

Software-Abhängigkeiten richtig beschreiben

Einführung in das Dependency Management für datenbasierte Projekte mit Open Source Tools

06.05.2021

Heinz-Alexander Fütterer

Referent für Forschungsdatenmanagement

<https://www.fu-berlin.de/forschungsdatenmanagement>

forschungsdaten
management

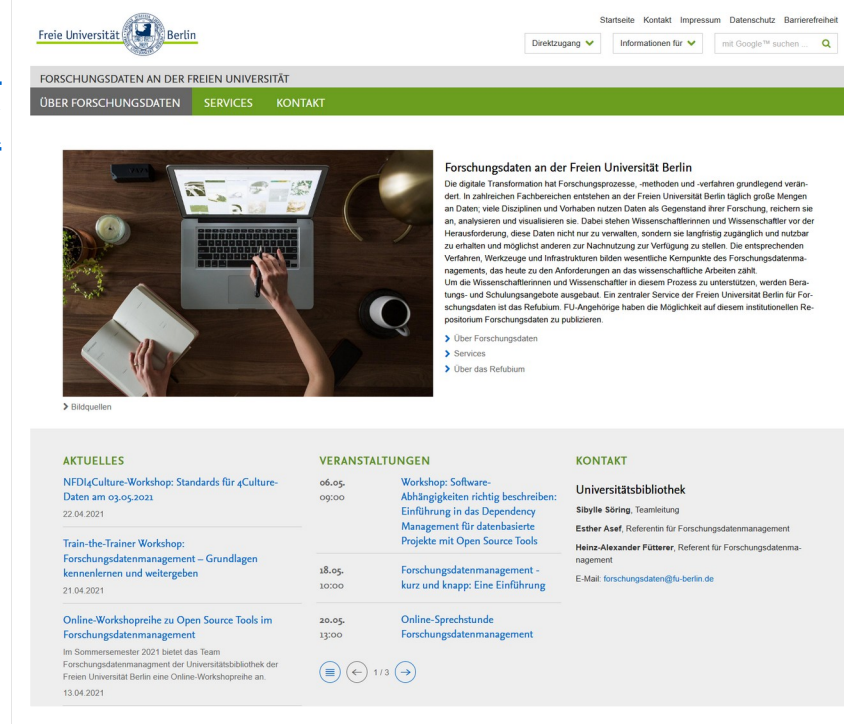


UNIVERSITÄTS
BIBLIOTHEK

0. WILLKOMMEN

Herzlich Willkommen

- Forschungsdatenmanagement-Team der Universitätsbibliothek
- Beratung
- Online-Sprechstunde
- Schulungen und Workshops
- Geplante Veranstaltungen



The screenshot shows the website of the Research Data Management team at the Free University of Berlin. The header includes the university logo and navigation links: Startseite, Kontakt, Impressum, Datenschutz, Barrierefreiheit. Below the header, there are links for 'FORSCHUNGSDATEN AN DER FREIEN UNIVERSITÄT', 'ÜBER FORSCHUNGSDATEN', 'SERVICES', and 'KONTAKT'. The main content area features a large image of a person working on a laptop, with a text block titled 'Forschungsdaten an der Freien Universität Berlin' explaining the digital transformation of research processes. Below this, there are three columns: 'AKTUELLES' (Current) with links to workshops and training sessions, 'VERANSTALTUNGEN' (Events) with a list of upcoming events, and 'KONTAKT' (Contact) with information about the library and contact details.

Forschungsdaten an der Freien Universität Berlin

Die digitale Transformation hat Forschungsprozesse, -methoden und -verfahren grundlegend verändert. In zahlreichen Fachbereichen entstehen an der Freien Universität Berlin täglich große Mengen an Daten, viele Disziplinen und Vorhaben nutzen Daten als Gegenstand ihrer Forschung, reichern sie an, analysieren und visualisieren sie. Dabei stehen Wissenschaftlerinnen und Wissenschaftler vor der Herausforderung, diese Daten nicht nur zu verwalten, sondern sie langfristig zugänglich und nutzbar zu erhalten und möglichst anderen zur Nachnutzung zur Verfügung zu stellen. Die entsprechenden Verfahren, Werkzeuge und Infrastrukturen bilden wesentliche Kernpunkte des Forschungsdatenmanagements, das heute zu den Anforderungen an das wissenschaftliche Arbeiten zählt. Um die Wissenschaftlerinnen und Wissenschaftler in diesem Prozess zu unterstützen, werden Beratungs- und Schulungsangebote ausgebaut. Ein zentraler Service der Freien Universität Berlin für Forschungsdaten ist das Refubium. FU-Angehörige haben die Möglichkeit auf diesem institutionellen Repository Forschungsdaten zu publizieren.

