EOSC Support Office Austria: Visionen, Bedürfnisse und Anforderungen an Forschungsdaten und -praktiken

Katharina Flicker (TU Wien), Thomas Hofer (Universität Innsbruck)

Dieses Interview ist auch als Download verfügbar: https://doi.org/10.5281/zenodo.7855137

Im Jahr 2015 führte die Vision eines föderierten Systems von Infrastrukturen zur Unterstützung der Forschung durch die Bereitstellung einer offenen, multidisziplinären Umgebung für die Veröffentlichung, Suche und Wiederverwendung von Daten, Werkzeugen und Diensten zum Start des Aufbaus der European Open Science Cloud (EOSC). Daher wurden Einrichtungen wie die EOSC Association auf europäischer Ebene und das EOSC Support Office Austria auf nationaler Ebene gegründet.

In diesem Rahmen und da Forschung schon immer im Mittelpunkt der EOSC stand, erheben wir Visionen, Bedürfnisse und Anforderungen an Forschungsdaten und -praktiken von Forschenden, die an öffentlichen Universitäten in Österreich tätig sind. Das folgende Interview wurde mit dem theoretischen Chemiker Thomas Hofer geführt:

„Die kritische Analyse und Auseinandersetzung mit Daten ist immer ein Muss."

KF: Würden Sie mir zu Beginn bitte Ihr Forschungsfeld beschreiben?

TH: Ich forsche auf dem Gebiet der Theoretischen Chemie, bzw. genauer im Bereich *Theoretical and Computational Chemistry*. In dieser Disziplin werden chemische Prozesse und Verbindungen nicht im Labor, sondern ausschließlich mit Berechnungen am Computer untersucht. D.h. ich arbeite unter anderem mit Quantentheorie und einer langen Liste abstrakter Berechnungsmethoden, um damit chemische Strukturen zu untersuchen.

KF: Mit welchen Daten arbeiten Sie?

> "Natürlich wären Qualitätskontrollen durchaus sinnvoll, aber bei der Fülle an Daten, die generiert werden, wüsste ich jetzt nicht, wie man das implementieren könnte."

TH: In diesem Feld arbeite ich vor allem mit Struktur- und Kristallographiedaten. Strukturdaten generieren wir mit Hilfe diverser Berechnungsmethoden selbst. Dabei werden 3D-Strukturen einzelner Atome im Raum bestimmt. Bei Kristallographiedaten handelt es sich um

Messungen chemischer Verbindungen, aus denen zuerst im Labor Kristalle gezüchtet werden, die dann mit Hilfe verschiedener Röngtenmethoden vermessen werden können. Für Kristallographiedaten gibt es etablierte Datenbanken, wie z.B. *The Materials Project*, die *Protein Data Bank PDB* oder jene des *Cambridge Crystallographic Data Centres*. Hier ist streng reglementiert, welche Daten überhaupt eingereicht werden dürfen, üblicherweise in bereits vorgefertigten Formaten. Im *The Materials Project* sind die Strukturen zwar stark vereinfacht, können allerdings durchaus als Startpunkt für Berechnungen dienen. *PDB* ist ein enormes Repository im Bereich der Life Sciences.

KF: Welche vorgefertigten Formate wären das?

TH: Das Bekannteste ist das CIF-File Format, was für *Crystallographic Information File* steht. Es handelt sich dabei um eine der gängigsten Strukturformate für Kristallographiedaten und wird häufig als Supplement zu einer Publikation oder schon direkt als Supplement-implementierung zum Download zur Verfügung gestellt.

> "Was ich mir noch am Ehesten vorstellen könnte, ist, Qualitätskontrollen der Daten als Teil des Peer-Reviewing-Prozesses zu implementieren (...)."

Auch für Strukturdaten würde es sich anbieten CIF-Files zu übernehmen. Die meisten Programme aus der Kristallographie sowie die meisten Workflows führen letztlich ohnehin zu einer dieser Dateien und sie zu erstellen ist aufgrund der vielen Regeln und Normen, wie dieses File auszusehen hat, in der Praxis doch recht aufwendig.

KF: Sie haben vorhin auf stark vereinfachte Daten, die zumindest als Startpunkt für Forschungsarbeiten dienen können angespielt. Gibt es in Ihrer Disziplin zentrale Qualitätsmerkmale im Zusammenhang mit Daten?

TH: Die kritische Analyse und Auseinandersetzung mit Daten ist immer ein Muss. In meiner Arbeitsgruppe haben wir ein Sprichwort: wenn man Daten nicht geprüft hat, dann werden sie falsch sein. Ich muss wirklich jeden einzelnen Schritt von Beginn an überprüfen, weil die Methodik unglaublich fehleranfällig ist.

Ein wesentliches Qualitätsmerkmal sind außerdem Peer-Reviewed Publikationen. Solche Publikationen umfassen die Daten, in der Form in der sie gemessen wurden, eine Schlussfolgerung und meist auch Screenshots. Trotzdem muss man auch hier die Daten penibel überprüfen – irgendein Hindernis tritt normalerweise immer auf, das teilweise aus Perspektive der Experimentellen Chemie irrelevant ist, aber in der Theoretischen Chemie in der Computerberechnung zu Fehlern führen kann.

