MASTER IN MANAGEMENT

MIM

MASTERS FINAL WORK

PROJECT

ENHANCED INFORMATION SYSTEMS AND DATA VISUALIZATION: DEVELOPMENT OF A TOOL FOR TRANSPARENT AND PARTICIPATORY WATER GOVERNANCE

LUCAS FARIA CHAUD

MARCH 2022

Master in Management
MIM

Masters Final Work
Project

Enhanced information systems and data visualization: development of a tool for transparent and participatory water governance

Lucas Faria Chaud

Juri
President: José Manuel Cristovão Veríssimo
Supervisor: Marta Pedro Varanda
Members: Winnie Ng Picoto

March 2022

ABSTRACT

There are several consequences to groundwater of overexploitation and intensive irrigation, such as depletion of water tables, saltwater intrusion, and pollution, which in addition to the decrease in rainfall, will pose even greater pressure on underground water resources, threatening food security and soils sustainability. Challenges to mitigate these risks vary from technical hydrogeological approaches to social engagement of stakeholders in participatory governance of water. To assist with the latter, and to programmatically provide inputs to the prior, in the regional area of the Campina de Faro Aquifer, in south Portugal, this project proposes a practical approach in building a technological framework to regular extract and concatenate data from multiple sources, providing meaningful insights through interactive visuals to engage stakeholders in participatory governance, as well as offering programmatic connectivity from its database to the research community. Further pathways to advance the tool to a production level with operational guidance are also discussed. Moreover, in consonance with a global effort for positive impact, this work relates to the United Nations' Sustainable Development Goal 6 – Clean Water and Sanitation for All.

Keywords: ICT, Data Management, Data Visualization, Cloud Infrastructure, Non-relational Databases, GIS, Groundwater, Participatory Management, SDGs, Sustainable Development Goals

INDEX

LIST OF FIGURES

1 INTRODUCTION

1.1 Project context and motivation

The project described in the next chapters as the main outcome of the final work of the Master in Management programme was inspired by, and is an unfolding development to, eGroundwater, a citizen science and ICT-based enhanced information systems project, part of the Partnership for Research and Innovation in the Mediterranean Area (PRIMA) programme, supported by the European Union's Horizon 2020 funding for research and innovation. The project initiative is also aligned with the global effort of the United Nations and the Sustainable Development Goals (SDGs), which presents a blueprint to achieve a better and more sustainable future for all, especially with the SDG 6, that specifies context, actions and measures to assure clean water and sanitation.

The overall objective of eGroundwater is to support sustainable participatory groundwater management through the design, testing, and assessment of enhanced information systems. The project aims to explore four different cases of study in the Mediterranean region: Algeria, Morocco, Spain, and Portugal. For the Portuguese use case, ISEG and The University of Algarve are the associated partners responsible for engaging stakeholders in the co-development of sustainable groundwater, integrating citizen science and ICT-based tools, in the Campina de Faro aquifer system, located in Algarve, Portugal’s southernmost region.

The Campina de Faro aquifer system is classified as poor groundwater quality status in the East and poor quantity status in the West. The Easternmost system of the aquifer is an area of significant agricultural pressures which lead to high input of nitrates in the aquifer during the last 50 years. On the other hand, the Westernmost system of the aquifer is a highly touristic area, famous for its golf courses, and has shown decreasing trends on groundwater levels over the last 20 years, and consequently, seawater intrusion (*Campina* *de Faro Aquifer | EGROUNDWATER*, n.d.). The main challenges in the region reside in raising awareness for the critical situation of water in Algarve, while invoking collaboration among private users, public authorities, and researchers towards thinking in innovative, structural, and participatory solutions for better water management to circumvent scarcity and low quality of water.

The idea of using technology and data visualization instruments for broadening and enhancing participation of different actors, especially civil society, dates back to the late ‘80s, when the ‘‘citizen panels’’ system was introduced by Dienel (1989), involving decision makers, experts, and citizens, in an attempt to unleash the public from being passive spectators, but taking part of decisions. A lot has been developed since then, not only in data collection and processing techniques to deal with the unprecedent growth on volume and variety of data availability, but also with visualization mechanisms to make the information understandable, attractive and, most important, accountable. In 2009, the Open Government Initiative launched by Barack Obama’s administration catalyzed the use of dashboards in several North American federal government agencies, embracing the principles of transparency, participation, and collaboration, while facing the challenges of, at external accountability level, being exposed to public scrutiny. In this context, it is paramount that both the dashboard performance measures, and the underlying data need to be publicly accessible for credible accountability (Ganapati, 2011).

Aligned with the principles described above, the Portuguese Environmental Agency through the Water Institute (INAG), in 1995, launched the National Information System of Hydric Resources (SNIRH), with the objective to process, validate and disseminate all information collected in the monitoring networks of the Agency and other entities. The system includes, in addition to the Database, multiple tools, from graphic and statistical analysis to the integration of simulation models and Geographic Information Systems (GIS), and it was awarded, in 1997, by the Portuguese Computing Institute (Prémio Descartes) which distinguishes original works in the field of Computer Science and the use of Information Systems and Technologies, with innovative aspects and relevant applications for society in general (Marques et al., 1999). However, while being extremely rich in data, the web portal and front-end provided by the SNIRH and available for the general public, is technologically outdated and lacks modern interactive, and flexible, interfaces in order to attract non-specialized audience and broadcast meaningful information. The contrast and duality of data availability and meaningful information is going to be central in the following chapters and it was paramount in the development of this project.

