OPEN SCIENCE, KNOWLEDGE SHARING AND REPRODUCIBILITY AS DRIVERS FOR THE ADOPTION OF FOSS4G IN ENVIRONMENTAL RESEARCH

Ionuț Iosifescu Enescu 1, Gian-Kasper Plattner 1, Leo Bont, Marielle Fraefel 1, Rolf Meile 1, Thomas Kramer 1, Lucia Espona-Pernas 1, Dominik Haas-Artho 1, Martin Hägeli 1, Konrad Steffen 1

1 Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Birmensdorf, Switzerland

(ionut.iosifescu, gian-kasper.plattner, leo.bont, marielle.fraefel, rolf.meile, thomas.kramer, lucia.espona, dominik.haas, martin.haegeli, konrad.steffen)@wsl.ch

Commission IV, WG IV/4

KEY WORDS: Open Science, Knowledge Sharing, Reproducibility, FOSS4G, EnviDat, Environmental Data Repository, Environmental Research, Research Data Publication, Research Data Management

ABSTRACT:

Support for open science is a highly relevant user requirement for the environmental data portal EnviDat. EnviDat, the institutional data portal and publication data repository of the Swiss Federal Research Institute WSL, actively implements the FAIR (Findability, Accessibility, Interoperability and Reusability) principles and provides a range of services in the area of research data management. Open science, with its requirements for improved knowledge sharing and reproducibility, is driving the adoption of free and open source software for geospatial (FOSS4G) in academic research. Open source software can play a key role in the proper documentation of data sets, processes and methodologies, because it supports the transparency of methods and the precise documentation of all steps needed to achieve the published results. EnviDat actively supports these activities to enhance its support for open science. With EnviDat, WSL contributes to the ongoing cultural evolution in research towards open science and opportunities for distant collaboration.

1. INTRODUCTION

EnviDat – www.envidat.ch – is the environmental data portal and repository of the Swiss Federal Institute for Forest, Snow and Landscape Research WSL. EnviDat actively implements the FAIR (Findability, Accessibility, Interoperability and Reusability) principles (Wilkinson et al., 2016) and provides a range of services in the area of research data management that were extensively described in Iosifescu et al. (2018a). With its capability to host and publish research data sets, EnviDat currently provides unified and managed access to more than 160 environmental data sets collected by WSL researchers and their research partners (Figure 1).

The existing datasets, software and additional research information, currently mainly originating from WSL, are accompanied by their documentation with appropriate metadata, including their geospatial footprint. Figure 2 shows an example of a published data set in EnviDat, where the research data is accompanied by proper citation information and Digital Object Identifiers (DOIs).

In accordance with the institutional data policy, WSL makes curated, quality-controlled, reusable and publication-ready research data accessible through EnviDat, therefore supporting worldwide data users with an efficient access to various environmental research data sets that were collected by WSL.
2. OPEN SCIENCE SUPPORT IN ENVIDAT

Open science has many aspects and it would be beyond the capabilities of EnviDat to address all aspects comprehensively. We thus focus on one core topic, i.e., knowledge sharing and reproducibility of published research. According to Baker (2016), we have an ongoing “reproducibility crisis”, including in the field of Earth and Environment sciences. Unavailable methods or code and missing raw data from original lab are mentioned as major factors for irreproducible results in this study. Publishing data with formal metadata is often not sufficient to effectively foster open science. For the goal of improving knowledge sharing and reproducibility of published research, the research methods, i.e. computations, need to be made transparent, too. Consequently, the sharing of software is a vital element for the proper understanding of the methodologies presented in research papers.

For this reason, EnviDat encourages WSL scientists to complement data publication with (i) description of research methods and (ii) open source software. An example is provided by Bont et al. (2018) with the opening of a methodology for optimizing the layout of cable roads needed for wood harvesting on steep slopes. It achieves realistic solutions with lower installation costs for cable roads layouts in Central European conditions. The publishing of this methodology is complemented with its implementation in the QGIS plugin “Seilaplan”, for everyone to use (Figure 3).

3. NEW DRIVERS FOR THE ADOPTION OF FOSS4G

Iosifescu et al. (2015) presented several reasons for the increasing adoption of free and open source software for geospatial (FOSS4G) in academic research, such as: the growing stability and maturity of the recent open source software packages, the faster bug-fixing turnover, the increasing availability of professional support, and the flexibility to change and repurpose the open software to tackle new research challenges. These reasons are still valid today, as corroborated by the fact that the EnviDat portal integrates several open software packages such as PostgreSQL/PostGIS, Apache Solr, CKAN and Vue.js in its software architecture.

