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

Katharina Flicker (TU Wien), Josef Küng (JKU)

Dieses Interview ist auch als Download verfügbar: https://doi.org/10.5281/zenodo.8051728

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 Informatiker Josef Küng geführt:

"Der Zugang zu Daten ist für die Forschung enorm wichtig."

KF: Womit beschäftigen Sie sich im Rahmen Ihrer Forschungstätigkeit?

JK: Derzeit beschäftige ich mich mit Zugriffskontrollmechanismen auf Daten, Data Analytics und Similiarity Queries.

KF: Mit welchen Daten arbeiten Sie bzw. wie arbeiten Sie mit Daten?

JK: Die Arbeit mit großen Datenbeständen ist eine häufige Anforderung. In einer unserer jüngeren Publikationen im Bereich der Zugriffskontrollmechanismen haben wir beispielsweise Performancemessungen durchgeführt. Dafür brauchen wir einen großen Datenbestand, um diesen als Testdaten verwenden zu können.

Große Datenbestände sind auch relevant, wenn es darum geht Methoden miteinander zu vergleichen. Wenn wir unsere Arbeit mit jener aus der Literatur vergleichen, ist es außerdem sinnvoll mit den gleichen Datensätzen zu arbeiten, damit wir einen direkten Vergleich durchführen können. Je nach Fragestellung suchen wir also nach geeigneten Datenbeständen oder versuchen Zugriff auf bestimmte Datenbestände zu bekommen.

KF: Wo und wie genau suchen Sie nach diesen Datenbeständen?

> "Die Arbeit mit großen Datenbeständen ist eine häufige Anforderung."

JK: Zum einen sind wissenschaftliche Publikationen hilfreich. Üblicherweise enthalten diese nämlich einen Link zu den Daten oder beschreiben die Daten, die verwendet wurden, zumindest. Zum anderen kommt es vor, dass wir Datenbestände aus bestimmten Anwendungsgebieten brauchen. Bei solchen Forschungsarbeiten handelt es sich aber meist um Kooperationen mit Firmen – beispielsweise wenn Methoden für eben diese Anwendungsgebiete optimiert werden sollen. In

diesem Fall versuchen wir Datenbestände auch von unseren Partnern zu bekommen.

Allerdings ist es nicht wirklich leicht an Datenbestände zu kommen, mit denen wir arbeiten können und wollen. Bei Datenbeständen, auf die wir durch Publikationen stoßen, ist es häufig auch ein Glücksfall, wenn wir tatsächlich genau die selben Daten verwenden können. Oft sind sie nämlich nicht verfügbar.

KF: Angenommen, der Datenbestand ist gefunden. Was macht diesen Datenbestand zu einem „guten" Datenbestand?

JK: Qualitätschecks müssen durchgeführt werden. Die meisten Datenbestände weisen irgendwo Lücken auf oder sind teilweise fehlerhaft. Die Daten vorzubearbeiten bzw. aufzubereiten, damit sie sauber und konsistent sind, ist oft ein enormer Aufwand.

KF: Welche Formen von Qualitätskontrollen gibt es?

> "Ein Referenzsystem dieser Art könnte – ebenso wie Qualitätskontrollen und Beschreibungen – das Vertrauen in die Datenqualität erhöhen."

JK: Überprüfen kann ich die Integrität des Datenbestandes, also seine Vollständigkeit, Konsistenz und Korrektheit. Die Vollständigkeit von Teilen des Datenbestandes kann man relativ leicht prüfen. Aber es gilt hier vorsichtig zu sein: Nehmen wir an, wir arbeiten mit Sensordaten, wobei es einige Ausreißer in den Daten gibt. Ob ich diese eliminieren will, oder nicht, kann durchaus von der eigentlichen Forschungsfrage abhängen. Korrektheit – also ob das, was in den Daten steht, auch tatsächlich der Realwelt entspricht – ist nur schwierig zu überprüfen. Außerdem gibt es ja auch künstlich erstellte Datensätze.

