BIG DATA ANALYTICS AND TRANSBOUNDARY WATER COLLABORATION CONSOLIDATION OF DATA AND APPLICATION OF BIG DATA TOOLS TO ENHANCE NATIONAL AND TRANSBOUNDARY DATA SETS IN SOUTHERN AFRICA THAT SUPPORT DECISION-MAKING FOR SECURITY OF WATER RESOURCES

Andrew Gemmell, Arnaud Sterckx, Claudia Ruz Vargas, Claudia Ruz Vargas, Ebrahiem Abrahams, Tyrel Flügel, Rui Hugman, Badisa Mosesane

USAID
FROM THE AMERICAN PEOPLE

science & innovation
Department:
Science and Innovation
REPUBLIC OF SOUTH AFRICA

GROUNDWATER MANAGEMENT INSTITUTE

IBM Research | Africa

SWP
SUSTAINABLE WATER PARTNERSHIP

USGS
science for a changing world

WATER RESEARCH COMMISSION

TT 844/20

Big Data Analytics and Transboundary Water Collaboration

Consolidation of Data and Application of Big Data Tools to Enhance National and Transboundary Data Sets in Southern Africa that Support Decision-Making for Security of Water Resources

Report to the
Water Research Commission

by

Andrew Gemmell[1], Arnaud Sterckx[2], Claudia Ruz Vargas[2], Claudia Ruz Vargas[2], Ebrahiem Abrahams[1], Tyrel Flügel[1], Rui Hugman[1], Badisa Mosesane[3]

[1]UMVOTO Africa (Pty) Ltd.
[2]International Groundwater Resources Assessment Centre
[3]University of Botswana

WRC Report No. TT 844/20

February 2021

Obtainable from
Water Research Commission
Private Bag X03
Gezina
PRETORIA, 0031

orders@wrc.org.za or download from www.wrc.org.za

The publication of this report emanates from a project entitled *Big Data Analytics and Transboundary Water Collaboration. Theme 1: Consolidation of Data and Application of Big Data Tools to Enhance National and Transboundary Data Sets in Southern Africa that Support Decision-Making for Security of Water Resources* (WRC Report No. K5/2880)

This report forms part of a series of four reports. The other reports are:

- *Imagining Solutions for Extracting Further Value from Existing Datasets on Surface and Groundwater Resources in Southern Africa* (WRC Report no. TT 842/20)
- *Localizing Transboundary Data Sets in Southern Africa: A Case Study Approach* (WRC Report no. TT 843).
- *Machine Learning Models for Groundwater Availability – Incorporating a Framework for a Sustainable Groundwater Strategy* (WRC Report no. TT 845/20)

DISCLAIMER

This report has been reviewed by the Water Research Commission (WRC) and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the WRC, nor does mention of trade names or commercial products constitute endorsement or recommendation for use

ISBN 978-0-6392-0237-2
Printed in the Republic of South Africa

Executive Summary

Effective management of groundwater resources requires ease of access to reliable data to support decisions. This is particularly important for transboundary systems where different environmental data are collected and managed by different institutions in different countries; thus, impeding the sustainable management of transboundary water resources.

Almost all real-world data and datasets suffer from incompleteness, inconsistencies and errors, and the water resource data received as part of this project were no different. Often these data issues are solved in an ad-hoc and manual manner requiring familiarity with the data. This report describes a programmatic approach that can be utilised to effectively deal with these issues. Further, a common repository for data storage and visualisation is introduced. The data that was collected as part of this assessment was provided by four organisations: Department of Water and Sanitation (DWS) Botswana, DWS South Africa, IMWI South Africa, and SADC-GMI.

The objective of this project was to utilise Machine Learning tools; however, the amount data available for our study area was insufficient for pure Machine Learning. As a result, the Big Data tools used focussed on statistics, visualization, and data comparisons using the Python programming language.

The project developed software tools to automate the collation and quality control of data from existing online databases into a format suitable for our primary repository, the existing SADC-GIP platform. The steps followed were data gathering, data assessment (e.g. detect outliers or any inconsistencies), data cleaning, and initial data quality control, all of which were undertaken programmatically using Python tools and documented in a Jupyter Notebook. This included work done during an internship at IBM by an intern from the University of Botswana. In addition, the data was shared to other online data platforms GGMN and GEMStat. The three platforms were selected as they each provide complimentary benefits.

There were four themes in the project, and our theme (Theme 1) provided the initial data to the other three themes. However, since the projects took place in parallel, Theme 1 was only able to share cleaned data to the other themes once they had started their analyses. Further, the objective was to incorporate data derived from the other three themes in SADC-GIP; however, this was not able to be included in the database during the project due to the parallel work.

As part of the project, a quality control system was formalised. This included protocols on designing a monitoring network, procedures for recording data (including metadata, logical and outlier checks), and a standard procedure for handling erroneous or dubious data.

To support groundwater data sharing between the Departments of Water Affairs of Botswana and South Africa about the Ramotswa transboundary aquifer protocols on groundwater data exchange were developed. The protocol could be taken up easily by other organisations undertaking projects on transboundary groundwater in the SADC region. The data sharing protocol included:

- Data and metadata needs for assessing transboundary aquifers
- Technical solutions to organize the exchange of data
- Frequency of Updates of the Transboundary Database
- Data Sharing Policy (SADC-GIP, GGMN, GEMStat)

Challenges and learnings from the project included:

- Memorandum of Understanding not in place at commencement of project; thus delay in receiving data from Primary Data Providers. Time required to source and negotiate access to data.
- There is a need to on-board data providers at the earliest stage.

- Limited/sparse primary data, especially temporal. Some data of uncertain fidelity / accuracy / providence (i.e. locations, missing metadata).
- Impacts of delay in internship/trainings on data analytics process.
- Parallel work, thus could not include other theme outputs in our dataset used at internship. Internships at differing times, limiting collaboration.

The way forward proposed by Theme 1 is to:

- Promote data sharing cross-border, and a shared data repository. Repository should be able to display surface water and groundwater data; including water quality, surface flows, and groundwater levels. Should allow visualisation, statistics and comparisons.
- Platforms should work together to increase cross-platform standardization, e.g. data requirements.
- Enhance in situ data with model outputs and satellite imagery “triangle”.
- Automated data processing techniques using Python to be adopted for other databases and data types to enhance a shared repository.

Table of Contents

Chapter **Description** **Page**

List of Tables

List of Figures

List of Abbreviations

DWS	Department of Water and Sanitation
e.g.	For example
GMI:	Groundwater Management Institute
ICWRGC	International Centre for Water Resources and Global Change
IGRAC	International Groundwater Resources Assessment Centre
IGS	Institute for Groundwater Studies
IWMI	International Water Management Institute
JSAP	Joint Strategic Action Plan
N.A.	Not Applicable
QA/QC	Quality assurance and quality control
RTBAA	Ramotswa Transboundary Aquifer Area
SADC	Southern African Development Community
SADC-GMI	Southern African Development Community Groundwater Management Institute
WRC	Water Research Commission

1. INTRODUCTION AND BACKGROUND

The Ramotswa dolomitic aquifer is a vital source of water for people and ecosystems in the Southern African Development Community (SADC) region. Located on the border between Botswana and South Africa, the aquifer is important for the upper part of the Limpopo River Basin, which is shared by Botswana, Mozambique, South Africa, and Zimbabwe. The Ramotswa aquifer makes important contributions to water security in the Gaborone Dam catchment and supplies water to communities on both sides of the border, primarily for domestic use and small-scale agriculture. The Ramotswa Transboundary Aquifer Area (RTBAA) encompasses a region beyond the strict boundary of the aquifer to include the surface waters that interact directly with the aquifer. The primary water resource institutions managing the groundwater resource – both with regards to quality and quantity – are the Botswana Department of Water and Sanitation, and the South Africa Department of Water and Sanitation (DWS). (IWMI, 2020).

1.1. Programme Overview

1.1.1. The Big Data Analytics and Transboundary Water Collaboration for Southern Africa

This research project, managed by the Water Research Commission of South Africa, is part of a series of four projects under the Big Data Analytics and Transboundary Water Collaboration for Southern Africa, bringing together key stakeholders in Water and Big Data sectors.

The Collaboration was first conceptualised in 2014 during the African Leaders Forum in Washington D.C., between USAID Global Development Lab and IBM Africa Research, which had opened its first hub in Nairobi (Kenya) in 2013, followed by the Johannesburg Lab in 2015. Since the early 2000s, the regional USAID mission for Southern Africa had been intensifying its regional support for transboundary water systems with both the Ramotswa Aquifer Project, involving Botswana and South Africa and the Resilience in the Limpopo River Basin Program (currently in its second phase with the Resilient Waters Programme, covering the entire Southern Africa region, with a focus on the Limpopo and Okavango River Systems). As part of this process, USAID had also been engaging with the Southern African Development Community (SADC): Groundwater Management Institute and the Department of Science and Innovation of South Africa to support knowledge and technological advancement in the region. The focus of this multi-agency collaboration was agreed as Big Data Analytics and Transboundary Water. On April 3 2017, the partners met with a multi-stakeholder regional group in a dynamic “Idea Jam” hosted by the IBM Africa Research Lab in Johannesburg. The objective was twofold:

- To answer the broad question “how best can big data analytics be used to enhance transboundary water management”, and
- To identity the research questions, which would have guided the projects.

Requiring the collaboration of at least five high profile government agencies and private institutions, it took over one year to move from the Idea Jam to the launch of the Call for Proposals in August 2018, and the awarding of the four research projects in January 2019.

1.1.1.1. The Collaboration: its partners and objectives

Currently, the Collaboration has seven partners, with a joint function for USAID Global Development Lab, Water Office, and Southern African Mission. The partners each contributed to the development of the research projects based on own technical and funding capacity, see **Figure 1-1**. The total funds provided by the Funding Partners to research directly amount to USD $ 500,000 (40%, 40%, 20%). IBM Africa contributed with the provision of the venue in Johannesburg, ad hoc, but more importantly, by sponsoring the internship programme to the five candidates from the research projects.

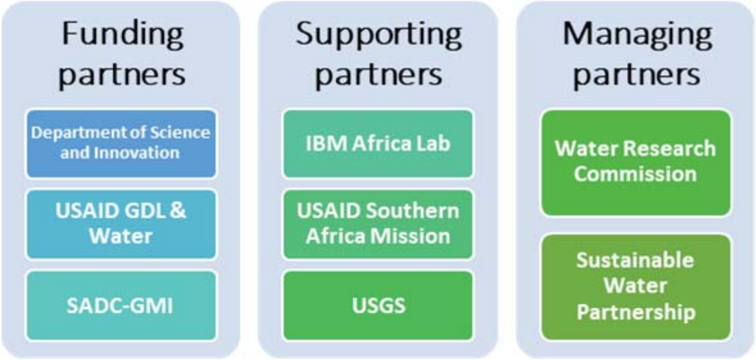

Funding partners
Department of Science and Innovation
USAID GDL & Water
SADC-GMI
Supporting partners
IBM Africa Lab
USAID Southern Africa Mission
USGS
Managing partners
Water Research Commission
Sustainable Water Partnership

Figure 1-1: Collaboration partners & functions

The Water Research Commission (WRC) is primarily tasked to oversee the financial and implementation management of the four research projects, as well as final reporting. The Sustainable Water Partnership (SWP) was called in by USAID in 2018 to act as the overarching Programme Coordinator, tasked with providing relation management, overall objective achievement, direction and positioning for the Collaboration in the region, and the fostering of a Community of Practice.

The United States Geological Society (USGS), IBM Research and SWP provided three sets of online training on issues pertaining to the focal topics of the Collaboration, which are now available on the Collaboration YouTube channel.

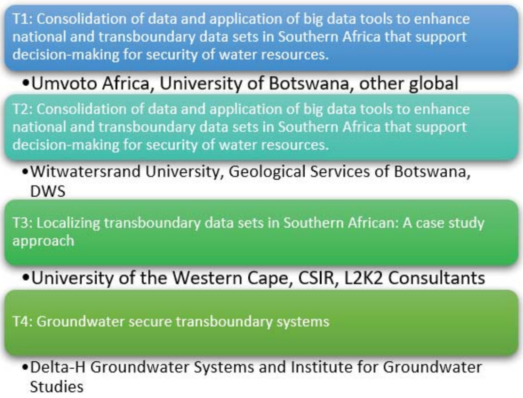
The Collaboration partners defined the objectives for this first phase of action, see **Table 1-1**. However, the long-term vision is to create a Community of Practice for research and innovation on Big Data for Water Security, building on the multi-donor environment which has proven successful.

Table 1-1: Collaboration goals and objectives

Goals	Objectives
Enhance current understanding of shared groundwater resources	Improve transboundary ground water management and collaboration
Provide big data skills development, capacity building and networking opportunities for Southern African researchers and their students	To foster multi-agency collaborative funding opportunities
To promote innovative thinking and application of Big Data Analytics to the Transboundary Water sector for integrated decision-making	To plant the seed for a growing community of pioneers in the use of Big Data Analytics for the study and management of Transboundary Water Aquifers

1.1.1.2. Research projects: funding and training

The four projects were awarded between December 2018 and January 2019, with a focus on a secondary river basin in the region: the Ramotswa, part of the Limpopo River Basin, spanning Botswana and South Africa. All the lead institutions of the project teams have partnered, see **Figure 1-2**, with Botswana government and private institutions, as well as other leaders in previous water programme in the area, such as UN-IGRAC (partner of Team 1) and IWMI, implementers of the Ramotswa 2 USAID Project.



T1: Consolidation of data and application of big data tools to enhance national and transboundary data sets in Southern Africa that support decision-making for security of water resources.

- Umvoto Africa, University of Botswana, other global

T2: Consolidation of data and application of big data tools to enhance national and transboundary data sets in Southern Africa that support decision-making for security of water resources.

- Witwatersrand University, Geological Services of Botswana, DWS

T3: Localizing transboundary data sets in Southern African: A case study approach

- University of the Western Cape, CSIR, L2K2 Consultants

T4: Groundwater secure transboundary systems

- Delta-H Groundwater Systems and Institute for Groundwater Studies

Figure 1-2: Title of the four thematic areas and projects

Despite working independently to address own project topics, the four research teams have progressively worked together to provide better integration for their outcomes. This process was led by the SWP in respect of providing a communication forum for the team leaders but was enhanced by the Internship Programme. The IBM mentors created a dedicated team and engaged the interns as individuals, as well as a group to help each other resolve new questions in coding and Machine Learning.

1.1.1.3. The future prospects

As the current phase is coming to an end with the closing of the four research projects, the Collaboration partners are already identifying new opportunities to build on the lessons learnt and address the gaps recognised in this preliminary work, enhance the partnership to include national and regional government stakeholders, as well as new funding partners.

The focus of the Collaboration will remain the nexus between Big Data Analytics and (Transboundary) Water Security, recognising the inter-relatedness of successful water management in both national and shared aquifers to both human development and environmental goals.

1.1.2. Background

Twenty-eight transboundary aquifers have been identified in the SADC region, some of which are heavily relied upon for drinking water supply, irrigation or livestock production. The sustainable management of these strategic water resources requires all stakeholders from the riparian states to cooperate, in first place the national (ground)water departments that are in charge of groundwater protection and development. At the core of transboundary cooperation is acquisition and sharing of data and information, which is necessary to develop efficient water management strategies.

Few transboundary aquifers in the SADC region have already been subject to cooperation between Member States. The first case study was the Stampriet Transboundary Aquifer System (STAS), shared between Botswana, Namibia and South Africa. Within the Groundwater Resources Governance in Transboundary Aquifers (GGRETA) project, funded by the Swiss Development Cooperation and with technical support of UNESCO-IHP and IGRAC, data were collected from the riparian states, harmonised, and used to make a first multidisciplinary assessment of the aquifer. To support the exchange of data between the three countries, a dedicated web-viewer was set up in the Global Groundwater Information System, the groundwater data sharing platform managed by IGRAC[1]. These activities were carried out from 2013 to 2015. During a second phase of the project (2016-2018), the STAS Multi-Country Cooperation Mechanism was operationalized and nested into ORASECOM, the Orange-Senqu River Commission. Data collected in phase I were moved to a new data sharing platform managed by

[1] https://www.un-igrac.org/global-groundwater-information-system-ggis

ORASECOM[2].

The second case-study was the Ramotswa transboundary aquifer, shared between Botswana and South Africa. Like for the STAS, a first multidisciplinary assessment was made in a first project phase (2015-2017), based on data collected from the states and harmonized. The project was funded by USAID and lead by IWMI. As a project partner, IGRAC provided a data sharing platform within the GGIS, the Ramotswa Information Management System (RIMS[3]), which was used by the Departments of Water Affairs of Botswana and South Africa to share groundwater data. Each country nominated a RIMS manager, responsible for data content and authorizations to access the data. This work was continued in the second phase of the project (2017-2019).

The importance of transboundary water cooperation and the prevalence of shared aquifers demands coordinated planning and joint action. Towards this, a Joint Strategic Action Plan (JSAP) was developed for the RTBAA (IWMI, 2019); the first strategic action plan developed for a transboundary aquifer in SADC. The JSAP utilized stakeholder consultations between 2016-2019 to determine an aspirational vision and specific objectives, targets and actions to address key issues related to the Ramotswa: The overall vision shared between Botswana and South Africa was **Water security and sustainable socioeconomic development in the RTBAA through joint research and management** (IWMI, 2019). The key JSAP actions were categorized into one of three components:

1. Managing water for sustainable use, availability and access;
2. Enhancing institutions and capacity; and
3. Expanding research and knowledge.

1.1.3. Theme Interaction

Each of the Themes have some degree of overlap with each other, including the focus on the Ramotswa system. As a result, during the project Reference Group meeting held from 28-29 January 2019, the gaps and synergies between each were discussed. The input that Theme 1 had to the other themes is summarised below:

Theme 2:

- Provided water quality data to Theme 2.
- It is intended to incorporate obtained citizen science and non-conventional data from Theme 2 within Theme 1 data set, as well as any modelled data.

Theme 3:

- Provide data for the Ramotswa Area to Theme 3, potentially to validate downscaling results.
- Data from Theme 4 and from Theme 3 for larger area to potentially validate downscaling results.
- It is intended to incorporate downscaled results in Theme 1 data set

Theme 4:

- It is intended for the dashboard of data storage system (Theme 1) to be guided by outputs of sustainability framework developed by Theme 4 (i.e. linked to sustainability indicators).
- Where possible, include modelling results in Theme 1 dataset

[2] gis.orasecom.org

[3] ramotswa.un-igrac.org

1.2. Motivation for the importance of this project

Effective management of groundwater resources requires ease of access to reliable data to support decisions. This is particularly important for transboundary systems where different environmental data are collected and managed by different institutions in different countries and also where data access and dissemination are very challenging because these datasets are diverse, often of variable quality, with inconsistent resolutions, being fragmented in nature and in many cases only accessible to a small community of users impeding the sustainable management of transboundary water resources.

Groundwater management decisions taken by an institution in one country can have repercussions on users or the environment in neighbouring countries. Transboundary aquifer management not only requires access to reliable data, but also transparency with regards to available data and collaborative decision taking between the various entities that rely on the system.

A common repository for data storage and visualisation creates a tool that allows all parties to access data on the aquifer system relevant to its management. By pooling data, water resource management can be more effective. Recognizing the need for a systematic collection and sharing of groundwater data, IGRAC took the initiative to develop the Global Groundwater Monitoring Network (GGMN) and is supporting SADC-GMI in managing the Southern African Development Community Groundwater Information Portal (SADC-GIP). In addition, the Global Freshwater Quality Database GEMStat provides scientifically-sound data and information on the state and trend of global inland water quality, which may assist in the management and development of transboundary systems. These visualisation platforms are described and compared in the following chapters.

1.3. Aims and Objectives

The overall aims and objectives of the Theme 1 approach was to:

- Identify **data holding organizations** and facilitate **data sharing** and collection. Focus on groundwater levels and quality
- Analyse, format and filter data for **compatibility**.
- Develop inter-operable **standard protocols** in data capture, quality control, storage and processing; including consistent approach regarding the parameters and units of measure.
- Upload of data to **platforms** to capture, store and process data. Includes selection of a shared data repository and make recommendations on long-term hosting.
- **Visualisation tools**, incl. RIMS, GGMN & GEMStat
- **Big Data Analytics** to improve existing datasets.

A training in the use of the selected database tools will be provided – date to be confirmed.

1.4. Project Team

1.4.1. Umvoto

Umvoto Africa [Pty] Ltd is an environmental consulting company based in Cape Town, South Africa. The company, established in 1992, is a multi-disciplinary earth and water science company that undertakes studies in water resource planning, water governance, integrated water resource management, discrete event and numerical modelling (hydrology, groundwater flow and contaminant transport modelling), hydrogeology, geology and seismology in addition to Disaster Risk Management and space geodesy. The company has a strong geospatial team, which provides in-house support to as well as undertaking purely GIS projects. Umvoto is recognized as an innovative, leading practitioner in the inter-disciplinary fields of integrated water resource planning, development, management, governance and modelling.

1.4.2. IGRAC

IGRAC is a UNESCO category 2 centre, specialises in regional- and transboundary-level assessment and monitoring of groundwater resources with a focus on among others managed aquifer recharge and groundwater governance. IGRAC has specific experience with the RAMOTSWA Transboundary Aquifer Area, and within the SADC in general. IGRAC will provide research contributions related to these items. One of IGRAC's flagship products is the Global Groundwater Information System (GGIS). It is a web-based Geographic Information System, which supports the storage, visualisation, analysis and sharing of groundwater data and information through map-based modules. This highly customisable system allows for the incorporation of new thematic and/or regional (project) modules. One of the products within the GGIS is the SADC-Groundwater Information Platform which includes the RAMOTSWA phase 1 and phase 2 projects. The Global Groundwater Monitoring Network (GGMN) is also part of the GGIS. GGMN has been developed to store, visualise and analyse time series data.

1.4.3. University of Botswana

The Hydrology and Water Resources department of the University of Botswana has experience in the RAMOTSWA system with the key individual being Professor Piet Kebuang Kenabatho. Professor Kenabatho has experience in the Ramotswa system, including as assessment of the effects of small dams in the upstream of Gaborone dam, and groundwater contamination in the Ramotswa wellfields in Botswana. In addition to this expertise, the individual for the IBM internship, Mr Badisa Mosesane was from the University of Botswana which aided in the transboundary nature of this project.

1.4.4. GEMS/Water

GEMS/Water, a programme of the United Nations Environment Programme (UN Environment), in cooperation with participating countries, is creating a unique global water quality monitoring network that provides water quality monitoring data to the global water quality database and information system called GEMStat. These data can be used for assessing status and trends in global inland water quality and tracking progress towards Goal 6 of the new Sustainable Development Goals. Contributions to this project by GEMS/Water included technical input and capacity building, as well as use of the GEMStat global water quality database and information system to represent water quality data.

1.5. Overview of methodology/approach to the project

The focus of the Theme 1 approach was on groundwater levels and quality. The broad approach followed is shown in **Figure 1-1** and summarised below:

- Identified data holding organizations and facilitate data gathering.
- Assessed data for compatibility.
- Cleaned data. Developed inter-operable standard protocols in data capture, quality control, storage and processing with a consistent approach to parameters and units of measure.
- Big Data Tools were used following the following steps
    - Gather
        - Code to extract and view all data in various formats (zip, shp, kmz, dbf, xls, csv, etc.)
        - Collect metadata describing the data
    - Assessing (visually and programmatically for quality and structural issues)
        - Groundwater data grouped into water levels, water quality, lithology using code
        - Data formatted using code to allow easy data analysis. This included preliminary data quality assessment (e.g. missing data, structural issues such as columns, rows, etc.)
    - Cleaning
        - Quality and structural issues were listed into defined tasks (i.e. remove duplicates, replace missing data with -9999/unknown, etc.), which was transformed into code
        - Each file was concatenated and structured according to RIMS format using code
    - Visualisation
        - Boreholes displayed graphically using code to assess the amount of data, density and areas where data exists. Data then ready for visualization in RIMS

1.6. Data Providers

Data was provided by four organisations, viz. DWS Botswana, DWS South Africa (i.e. Geo-Water Affairs North West and Geo-Water Affairs Pretoria, IMWI South Africa, and SADC-GIP). The data provided by each is summarized as follows in **Table 1-2**:

The Ramotswa project identified the priority data providers as the Botswana Department of Water Affairs and the South Africa Department of Water and Sanitation. Relationships with these departments have been developed over time by IWMI, SADC-GMI and IGRAC.

Collaborators from the Botswana Department of Water Affairs with experience specific to the Ramotswa system have been identified. Collaborators include Thato Sethloboko (Head of Groundwater Division); Keodumetse Keetile (focal person towards the RAMOTSWA-2 project); and Bochengedu Somolekae working on a World Bank funded project for Botswana to improve access to their groundwater data. The data typically comes at no cost to research organisations.

As part of the RAMOTSWA-2 project, staff within the South African Department of Water Affairs have been identified, including Mxolisi Mukhawana and Fhedzisani Ramusiya.

