Automating Dataverse with pyDataverse

Migrations and Testing

Stefan Kasberger

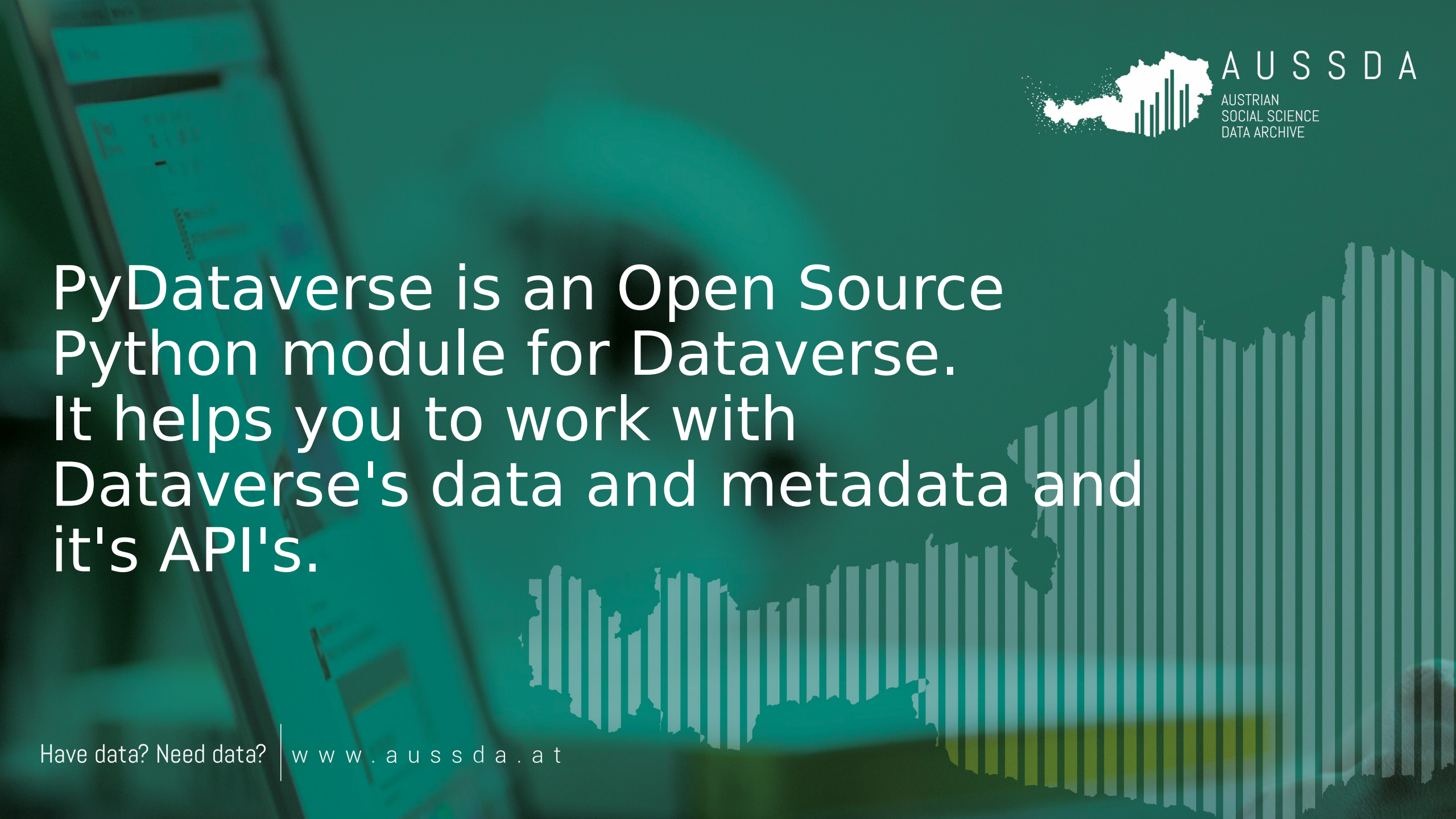
AUSSDA - The Austrian Social Science Data Archive