1.2 Objectives and project relevance

The central proposal of the project resides on building a technological framework to gather, post-process and centralize data related to the Campina de Faro Aquifer from relevant data sources, including time series from SNIRH and the future outcome database of eGroundwater, bonding and enabling the development of meaningful and interactive online visuals to engage stakeholders in participatory governance. Furthermore, it aims to provide easy connectivity for programmatic and non-programmatic access of the underlaying data for university students, researchers, among others, unleashing potential for enhanced technical analysis that depends on groundwater measures in the region.

The overarching objective is to create a low-cost go to place to find all relevant information regarding the Aquifer, creating long lasting relationship with the community. Thus, it is paramount that the project continues to be in constant evolution, opened to new possibilities and expanding its capabilities regarding data connections (e.g., soil moisture, overland flow, weather forecast, etc.). Although the details of these possible expansions are not being detailed in this work, proposed ways forward are discussed in the Conclusion chapter.

1.3 Project structure

This work is divided into six chapters. Introduction, the first, unveils the context, motivation, relevance, and objectives of the project. The second, Literature Review, brings a brief reflection and analysis on how the use of information and communication technologies as well as data visualization have been used to create a culture of transparency in participation processes and policy making, enriching water governance management. In the following chapters, a deep dive in the tool developed, as a product of the project, is provided: Chapter 03 is dedicated to the design and implementation of the infrastructure to host the database and the automatic extract-transform-load (ETL) mechanism, the available data format and the logic to web-scrapping SNIRH, the database structure including its collections and data-types, the programming languages and libraries to be used, the ODBC connector for the data visualization tool, the visuals to be constructed, and, finally, considerations regarding quality assurance and control, security and administration. Chapter 04 is devoted to the operational challenges of the solution as well as key success factors and recommendations for the tool beyond the time constraint of this project. Lastly, Chapter 05 concludes the work with achievements and proposed ways forward to create broader impact.

2 LITERATURE REVIEW

2.1 The use of technologies to create a culture of transparency

Information asymmetry can be defined as the imbalanced relation in which one party has more information than another party. Not only the impact in the financial sector and implications of imperfect information were considered in the Nobel Laureate Joseph Stiglitz (1990) in his revised Theory of The Firm, but also consequences regarding moral hazards, monopoly of information, and adverse selection have been extensively discussed (Aboody & Baruch, 2000), with a clear understanding that while there is asymmetry there is no full transparency. Although not being the focus of this project, corruption, defined as the abuse of entrusted power for personal gain or specific group interest, and therefore, a product of information asymmetry, has been in the heart of the development and use of new technologies to promote transparency. This is especially true when considering the latest public approaches to the adoption of e-government model, which is recognized as a tool to reinvent the public sector by transforming internal government work processes and external relationships with citizen participation (Chul et al., 2008). Hence, any available mechanism that one can screen for information to reduce the impact of asymmetric information on decisions, will also benefit public engagement towards promoting good governance, strengthening collective-oriented actors, and reducing the asymmetry (Marwala & Hurwitz, 2015).

As a result of the above, achieving accountability through transparency is also a key outcome to succeed in chasing collaboration and engagement. However, it must be accompanied by the possibility to intervene, as only viewing data without having the possibility to take actions based on the result has limited use and might result in abandonment and disengagement (Matheus et al., 2020).

2.2 Environmental and participatory decision making based on data visualization techniques

Geographic information systems (GIS) are the most common structures used to analyze and display information of spatially distributed phenomena for the purpose of inventory, decision making and/or problem solving (Goodchild et al., 1993). Nevertheless, the notion of GIS used as an enabling technology to enrich public participation in local decision making affecting the use of land and other natural resources is still under development. There is still a certain level of skepticism that adds to the hurdle of further development, under the superficial notion that GIS knowledge is constrained to experts and institutions, while using everyone's data.

In an effort to unveil this notion, Jankowski (2009) discusses how visualization of spatial data can empower groups of people to participate in decisions shaping their communities and promoting sustainable use of natural resources, as well as how GIS becoming a tool of participatory democracy could be used to reinvigorate traditional models of citizen involvement in making decisions about the use of public, natural resources, more specifically to public participation in water resources planning. Sieber (2008) has established a framework to evaluate the breadth and depth of current and future of participatory GIS, showcasing how it incorporates local knowledge, integrates, and contextualizes spatial information, allows participants to dynamically interact with input, analyzes alternatives, and empowers individuals and groups.

Al-Kodmany (1999) analyzed the impact of different visualization tools in different phases of a participatory planning process in Chicago's Pilsen neighborhood, attesting the effectiveness of the tools in empowering residents to plan and design for the future of their community, progressing toward the combination of computer-based visualization techniques with the art of designing with the people. The benefit of visual analytics, visualization and geo-visualization techniques in several steps of the participatory processes is also highlighted by Marzouki et al. (2017), demonstrating the impact of using technology to represent citizens' living context through spatial, temporal and semantic dimensions.

As a pragmatic balance, Meitner et al. (2005) and Sheppard (2001) throw lights on the fact that visualizations alone are not enough for enhanced decision-making, and that significant risks may be incurred from inappropriate applications. Testing, measuring, and verifying the benefits and risks of using visualization tools in decision support contexts must be a constant throughout the process.

