December 2019

Newsletter of the *Digital Earth* Project

Contributions of Forschungszentrum Juelich to Digital Earth

This newsletter presents project contributions and broader activities of Forschungszentrum Juelich conducted in the framework of Digital Earth.
Contact: Andreas Petzold, a.petzold@fz-juelich.de

New perspectives on quality assurance and quality control of environmental observation data

Martin G. Schultz[1], Ralf Kunkel[2] and Andreas Petzold[3]
[1] Jülich Supercomputing Centre, [2] Institute of Bio- and Geosciences 3 - Agrosphere, [3] Institute of Energy and Climate Research 8 - Troposphere, Forschungszentrum Jülich GmbH, Jülich, Germany

The current developments towards the establishment of a culture of FAIR (Findability, Accessibility, Interoperability, Reusability) and open data puts high requirements on individual scientists as well as on research institutions operating data repositories, concerning all aspects of data management, curation and stewardship. In that context, the quality of data is a crucial, albeit not explicitly mentioned requirement of data FAIRness as it is important to ensure reusability of observation data. Documented data quality is essential to build trust in users and allow meaningful selection of data for specific data use cases. This has become even more important with the introduction of low cost sensors and citizen science in many areas of environmental sciences.
While traditional repositories, hosting data from scientific or regulatory monitoring initiatives or from scientific field campaigns could usually rely on a more or less rigid quality assurance chain, environmental monitoring in the future will include much less well characterized instruments that are employed with less stringent operating procedures. Modern data science methods may offer the potential to fuse such diverse datasets, so that some valuable information can be extracted also from lower quality measurements, but in order to validate and interpret the results from such analyses it is essential to know the quality of that data.
Quality assurance can be defined as "part of quality management focused on providing confidence that quality requirements will be fulfilled", whereas quality control is essentially the "inspection" component of quality assurance, i.e. "part of quality management focused on fulfilling quality requirements". [1] From the data management and data curation perspective, the focus is clearly on quality control, whereas a research network or infrastructure must of course also deal with other aspects of quality assurance, such as ensuring traceable calibration chains, defining standard operating procedures, etc.
Similar quality control issues occur in different monitoring domains and across different types of data. Recently, a number of projects and initiatives have begun to harmonize data quality control efforts (e.g. ENVRI-FAIR and NFDI) and to develop software tools that can assist in the quality control across (environmental) research domains. In this newsletter, some recent developments with respect to quality control in the context of Digital Earth are presented. This topic is the focus of Forschungszentrum Jülich in the Digital Earth initiative.

[1] https://asq.org/quality-resources/quality-assurance-vs-control

AutoQC4Env: A Novel Quality Control Tool for Data Ingestion Workflow, Using a Statistical Model

Najmeh Kaffashzadeh, Felix Kleinert, Martin G. Schultz, Lukas Leufen, and Sabine Schröder
FZ Jülich, Jülich Supercomputing Centre

As the data volume grows larger and more complex, the automatization of data quality control (QC) has become of primary concern. Although various QC tools and software have been developed, most of them are based on subjective screening which can be error-prone. Furthermore, some of them have been created for a particular dataset or purpose. The challenge is to provide techniques and tools that explicitly target the automatization of QC throughout the whole data ingestion cycle, thereby allowing a data centre to check the data QC before the data publishing process.
At the Forschungszentrum Jülich, we developed a novel quality control tool (AutoQC4Env), in which a sequence of robust statistical tests is implemented. The tool has been presented at the Digital Earth workshop (QA-QC) on 25 February 2019 and it has since been developed further. The statistical tests are categorized into several sub-groups (G0, G1, G2 and G3) with increasing complexity. A probability-score was introduced as an indicator for the plausibility of individual data points in an arbitrary environmental time series. All test parameters can be easily (re)configured via editing plain text files in JSON format. This tool was implemented in a Python framework with modern software architecture. We supply a tool-chain that automates large parts of the framework, thus allowing it to be continuously applied throughout the data life cycle with little effort.
As a case study, the AuthoQC4Env was embedded in the data ingestion workflow of the TOAR database (Figure 1).