[Über Forschungsdaten](#)
[Services](#)
[Über das Refubium](#)

AKTUELLES

NFD4Culture-Workshop: Standards für 4Culture-Daten am 03.05.2021
22.04.2021

Train-the-Trainer Workshop: Forschungsdatenmanagement – Grundlagen kennenlernen und weitergeben
21.04.2021

Online-Workshopreihe zu Open Source Tools im Forschungsdatenmanagement
Im Sommersemester 2021 bietet das Team Forschungsdatenmanagement der Universitätsbibliothek der Freien Universität Berlin eine Online-Workshopreihe an.
13.04.2021

VERANSTALTUNGEN

06.05. 09:00 Workshop: Software-Abhängigkeiten richtig beschreiben: Einführung in das Dependency Management für datenbasierte Projekte mit Open Source Tools

18.05. 10:00 Forschungsdatenmanagement - kurz und knapp: Eine Einführung

20.05. 13:00 Online-Sprechstunde Forschungsdatenmanagement

KONTAKT

Universitätsbibliothek
Sibylle Söring, Teamleitung
Esther Asef, Referentin für Forschungsdatenmanagement
Heine-Alexander Fütterer, Referent für Forschungsdatenmanagement
E-Mail: forschungsdaten@fu-berlin.de

Online-Workshopreihe zu Open Source Tools im FDM

- [Online-Workshopreihe](#): Drei Termine im Sommersemester 2021
- Donnerstags von 09:00 – 11:00 Uhr
- Open Source Tools mit Schwerpunkt auf der Programmiersprache Python

#	Datum	Titel
1	06.05.2021	Software-Abhängigkeiten richtig beschreiben: Einführung in das Dependency Management für datenbasierte Projekte mit Open Source Tools
2	10.06.2021	Einführung in die automatisierte Validierung von Forschungsdaten mit Open Source Tools
3	01.07.2021	Einführung in das Project-Templating mit Open Source Tools

Heute: Software-Abhängigkeiten richtig beschreiben

Was heute stattfindet:

0. Willkommen
1. Einleitung
2. Demo und praktische Übungen
3. Q&A
4. Evaluation

Was heute nicht stattfindet:

- Grundlagen der Programmierung in Python (siehe z.B. [Kurse im Weiterbildungszentrum](#))
- Software-Zitation (siehe z.B. [How to cite and describe software](#))
- Best Practices: Software-Engineering
- Best Practices: Packaging

Kurze Vorstellung

WHO How
WHEN ? WHAT
WHERE WHY



Bildquelle: [Pixabay](#)

- Wie heiße ich?
- Aus welchem Fachbereich komme ich?
- Warum bin ich heute hier?
- Welche Vorkenntnisse bringe ich mit?

EINLEITUNG

Motivation und Kontext I: Gute wissenschaftliche Praxis



Leitlinien zur Sicherung
guter wissenschaftlicher Praxis

Kodex

DFG

Leitlinie 7: Phasenübergreifende Qualitätssicherung

„**Die Herkunft von** im Forschungsprozess verwendeten Daten, Organismen, Materialien und **Software wird kenntlich gemacht** und die Nachnutzung belegt; die Originalquellen werden zitiert“

Leitlinie 13: Herstellung von öffentlichem Zugang zu Forschungsergebnissen

„Selbst programmierte Software wird unter Angabe des Quellcodes öffentlich zugänglich gemacht. **Eigene und fremde Vorarbeiten weisen Wissenschaftlerinnen und Wissenschaftler vollständig und korrekt nach.**“

Deutsche Forschungsgemeinschaft. 2019. „Leitlinien zur Sicherung guter wissenschaftlicher Praxis (Kodex)“. <https://zenodo.org/record/3923602>.

Siehe auch: <https://wissenschaftliche-integritaet.de/?s=software>

Motivation und Kontext II: Good Practice

Make dependencies and requirements explicit (2g). This is usually done on a per-project rather than per-program basis, i.e., by adding a file called something like requirements.txt to the root directory of the project or by adding a "Getting Started" section to the README file.

Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal TK (2017) Good enough practices in scientific computing. PLoS Comput Biol 13(6): e1005510. <https://doi.org/10.1371/journal.pcbi.1005510>

Motivation und Kontext III: Simple Rule

Rule 3: Archive the Exact Versions of All External Programs Used

In order to exactly reproduce a given result, it may be necessary to use programs in the exact versions used originally. Also, as both input and output formats may change between versions, a newer version of a program may not even run without modifying its inputs. Even having noted which version was used of a given program, it is not always trivial to get hold of a program in anything but the current version. **Archiving the exact versions of programs actually used may thus save a lot of hassle at later stages.** In some cases, all that is needed is to store a single executable or source code file. In other cases, a given program may again have specific requirements to other installed programs/packages, or dependencies to specific operating system components. To ensure future availability, the only viable solution may then be to store a full virtual machine image of the operating system and program. As a minimum, you should note the exact names and versions of the main programs you use.

Sandve GK, Nekrutenko A, Taylor J, Hovig E (2013) Ten Simple Rules for Reproducible Computational Research. PLoS Comput Biol 9(10): e1003285. <https://doi.org/10.1371/journal.pcbi.1003285>

Was sind Software Dependencies?