2.3 Practical design and implementation of web-based data management

The review of the literature available for practical approaches to design and implement web systems as a supporting tool for enhanced water management is focused in four main projects: The first, and more technologically advanced, is the Jeferson Project at Lake George, in the New York state protected Adirondack region, and counts with the collaboration of scientists, engineers and technologists from IBM Global Research and Rensselaer Polytechnic Institute, combined with the advocates and program managers at the Lake George Association (LGA). Then, the aspects of the development of a 3D Water Atlas for Surat Basin, a geological basin in eastern Australia. Still in Australia, but for a much broader area, challenges, and lessons of designing and deploying a web-based groundwater data management system are discussed. And, finally, the development of a visualization tool for integrated surface water and groundwater modeling in the Heihe River Basin, in China.

2.3.1 The Jeferson Project at Lake George

The goal of the project is to understand and anticipate the effects of road salt, invasive species, nutrient loading and runoff, and insecticide impacts on wetland communities, decreasing Lake George water quality (Stoler et al., 2017). To be able to accomplish those, knowledge around data gathering and integration mechanisms to link spatial data and hydrologic network was produced, including the development of new programming packages in R language, to evaluate the spatial data quality for the lake and stream sites available through the U.S. Water Quality Portal (Winslow et al., 2018).

The project counts with the application of Internet of Things to build a Smart Sensor Network, which incorporates over 50 sensor platforms with more than 500 sensors, sending live data, including water depth, temperature, salt, and weather conditions all over the lake, to be post processed and modelled at a cyberinfrastructure, with the capacity to collect more than nine terabytes of data per year that in turn generate 73 terabytes of data by computer models. The result is then displayed in an interactive dashboard publicly available, enabling live and historical data exploration, that is being used to inform science-to-solutions approaches to protect Lake George (*The Jefferson Project Data Dashboard at Lake George*, n.d.).

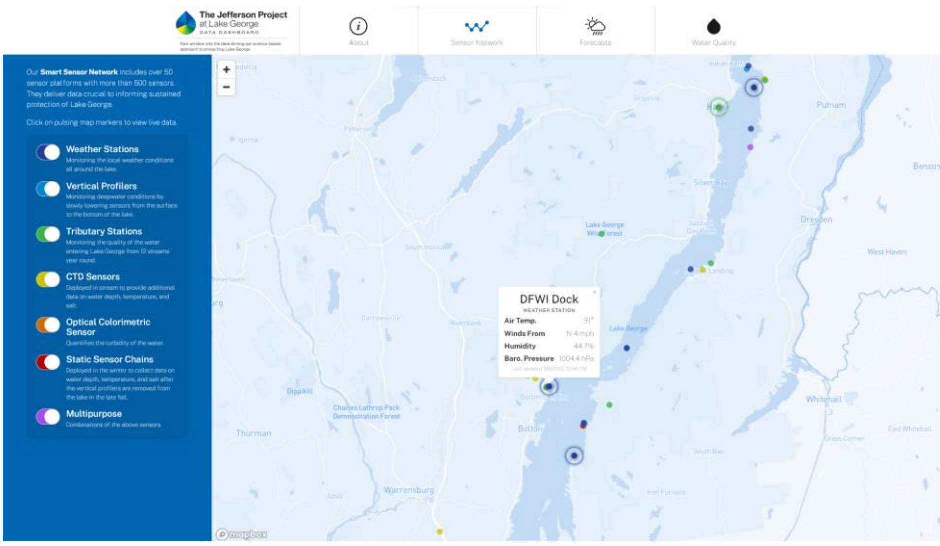


Figure 1 - The Jefferson Project Smart Sensor Network Dashboard

Claiming to have built "The Smartest Lake in the World", the project fostered over 60 scientific publications and various international congress participation, empowering breakthrough research on globally significant issues impacting watershed ecology, including harmful algal blooms, invasive species, and the compounding effects of climate change (*The Jefferson Project: Building The Future of Freshwater Protection | Lake George Association*, n.d.), while raising public awareness and increasing the perception of their role in preserving the Lake.

2.3.2 3D Water Atlas

Aligned with the purpose of broadening the access and decentralizing the capabilities of presenting three-dimensional data from few private software companies, 3D Water Atlas combines the use of an open-source web-based platform, tailored with common programming languages, including HTML, PHP, Python and JavaScript, with a PostGreSQL database and PostGIS, for storage and displaying of spatial information. Groundwater data from several sources across the Surat Basin, relating mainly quantity and quality of water, is ingested and mapped to a common model.

The information is then presented in a 3D scene for web access, where Groundwater bores are shown against a map background, with thematic visualization of the depth of the bore and additional information about can be accessed through a pop-up window, with tabs for different information and enhanced diagrams and temporal plots, enriching and facilitating the user experience when navigating through the tool.

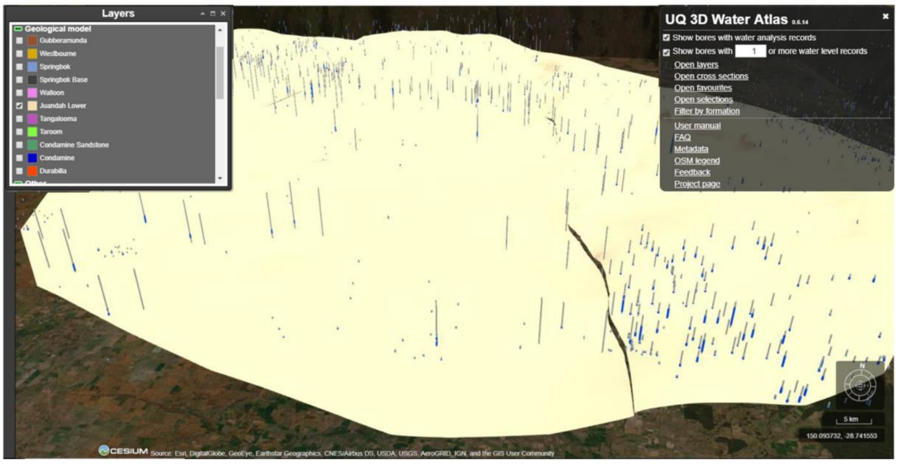


Figure 2 - View of the 3D Water Atlas showing groundwater bores and a layer of the geological model