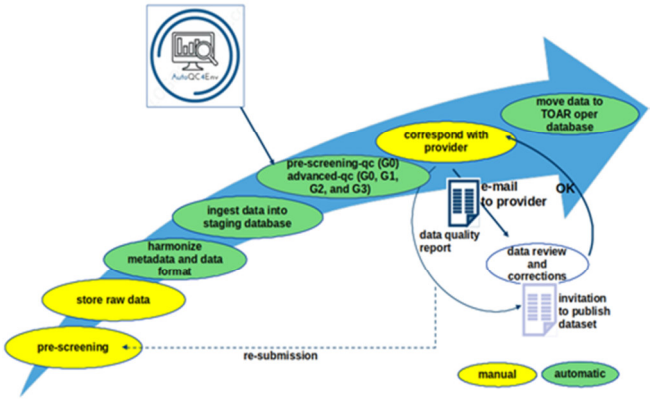

Figure 1 AutoQC4Env tool embedded in the TOAR data ingestion workflow.

The software is called automatically from the workflow manager as soon as the data have been imported into a special staging database. In the future, it would also become possible to invoke the tool from a web interface, either to double-check a new data series or to apply statistical tests to data that are already in the public database. The AutoQC4Env framework has been written

with a broad array of users and environmental time series in mind. It can be applied as a stand-alone tool or easily integrated into python-based workflows. The alpha-version of the AutoQC4Env tool has been released in May 2019 within the Digital Earth community.

Near Real Time Processing Tool in IAGOS

Susanne Rohs and Mihal Rütimann
FZ Jülich Institute of Energy and Climate Research 8 - Troposphere

IAGOS (In-service Aircraft for a Global Observing System[2]) is a European research infrastructure which operates compact instruments on board of passenger aircraft. In long-term routine in situ observations of atmospheric chemical composition (O_3, CO, NO_x, NO_y, CO_2, CH_4), water vapour, aerosols, clouds, and temperature are conducted. A data set is achieved, which is unique in terms of spatial and temporal coverage. Access to data is free and open for the global scientific community.
One of the longest data series refers to water vapour and relative humidity which are key to numerical weather prediction, atmospheric radiative forcing and cloud formation processes.
The Copernicus Atmosphere Monitoring Service CAMS [3] requests this data in near-real-time (NRT) for the continuous validation of the chemical transport models operated by CAMS. These models provide quality-controlled information related to air quality and health, solar energy, greenhouse gases and climate forcing, everywhere in the world. For CAMS, NRT-transmission of the water vapour data, i.e. with a maximum of 3 days difference between measurement and automated data transmission, is specifically requested. The automated processing of data in NRT as operated by IAGOS is a central contribution to Digital Earth.
Due to the large amount of data generated daily by IAGOS and the required speed of data evaluation for NRT data provision, an automated data processing and quality control (QC) of the data is required. For this purpose, a modular program package was developed, which transfers the raw data into quality-controlled NetCDF files (see Figure 2a).

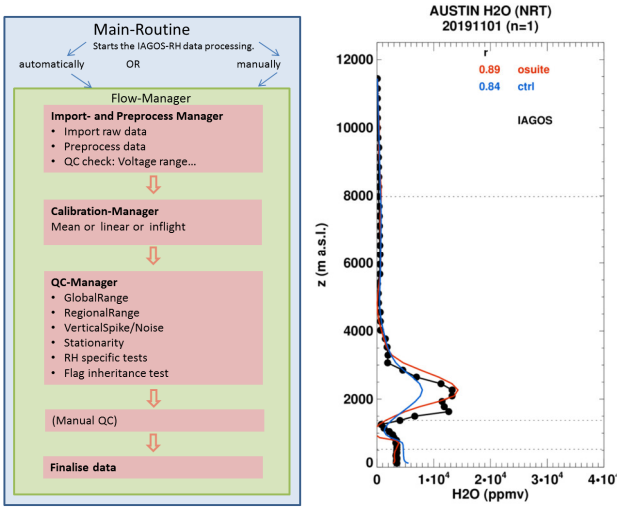

Figure 2 a) Toolbox for H2O data processing (currently written in MATLAB); b) Example of "Profile of the day" at www.iagos-data.fr/cams/nrt_profiles_h2o.php.

After the reading of the data and first quality tests, the relative humidity over water and the relative humidity over ice are calculated from the raw data using the calibration coefficients determined from the associated calibrations. This is followed by a comprehensive automated review of the data quality.
The NRT provision of H_2O data is operational since November 2019, see Figure 2b. This achievement is a major success for IAGOS. In the framework of Digital Earth we have started to optimize the NRT data processing tool by harmonizing data quality control through the use of the AutoQC4Env Python framework for automated QC, and by designing and implementing a modularly structured data base. The results will be made accessible to Digital Earth.

