Geophysical Research Abstracts
Vol. 18, EGU2016-17783, 2016
EGU General Assembly 2016

Open data used in water sciences - Review of access, licenses and understandability

Esa Falkenroth (1), Emma Lagerbäck Adolphi (2), and Berit Arheimer (1)
(1) SMHI, Hydrological Research, Norrköpping, Sweden (berit.arheimer@smhi.se), (2) Uppsala University

The amount of open data available for hydrology research is continually growing. In the EU-funded project SWITCH-ON (Sharing Water-related Information to Tackle Changes in the Hydrosphere – for Operational Needs: www.water-switch-on.eu), we are addressing water concerns by exploring and exploiting the untapped potential of these new open data. This work is enabled by many ongoing efforts to facilitate the use of open data. For instance, a number of portals provide the means to search for open data sets and open spatial data services (such as the GEOSS Portal, INSPIRE community geoportal or various Climate Services and public portals). However, in general, many research groups in water sciences still hesitate in using this open data. We therefore examined some limiting factors.

Factors that limit usability of a dataset include: (1) accessibility, (2) understandability and (3) licences. In the SWITCH-ON project we have developed a search tool for finding and accessing data with relevance to water science in Europe, as the existing ones are not addressing data needs in water sciences specifically. The tool is filled with some 9000 sets of metadata and each one is linked to water related key-words. The keywords are based on the ones developed within the CUAHSI community in USA, but extended with non-hydrosphere topics, additional subclasses and only showing keywords actually having data.

Access to data sets: 78% of the data is directly accessible, while the rest is either available after registration and request, or through a web client for visualisation but without direct download. However, several data sets were found to be inaccessible due to server downtime, incorrect links or problems with the host database management system. One possible explanation for this could be that many datasets have been assembled by research project that no longer are funded. Hence, their server infrastructure would be less maintained compared to large-scale operational services.

Understandability of the data sets: 13 major formats were found, but the major issues encountered were due to incomplete documentation or metadata and problems with decoding binary formats. Ideally, open data sets should be represented in well-known formats and they should be accompanied with sufficient documentation so the data set can be understood. The development efforts on Water ML and NETCDF and other standards could improve understandability of data sets over time but in this review, only a few data sets were provided in these formats. Instead, the majority of datasets were stored in various text-based or binary formats or even document-oriented formats such as PDF. Other disciplines such as meteorology have long-standing traditions of operational data exchange format whereas hydrology research is still quite fragmented and the data exchange is usually done on a case-by-case basis. With the increased sharing of open data there is a good chance the situation will improve for data sets used also in water sciences.

License issue: Only 3% of the data is completely free to use, while 57% can be used for non-commercial purposes or research. A high number of datasets did not have a clear statement on terms of use and limitation for access. In most cases the provider could be contacted regarding licensing issues.