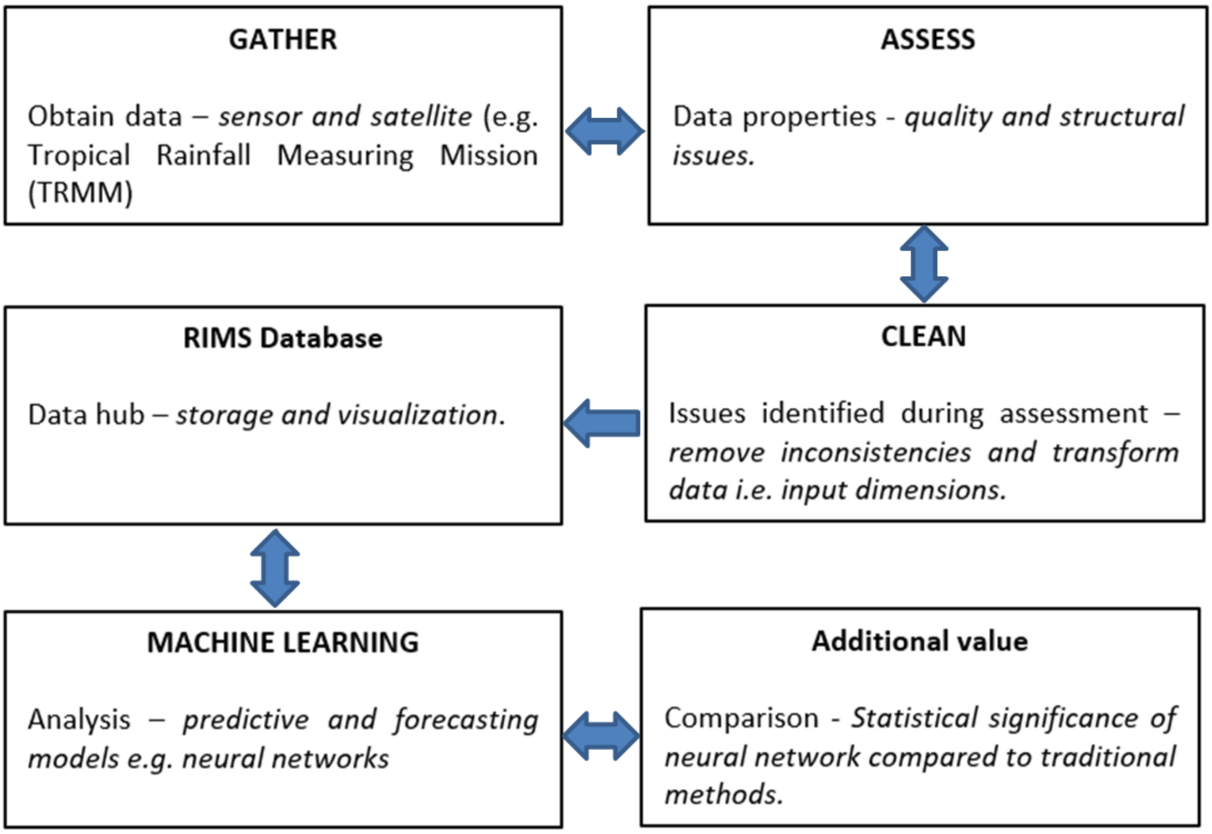

GATHER
Obtain data – *sensor and satellite* (e.g. Tropical Rainfall Measuring Mission (TRMM)
ASSESS
Data properties - *quality and structural issues.*
CLEAN
Issues identified during assessment – *remove inconsistencies and transform data i.e. input dimensions.*
RIMS Database
Data hub – *storage and visualization.*
MACHINE LEARNING
Analysis – *predictive and forecasting models e.g. neural networks*
Additional value
Comparison - *Statistical significance of neural network compared to traditional methods.*

Figure 1-3: Project Flow Process

The rationale for the establishment of the SADC-GMI was based on the importance of groundwater in the region and the need to set up a "Centre of Excellence" for groundwater management and groundwater dependent ecosystems in the region, and to have an institution that will serve as an interlocutor with national, regional and international groundwater initiatives and institutions. The institute is hosted by the University of the Free State in Bloemfontein, South Africa on behalf of, and under the strategic guidance of the SADC Secretariat, Directorate of Infrastructure's Water Division, in Gaborone, Botswana. SADC-GMI is currently implementing multiple projects in the SADC Region in partnership with key players in the water sector. This includes the mandate to pick up the lessons learnt and best practices from the past Ramotswa investigations for possible replication in the remaining more than 25 transboundary aquifers in the SADC region.

The International Water Management Institute (IWMI) is a non-profit, scientific research organization focusing on the sustainable use of water and land resources in developing countries. IWMI works in partnership with governments, civil society and the private sector to develop scalable agricultural water management solutions that have a real impact on poverty reduction, food security and ecosystem health. IWMI is a member of CGIAR, a global research partnership for a food-secure future.

IWMI and IGRAC are both involved in the still running RAMOTSWA. Previously, the two organisations already cooperated while developing the training manual on 'Integration of Groundwater Management into Transboundary Basin Organizations in Africa'.

Table 1-2: List of data providers, files and summary description of data provided.

Data provider	Files received (Original)	Summary
IMWI South Africa	Copy of Ramotswa Groundwater-monitoring data	Groundwater level time series data and water quality measurements. Water quality (physio-chemical) parameters include electrical conductivity, total dissolved solids, fluoride and nitrate measurements.
SADC-GIP	gip_BHdata_botswana.xls, gip_BHdata_southafrica_northwest.xls	Consists of borehole information (e.g. water levels and water quality parameters) and lithology (e.g. rock type). Water level information includes water depth, water (mamsl), etc. Water quality includes pH, electrical conductivity, total dissolved solids, bromine, chloride, fluoride, nitrites, nitrates, phosphate, sulphates. Lithology includes rock type, etc. The data contained in these files are described in the SADC Groundwater Information Portal (SADC-GIP) – Data Overview report, 2017.
Botswana Department of Water and Sanitation	BotsRamsbhs.shp	Borehole construction and yield information. Construction information include depth of borehole, start-end date of construction, hole diameter, etc.
South African Department of Water and Sanitation	BasicInfo, Lithology and WaterLevels	Basic borehole information, water levels and lithology. Basic information includes borehole name and location/municipality/region. Water level information includes specific yield, water level status and lithology provide information of surface geology, depth to bottom and depth to top, formation name and lithology name.
IGRAC	Non-RIMS data	

1.7. Case Study overview – Ramotswa Aquifer

The study areas were not specified by WRC during the proposal stage, leaving the selection of transboundary systems to the bidders. These transboundary systems were then further refined during the project Reference Group meeting (28-29 January 2019). Each of the themes selected the Ramotswa transboundary system as the primary study areas, with Theme 3 and 4 also proposing the Shire transboundary system as a secondary comparison sites. The Ramotswa transboundary system comprises shared water resources between Botswana and South Africa. While the Shire transboundary system comprises shared water resources between Malawi and Mozambique.

The Ramotswa aquifer is shared between Botswana and South Africa. There are various multilateral agreements supporting cross-boundary access to water between Botswana and South Africa, including the United Nations Human Right to Water and Sanitation (Resolution A/RES/64/292) dated July 2010; the Revised Protocol on Shared Watercourses in the Southern African Development Community (SADC, 2000) and the Draft Articles on the Law of Transboundary Aquifers (United Nations resolutions A/RES/63/124 (2008) and A/RES/66/104 (2012).

Ramotswa, in the South-East District of Botswana, south-west of the capital Gaborone (Figure 1-4).

The majority of the information in the following sections is derived from the IWMI publication *Resilience in the Limpopo Basin: The Potential Role of the transboundary Ramotswa Aquifer. Baseline report – 15th June 2016.*

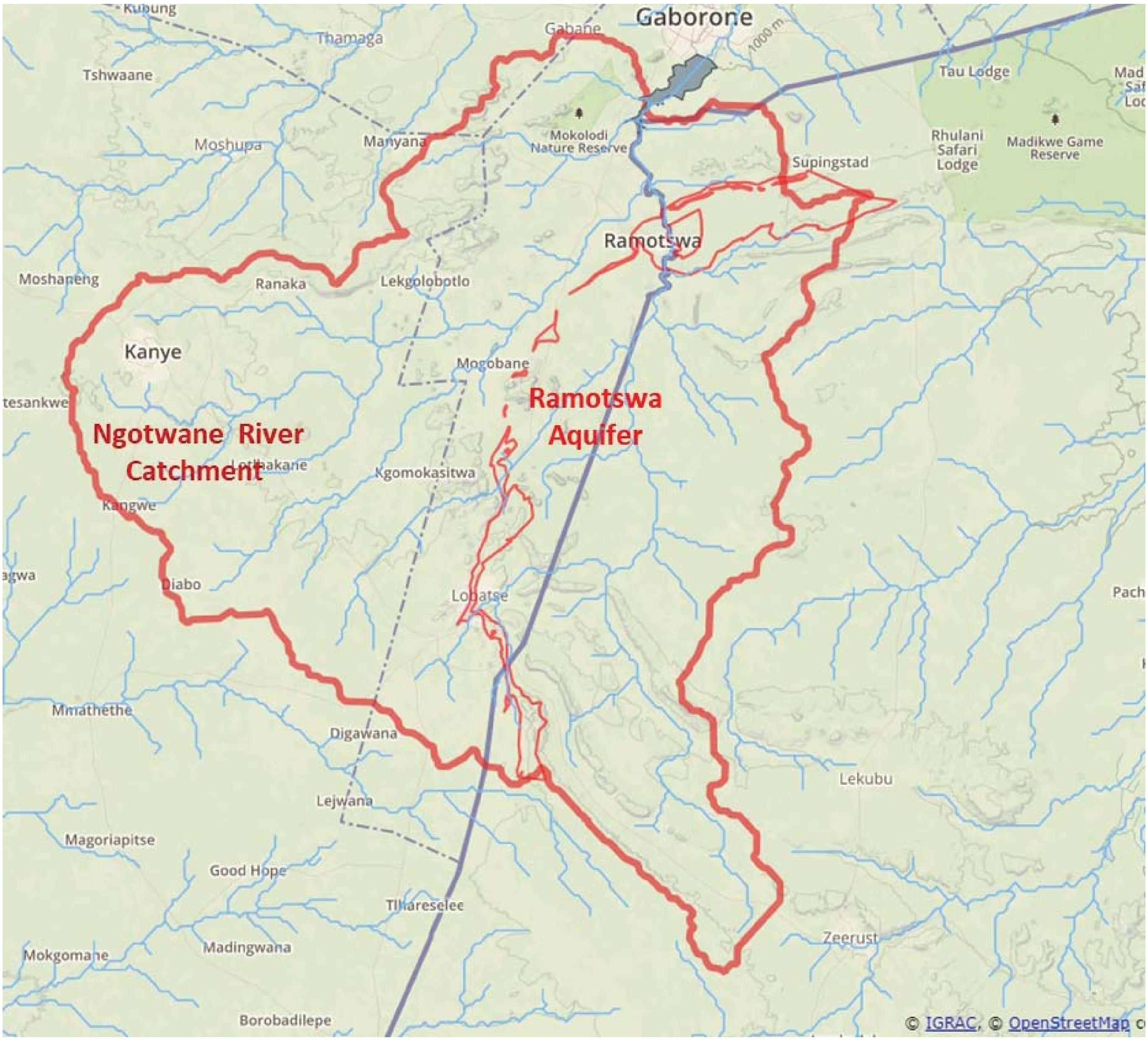


Figure 1-4: Location of outcrops of Ramotswa aquifer and RAMOTSWA project study area (the catchment of Gaborone dam reservoir, within the Ngotwane River Catchment). Source: RIMS

1.7.1. Hydrology and Topography

The Ramotswa aquifer is located in the Upper Limpopo River Basin, in the Marico and the Ngotwane sub-catchments. The Ngotwane River catchment is about 18,200 km^2 and starts in South Africa, while most of the catchment area is in Botswana (90% of the total catchment area). It is an ephemeral river flowing roughly north-eastward with river flow directly dependent on rainfall. The Marico River catchment is about 13,200 km^2 and starts in South Africa where most of the catchment area is situated (92%). It is a perennial river flowing roughly northward. The two main tributaries (Klein Marico and Groot Marico rivers) are located in the upper catchment and are fed by a number of springs within the Groot Marico dolomitic aquifer compartment.

The topography is fairly hilly, which is typical for southeast Botswana. The altitude of the plains ranges from 1,000 to 1,050 meters (m) above mean sea level (amsl), whereas the altitude of the surrounding hills varies from 1,068 to 1,189 mamsl. The hills and escarpments are remnants of erosion cycles which began in the Tertiary period. River infiltration into the aquifers occurs beneath the Ngotwane riverbed.

The climate is semiarid. Rainfall is strongly seasonal, with most of it occurring as thunderstorms during the summer period between October and April. Mean annual rainfall ranges from 450 to 600 mm and decreases from the eastern to the western side of aquifer area. Mean annual temperature ranges between 18 and 20°C. Maximum and minimum temperatures are experienced in January and July, respectively. Climate projections foresee not only uncertain trends in rainfall volumes in the future, but a likelihood of an increase in intensity of rainfall events.

1.7.2. Geology and Hydrogeology

The Ramotswa Aquifer is within the western part of the Transvaal Supergroup. The three sequences of the Transvaal Supergroup include, in stratigraphic succession from the oldest to the youngest:

- The Chuniespoort Group is typified by basal quartz arenites, thick dolomite, and iron formations.
- The Pretoria Group is typified by alternating mudrocks and sandstones, with volcanic horizons, diamictites, and conglomerates.
- The Rooiberg Group is composed of volcanics, mudrocks, sandstones, and felsites

Within the study area, the Ramotswa Aquifer is within the Pretoria Group and the Chuniespoort Group, the latter being the major water-bearing unit for the area. The main water-bearing unit of the Ramotswa Aquifer corresponds to the Malmani Subgroup (Chuniespoort Group) which is the dolomitic part of the Ramotswa Aquifer, called the “Ramotswa Dolomite”. The Ramotswa Dolomite has been intruded by three vertical to near vertical dolerite dike swarms. The hydraulic conductivity is enhanced by fracturing and karstification. The source of recharge to the upper karst zone is thought to be direct rainfall infiltration and the Ngotwane River.

As a complement, another water-bearing unit of the Transvaal supergroup is the Timeball Hill Formation of the Pretoria Group, called Lephala Formation in Botswana. This is located within the shales, quartzites and conglomerates. The Lephala Aquifer is not the focus of the study because it is not the main water-bearing unit of the Ramotswa Aquifer.

1.7.3. Groundwater Levels

Analysis of the static water level data in the study area shows that, generally, the depth to the groundwater table is shallower in the northern portion of the study area, where the average water level is approximately 24 m below the ground surface. In the southern portion, the average water level is approximately 41 m below ground surface. Data on water yield show that most boreholes in and around the study area have yields of 5 l/s (18m^3/hr) or less. Well yields are variable and increase in areas of karstification.

1.7.4. Groundwater Use

The Ramotswa aquifer includes the settlements of Ramotswa Town, Ramotswa Station (Taung), Boatle and the surrounding area in Botswana, and the Ramotshere Moiloa Local Municipality within the Ngaka Modiri Molema District Municipality in the Northwest Province of South Africa. In Botswana, the areas in the vicinity of the aquifer are dominated by urban development with high economic growth and an increase in population. In South Africa, there is stalling economic growth and low population density; however, the aquifer attracts considerable interest from national water planners because of rising water demands on the South African side, including the cities of Mafikeng, Zeerust and Lichtenburg.

Most of the water from the Ramotswa Aquifer is presently being abstracted in the Ramotswa Wellfield in Botswana. This wellfield was decommissioned in the mid-1990s because of issues with pollution but has been re-opened in 2014 as a results of reoccurring water shortages in the water supply for the Ramotswa – Gaborone area. Urban wastewater is presently reused for peri-urban crop production and watering of golf courses.

In the Botswana part of the study area, there are four wellfields used for domestic and industrial purpose: Ramotswa Wellfield, Lobatse Wellfield, Pitsanyane Wellfield and Woodland Wellfield. In South Africa, there are 718 boreholes recorded in the RAMOTSWA study area, but according to DWS time-series data on groundwater levels is only available for 2 boreholes are available. From the recorded 3,238 boreholes in the wider Ramotshere Moiloa Local Municipality only 499 boreholes (15.5%) have a specified purpose. Water from the boreholes is used for both agricultural (both crops and livestock) and domestic purposes. For those boreholes where a purpose was indicated, more than 70% of the boreholes in use supply agricultural activities. Most records do not indicate any purpose.

1.7.5. Groundwater Quality

The primary source of groundwater impact at the Ramotswa aquifer in the town of Ramotswa and the immediate vicinity of the Ramotwa wellfield is from pit latrines, resulting in increased nitrate concentrations. Due to the nitrate contamination, the Ramotswa Wellfield was decommissioned in 1996. Groundwater abstraction recommenced in 2014 as an emergency water source due to a drought. There is limited water quality data for South African portion of the aquifer.

2. DATA ANALYTICS AND BIG DATA

2.1. Overview and introduction to the key concepts relating to Big data and Data Analytics

> *Big data is a term applied to data sets whose size or type is beyond the ability of traditional relational databases to capture, manage, and process the data with low-latency (lag) – IBM*

Whilst big data has become a strong focus of global interest, increasingly attracting the attention of academia, industry, government and other organizations (Li et al., 2016), its rapid evolution in both public and private sectors, has resulted in a lack of common understanding and nomenclature of the concept to develop. For example, there is little consensus around the fundamental question of how big the data has to be to qualify as 'big data' (Gandomi and Haider, 2015). Furthermore, the terms 'Big Data Tools' or 'Big Data Analytics' are often used interchangeably without a clear definition of what they mean. In practice, these terms are often used to describe what is essentially 'data science' (the study of the generalizable extraction of knowledge from data) as applied to 'big data'. Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyse actual phenomena" with data (Hayashi, 1998a).

We therefore outline below what we will consider as 'big data' for this project, a necessary step to be able to assed the usefulness and selection of Big Data analytical 'tools' to a transboundary aquifer and finally review their potential for the processing and analysis of groundwater data using the Ramotswa aquifer as a case study.

2.1.1. Big Data

Two relevant proposed definitions of big data are:

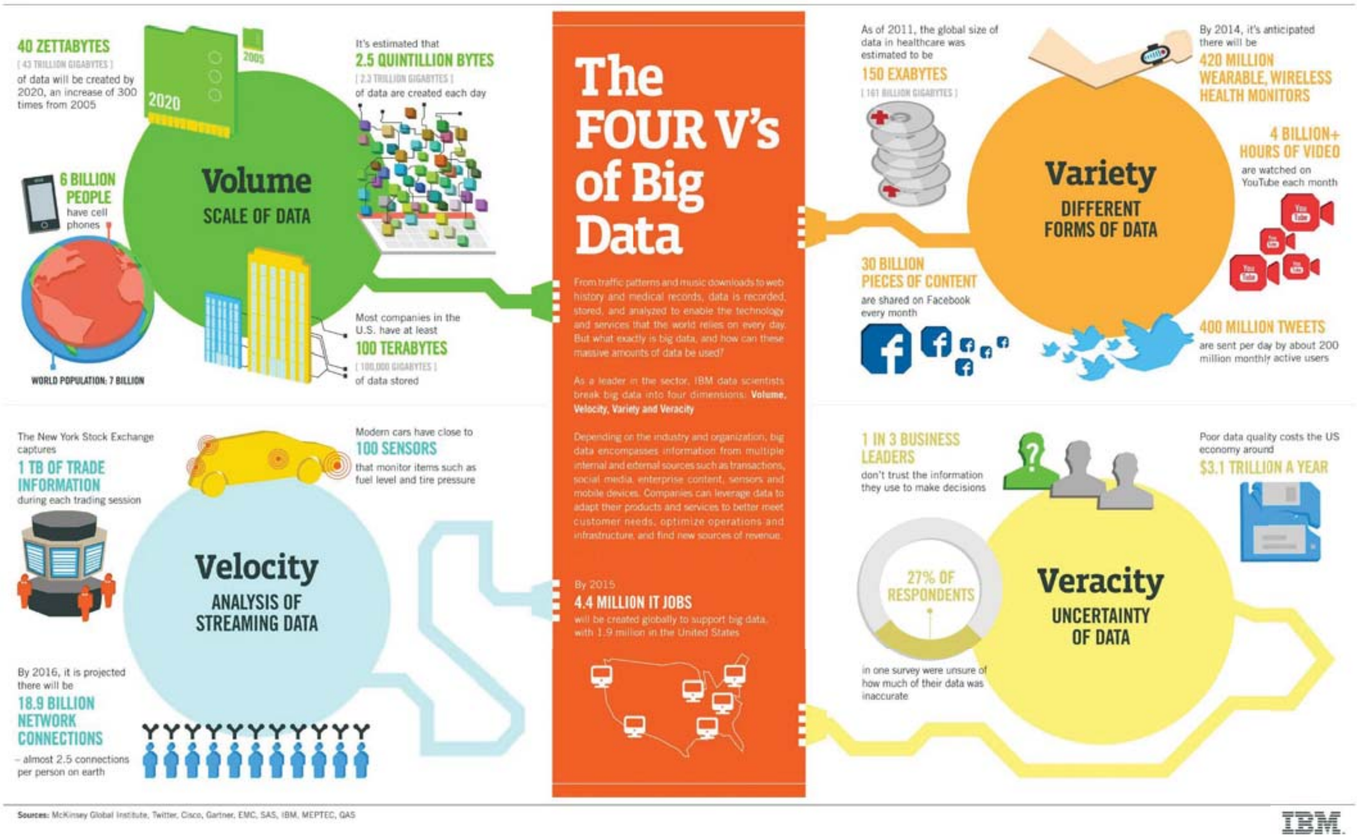
- "Big data is a term applied to data sets whose size or type is beyond the ability of traditional relational databases to capture, manage, and process the data with low-latency (lag). And it has one or more of the following characteristics – high volume, high velocity, or high variety." – IBM
- "Big data is a term used to describe data that is high volume, velocity, and variety; requires new technologies and techniques to capture, store, and analyse it; and is used to enhance decision making, provide insight and discovery, and support and optimize processes." – Adamala (2017)

The common stated characters of Big Data from the definitions above are:

1. Volume: the amount or quantity of data that has been created from all the sources. This not only includes mass storage (e.g. number of terabytes) but also data density (number of data points).
2. Velocity: the rate at which data is created, stored, analysed and processed. For example: the velocity of big data in remote sensing involves not only generation of data at a rapid growing rate, but also efficiency of data processing and analysis. In other words, the data should be analysed in a (nearly) real or a reasonable time to achieve a given task, e.g. flood predictions and warnings, where seconds can save hundreds or thousands of lives
3. Variety: the diversity of types of data, including structured, semi-structured, and unstructured (images) data sets depending on the sort of structure. For example, remote sensing data consist of multisource (laser, radar, optical, etc.), multi-temporal (collected on different dates), and multi-resolution (different spatial resolution) data, as well as data from different disciplines depending on several application domains.

Transboundary water resource data can be classified as ‘Big data’ due to its (1) **Variety** (for example data is collected on a diverse set of parameters, across different temporal and spatial scales; data types that provide integrated information can be diverse, from point-samples to satellite imagery; data can come from disciplines ranging from hydrological to social sciences, as well as institutions/entities), (2) its **Volume** (for example data can be collected from hundreds or thousands of locations, with long and high resolution time-series; the advent of easily accessible open-source satellite imagery can lead to large volumes of data collated; the potential for citizen science to contribute data will ideally lead to large volumes of additional data) and (3) its **Velocity** (for example telemetry allows for remote collection of data being transmitted from sensors in the field in near real-time).

An additional factor, as shown in **Figure 2-1** is **Veracity**: Uncertainty of data (i.e. how to deal with questionable data quality.


The FOUR V's of Big Data
Volume
SCALE OF DATA
40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2020
2005
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored
Variety
DIFFERENT FORMS OF DATA
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
Velocity
ANALYSIS OF STREAMING DATA
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth
Veracity
UNCERTAINTY OF DATA
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States
IBM.

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

Figure 2-1: The Four V’s of Big Data (IBM)

All this data proves a processing and storage crisis for enterprises. This is because manual approaches and traditional tools fail to scale to what is required for big data; hence the requirement of newer and improved approaches (and tools) is necessary to cope with this data. **Figure 2.2** shows a simple example of a customized data pipeline designed to transform unstructured datasets into a standardized format before being uploaded, visualised and interpreted in a visualisation platform. These large volumes of data collected has the potential to reveal useful trends and patterns. For organizations, both private and public, it is more important than ever to exploit the insights that this data holds to improve understanding of processes of a system or systems and potential risks that could potentially occur so that mitigation measures can be implemented in advance.