- In der Regel schreiben Sie nicht die gesamte Software, die in Ihrem Projekt eingesetzt wird
 - Und sie müssen es auch nicht!
- Sie verwenden bestehende Pakete, Bibliotheken und Frameworks und importieren Funktionalität daraus
- Immer wenn Sie in einem Python-Modul eine Zeile wie **import pandas as pd** oder **import numpy as np** verwenden, die sich auf Software bezieht, die sich nicht geschrieben haben
 - handelt es um sich eine Software Dependency



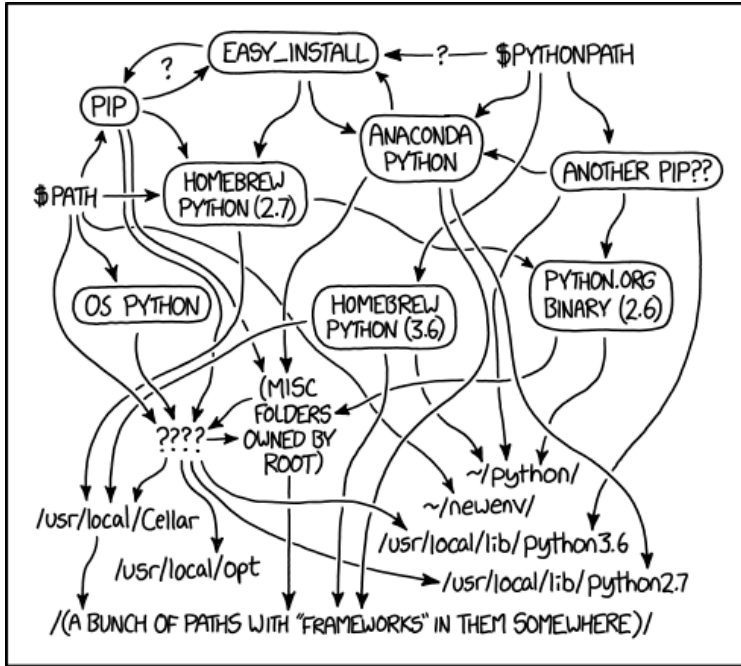
Bildquelle: Wikimedia Commons

Was ist Dependency Management?

- Die explizite Nennung von exakten Versionen verwendeter Software Dependencies
- Bestenfalls orientiert an Best-Practices (nicht schriftlich im Anhang eines Papers)
- Ermöglicht einfache Installation der Software Dependencies durch Tooling:
 - package.json ([npm](#) für Node.js/JavaScript)
 - [pom.xml ([Maven](#) für Java)]
 - [\[renv](#) für R]
 - [\[Dr. Watson](#) für Julia]

Siehe auch: <https://coderefinery.github.io/reproducible-research/03-dependencies/>

Fokus-Thema: Dependency Managment Tools für Python



MY PYTHON ENVIRONMENT HAS BECOME SO DEGRADED
THAT MY LAPTOP HAS BEEN DECLARED A SUPERFUND SITE.

Bildquelle: [xkcd](#)

Teil des Workshops

- Pip (requirements.txt)
- Virtualenv
- Pipenv (Pipfile, Pipfile.lock)
- Poetry (pyproject.toml, poetry.lock)

Nicht behandelt

- [Conda]
- [pip-tools]
- [Pyenv]
- [Docker]

ÜBUNGEN

Referenzen und Literaturhinweise

- Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal TK (2017) Good enough practices in scientific computing. PLoS Comput Biol 13(6): e1005510. <https://doi.org/10.1371/journal.pcbi.1005510>
- Sandve GK, Nekrutenko A, Taylor J, Hovig E (2013) Ten Simple Rules for Reproducible Computational Research. PLoS Comput Biol 9(10): e1003285. <https://doi.org/10.1371/journal.pcbi.1003285>
- Wilson, Damien Irving, Kate Hertweck, Luke Johnston, Joel Ostblom, Charlotte Wickham, and Greg. *Research Software Engineering with Python*. *merely-useful.tech*, <https://merely-useful.tech/py-rse/>
- Community, The Turing Way, u. a. *The Turing Way: A Handbook for Reproducible Data Science*. v0.0.4, Zenodo, 2019. [DOI.org](https://doi.org/10.5281/ZENODO.3233986) (Datacite), doi:10.5281/ZENODO.3233986.
- Software und Data Carpentry-Kurse
 - <https://software-carpentry.org/lessons/>
 - <https://datacarpentry.org/lessons/>

Vielen Dank für Ihre Teilnahme

Heinz-Alexander Fütterer

Referent für Forschungsdatenmanagement

Kontakt

E-Mail: forschungsdaten@fu-berlin.de

Web: <https://www.fu-berlin.de/forschungsdatenmanagement>

Bildquellen sind auf jeder Folie angegeben, Piktogramme von Microsoft Powerpoint ohne Copyright or Lizenz-Angabe:
<https://support.microsoft.com/de-de/office/einfügen-von-piktogrammen-in-microsoft-office-e2459f17-3996-4795-996e-b9a13486fa79>