Transition to a new cloud-based data platform for statistics production (Dapla)

- With a particular focus on geospatial data and GIS

Svein Reid and Jørn K. Undelstvedt

Statistics Norway

Abstract

Rapid technological development and high external expectations mean that Statistics Norway constantly must further develop and improve products and services. At the same time, the financial framework will be limited. There is therefore a need to streamline and renew, and to release resources that can make room for development.

In the coming years, the development work at Statistics Norway will primarily concentrate on the transition to new cloud-based data platform for statistics production (Dapla). At the same time, we will continue the work to adapt our products and services to the agreed target groups, and we will work for more efficient collection, use and sharing of data. The main reasons why Statistics Norway decided a few years ago to develop a new cloud-based data platform for statistics production are:

- to be better equipped to meet increasing user expectations and take advantage of technological developments
- to have increased opportunities for employees to contribute to innovation, streamlining and continuous improvement of statistics production processes
- to get better opportunities to reuse and share code, methods and data
- to be able to increase the use of machine-to-machine data capture and handle changes at register owners more quickly
- to provide better information security
- a need for more scalable data storage and a better ability to handle larger amounts of data

The roadmap for development provides an overview of prioritized themes that Statistics Norway has identified as decisive to realize the goals in the strategy. The objective for the development work is that all statistical and research products are going to be modernized and produced at Dapla, and on-premises systems to be turned off.

As for statistics production based on geospatial data and GIS tools, Dapla introduces a transition from proprietary software to open-source software. The working modality is more code-based and much less “model building”. We use Onyxia – a platform for deploying open-source software (like QGIS) in virtual environments through Kubernetes, developed by The French National Institute of Statistics and Economic Studies (INSEE). Code, mostly Python in Jupyter-Lab, is put on GitHub. Pandas and GeoPandas are Python libraries. Data storage in Google Cloud. The geospatial data are in (geo)parquet.

The presentation/talk will introduce some challenges in the transition process and its possible impact on the quality of statistical products and publications.

Keywords: cloud-based data platform, statistics production, geospatial data