Redet über die Daten!

Forschungsdatenmanagement und Hochschullehre in der Physik und darüber hinaus

Ein Kommentar von Philipp Jaeger [†,*] *und Janice Bode* [‡]

25.05.2021

Abstract. Die Digitalisierung der Gesellschaft und der modernen wissenschaftlichen Arbeitsweise wurden im Physikstudium lange vernachlässigt. Vor dem Hintergrund aktueller Initiativen zu Open Science und *Forschungsdatenmanagement* (FDM) auf nationaler und europäischer Ebene werden verschiedene Sichtweisen auf zukunftsweisende FDM-Systeme diskutiert. Weiter wird aufgezeigt, dass mit vergleichsweise wenig Aufwand Kompetenzen ins Physikstudium integriert werden können, die eine innovative Nutzung von Forschungsdaten fördern.

Im Physikstudium wird der Umgang mit allerlei mathematischen Konzepten, Theorien und komplexen Messgeräten erlernt. In den Praktika wird gemessen, abgespeichert, ausgewertet, Messfehler analysiert – wir alle erinnern uns. Und doch steht der sprichwörtliche Elefant mitten im Raum, über den nicht gesprochen wird: die Daten selbst.

Aufgrund der fortschreitenden Digitalisierung stehen uns heute mehr Daten und Informationen jederzeit zur Verfügung als jemals zuvor, etwa in Online-Enzyklopädien, Nachrichtenportalen, Suchmaschinen, etc. Auch in der Forschung nimmt die Bedeutung großer Datenmengen zu und manifestiert sich in Initiativen wie der *European Open Science Cloud* [1] (EOSC) oder in Deutschland in der *nationalen Forschungsdateninfrastruktur* [2] (NFDI).

Doch zurück zum physikalischen Grundpraktikum, das im Folgenden als Beispiel dienen soll. Selbst bei sehr einfachen Experimenten müssen Daten so abgelegt werden, dass sie im Nachhinein für den oder die Studierende interpretierbar sind – was wurde wie und mit welchem Instrument gemessen, wie groß sind die Messfehler, aber auch externe Faktoren, wie beispielsweise die Raumtemperatur oder der Luftdruck bei Experimenten zur Thermodynamik.

Digitalisierung des Studiums: Forschungsdatenmanagement im Grundpraktikum

Anhand dieser Metadaten kann später nachvollzogen werden, wo bei der Durchführung des Experiments Probleme aufgetreten sind. Je umfangreicher diese Metadaten sind, desto wertvoller wird der Datensatz insgesamt, etwa weil er auch für Dritte nachvollziehbar wird, wenn eine ausreichend genaue Beschreibung der Messapparatur beiliegt. Werden solche Datensätze – nennen wir sie Datenobjekte – in einer Form abgelegt, die sie anhand der Metadaten vergleichbar macht, kann eine Auswertung der kumulierten Daten völlig neue Antworten liefern, z.B. in Bezug auf temperaturbedingte jahreszeitliche Schwankungen der Resultate oder Materialermüdung in mechanischen Komponenten der Apparatur.

Eine notwendige Voraussetzung für die Nachnutzung von Daten und damit für ein funktionierendes *Forschungsdatenmanagement* (FDM) ist, dass sie auch nach Jahren noch auffindbar sind. Umfangreiche Metadaten ermöglichen in Kombination mit offenen Lizenzen einen leichten Zugang. Durch die Verwendung standardisierter Dateitypen wird gewährleistet, dass die Daten leicht mit unterschiedlichen Anwendungen bearbeitet werden können und der Zugang möglichst nicht von proprietärer Software abhängt. Kurz zusammengefasst sollten Daten, wo immer möglich, den oben

[†] Department of Physics and Astronomy, University of Manitoba, und Fachgruppe Physik, Fakultät 4, Bergische Universität Wuppertal
[‡] Institut für Theoretische Physik, Westfälische Wilhelms-Universität Münster
[*] Korrespondezadresse: jaeger@jdpg.de

beschriebenen *FAIR-Prinzipien*,[3] (nach den englischen Begriffen *findable, accessible, interoperable* und *reusable*) entsprechen.