Enabling Quality Control of Sensor Web Observations in TERENO

Ralf Kunkel, Anusurija Devaraju and Jürgen Sorg
FZ Jülich Institute of Bio- and Geosciences 3 - Agrosphere

The rapid development of sensing technologies had led to the creation of large amounts of heterogeneous environmental observations. The Sensor Web provides a wider access to sensors and observations via common protocols and specifications. Observations typically go through several levels of quality control, and aggregation before they are made available to end-users. Raw data are usually inspected, and related quality flags are assigned. Data are gap-filled, and errors are removed. New data series may also be derived from one or more corrected data sets.
Quality measures (e.g., accuracy, precision, tolerance, or confidence), the levels of observational series, the changes applied, and the methods involved must be specified to use the data series. It is important that this kind of quality control information is well described and communicated to end-users to allow for a better usage and interpretation of data products.
Using the TERENO initiative[4] as a starting point a quality control framework for processing and assessing environmental time series within the TERENO data infrastructure has been developed. Custom data workflows are defined for the different ways data are imported into the infrastructure. A scheme to describe the data processing levels specifies the underlying data processing, assessment and accessibility. A two-tiered quality flag scheme is adapted to represent the different flag systems of different sensing applications. The observation data model and the web services used for data provision (SOS) are modified, so that observation data with metadata of quality control are accessible in the Sensor Web. Another application utilizing these is the customized Sensor Web. The framework will be applied and extended within the Digital Earth project.

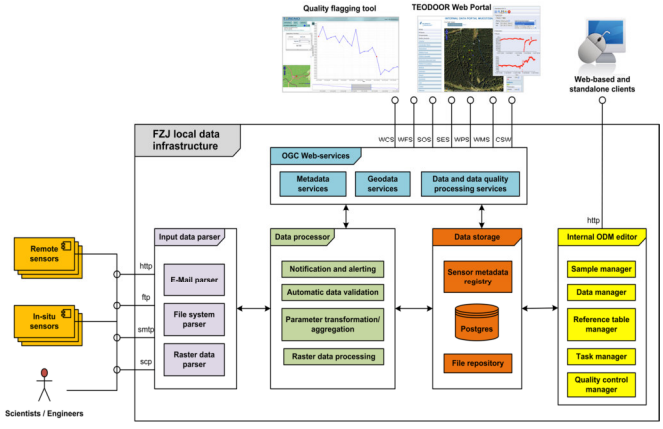

Figure 3 Overview of the TERENO Times Series Management System incorporating the quality control framework.

[2] https://www.iagos.org
[3] https://atmosphere.copernicus.eu/
[4] https://www.tereno.net

Crossing the Boundaries of Applied Earth System Science in ENVRI-FAIR

Andreas Petzold and Daniela Franz
FZ Jülich Institute of Energy and Climate Research 8 - Troposphere

The EU project ENVRI-FAIR[5] builds on the Environmental Research Infrastructure (ENVRI) community that includes principal European producers and providers of environmental research data and research services. The ENVRI community integrates the four subdomains of the Earth system - Atmosphere, Ocean, Solid Earth, and Biodiversity/Terrestrial Ecosystems. The environmental research infrastructures (RI) contributing to ENVRI-FAIR have developed comprehensive expertise in their fields of research, but their integration across the boundaries of applied subdomain science is still not fully developed. However, this integration is critical for improving our current understanding of the major challenges to our planet such as climate change and its impacts on the whole Earth system, our ability to respond and predict natural hazards, and our understanding and preventing of ecosystem loss.

ENVRI-FAIR targets the development and implementation of the technical framework and policy solutions to make subdomain boundaries irrelevant for environmental scientists. Harmonization and standardization activities across disciplines together with the implementation of joint data management and access structures at RI level facilitate the strategic coordination of observation systems required for truly interdisciplinary science. ENVRI-FAIR will finally create an open access hub for environmental data and services provided by the contributing environmental RIs, utilizing the European Open Science Cloud (EOSC) as Europe´s answer to the transition to Open Science.
FAIRness assessment tools, standardization and harmonization approaches developed in ENVRI-FAIR will directly contribute to respective Digital Earth activities, ensuring the coherent developments in the national and European open science efforts.

[5] https://envri.eu/home-envri-fair/