Materials

GitHub: AUSSDA/dataverse2021_automation-with-pydataverse

bit.ly/3pMnjPv

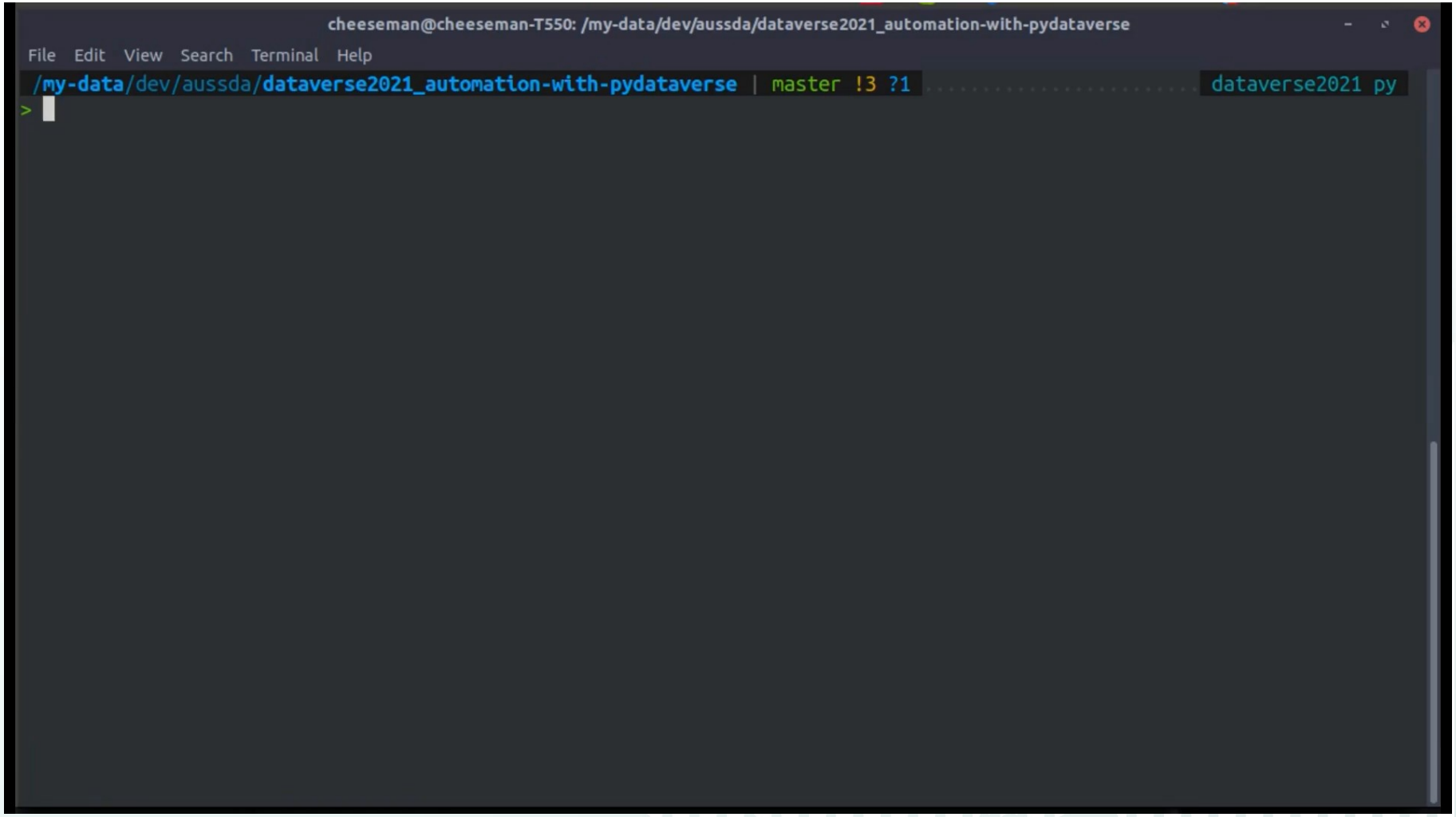

AUSSDA
AUSTRIAN
SOCIAL SCIENCE
DATA ARCHIVE
pyDataverse
Have data? Need data?
www.aussda.at


AUSSDA
AUSTRIAN
SOCIAL SCIENCE
DATA ARCHIVE
PyDataverse is an Open Source Python module for Dataverse. It helps you to work with Dataverse's data and metadata and it's API's.
Have data? Need data?
www.aussda.at

pyDataverse

- Goal: Migrations, Automation (Testing), Microservices, Data Science
- Python + Pip
- Latest: v0.3.1
- Open Source: MIT
- Tested and Docs
- GitHub: gdcc/pydataverse
- Funded: AUSSDA + SSHOC

Models + API


cheeseman@cheeseman-T550: /my-data/dev/aussda/dataverse2021_automation-with-pydataverse
File Edit View Search Terminal Help
/my-data/dev/aussda/dataverse2021_automation-with-pydataverse | master !3 ?1
dataverse2021 py
>

Jupyter Notebook

[1]:
```
import io
import pandas as pd
from pyDataverse.api import DataAccessApi
from pyDataverse.api import NativeApi
```