FAIRe Datenobjekte (FDO)[4] sind wie am oben stehenden Beispiel gezeigt der elementare Baustein nachhaltiger Datennutzung. Der Umgang mit Daten wird klassischerweise im Zuge des Grundpraktikums erstmals ausführlicher gelehrt. Diese Gelegenheit sollte genutzt werden, um frühzeitig den Mehrwert gemeinsamer Datennutzung zu vermitteln und dem derzeit weit verbreiteten Datenprotektionismus vorzubeugen. Die *Zusammenkunft aller deutschsprachigen Physikfachschaften* (ZaPF) hat diesbezüglich 2020 Vorschläge vorgelegt[5], die auch der Diskussion um Open Science Rechnung tragen.

Technische Möglichkeiten von FDO

Wie bei de-facto-Standards zum Datenaustausch im Internet – etwa HTML-Seiten für formatierten Text oder JSON für komplexere Objekte – können FDO ein sehr abstraktes Vehikel für jede Art von Daten sein. Dies gilt insbesondere auch für große Datenmengen, die man in der Regel nicht bei der ersten Anfrage an den Server übertragen möchte, da oft nur ein kleiner Teil davon wirklich beim Client gebraucht wird. Sowohl HTML als auch JSON können Spezifizierungen eines FDO-Standards sein, ebenso wie sehr große Dateien auf einem Object-Storage-Dateisystem, in denen Metadaten und das eigentliche Objekt schon konzeptuell getrennt sind und nicht notwendig auf derselben Festplatte oder demselben Rechner gespeichert sind.

Intelligente Speicherungs- und Cachingmechanismen können dabei für eine hohe Verfügbarkeit und durch dezentrale Datenaufbewahrung Ausfallsicherheit gewährleisten. Dabei gehen einzelne Repositorien in so genannten *Data Lakes* auf, deren Struktur der des Internets selbst ähnelt. Durch geeignet abgesicherte Identifizierungscodes bleibt auch bei verteilter Datennutzung nachvollziehbar, welche Forschenden einen Datensatz zur Verfügung gestellt haben, sodass dieser bei daraus resultierenden Publikationen über einen *Digital Object Identifier* (DOI)[6] zitiert werden kann.

Wissenschaftspolitisch hätte dies den Nebeneffekt, dass sich FDO mittelfristig neben Veröffentlichungen in anerkannten Peer-Review-Zeitschriften als eine weitere Währung für akademischen Erfolg etablieren könnten. Hier bleibt die Frage zu klären, inwieweit eine Qualitätssicherung und Kuration von FDO notwendig ist und wie diese aussehen kann, ohne den freien Zugang zum Data Lake unnötig zu verkomplizieren, vor allem aber ohne hohe Kosten für dessen Nutzung zu verursachen – es sei an die laufende Debatte um *predatory Journals* und die Rolle der großen Wissenschaftsverlage erinnert. Positive Beispiele sind etwa Zenodo[7] oder das *Open Science Framework* (OSF)[8]. Es dürfen in diesem Prozess auf keinen Fall Institutionen entstehen, die eine *„Gate Keeper“*-Funktion haben. Auch diese Diskussion findet – z.B. durch die Monopolstellung großer Digitalkonzerne befeuert – bereits in den Medien statt. Lehren daraus sollten also schon in der Konzeption zukünftiger Infrastrukturen berücksichtigt werden.

Transdisziplinäre Ausgestaltung von FDM-Systemen

Aufmerksamen Lesenden dürfte nicht entgangen sein, wie universell FDO einsetzbar sind. Wie oben beschrieben können viele der heute üblichen Standards zur Datenübertragung im Internet leicht so modifiziert werden, dass sie den FAIR-Prinzipien entsprechen – es ist also nur folgerichtig, die Umsetzung von FDM-Systemen jenseits des Kontexts und der Anwendungsfälle einzelner Disziplinen zu betrachten. Damit rücken automatisch der Forschungsgegenstand und die eingesetzten Methoden in den Vordergrund. Bei Letzteren kommen zu den etablierten Methoden der Fachwissenschaft die der technologiegestützten Datenanalyse hinzu, wie etwa der Einsatz künstlicher Intelligenz zur Transkription und strukturierten Repräsentation von Texten oder die automatisierte Suche nach statistischen Anomalien der zugrunde liegenden Daten.

