, 2022). Zuverlässige und nachhaltig nutzbare Software ist daher wesentlich für die Reproduzierbarkeit von Forschungsergebnissen in allen Disziplinen.

3. Forschungsdatenmanagement

Der systematische Umgang mit Forschungsdaten und -software entlang aller Schritte im Forschungsprozess verlangt die Berücksichtigung verschiedener Aspekte. Dazu zählen neben technischen Lösungen auch organisatorische und rechtliche Fragen.

Auf verschiedenen Ebenen sind bereits Policies erlassen, die einen Rahmen für den Umgang mit Forschungsdaten aufspannen – durch Geldgeber, durch Fachgesellschaften, durch Institutionen und durch Publikationsplattformen. Dabei sind die Regelungen jedoch weder strukturell noch inhaltlich konsistent; hier bedarf es weiterer Konsolidierungen (Hrynaszkiewicz et al., 2020). Die jeweils formulierten Erwartungshaltungen sind zudem mit Unterstützungsangeboten für die in der Forschung tätigen Akteure zu untersetzen.

Das betrifft u.a. nutzbare Infrastrukturen und Werkzeuge. Für verschiedene Schritte im Forschungsprozess gibt es bereits Tools, die das Management von Forschungsdaten -und -software unterstützen. Dazu zählen die Verwaltung von Datenmanagementplänen, das Führen elektronischer Laborbücher oder die Erstellung von Metadaten. Die Vielfalt ist groß, die Integration in ein durchgängig nutzbares Ökosystem jedoch noch dürftig. Selbst für ein minimales Ni-

veau an Interoperabilität – auf der Ebene von Dateiformaten und Metadaten – fehlen bisweilen noch Bausteine.

Weitere Unterstützung ist durch gezielte Kompetenzentwicklung sowohl bei Forschenden als auch bei Studierenden erforderlich. Hier existieren bereits Kataloge der zu vermittelnden Kompetenzen (Petersen et al., 2022), Materialsammlungen und komplette Bildungsangebote zu ausgewählten Bausteinen (Biernacka und Schulz, 2022) sowie empirische Forschung zur Wirksamkeit verschiedener Ansätze (Wiljes & Cimiano, 2019). Die systematische Integration in Curricula an Hochschulen steht jedoch noch aus.

Eine deutliche Verbesserung des Bildes dürfte sich durch nationale Initiativen wie die NFDI[1] oder durch internationale Ansätze wie die EOSC[2] ergeben. Disziplinspezifische Maßnahmen werden konkrete Bedarfe in den Blick nehmen und spezifische Lösungen bereitstellen. Für die Informatik wird dies bspw. das Konzept des Research Data Management Containers (Goedicke & Lucke, 2022) umfassen, der neben Forschungsdaten auch die damit verbundene Software mitsamt ihrer Ausführungsumgebung aufnehmen und nachhaltig verfügbar machen soll. Davon werden auch andere Disziplinen profitieren.

4. Auf dem Weg zu Open Science

Wenn wir die Nachvollziehbarkeit von Forschungsergebnissen als Herausforderung ernst nehmen, dann sind offene Daten („Open Access“ oder „FAIR data“) und offene Software („Open Source) nur ein Teil der benötigten Lösung. Sie betreffen nur die in der Forschung entstehenden oder genutzten Artefakte.
Öffnet man darüber hinaus das Verständnis von reproduzierbarer Forschung auch für den Prozess
(die angewandte Methodik oder die durchgeführte Begutachtung), so gelangt man zum Konzept von „Open Science“ (Vicente-Saez & Martinez-Fuentes, 2018). Hier schließen Ideen wie „Citizen Science“ oder „partizipative Forschung“ an (Unger, 2014), die zugleich mehr Transparenz und Innovation versprechen. Praktiken ändern sich jedoch nur langsam, und daher sind gezielte Mechanismen für Education & Training und zum Community Development von herausragender Bedeutung für den nötigen Kulturwandel.

