Data visualization and processing with Jupyter Notebook in SeaDataCloud Virtual Research Environment

Jani Ruohola (jani.ruohola@ymparisto.fi)[1], Sri Harsha Vathsavayi[2], Seppo Kaitala[1], Christopher Ariyo[2]

[1]Finnish Environment Institute SYKE, Finland

[2]CSC IT Center for Science, Finland

Introduction

As reproducibility is an essential part of scientific research, the failure to reproduce experiments can lead to unreliable results and false scientific findings. Currently, irreproducible research is a problem across all domains of science[1]. We are facing a similar replication crisis in marine research as well. The complex and rapidly changing nature of computer environments and software libraries can create obstacles for the computational reproducibility of research (Fig. 1). In order to reproduce experiments, we need to capture the data, software environment and the whole workflow. This study demonstrates the use of notebooks and virtual research environments (VREs) to create reproducible experiments.

```
/usr/local/lib/python3.5/dist-packages/numpy/ma/core.py:6447: MaskedArrayFutureW
arning: In the future the default for ma.minimum.reduce will be axis=0, not the
current None, to match np.minimum.reduce. Explicitly pass 0 or None to silence t
his warning.
```

Fig 1. Various warnings and errors are a common sight when analyzing data. As the tools used for data analysis become more diverse, it can be a tedious task to try to reproduce other scientists' workflows. Even reusing one's own code may be difficult, since different versions of packages and modules may cause unexpected errors.

Methods

SeaDataCloud Virtual Research Environment (SDC-VRE) is a work-in-progress service that aims to facilitate collaboration and aid in research by integrating data from different sources with tools such as Jupyter Notebooks, EUDAT B2DROP[2] and DIVA[3]. Data exists in the user's B2DROP account, which is integrated with the VRE (Fig. 2). Data processing, analysis and visualization is conducted within Jupyter Notebook and the results can be distributed for example as netCDF files. The computational environment exists within the VRE in a separate Docker container, which offers OS-level virtualization and user-space isolation. Moreover, the data and notebooks can be securely shared with the research groups via B2DROP for collaboration and for reproducing the workflow.

An example workflow for reproducible research in SDC-VRE was tested within a virtual machine (VM) provided by CSC, which is leading the technical work of data infrastructure in the SeaDataCloud project. A Docker container running Jupyter Notebook was created to the VM and the B2DROP account containing the data was mounted to it by using the WebDAV API of B2DROP. The Jupyter server was then accessed with a web browser.

The test dataset used in this study was FerryBox measurements from M/S Silja Serenade, cruising between Helsinki and Stockholm on the Baltic Sea. A Python class for processing ODV text data format was written, with methods to plot the data in scatter or contour plots. Interpolation was done by using SciPy's interpolation sub-package.

Results

An introductory data visualization notebook was successfully created by connecting data from a personal B2DROP account to the VRE (Fig. 3). The notebook server could then easily be accessed with the IP address and used as a demonstration tool. Creating new Docker containers in the VM via SSH was working and although useful for testing purposes, it is not needed by end-user in the final VRE. The created notebook demonstrated the use of custom data classes in Python and was considered useful for domain scientists with little programming experience, while simultaneously acting as a showcase for promoting the use of open-source tools.

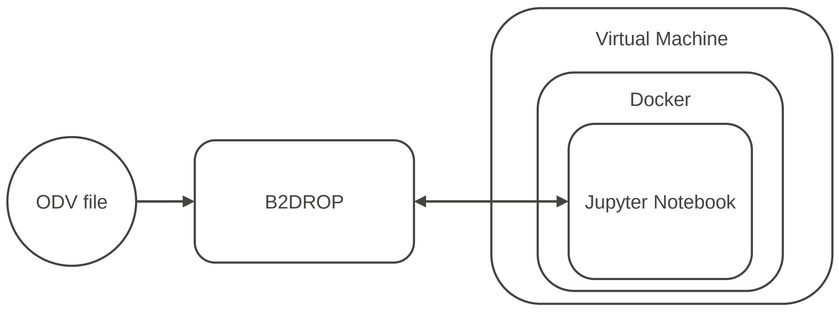

Fig 2. Schematic workflow for the SDC-VRE. ODV file exists in the user's B2DROP account, which is connected to the VM hosting the Docker container (Figure adopted from *Specification of visualization services and development plan*, SDC deliverable D10.13).

Conclusions

The purpose of this study was to test the feasibility of setting up and using a Dockerized computational environment within the work-in-progress SDC-VRE. Jupyter Notebook proved to be a useful tool for documenting and distributing the data processing and visualization workflow. By including rich-text annotations and figures to the notebook, a tutorial-like document was created, considered beneficial especially to the non-programmers. In the era of increasing importance of data-driven research, these tools are invaluable. This approach also seems to offer benefits for more traditional hypothesis-driven experimental research, since the documentation of the computational part is as important as the documentation of the experimental methods. Thus, SDC-VRE was considered to be a significant improvement in the marine scientist's workbench.