According to Wolhuter et al. (2020), the ability to view and explore the data that goes into groundwater models via the 3D Water Atlas increased the community confidence in groundwater modeling and can potentially influence management decisions based on the modeling results. The project also convenes the increasing need for aggregational platforms that bind different data sources and provide a new understanding of groundwater systems, giving visibility to management decisions and identifying possible interventions. As discussed in previous sections, the augmentation of public users on the knowledge of the system, contributes to elevating trust in the water management process and promotes transparency.

2.3.3 Web-based groundwater data management system in Australia

The paper by Iwanaga et al. (2013) depicts the development process through different phases of a web-based system built for the National Centre for Groundwater Research and Training (NCGRT) in Australia, exploring the minutiae since requirements identification and analysis, to system design and implementation of groundwater cyber infrastructure at six sites throughout the country, while ensuring connectivity with Australia’s existing National Groundwater Information System (NGIS). The lessons learnt from this practical approach is especially useful for one of the main objectives of the project defended by this Final Work, which is aggregate already collected and publicly available data from national systems to a recent designed framework.

Figure 3 - NCGRT DMS - Sensor information being displayed on selection of a map feature

One key aspect unveiled by the authors is that clear and constant communication among all stakeholders of the project (developers, data providers, and end users) is paramount to a successful and smoother development, as well as more appealing for future adoption of the proposed tool. Despite being quite an expected conclusion, it is easy to be overlooked especially by the team building the infrastructure and the pieces of code, where the speed of development cannot be an isolated measure without quality and applicability.

2.3.4 Heihe River Basin Tool

Another outstanding practical example of integrating existing tools and data sources to build exceptional visualization systems can be found in the work of Tian et al. (2016), where all the scientific data of Heihe River Basin, in China, relating surface water and groundwater were brought together into a comprehensive system, called IHM3D, facilitating the interpretation and validation of modeling results. Although the end product of the aforementioned work is much more targeted to hydrogeologist specialist rather than public audience, it presents valuable insights in the process of designing, deploying and managing visualization systems for nature-based purposes with largely different data structures. Furthermore, promoting data and model sharing in the water resources research community, to facilitate new studies in understanding the hydrological cycle in a given region, is also a desirable outcome of this Master's Final Work.

The figure below illustrates the design principle and architecture of the tool and the variety of information that could be bonded to build the desired visualization, which is quite inspirational for further development and achievement goals of this proposed Final Work. It encompasses a base terrain 3D layer, GIS layers, model inputs and outputs, including model spatial structures (computational grids, subsurface cross-sections, drainage systems, pumping well locations), as well as distributed parameters (soil features, hydrogeological properties, river characteristics) and meteorological data:

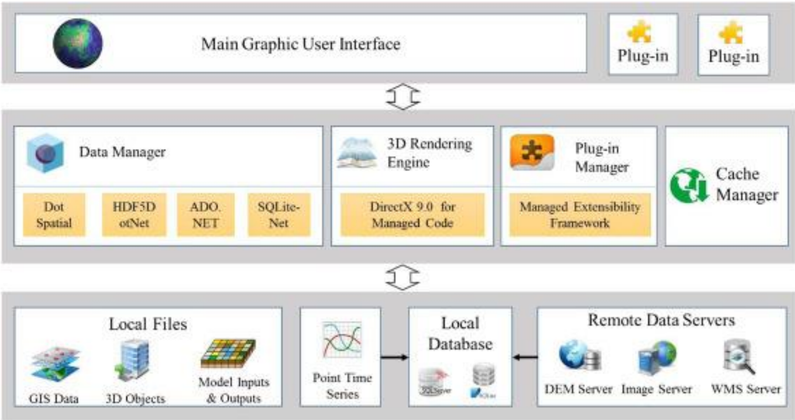


Figure 4 - System architecture of IHM3D

3 Design and Implementation

The upcoming chapters will be dedicated to the actual design, implementation and operations of the tool that was developed for this project. Details for each element can be found in each corresponding section and topics that are still not fully deployed or are moonshots to be achieved in an ideal plan for the near future, will be then discussed in the Conclusions and Proposed Way Forward chapter.

3.1 Methodology

The initial concept and discussions of which sets of data could be potentially useful to the broader community to access, understand and be provoked by the intention to monitor and protect their local natural resources, happened in the context of the eGroundwater project and specifically under the project’s work package number three, which aims at designing, deploying and testing a platform to provide more efficient and sustainable groundwater management. The platform should include a smartphone application to easily sharing and visualizing data. It also should be tailored to the case studies in Morocco, Algeria, Portugal and Spain. To achieve this central goal, the project’s work package has 5 sub-objectives:

- Engaging stakeholders in the project, to ensure that the tool responds to their needs.
- Developing a shared vision of groundwater resources, use and management issues between stakeholders and scientists of the project team; identify information gaps; identify innovative tools that can be deployed to bridge those gaps; and develop terms of reference for the ICT platform to be developed in each case study.
- Developing a generic prototype ICT Platform to collect GW data provided by farmers (citizen science approach), earth observation technologies and groundwater sensors.
- Developing models and routines to assess and forecast irrigation water needs through remote sensing (drones and satellite images), crop models and weather forecasts
- Co-developing with stakeholders, implementing, assessing and improving the prototype in case studies

