Ensuring data trustworthiness within SMART Monitoring of environmental processes

Uta Koedel\textsuperscript{1}, Peter Dietrich\textsuperscript{1,3}, Philipp Fischer\textsuperscript{2}, and Claudia Schuetze\textsuperscript{1}

\textsuperscript{1}UFZ- Helmholtz Centre for Environmental Research GmbH, Monitoring-and Exploration Technologies, Leipzig, Germany (uta.koedel@ufz.de)
\textsuperscript{2}AWI- Alfred Wegener Institute
\textsuperscript{3}Eberhard-Karls-University of Tübingen

The term SMART Monitoring was also defined by the project Digital Earth (DE), a central joint project of eight Helmholtz centers in Earth and Environment. SMART Monitoring in the sense of DE means that measured environmental parameters and values need to be specific/scalable, measurable/modular, accepted/adaptive, relevant/robust, and trackable/transferable (SMART) for sustainable use as data and improved real data acquisition. SMART Monitoring can be defined as a reliable monitoring approach with machine-learning, and artificial intelligence (A.I.) supported procedures for an “as automated as possible” data flow from individual sensors to databases. SMART Monitoring Tools must include various standardized data flows within the entire data lifecycle, e.g., specific sensor solutions, novel approaches for sampling designs, and defined standardized metadata descriptions. One of the SMART Monitoring workflows’ essential components is enhancing metadata with comprehensive information on data quality. On the other hand, SMART Monitoring must be highly modular and adaptive to apply to different monitoring approaches and disciplines in the sciences.

In SMART monitoring, data quality is crucial, not only with respect to data FAIRness. It is essential to ensure data reliability and representativeness. Hence, comprehensively documented data quality is essential and required to enable meaningful data selection for specific data blending, integration, and joint interpretation. Data integration from different sources represents a prerequisite for parameterization and validation of predictive tools or models. This data integration demonstrates the importance of implementing the FAIR principles (Findability, Accessibility, Interoperability, and Reusability) for sustainable data management (Wilkinson et al. 2016). So far, the principle of FAIRdata does not include a detailed description of data quality and does not cover content-related quality aspects. Even though data may be FAIR in terms of availability, it is not necessarily “good” in accuracy and precision. Unfortunately, there is still considerable confusion in science about the definition of good or trustworthy data.

An assessment of data quality and data origin is essential to preclude the possibility of inaccurate, incomplete, or even unsatisfactory data analysis applying, e.g., machine learning methods, and avoid poorly derived, misleading or incorrect conclusions. The terms trustworthiness and representativeness summarise all aspects related to these issues. The central pillars of
trustworthiness/representativeness are validity, provenience/provenance, and reliability, which are fundamental features in assessing any data collection or processing step for transparent research. For all kinds of secondary data usage and analysis, a detailed description and assessment of reliability and validity involve an appraisal of applied data collection methods.

The presentation will give exemplary examples to show the importance of data trustworthiness and representativeness evaluation and description, allowing scientists to find appropriate tools and methods for FAIR data handling and more accurate data interpretation.