Die rasante Digitalisierung der letzten Jahrzehnte zeigt die Notwendigkeit eines solchen Konzepts im Umgang mit Daten. So hat sich gemäß Moore’schen Gesetz[9] die Transistordichte in elektronischen Komponenten zwischen

1965 und 2020[10] etwa alle zwei Jahre verdoppelt. Die Leistung von modernen Supercomputern liegt im ExaFLOP-Bereich[11,12] und Smartphones haben mehr Rechenleistung als den Apollo-Mondmissionen der späten 1970er Jahre zur Verfügung stand. Im internationalen Vergleich liegt Deutschland auch in Sachen FDM im Mittelfeld: In Nordamerika müssen Forschungsprojekte in naher Zukunft FDM-Konzepte vorlegen, während in anderen EU-Ländern viel stärker mit offenen Lizenzen gearbeitet wird als hierzulande[13]. Sowohl die deutliche Zunahme der zur Verfügung stehenden Rechenleistung und der Datenmengen als auch mindestens vergleichbare Bestrebungen im internationalen Vergleich zeigen, dass ein neuer Umgang mit Forschungsdaten erforderlich ist.

Die NFDI riskiert in ihrer aktuellen Struktur aus voneinander weitgehend unabhängigen Konsortien der Einzeldisziplinen, eine absehbare Entwicklung nicht ausreichend vorweg zu nehmen. Die Querschnittaspekte der Konsortien der einzelnen Disziplinen – von Überlegungen zum wissenschaftstheoretischen Hintergrund über die konkrete Umsetzung einer FDM-Infrastruktur bis zu tragfähigen Konzepten zur digitalen Transformation in der Breite der akademischen Community und der Gesellschaft als Ganzem – müssten stärkere Berücksichtigung finden. Zum jetzigen Zeitpunkt hätte das *Bundesministerium für Bildung und Forschung* (BMBF) die Chance diese Prozesse mit zusätzlichen Mitteln anzustoßen und den Wissenschaftsstandort Deutschland in Sachen Digitalisierung wieder in eine Führungsrolle zu bringen.

Derart tiefgreifende Veränderungen sind nicht in der Breite realisierbar, wenn sie allein von den NFDI-Konsortien umgesetzt werden sollen. Stattdessen können alle zukünftigen Forschenden mit deutlich geringerem Aufwand schon im Studium erreicht werden, wie oben im Fall des physikalischen Grundpraktikums skizziert. Dabei muss gewährleistet werden, dass alle Studierenden, die das Grundpraktikum absolvieren, in gleichem Maße von den Neuerungen profitieren. Insbesondere für Studierende, die das Praktikum im Nebenfach oder als Importmodul absolvieren, muss ebenfalls ein Mehrwert entstehen, etwa durch die Betonung methodischer anstatt fachlicher Kompetenzen. Mit dieser Schwerpunktsetzung wird ein transdisziplinärer Arbeitsansatz implizit mit gestärkt, bei dem, im Gegensatz etwa zur interdisziplinären Forschung, der Gegenstand der Forschung und die verwendeten Methoden ins Zentrum rücken und die unterschiedlichen Sichtweisen einzelner Disziplinen in den Hintergrund treten. Voraussetzung hierfür ist die strukturelle Integration von FAIRem FDM in den Lernzielen von Studiengängen und in möglichst verschiedenen Modulen im gesamten Studium.

Ein Realitätscheck

Die gesetzlichen Regelungen sehen vor, dass die Qualifikationsziele von Studiengängen dem Abschlussniveau entsprechen und zu einer wissenschaftlichen Tätigkeit bzw. einer qualifizierten Erwerbstätigkeit befähigen[14]. Insbesondere folgt daraus, dass in der Studiengangentwicklung nicht nur der Status quo Berücksichtigung finden muss, sondern auch absehbare Entwicklungen der kommenden Jahre vorweggenommen werden müssen.