Leider gibt es oft keine Qualitätskontrolle der Daten an sich. Und das wiederum könnte durchaus ein Hindernis für viele Forschende sein, die sich mit der Bereinigung der Daten nicht befassen können, weil Kompetenzen im Bereich der Computer Sciences und Data Manipulation fehlen. In meiner Arbeitsgruppe schreiben wir beispielsweise

durchaus Programme / Skripte, um Daten zu bereinigen. Wenn Forschende lediglich Analysesoftware im Anwendungsbereich verwenden, wird das natürlich schwieriger bzw. zum Hindernis, um überhaupt mit den Daten zu arbeiten.

KF: Könnte man Qualitätskontrollen einführen bzw. wie könnten diese gestaltet werden?

TH: Das ist schwierig. Natürlich wären Qualitätskontrollen durchaus sinnvoll, aber bei der Fülle an Daten, die generiert werden, wüsste ich jetzt nicht, wie man das implementieren könnte. Ich glaube nicht, dass das automatisierbar wäre und für *Human Intervention* bräuchte man wirklich Expertinnen und Experten auf den einzelnen Gebieten.

Was ich mir noch am Ehesten vorstellen könnte, ist, Qualitätskontrollen der Daten als Teil des Peer-Reviewing-Prozesses zu implementieren und gleichzeitig strengere Standards einzuführen. Aber da stellen sich dann wieder Fragen danach, ob man ReviewerInnen diesbezüglich überhaupt in die Pflicht nehmen kann oder welche Konsequenzen bei Nicht-Einhaltung der Standards drohen. Aktuell bin ich jedenfalls nur bereit Daten zu vertrauen, die im Kontext einer peer-reviewed Publikation auch veröffentlicht wurden bzw. die wir selbst mühsam nachkontrollieren.

“Aktuell bin ich jedenfalls nur bereit Daten zu vertrauen, die im Kontext einer peer-reviewed Publikation auch veröffentlicht wurden bzw. die wir selbst mühsam nachkontrollieren.“

KF: Würden Sie das bitte genauer ausführen?

TH: Gerne. Im Idealfall würden nur Daten veröffentlicht und akzeptiert werden, die in einem Journal mit peer review veröffentlicht wurden. Beispielsweise bin ich kein Fan von Daten, die im Zusammenhang mit Pre-Prints erschienen sind – obwohl es einen Trend dazu zu geben scheint. Viele dieser Pre-Prints bleiben leider in genau dieser Stufe hängen und werden nie in einem anerkannten Journal veröffentlicht. Hier stellt sich schon die Frage, was mit diesen Publikationen nicht stimmt und wie zuverlässig Daten aus diesen Artikeln sind, die von Reviewern nicht akzeptiert wurden. Leider gibt es manchmal auch keine Möglichkeit Pre-Prints von tatsächlichen Veröffentlichungen auf einen Blick zu unterscheiden. Gäbe es eine Art Markierung, die den accepted-Status eines Papers unmittelbar sichtbar macht, wäre das meiner Meinung nach schon sehr praktisch.

Natürlich ist es möglich auch mit Daten aus Pre-Prints zu arbeiten, es ist aber besonderes Augenmerk auf die Überprüfung, Bereinigung und Korrektur der Daten zu legen. Ich persönlich würde aber mit solchen Daten nicht arbeiten oder Pre-Prints zitieren wollen, da meine Forschung dann durchaus einen Qualitätsmangel aufweisen würde.

KF: Und wenn offen gelegt werden würde, warum diese Artikel im Pre-Print-Status hängen geblieben sind?

TH: Vermutlich könnte man Forschende zwingen, Gutachterkommentare für die Ablehnung öffentlich zu machen. Aber ich denke nicht, dass Forscherinnen und Forscher das wollen würden – ich zumindest würde das nicht wollen. Es wäre auch zu hinterfragen, wie sinnvoll das am Ende dann wäre. Beispielsweise könnten Artikel auch einfach auf der Pre-Print Stufe bleiben, wenn Forschende ihre Arbeitsgruppe verlassen und sich für eine bestimmte Publikation einfach nicht mehr zuständig fühlen.

KF: Ich verstehe.

TH: Es gibt im Übrigen ein Thema, das mich – aus meiner Disziplin kommend – im Zusammenhang mit der EOSC beschäftigt.

KF: Das wäre?

TH: Die EOSC sollte ein multidisziplinäres Umfeld sein, in dem wir als Forschende Daten veröffentlichen, finden und wiederverwenden können. Ich frage mich also, welche Erwartungen / Forderung an die Forschungsgemeinde herangetragen werden. Welche Daten sollen wir in welcher Form teilen bzw. was würden wir im Kontext der Wiederverwendung von Daten in welcher Form finden? Wie lange würden welche Daten aufbewahrt werden? Wie soll der Zugriff geregelt werden? Wie erfolgt die Kommunikation bzw. das Feedback zwischen Forschenden, die Daten hochladen und jenen, die die Daten dann wiederverwenden? Das sind alles wesentliche Usability-Fragen, die ich gerne klar beantwortet hätte. Für mich ist das alles zum gegenwärtigen Zeitpunkt leider sehr undurchsichtig.

KF: Gute Punkte. Vielen Dank für das Interview.

Dr. Thomas S. Hofer promovierte im Fach Chemie an der Universität Innsbruck. Nach Auslandsaufenthalten an der ETH Zürich und der University of Cambridge habilitierte er sich im Bereich Theoretische Chemie und Computerchemie. Seit 2011 ist Dr. Hofer assoziierter Professor am Institut für allgemeine, anorganische und theoretische Chemie der Universität Innsbruck. Der Schwerpunkt seiner Forschung liegt im Bereich der computerunterstützten Materialwissenschaften, mit speziellem Fokus auf funktionelle Nanomaterialien und neuartigen Energietechnologien.