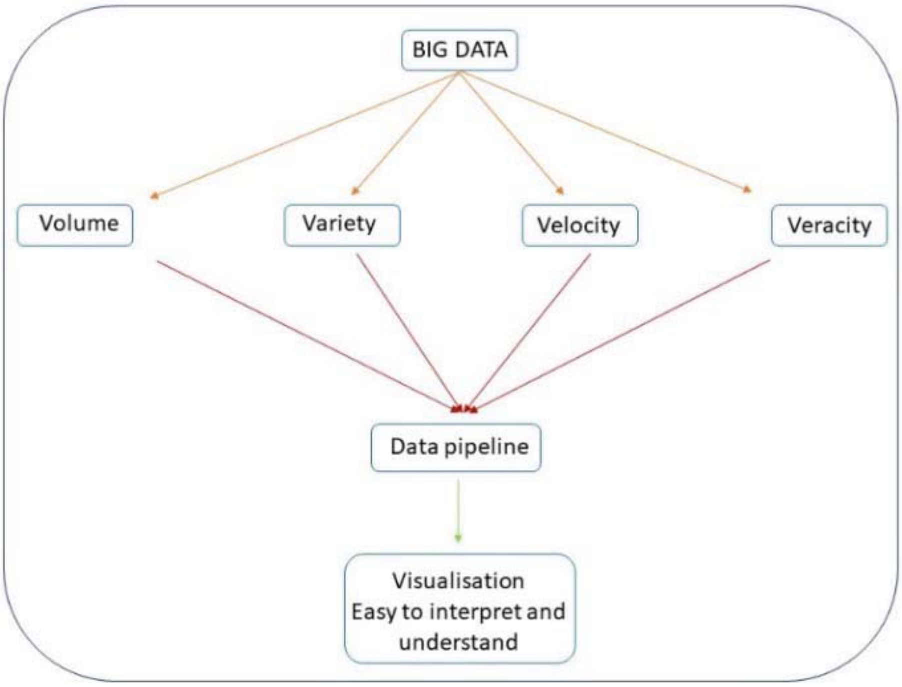


Figure 2.2: **Data pipeline developed to transform raw data into a standardized format before being uploaded to a visualisation platform such SADC-GIP to gain intelligence of a process/system (IGRAC, 2020).**

2.1.2. Big Data Analytics

The value of data is obtained when it is being applied to support or drive decision making. Thus, to gain value from Big data, the challenge becomes acquiring, processing and analysing large quantities of complex and diverse data into useful insights in a timely manner. This process can be divided into two parts: (1) data management and (2) data analysis. Data management involves processes and supporting technologies to acquire and store data and to prepare and retrieve it for analysis. Here we use the term data analysis, in which we include predictive analysis, to refer to techniques used to analyse and acquire intelligence from data.

Data analytics is a term often used interchangeably with data science, data mining, knowledge discovery and others – with no clear distinction. All of these terms refer to extracting useful information to inform decision making from a processed dataset. The analysis of extensive quantities of data and the need to grasp value out of individual behaviours require processing methods that go beyond the traditional statistical techniques (De Mauro, Greco and Grimaldi, 2015). The large volumes of heterogeneous data require automated or semi-automated analysis techniques to extract value within a useful timeframe. As such, the analysis of Big data tends to draw on techniques that can combine statistical analysis, optimization and machine learning to identify patterns, detect anomalies and build predictive models. Some of the Big Data analytical techniques used in Water Resources are summarised in Appendix A (Adamala, 2017). The various techniques can be grouped into four main types:

1. **Descriptive:** Explains what has happened in the past based on data presented in graphics and report, but not why it happened and what might happen in the future.
2. **Diagnostic:** Based on the descriptive analytics, seeks to understand the reasons why any given event took place in the past.
3. **Predictive:** Uses past data to model future outcome – predicts what could happen?
4. **Prescriptive:** Uses optimization to advise on how best to manage a system to obtain a desired result.

2.2. Big Data and Analytics Tools Applicable to Project

The aim of Theme 1 is to consolidate transboundary water data and to apply data science approaches to enhance national and transboundary data sets that support water resources security decision-making. As such, we focus on techniques that address the first three types (Descriptive, Diagnostic and Predictive) to describe the groundwater system, understand patterns and correlations in the data, as well as make predictions about future events. The methodologies that will be investigated for use during this project fall into the broad categories of machine learning, statistics and visualisation.

2.2.1. Machine Learning

Machine learning encompasses a range of techniques that apply algorithms and statistical models to allow computer systems to perform a specific task, without explicit programming, by inputting data to generate outputs. Unlike pre-programmed solutions, machine learning eliminates the need for an operator to continuously code or analyse data. Once programmed it automatically updates and learns from new sources of information. Machine learning tasks can be classified into several broad categories.

- In supervised learning, the algorithm builds a mathematical model from a set of data that contains both the inputs and the desired outputs. For example, if the task were determining whether a groundwater level was realistic, the training data for a supervised learning algorithm would include input with and without that realistic data values (the input), and each input would have a label (the output) designating whether it was realistic or not. Supervised learning can further be categorised into classification, regression and forecasting. Examples of supervised learning include Naïve Bayes Classifier, linear regression, random forest and Support Vector Machines.
- In contrast to supervised, unsupervised learning has no operator or programmer facilitating the learning process. In unsupervised learning the algorithm builds a mathematical model from a set of data which contains only inputs and no desired output labels. Unsupervised learning algorithms are used to find structure in the data, like grouping or clustering of data points. Unsupervised learning can discover patterns in the data, and can group the inputs into categories, as in feature learning. Unsupervised learning can be divided into two categories, clustering and dimension reduction (or association). Examples of unsupervised learning are K-Mean clustering, *a priori* algorithm, Principal Component Analysis (PCA), and density based spatial clustering.
- Reinforcement learning teaches the machine trial and error, learns from past experiences and adapts its approach in response to the situation to achieve the best results. Examples of reinforcement learning are Artificial Neural Networks (ANN), Deep Q-Networks (DQNS), and Deep Deterministic Policy Gradients (DDPG).

2.2.2. Statistics

Two main statistical methods are used in data analysis: descriptive statistics, which summarize data from a sample using indexes such as the mean or standard deviation, and inferential statistics, which draw conclusions from data that are subject to random variation (e.g. observational errors, sampling variation).

In statistics, imputation is the process of replacing missing data with substituted values. Imputation requires extensive knowledge of patterns in data. This is Important from a groundwater perspective as these patterns (complex behaviour) may vary from one location to another – as a result of environmental characteristics (Moravcík, 2016). Traditional methods such as regression requires individual approaches for each case. Increase in the amount of data, the process of gap filling becomes more demanding on human resources – eliminate the need for someone to continuously code or analyse data themselves to solve a problem (Cagala, 2017). There is a need for automated tools to help with this process, and Machine Learning is one such methodology.

2.2.3. Visualisation

Data visualization refers to the techniques used to communicate data or information by encoding it as visual objects (e.g. points, lines or bars) contained in graphics – ranging from static printed form to real-time interactive digital four-dimensional representations. To convey ideas effectively, both aesthetic form and functionality need to go hand in hand, providing insights into a rather sparse and complex data set by communicating its key-aspects in a more intuitive way. Data visualisation is an essential component to make insights derived from data analysis understandable and usable to the end user, and may be tailored according to the end user in mind, e.g. a geohydrologist will assess data differently compared to a municipal water manager.

Visualization platforms are an essential element of any Big Data and analytics strategy. The most up-to-date, real-time information and advanced analytics solutions becomes redundant if the results are not communicated clearly and effectively to those who make decisions from it. Data visualization tools provide an easy way to create visual representations of large datasets allowing to detect patterns, trends, and outliers in data. When dealing with datasets that include millions of data points, automating the process of creating a visualization, makes the process significantly more efficient and easier. These data visualizations can then be used for a variety of purposes: dashboards, annual reports, and virtually anywhere information needs to be interpreted immediately.

Processing and ingestion of large datasets can be challenging. In some cases, datasets are so large that it becomes impossible to discern anything valuable from them. This is where data visualisations platforms form a key part of the process. Visualisation platforms are essentially computer products created to visualise (and analyse data) through an interface. Examples of big data visualisation products include Tableau, Hadoop and Spark. There are also domain specific platforms (i.e. SADC-GIP, GGMN, and GEMS) as are the subject of this report.

The objective of this project was to utilise Machine Learning tools. However, the amount data available for our study area was insufficient for pure Machine Learning. As a result the Big Data tools used were statistics, visualization, and data comparisons. This was achieved largely via Python programming language to automate the process.

The subsequent sections provide an overview of each of these visualizations tools and data policies associated with each platform. Readers are advised to consult the relevant documentation provided and links to webpages in text for further details on respective database.

3. DATA AND VISUALISATION PLATFORMS

Data need to be stored in a structured way, in digital formats that can be easily processed to enable efficient and cost-effective access, retrieval and processing for future studies. The choice of database software should be based on the expected amount of data to be stored as well as available human capacity and skills to manage the data (SADC-GMI, IGRAC, IGS, 2019).

Advanced server-based database solutions provide the most robust solutions but come at a high cost. For countries with a limited amount of groundwater data, limited resources and human capacity, well designed spreadsheets can be an adequate and highly cost-effective alternative. If the data are well-structured, migrating the data to more advanced database systems at a later stage when needed, is relatively straightforward. However, no matter the database option, the data providers need to develop a national naming convention for groundwater sites and standardised unique naming of all database parameters. This is especially important for transboundary data sharing (SADC-GMI, IGRAC, IGS, 2019).

For transboundary datasets there is a need for a shared storage location accessible by all, without danger of being made country-specific, or being made redundant. As a result, country data is often shared with neutral bodies such as the International Groundwater Resource Assessment Centre – IGRAC (that works under the auspices of the World Meteorological Organisation – WMO) and the global water quality database GEMStat (hosted, operated, and maintained by the International Centre for Water Resources and Global Change – ICWRGC). Data in such databases can be used for status evaluation, policy making, research purposes or within the scope of education and training initiatives.

To assist in simplified ingestion of existing data to the selected database option, automated data ingestion should be considered. This converts the data provider's data into a format that is suitable for the database upload. This can be done using excel macros (Sterckx et al., 2019) or programming tools (Abrahams et al., 2020).

Visualization platforms are an essential element of any data strategy. The most up-to-date, real-time information and advanced analytics solutions becomes redundant if the results are not communicated clearly and effectively to those who make decisions from it. Data visualization tools provide an easy way to create visual representations of large datasets allowing to detect patterns, trends, and outliers in data. Broadly these should allow for the representation of spatial variability (e.g. maps), temporal variability (e.g. time-series) and statistics (e.g. box-and-whisker plots, Durov plots, etc.).

3.1. Southern African Development Community Groundwater Information Portal (SADC-GIP).

3.1.1. Overview

The proposed Theme 1 primary database changed from the Ramotswa Information Management System (RIMS) which was originally proposed and utilised, to the Southern African Development Community Groundwater Information Portal (SADC-GIP[4]). The SADC-GIP is an initiative of the Southern African Development Community Groundwater Management Institute (SADC-GMI). Notable is that this change did not influence the Theme 1 approach. The required data formats are similar, and the only difference is that the SADC-GIP doesn't support the upload of CSV files, rather a shapefile format which can be done with existing algorithms in Python or R (or via GIS platforms such as QGIS and ARCGIS).

The Limpopo Watercourse Commission (LIMCOM) decided in 2019 to address conjunctive water management in the three transboundary aquifers identified in the Limpopo basin: Tuli-Karoo, Limpopo and Ramotswa. LIMCOM also signed a memorandum of understanding with the Southern African Development Community Groundwater Management Institute (SADC-GMI), by which SADC-GMI will

[4] www.sadc-gip.org

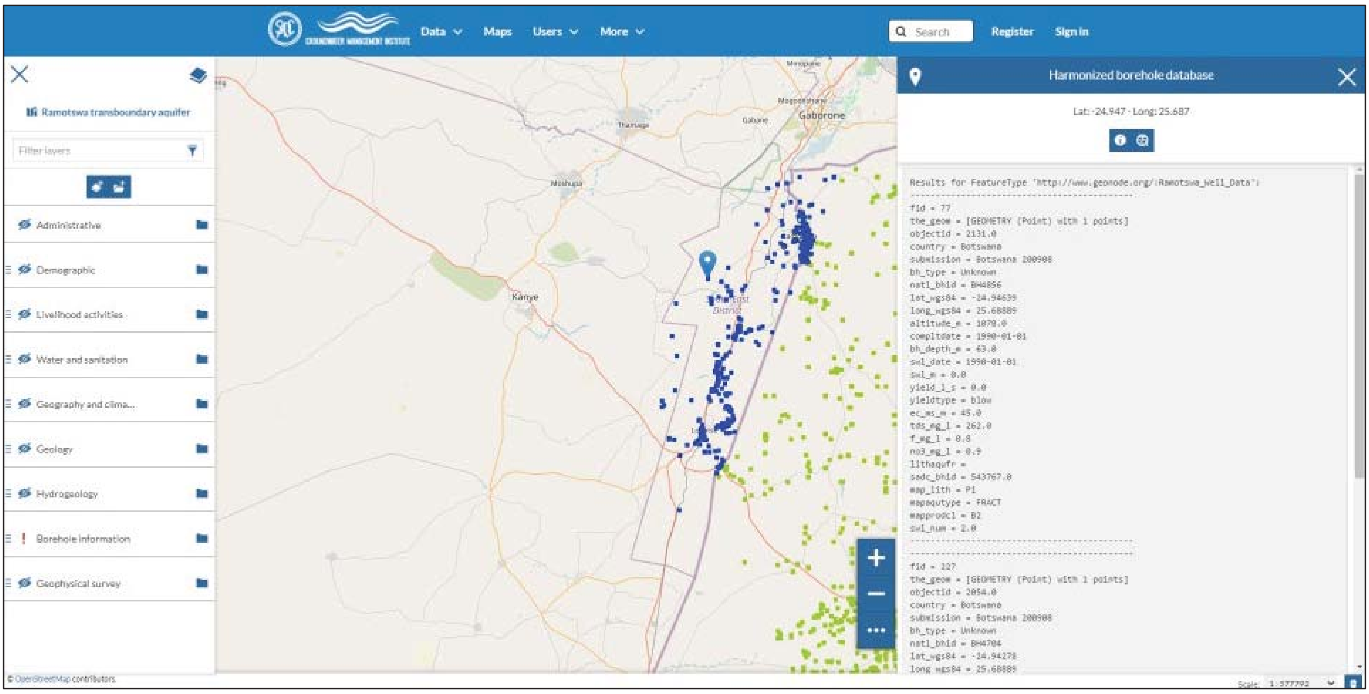
support LIMCOM for groundwater issues, including groundwater data and information sharing[5][6]. Therefore, the decision was made by these role-players to migrate the RIMS data to the SADC-GIP managed by SADC-GMI. RIMS will then be phased out. The plan is that data from the Tuli-Karoo and Limpopo aquifers will also be shared in the SADC-GIP.

SADC-GIP is a web-based spatial data infrastructure platform where multiple users can share data and information on groundwater resources in the SADC region (**Figure 3-1**). By providing easy access to groundwater data and information, SADC-GIP enables all stakeholders to participate actively and in an informed manner in the sustainable management of groundwater resources.

The SADC-GIP was first launched in 2017. It was nested in the Global Groundwater Information System (GGIS) and was managed by SADC in cooperation with IGRAC. A new version of the SADC-GIP was launched in 2020, whose main features and functionalities are summarized in **Figure 3-2**. The SADC-GIP allows users to easily retrieve data in a catalogue and to explore data layers in map viewers. Data stored on external servers can also be retrieved in the SADC-GIP. If allowed by the data provider, data can be downloaded. Anyone can register in the SADC-GIP to access restricted data, upload data, create map viewers and join user groups.

This 2020 version of the SADC-GIP is based on a free and open source application called GeoNode. GeoNode is extensively documented online, including a user guide[7] that covers all features and functionalities of GeoNode, being the main reference for any user who wants to master the SADC-GIP.

The following sections in this chapter are derived from the User Manual of the SADC-GIP (version 1.0, June 2020).



```
Results for FeatureType 'http://www.geonode.org/:Ramotswa_Well_Data':
--------------------------------------------------------------
fid = 77
the_geom = [GEOMETRY (Point) with 1 points]
objectid = 2131.0
country = Botswana
submission = Botswana 200908
bh_type = Unknown
natl_bhid = BH4856
lat_wgs84 = -24.94639
long_wgs84 = 25.68889
altitude_m = 1070.0
compltdate = 1990-01-01
bh_depth_m = 63.0
swl_date = 1990-01-01
swl_m = 0.0
yield_l_s = 0.0
yieldtype = blow
ec_ms_m = 45.0
tds_mg_l = 262.0
f_mg_l = 0.8
no3_mg_l = 0.9
lithaqufr =
sadc_bhid = 543767.0
map_lith = P1
mapaqutype = FRACT
mapprodcl = B2
swl_num = 2.0
--------------------------------------------------------------
--------------------------------------------------------------
fid = 127
the_geom = [GEOMETRY (Point) with 1 points]
objectid = 2054.0
country = Botswana
submission = Botswana 200908
bh_type = Unknown
natl_bhid = BH4704
lat_wgs84 = -24.94278
long_wgs84 = 25.68889
```

Figure 3-1: A screenshot of the SADC-GIP platform web page, showing a map of Botswana boreholes (blue dots) and South Africa boreholes (green dots) the Ramotswa area in the, the various layers available to the left, and selected borehole information to the right.

[5] https://gripp.iwmi.org/2019/04/01/the-limpopo-watercourse-commission-limcom-in-southern-africa-launches-its-first-ever-groundwater-committee/

[6] https://sadc-gmi.org/2019/03/25/limpopo-river-basin-course-commission-limcom-and-sadc-groundwater-management-institute-launch-the-limcom-groundwater-committee/

[7] https://docs.geonode.org/en/master/usage/index.html

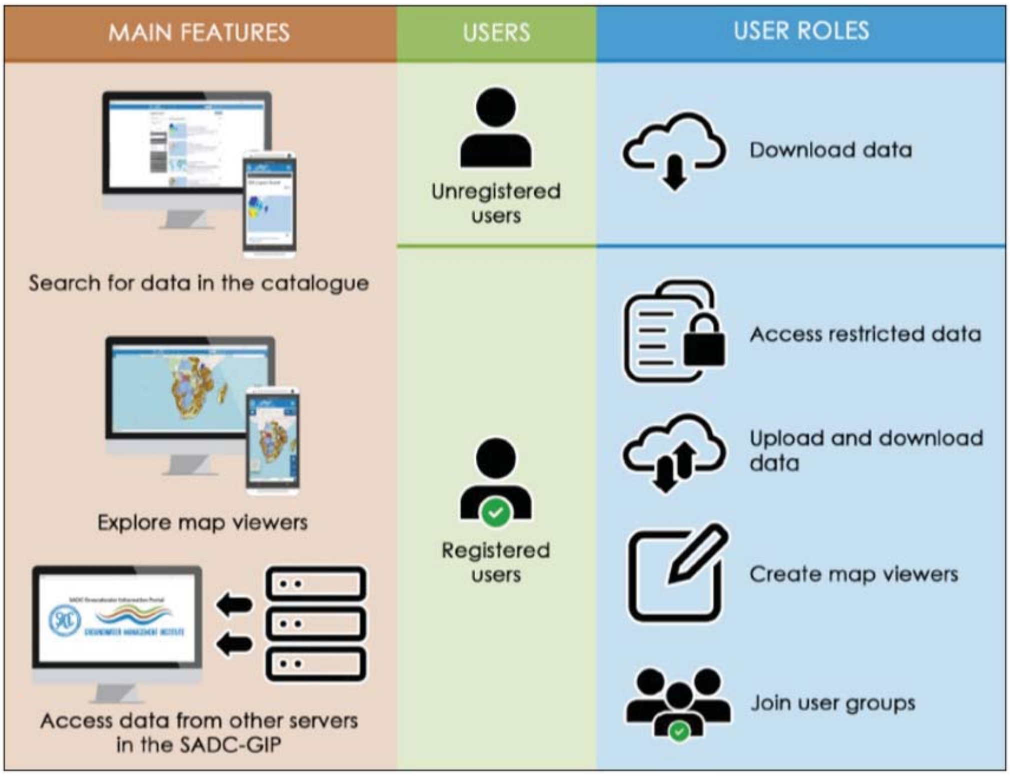


Figure 3-2: Main features and functionalities of the SADC-GIP

3.1.2. Data preparation

3.1.2.1. Data import

Registered users can upload layers and documents in the SADC-GIP. Layers can be uploaded in raster (GeoTiff) and vector (shapefile) formats. For shapefiles, it is necessary that the various extension files are uploaded and have the same name. Documents can be uploaded in of the following file formats:

.doc
.docx
.gif
.jpg
.pdf
.png
.ppt
.pptx
.txt
.xls
.xlsx
.xml
.ods
.odt
.odp
.doc
.sld
.tif
.tiff
.pdf
.gz
.qml
.txt

Data need to be checked before being entered into the SADC-GIP. The process of tracing erroneous data is provided in the *SADC Framework for Groundwater Data Collection and Data Management* (SADC-GMI, IGRAC and IGS, 2019) and summarised in Deliverable 3 Chapter 4 of this project (Umvoto, 2019). An automation process for data quality control and data cleaning is provided in Deliverable 6 of this project.

Quality assurance and quality control procedures include:

- Metadata checks (e.g. borehole ID, coordinates and coordinate system, units, elevation, data/time);
- Logical checks (e.g. data types, consistency, constraints, cross-reference);
- Outlier detection (e.g. comparison with previous data); and,
- Data validation flow (i.e. flagged data or data identified as suspicious).

3.1.2.2. Data export

If allowed by the data provider, Data can be downloaded from the SADC-GIP in a number of ways. Data export is more advanced than in the previous version: data can be downloaded in different formats, it is also possible to download metadata (also in different formats), print and export map views, etc.

3.1.3. Data visualisation

SADC-GIP contains a **data** section. This includes map layers (Layers) and documents (Documents). Layers are Geographical Information System (GIS) vector and raster files, while documents encompass Microsoft Word and PDF documents, images and spreadsheets, among others. The Data section also contains a list of external data servers such as GeoServer Web Map Service, WMS server for WFP GeoNode and GeoNode (Web Map Service) connected to the SADC-GIP (remote services), whose data can be accessed within the SADC-GIP. Catalogues can be searched or filtered on several metadata, for instance on Keywords, Extent or Groups.

If authorized by the data provider, data and metadata can be downloaded. Requests to download data have to be addressed to the data providers (as identified in the metadata).

SADC-GIP contains a **map** section. This section is a catalogue of maps. Maps are sets of data layers combined in a viewer, with a layer tree on the left-hand side, where the legends are available. The order of the layers and their transparency can be modified. As in the Layers and Documents sections, maps can be searched and filtered on several metadata.

The home page of the SADC-GIP also contains a set of links to other relevant databases and websites where to get groundwater related data and information, under Additional Resources.

3.1.3.1. Map interface

Maps can be easily created through the use of the add layers option. Once created, maps can be assigned permissions, just like data, by which authors can specify who else can visualize, download, or edit the map. When creating a map, it is important to ensure consistency between the permissions of the map and the permissions of the layers in the map. For instance, it is not recommended to create a map available to anyone if it contains mostly restricted layers, because many users will not be able to see the layers it contains.

3.1.3.2. Analytics

It is possible to produce some basic statistics in maps, but it is not the focus of the platform, this platform mainly being the display and distribution of spatial/geographical data

3.1.4. Data policy

Anyone can register in the SADC-GIP to upload data, get access to data with restricted access, create maps or join user groups. The registration form can be accessed at the right of the menu bar in the home page. After registering, users have the possibility to create a user profile. It is mandatory to provide at least your name in order to be identifiable by other users and nameless profiles are deleted by the administrators.

Registered users can form groups, e.g. per project, per region, and per thematic area. Groups facilitate the interaction between group members and the sharing of data with restricted access. Data published under a group can also be easily filtered in the data catalog. Groups need to be created or authorized by the administrators. This is to avoid the proliferation of groups and keep the SADC-GIP in order.

Per default, data can be seen and downloaded by anyone. When uploading data, users can change the permissions settings and determine which users or user groups will be allowed to see, download and edit the data. These settings can be updated anytime by the data provider.

After uploading the data, data providers are invited to fill in metadata (under Editing Tools). It is recommended to fill in as many metadata fields as possible, to allow other users to find the data easily and to use it properly. It is mandatory to fill in the following fields:

- Project title.
- Abstract. Explain what the data represent, how they were collected/created, when and by whom. If possible, provide a preferred citation.
- Free-text keywords. If possible, try to use one of the keywords already available in the list provided. If not, create a new keyword that is neither too vague (e.g. “hydrogeology”), nor too specific (e.g. “transient 3D modelling”). Make sure you are not creating duplicates (e.g. “recharge” and “groundwater recharge”, “over-abstraction” and “over-pumping”). It is recommended to not use geographic keywords, as these can be identified under Regions (see below). For data from specific aquifers, it is recommended to create a dedicated group and to specify it under Group (see below).
- Group. Groups can be created for thematic or regional groundwater assessment projects, for instance for transboundary aquifers. To create a group, contact the administrators. It is very easy to filter the data from such projects when the corresponding group is properly indicated in the metadata.
- Project image to use as a thumbnail.
- Preferred language.
- Regions: Identify the countries covered by the data. For SADC-wide data, one can select “SADC region”.

In addition, it is possible to specify restrictions or a license to data. If the appropriate restriction or license is not available in the list, the administrators can be contacted (gip@sadc-gmi.org) to add it. Importantly data size is limited to around 100 MB per a file. For uploading larger files, administrators need to be contacted.