[2]:
```
# Establish API connection and retrieve content of the dataset
native_api = NativeApi("https://data.aussda.at")
resp = native_api.get_dataset("doi:10.11587/P5YJ0O")
# Get all datafiles related information
datafiles = resp.json()["data"]["latestVersion"]["files"]
# Let's dig further and display the available files
for df in datafiles:
    filename = df["dataFile"]["filename"]
    datafile_id = df["dataFile"]["id"]
    print(f'Filename is "{filename}", datafile ID is "{datafile_id}"')
```

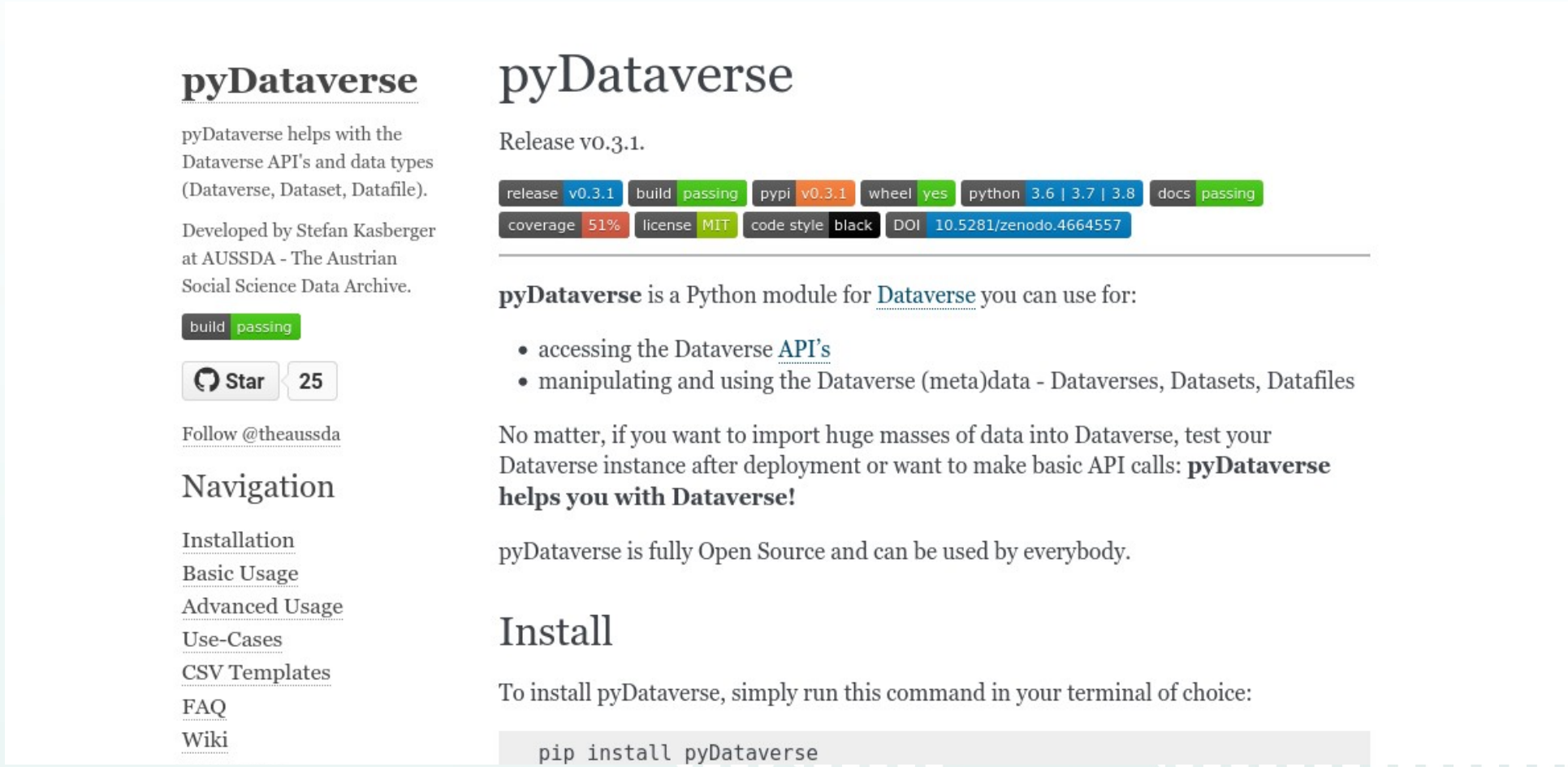
```
Filename is "10095_da_de_v2_0.tab", datafile ID is "4025"
Filename is "10095_da_de_v2_0.zip", datafile ID is "4024"
Filename is "10095_da_de_v2_0.zsav", datafile ID is "4026"
Filename is "10095_mr_en_v2_0.pdf", datafile ID is "4031"
Filename is "10095_om_de_v2_0.zip", datafile ID is "4028"
Filename is "10095_qu_de_v2_0.pdf", datafile ID is "4029"
Filename is "10095_vi_de_v2_0.tab", datafile ID is "4032"
```

