Digitale Forschungswerkzeuge

Nachhaltigkeit für Software und Daten

| WOLFGANG MAUERER | STEFANIE SCHERZINGER |

Die wissenschaftliche Reproduktionskrise hat den Blick auf digitale Forschungswerkzeuge intensiviert. Auch wenn der Mehraufwand für Reproduzierbarkeit und Zugänglichkeit zunehmend anerkannt wird, existieren noch Defizite in der Umsetzung, wenn es darum geht, die Datenbasis und Forschungswerkzeuge verfügbar zu machen.

Reproduzierbarkeit zielt auf eine möglichst objektive, unabhängige Überprüfung von Resultaten durch Dritte, mit dem gleichen experimentellen oder methodischen Ansatz der Originalstudie. Noch anspruchsvoller ist Replizierbarkeit – anderes Team, anderer Ansatz, gleiche Fragestellung. Beide Eigenschaften sind wesentliche Qualitätskriterien vieler empirischer Felder und ermöglichen einen großen Schritt in Richtung wissenschaftlicher Nachhaltigkeit. Auch wenn die Konzepte in Lebenswissenschaften, Medizin und anderen Disziplinen mit ausgeprägter individuenspezifischer Heterogenität oder bei der Untersuchung singulärer Ereignisse problematisch sein können, dürften die Voraussetzungen dank omnipräsenter elektronischer Datenverarbeitung niemals besser gewesen sein, zumindest in den konsolidierenden Phasen des Forschungsprozesses. Die Reproduktionskrise zeigt, dass es derzeit allerdings noch an Umsetzungskompetenz mangelt.

AUTOREN

Wolfgang Mauerer ist Professor für Informatik an der OTH Regensburg sowie Direktor des Regensburg Center for Artificial Intelligence.

Stefanie Scherzinger ist Inhaberin des Lehrstuhls für Informatik mit Schwerpunkt Skalierbare Datenbanksysteme an der Universität Passau.

Vorbild Industrie

In der Industrie ist Reproduzierbarkeit eine oft selbstverständliche Anforderung: Konstruktionsschritte für Maschinen müssen Jahrzehnte nach der Anfertigung von Bauplänen exakt wiederholbar sein und komplexe Programmcodes müssen von menschen- in maschinenlesbare Form transformierbar bleiben. Die Industrie hat starkes Interesse und hinreichend finanzielle Mittel, um die dazu notwendige Software selbst über Jahrzehnte einsatzfähig zu halten; davon kann die akademische Forschung profitieren. Proprietäre Systeme, die häufig zur Archivierung von digitalen Artefakten und den zugehörigen Werkzeugen enthusiastisch beworben werden, stellen aus Sicht der Autoren ein zu großes Risiko für den langfristigen Zugriff dar, da sie erfahrungsgemäß exakt bis zum Ende der Projektfinanzierung gepflegt werden.

Kommunale Offenheit

Quelloffenheit ist das Schlagwort der Stunde: Open Source-Codes, die neben Konfigurationsmanagement und detaillierter Dokumentation des Forschungsprozesses auch vollautomatisierte Bau-, Analyse- und Auswertungsumgebungen ermöglichen und Software-Werkzeuge zur Visualisierung von Daten und der finalen Produktion von Artikeln bereitstellen, werden insbesondere in Naturwissenschaft und Technik auf breiter Front eingesetzt. Aber es sind weitere Mechanismen verfügbar, die (effektiven) Determinismus in der wissenschaftlichen Produktionskette ermöglichen. Sie erlauben die Abstraktion von spezifischer Hardware und verwalten komplexe Abhängigkeiten zwischen Softwarepaketen und den zahlreichen notwendigen externen Softwarebibliotheken so, dass eine Wiederauswertung auch auf einer einsamen Insel ohne Internetanbindung (und daher auch 20 Jahre in der Zukunft) möglich ist. In Zeiten gigabyteschwerer, sich schnell inkompatibel verändernder Basisplattformen ist dies kein technisch einfaches, aber dennoch machbares Unterfangen. Institutionell verankerte Dienste wie z.B. Zenodo halten Reproduktionspakete über einen stabilen Link (DOI) langfristig verfügbar. Unter der Voraussetzung, dass die Pakete in sich abgeschlossen sind, also keine externen Abhängigkeiten auf Software-Bibliotheken besitzen oder von Internetressourcen abhängen, bleibt ihre Funktionsfähigkeit bewahrt.