3.2. Global Groundwater Monitoring Network (GGMN)

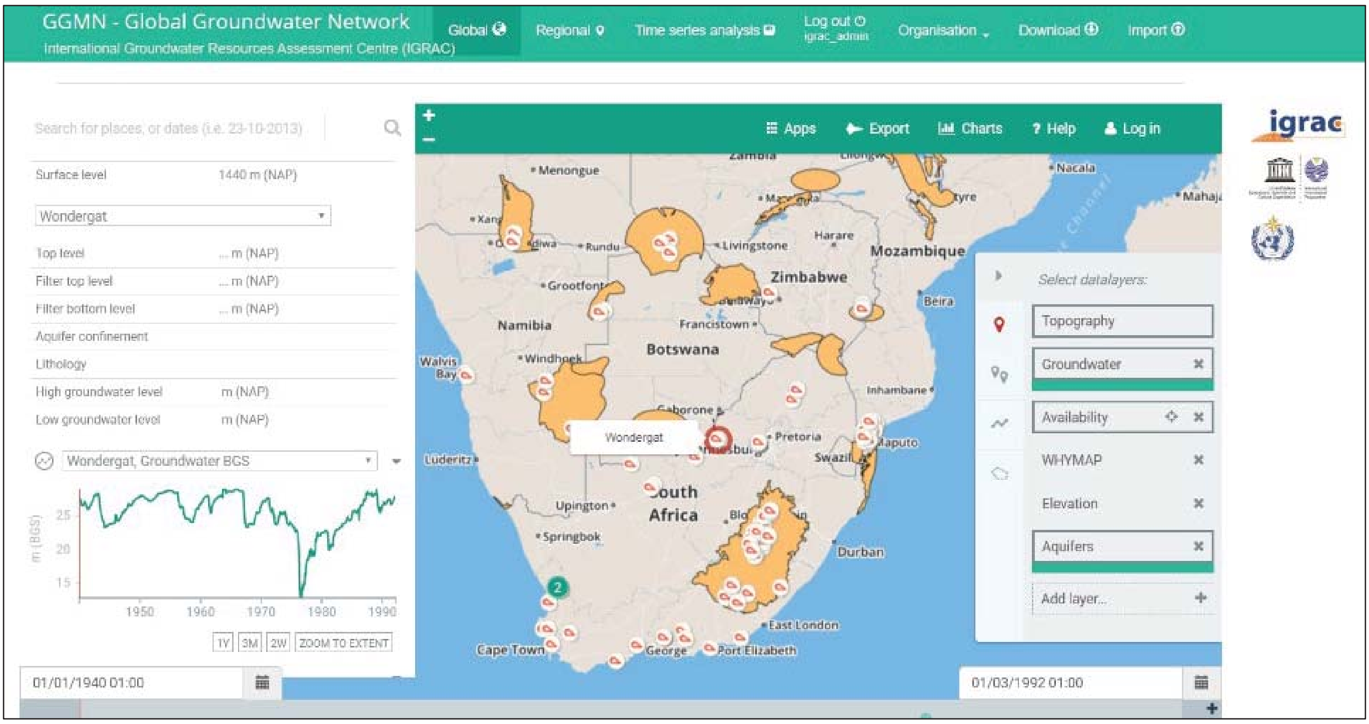
3.2.1. Overview

The Global Groundwater Monitoring Network (GGMN, https://ggmn.un-igrac.org/) is a UNESCO (United Nations Educational, Scientific and Cultural Organization) and WMO (World Meteorological Organization) programme, implemented by IGRAC and supported by various global and regional partners.

GGMN is a participative, spatial web-based (**Figure 3-4**) network of networks, developed to assist the aggregation procedure, as well as the gathering and dissemination of groundwater information. The platform was set up to improve quality and accessibility of groundwater monitoring information and subsequently knowledge on the state of groundwater resources across the globe.

GGMN is a user-friendly tool enabling access to groundwater relevant data and information for stakeholders responsible for the assessment, management and governance of groundwater resources. The main components of GGMN can be seen in **Figure 3-4**.

The GGMN Programme consists of two components: The GGMN Tools (Portal and Mobile App), and the GGMN People Network. The GGMN portal has a protected environment and public view. The protected and public view either allow or restrict access to data depending on the permission granted by the data provider. The platform allows multiple users opportunity to share data and information on groundwater resources. The availability of a common information system facilitates transparency and cooperation between aquifer states and provides a tool for all stakeholders – also non-governmental – involved in the governance of the aquifer.



Surface level 1440 m (NAP)

Wondergat

Top level ... m (NAP)

Filter top level ... m (NAP)

Filter bottom level ... m (NAP)

Aquifer confinement

Lithology

High groundwater level m (NAP)

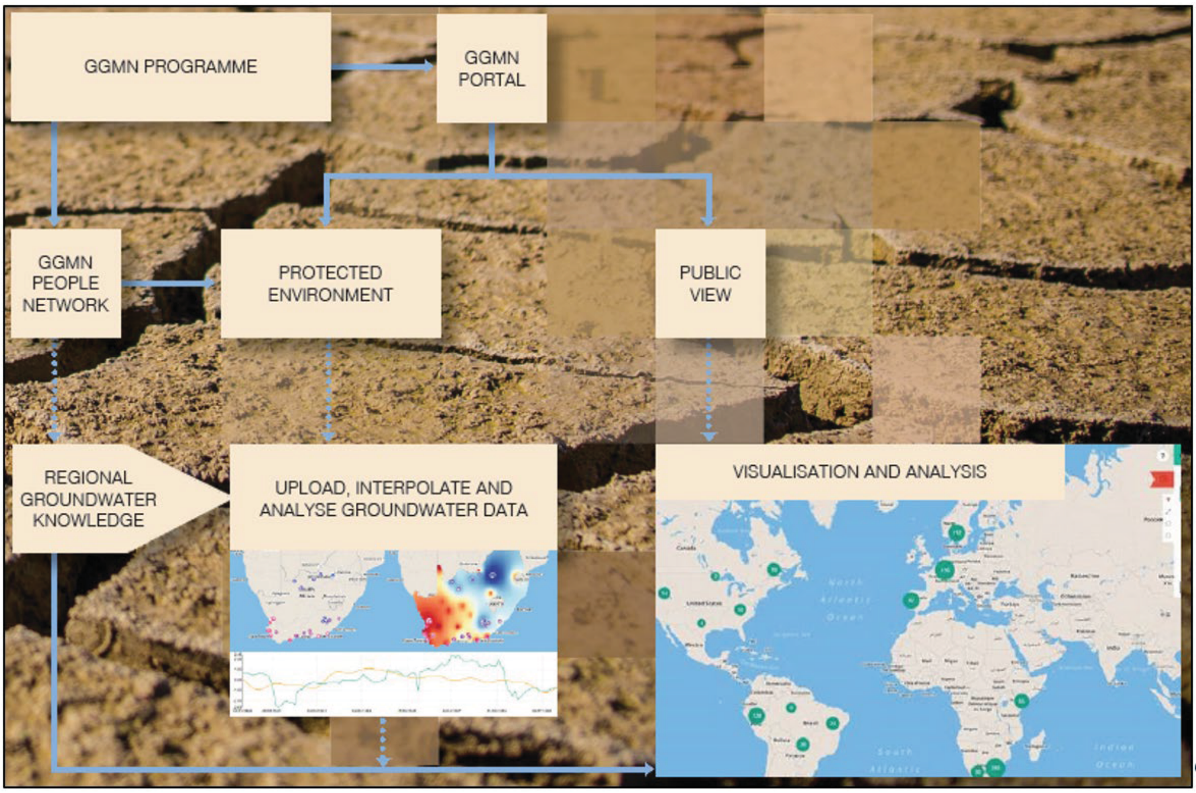
Low groundwater level m (NAP)

Wondergat, Groundwater BGS

01/01/1940 01:00

01/03/1992 01:00

Figure 3-3: Screenshot of the GGMN web page, showing a map of study areas central to the image, data layers to the right, and a time-series visualisation of the Wondergat groundwater sampling location to the left.



GGMN PROGRAMME

GGMN PORTAL

GGMN PEOPLE NETWORK

PROTECTED ENVIRONMENT

PUBLIC VIEW

REGIONAL GROUNDWATER KNOWLEDGE

UPLOAD, INTERPOLATE AND ANALYSE GROUNDWATER DATA

VISUALISATION AND ANALYSIS

Figure 3-4: Main components of GGMN.

3.2.1.1. GGMN Tools: Portal and Mobile App

The GGMN portal assists in the spatial and temporal analysis of monitoring data. Groundwater level data and changes occurring in groundwater levels can be displayed on a global or regional scale.

The GGMN Portal enables the user to upload, interpolate and analyse groundwater data. Additional data layers and information are available to assist in understanding of regional variability of groundwater levels.

Moreover, IGRAC launched in 2018 the GGMN Groundwater Monitoring app. This app enables users to georeference and register groundwater monitoring stations and groundwater level monitoring data, with the option of submitting the collected data to the GGMN Portal. The app can be used in two different ways, either as an independent monitoring tool or in connection with GGMN. It also works offline, and it is available for Android devices via the Google Play Store.

Additionally, the GGMN Portal provides a plugin to work locally on groundwater monitoring data using QGIS to process data offline. QGIS is an open source GIS that contains a variety of functionalities to analyse the data and create spatially interpolated groundwater level maps. This provides the power and accessibility of desktop applications, while still using an online database and online viewer.

3.2.1.2. GGMN People Network

The GGMN relies on the participation of groundwater specialists with knowledge of regional hydrogeology to populate the network. Also, regional (spatial) interpolation of groundwater point measurements is much more than a numerical interpolation and averaging process. It needs to be carried out by regional experts with a clear understanding of local hydrogeological conditions, existing monitoring practices, historic developments, socio-economic changes and other relevant factors. Therefore, establishing a network of regional groundwater specialists is the key task of the GGMN.

3.2.2. Data preparation

3.2.2.1. Data import

Groundwater monitoring data can be imported to the GGMN portal via the import screen. To import data, one requires administrative privileges. The import portal consists of two different components:

1. Import an ESRI Shapefile with the locations of the time series; and,
2. Import a comma-separated values (CSV) file with groundwater levels time series.

The section below describes the structure of the groundwater level time-series file to be uploaded. Time series can be uploaded through a 4-column CSV. Multiple time series can be uploaded in one CSV-file, by adding more rows. The CSV should not contain a header. The first column is the date and time of the measurement, in ISO 8601 format, second column contains the unit ID ("GWmMSL" for groundwater levels above the mean sea level, and "GWmBGS" for groundwater levels below the ground surface), the third column is the value (measured groundwater level). The last column is the ID of the measuring station.

After the station has been uploaded to the portal, one can add the groundwater measurements belonging to that station. The same quality control procedure followed by SADC-GIP (**Section 3.1.2.1**) is applied to GGMN data.

3.2.2.2. Data export

All the data of the GGMN-portal can be downloaded for further use. The data of the selected organization will be downloaded to a CSV-file, if the user has the appropriate rights to do so. In the file the station information and time series data are presented. The Universal User Identification (UUID) column can be used to relate the correct time series with the correct station. Non-registered users can download public time series per station using the button "Export".

3.2.3. Data visualisation

3.2.3.1. Viewer interface

There are three view options on the GGMN platform: Global, Regional and Time Series analysis. The global map provides insight on availability of groundwater monitoring data through space and time and adds extra data layers and information in a broader water-related context. On the map, the data availability and the locations of the measuring stations are visible. A worldwide overview gives direct information on the number of countries involved in the GGMN program.

To obtain an overview, groundwater station locations are displayed by default on the map and correspond to the layer 'groundwater' in the data menu. Green circles are used to denote the 'availability' of data, the number of monitoring stations aggregated in certain areas. Information on the monitoring stations and the corresponding time series are presented by selecting the monitoring station icon on the map. A textbox with the station information is displayed. If the time series are publicly available, the time series will appear.

Other layers include topography, landcover, soil, Digital Elevation Model (DEM), aquifers including the transboundary aquifers of the world map (IGRAC, 2015) and World-wide Hydrogeological Mapping and Assessment Program (WHYMAP) layer.

The regional map gives access to the screen that is focused on analysing more local sets, to display, mean, range and changes in groundwater levels taking a regional perspective. The idea of the Regional viewer is to show at once how groundwater levels are distributed spatially.

Groundwater data are stored either as meters below ground surface or above mean sea level. In the regional portal the user is able to visualize statistics like mean, maximum, minimum, range (max-min) and difference (mean last-first year) for both, in the wells where there are data stored.

3.2.3.2. Analytics: FREQ-Tool (time-series analysis)

The FREQ-tool consists of a time-series analysis tool for groundwater level data followed by analysis of optimal monitoring frequency. Time series analysis module assists in the understanding of the functioning of the groundwater system. The characteristics of a time series can be described by its mean, variance, autocorrelation and stationarity.

- Mean: a measure of the central tendency of a groundwater time series around which the groundwater level fluctuates
- Variance and standard deviation: an indication of the dispersion or the spread of the time series around the mean
- Autocorrelation: a groundwater level at time *t* is dependent on the previous values due to the regulation function of the storage of a groundwater system.
- Stationarity: If the statistical properties of time series do not change with time they are called *stationary*. Groundwater time series are non-stationary, due to the existence of a trend or periodic components

Autoregressive moving average (ARMA) models are used extensively in hydrology for modelling hydrological time series. One assumption to apply these models is that the time series is stationary.

Non-stationary time series (e.g. groundwater time series) should be transformed into stationary series before ARMA models can be applied.

The residuals of groundwater time series obtained by subtracting trend and/or periodic components are usually stationary and the ARMA models can be applied to them. Trends can be caused by excessive pumping, irrigation or climate change, while periodic fluctuations are often the result of seasonally varying rainfall or seasonal varying abstraction.

The time series analysis tool of GGMN is therefore a step-by-step procedure to identify trends, periodic fluctuations and autoregressive model. These components together form the additive model. Based on the time series analysis, an identification of the optimal monitoring frequency can be obtained. Monitoring frequency is one of the key parameters for groundwater monitoring network design.

The GGMN operates according to principles of the WMO and UNESCO with the aim of encouraging the widespread use of hydrological data for national, regional and global studies. Members and other data providers are encouraged to contribute to improve quality and accessibility of groundwater monitoring information, by contributing to the GGMN quality controlled, hydrological data, together with meta-information on the monitoring stations. Nevertheless, data providers can always determine the level of data accessibility by third parties: open, limited or restricted access.

Data in the protected environment are only accessible for authorised users. Ownership of the data uploaded to GGMN remains with the data provider. A user can be authorized for multiple organizations. GGMN Portal can be configured for one organization, for multiple organizations in one country, multiple organizations sharing an aquifer, or it can be used within projects which aim to collect groundwater data on a regional scale from various sources. If the privacy settings are adjusted from public to private (via IGRAC) the time series are visible only after logging in. User account and authorization credentials can be obtained via IGRAC.

Time series are visible for organisations that accepted to make them publicly available. Logging in gives the possibility to see the time series from the organization the user belongs to. The GGMN portal has a public view mode that is meant for the general public, including researchers, consultants, teachers, policy makers and NGOs. From the publicly available data, changes in groundwater level point measurements can be visualised over time on a regional and a global scale.

3.2.4. Data policy

In the case of GGMN, IGRAC is in charge of managing the requests to access the private workspaces. In order to be able to do so, a Licence Agreement should be signed between the Departments and IGRAC before storing private data in GGMN. In this agreement, the Departments (data providers) determine the level of data accessibility by third parties, as follows:

- Open access: Groundwater level data and metadata on monitoring stations are publicly available in GGMN
- Limited access: Only metadata on monitoring stations are visible in GGMN. If information on groundwater level data is requested by third parties, IGRAC will proceed according to the Principles of dissemination of data.
- Restricted access: Groundwater level data and metadata on monitoring stations in GGMN are available only to the data provider. If information on groundwater level data is requested by third parties, IGRAC will proceed according to the Principles of dissemination of data.

The principles of dissemination of data are:

1) Data submitted to GGMN and stated as ‘open access’ are considered public domain and can be accessed via the GGMN portal by the public.
2) Data submitted to GGMN and stated ‘limited access’ can be disseminated for non-commercial use only upon an official request to IGRAC at no cost (after the party signs the User Declaration).
3) Data submitted to GGMN and stated "restricted access" will not be disseminated; however, the name of the data provider will be sent to eligible institutions upon request.

3.3. Global Environment Monitoring System for Freshwater (GEMStat).

3.3.1. Overview

The UNEP Global Environment Monitoring System for Freshwater (GEMS/Water[8]) was established to collect worldwide water quality data from a network of government nominated individuals and organizations to support scientific assessments and decision-making processes in global inland water quality. The water quality data of ground and surface waters is collected within the global water quality database and information system UNEP GEMStat[9] that provides a global overview of the condition of water bodies and the trends at global, regional and local levels (**Figure 3-5**).

The GEMS/Water Data Centre (GWDC) within the International Centre for Water Resources and Global Change (ICWRGC[10], and in cooperation with the German Federal Institute of Hydrology (BfG[11]) in Koblenz, Germany hosts and maintains GEMStat as a German contribution to GEMS/Water. It coordinates the data-related activities focusing on collecting and quality-assuring the water quality monitoring data provided by the GEMS/Water Global Monitoring Network.

At present, the growing database contains more than 7 million entries for rivers, lakes, reservoirs, wetlands and groundwater systems from 75 countries and approximately 5700 stations. Overall, data is available for the time period from 1965 to 2019 and about 300 parameters.

UNEP GEMS/Water as well as the World Water Quality Alliance working towards a global water quality assessment (mandated by Resolution UNEP/EA3/10 from 2017) are collaborating on Theme 1. Data generated/discovered/harmonized in the run of the project will be fed into UNEP GEMStat.

The data that is collected is mostly provided by governmental partners in water agencies or other organizations in charge of water quality monitoring or environmental data management. In some cases there are collaborating partners that have collected water quality data within research projects and want to share it more widely. Countries and organizations provide water quality data voluntarily from their own monitoring networks.

The water quality data available in GEMStat is used for status evaluation, policy-making, and research purposes or within the scope of education and training initiatives.

At its core is the commercial water data management system WISKI 7 that is used to manage both the water quality monitoring data collected by the GWDC as well as the discharge data collected by the Global Runoff Data Centre (GRDC). The system consists of a backend with an Oracle RDBMS for data storage and a Java-based middleware for time series and water quality sample data processing and provisioning through Representational State Transfer Application Programming Interface (REST API) allowing for a wide range of data quality checks and analysis calculations that are used to assure the quality of the data and develop derived products such as reports, statistics and maps. The frontend web portal allows end users to visualize and/or download the data for their own use, as seen in **Figure 3-6**.

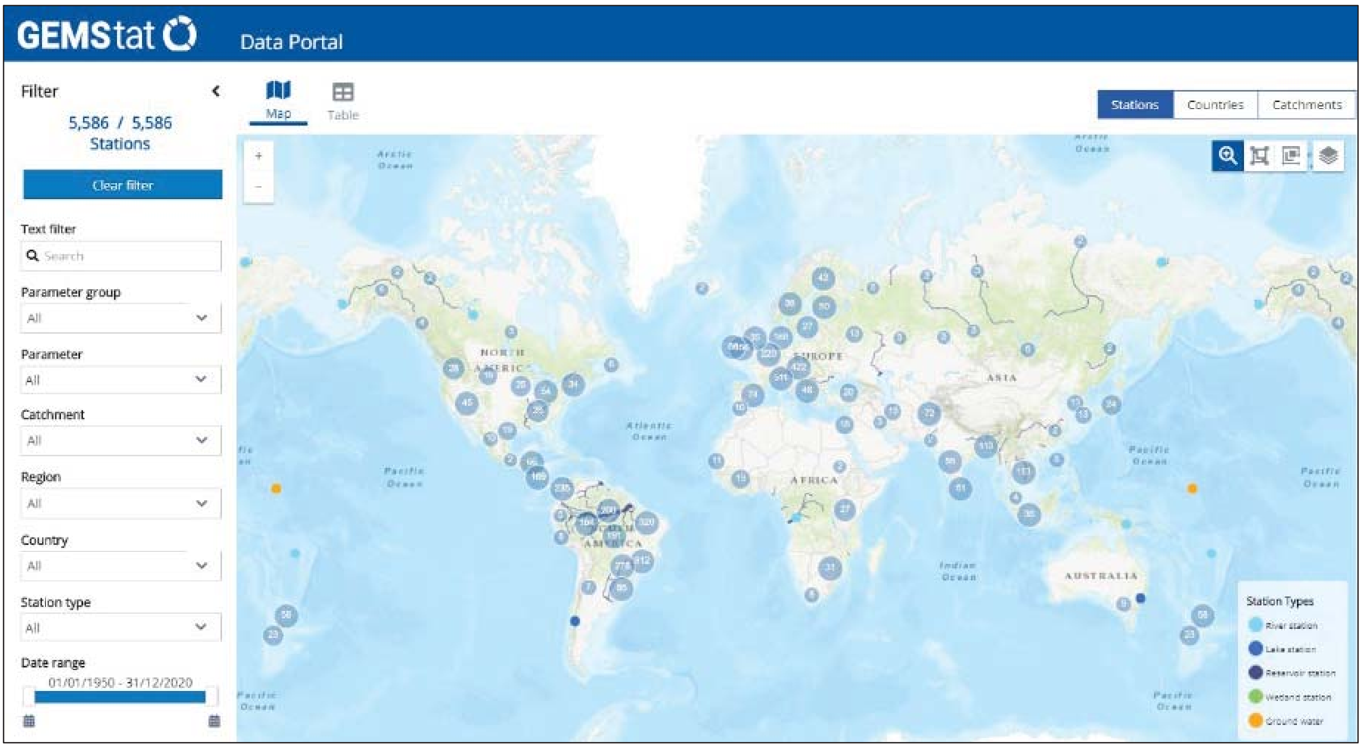
Additionally, the spatial data infrastructure of the Federal Institute of Hydrology is used to provide interactive maps and dashboards providing access to different aspect of the data.

[8] https://www.unenvironment.org/explore-topics/water/what-we-do/monitoring-water-quality

[9] (https://gemstat.org

[10] https://www.waterandchange.org/en/

[11] https://www.bafg.de/



Filter

5,586 / 5,586 Stations

Clear filter

Text filter: Search

Parameter group: All

Parameter: All

Catchment: All

Region: All

Country: All

Station type: All

Date range: 01/01/1950 - 31/12/2020

Figure 3-5: Screenshot of the GEMStat web page, showing where data is available.

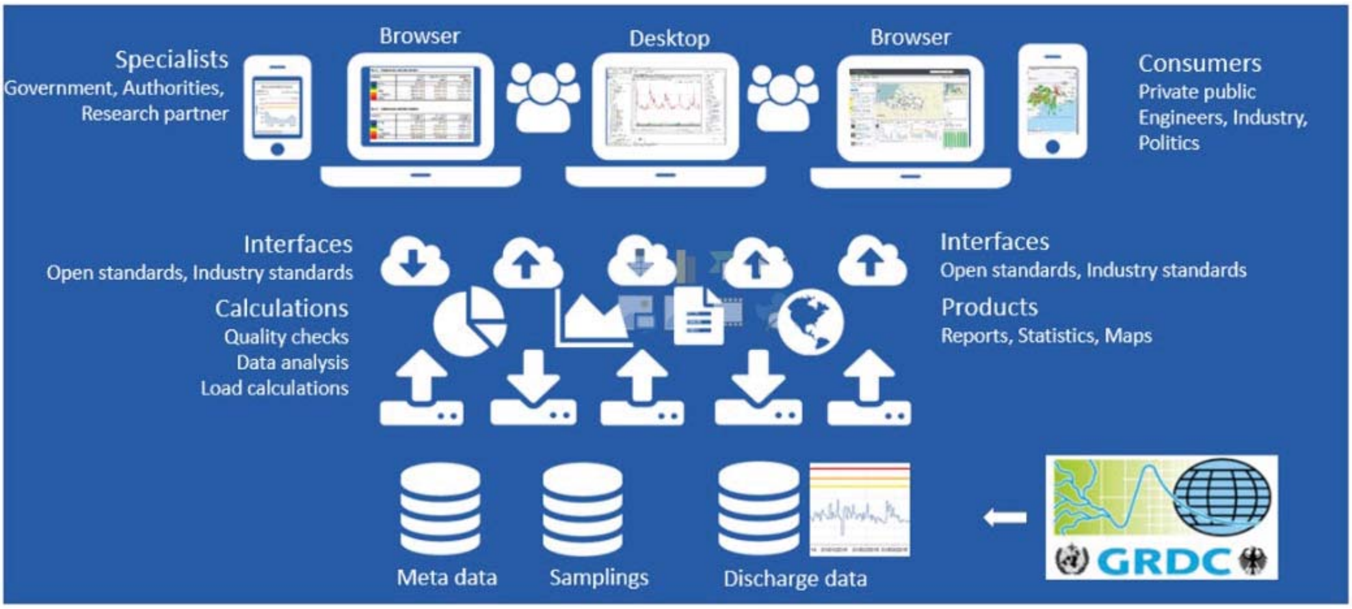


Figure 3-6: GEMStat architecture.

GEMStat contains more than 7 million entries from approximately 5700 stations in 75 participating countries. The greatest coverage of stations is currently in Latin America and the Caribbean. The highest number of sample values is currently available from Europe. Currently, the largest number of the data by far comes from river stations, followed by data from lakes and groundwater.

Overall, data is available from 1965 to 2019, covering a total time frame of 54 years. The longest time series are currently available from Europe, North America and Asia and the Pacific.