Acknowledgements

EUDAT, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065

SeaDataCloud project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 730960.

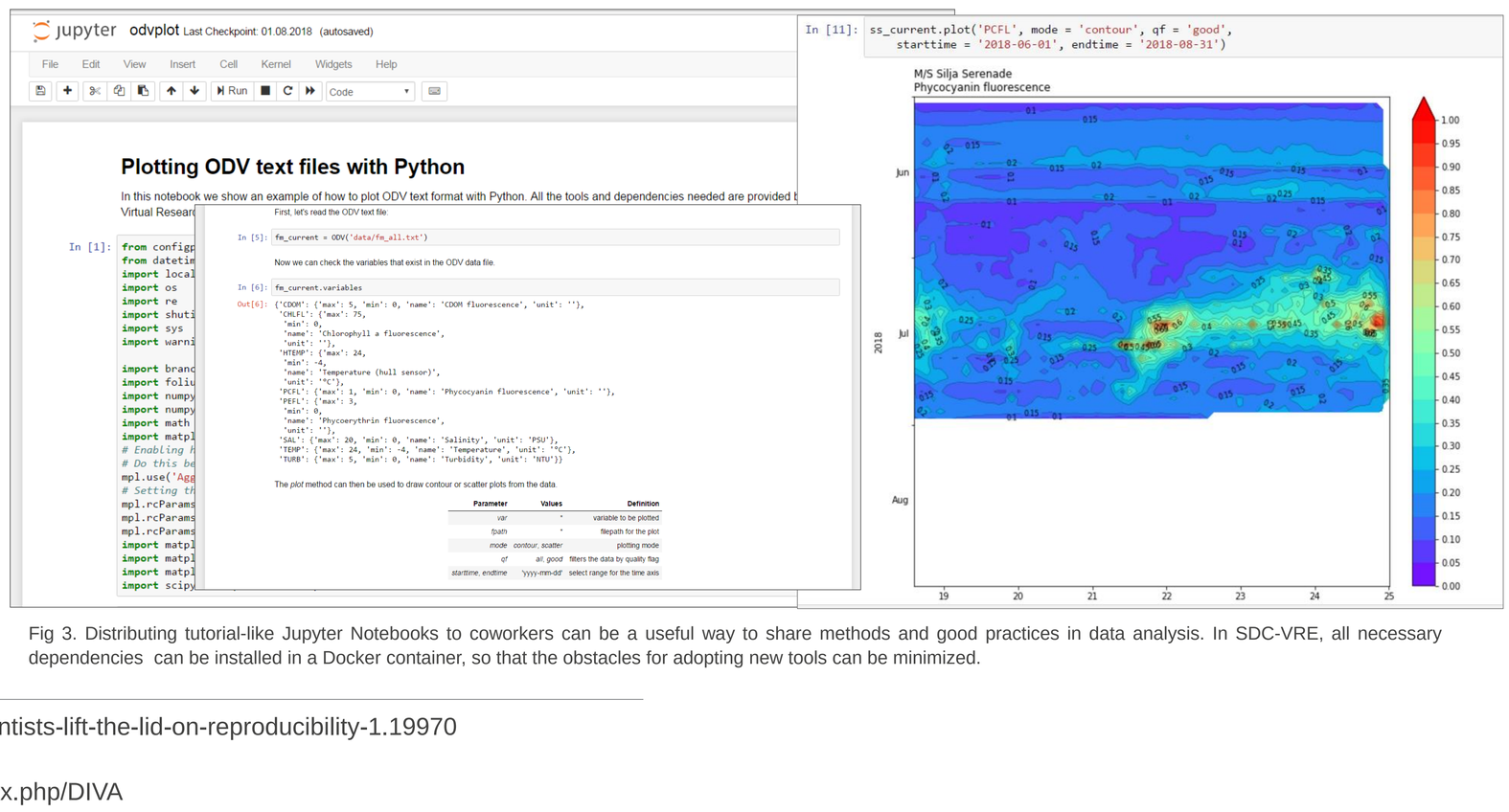

Fig 3. Distributing tutorial-like Jupyter Notebooks to coworkers can be a useful way to share methods and good practices in data analysis. In SDC-VRE, all necessary dependencies can be installed in a Docker container, so that the obstacles for adopting new tools can be minimized.

[1]https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

[2]https://eudat.eu/services/b2drop

[3]http://modb.oce.ulg.ac.be/mediawiki/index.php/DIVA