Quid pro quo

Durchgängige Reproduzierbarkeit bringt Vorteile für die Wissenschaftsgemeinschaft: Kollektive Erkenntnis – die Summe wissenschaftlicher Publikationen – gewinnt an Wert, wenn neben publizierten Artikeln auch der detaillierte Weg zu den vorgelegten Ergebnissen verfügbar ist.

Der Wunsch nach Herausgabe von Forschungsartefakten basiert meist auf dem Bemühen, das Zustandekommen publizierter Ergebnisse besser zu verstehen und einen fairen Vergleich mit

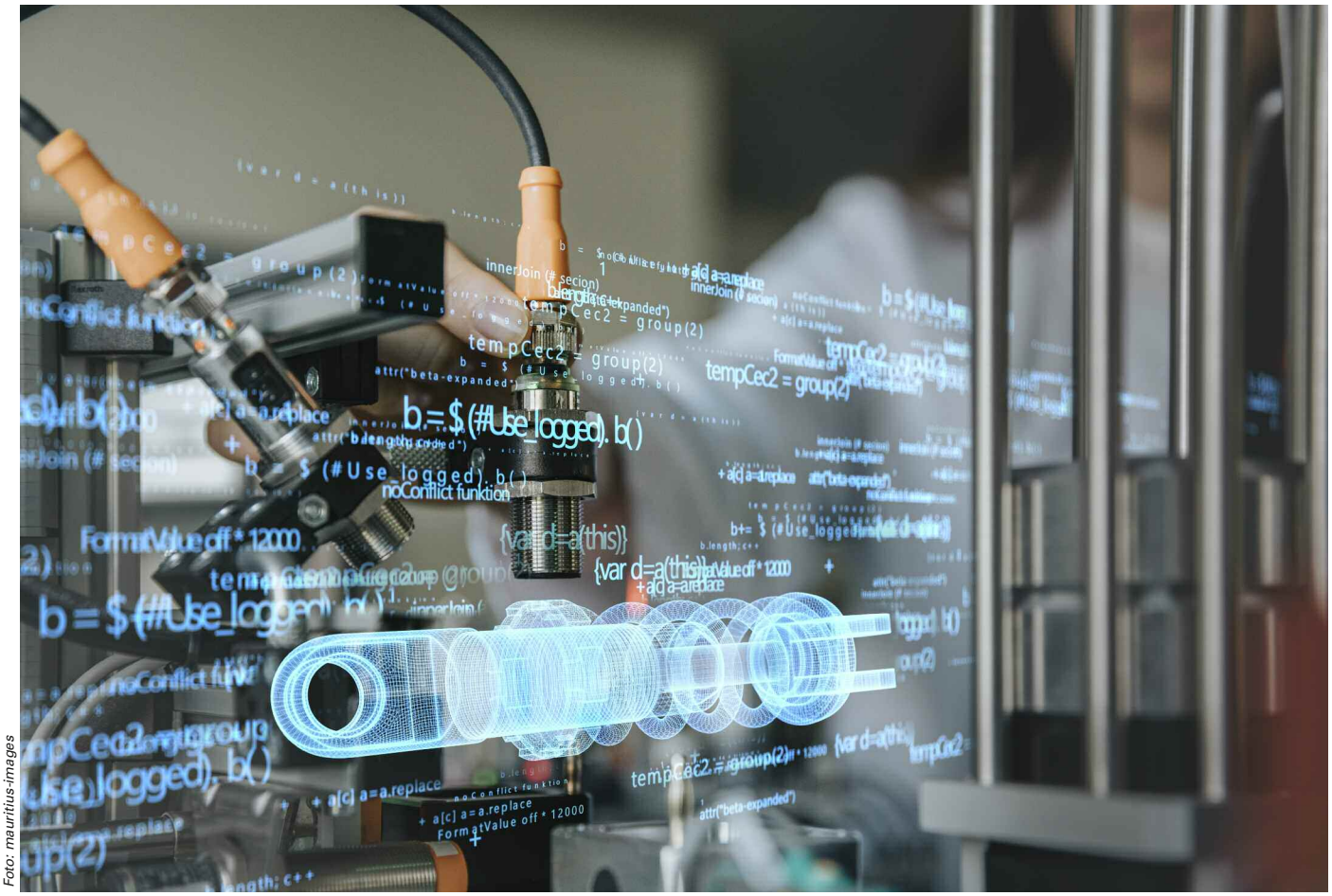

Foto: mauritius-images

den eigenen Resultaten ziehen zu können. Dass die Freigabe mitunter verwehrt wird, mag Bedenken entspringen, als Steigbügelhalter der Konkurrenz ausgenutzt zu werden. Hinzu kommt, dass hinter kommerziellen Bezahlschranken nicht-verfügbare Artikel als publiziert akzeptiert werden, während frei verfüg- und nutzbare Open Source-Supplemente gerne als unpubliziertes (und schwer zitierbares) Nebenprodukt abgetan werden. Tatsächlich bedeutet sauberes Reproducibility Engineering einen Mehraufwand von bis zu 30 Prozent. Ob in einer publish or perish-Umgebung die Zeit nicht besser in einen zusätzlichen Eintrag für die Publikationsliste investiert wäre?

Tatsächlich zeigt die praktische Erfahrung, dass Reproduzierbarkeit ein nicht zu unterschätzender Vorteil sein kann: Ein Wechsel auf neue IT-Infrastruktur ist z.B. dann kein Problem, wenn es mehrere Rechner gibt, auf denen eine Auswertungspipeline lauffähig ist, oder mehrere Personen im Team diese mit geringem Aufwand automatisiert aufsetzen können. Darüber hinaus wird auch eine standortübergreifende Zusammenarbeit einfacher, und ein Generationswechsel bei Promovenden kann glatt über die Bühne gehen. Zudem gibt es Anzeichen dafür, dass sich früher oder später alle Arbeitsgruppen dem Thema Reproduzierbarkeit werden stellen müssen, ungeachtet einer kurzfristigen Kosten-Nutzen-Rechnung.