Within the work package development, a series of meetings with the technical team involved in eGroundwater, centrally orchestrated by the Bureau de Recherches Géologiques et Minières from France, with local participation of ISEG and the University of Algarve, were held to discuss and build the possible relevant scenarios to the Portuguese use-case. While not reproducing the completing minutes of each meeting, the following dates can be appointed as key in the discussion of data sources available and decisive to the objective of this Masters Final Work:

- Initial approach and overall vision: September 18, 2020
- In-depth discussion of the proposed diagram for connecting data-sources in other case studies: October 12, 2020

- In-depth session for the evolution of the Work Project 3 among all case studies: December 15, 2020
- Joint session with the Portuguese Environmental Agency: April 8, 2021
- Webinar session with Campina de Faro stakeholders, named "Conversas C@mpina - Sharing the voices of the solution ", with the objective of explaining the state and use of the aquifer, giving rise to reflections on the need to advance the discussion of solutions, and how it affects the stakeholders: June 17, 2021

3.2 Proposed Architecture

Considering the project aspiration and the various inputs collected during the above conversations, the project described in this Final Work was then designed. The picture below shows the high-level architecture for the solution proposed, depicting the main links between data sources and systems, in different environments, to illustrate the planned information flow:

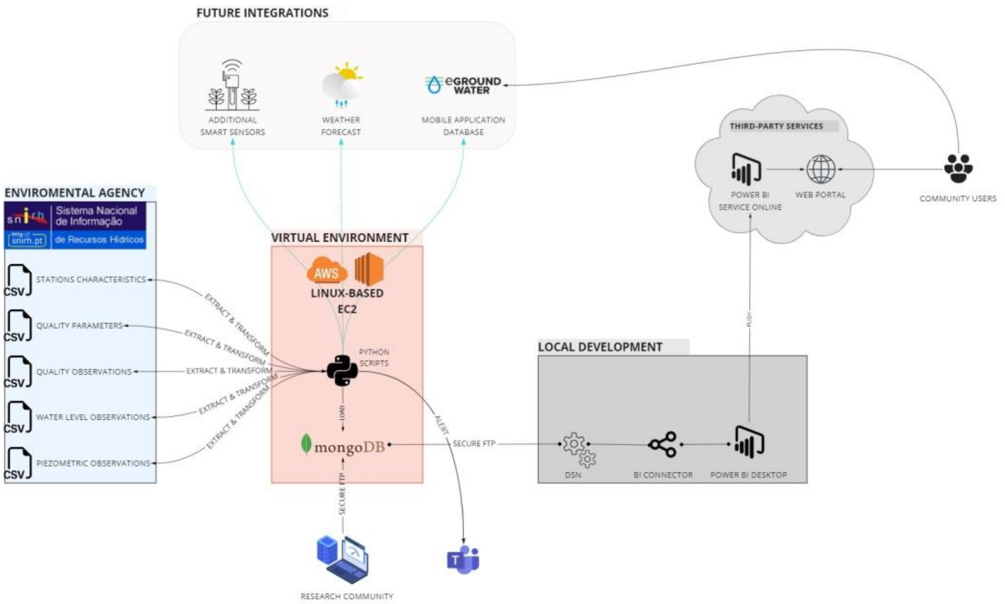


Figure 5 - Tool High-Level Architecture

3.3 Environmental Agency – Data Availability and Connectivity

As briefly described in the introductory chapter, Portuguese Environmental Agency provides access to water measurements taken by public agents in the National Information System of Hydric Resources (SNIRH) tool. This section will discuss the type of information available, the historical series and frequency of update, how the portal currently provides access to the underlaying data and the key elements for the object of study, the Campina de Faro aquifer's quantity and quality of water.

3.3.1 Stations Characteristics Data

The most fundamental element to collect and store for further data correlation, unlocking more precise classifications of portions of the aquifer, is the location of the stations where measures are and were taken, accompanied by basic identification items. These registers can be extracted from SNIRH, where there are 33 stations indexed for Piezometric and Water Level measures and 26 stations indexed for Groundwater Quality measures, representing 57 distinct stations of interest within the boundaries of the Campina de Faro Aquifer.

For both sets, the following information can be found:

"CODIGO": The Station Unique Code Identifier
"NOME": The Station Name Identifier
"DISTRITO": The district where the Station is located
"CONCELHO": The council where the Station is located
"FREGUESIA": The parish where the Station is located
"BACIA": The basin where the Station is located
"ALTITUDE (M)": The altitude in meters where the station is located
"COORD_X (M)": Longitude of the Station following the Portuguese Military Coordinates' System
"COORD_Y (M)": Latitude of the Station following the Portuguese Military Coordinates' System
"SISTEMA AQUIFERO": The Aquifer system where the station is located
"ESTADO": The working status of the Station (if in service, abandoned, destroyed etc.)

It is worth noting that, for the BI Tool designed to be used in the project, all coordinates need to be transformed to the World Geodetic System (WGS84), in a decimal degree format. The transformation method will not be discussed in this work, but its result will be carried by two additional fields in each station register (DEC LAT and DEC LONG).

Updates to this type of data are not periodic nor frequent, hence, programmatic access and Extract-Process-Load (ETL) mechanisms to pull this specific data will not be the focus, even with the poor user experience to access and download (in form of comma-separated-value files) the information from SNIRH.

