Contribution ID: **14** Type: **not specified**

Ilhan Mutlu (COMPBC): From Incomplete to Insightful: Curating Spatially Annotated Monitoring Data with CleanGeoStreamR

Wednesday 7 May 2025 11:45 (15 minutes)

Authors: Ilhan Mutlu, Jana Schor, Jörg Hackermüller

Abstract: "Environmental monitoring data is frequently compromised by inconsistent, incomplete, or erroneous spatial metadata, which restricts its effective utilization in research and informed decision-making processes. Recently, we have developed the CleanGeoStreamR to overcome such challenges in an automated way. CleanGeoStreamR utilizes an automated pipeline that normalizes textual entries, corrects spatial annotations, and systematically fills critical data gaps. By incorporating advanced data normalization and reverse geocoding techniques, the tool significantly enhances metadata quality while simultaneously increasing the volume of reliable data available for comprehensive analytics and artificial intelligence applications. For instance, we have automatically filled more than 107 million data gaps in the NORMAN surface waters database with CleanGeoStreamR. This corresponds to approximately 37% of the already existing data gaps.

Recent advancements in the CleanGeoStreamR tool have expanded its flexibility and adaptability, enabling broader applicability across various environmental monitoring datasets. Originally developed to address inconsistencies and gaps in spatial metadata—particularly in the surface water data set—CleanGeoStreamR now supports a more generic and configurable approach to data curation. Users can define both core and optional columns to be retained and tailor the curation workflow through editable configuration files. These enhancements make it possible to adapt the curation pipeline to other datasets, including but not limited to biota, sewage water, and wastewater monitoring, etc.

By offering a transparent, reusable, and highly customizable framework, CleanGeoStreamR facilitates harmonized data preparation across institutions and projects. Its open-source nature and adherence to FAIR principles make it an ideal tool for collaborative research and cross-organizational data integration efforts, fostering consistent and high-quality environmental datasets at scale."

Session Classification: Session II: Data science and modelling , Chair: tbd