In this contribution, we discuss two novel drivers for the adoption of FOSS4G in environmental research that arose in connection with open science, namely knowledge sharing and reproducibility. We illustrate the larger issue of supporting open science in an environmental data portal such as EnviDat, with the example of Fraefel (2018). The Jupyter notebook for calculating road densities in the neighborhood of sample plot locations was based on the ArcPy site package, allowing the use of ESRI ArcGIS geoprocessing tools in Python. However, although the Python code (through the Jupyter notebook) as well as the corresponding datasets (publicly available in EnviDat) are open, a user would need a proprietary installation with a valid license in order to run it.

We argue that independent research replication at peer-review and after publication is facilitated by (i) the immediate availability of FOSS4G, (ii) the absence of software licensing issues and (iii) the openness of the code even for older versions of a software. Moreover, researchers producing their own code can expect a wider distribution of the produced software, and though it, their knowledge and expertise. Finally, we would like to note, that these new drivers, at the moment, can be observed in academia. These have not been proven valid for industry, administration or any other domain that is not pursuing academic open science.

4. END TO END REPRODUCIBILITY EXAMPLE FOR ACADEMIC RESEARCH WITH FOSS4G

In EnviDat, open science is supported by the combined publication of bundles of datasets and software as detailed in the next demonstrative use case. The EnviDat record from Bont et al. (2019) demonstrates a fully reproducible preprocessing workflow that generates several output rasters. The notebook presented in Figure 4, extracts harmonized output rasters with the same extent. The extent is given by a polygon vector dataset (Perimeter). These output rasters, such as obstacles, aspect, slope, forest cover, can serve as input data for later computations related to forest accessibility and wood harvesting questions.

The obstacles output is obtained by transforming line vector datasets (railway lines, high voltage power lines) to raster. The forest dataset is derived from a vector dataset. The aspect and slope are both derived from a digital elevation model (DEM). The Python code defines several functions for preprocessing, for example in order to convert raster datasets into ASCII files, to turn an array into a raster or to convert vector data into raster data. The input data is automatically downloaded and extracted from the corresponding EnviDat online public resources. Then, the code will use the GDAL functionality in order to perform the needed geoprocessing, such as resampling of the DEM, computing the slope and aspect from the resampled DEM or generating the obstacle raster. Finally, the computed rasters are written to disk.
The DEM sample for the selected forest range is based on the original DEM source reprojected, clipped, and resampled to 25 meters using QGIS. More information about this example (including data and notebook) is available at EnviDat at doi: 10.16904/envidat.75.

For demonstration purposes, the notebook uses fully open vector data for railways, forests and power lines, as well as an open DEM for a small area around a sample forest range in Europe (Germany, Upper Bavaria, Kochel Forest Range, some 70 km south of München, at the edge of the Bavarian Alps). The data of this example can also be used with additional open methodologies and software such as the QGIS plugin Seilaplan (Bont et al. 2018) for optimizing the geometric layout of cable roads or with additional open software for computing the forest accessibility for wood harvesting.

For demonstration purposes, the notebook uses fully open vector data for railways, forests and power lines, as well as an open DEM for a small area around a sample forest range in Europe (Germany, Upper Bavaria, Kochel Forest Range, some 70 km south of München, at the edge of the Bavarian Alps). The data of this example can also be used with additional open methodologies and software such as the QGIS plugin Seilaplan (Bont et al. 2018) for optimizing the geometric layout of cable roads or with additional open software for computing the forest accessibility for wood harvesting.

For demonstration purposes, the notebook uses fully open vector data for railways, forests and power lines, as well as an open DEM for a small area around a sample forest range in Europe (Germany, Upper Bavaria, Kochel Forest Range, some 70 km south of München, at the edge of the Bavarian Alps). The data of this example can also be used with additional open methodologies and software such as the QGIS plugin Seilaplan (Bont et al. 2018) for optimizing the geometric layout of cable roads or with additional open software for computing the forest accessibility for wood harvesting.