3.3.2 Water Level and Piezometric Data

An important distinction to made regards the definition of Piezometric Level and Water Level of an aquifer. Piezometric Level is related to the level at which the water in an aquifer is at atmospheric pressure. It coincides with the water table of a free aquifer. In confined aquifers, the piezometric level is higher than the aquifer ceiling, and there may be areas where it is higher than the topographic surface. Water Level is then defined as the height of the free surface of a body of water in relation to a reference plane, usually the altitude of the station related to the sea level. While piezometric measures are most used by hydrogeologists to map and understand the behavior of an aquifer, relative water levels resonate better with the broader community, given the common understanding of the lower the values, the worst.

Both measures are available to the same stations and at same extent, which is possible considering that the altitude of each measurement point is known. For these series, SNIRH has impressive historical measurements, that date back to November 1973 and it is being continuously updated until today. With very few exceptions, the measures are taken every month by public agents in the active (in service during the correspondent time) boreholes and then uploaded to the system, also in a monthly basis.

Methods and techniques to web-scrap the data will be discussed in the dedicated Python section further in this Work, however, it is important to point out the current critical limitation of SNIRH in exporting only 50 registers at a time. Which is quite cumbersome for regular users that would like to work with the data or simply having an overview of the status of those parameters in a given aquifer system. For example, even for a single month, in the current system, it is not possible to gather the piezometric and water quality measures for all the 33 boreholes in Campina de Faro at once, since this would lead to 66 measures, therefore above the current limitation.

The desired output is a simple register per location, with the timestamp of the measure, the target measurement station (borehole), and the piezometric value and water level in meters.

3.3.3 Water Quality Data

Regarding Water Quality, SNIRH has keeps record of 107 parameters for each collection over the set of 26 stations. The first record dates to April 1995 and the most recent was input into the system by April 2021. However, frequency of new inputs is not well defined nor the coverage of boreholes in each measurement set.

A complete list of the quality parameters is provided in the Annex. The key indicators chosen in partnership with eGroundwater's specialists, which can cover for the most relevant classes for quality (salinity, light metals, microbiology, pesticides, and fertilizers), were: Aluminum, Total Arsenic, Total Lead, Fecal Coliforms, Total Cadmium, E. Coli, Phosphate (mg/L PO4), Total Mercury, Total Pesticides, and Field pH. The desired output would also contemplate the timestamp of the input and the target station from where the quality analysis relates to.

3.4 Virtual Environment

Having the tool infrastructure hosted in a public cloud was the preferred option to carry this project, creating a long-lasting engine that can be perennialized after this project is over. High level discussions regarding cloud security and network constraints, if this project is incorporated by any institution, are further presented in the corresponding sector.

Due to the experimental character at this stage of the Work, free-tier services of public cloud providers were the best option to proceed, to not carry-on additional costs without demonstrating all the capabilities and potential of the end product. Hence, Amazon Web Services (AWS) was chosen as the cloud infrastructure provider. AWS is one of the world's most comprehensive and broadly adopted cloud platform, offering hosting options from data centers globally, besides a 12-month free tier of resizable compute capacity, limited to 750 hours per month, is also available. For this project, AWS region eu-central-1, in Frankfurt, was the preferred data center to host the solution, once the latency to Portugal is likely to be lower than other regions.

3.4.1 Linux Elastic Computing Instance

Regarding computing capacity, a t2.micro instance was spun-up for the project. T2 instances are one of the lowest-cost options and are ideal for low-latency interactive applications, small and medium databases, virtual desktops, development, build and stage environments, code repositories, and product prototypes. T2.micro falls under the 12-month free tier and counts with one virtual central processing unit (vCPU) and 1 Gibibytes (GiB) of Random Access Memory (RAM). Attached to the instance, AWS offers a generic purpose Elastic Block Store (EBS) of 8 GiB, with a 1 GB of snapshot storage. Further optimization and improvements suggestions are made in the Operations Chapter, such as building a redundant instance, configuring a load balancer and on-demand capacity building.

As for the Operational System, a long-term service version, the Focal Fossa 20.04, of Linux Ubuntu was selected as the best option, as it was launched in 2020 and has extended security maintenance guaranteed until 2030, besides the fact that Ubuntu distribution of Linux is made available freely by the community to the community, honoring the origins of the African concept of ‘I am what I am because of who we all are’, that gives name to the distribution.

3.4.2 Data Extraction, Transformation, and Loading with Python

Python was the programming language chosen to perform the Extract-Transformation-Load functionality of the solution. Therefore, Python 3.8.10 release was installed in the computing instance, and in order not to waste instance capacity, the inbound rules in the security group was adjusted to allow SSH (Secure Shell protocol) on port 22 from a local environment IP, enabling local development with remote execution.

3.4.2.1 Historical Data

The first programmatic task was to circumvent SNIRH data download limitation, without losing information, and gather all the available measures since 1973. To be able to execute it, a web-scrapping technique was used within a loop routine throughout the years for all Stations for Piezometric Values and Water Levels. For Quality Measures, to collect all the 107 parameters for every 26 stations during all the years, the approach was also to web-scrap the data but looping first each parameter for a given station in all years, and then loop all the stations where measurements were available.

