Geophysical Research Abstracts
Vol. 17, EGU2015-12303, 2015
EGU General Assembly 2015

Review of access, licenses and understandability of open datasets used in hydrology research

Esa Falkenroth (1), Berit Arheimer (1), and Emma Lagerbäck Adolphi (2)
(1) Swedish Meteorological and Hydrological Institute, Sweden, (2) Uppsala Universitet, Sweden

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), 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 (such as the GEOSS Portal and the INSPIRE community geoportal) provide the means to search for such open data sets and open spatial data services. However, in general, the systematic use of available open data is still fairly uncommon in hydrology research.

Factors that limits (re)usability of a data set include: (1) accessibility, (2) understandability and (3) licences. If you cannot access the data set, you cannot use if for research. If you cannot understand the data set you cannot use it for research. Finally, if you are not permitted to use the data, you cannot use it for research.

Early on in the project, we sent out a questionnaire to our research partners (SMHI, Universita di Bologna, University of Bristol, Technische Universiteit Delft and Technische Universitaet Wien) to find out what data sets they were planning to use in their experiments. The result was a comprehensive list of useful open data sets. Later, this list of data sets was extended with additional information on data sets for planned commercial water-information products and services. With the list of 50 common data sets as a starting point, we reviewed issues related to access, understandability and licence conditions.

Regarding access to data sets, a majority of data sets were available through direct internet download via some well-known transfer protocol such as ftp or http. 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 data sets 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.

Regarding understandability of the data sets, the issues encountered were mainly 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. Furthermore, machine-readable format would be preferrable. Here, the development efforts on Water ML and NETCDF and other standards should improve understandability of data sets over time but in this review, only a few data sets were provided in these wellknown formats. Instead, the majority of datasets were stored in various text-based or binary formats or even document-oriented formats such as PDF. For some binary formats, we could not find information on what software was necessary to decipher the files. Other domains 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 in hydrology research.

Finally, regarding licensce issue, a high number of data sets 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.