Trendwende in Sicht

Ein entsprechender Trend ist nicht nur in der Informatik zu beobachten: Manche großen, internationalen Konferenzen erbitten bei Einreichungen die zur Reproduktion nötigen Artefakte, und erfolgreich reproduzierte Artikel werden gesondert ausgezeichnet. Erste Konferenzen (z.B. VLDB oder FSE) verlangen sogar eine Rechtfertigung, wenn Datenbasis und Forschungswerkzeuge nicht verfügbar gemacht werden. Auch Konferenzen auf nationaler Ebene (z.B. die GI-Konferenz BTW) richten Reproduktions-Komitees ein. Eine Wende von der Kür zur Pflicht scheint angestoßen.

Keine Ausnahmen

Neue Rechnertechnologien wie Quantencomputer bringen mit sich, dass die kommerziellen Betreiber der komplexen Maschinen die Kontrolle über die Ausführungsumgebung besitzen und beispielsweise Konfigurationen unangekündigt ändern können. Zudem produziert die Berechnung selbst inhärent stochastische Resultate – eine Herausforderung, die auch bei Monte-Carlo-Simulationen oder KI-Algorithmen auftritt. Selbst in solchen Umgebungen können Reproduktionspakete geschnürt werden, die neben den eigentlichen Eingabedaten auch Resultate aus proprietären Zwischenschritten, stochastischen Modellen und dergleichen enthalten.

Zukünftig Pflicht statt Kür

Aus Sicht der Autoren ist es absehbar, dass Inhalte wie Forschungsdatenmanagement in den akademischen Curricula verankert werden müssen; zudem sollte Reproduzierbarkeit ein überprüfbares und bewertetes Kriterium bei Abgabe von Abschlussarbeiten sein, vom Bachelor bis zur Promotion. Auch in Berufungskommissionen sollte das Thema stärker Beachtung erfahren. Die Informatik kann hier mit methodischen Serviceangeboten zu digitalen Forschungswerkzeugen einen wichtigen Beitrag leisten, um die Wissenschaft gestärkt aus der Reproduktionskrise herauszuführen.