The repetitive web-scrapping emulates hundreds of human queries in SNIRH to pull the underlaying data and can be achieved through the composition of a target URL that SNIRH uses to provide csv files to users. To better understand the syntax and logic of the URL, the address can be decomposed by the following:

$$Target\ URL = base\_URL + stations\_of\_interest + measurement\_parameter + time\_window + format$$

Where:

- base_URL: A fixed SNRIH string, defined by the following snippet "https://snirh.apambiente.pt/snirh/_dadosbase/site/paraCSV/dados_csv.php?sites="
- stations_of_interest: List of stations to be queried, separated by a comma
- measurement_parameter: Defines which measurement to be queried, being:
  - Piezometric Value: pars=100290981
  - Water Level: pars=2277
  - Quality: Individual values to be looped for each one of the 107 parameters. A complete list of these values is also available in the Appendix.
- time_window: Initial (tmin) and final (tmax), in dd/mmm/yyyy format, dates to be queried, also to be looped.
- format: Indicates the format to pull the information, in the desired case, csv format.

A simple representation to be written in code, representing in bold the variables to be looped, can be showed as the following:

$$url = "https://snirh.apambiente.pt/snirh/\_dadosbase/site/paraCSV/dados\_csv.php?sites = " + \boldsymbol{stacd} + "\&pars = " + \boldsymbol{parid} + "\&tmin = " + \boldsymbol{isdate} + "\&tmax = " + \boldsymbol{ifdate} + "\&formato = csv"$$

After each iteration of the loop, the information is then transformed and temporarily stored in a data frame available through Python's pandas library, for a more comprehensive understanding of the information and eliminating unwanted columns and rows. Finally, Python's pymongo library is used to grant the capabilities needed to insert the records in the corresponding MongoDB collection. Details on how the database is configured and the schemas set to the collections, are discussed in section 3.2.3.

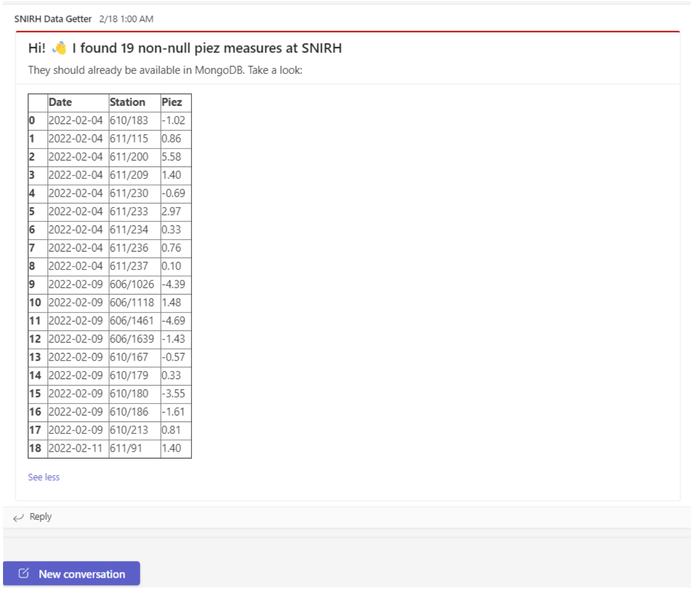
The data gathering process described above results in: 2.7M registries with Piezometric data and Water Level data, and 998 documents with Water Quality data, comprehending all data available in SNIRH for the Campina de Faro Aquifer, until the date the queries were executed.

3.4.2.2 Daily Routine for Assessing New Data

Having the database collections filled with the historical data, the task to be completed next was to create a daily routine to assess SNIRH and look for new piezometric, level or quality data, compared to the information already stored in the database. If a new input is found, the data is pulled, transformed to the same format of the historical data, and appended to the corresponding collection in the database.

Therefore, the logic is quite similar to the one described before, to web-scrap any new data available, and is not going to be described again in this section. The repetitive routine is then achieved through Linux Cron functionality, which is a built-in daemon Linux utility that runs processes on the system at a scheduled time. For this project, three routines were scheduled to check and collect piezometric, water level, and quality measures, offset in execution by one hour each to avoid concurrency of resources, besides a monthly backup of the database, on the first day of each month.

Additionally, whenever a new input is found, a webhook from the virtual environment to a given Microsoft Teams' team is triggered, using Python's pymsteams library, to let any stakeholder, which can be added to the Team as Guest user, that there's new data available in the database. This is especially useful for the research community that can be warned, with shorter delay, when new measures are posted, instead of having to regularly browse SNIRH, considering that there is no official date for the measures to be published.



	Date	Station	Piez
0	2022-02-04	610/183	-1.02
1	2022-02-04	611/115	0.86
2	2022-02-04	611/200	5.58
3	2022-02-04	611/209	1.40
4	2022-02-04	611/230	-0.69
5	2022-02-04	611/233	2.97
6	2022-02-04	611/234	0.33
7	2022-02-04	611/236	0.76
8	2022-02-04	611/237	0.10
9	2022-02-09	606/1026	-4.39
10	2022-02-09	606/1118	1.48
11	2022-02-09	606/1461	-4.69
12	2022-02-09	606/1639	-1.43
13	2022-02-09	610/167	-0.57
14	2022-02-09	610/179	0.33
15	2022-02-09	610/180	-3.55
16	2022-02-09	610/186	-1.61
17	2022-02-09	610/213	0.81
18	2022-02-11	611/91	1.40

Figure 6 - Teams Webhook Bot, letting users know that new piezometric measures are available for 19 stations, the date they were taken and that was published in SNIRH at February 18, 2022, the same date of the post on Teams.

A copy of all the python scripts used can be found, under request, in the following GitHub repository: *https://github.com/lucaschaud/eGW.git*.

3.4.3 Non-Relational Database – MongoDB Structure & Connectivity

MongoDB Community Edition was chosen to be this project database provider, once it is a free, open-source database software that is being widely used in modern applications. It is a non-relational database, which means that it stores data in a non-tabular document-based structure, enhancing the side-by-side digestion and organization of various types of data, making non-relational databases much more flexible than relational databases.