KF: Da wir gerade über Daten sprechen sowie vor dem Hintergrund der EOSC, die ja als föderiertes System von Infrastrukturen, in dem Forscher unter anderem Daten veröffentlichen, finden und wiederverwenden können, gedacht wird: was würde das Vertrauen in die Qualität von Daten eröhen?

JK: Ich denke, einerseits müssten Standardmethoden zur Qualitätsüberprüfung zur Verfügung gestellt werden. Andererseits müssten Testdaten bzw. Datensätze gut beschrieben werden. Eine gute Beschreibung beinhaltet, ob ein Datensatz bereits Qualitätskontrollen durchlaufen hat, welche Qualitätskontrollen das waren sowie das Ergebnis. Ebenfalls praktisch kann es sein, klar erkenntlich zu machen, wenn ein Datensatz keine Qualitätskontrollen durchlaufen hat. Manchmal ist es nämlich notwendig auch mit solchen Datensätzen zu arbeiten – wenn ich beispielsweise gegen qualitativ weniger hochwertige Datenbestände robuste Methoden entwickeln will.

KF: Wäre eine Art Gütesiegel als Qualitätsmerkmal sinnvoll?

JK: Nein. Ich wäre als Wissenschaftler mit einer guten Beschreibung, welche Qualitätsprüfungen dieser Datensatz durchlaufen hat, wesentlich glücklicher als wenn es da jetzt ein Ranking gäbe, von dem ich dann dann wieder nicht genau weiß, was das entsprechende Gütesiegel im Detail bedeutet.

Das Kommentieren von Datensätzen fände ich dagegen schon sinnvoll. Das heißt, verwendet jemand einen bestimmten Datensatz, muss kommentiert bzw. bewertet werden: Wer hat den Datenbestand wofür benutzt? Was sind die Erfahrungen mit dem Datenbestand? Wurden Fehler gefunden und wenn ja, welche? Welche wissenschaftlichen Erkenntnisse wurden damit generiert? Ein Referenzsystem dieser Art könnte – ebenso wie Qualitätskontrollen und Beschreibungen – das Vertrauen in die Datenqualität erhöhen.

KF: Ich wäre mit den Fragen durch. Gibt es von Ihrer Seite irgendetwas, das sie anmerken möchten?

> "Was den Zugang zu Daten betrifft, müsste außerdem auf eine Gleichstellung aller Forscherinnen und Forscher geachtet werden."

JK: Ja. Der Zugang zu Daten ist für die Forschung enorm wichtig. Vor allem, wenn die Datensätze bereits aufbereitet wurden bzw. von guter Qualität sind. Da der Aufwand, um Datensätze aufzubereiten sehr hoch ist, stellen solche Datensätze auch einen bestimmten Wert dar. Im Zusammenhang mit Daten aus industrieller Forschung bedeutet das, dass man sich überlegen müsste, was der Benefit für Unternehmen sein könnte, die Daten zur Verfügung stellen. Daten mühsam und teuer aufzubereiten und dann gratis oder günstig zu teilen, ist nämlich sicher keine Motivation.

Was den Zugang zu Daten betrifft, müsste außerdem auf eine Gleichstellung aller Forscherinnen und Forscher geachtet werden. Ansonsten könnte es passieren, dass einige privilegierte Forschungsgruppen mit Zugang einen klaren Vorteil haben gegenüber anderen Forschungsgruppen aus möglicherweise wirtschaftsschwächeren Ländern. Das würde ich mir zumindest innerhalb der EU bei von der Europäischen Kommission finanzierten Projekten wünschen. Letztendlich ist das aber eine politische Entscheidung.

KF: Vielen Dank für das Interview.

Josef Küng is associate professor at the Institute of Application Oriented Knowledge Processing (FAW) at Johannes Kepler University Linz (JKU). He holds a Ph.D. in Informatics and got his habilitation for Applied Computer Science there in 2000. His core competencies cover Information Systems, Security in Information Systems, Knowledge Based Systems, Semantic Technologies and Similarity Queries where he has published a fair number of scientific papers. He has been and still is member of program committees (also in chairing and editing functions) and steering committees of recognized scientific journals and conferences (e.g. DEXA, International Journal of Web Information Systems). Additionally, he is very active in co-operation projects with partners from industry, business and administration.