For demonstration purposes, the notebook uses fully open vector data for railways, forests and power lines, as well as an open DEM for a small area around a sample forest range in Europe (Germany, Upper Bavaria, Kochel Forest Range, some 70 km south of München, at the edge of the Bavarian Alps). The data of this example can also be used with additional open methodologies and software such as the QGIS plugin Seilaplan (Bont et al. 2018) for optimizing the geometric layout of cable roads or with additional open software for computing the forest accessibility for wood harvesting.

For demonstration purposes, the notebook uses fully open vector data for railways, forests and power lines, as well as an open DEM for a small area around a sample forest range in Europe (Germany, Upper Bavaria, Kochel Forest Range, some 70 km south of München, at the edge of the Bavarian Alps). The data of this example can also be used with additional open methodologies and software such as the QGIS plugin Seilaplan (Bont et al. 2018) for optimizing the geometric layout of cable roads or with additional open software for computing the forest accessibility for wood harvesting.

For demonstration purposes, the notebook uses fully open vector data for railways, forests and power lines, as well as an open DEM for a small area around a sample forest range in Europe (Germany, Upper Bavaria, Kochel Forest Range, some 70 km south of München, at the edge of the Bavarian Alps). The data of this example can also be used with additional open methodologies and software such as the QGIS plugin Seilaplan (Bont et al. 2018) for optimizing the geometric layout of cable roads or with additional open software for computing the forest accessibility for wood harvesting.

For demonstration purposes, the notebook uses fully open vector data for railways, forests and power lines, as well as an open DEM for a small area around a sample forest range in Europe (Germany, Upper Bavaria, Kochel Forest Range, some 70 km south of München, at the edge of the Bavarian Alps). The data of this example can also be used with additional open methodologies and software such as the QGIS plugin Seilaplan (Bont et al. 2018) for optimizing the geometric layout of cable roads or with additional open software for computing the forest accessibility for wood harvesting.

The data of this example (including data and notebook) is available on the Austrian Open Data Portal at http://data.opendataportal.at/dataset/dtm60_germany. This original DEM source was reprojected, clipped, and resampled to 25 meters using QGIS. More information about this example (including data and notebook) is available at EnviDat at doi: 10.16904/envidat.75.

5. CONCLUSIONS AND OUTLOOK

Supporting reproducibility of research is a complex issue that can be facilitated by the adoption of FOSS4G. We would like to stress that reproducibility in science also consists of transparency of methods and accurate documentation of all steps necessary to arrive at the published results. Here, open source software can play a key role.

With EnviDat, WSL fosters international research cooperation in the field of environmental science and contributes to the ongoing cultural evolution in research towards openness, shared data and opportunities for distant collaboration. By openly publishing open software (e.g. as Jupyter notebooks) alongside research data sets, researchers can contribute to mitigate the issues generally described as “reproducibility crisis”. As researchers can combine open source code, detailed text-based descriptions and rich research output in Jupyter notebooks, they will benefit from improved means for opening and documenting their research methods.

In the coming years, EnviDat aims for a deeper integration of user-uploaded Jupyter Notebooks. The development of a generic solution that would offer WSL researchers the opportunity to access their code from anywhere on the existing EnviDat infrastructure is underway. After small-scale proof-of-concept work has been performed that allowed for execution of Jupyter notebooks, we now work on getting such EnviDat-hosted notebooks compatible with the WSL High-Performance Computing (HPC) Linux Cluster. The infrastructure consists of several computing nodes which are connected through a low-latency and high-bandwidth network. Future work will address the issue of providing external access to the current JupyterHub/JupyterLab beta installation on the HPC cluster. Notebooks hosted in EnviDat could then be executed from anywhere, allowing reviewers of WSL research to request access and interactively execute the published code during time windows of low cluster utilization.

These developments are expected to strengthen WSL’s commitment to accessible research data in order to advance science and to further evolve EnviDat towards a next-generation environmental data repository. Combined with good and well-established data management practices, these will help provide a better overview over the individual steps of a research process, thus further enhancing the reusability of research results, know-how and expertise.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the support of the WSL IT department, in particular Markus Reinhardt and Stephan Röthlisberger. Furthermore, we would like to thank Matthias Häni and other members of the EnviDat User Group for all their inputs, feedback and support given to the EnviDat program.
REFERENCES

Baker, M., 2016: 1,500 scientists lift the lid on reproducibility. *Nature*, 533, 452–454 (26 May 2016). doi.org/10.1038/533452a