Typische Physikstudiengänge sind sehr stark inhaltsbasiert konzipiert – es werden die traditionellen Themen Mechanik, Elektrodynamik, Optik, Quantenmechanik, etc. gelehrt. Die meisten Studiengänge sind auf den Empfehlungen der *Konferenz der Fachbereiche Physik* (KFP) zum Bachelor- und Masterstudium[15] von 2010 bzw. dem vorhergehenden Papier zur Umstellung der früheren Diplomstudiengänge[16] von 2005 aufgebaut, die sich wiederum auf eine lange Tradition der inhaltlichen Gestaltung von Physikstudiengängen berufen. In der Astronomie oder Teilchenphysik sind Experimente häufig zu groß oder kostenintensiv, um sie an anderer Stelle nachzubauen. So kommt beim ATLAS-Experiment am CERN bereits eine FDM-Lösung zum Einsatz[17]. Hier werden kontinuierlich Daten akkumuliert und damit die statistische Signifikanz der Messungen verbessert. Ähnliche Repositorien sind z.B. mit dem MAST-Archiv[18] in der Astronomie etabliert. Das Verfügbarmachen experimenteller Rohdaten ist hier von entscheidender Bedeutung, damit die Datenanalyse vollständig nachvollziehbar ist und somit ein Mindestmaß an Reproduzierbarkeit gewährleistet werden kann.

Diese Beispiele zeigen, dass sich die Rahmenbedingungen des Physikstudiums und des Forschungsalltags funda-

mental verändert haben. Allerdings scheint dies bisher keine wesentliche Veränderung der Art, wie Physik in Schule und Hochschule unterrichtet wird, bewirkt zu haben. Man muss fragen, ob die Ziele physikalischer Bildung nicht mittlerweile an den gesellschaftlich relevanten Themen vorbeigehen. Eine digitale und interaktive Gestaltung der Praktika, sodass Studierende neben der Arbeit mit Daten auch die Entwicklung experimenteller Aufbauten erlernen, ist dringend notwendig. Außerdem wäre eine Nutzung – bevorzugt quelloffener und frei lizenzierter – digitaler Tools, die dazu geeignet sind, kollaboratives und studierendenzentriertes Lernen zu fördern, wünschenswert. Damit würde auch die gesetzlich vorgesehene kompetenzorientierte Überprüfung von Lernzielen[14] deutlich vereinfacht.

Die niedrig hängenden Früchte

Die tiefer gehenden Aspekte von FDM-Systemen bedeuten weitreichende Umstrukturierungen in Forschung und Lehre. Allerdings können, wie bereits am Beispiel des physikalischen Grundpraktikums beschrieben, erste Schritte sehr leicht gegangen werden. Etwa könnten Studierendengruppen leicht unterschiedliche Versuche durchführen, ihre Messdaten miteinander teilen und dann jeweils eine Auswertung unter Berücksichtigung aller Messdaten anfertigen, ohne, dass die Bewertung der individuellen Leistungen der Studierenden im prüfungsrechtlichen Sinne beeinträchtigt wäre. Hierbei wird nicht nur wie bisher das Experimentieren und Protokollieren der Ergebnisse erlernt, sondern auch das Aufbereiten der eigenen und der Umgang mit fremden Daten. Darüber hinaus sollte ein kritischer Umgang mit den eigenen Daten hinsichtlich der Ursachen von statistischen und systematischen Abweichungen, der Überprüfung theoretischer Erwartungen, und der Aussagekraft der durchgeführten Messreihen gefördert werden.

Weiter könnten allein stehende Module, beispielsweise in Master-Studiengängen im Wahl- oder Wahlpflichtbereich, geschaffen werden. Dies öffnet die Möglichkeit, in der Konzeption und Durchführung des Moduls verschiedene Disziplinen zu beteiligen und deren jeweilige Rezeption von Forschungsdaten zu berücksichtigen und gleichzeitig übergreifende Konzepte wie FDO, Repositorien, Metadaten und die FAIR-Prinzipien thematisieren. Soweit diese auf konkrete, interdisziplinäre Beispiele angewendet werden, wird in einem solchen Modul sehr einfach eine objekt- und methodenbasierte, transdisziplinäre Arbeitsweise vermittelt. Für Promovierende und Forschende in den ersten Karriereabschnitten können passende Angebote in Form von Summer Schools geschaffen werden. Diese sollten sowohl fachspezifische Inhalte als auch Querschnittthemen im zuvor beschriebenen Sinne beinhalten.