In total, about 300 parameters are available, which are classified into a hierarchical system of groups and subgroups. Currently, most GEMStat data falls within the category of chemical parameters, followed by physical parameters, while biological parameters represent only a minor proportion. The largest proportion is thereby contributed by inorganic compounds and nutrients.

3.3.2. Data preparation

3.3.2.1. Data import

The current method of reporting water quality data to GEMStat is based on e-mail communication. Data providers are requested to use the excel templates provided on the website to report stations, analysis methods and water quality values[12]. In GEMStat various quality control checks are performed, e.g.:

- Major ion balance (the sum of cations should be equal to the sum of anions in a solution).
- Comparison of electrical conductivity and total dissolved solids (being directly proportional).
- Calculated versus measured total dissolved solids (calculated TDS should be equal or less than measured TDS, since some ions are often not included in the analysis).
- Anomalous results (e.g. If nitrate is present in the absence of dissolved oxygen, the value for one or the other is likely to be incorrect, since nitrate is rapidly reduced in the absence of oxygen).
- Scientifically proven (e.g. total iron must be greater than dissolved iron).

3.3.2.2. Data export

To download the monitoring data or aggregated statistics, the user provides personal details, the purpose of the download and consent to the data policy. After submitting the download request, the GEMStat system extracts the data from the database into Excel spreadsheets and sends a link to the data download to the email address of the user[13]. Currently, the download is limited to a maximum of 500 stations. For larger datasets, such as global data a custom request is required.

3.3.3. Data visualisation

3.3.3.1. Viewer interface

The GEMStat website provides access to all data and derived. The main entry point for exploring the available data and visualizing it is the data portal[14]. Furthermore, the GWDC is developing interactive web maps to visualize the quality status of water resources at different spatial scales using the ESRI ArcGIS Spatial Data Infrastructure maintained by the BfG[15].

The GEMStat data portal allows users to search for available data and visualize or download the data. Data can be filtered by parameters, catchments, regions, country and stations types as well as temporally. The system provides access to the actual monitoring data at station level, as well as aggregated to catchments or countries.

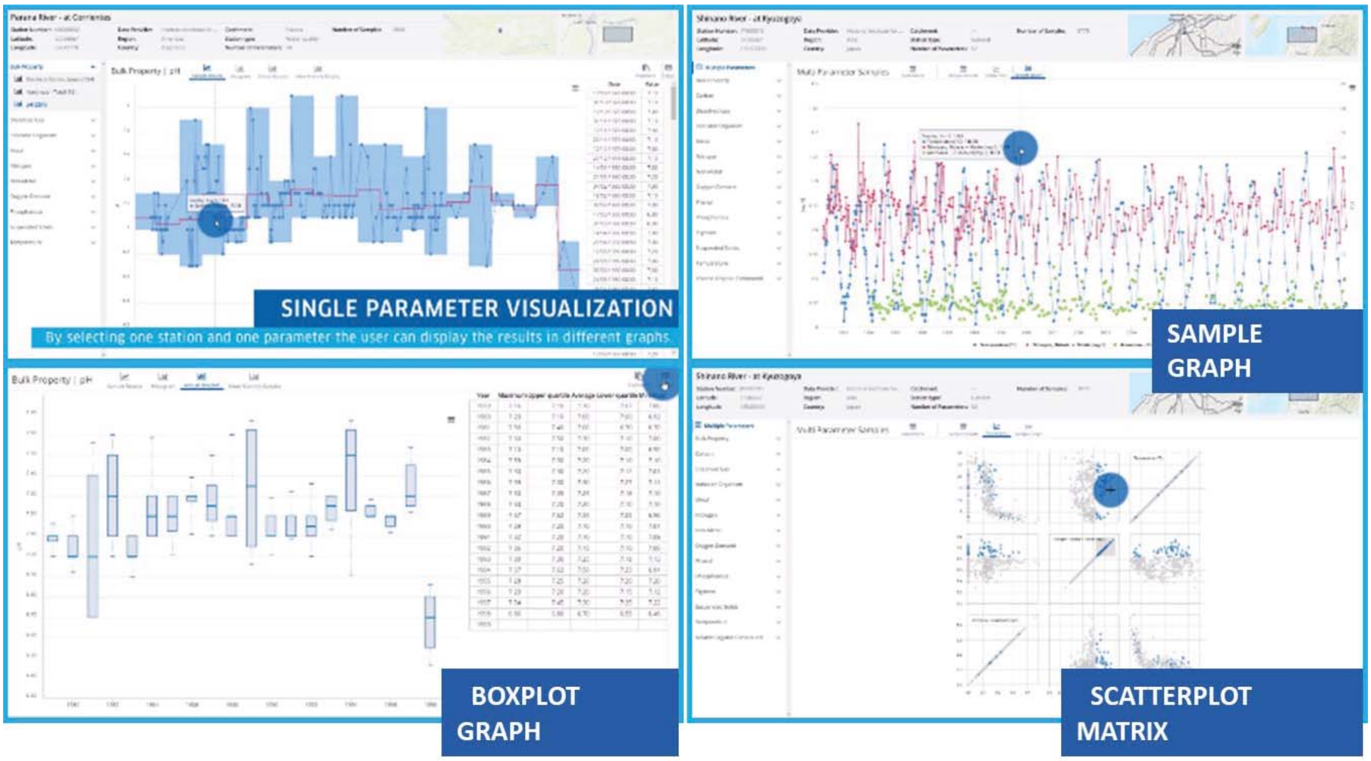
The data portal offers different types of data visualizations to explore the monitoring data. Users can either chose a single station and parameter and plot the time series, histogram of data distribution or boxplots at yearly of mean monthly intervals. Alternatively, users can select different parameters to compare them to each other through sample graphs or scatter plots (**Figure 3-7**).

[12] https://gemstat.org/data/data-submission/

[13] https://gemstat.org/data/data-portal/

[14] https://portal.gemstat.org

[15] https://gemstat.org/data/maps



By selecting one station and one parameter the user can display the results in different graphs.

Figure 3-7: Example of data visualisation options within GEMStat

3.3.3.2. Analytics

The Statistics Portal provides a convenient way of viewing aggregated statistics of water quality data available in GEMStat. The user can select water quality parameters of interest, a desired level for spatial aggregation, as well as the individual spatial element for the spatial aggregation. Additionally, the user can include or exclude types of water bodies.

In GEMStat, the statistical data can be summarized visually using various graphs such as boxplots, scatter plots, sample graph and single parameter visualization

3.3.4. Data policy

To get data from the Data portal, stations or aggregated data at country or catchment level have to be selected, and to start with the download process, a contact form has to be filled, providing contact information and other details. Then, a download link for the requested data will be sent to the email address provided. A maximum of 500 stations can be downloaded in this way. To request larger datasets, a Custom Request has to be sent through the website (https://gemstat.org/custom-data-request/).

In order to get datasets from GEMStat that are not available on-line through the Data Portal, a request made by non-commercial and non-private persons will be filled at no cost, depending on data availability, GEMS/Water Programme workload and the data set requested.

Normally, requests will be queued with priority given to those related to joint activities with UN Environment Programme area and other UN agencies.

When data providers share their data with GEMStat, they remain the owner of the data and define how the data is being used and re-shared. The GEMS/Water data policy has three different levels of sharing:

- **Open**: Data is publicly available
- **Limited**: Data is shared on request for non-commercial research
- **Restricted**: Data is not shared but used for UN assessments and data products

Datasets defined as "Restricted" will not be distributed, but information about the data sets (including the Data Provider's name and summary statistics on a country level) could be sent to eligible institutions upon request. Users may not transfer, sublicense, rent, lease or sell data obtained from GEMStat. Users are required to submit a signed user declaration in order to obtain the data.

3.4. Comparison between SADC-GIP, GGMN and GEMStat Storage and Visualisation Platforms.

A comparison of the main features of each platform is summarized in **Table 3-1**. It is evident from the table that each platform has its own strengths and weaknesses.

Table 3-1: Comparison of the main features across the three storage and visualisation platforms.

Group	Feature	SADC-GIP	GGMN	GEMStat
Data	Water quality	Yes	No	Yes
	Water levels	Yes	Yes	No
	Other data	Yes	No	No
Privacy and sharing	Restrictions	Yes	Yes	Yes
	Public	Yes	Yes	Yes
	Data sharing	Yes	Yes	Yes
	Upload and Download	Yes	Yes	Only download
Content	Temporal	No *)	Yes	Yes
	Spatial	Yes	Yes	Yes
	Metadata	Yes	Yes	Yes
Application	Desktop	Yes	Yes	Yes
	Mobile	Yes	Yes	Yes
Analytics	Statistical	No	Yes	Yes
	Visualisation	Yes	Yes	Yes
	Quality control	Manual checks	No	Yes

* Currently under development

The three platforms assessed have a few things in common:

- The user-friendly interface and ease of use. These platforms were designed to ensure that the interface contain elements that are easy to access, understand, and use to facilitate those actions. Broadly, the interface elements include but are not limited to: input controls (e.g. buttons, text fields, checkboxes, dropdown lists, list boxes, toggles, and date field), navigation components (e.g. slider, search field, icons) and informational components (e.g. tooltips, icons, progress bar, notifications). Users therefore require only basic level of expertise to operate the systems.
- Functionality for data sharing, which is suitable for individuals/groups working within the same system (i.e. a transboundary aquifer). Although each of these platforms promote data sharing, it has advanced security features which allows different levels of access to data to be set according to the needs of individual users. Any imports or exports to date are largely still a manual process.

SADC-GIP is hosted online and can handle multiple sets of data in a single visualization platform including groundwater levels (currently under development), groundwater quality, discharge estimates and lithological data. This is a significant advantage over GGMN and GEMStat as SADC-GIP accounts for additional groundwater related data. The platform has a range of import options (formats) available, ranging from CSV files to shapefiles (vector) and GeoTiff (raster). Similar to GGMN, data providers are responsible for the quality of the data provided. Administrators of the SADC-GMI check if data uploads meet the requirements in terms of data quality and metadata completeness. The data stored on SADC-

GIP can be visualized spatially and exported in a variety of formats (e.g. PDF, PNG, and JPEG). Although the SADC-GIP is not primarily designed for storing and analysing time series, a time series data functionality is in development that would allow visualizing monitoring data in x-y charts. Additionally, it allows users the opportunity to do some basic statistics but this is not the purpose of the platform.

GGMN was developed specifically for the storage, visualization, analysis and sharing of groundwater level time series. GGMN has two options available, the Portal and the mobile app., which allows groundwater managers access to information using their smartphones. Embedded within its spatial and temporal mapping capabilities, managers have access to statistical reports to assess trends and patterns. The ability to have access to this information is important in terms of decision making and managing the system. For example, to improve management of an aquifer requires assessing trends and patterns. Access to this information provides the ability to not only assess but improve decisions making. GGMN features a drag-and-drop visualization tool that allows the creation of visualizations of data for infographics using simple line and trend graphs. GGMN has the option to export and process data using other platforms such as QGIS. This means that point measurements can be combined with proxy information and personal expertise to create groundwater level maps. The platform gives a global and regional overview of groundwater level data, which can be shared. For shared systems such as transboundary systems, this is important as it allows for the holistic management of resources between member states using a standardized platform. However, options are available to limit or restrict access to data via the privacy settings before import is made. If data is made accessible, the data can be viewed in the public viewer. The data provider is responsible for the quality of the dataset being uploaded.

GEMStat was developed specifically for the storage, visualisation and analysis of global surface water and groundwater quality data. There are multiple import options available ranging from excel files to databases. The platform has spatial and temporal mapping capabilities for water quality data (both surface and groundwater). Output options include multiple chart formats as well as data formats for sample and aggregated data (CSV, XLSX, XML, and JSON). GEMStat has two main features embedded in the system. Firstly, the system incorporates an automated quality control of the imported data. The quality control tool is based on Analytical Methods for Environmental Water Quality and state of the art checks to ensure the provision of quality-controlled datasets. Secondly, the option of exporting advance statistical analysis for assessment and reporting. Due to the mandate of UNEP GEMS/Water covering freshwater quality, the GEMStat database only contains water quality monitoring data and lacks additional groundwater related data although the underlying system is capable of handling many types of environmental point and gridded monitoring data. The quality of groundwater is influenced by a range of environmental factors. For example, groundwater chemistry of aquifers is influenced by water-rock interaction. Inclusion of groundwater related data is necessary to improve the analysis of the system in question.

3.5. Other Data Storage Options

3.5.1. Excel

A simple solution would be to store the data in Excel files. The database could be structured in several datasheets:

1. spreadsheet containing data points
2. spreadsheet containing stratigraphy records
3. spreadsheet containing well designs
4. spreadsheet containing groundwater level monitoring data
5. spreadsheet containing groundwater quality monitoring data
6. spreadsheet containing groundwater abstraction monitoring data

In Excel 10, each spreadsheet is limited to 1,048,576 rows and 16,384 columns of data. It is not likely that the number of data points will once exceed 1 million. Estimates in how many years of monitoring the maximum capacity of the spreadsheets will be reached, depending on the number of monitoring stations and the frequency of monitoring are shown in **Table 3-2**. As shown in this table, there is a risk to reach the limit only if automatic data loggers are installed and measurements are taken every hour or more.

Table 3-2: Estimated number of years for monitoring data to reach data limit in Excel spreadsheets.

		Frequency (once every)				
		year	**month**	**day**	**hour**	**minute**
Number of monitoring stations	**10**	10.4858	8.738	287	12	0
	100	10.486	874	29	1	0
	200	5.243	437	14	1	0
	500	2.097	175	6	0	0
	1000	1.049	87	3	0	0

Excel files should be stored on a server or in the cloud to allow both departments to access it. The best option would be to use one server in any of the two departments, if any. That server should provide enough security. Otherwise, a server provider might be contracted. Alternatively, cloud services can be used, whose costs usually increase with the storage capacity (e.g. Dropbox, Google Drive).

Implementing an Excel transboundary database would be an easy and cheap solution. Main disadvantages are the limited possibilities to implement QC checks and the lack of indexing of the data. It would be a satisfying solution as long as the number of monitoring data remains steady.

3.5.2. Server-based relational database

3.5.2.1. Introduction

Server-based relational databases are more advanced than the other solutions, as shown in **Table 3-3**. For instance, they have a much higher storage limit, and allow storing plenty of parameters and connections between parameters. QC checks can be implemented and changes in the database can be tracked. Multiple users can access the database simultaneously. Also, relational databases can be enhanced with data visualization and analysis tools.

A significant benefit of a server-based relational database would be to establish an automatic update of the transboundary database via APIs[16]. This would ease the work in terms of data upload and would allow data users to always use up-to-date data. However, setting up an API would be possible only if the departments' databases support this technology.

The downside of server-based relational databases is of course the cost for development, server and maintenance. It also requires appropriate training of the staff in the departments to use it. The departments should also agree on where to store the server, i.e. in which country.

A server-based relational database represents a considerable investment, maybe not needed for the Ramotswa transboundary aquifer. In the current situation of data collection, data exchange could be handled efficiently with the existing solutions described above. Investing in a server-based relational database would be profitable if the departments install more monitoring boreholes equipped with data loggers, or if groundwater data exchange is upscaled to the Limpopo River Basin and managed by LIMCOM.

[16] An Application Programming Interface (API) is a subroutine allowing computer applications to communicate with each other, e.g. one database application can extract data from another database.

Table 3-3: Overview of database functionalities and organisational requirements for different database options (from SADC-wide Framework).

	RIMS	GGMN	GemStat	Excel	Relational database
Current availability	Yes	Yes	Yes	No	No
Already in use	Yes	No	No	No	No
Data	Maps (geospatial data) and documents	Groundwater level monitoring data	Groundwater quality monitoring data	All	All
Database model	Flat	Flat	Flat	Flat	Relational
Query/filtering functionality	Basic	Basic	?	Basic	Advanced
Indexing of data	No	?	?	No	Yes
Logical checks on data entry	No	No	?	Possible	Advanced
Data quality control process	Basic	?	?	Basic	Advanced
Backup	Yes	Yes	?	Needs to be implemented if server-based	Needs to be implemented
Automatic update	No	Possible	?	?	Possible
Users interface	Yes	Yes	Yes	No	Yes (custom-made)
Users roles	Advanced	Basic	?	Basic	Advanced
Additional tools (processing, analysis, visualisation, reporting)	Advanced	Advanced	Advanced	No	Advanced (custom made)
Cost	None	None	None	Average cost: server/cloud & possibly maintenance	Important cost: Server, Development of the software & Maintenance
Human capacity requirement, need for training	No (already done)	To be provided in this project	To be provided in this project	Yes (server/cloud solutions, Excel)	Yes (server solutions, database software)

3.5.3. Example: Groundwater Markup Language

Groundwater related data is one example of data with a complex structure: groundwater is stored in aquifers, which are hydrogeological units. Aquifers have different properties, as permeability and conductivity. There are also certain elements that interact with aquifers, for example, a monitoring well measuring the groundwater level. The groundwater level is measured in time, and in one monitoring station the level will be measured several times. There could be other devices measuring hydrogeological properties, such as flow rate and chemical composition. Moreover, there could be other types of (groundwater) elements interacting with the aquifer and monitoring well, or connected to them somehow, such as a borehole log, injection wells, springs, etc. Data and metadata from all these sources need to be stored, considering how they relate to each other

Several tables would be needed to represent the complex scenario described above, for instance:

- One table listing all data points (wells, boreholes, springs), including coordinates, hydrogeological units, status, and other characteristics (metadata)

- One table listing all hydrogeological units and associated characteristics (permeability, conductivity, chemical composition, etc.)
- One table per data point of one kind with time series associated (for example, one table with time series of all groundwater level monitoring wells, another table with time series of flow rate of all springs, and so on)

This way of organising data is acceptable when it is the first attempt to do so, and when existing databases are being used for this purpose. This is exactly what is suggested to do as a first step for this project (to use RIMS for map-type data, GGMN for time series of groundwater levels, and GEMStat for groundwater quality, all of them based on tabular data entries). However, when the plan is to create a database that will be used to store and exchange all kind of data and that is future proof, another type of structure needs to be used.

There is a type of data format called XML (Extensible Markup Language). XML, just as CSV for instance, allows data to be stored and transferred, but also, is useful to handle data types with a complex structure, such as groundwater data. There are different types of XML files whose structure depends on the kind of information that needs to be described.

The Groundwater Markup Language (or GWML2), a type of data format based on XML, is a data standard specifically created for storage and transfer of complex groundwater data. Technically speaking, it includes the groundwater part of WaterML2 (a type of XML structure designed for water applications) and consist of data structures and encoding guidelines for groundwater data (Brodaric et al., 2018a). Groundwater data conforming to GWML2 are encoded in GML-conformant XML documents, where GML stands for Geography Markup Language (type of XML to express geographical features).

GWML2 is designed in a way that covers all hydrogeological units, units, fluid bodies, voids, fluid flow, water wells and associated elements. The GWML proposes a comprehensive schema to describe groundwater data and metadata, as well as relationships between data. **Figure 3-8**, extracted from Brodaric et al. (2018b), shows a simplified version of the conceptual schema of GWML2, using UML (unified modelling language).

In a transboundary scenario, where two countries use a database designed following GWML2, countries are able to send and receive data in a seamless way via OGC[17] standards as WMS, WFS, SOS and WPS. The transfer is seamless as all metadata fields are correspondent, or in simple words, in each database there is a "slot" for each type of data with its correspondent metadata and relations associated.

One important disadvantage of GWML2 is its complexity. For example, the content of a GWML2 file cannot be easily understood via Notepad, as it could be a CSV file. A data specialist with knowledge in HTML and Java Script is needed to construct the database in GWML2, which will be accessed later by groundwater specialists either via browser or specialised software. One way of opening such file is by using HTML, the language used to display elements in a web browser.

Several organisations have tested the implementation of GWML2, e.g. the Geological Survey of Canada (GSC), United States Geological Survey (USGS), Bureau of Meteorology (BOM, Australia), among others (Brodaric et al., 2018 b). These organisations have been able to exchange groundwater information from different data providers and incorporate GWML2 into their operational data delivery mechanisms (Brodaric et al., 2018 a).

The GWML could be used in the future by the departments to develop dedicated relational databases for storing groundwater data. When stored in the same GWML format, groundwater data could be easily exchanged between the departments using OGC services.

[17] Created in 1994, the Open Geospatial Consortium (OGC) regroups more than 500 commercial, governmental, non-profit and research organizations worldwide, aiming at developing and implementing open standards supporting the interoperability of geospatial data and geospatial applications. The OGC proposes formats and protocols for sharing geospatial data that are widely used in Spatial Data Infrastructures (SDI). The exchange of maps in the RIMS reposes on such services.

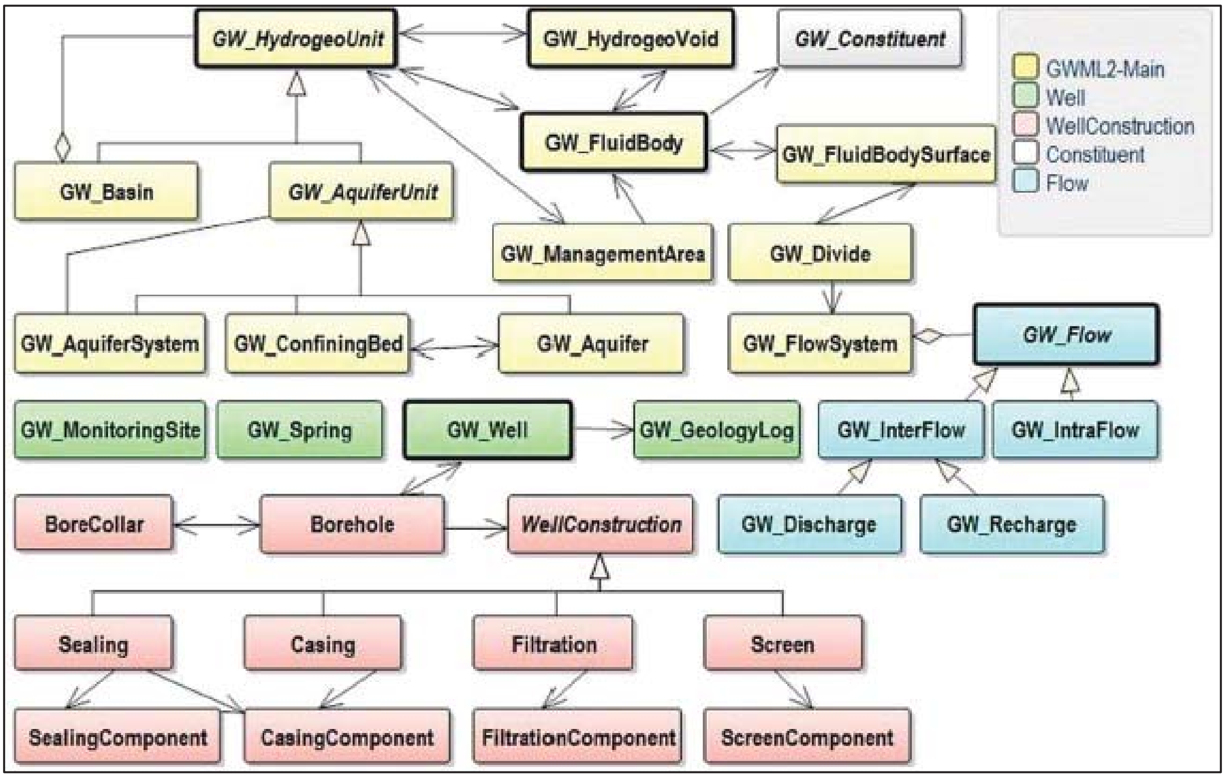


Figure 3-8: Simplified UML representation of the GWML 2 conceptual schema (from Brodaric et al., 2018 b)

The frequency of updates of the transboundary database is summarised in **Figure 3-9**. A yearly update of the transboundary database should be a minimum.

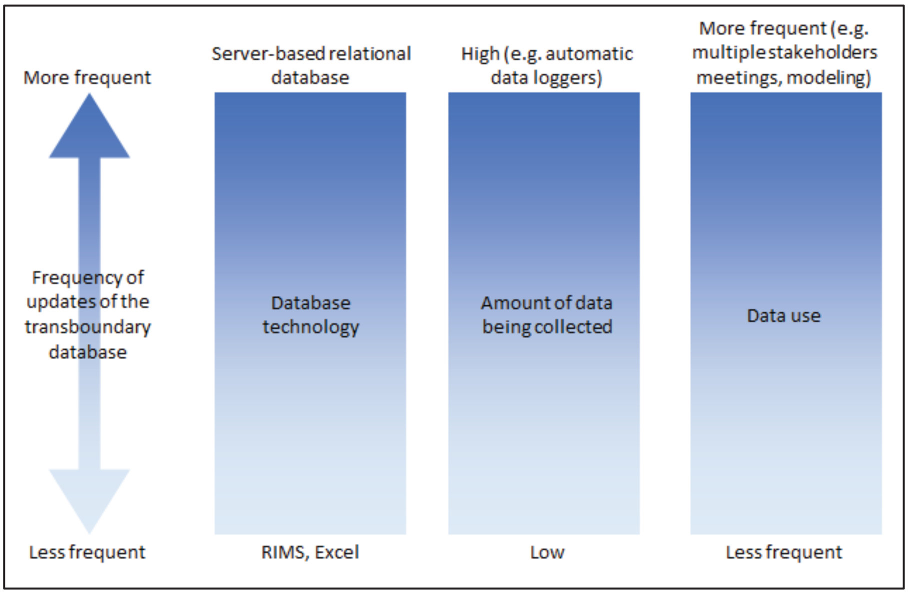

More frequent
Frequency of updates of the transboundary database
Less frequent
Server-based relational database
Database technology
RIMS, Excel
High (e.g. automatic data loggers)
Amount of data being collected
Low
More frequent (e.g. multiple stakeholders meetings, modeling)
Data use
Less frequent

Figure 3-9: Factors to consider when determining the frequency of updates of the transboundary database.

