Die Wahl der Mittel – Jupyter-Notebooks als Forschungsinfrastruktur

Jung, Kerstin
kerstin.jung@ims.uni-stuttgart.de
Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung

Hein, Pascal
pascal.hein@ilw.uni-stuttgart.de
Universität Stuttgart, Institut für Literaturwissenschaft

Blessing, André
andre.blessing@ims.uni-stuttgart.de
Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung

Hess, Jan
jan.hess@dla-marbach.de
Deutsches Literaturarchiv Marbach

Kushnarenko, Volodymyr
volodymyr.kushnarenko@hlrs.de
Höchstleistungsrechenzentrum (HLRS), Universität Stuttgart

Mit Python als vielgenutzter Programmiersprache in den Digital Humanities[1] steigt auch der Bedarf an Möglichkeiten zur nachhaltigen Weitergabe und Wiederverwendbarkeit von Python-Quellcode. Softwareentwicklungsnahen Lösungen wie der Verfügbarmachung über Versionsverwaltungsrepositorien oder dem Einpflegen in eine Paketverwaltung wurde der ‚Notebook'-Ansatz[2] zur Seite gestellt, der Dokumentation, Ausführung und Visualisierung verzahnt und eine Aufbereitung für verschiedene Zielgruppen ermöglicht.

Im Forschungskontext werden solche Notebooks daher verwendet, um auf einer (Web-)Seite Datensätze einzulesen, zu analysieren, visualisieren und die verwendete Methodik zu erläutern, ohne dies auf verschiedene Orte oder Zugänge verteilen zu müssen. In der (Nach-)Nutzung können z. B. Parameter in der Analyse oder Visualisierung direkt im Browser verändert werden und eine Anpassung ohne Programmierkenntnisse oder -erfahrung ermöglichen. Die Notebook-Dateien können wiederum über entsprechende Softwareentwicklungs-Repositorien zur Verfügung gestellt werden, was Anpassungen für weitere Datensätze oder Forschungsfragen erlaubt. Jupyter-Notebooks sind dabei als JSON-Dokumente strukturiert verarbeitbar.

Im Rahmen unseres Projekts geht es uns um die Möglichkeit, Jupyter-Notebooks so zur Verfügung zu stellen, dass sie für eine sehr heterogene Nutzendengruppe (u. a. Autor*innen, Forschende, Schüler*innen) einen Mehrwert bedeuten.[3] Wir möchten verschiedene Ebenen der Vorkenntnisse bedienen und gleichzeitig ermöglichen, eigene Forschungsfragen einzubringen. Im Beitrag soll aber auch der infrastrukturelle Aufwand verdeutlicht werden, der hinter der Möglichkeit nachhaltig ausführbarer Notebooks für die Forschung steht und auf einen Ausgleich zwischen Flexibilität der Nutzung und Sicherheit des Angebots hinausläuft. Letztendlich möchte der Beitrag die Diskussion befördern, inwieweit die Community von einer forschungsgetriebenen, unabhängigen und nachhaltigen Infrastruktur zum Umgang mit Jupyter-Notebooks profitieren würde, da ein entsprechendes Vorgehen für individuelle Projekte weniger umsetzbar ist.

Zu den Vorteilen der Bereitstellung von Zugängen zu Daten und Analysen durch Notebooks gehören (i) die Möglichkeit, ein Angebot an eine breite Nutzendengruppe zu machen: Je nach Aufbereitung der Notebooks (interaktive Elemente wie Dropdown-Menüs oder Range-Sliders sind möglich) können sie fast ohne Vorkenntnisse mit Python betrieben werden und an das individuelle Forschungsinteresse angepasste Ergebnisse produziert werden, (ii) dass die technischen Voraussetzungen, z. B. benötigte Pakete, im Notebook selbst spezifiziert sind. Diese Vorteile kommen allerdings nur in einer konfigurierten Ausführungsumgebung zum Tragen. Werden nur die Jupyter-Notebook-Dateien bereitgestellt, setzt das bei den Nutzenden Kenntnisse in Python, Bash o. Ä. sowie im Umgang mit Jupyter voraus. Oft sind Pakete in aufeinander abgestimmten Versionen erforderlich oder in Abhängigkeit vom Betriebssystem verfügbar, so dass nur eine vorkonfigurierte Umgebung den Nutzenden tatsächlich die technischen Hürden abnimmt.

So stellt sich die Frage, in welchem Rahmen ausführbare Jupyter-Notebooks zur Verfügung gestellt werden können. Der Betrieb einer zugänglichen Ausführungsumgebung (‚Hub') setzt Hardware, Administrations- und Wartungskapazitäten voraus. Eine Nutzungsverwaltung (Vergabe und Pflege von Accounts, Monitoring von Speicher- und Rechenkapazitäten) ist dabei ebenso unerlässlich wie Aktualisierungen mittels Updates auf Ebene von Maschine, Hub und Paketen und damit verbundene Wartungsarbeiten durch Abhängigkeiten in den Notebooks. Der Betrieb einer nachhaltigen Ausführungsumgebung setzt dies für einen längeren Zeitraum voraus, so dass die Idee der eigenen Ausführungsumgebung den Rahmen eines Forschungsprojekts oft übersteigt. Des Weiteren muss der Sicherheitsaspekt berücksichtigt werden, da es sich bei ausführbaren Jupyter-Notebooks um ausführbaren Quellcode handelt, der gewollt oder ungewollt Schaden am eigenen oder an externen Systemen verursachen kann.