Entsprechend der vorzunehmenden Änderungen müssen die vorgesehenen Lernziele angepasst und um entsprechende digitale und datenspezifische Kompetenzen und Qualifikationsziele erweitert werden. Das „Hochschulforum Digitalisierung“ hat dazu unter Beteiligung der *Hochschulrektorenkonferenz* (HRK) ein Diskussionspapier[19] vorgelegt. Nach Abschlussniveaus gestufte Vorschläge der ZaPF zur Integration von FDM-Inhalten in Physikstudiengänge[20] werden zeitnah veröffentlicht. Auf dieser Basis können im nächsten Akkreditierungszyklus entsprechende Lernziele in Studiengänge integriert werden.

Auf diese Weise entsteht ein Bewusstsein für einen FAIRen Umgang mit Daten. Gleichzeitig wird mit niedrigem Aufwand ein großer Personenkreis erreicht. Dies wird auch die langfristigen Ziele der NFDI – die Schaffung breit angelegter Repositorien und Data Lakes – deutlich vereinfachen, da der erste Schritt zu einem systematischen FDM bereits getan ist. Vor allem aber können Studierende und Forschende durch eine vereinfachte Kontrollmöglichkeit ihrer eigenen Ergebnisse und bessere Nachvollziehbarkeit und Weiternutzbarkeit von Publikationen unmittelbar von zitierbaren Datensätzen und deren vereinfachter Vergleichbarkeit profitieren.

Referenzen

[1] https://eosc-portal.eu/about/eosc

[2] https://www.dfg.de/foerderung/programme/nfdi/

[3]Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. *The FAIR Guiding Principles for scientific data management and stewardship*. Sci Data **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

[4]De Smedt, K., Koureas, D., & Wittenburg, P. (2020). *FAIR digital objects for science: from data pieces to actionable knowledge units*. Publications, **8(2)**, 21.

[5]*Positionspapier zu FAIR und Open Data im physikalischen Praktikum*, ZaPF, 2020 https://zapfev.de/resolutionen/wise20/opendata/opendata.pdf

[6]DOI Handbook, https://dx.doi.org/10.1000/182

[7]Zenodo Open Library, OpenAIRE/CERN https://zenodo.org/

[8]Open Science Framework, Center for Open Science https://osf.io/

[9]G. E. Moore, *Cramming more components onto integrated circuits, Reprinted from Electronics*, volume **38**, number 8, April 19, 1965, pp.114 ff., in IEEE Solid-State Circuits Society Newsletter, vol. 11, no. 3, pp. 33-35, Sept. 2006, https://dx.doi.org/10.1109/N-SSC.2006.4785860

[10]Ob Moore's Gesetz noch gilt und wie lange es Gültigkeit behalten wird ist umstritten, siehe z.B. https://arxiv.org/abs/1511.05956

[11]FLOP: *floating point operation per second* – Anzahl and Rechenoperationen pro Sekunde. 1 ExaFlop = 10^{18} FLOP

[12]T. Ishikura, *No contest: Japan's Fugaku again fastest supercomputer*, The Ashai Shimbun, 2020, http://www.asahi.com/ajw/articles/13938448

[13]RfII – Rat für Informationsinfrastrukturen, *Entwicklung von Forschungsdateninfrastrukturen im internationalen Vergleich*, Göttingen (2017). https://rfii.de/?p=2346

[14]Studienakkreditierungsstaatsvertrag Art. 2 Abs (3), https://www.akkreditierungsrat.de/sites/default/files/downloads/2019/Studienakkreditierungsstaatsvertrag.pdf

[15]https://www.kfp-physik.de/dokument/KFP_Handreichung_Konzeption-Studiengaenge-Physik-101108.pdf

[16]https://www.kfp-physik.de/dokument/Empfehlungen_Ba_Ma_Studium.pdf

[17]ATLAS Open Data, https://atlas.cern/resources/opendata

[18]*Mikulski Archive for Space Telescopes* (MAST), https://archive.stsci.edu/about-mast

[19]Hochschulforum Digitalisierung. *20 Thesen zur Digitalisierung der Hochschulbildung*, Arbeitspapier Nr. **14**, Berlin, 2015. https://hochschulforumdigitalisierung.de/sites/default/files/dateien/HFDAPNr14_Diskussionspapier.pdf

[20]*Einbindung von Forschungsdatenmanagement in der Lehre*, ZaPF, 2021. https://zapfev.de/resolutionen/sose21/fdm/fdm.pdf