4. DATA PROCESSING PROCEDURE

To make data valuable and accessible requires techniques to gather (extract) data from a number of sources, organize (transform) the data in a uniform way and centralize the data into a single repository (load). To achieve this, the project leverages off two concepts namely, the Extraction, Transformation and Loading (ETL) of data and Exploratory Data Analysis (EDA) during the process of data management.

4.1. Tools

4.1.1. Extract, Transform and Load (ETL)

ETL is a type of data integration process referring to three distinct, but interrelated steps to ensure data usability.

- **Extract**: the initial step of the process is to collect raw data from an array of sources.
- **Transform**: Gathered raw data, the raw data undergoes transformation by applying specific rules and regulations to achieve standardization, deduplication, verification, sorted, or any other custom tasks required.
- **Load**: The transformed data is loaded in storage facility by executing the task via the command line or GUI interface. The final product is an ETL pipeline used to synthesize data from multiple sources that is then used for reporting and analytics to assist decision making. A schematic of the ETL pipeline or process can be seen in Figure 4-1.

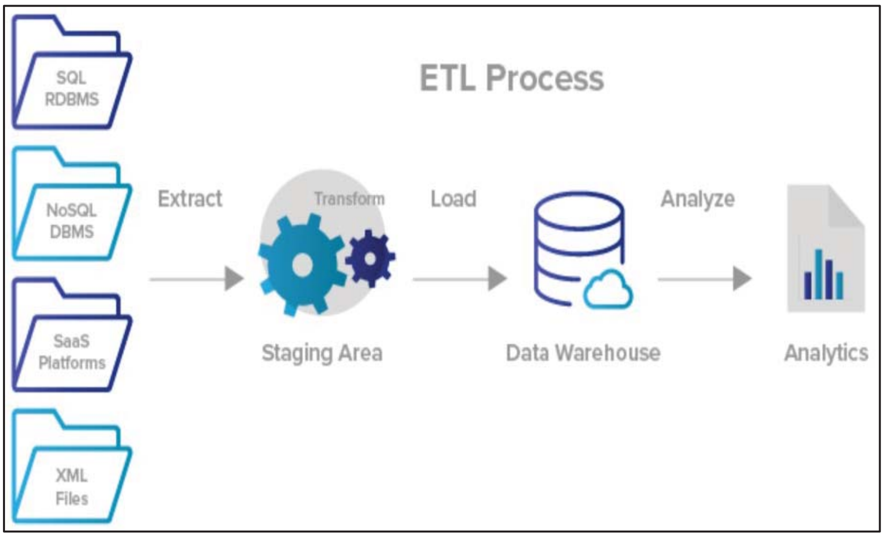


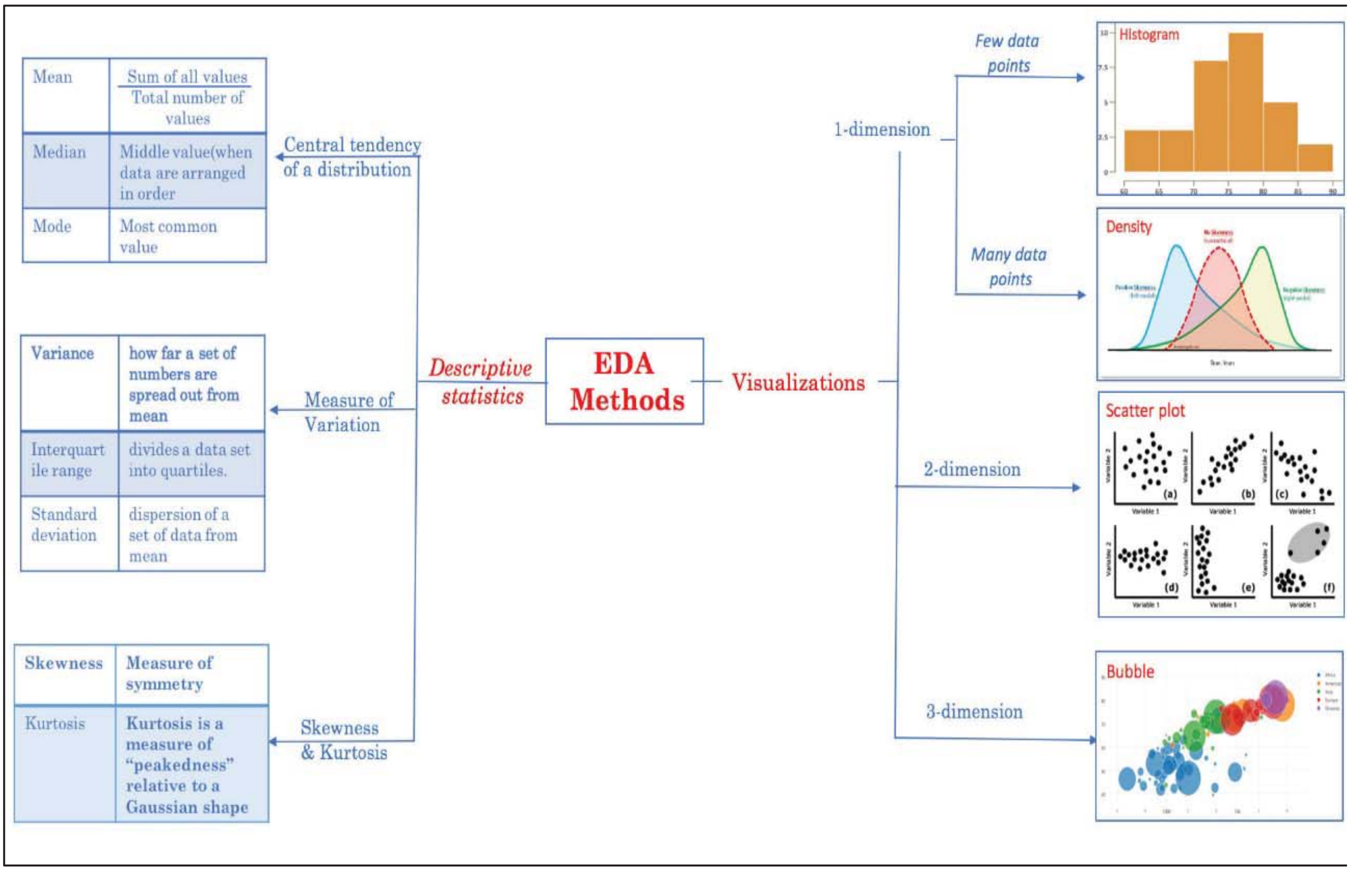
Figure 4-1: A Schematic of the ETL process. Data is extracted from various sources, transformed into a single cohesive dataset and uploaded to a storage facility for analytics and reporting (https://www.xplenty.com/blog/etl-data-warehousing-explained-etl-tool-basics/)**.**

4.1.2. Exploratory Data Analysis (EDA)

Data visualization is a key part of data processing and form part of the 'Transform' step of ETL. Summarising data or information visually makes it easier to identify patterns and trends compared to inspecting spreadsheets (**Figure 4-2**). With the rise of big data, the ability to interpret increasingly larger batches of data is extremely important, especially where near-real time data is required for decision making. Exploratory Data Analysis (EDA) employs a variety of techniques, with the help of summary statistics and graphical representations, to:

- detect outliers and anomalies;

- uncover underlying structure;
- test underlying assumptions;
- extract important variables;
- develop parsimonious models; and
- maximise insight into a dataset.



Mean	Sum of all values / Total number of values
Median	Middle value(when data are arranged in order
Mode	Most common value

Variance	how far a set of numbers are spread out from mean
Interquart ile range	divides a data set into quartiles.
Standard deviation	dispersion of a set of data from mean

Skewness	Measure of symmetry
Kurtosis	Kurtosis is a measure of "peakedness" relative to a Gaussian shape

Figure 4-2: A schematic of Exploratory Data Analysis (EDA). Once the ETL process is completed, the data is used to perform various analytics and visualized for reporting (https://www.fiverr.com/ivancui/do-exploratory-data-analysis)**.**

Most EDA techniques are graphical in nature, as the purpose of EDA is to explore data to reveal its structures and patterns quickly and to readily gain new insight into the data. While there are a few quantitative EDA techniques, the simplicity of the graphical techniques allows for quicker assessment of the data, and relies on various techniques of:

- Plotting raw data, such as data traces, histograms, bi-histograms, probability plots, lag plots and block plots.
- Plotting simple statistics, such as mean plots, standard deviation plots, box plots, and main effects plots of the raw data.
- Positioning such plots to maximize our natural pattern-recognition abilities, such as using multiple plots.

In this project, the ETL and EDA processes described above were used to process and interpret the data received. The EDA process was used in the Transformation stage of the ETL pipeline to gain insight and better understand the data.

4.1.3. Data processing

Data processing (sometimes referred to as 'data wrangling') refers to the process of cleaning, restructuring, and enriching the raw data available into a more usable format. It goes through the generic steps of:

(1) collation,

(2) assessment,

(3) cleaning (or formatting),

(4) validation (or quality control) and finally

(5) uploading to the database.

Data from each of the providers was often stored with different structures and file types and therefore it was necessary to pre-process this data into a single coherent schema to RIMS requirements for upload. Note that a RIMS-ready format was used and this remains valid, despite the migration to SADC-GIP. During the ETL, the "Transform" step allows for the identification and correction of incomplete, inconsistent, and erroneous data this is often present in real world datasets. **Figure 4-3** provides a schematic illustration of the entire process. The current document describes the steps taken from multiple datasets to a single dataset, and then quality control the final dataset before being uploaded to RIMS.

All operations to sort and clean data into a structured format before being stored on RIMS were carried out programmatically within a Python environment. This produces a 'paper trail' of changes from the received data to the final output dataset, and allows for easy replication at a future date for additional data similarly formatted. The final output is a 'RIMS-Ready' dataset that can be uploaded to the online RIMS database. As permissions are under negotiation with the primary data providers and not received yet, the data may only be uploaded once a formal agreement is reached.

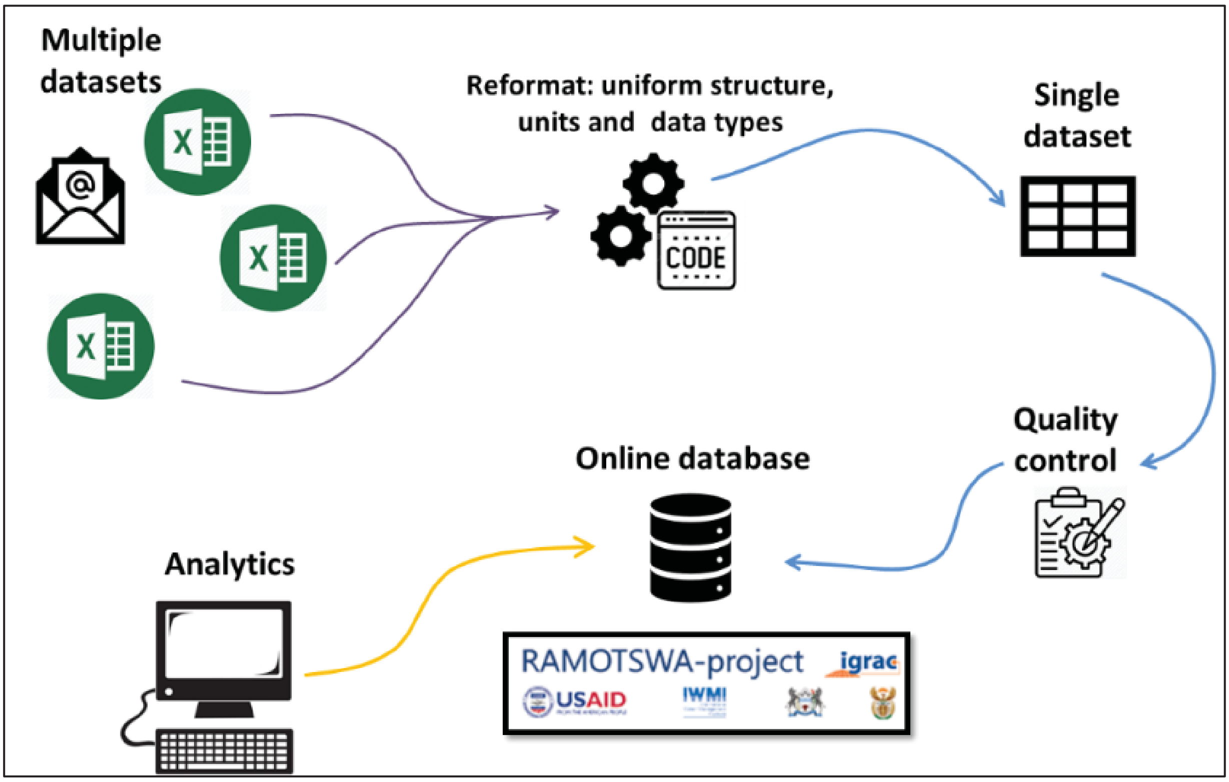


Figure 4-3: Schematic diagram illustrating the general data flow process.

4.1.4. Python tools and setting up Python environment

As the datasets being dealt with for this project were of varied quality and structure, an initial round of data pre-processing and cleaning was required. This was done using Python 3 was used in a Jupyter[18] notebook environment.

Python is one of the main programming languages for a variety of data science applications. It has libraries for data loading, wrangling, visualization, statistics, machine learning and many more. These libraries provide a large array of general- and special-purpose functionality such that pre-processing data for application in machine learning algorithms can be performed using a single interface, such as a Jupyter Notebook. Using the Anaconda distribution environment Python 3 was installed in a Jupyter notebook environment.

Jupyter Notebooks provide for easy documentation of data processing steps and allow for easy sharing of data analysis projects. An additional advantage of Jupyter Notebooks is that the documentation can be shared with non-python users as HTMLs and PDFs allowing for enhanced data sharing. The assessment focused on structural (i.e. rows, columns) and quality (i.e. missing, erroneous values) issues, with the overall aim of transforming the data received into a 'RIMS-ready' format. Changes were captured in the Jupyter Notebook and defined into code cleaning tasks.

The Python programming language contains a variety of built-in functions to read and manipulate data. The main packages (tools) used for assessment, manipulation, cleaning and visualization are listed in **Table 4-1**. These packages for datasets to be cleaned and formatted into the required structure for upload to RIMS, with the inputs of these correctly structured and cleaned datasets being shown in **Table 4-1.**

The main libraries used during pre-processing were the Pandas and Matplotlib packages. Pandas is an open source library, developed for data handling and analysis, being built around a data structure called a Data frame; which is a table, like an MS Excel spreadsheet. Pandas can ingest from a great variety of file formats and databases, like SQL, Excel files, and comma-separated values (CSV) files. Pandas provide a range of methods to modify and operate tables; in particular, it allows SQL-like queries and joins of tables. Matplotlib allows for scientific plotting in Python, providing functions for making visualizations such as line charts, histograms and scatter plots. These built-in functions, however, are limited, and therefore the project developed tailored functions that leverages off more sophisticated programs called modules (Table 4-1).

Table 4-1: List of packages used during the assessment, wrangling and cleaning of data.

Use	Packages	Link
Data processing and manipulation packages	Pandas	https://pandas.pydata.org/pandas-docs/stable/
	Numpy	https://numpy.org/doc/
Data visualization and exploration	Matplotlib	https://matplotlib.org/3.1.1/contents.html
	Seaborn	https://seaborn.pydata.org
For systems management	Os, glob, shutil	https://docs.python.org/3/library/os.html
Accessing outlook emails	extract_msg, win32com.client	https://pypi.org/project/extract-msg/
Accessing databases	Simpledbf	https://pypi.org/project/simpledbf/
zipfile	Accessing zip files	http://effbot.org/librarybook/zipfile.htm
Scipy	Statistical package	https://www.scipy.org/

[18] https://jupyter.org/

4.2. DATA CLEANING PIPELINE PROCEDURE

The follow sections provide a summarized description of the processes taken during each step of the data cleaning procedure. **Appendix 1** details the various steps taken, and assessments carried out, with the respective code for each file in the form of the Jupyter Notebook.

4.2.1. Collating

NGA and WMS data was provided by the Department of Water and Sanitation in .zip files. Each .zip file was extracted programmatically within a Python environment. The project could not automatically download data from NGA due to the need for an API which grants permissions for accessing the database.

4.2.2. Assessing

The assessment phase served to identify quality (content) and structural (tidiness) issues with a dataset. Quality issues are usually related to data type or incorrect information within the dataset. Structural issues are related to the organisation of the data (i.e. each variable forms a column, each observation forms a row, each type of observational unit forms a table). Assessment is done through a combination of visual (i.e. scrolling through the data, checking column names, etc.) and programmatic approaches (summary statistic checks with the use of pandas .info() and .describe() methods, plotting, etc.).

The visual and programmatic assessment identifies quality and structural issues needing to be addressed during the cleaning phase. This included checking for consistency in data types, including:

- Whether the input data points are of numeric, string, datetime, etc. data types,
- missing or invalid values (having impossible values such as negative borehole depth; depth to groundwater level deeper than the total depth of the monitoring well, etc.)
- inaccurate and/or inconsistent values (i.e. units of measurement).

The processes used to identify issues are documented in a Jupyter Notebook. Identified issues with each individual dataset are summarised below:

- Structural issues, i.e. where columns are variables/parameters and each row represent an observation point. Unstructured data sets make it difficult to perform analysis with the data. For example, machine learning packages in Python require that the data be set up as X (input) and Y (output) columns variables, and each observation as rows. Additionally, this is a requirement for data sets being uploaded to RIMS.
- Columns of the same value named differently to what is required by RIMS. For example, Station name to borehole id (bh_id) as in RIMS. To ensure consistency, names were changed as per RIMS requirements.
- White spaces at the end of station names (alias for borehole ID), e.g. 'A1N0001 ' instead of 'A1N0001'.
- Date field not captured in the required format, e.g. 19980904 instead of '**%Y-%m-%d** (i.e. 1998-09-04). Additionally, measurements with missing times and dates, which was replaced an 'unknown' value. As per RIMS requirement, all missing non-numeric data points should be replaced with 'unknown' values (-9999 for numeric data points).
- Quality and datatrans columns in water level data sets captured as codes (e.g. 26), while descriptions of each code represented in separate files. Furthermore, water level entries captured as negative (e.g. -28.93), while RIMS requires absolute values.
- Coordinates (latitudes and longitudes) for each borehole ID/station were stored in a separate file. Data points with no coordinates hold little value.
- No source and contact information columns provided, for tracing where the data came from.

4.2.3. Cleaning

Each individual dataset required unique steps for cleaning and restructuring to match the RIMS upload format. No changes were made to the original data files, merely to the **DataFrame** within the Python environment.

Overall, initial steps were to structurally align rows as properties or samples, and columns as features in each dataset. Aligned features (or columns) were renamed to match the RIMS template, ensuring consistency of data types and units. Rows or columns having no data, i.e. blanks or no data-points such as **'ND'** or **'NO DATA'** were changed to '**-9999'** if numeric or **'unknown'** if text. For all sites latitude and longitude were mapped across separate files and were concatenated into a single longitude latitude column and, using Pandas datetime module, dates converted to RIMS specific formatting, i.e. **yyyy-mm-dd.** Additionally, negative water levels were changed to absolute values and descriptions were added to water quality codes to understand what code 26 means using **'Quality Code Description.xlsx'** and **'Quality codes.xls'** using a mapping function to convert code number to description of code.

Finally, all files were concatenated into a single **DataFrame**, and further inspected for any structural issues. This identified that new data features (i.e. columns) were included, which are not yet accounted for in the RIMS upload format. At this stage these additional columns were kept within the **DataFrame** for legacy purposes, however they can be dropped at the final stage prior to uploading to RIMS. The cleaning steps taken for each individual dataset are documented in Appendix 1.

4.2.4. Quality control

RIMS does not currently allow for data quality control once data is uploaded; creating the need for such a step to be done prior to upload. A preliminary quality control estimator using standard Python functionality was developed, which iterates through a standard RIMS-ready template (rows, columns), and identifies erroneous data values based on conditional statements. Erroneous data entries were flagged and captured for further inspection. The entire process was carried out in a Python environment and documented in a Jupyter Notebook.

The **data_quality_control_check.py** script handles the preliminary data quality control check on the final datasets generated. These steps are outlined below:

- The algorithm reads individual files and checks whether the minimum required columns in that file are present. If not, it prints the message to the console showing missing required columns. The required columns are obtained from the RIMS template, i.e. **bh_id, country, latitude, longitude**. If required columns in the file are missing, then the file processing is not undertaken. Currently, it only prints a message to the console or terminal, https://docs.microsoft.com/en-us/windows/terminal/. The next step would be to develop a simple Graphical User Interface (GUI) to display these messages.
- The same checking of required columns is done on the final dataset and output the names of columns that are missing.
- The script also enforces the datatypes; it explicitly loads the columns with the assigned datatypes from the TWC datatypes dictionary, the dictionary uses a list of column names and the specific data type the column should be.

In should be noted that the conditional statements and logic checks have not yet been defined. Sterckx et al. (2019) as part of Theme 1 Deliverable 3 (Quality Control System Report) of this project outlines suggested criteria; however these have not yet been implemented programmatically. The criteria and thresholds will need to be developed, agreed upon by stakeholders and implemented at a later stage.

4.3. DATA CLEANING VISUALISATIONS.

Data visualization refers to the techniques used to communicate data or information by encoding it as visual objects (e.g. points, lines or bars) contained in graphics – ranging from static printed forms to real-time interactive digital four-dimensional representations. To convey ideas effectively, both aesthetics and functionality need to go together to ensure key aspects are communicated in a more intuitive way. Data visualisation is an essential component to make insights derived from data analysis understandable and usable to the end user and may be tailored according to the end user in mind, e.g. a geohydrologist will assess data differently compared to a municipal water manager.

4.3.1. Cleaning Ramotswa aquifer time series groundwater levels

Quantitative analysis often involves working with time series data in various forms. A time series is an ordered sequence of data that typically represents how some quantity changes over time. Examples of such quantities could be water quality changes or water level fluctuations in a borehole over time. The pandas and Matplotlib packages in Python are a powerful suite of optimized tools that can produce useful analyses of time series data with a few lines of code.

Data cleaning is one of the most critical steps before any analytics can be performed on the data. However, with large datasets this process becomes time consuming and prone to error when done manually. Effectiveness is increased by using Python, to more easily assess data errors (**Figure 4-4**) reliably transform and clean data with a few lines of code. Visualising each step of cleaning offers advantages to the cleaning process. As seen in **Figure 4-5** temporal static water levels contained both -9999 values (top plot) as well as anomalies (middle plot) and once removed, a clean dataset of static water levels and visualization was produced (bottom plot). This procedure was undertaken using Pandas to format datetime index and Matplotlib library to plot changes in water level over time. This process was described in Theme 1 Deliverable 2 of this project (Umvoto, 2019).

