EMS Annual Meeting Abstracts
Vol. 18, EMS2021-43, 2021
https://doi.org/10.5194/ems2021-43
EMS Annual Meeting 2021

Data rescue and digitization through image recognition

Hans Olav Hygen, Christoffer Artturi Elo, and Herdis Motrøen Gjelten
Norwegian Meteorological Institute, Oslo, Norway (hans.olav.hygen@met.no)

MET Norway has, like many other NHMS, more than a century of data, where a large portion of early measurements are not digitized. At MET Norway these data were stored on microfilm, which was smouldering away. In the autumn of 2020 were we able to scan all of these microfilms, which produced 1.3 million pictures with 1 to 6 datasheets per image. Each image contains daily data from one month. These images were produced automatic, and basically without metadata.

Moving these images towards data requires multiple steps, and a manual job in-house in MET would be too costly. In other projects at MET Norway image recognition has been applied to e.g. cloud classification. Based on this, image recognition was applied. We apply this in a multi-stage approach to ensure the quality of every stage and gain confidence in the technology. The first stage is to make a catalogue of which stations is represented in each image. The next stage is to capture the metadata such as the year and month of each image/datasheet. Stage three is to extract the data from each image and prepare it for MET Norway's data storage and distribution. As part of this, the regular quality control from MET Norway will be performed on the dataset, thus ensuring that the quality is up to MET Norway's regular standard.

We are still in the early stages of the project. As stated are the microfilms scanned, and we have used image recognition to create a catalogue of which image belongs to which stations. We have also extracted which year is represented in each picture. This was done due to this information been in printed letters. The observations are handwritten and have so far been to be harder to extract. An internal website has been established to monitor the progress from the image recognition and limited manual corrections.