

Digitization of analogue records in time (DO IT!)

Dr.-Ing. Müller-Thomy, Hannes

TU Braunschweig, LWI-Department of Hydrology and River Basin Management

16.06.2023

6 months

Abstract

Thousands years of analogue records exist in earth system sciences, but are still an unburied treasure in the age of digitization. The digitization of these data needs manpower – but not necessarily with scientific background. Breaking down the digitization to an easy applicable smartphone application enables the involvement of citizen scientists to overcome the manpower bottleneck. This app can be applicable to all kind of analogue data and hence useful for numerous scientific fields.

I. Introduction

Records of environmental observations exist for decades before the digital era either as hand-written values or registrations from measuring instruments. Unfortunately, these data are most often not digitized yet and hence are not accessible, although it would be highly valuable for long-term analyses, trend detections or similar. Usually the data are digitized manually by internships, student assistants or employees if there is some spare time. Due to the required manpower, this way of digitization will take several decades to be completed. Furthermore, the quality of paper records can degrade significantly over time. In many places, there is a risk of losing these unique data in the cellars of public authorities. The novelty of our approach is to enable the involvement of numerous volunteers without scientific background (citizen scientists =CS) by breaking down the time-consuming part of the digitization to an easy applicable smartphone application. By providing scans of the analogue data the CS can type the hand-written numbers into a standardized table or retrace the recorded graph with their fingers. Beside the digitalized data, the most important outcome of the project will be the developed smartphone app, which can flexibly handle different types of data and thus be transferred to other scientific fields. The PI has worked with hydrological data during his study of hydrology (2004-2010), in numerous hydrologic projects during (2010-2016) and after his promotion (2016-today) at TU Dresden, Leibniz Universität Hannover, TU Wien and TU Braunschweig. Furthermore, the PI has lectureships on ‘Citizen Science in Hydrology’ and ‘Data analysis for hydrologic-hydraulic modelling’ at TU Braunschweig, which proves his knowledge of all aspects of citizen science and data interpretation.

II. Incubator Project description

The workflow of the project is shown in Fig. 1 (left), extended by the workflow in follow-up applications (Fig. 1, right). For the incubator project already existing scans of analogue data will be used. For the initial development of the smartphone app including its structure and routines three months will be required. Two more months will be necessary to identify pitfalls and unforeseen user-based issues with a group of volunteering CSs (e.g. rephrasing explanations for better/not misleading information). The sixth month will be used for the final documentation for GitHub, launching the app and spreading the word to create awareness of the community (platforms: e.g. eu-citizen.science/, buergerschaffenwissen.de/projekte) and the institutions with analogue records, and implement additional scans if they exist already.

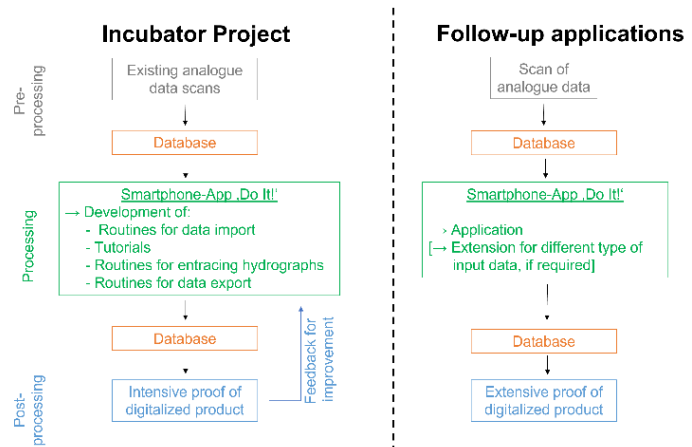


Fig. 1: Workflow of the Incubator Project (focus: smartphone app development) and for follow-up applications

How does the digitization work?