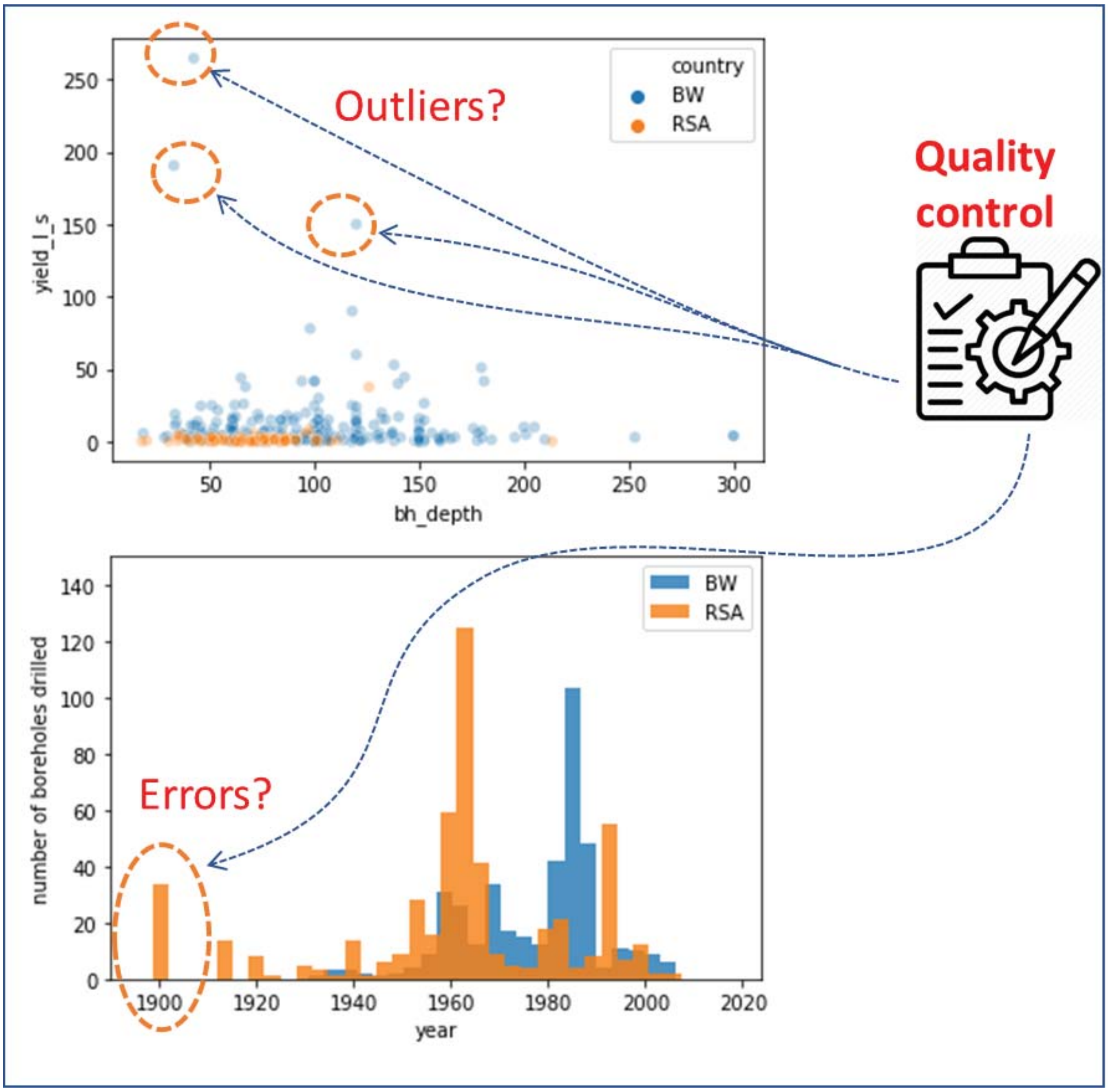


Figure 4-4: An example of the visualisation of the number of boreholes drilled versus time for Botswana (blue) and South Africa (orange). Notable is are data outliers in the top graph, and date errors causing a clustering of results on 1 January 1900.

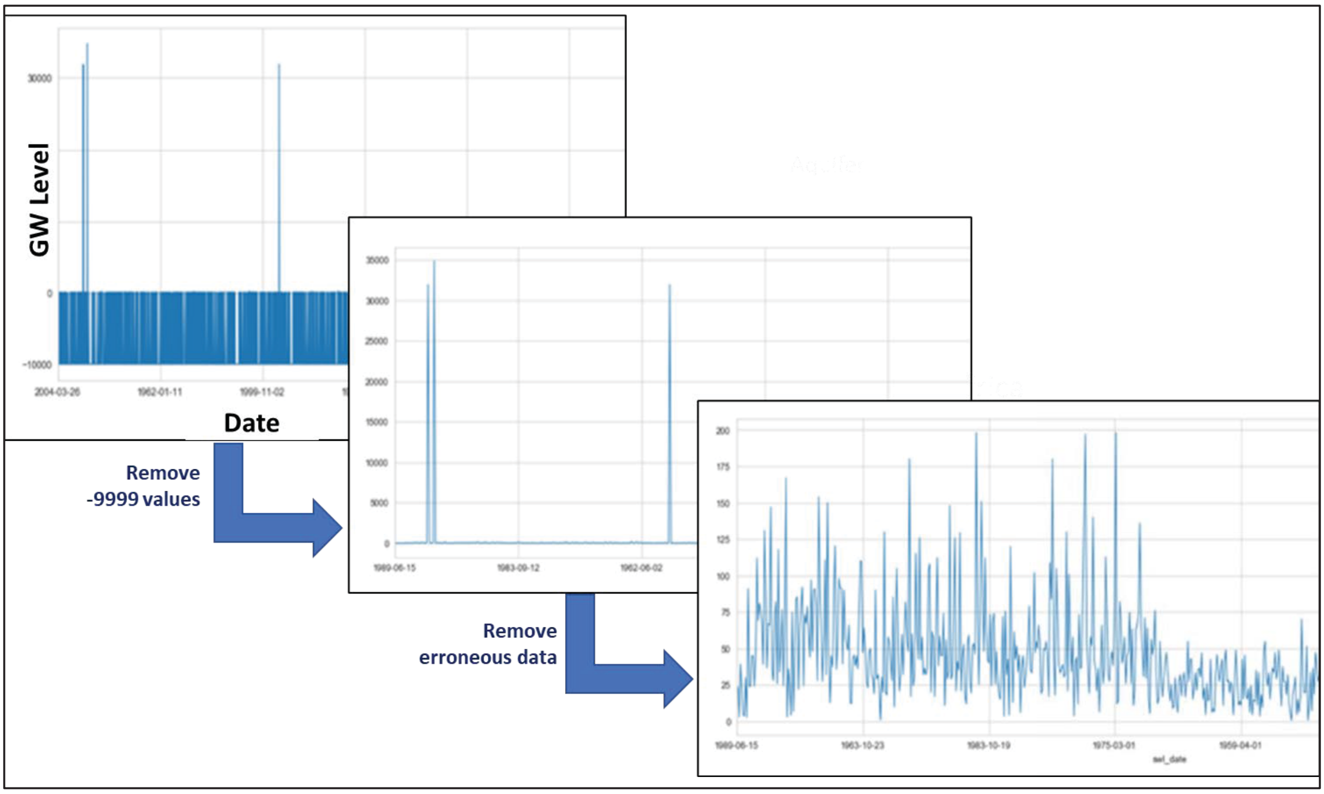


Figure 4-5: Cleaning static groundwater levels. The graph shows that the static water levels contained -9999 (top) and outliers (middle). The bottom graph shows the results for cleaned static water level data.

4.3.2. Borehole locations in the Ramotswa

Numeric data that identifies the geographical location of a physical object according to a geographic coordinate system is termed spatial, geospatial, or geodata. An example of a kind of spatial data are the coordinates of an object such as latitude, longitude, and elevation. Geographic Information Systems (GIS) or other specialized software applications can be used to access, visualize, manipulate and analyses geospatial data. GeoPandas: extends the datatypes used by pandas to allow spatial operations on geometric types. Below, the Geopandas library was used to plot the borehole IDs (bh_id) to have a better visual of where the stations are located (**Figure 4-6**). Keys for each borehole ID were assigned to determine which bh_id points belong to the which coordinates. Plotting these points on a map provides an understanding of the datasets and where the data points are located spatially.

4.3.3. Simple imputation technique

As with many real-world datasets, the data received contained missing values. Although, there exist more robust techniques for imputing missing values, **Figure 4-7** shows a simple example of an approach to determine the strength and character of the relationship between elevation data received and SRTM 90 m resolution elevations extracted in GIS. Because the elevation data received contain missing values, the technique in **Figure 4-7** can be used to fill in these missing values.

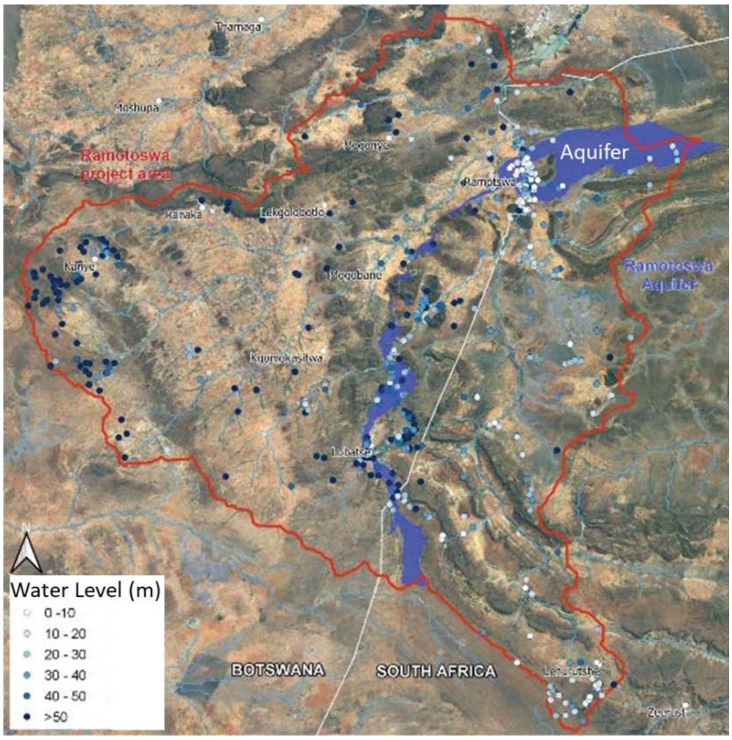


Figure 4-6: Plotting borehole data (in this case water levels) in geographic space. The darker points locations indicate borehole locations with greater depth to water.

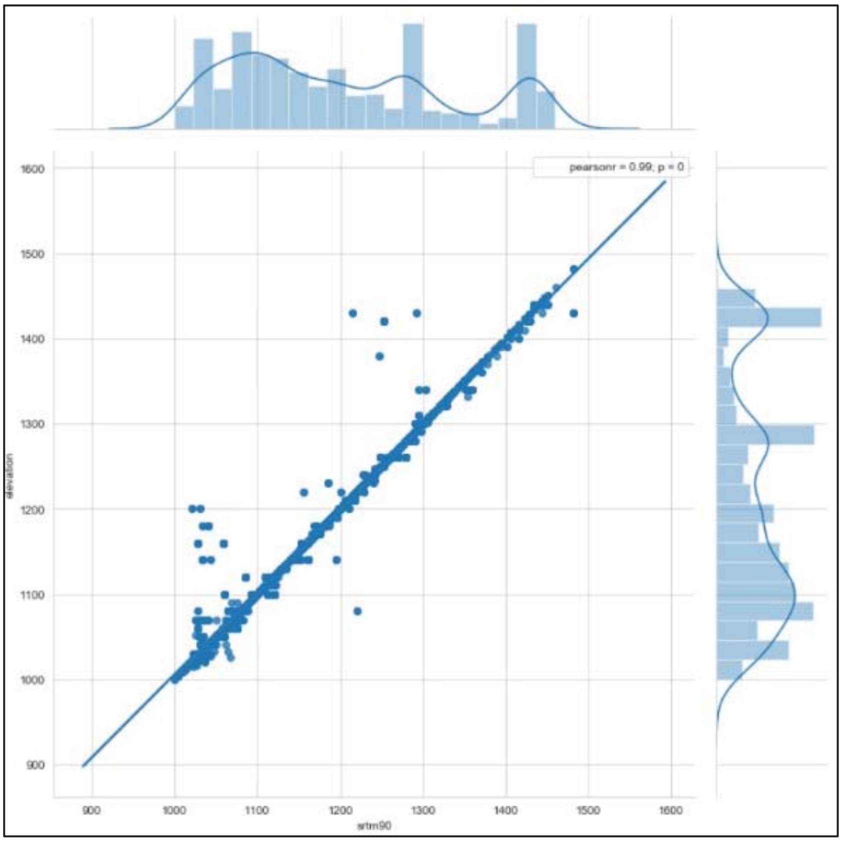


Figure 4-7: Regression analysis for imputing missing values.

4.4. How to Run the Toolkit

Data cleaning tasks that were developed in the Jupyter Notebook were converted into executable Python modules. There are several modules (**Table 4-1**) built to modularize the different data consolidation and cleaning tasks with each python module/script doing a specific task in the pipeline – all data files and modules are stored in the same folder or repository: **TWC_DataAnalysis**.

Currently, a copy of the repository is needed on the local machine of the user. Future work should consider uploading the pipeline package to Python Package Index (**PyPi,** https://pypi.org/)**,** which can then be installed via **pip** (https://packaging.python.org/tutorials/installing-packages/) install command in Python. Uploading the final pipeline package to PyPi removes the need for the user to download individual packages to their local machine. One uploaded, to install the pipe the command **pip install <package_name>** should be executed from the command line.

All the scripts are called inside the **main.py** (**Appendix 3**) script, which is executed first and to call in the relevant scripts. To run the **main.py** script, open the terminal on the computer then change the directory to the project directory, in this case **TWC_datasets** achieved by running the following command: **$ cd TWC_Datasets,** i.e. executing the task from the command line. Once inside this directory, execute the script by running: **$/main.py** this will run the main script initiating the data cleaning process. The main module will call the specific python script that does the first task in the data cleaning process, find files, read water levels and quality files, builds dataframes, manipulate data on the dataframes, transform the fields, check datatypes, check outliers, check thresholds and lastly display the output to the console.

5. QUALITY ASSURANCE

5.1. Overview

The ultimate factor that determines the good quality of data is its usability. Clearly, data is useful when it is not erroneous, but also, when it is relevant and sufficient: when the sampling frequency is enough to identify trends, when measurements are taken at enough locations, or in the specific case of groundwater quality, when all the necessary parameters needed for a certain purpose are monitored.

If the purpose of the network is not well defined, if it is not clear who will benefit from the data, if the measurements are not taken with an appropriate frequency or if the network does not have a sufficient number of wells, it is very likely that the monitoring programme will not achieve the desired results. These key aspects are explained below in more detail.

To ensure the accuracy of data, standard quality assurance and quality control (QA/QC) measures are required. This consist of a series of measures and procedures that must be applied from the beginning of the data acquisition plan.

QA/QC measures for the **monitoring** network itself includes (Sterckx and Ruz Vargas, 2019; SADC-GMI, IGRAC, IGS, 2019):

- **Purpose**, of the network to make clear who will benefit from the data.
- **Roles and responsibilities** including in the field, laboratory and office. Data quality can be improved through:
  - Certification of professionals involved in groundwater data collection;
  - Vocational training programs for groundwater technicians;
  - Adhering to standard guidelines and sufficient training of field technicians;
  - Using modern technologies such as digital field forms/mobile Apps;
  - Using automated data collection devices; and,
  - Routine data checks.
- Appropriate **frequency** and a sufficient **number of wells** to achieve the desired results.
- Appropriate well **location** based on
  - Purpose: e.g. aquifer dynamics requires areas with limited human influence, while water quality requires wells source zones and pathways); and,
  - Installation depth, targeting aquifer(s) of interest.

There are several consistency checks to be performed to test the quality of data. Metadata checks, logical and outlier checks are fundamental, both for groundwater quality and quantity (Sterckx and Ruz Vargas, 2019):

- **Metadata Checks**, ensuring at least the following is provided:
  - Well identifier;
  - Coordinate system;
  - Units;
  - Date and time of sampling; and,
  - Elevation (especially when measuring groundwater level).
- **Logical Checks:**
  - Data type: e.g. are numbers or text received for corresponding data type (e.g. text for geology, and number for water level)?
  - Constraint checks: e.g. is pH between 0-14? Are mapped coordinates within expected area?
  - Consistency: e.g. depth to groundwater level cannot be deeper than the total depth of the monitoring well. For water quality, there are standard checks that can be used such as

major ion balance, ratio of electrical conductivity to total dissolved solids, scientifically impossible results (such as elevated nitrate in anaerobic waters)

  - Cross-Reference: e.g. are measured results comparable to previous data from the same data point, or with contemporary data from neighbouring data points?

- **Outlier checks:** This can include the use of visualisations to chart results (e.g. Box-and-Whisker Plots) or software to flag outliers for scrutiny.

Flagged data, or data identified as suspicious, must be traced back to identify errors in the collection or data transfer processes, and such errors should be rectified, if possible. If it is determined that the procedures followed were correct and no other possible errors are found, the data should be accepted in the database. However, if it is not possible to assure that there were no errors in the collection and data transfer processes, the data should remain flagged.

The data validation flow (Figure 5-1) taken from the *SADC Framework for Groundwater Data Collection and Data Management* (SADC-GMI, IGRAC and IGS, 2019) provides a data flow diagram, from field collection to final database storage, including various QA and QC processes.

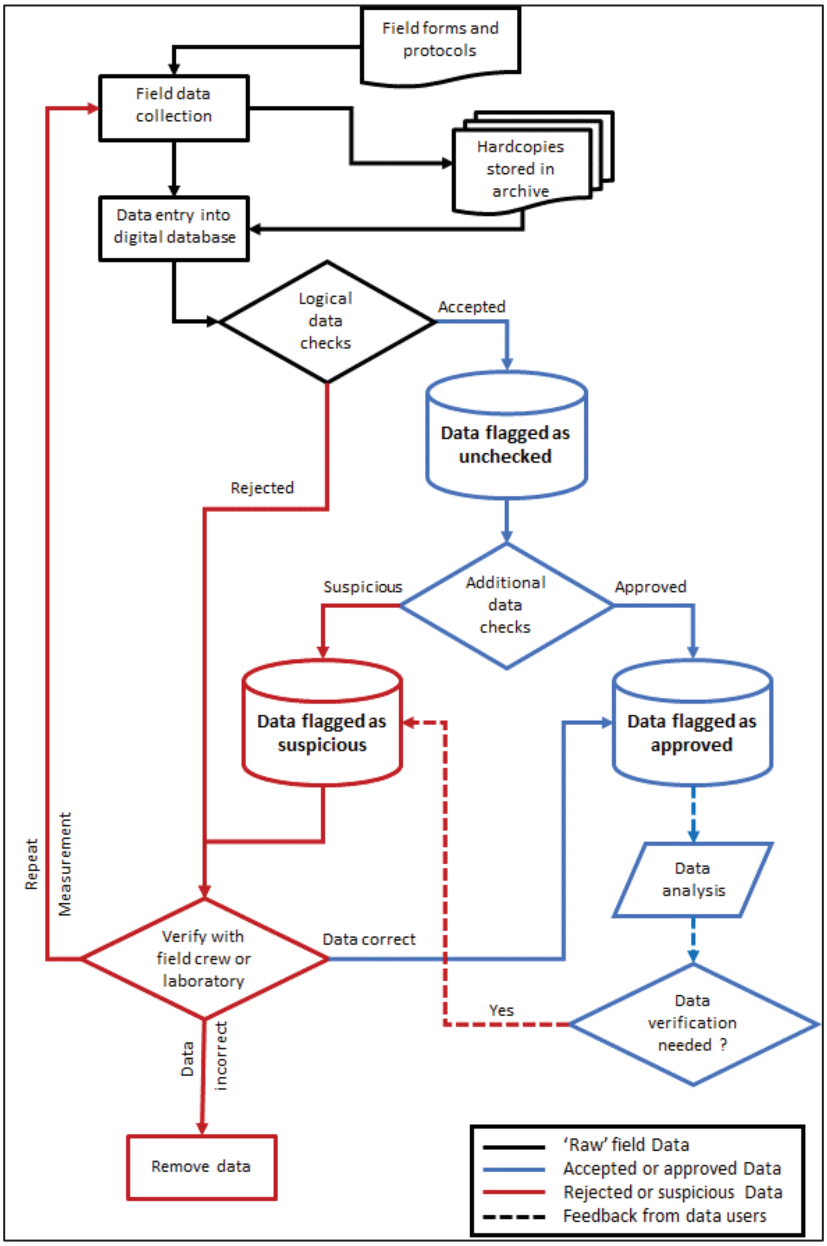

Field forms and protocols
Field data collection
Hardcopies stored in archive
Data entry into digital database
Logical data checks
Accepted
Data flagged as unchecked
Rejected
Suspicious
Additional data checks
Approved
Data flagged as suspicious
Data flagged as approved
Data analysis
Repeat
Measurement
Verify with field crew or laboratory
Data correct
Yes
Data verification needed ?
Data
incorrect
Remove data
'Raw' field Data
Accepted or approved Data
Rejected or suspicious Data
Feedback from data users

Figure 5-1: Data QC flowchart for data management (SADC-GMI, IGRAC and IGS, 2019)

5.2. Handling Erroneous and Dubious Data

When data validation checks identify erroneous or dubious data, these data should be at least flagged and maybe deleted from the database. While both database managers and data users can perform data validation checks, only database managers should be able to go through this process.

5.2.1. Save a Copy of the Database

It is important to save a copy of the database before flagging and possible deleting data. In case of mishandling, the database could be restored from the backup.

Data stored in the SADC-GIP are saved on an external server. Data will be erased only if the data file is deleted on the server. It is possible to store previous versions of data files in the SADC-GIP. If the database is not saved, a copy should be made and stored locally.

Data from SADC-GIP can be downloaded as ESRI shapefile (.shp) or Excel (.xls) directly from the viewer.

Data can be downloaded from the GGMN in .csv format. By clicking on the button Download, the user can get all data from the current organisation (in this case, from the organisation “Ramotswa Aquifer”). By selecting one station and clicking on the button Export, the user can get the complete time series of the selected well.

Data from GEMStat can be requested via direct email to gwdc@bafg.de or using the online data-request form from https://gemstat.org/custom-data-request/. In both cases, it should be stated which parameters are needed. Data is not to be used for commercial purposes and not be transferred to third parties.

5.2.2. Flag Erroneous or Dubious Data

There are different ways of doing this depending on the database. In the SADC-GIP, data files should be uploaded with an additional column for flagging dubious data. Data cannot be flagged in the SADC-GIP itself, but in a GIS software or in Excel (the SADC-GIP supports both Excel files and shapefiles).

In the GGMN, it is necessary to use another workspace to store the flagged data, for example, “Ramotswa Aquifer – flagged data”. If this workspace does not exist yet, the administrator of the workspace has to send a request to the service desk to create it. The suspicious data has to be stored here, by uploading a local copy of the suspicious dataset. Stations are uploaded as ESRI shapefiles, and timeseries as “csv” files.

In GEMStat, data can be flagged during the upload or afterwards. Data can be flagged as “Fine”, Suspicious” and “Poor”. This flag is also included in exports to answer data requests.

5.2.3. Remove Erroneous Data from the Database

Once the suspicious dataset is flagged, the copy that is not flagged has to be removed from its original place.

In the case of GGMN, this is done by sending a request to the service desk indicating which stations or time series should be removed. The handier way to do this is to request the deletion of the whole time series (per station), and then upload again the time series but without including the suspicious measurements.

In the SADC-GIP, new data files can be uploaded in which erroneous data have been removed. Precedent versions of the data files can be deleted from the SADC-GIP or kept for record or backup.

To remove data in GEMStat, a request should be sent to gwdc@bafg.de.

5.3. Implementing Quality Assurance/Quality Control Procedures

Most of data validation checks can be automatized. These checks are available in many advanced relational database software programs, but such databases are expensive and require database developers. Checks can also be programmed in Excel using macros[19].

More importantly, the staff in charge of the database have to be adequately trained. Appropriate training and capacity development were identified as the primary solution to guarantee good quality data in the *SADC Framework for Groundwater Data Collection and Data Management* (SADC-GMI, IGRAC and IGS, 2019). For database management, responsible staff should have enough information and communications technology (ICT) skills and be provided ad-hoc training on the software in use. In 2018, multiple dedicated trainings on the RIMS and GIS software were organised with the technical staff of the national water departments during the Ramotswa 2 project. More training could be provided on GIS and SDI. Specific training could also be provided on the use of Excel. Even frequent Excel users often lack a sound understanding of Excel functioning, data types, basic functions, let alone advanced functions and macros development. Training could be provided as a few days' workshop (like in the Ramotswa 2 project) or online, depending on the budget available and the needs of the departments.

When using RIMS, GGMN and GEMStat as the transboundary databases, data QA/QC will depend mainly on the implementation of good practices and other measures (suggested in this report) since these databases do not have the capabilities to do so automatically. To assist in simplified ingestion of existing data to the selected database option, automated data ingestion should be considered. This converts the data providers data into a format that is suitable for the database upload. This can be done using excel macros (Sterckx et al., 2019) or programming tools as developed by Theme 1 as part of this project.

[19] A macro is programming code that runs in Excel environment and helps automate routine tasks.

6. DATA SHARING

Access to sufficient data, information and knowledge is recognized as one of the first conditions to good groundwater governance[20]. Data access allows stakeholders to participate actively in the sustainable management of groundwater resources, and thus data should be made publicly available as far as possible.

Few transboundary aquifers in the SADC region have already been subject to cooperation between Member States. The first case study was the Stampriet Transboundary Aquifer System (STAS), shared between Botswana, Namibia and South Africa. The second case-study was the Ramotswa transboundary aquifer, shared between Botswana and South Africa. In both case-studies, a first multidisciplinary assessment was made, based on data provided by the countries and harmonized. The role of data exchange platforms was instrumental. Despite the success of these two projects, some shortcomings in data sharing were identified that need to be addressed to advance transboundary cooperation and management of groundwater resources. Experience (in particular from the RIMS) shows that the collection of data was limited after all readily available data were collected (the “low-hanging fruits”). The national departments did not fully engage in updating the datasets as more data were collected in their country. There was also some confusion as of what data should be collected and who could access the data. Decisions on these matters were made on a case-by-case basis. Another challenge was the harmonization of data collected from different databases, in different formats. Discrepancies appear because organisations and countries have different practices, standards, terminology and computer systems. The use of protocols would considerably improve the flow of data exchange between the national departments and facilitate the harmonisation of data.

For transboundary datasets there is a need for a shared storage location accessible by all, without danger of being made country-specific, or being made redundant. As a result, country data is often shared with neutral bodies such as the International Groundwater Resource Assessment Centre – IGRAC (that works under the auspices of the World Meteorological Organisation – WMO) and the global water quality database GEMStat (hosted, operated, and maintained by the International Centre for Water Resources and Global Change – ICWRGC). Data in such databases can be used for status evaluation, policy making, research purposes or within the scope of education and training initiatives.

