TERMS OF REFERENCE FOR A SOLICITED WRC PROJECT

KEY STRATEGIC AREA:	**Water Resources and Ecosystems**
THRUST:	**6**
PROGRAMME:	**1**
Title:	**Text mining to enhance hydroinformatics**

***General* Objectives:**
To develop a framework and tools to extract water and related data and information from online platforms.

Specific:

- Framework and guidelines to extract water, weather and related data and information from online platforms.
- Develop and/or recommend suitable tools for data and information extraction.
- Develop an intermediary database to store and curate extracted information– including quality control protocols.

Rationale:

The successful management of water resources is founded on the maintenance of long-term water databases and the accessibility of these data to the public. Social media, media and other crowdsourcing platforms are rich in both qualitative and quantitative data collected by individual citizens and social groups. In addition, other sources of data are from formal, regulatory, and statutory sources obtained from ground-based and remotely sensed sensors. Recently, the WRC has invested in programmes to standardise citizen science approaches for surface water, groundwater, and weather[1]. The intention of this project is not to duplicate, but to build on the work already done. The citizen science approaches will develop towards mobile app-based data and information collection. Social media and general online data and information can be collected to enhance our understanding of water, weather and related phenomena and serve as a mechanism to improve the public's understanding and awareness of the role and impact of water in the environment. During extreme events or disasters it is not always possible to gather data and information in real-time but people often post their observations and photos of such

[1] See for example Nuapia et al. 2021. Imagining Solutions for Extracting Further Value from Existing Datasets on Surface and Groundwater Resources in Southern Africa. WRC Report No. TT 842/20. Available from WRC Knowledge Hub: www.wrc.org.za

events. They may even share data from their home-based instruments (raingauge, thermometers, etc.).

The main approach will be through text mining - text mining is the process of deriving high-quality information from text. It involves "the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources"[2]. These sources can be groups from interested parties such as groups set up during droughts, to track weather and climate or groups set up by institutions with special instructions such as hashtags and tagging for discovery. Engaging different groups to harmonise and standardise hashtags for southern African contexts and discovery.

Deliverables:

1. Inception report
2. Text mining framework and preliminary guidelines
3. Text mining tools and approaches
4. Pilot and demonstrations on using text mining approaches, and curation in intermediary database
5. Handover report to Water Research Observatory project for incorporation

Time Frame:
4 months

Total Funds Available:
R 500 000

[2] Wikipedia