Fig. 2 a shows a scanned stage hydrograph (water level over time) as an example of the dataset provided by the LUBW (Landesanstalt für Umwelt Baden-Württemberg). The app provides the scan to the CS for digitizing the stage hydrograph. Therefore, the temporal scale (x-axis) and quantitative scale (y-axis) have to be identified. The CS marks the lower (point of origin) and upper bounds of the scales (max(x) and max(y)), which enables to relate the subsequent retracing of the stage hydrograph to these axes. Afterwards the CS will be asked for the values of both, lower and upper bounds. Subsequently, the CS follows for the retracing the stage hydrograph with its finger, and the smartphone app translates this movement in relation to the lower and upper bounds into absolute stage values for the single time points.

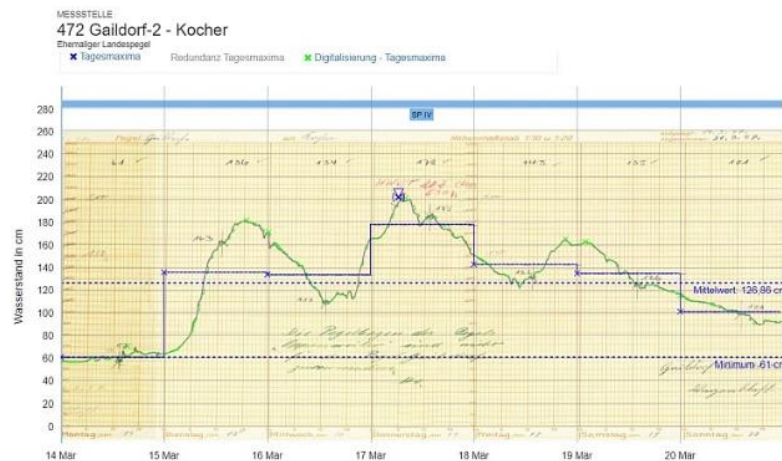


Fig. 2: Scan of a stage hydrograph (green) including manual notes (red, black) and resulting statistics (blue) (LUBW homepage, 09.06.2023, 14:00 CET)

The quality of the digitalized data will be ensured on different levels:



Individual level:

- i) When starting the smartphone app for the first time, the CS has to join a tutorial as training session in the smartphone app. This ensures the correct application of the provided tools and compares the digitization results of the first e.g. three scans with the correct values, so the CS gets a feedback immediately.
- ii) The CS can always choose a region of interest to digitize data from e.g. his home region first. This can also increase the motivation of the CS.
- iii) After the digitization of each scan the CS will be provided with the scan itself and the digitized stage hydrograph. The CS will be asked if the retrace is ready for submission or should be modified before.

Community level:

- iv) Each scan will be handed out several times. A minimum of three digitized versions enables statistical comparisons. If the deviations among the versions are too high (tolerance values are determined by the data provider, e.g. maximum difference per time step and/or maximum of cumulative difference over a certain period), it will be handed out again. If deviations will remain after one or two additional digitization runs, the scan will be flagged for identification on expert level. If no deviation occurs, the digitized data set will become part of the digital database.
- v) There will be gamification elements for maintaining the motivation of the CS in the long-term. By earning status points for each scan digitized, the CS can distinguish as champions of the month. The implementation of other gamification elements will follow at a later stage as well.

Expert level:

- vi) The original data provider are referred to as experts. They will analyze why certain scans from the community level got flagged to provide a feedback to the smartphone app development team (or to improve their scan procedures). Additionally, experts can define gauge-related thresholds, to get alerted if a digitized scan exceeds the threshold for detailed analysis in the post-processing stage (e.g. rare flood event).

III. Relevance for the NFDI4Earth

The expected direct profiteers are the federal agencies, weather services, and many other institutions with years of analogue records, which won't be digitized otherwise. Based on the FAIR principles the data will be made open accessible, so every institution with interest in the data benefits. As repository the NFDI-repository OneStep4All seems promising to provide the digitized data for the point of its original measuring. From the results based on the extended data bases (e.g. improved flood risk management plans) all citizens will profit by the digitized data indirectly. Science itself profits by the involvement of CS with probably no scientific background since their involvement enables scientific communication.

IV. Deliverables

The main deliverable will be the developed smartphone app. The code will be open-source published on GitHub. The digitized data will be published open-accessible according to the FAIR principles. The smartphone app includes example data used in the tutorials and data for digitizing. It will be applicable to all kind of analogue data and hence useful for numerous scientific fields.