Mit dem Service Colaboratory[4] bietet Google an , Notebooks einzurichten, Pakete dafür dauerhaft zu installieren und diese Notebooks ggf. mit Zugangsbeschränkung zu veröffentlichen. Dies zeigt die technische Möglichkeit, einen Notebook-Service mit großen Gestaltungsmöglichkeiten für die Nutzenden zu hosten. Allerdings kann diese Lösung für die Forschung nicht als Standard vorgeschlagen werden – allein aufgrund der problematischen Speicherung aller Daten, aber auch weil hier aufgrund des unvorhersehbaren Umgangs mit den eigenen Diensten die Nachhaltigkeit nicht gesichert werden kann.

Eine Alternative hierzu kann der Betrieb einer stark restringierten Ausführungsumgebung sein, die zwar die vorhandenen Notebooks abspielen kann und Nutzende ggf. aus vorgegebenen Parametern wählen lässt, Forschenden aber kaum Flexibilität bezüglich einer eigenen Exploration oder Einbindung weiterer Pakete ermöglicht.

Sofern spezifische technische Expertise angenommen werden kann, ist eine weitere Möglichkeit, Docker-Container zum Download zur Verfügung zu stellen oder eine detaillierte Dokumentation zur Nutzung eines Notebooks innerhalb einer integrierten Entwicklungsumgebung zu liefern. Die Zielgruppe wird damit allerdings auf Nutzende der entsprechenden Infrastruktur einschränkt.

Notebooks, die über bestimmte Repositorien öffentlich zur Verfügung gestellt werden, können über Binder[5] in eine Ausführungsumgebung gebracht werden. Dabei sind vor allem forschungsbezogene Repositorien wie Zenodo von Interesse, die gute Voraussetzungen für die langfristige Verfügbarkeit auf geschützten Servern bieten. Forschungsgetriebene Ansätze wie von GESIS[6] mit Binder (Bleier und Erdogan 2020) sowie Forschungsumgebungen mit Nutzendenverwaltung wie das DHVLab[7] und DH2go[8] (Heckelen et al. 2022), die die Ausführung von Jupyter-Notebooks erlauben, ermöglichen ggf. auch den Umgang mit spezifischeren Daten, sind aber ggf. auf Nutzende aus bestimmten Fachbereichen oder Institutionen beschränkt.

Ein entsprechender Ansatz für die breite Forschungscommunity wäre ein großer Gewinn bezüglich der Verfügbarmachung, Nachnutzung und Dokumentation von Forschungsmethoden und -ergebnissen.

Fußnoten

1. Vgl. die Wellen der Sprachen Fortran, Prolog, Perl und Python in der Korpusstudie von Burghardt et al. (2022).
2. Jupyter-Notebooks, https://jupyter.org/ (zugegriffen: 15. Dezember 2022).
3. Im Projekt SDC4Lit (Science Data Center for Literature, https://www.sdc4lit.de/, zugegriffen: 15. Dezember 2022) werden u.a. literarische Blogs zur Untersuchung zur Verfügung gestellt. Dabei kommen z.B. Tools zu Linkextraktion und Graphaufbau zum Einsatz, die über Jupyter-Notebooks bereitgestellt werden können.
4. https://colab.research.google.com/ (zugegriffen: 15. Dezember 2022).
5. https://mybinder.org/ (zugegriffen: 15. Dezember 2022).
6. https://notebooks.gesis.org/binder/ (zugegriffen: 15. Dezember 2022).
7. https://dhvlab.gwi.uni-muenchen.de/ (zugegriffen: 15. Dezember 2022).
8. https://dh2go.ilw.uni-stuttgart.de/ (zugegriffen: 15. Dezember 2022).

Bibliographie

Bleier, Arnim und Kenan Erdogan. 2020. *"A Persistent BinderHub: Democratizing Access to Computational Resources in the Social Sciences." JupyterCon2020 Online Conference*.

Burghardt, Manuel, Jan Luhmann und Andreas Niekler. 2022. *"Tools as Epistemologies in DH? A Corpus-Based Exploration."* In *Digital Humanities 2022. Conference Abstracts,* 144-146. Tokyo, Japan.

Heckelen, Malte, Claus-Michael Schlesinger und Fabienne Burkhard. 2022. *"Dh2go - Lehr- Und Lernumgebung Für Die Digital Humanities."* In *DHd2022: Kulturen des digitalen Gedächtnisses. Konferenzabstracts.* Zenodo. https://doi.org/10.5281/zenodo.6328013