Jupyter Notebook

[3]:
```
# To download data, we need the DataAccess API
da_api = DataAccessApi(base_url)
# Let's select the first datafile
datafile_id = "4025"
# Download the datafile
resp = da_api.get_datafile(datafile_id)
# Turn the content into a Pandas DataFrame
data = io.StringIO(str(resp.content,'utf-8'))
data = pd.read_csv(data, sep="\t")
print(data.head(1))
```

```
             version                   doi  \
0  2.0 (2021-03-31)  doi:10.11587/P5YJ0O

                                        CITATION  \
0  Kittel, Bernhard; Kritzinger, Sylvia; Boomgaar...

                            FUNDING_ACKNOWLEDGEMENT  RESPID  ENTRY_WAVE  \
0  Data collection has been made possible by COVI...  990001          14

   W1_PANELIST  W2_PANELIST  W3_PANELIST  W4_PANELIST  ...  W15_Q82A6  \
0          NaN          NaN          NaN          NaN  ...        NaN

   W15_Q82A7  W15_Q82A8  W15_Q82A9  W15_Q83  W15_Q84A1  W15_Q84A2  W15_Q84A3  \
0        NaN        NaN        NaN      NaN        NaN        NaN        NaN
```

Docs



pyDataverse

pyDataverse helps with the Dataverse API's and data types (Dataverse, Dataset, Datafile).

Developed by Stefan Kasberger at AUSSDA - The Austrian Social Science Data Archive.

build passing

Star 25

Follow @theaussda

Navigation

Installation
Basic Usage
Advanced Usage
Use-Cases
CSV Templates
FAQ
Wiki

pyDataverse

Release v0.3.1.

release v0.3.1 | build passing | pypi v0.3.1 | wheel yes | python 3.6 | 3.7 | 3.8 | docs passing
coverage 51% | license MIT | code style black | DOI 10.5281/zenodo.4664557

pyDataverse is a Python module for Dataverse you can use for:

- accessing the Dataverse API's
- manipulating and using the Dataverse (meta)data - Dataverses, Datasets, Datafiles

No matter, if you want to import huge masses of data into Dataverse, test your Dataverse instance after deployment or want to make basic API calls: **pyDataverse helps you with Dataverse!**

pyDataverse is fully Open Source and can be used by everybody.

Install

To install pyDataverse, simply run this command in your terminal of choice:

```
pip install pyDataverse
```

AUSSDA
AUSTRIAN
SOCIAL SCIENCE
DATA ARCHIVE

Migrations

Have data? Need data? | www.aussda.at

Migrations

- Dataverse 2 Dataverse
- NESSTAR 2 Dataverse
- GESIS Dspace 2 Dataverse
- Paper 2 Dataverse

} CSV templates 2 Dataverse

More than 1.100 datasets and 1.900 datafiles imported at AUSSDA!

=

Saved more than 600 hours of boring work!

