

Open Data – Open Software: Implementing Geospatial Requirements in EnviDat with an Open-Source Stack

Ionuț Iosifescu Enescu^{a, *}, David Hanimann^a, Dominik Haas-Artho^a, Marius Rüetschi^a, Dirk Nikolaus Karger^a, Gian-Kasper Plattner^a, Martin Hägeli^a, Rebecca Kurup Buchholz^a, Lucia de Espona^a, Niklaus E. Zimmermann^a, Loïc Pellissier^{a, b}

^a Swiss Federal Research Institute WSL, ionut.iosifescu@wsl.ch, david.hanimann@wsl.ch, dominik.haas@wsl.ch, marius.rueetschi@wsl.ch, dirk.karger@wsl.ch, gian-kasper.plattner@wsl.ch, martin.haegeli@wsl.ch, rebecca.kurup@wsl.ch, lucia.espona@wsl.ch, niklaus.zimmermann@wsl.ch

^b ETH Zurich, loic.pellissier@usys.ethz.ch

* Corresponding author

Keywords: open data, open software, environmental data, very large geodata, EnviDat

Abstract:

EnviDat is the environmental data portal of the Swiss Federal Research Institute WSL providing unified and managed access to environmental monitoring and research data (accessible at www.envidat.ch). EnviDat actively promotes good practices for Research Data Management (RDM) and Open Science at WSL and supports scientists with the formal publishing of environmental datasets from forest, landscape, biodiversity, natural hazards and snow and ice research. With EnviDat, WSL actively strives to increase the proportion of FAIR (Findable, Accessible, Interoperable, Reusable) environmental data that is easily available for reuse by researchers and the public. EnviDat provides Document Object Identifiers (DOIs), appropriate citation information, and formal publication for registered research data. Currently EnviDat hosts and disseminates around 250 DOI-ed environmental data sets collected by WSL researchers and their research partners.

In addition to data publication and dissemination, EnviDat is moving towards integrating more comprehensive visualization functionalities, including semi-automated (meta)data visualization on interactive Web maps. This is driven bottom-up by the diverse set of WSL researchers' requirements for a more specialized environmental data portal. Publishing and visualizing geodata is, however, not without challenge. Among other aspects, EnviDat has to optimally manage and publish large environmental geodata sets, such as remote sensing data or climate model outputs. For example, our researchers wish to provide unified access, through EnviDat, to modified Copernicus data for Switzerland (www.copernicus.eu). Furthermore, newer versions of the "climatologies at high resolution for the Earth's land surface areas" (CHELSA; Karger et al., 2017), are also made available through EnviDat.

In order to create the capability for efficiently managing and organizing large datasets such as (i) processed Sentinel-2 (S2) imagery for Switzerland or (ii) CHELSA, we have been prototyping a generic, hypercube-based data organization and visualization (Iosifescu et al., 2015). It is specifically adapted and simplified for the management of gridded time-series of hundreds of multi-band S2 images and thousands of climate model output layers for a single thematic variable. Since we could not identify any out-of-the-box software that can efficiently implement the requirements for the management and visualization of such large time-series, we have turned to free and open-source software (FOSS) because of its maturity, flexibility, deployment speed, and cost-effectiveness. Open-source software is being increasingly adopted in academia and research because it can play a key role in the proper documentation of data sets, processes and methodologies. Furthermore, open-source software allows better reproducibility and transparency of methods when following best practices in Open Science (Iosifescu et al., 2019). Likewise, EnviDat serves as an example for WSL researchers by using and implementing as much open software as possible, with the EnviDat source code openly available on GitHub (<https://github.com/EnviDat/>).

A limited proof-of-concept for map-based visualization of climate data is being implemented based on the new CORE (Cloud Optimized Raster Encoding) format for gridded data and the following open-source software stack: PostgreSQL (www.postgresql.org), PostGIS (<https://postgis.net>), PostGREST (<https://postgrest.org>), GDAL (<https://gdal.org>), QGIS Server (<https://www.qgis.org>; https://docs.qgis.org/3.18/en/docs/server_manual/), Solr (<https://solr.apache.org>), rclone (<https://rclone.org>), FFmpeg (<https://www.ffmpeg.org>), Python3 (www.python.org), Leaflet (<https://leafletjs.com>) and Vue.js (<https://vuejs.org>). Each of these software components has a role in building an institutional Spatial Data Infrastructure (SDI) targeted at making large and open environmental research data sets easily available for reuse. With the gradual development of such an SDI at WSL, EnviDat will further improve its capabilities for presentation, documentation and exchange of scientific information. Consequently, we make a contribution to the ongoing evolution in research towards an Open Research Data and Open Science culture, enabling enhanced opportunities for national and international scientific collaboration and exchange. We hope EnviDat can serve as a model for other FAIR platforms and repositories that are being specifically developed for the management and publication of environmental data.