[1] https://www.nfdi.de/
[2] https://eosc-portal.eu/

LITERATUR

1. ACM - Association for Computing Machinery (2022). Artifact Review and Badging, Version 1.1, https://www.acm.org/publications/policies/artifact-review-and-badging-current

2. Baker M (2016). 1,500 scientists lift the lid on reproducibility. Nature. 533 (7604): 452–4. https://doi.org/10.1038/533452a

3. Biernacka K, Schulz S (2022). Forschungsdatenmanagement in der Informatik. Berlin: Logos. https://doi.org/10.30819/5490

4. BMBF und BMEL (2020): Nationale Bioökonomiestrategie. Bundesministerium für Bildung und Forschung und Bundesministerium für Ernährung und Landwirtschaft, www.bmel.de/SharedDocs/Downloads/DE/Broschueren/nationale- biooekonomiestrategie-langfassung.html

5. Chue Hong et al. (2022). FAIR Principles for Research Software (FAIR4RS Principles) (1.0). RDA FAIR4RS WG. https://doi.org/10.15497/RDA00068

6. DFG - Deutsche Forschungsgemeinschaft (2022). Guidelines for Safeguarding Good Research Practice. Code of Conduct, Version 2. https://doi.org/10.5281/zenodo.6472827

7. Eklund A, Nichols TE, Knutsson H (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. PNAS 113(28): 7900-5. https://doi.org/10.1073/pnas.1602413113

8. Goedicke M, Lucke U (2022). Research Data Management in Computer Science - NFDIxCS Approach. Proceedings INFORMATIK 2022, Bonn: GI, 1317-1328. https://doi.org/10.18420/inf2022_112

9. Hrynaszkiewicz I, Simons N, Hussain A, Grant R, Goudie S (2020). Developing a Research Data Policy Framework for All Journals and Publishers. Data Science Journal, 19(1), 5. http://doi.org/10.5334/dsj-2020-005

10. Ioannidis JPA (2005). Why Most Published Research Findings Are False. PLoS Med 2(8): e124. https://doi.org/10.1371/journal.pmed.0020124

11. Lucke U (2022). The role of Infrastructure for Software in Open Science. Proc. Open Science European Conference (OSEC), Paris: OpenEdition Press, 177-182. https://doi.org/10.4000/books.oep.15829

12. NFDI4Agri (2019). Research Data Infrastructure for Agricultural and Soil Sciences. Position Paper. https://www.nfdi4agri.de/phocadownload/documents/NFDI4Agri_PositionPaper_0.65.pdf

13. Nosek BA; Ebersole CR; DeHaven AC; Mellor DT (2018). The preregistration revolution. Proceedings of the National Academy of Sciences. 115 (11): 2600–2606. https://doi.org/10.1073/pnas.1708274114

14. Nüst D; Granell C; Hofer B; Konkol M; Ostermann FO; Sileryte R; Cerutti V (2018). Reproducible research and GIScience: an evaluation using AGILE conference papers. PeerJ 6:e5072. https://doi.org/10.7717/peerj.5072

15. Open Science Collaboration (2015). Estimating the reproducibility of psychological science. Science 349(6251). https://doi.org/10.1126/science.aac4716

16. Petersen B, Engelhardt C, Hörner T, Jacob J, Kvetnaya T, Mühlichen A, Schranzhofer H, Schulz S, Slowig B, Trautwein-Bruns U, Voigt A, Wiljes C (2022). Lernzielmatrix zum Themenbereich Forschungsdatenmanagement für die Zielgruppen Studierende, PhDs und Data Stewards, Version 1. Zenodo. https://doi.org/10.5281/zenodo.7034478

17. Plesser HE (2018). Reproducibility vs. Replicability: A Brief History of a Confused Terminology. Front. Neuroinform. 11:76. https://doi.org/10.3389/fninf.2017.00076

18. Riedel C, Geßner H, Seegebrecht A, Ayon SI, Chowdhury SH, Engbert R, Lucke U (2022). Including Data Management in Research Culture Increases the Reproducibility of Scientific Results. Proceedings INFORMATIK 2022. GI: Bonn. 1341-1352. https://doi.org/10.18420/inf2022_114

19. Stagge, J.; Rosenberg, D.; Abdallah, A. et al. (2019). Assessing data availability and research reproducibility in hydrology and water resources. Sci Data 6, 190030. https://doi.org/10.1038/sdata.2019.30

20. Unger, H (2014). Partizipative Forschung. Springer. https://doi.org/10.1007/978-3-658-01290-8

21. Vicente-Saez R; Martinez-Fuentes C (2018). Open Science now: A systematic literature review for an integrated definition. Journal of Business Research. 88: 428–436. https://doi.org/10.1016/j.jbusres.2017.12.043

22. Wiljes C, Cimiano P (2019). Teaching Research Data Management for Students. Data Science Journal 18(1): 38. https://doi.org/10.5334/dsj-2019-038

23. Wilkinson et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

GESELLSCHAFT FÜR INFORMATIK

WWW.DIGITALGREENTECH.DE