A common concern from data providers is retaining data ownership and receiving appropriate citation for data used by others (to record data/project impact). Readily available licenses like Creative Commons licenses[21] can be used to specify different options in terms of attribution, processing or application.

This includes making data freely available, or specifying selection(s) of the following:

- Credit must be given to the creator;
- Adaptations must be shared under the same terms;
- Only non-commercial uses of the work are permitted;
- No derivatives or adaptations of the work are permitted.

Exceptions to data sharing are unreliable data that do not meet enough quality standards and work in progress which should not be shared publicly, but only between the departments. In addition, sensitive data collected by the departments and data from external stakeholders typically requires restricted sharing (Sterckx et al., 2019).

[20] See for instance the conclusions of the worldwide programme Groundwater Governance – A Global Framework for Action (http://www.groundwatergovernance.org/). Based on a global diagnostic of groundwater governance across the world, the project defined general conditions for good governance, among which availability and access to data.
[21] https://creativecommons.org/

6.1. Data Needs

The data required for integrated assessment of groundwater resources that can be used to describe the general state of transboundary groundwater resources includes (SADC-GMI, IGRAC, IGS, 2019):

- Physiography and climate (temperature, precipitation, evapotranspiration, land use, topography, surface water);
- Aquifer geometry;
- Hydrogeological characteristics and parameters;
- Environmental aspects (e.g. pollution and pollution sources);
- Socio-economic aspects (e.g. Statistical information on populations, water sources, water supply); and,
- Legal and institutional aspects (e.g. transboundary and domestic legal and institutional frameworks).

Not all data are required under all circumstances and the list of parameters above is not comprehensive. The importance of the list is that it highlights the multitude of data required for integrated management of groundwater resources and acknowledges that not only physical parameters are required on a regular basis (SADC-GMI, IGRAC, IGS, 2019).

The data needs from these sources for various objectives are outlined in **Table 6-1**. Broadly, water monitoring can include **groundwater observation wells, groundwater pumping wells, springs and surface water** observations. Water data measurements associated with these can be divided into three types: **levels, discharge and quality**.

Focussing on the measurement of groundwater levels, discharge and quality, **Table 6-2** summarises the primary data needs to provide a robust data set. Included in this is metadata which is the "data about data."

Table 6-1: Example of data needs from different data sources for specified objectives (SADC-GMI, IGRAC, IGS, 2019)

Monitoring Objectives	Groundwater observation wells			Groundwater pumping wells			Springs			Surface water observation points		
	Levels	Discharge	Quantity	Levels	Discharge	Quantity	Levels	Discharge	Quantity	Levels	Discharge	Quantity
Groundwater development												
Groundwater system characterization	XX	N.A.		X			X			X		
Groundwater potential for development (quantity and quality)	XX	N.A.	XX		XX	XX		XX	XX		XX	XX
Best locations for well fields	XX		XX			XX			X			(X)
Control and Protection												
Trends of over-exploitation	XX	N.A.		X	XX			XX			XX	
Nature conservation	XX	N.A.			XX		X	XX			XX	
Saline water intrusion	X	N.A.	XX*		XX	XX*				X	X	(X)
Land subsidence	X	N.A.			XX							
Contamination of aquifers		N.A.	XX			XX			XX			XX

X = desirable data, XX = necessary data, XX* = mainly chloride; N.A. not applicable

Table 6-2: Primary groundwater monitoring data needs (after Sterckx *et al.*, 2019).

Data Type		Description
Metadata	**Data about the point data**	• Data point type • Original identifiers and unique transboundary database identifier • Country • Physical address • Location • Elevation • Status
	Data about the data provider	• Data providers • Data transfer date • Method of data transfer
Borehole drilling data		• Completion date • Depth (of water strike) • Stratigraphic log • Design: depth, altitude of screen intervals, material, pump • Pumping test: including yield and date Ideally this data is provided to water resource managers directly by the borehole drillers
Monitored data	**Groundwater level**	• Unit: Length. • Reference point to start measuring (e.g. sea level or ground surface, or casing edge) • Date of recording (preferably including year, month, day, hour, minute and second) • Method of measurement: manually or automatically. • If data loggers used: depth to probe, atmospheric correction, etc.)
	Groundwater quality	• Key groundwater quality parameters include: ○ pH ○ Temperature ○ Electrical conductivity ○ Major anions (i.e. nitrate, sulphate, chloride, bicarbonate alkalinity) and cations (i.e. calcium, magnesium, potassium and sodium) ○ Biological parameters (e.g. *E. coli*) ○ Hardness ○ Heavy metals ○ Dissolved oxygen • Name of parameter, • Unit (of concentration) • Method of measurement • Name of laboratory • Relevant dates • Automatic measurement devices linked to telemetry are available for various parameters
	Groundwater abstraction/ flow	• Unit (volume/time) • Method of measurement: flowmeter, bucket/chronometer • Date (preferably including year, month, day, hour, minute and second)

Access to sufficient data, information and knowledge is recognized as one of the first conditions to good groundwater governance[22]. Data access allows stakeholders to participate actively in the sustainable management of groundwater resources, and thus data should be made publicly available as far as possible.

[22] See for instance the conclusions of the worldwide programme Groundwater Governance – A Global Framework for Action (http://www.groundwatergovernance.org/). Based on a global diagnostic of groundwater governance across the world, the project defined general conditions for good governance, among which availability and access to data.

A common concern from data providers is retaining data ownership and receiving appropriate citation for data used by others (to record data/project impact). Readily available licenses like Creative Commons licenses[23] can be used to specify different options in terms of attribution, processing or application.

This includes making data freely available, or specifying selection(s) of the following:

- Credit must be given to the creator;
- Adaptations must be shared under the same terms;
- Only non-commercial uses of the work are permitted;
- No derivatives or adaptations of the work are permitted.

Exceptions to data sharing are unreliable data that do not meet enough quality standards and work in progress which should not be shared publicly, but only between the departments. In addition, sensitive data collected by the departments and data from external stakeholders typically requires restricted sharing (Sterckx et al., 2019).

7. CHALLENGES ENCOUNTERED

- Delay in receiving data from Primary Data Providers due to Memorandum of Understanding not in place at commencement of project.
- There is a need to on-board data providers at the earliest stage. This includes
    - relevant interest and buy-in from Governmental Department Directors,
    - identification of key outcomes and products the data providers expect or would appreciate by project conclusion,
    - identification of mandated department officials to assist in the process,
    - data-sharing permissions
- Limited/sparse primary data, especially temporal.
- Some data of uncertain fidelity/accuracy/providence (i.e. locations, missing metadata).
- Big data analytic options suitable to data at hand. Any new data needs new tools.
- Machine learning requires more data than currently available.
- Impacts of delay in internship/trainings on data analytics process.
- Permissions to automatically capture NGA data could not be finalised.
- Parallel work, thus could not include other theme outputs in our dataset used at internship.

[23] https://creativecommons.org/

8. PROJECT CONCLUSIONS AND RECOMMENDATIONS

Almost all real-world data and datasets suffer from incompleteness, inconsistencies and errors, and the water resource data received as part of this project are no different. Often these data issues are solved in an ad-hoc and manual manner requiring familiarity with the data. This report describes a programmatic approach that can be utilised to more effectively deal with these issues.

As part of the project, Umvoto and IGRAC formalized a quality control system to be applied to the datasets for the Ramotswa Aquifer System. This included protocols on data quality assurance to limit errors from happening during data acquisition as well as data quality control procedures for checking, validating (and where possible, correcting) data after data acquisition. After acquisition and prior to uploading and using for shared decision making, data has to undergo several phases of processing and cleaning which was undertaken using a data cleaning pipeline toolkit developed by Umvoto and IGRAC. The steps followed were data gathering, assessment (e.g. detect outliers or any inconsistencies), cleaning, and initial quality control, all of which were undertaken programmatically and documented in a Jupyter Notebook. The outcome is a 'RIMS-Ready' dataset, that can be uploaded to the online information management system. Although RIMS was migrated to SADC-GIP, this format remains valid.

Data QA and QC consist of a series of measures and procedures that must be applied from the beginning of the data acquisition plan. This means that a correct design of the groundwater monitoring network and a proper planning of the field campaigns are crucial to assure the capture of relevant and sufficient data.

Guidelines have to be followed during data collection, depending on the parameters of interest (groundwater quantity or quality), but this alone is not sufficient to ensure quality. Various quality checks have to be performed once in the office, to make sure that erroneous data are not being entered in the database, and that suspicious data are flagged for further analyses. The input of an experienced hydrogeologist that is familiar with the study area is key throughout the whole process.

Additionally, there are several consistency checks to be performed to test the quality of data. Metadata checks and logical checks are fundamental, both for groundwater quality and quantity.

SADC-GIP, GGMN and GEMStat, the online platforms chosen to store GIS data, groundwater level timeseries and groundwater quality data respectively, do not have their own automated QA/QC systems. Therefore, it is the responsibility of the staff in charge to implement external QA/QC measures, as the ones described in this report.

With regards to the Groundwater Data Sharing Protocol developed, there are the following recommendations:

- Due to the relatively short duration of the project *"Consolidation of Data and Application of Big Data Tools to Enhance National and Transboundary Data Sets in Southern Africa that Support Decision-Making for Security of Water Resources"*, of which this report is an output, it was not feasible to engage the Departments of Water Affairs of Botswana and South Africa in the production of this protocol. Experiences from previous and ongoing collaboration with the Departments of Water Affairs on groundwater data sharing in the Ramotswa transboundary aquifer area helped the project team to create this protocol but the direct input or feedback from the Departments could not be sought, unfortunately. Therefore, this protocol does not constitute a final document to be adopted right away by the Departments. It rather provides guidance for the Departments to create and adopt a protocol of their own. The Departments are invited to improve the present document into a final protocol that they can enforce.

- The Departments might decide to develop a protocol specific to the Ramotswa transboundary aquifer area, or they could agree on one groundwater data sharing protocol for several (if not all) aquifers shared between Botswana and South Africa. In that wider perspective, River Basin Organisations which are concerned by transboundary aquifers might be invited at the discussion table. For instance, LIMCOM and ORASECOM aim at improving the governance of transboundary groundwater. As mandated transboundary organisations, they might play an important role in the sharing of groundwater data between the countries.

- Next to a protocol on groundwater data sharing, the Departments will also need to plan regular meetings to supervise the enforcement of cross-border data exchange (according to the protocol), translate the data being collected into relevant information on transboundary groundwater resources, and elaborate management strategies to address transboundary issues.
- In a subsequent phase (medium-term horizon), the Departments could bring transboundary cooperation to the next level, by coordinating groundwater data collection on both sides of the border. They could agree on the data to be collected, the location and the frequency of groundwater monitoring data collection and the methods/formats to collect the data. This would facilitate the exchange and harmonization of datasets collected in each country. In terms of groundwater monitoring, the design of dedicated transboundary monitoring programs would ensure that all transboundary components of groundwater resources are adequately captured. Currently, this might be not be the case because national groundwater monitoring programs tend to address parts of transboundary aquifers only nationally, without considering their transboundary nature. A dedicated task team could identify where additional data would be needed, in complement to the monitoring data being collected by the countries.

Theme 1's aim for the visualisation tools was to assess the applicability of SADC-GIP, GGMN and GEMStat as visualisation tools for the Ramotswa Transboundary Aquifer. These three platforms were selected as they provide various benefits: SADC-GIP provides spatial data on a range of groundwater and groundwater related variables, GGMN focuses on the spatio-temporal groundwater level characteristics by providing advance statistical analysis including autoregressive models, and GEMStat is a global water quality database with unique data visualisation (graphs/plots) and statistical reports. The SADC-GIP is a SADC-GMI initiative under their management, and this is supported by IGRAC.

It is recommended that options be sought to incorporate the visualisations in any future tools for the Ramotswa Aquifer. By harnessing the existing power of these tools, this can leverage each of their strengths with regards to water resource data management. So as to promote the use of such platforms, future recommendations should consider the following:

- Data providers should move towards storing and making their data easily accessible. This includes governmental departments mandated with water resource management, such as the Departments of Water and Sanitation in both South Africa and Botswana. This would allow for the automation of data extraction and loading into visualization platforms, accelerating or even automating the process of going from raw data to useful information.
- Platforms providers such as SADC-GIP, GGMN and GEMStat should work together to increase cross-platform standardization, e.g. data requirements.

Further software development work should consider uploading the toolkit to **PyPi** for users to install via Python's **pip** installation command line. However, before uploading the tool further development is required and should include:

- Datasets column naming convention at the point of data collection by entities that collect data;
- Standard format for storing data;
- Metadata for all columns present in the final datasets;
- Informed and robust data quality control check standard;
- Techniques for handling missing information like data imputation using Generative Adversarial networks and other statistical methods;
- Tasks using machine learning for enhancing datasets, e.g. time series predictions; and
- A requirements specification document to guide the development of automating data downloads scripts to specify APIs, metadata, database structures, storage capacity, what to download, how often downloads should be done, how to handle big data in the near future.

Protocols to guide water resource managers in aspects of transboundary groundwater data management, such as the type of data to measure (including metadata needs), frequency of measurements, quality assurance and quality control procedures, data sharing, and data storage and visualisation are required.

This could include:

- Aquifer data management
    - Transboundary programmes of groundwater data collection and data management should be implemented gradually, starting with an initial assessment based on gathering, harmonising, and combining existing datasets from the countries, before moving on to joint monitoring.
    - Even though data may be collected through national organisations, transboundary and international institutions such as Lake and River Basin Organisations (L/RBOs) or SADC-GMI are well positioned to provide support to or even lead transboundary programmes of groundwater data collection and data management. Long-term transboundary cooperation requires some degree of formalisation (e.g. a memorandum of understanding, a joint action plan or a treaty) (SADC-GMI, IGRAC and IGS, 2019).
    - Access to sufficient data, information and knowledge is recognized as one of the first conditions to good groundwater governance. So that stakeholders can participate actively in the sustainable management of groundwater resources, data should be made publicly available as far as possible.
    - Visualization platforms are an essential element of any data strategy to create visual representations of large datasets allowing to detect patterns, trends, and outliers in data.
- Data collection and storage
    - Groundwater data requirements differ depending on the specified objectives. Generally data types needed are metadata (the data on data); borehole drilling data; and groundwater levels, quality and abstraction/flow.
    - Data need to be stored in a structured way, in digital formats that can be easily processed to enable efficient and cost-effective access, retrieval and processing for future studies. The choice of database software should be based on the amount of data to be stored as well as available human capacity and skills to manage the data. For transboundary datasets there is a need for a shared storage location accessible by all, without danger of being made country-specific, or being made redundant.
    - To assist in simplified ingestion of existing data to the selected database option, automated data ingestion should be considered (e.g. using excel macros or programming tools).
- Data quality and ownership
    - To ensure the accuracy of data, standard quality assurance and quality control (QA/QC) measures are required. This consist of a series of measures and procedures that must be applied from the beginning of the data acquisition plan. This includes QA/QC of the monitoring network itself, as well as the measured data.
    - A common concern from data providers is retaining ownership of data and receiving appropriate citation for data shared. Readily available licenses like Creative Commons licenses can be used to specify different options in terms of attribution, processing, or application.

9. REFERENCES

Abrahams, E., Gemmell, A., Hugman, R., Flügel, T., Sterckx, A., Ruz Vargas, C. and Mosesane, B. (2020). Data processing tools and manual: Big Data Analytics and Transboundary Water Collaboration – Theme 1: Consolidation of Data and Application of Big Data Tools to Enhance National and Transboundary Data Sets in Southern Africa that Support Decision-Making for Security of Water Resources. Prepared by Ebrahiem Abrahams, Andrew Gemmell, Rui Hugman, Tyrel Flügel, of Umvoto Africa Pty (Ltd.); Arnaud Sterckx and Claudia Ruz Vargas of the International Groundwater Resource Assessment Centre (IGRAC); Badisa Mosesane, University of Botswana. Version 1; Report No. 942/06

Altchenko, Y., Genco, A., Pierce, K., Woolf, R., Nijsten, G., Ansems, N., Magombeyi, M., Ebrehim, G., Lautze, J., Villholth, K.G., Lefore, N., Modisha, R., Baqa, S., McGill, B., Kenabatho, P. (2017). Resilience in the Limpopo Basin: The Potential Role of the Transboundary Ramotswa Aquifer. Hydrogeology Report IWMI, Pretoria, South Africa.

Altchenko, Y., Lefore, N., Villholth, K.G., Ebrahim, G., Genco, A., Pierce, K., Woolf, R., Mosetlhi, B.B.T., Moyo, T., Kenabatho, P. and Nijsten, G. (2016). Resilience in the Limpopo basin: the potential role of the transboundary Ramotswa aquifer. Baseline report. 15 June 2016.

Basson, M.S., Van Niekerk, P.H and Van Rooyen, J.A. (1997). Overview of Water Resources Availability and Utilization in South Africa. Department of Water Affairs and Forestry, Pretoria

Beger, K. (2001). Environmental Hydrogeology of Lobatse; South East District, Republic of Botswana, s.l.:

Brodaric, B., Boisvert, E., Chery, L., Dahlhaus, P., Grellet, S., Kmoch, A., Létourneau, F., Lucido, J., Simons, B. and Wagner, B. (2018a). Enabling global exchange of groundwater data: GroundWaterML2 (GWML2). https://doi.org/10.1007/s10040-018-1747-9

Brodaric, B., Boisvert, E., Dahlhaus, P., Grellet, S., Kmoch, A., Létourneau, F., Lucido, J., Simons, B. and Wagner, B. (2018b). The conceptual schema in geospatial data standard design with application to GroundWaterML2. https://doi.org/10.1186/s40965-018-0058-3

De Mauro, A., Greco, M. and Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics, AIP Conference Proceedings, 1644, pp. 97-104. doi: 10.1063/1.4907823.

Department of Geological Survey Botswana, Environmental Geology.

Gandomi, A. and Haider, M. (2015). 'Beyond the hype: Big data concepts, methods, and analytics', International Journal of Information Management. Elsevier Ltd, 35(2), pp. 137-144. doi: 10.1016/j.ijinfomgt.2014.107.

GEMStat (2020). GEMStat Data submission Guide for Data Providers Portal – User Manual. February 2020. Prepared by UNEP GEMS/Water Data Centre. Version 02.

Global Groundwater Monitoring Network: https://ggmn.un-igrac.org/

Groundwater Markup Language (GWML2). Overview, downloads and official schemas. https://www.opengeospatial.org/standards/gwml2

Hayashi, C. (2013). What is Data Science? Fundamental Concepts and a Heuristic Example, in Hayashi C., Yajima K., Bock HH., Ohsumi N., Tanaka Y., B. Y. (ed.) Data Science, Classification, and Related Methods, Data Analysis, and Knowledge Organization. Tokyo: Springer, pp. 40-51. doi: 10.1007/978-4-431-65950-1_3.

IGRAC (2015). Guidelines for Multidisciplinary Assessment of Transboundary Aquifers. Draft version. https://www.un-igrac.org/sites/default/files/resources/files/Guidelines%20for%20TBA%20Assessment%2020150901.pdf

IGRAC (2016). Ramotswa Information Management System – User Manual. December 2016. Prepared by Arnaud Sterckx and Claudia Ruz Vargas of the International Groundwater Resource Assessment Centre (IGRAC). Version 4.

IWMI. Resilience in the Limpopo Basin: The Potential Role of the transboundary Ramotswa Aquifer. Baseline report – 15th June 2016.

IGRAC (2018). Global Groundwater Monitoring Network Portal – User Manual. May 2018. Prepared by Nelson & Schuurmans for the International Groundwater Resource Assessment Centre (IGRAC). Version 4.7. https://ggmn.lizard.net/media/manuals/Manual_GGMN_UN-IGRAC.pdf

IGRAC, Nelen & Schuurmans (2018). GGMN Portal Instruction Manual. Available at: https://ggmn.lizard.net/media/manuals/Manual_GGMN_UN-IGRAC.pdf

International Water Management Institute (IWMI) (2019). Joint Strategic Action Plan for the Ramotswa Transboundary Aquifer Area. Supported by the United States Agency for International Development (USAID). Pretoria, South Africa: International Water Management Institute.

International Water Management Institute (IWMI) (2020). Project Brief: Transboundary Water Management in Southern Africa – Joint Strategic Action Plan for the Ramotswa Transboundary Aquifer Area.

Lawrie, K.C., Brodie, R.S., Tan, K.P., Gibson, D., Magee, J., Clarke, J.D.A., Halas, L., Gow, L., Somerville, P., Apps, H.E., Christensen, N.B., Brodie, R.C., Abraham, J., Smith, M., Page, D., Dillon, P., Vanderzalm, J., Miotlinski, K., Hostetler, S., Davis, A., Ley-Cooper, A.Y., Schoning, G., Barry, K. and Levett, K. (2012). BHMAR Project: Data Acquisition, Processing, Analysis and Interpretation Methods. Geoscience Australia Record 2012/11. 826p.

Li, S. et al. (2016). Geospatial big data handling theory and methods: A review and research challenges, ISPRS Journal of Photogrammetry and Remote Sensing. International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS), 115, pp. 119-133. doi: 10.1016/j.isprsjprs.2015.10.012.

Miller, H.J. and Hanz, J. (2009). Geographic Data Mining and Knowledge Discovery: An Overview Geographic Data Mining and Knowledge Discovery, second ed. CRC Press, pp. 1-26.

Modisha, R.C.O. (2017). Investigation of the Ramotswa Transboundary Aquifer Area, groundwater flow and pollution. University of the Witwatersrand. Master of Science in Hydrogeology

Pietersen, K., Beekman, H.E. and Holland, M. (2011). South African Groundwater Governance Case Study. Report prepared for the World Bank in partnership with the South African Department of Water Affairs and the Water Research Commission. WRC Report No. KV 273/11, ISBN 978-1-4312-0122-8. 89 pp.

Ramotswa Information Management System: www.Ramotswa.un-igrac.org/

Ranganani, R.T., Gotlop-Bogatsu, Y., Maphanyane, J. and Tladi, B. (2001). Hydrochemical and Geophysical Evaluation of Groundwater Pollution in the Ramotswa Wellfield, SE Botswana. BIE2001 Technical Papers, pp.193-200.

SADC-GMI and IGRAC (2020). Southern African Development Community Groundwater Information Portal – User Manual. Version 1.0, June 2020.

SADC-GMI, IGRAC, IGS (2019. SADC Framework for Groundwater Data Collection and Data Management. SADC-GMI report: Bloemfontein, South Africa.

Sirisha Adamala. An Overview of Big Data Applications in Water Resources Engineering. Machine Learning Research. Vol. 2, No. 1, 2017, pp. 10-18. doi: 10.11648/j.mlr.20170201.12

Staudt, M. (2003). Environmental hydrogeology of Ramotswa. Report by the Environmental Geology Division. Dept of Geological Survey, Lobatse, Botswana

Sterckx, A., Ruz Vargas, C. (2019a). Big Data Analytics and Transboundary Water Collaboration – Theme 1 Quality Control System Report. Prepared by Arnaud Sterckx and Claudia Ruz Vargas of the International Groundwater Resource Assessment Centre (IGRAC). Version 0.1; Report No. 942/29112019

Sterckx, A., Ruz Vargas, C. (2019b). Big Data Analytics and Transboundary Water Collaboration – Theme 1 Groundwater Data Sharing Protocol. Prepared by Arnaud Sterckx and Claudia Ruz Vargas of the International Groundwater Resource Assessment Centre (IGRAC). Version 0.1; Report No. 942/30122019

Sterckx, A., Ruz Vargas, C. and Gemmell, A. (2019). Groundwater Data Sharing Protocol: Big Data Analytics and Transboundary Water Collaboration – Theme 1: Consolidation of Data and Application of Big Data Tools to Enhance National and Transboundary Data Sets in Southern Africa that Support Decision-Making for Security of Water Resources. Prepared by Arnaud Sterckx and Claudia Ruz Vargas of the International Groundwater Resource Assessment Centre (IGRAC) and Andrew Gemmell of Umvoto Africa. Version 1.1; Report No. 942/30122019

Taylor, R., Koussis, A., Tindimugaya, C. (2009). Groundwater and climate in Africa – A Review. Hydrological Sciences 54(4): 655-664.

UN Environment GEMS/Water Data Centre, July 2019. GEMStat Data Submission Guide for Data Providers, Version 01. Available at: https://gemstat.org/data/data-submission/

United States Environmental Protection Agency (US EPA). (2007). Long-Term Groundwater Monitoring Optimization Clare Water Supply Superfund Site Permeable Reactive Barrier and Soil Remedy Areas Clare, Michigan. EPA 542-R-07-010. Available from: https://frtr.gov/pdf/PRB%20and%20Soil%20Remedy%20Areas%20Final%20Report.pdf

World Meteorological Organization (2008). Guide to Hydrological Practices Vol I. 6th edn, WMO-No. 168. 6th edn.

Van der Gun, J. (2018). Data, Information, Knowledge and Diagnostics on Groundwater. In Advances in Groundwater Governance, Karen G. Villholth, Elena Lopez-Gunn, Kirstin Conti, Alberto Garrido, Jac Van Der Gun (eds). CRC Press, Leiden. ISBN 9780367890100.