AUSSDA
AUSTRIAN
SOCIAL SCIENCE
DATA ARCHIVE

Dataverse Tests

Have data? Need data? | www.aussda.at

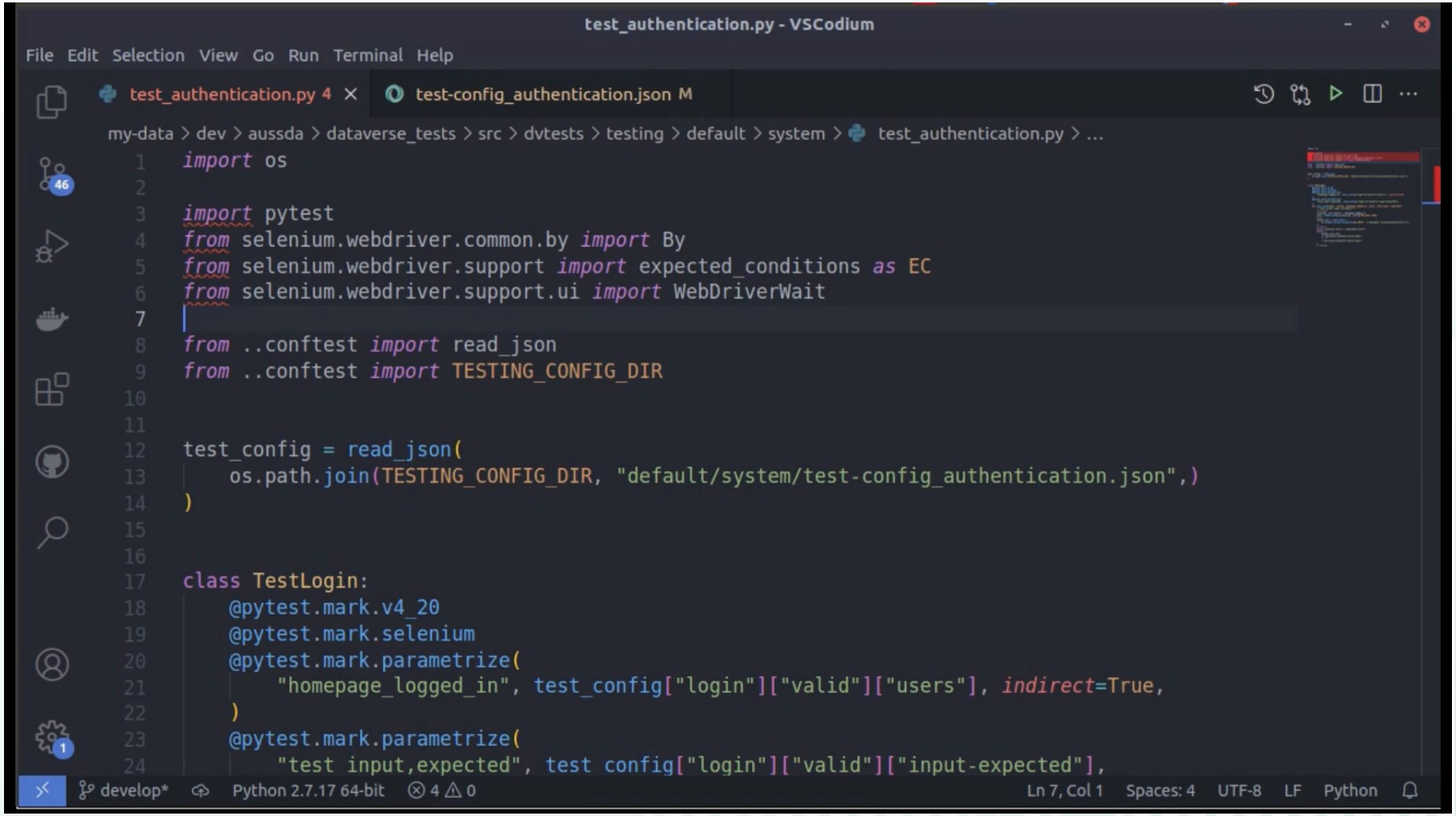
Dataverse Tests

- Goal: Dataverse Tests for DevOps - test your newly setup or updated or running Dataverse installation → optimized for Jenkins.
- PyDataverse, Pytest + Selenium
- Open Source: MIT
- GitHub: gdcc/dataverse_tests
- new → must be used to become more stable
- A few flexible and modular tests, easy to extend
- Settings Management
- Configs: tests, Dataverse installations, users
- Funded: AUSSDA + SSHOC

Tests

- Basic: API, OAI-PMH, Sitemap, robots.txt, DOI URL, ToU
- Frontend: Create Dataverse, Search
- Authentification: normal, Shibboleth
- Data completeness

Test: Shibboleth Login



Utils

- Helper functions
- CLI
- Functions:
  - Create testdata
  - Remove testdata
  - Collect data
  - Create users

Utils: Create testdata

- :root
  - dataverse-1
    - dataset-1
      - file: essay.txt
      - file: doc.pdf
    - dataset-2
      - file: data.csv
  - dataverse-2
    - dataverse-3
    - dataset-3
      - file: data.tab
    - dataset-1

Dataverse Testdata

- Goal: Collection of quality metadata
- GitHub: aussda/dataverse_testdata
- Funding: AUSSDA
- Other repo:
  - GitHub: IQSS/dataverse-sample-data


AUSSDA
AUSTRIAN
SOCIAL SCIENCE
DATA ARCHIVE
Get Involved!
Have data? Need data?
www.aussda.at

Community wins!

- An Open Source project is only as good, as it's community!
- The projects are not mine. They are community projects, and can only sustain if others join in.
- How to contribute
  - Use it
  - Create issues
  - Contribute

AUSSDA
AUSTRIAN SOCIAL SCIENCE DATA ARCHIVE

Thanks!

For questions please contact:

stefan.kasberger@univie.ac.at

@stefankasberger

www.aussda.at