For this project, MongoDB 5.0 Community Edition was installed in the Linux Computing Instance using the default apt package manager from mongodb-org, once it is the latest and most stable version supported by Ubuntu 20.04. By default, MongoDB uses the Transmission Control Protocol (TCP) port number 27017, hence, this is the port in which python scripts will be connecting to read and write data. Additionally, to allow remote connectivity this port was also opened within the inbound security rules in AWS console management.

Remote connectivity is paramount for raising the use of the solution by the research community, for different purposes data collection and customized interests, once they can use remote MongoDB structure with their local environments, or even having a local copy of Mongo Compass, an interactive graphical user-interface (GUI) tool for querying, optimizing, and analyzing MongoDB data, that allows users to easily get key insights, drag and drop to build pipelines in a single, centralized interface.

As already mentioned, MongoDB stores data records as documents (specifically Binary JavaScript Object Notation documents) which are, then, gathered in collections. A database stores one or more collections of documents. For this project, one database named “eGW” was configured, with six main collections to store documents:

- [Observations] Piezometers: To store piezometric measures of every station available. Each document has an Object ID, the timestamp of the measure, the Station ID and the value of the piezometric value.

- [Observations] Quality: To store quality parameters measures of every station available. Each document has an Object ID, the timestamp of the measure, the Station ID and 107 parameters with the corresponding value.

- [Observations] Water Level: To store piezometric measures of every station available. Each document has an Object ID, the timestamp of the measure, the Station ID and the value of the water level value.

- [Parameters] Quality: To reference and associate all the quality parameters with an ID used by SNIRH. Each document has an Object ID, the quality parameter ID, the quality parameter name.

- [Stations] Characteristics: To reference all the unique 56 measurement stations/boreholes. The structure of each document has been shown in section 3.1.1.

- [Stations] Quality: A list of the stations that are exclusively used to perform water quality parameters measurements. Each document has an Object ID, a Station Code and a Station ID.

The snippet below showcases an example of how easy researchers can query a segmented and target data from the remote database using Python, without having to worry about local database management or consuming local resources. The example returns the data related to the latest measures of the quality parameter ‘pH – campo’ for four different stations '606/1019', '606/1089', '606/434', and '610/168':

```
from pymongo import MongoClient
import pandas as pd

client = MongoClient ('mongodb://user:password@serverip:27017/')
db = client.eGW
col_par = db['[Parameters] Quality']
col_obs = db['[Observations] Quality']
col_car = db['[Stations] Characteristics']


vpar = 'pH - campo'
df = []
lst_sta = ['606/1019','606/1089','606/434', '610/168']

for i in range(len(lst_sta)):
    var_sta = lst_sta[i]

    filter = {'CODIGO': var_sta}
    result = col_car.find(filter=filter)

    for doc in result:
        vlat = doc['DEC LAT']
        vlong = doc['DEC LONG']

    filter={'Station ID': var_sta}
    sort=list({'Date': -1}.items())
    result = col_obs.find(filter=filter, sort=sort, limit=1)

    for doc in result:
        vobs = doc[vpar]
        vdate = doc['Date']

    df.append([vdate, var_sta, vlat, vlong, vobs])

df = pd.DataFrame (df, columns=['Date', 'Station', 'Latitude', 'Longitude', vpar])


Out:

Date      Station                     Latitude          Longitude         pH - campo
0         2020-10-27 10:20:00606/1019 37.06109194       -8.031378611      7.1
1         2020-10-27 09:30:00606/1089 37.05767333       -8.064763333      7.6
2         2020-10-27 12:00:00606/434  37.06136611       -7.969786667      7.1
3         2020-10-27 11:32:00610/168  37.04109167       -8.015466111      7.4
```

Further developments to build a user-friendly frontend customized for researchers are not going to be discussed in this Work, but it is certainly a useful unfold of the project.

3.5 Local Environment

Local Environment is going to be used only as an intermediate, temporary environment to host Power BI Desktop, as a dashboard builder tool, and to publish the visuals to an Online Service, passing the credentials along the publication to allow a direct connection from the database to the online dashboard, permitting constant refresh. Hence, it was not worth for the project to build a dedicated environment to this purpose, nor waste valuable computing resources in the existent cloud instance.

The local environment used was a simple Windows 10 workstation with 64-bit operating system, x64-based processor with 4 GB of installed RAM.

3.5.1 DSN and ODBC BI Connector

As shown by Figure 5 and mentioned above, Power BI was the selected tool to build the dashboards and visual reports to the project. However, Power BI is designed to work with tabular, row-and-column data rather than non-relational structures. To enable the connectivity among both, MongoDB offers an OBDC Connector which allows querying MongoDB data with traditional SQL methods.

Additionally, a system Data Source Name (DSN) for the BI Connector's mongosqld process is also needed to be added to the Windows station. A DSN is a saved configuration which describes a database connection to be used by an ODBC driver. Once the DSN is created for the BI Connector, it is possible to configure a wide range of SQL clients and BI tools, such as Power BI, to use the DSN and import data from MongoDB.

3.5.2 *Designing Visuals with Power BI Desktop*

An important disclaimer to be made ahead of the following section is that the visuals presented in this Work are not necessarily accurate regarding the technicalities nor precise mathematical approach of hydrogeology. They were conceived to showcase the visual possibilities of presenting the data and the best design to resonate with a broader audience.

The relationship between MongoDB collections is defined using the Station ID as key, unique and common identifier for all. Therefore, three main visuals were designed to show Piezometric, Water Level and Quality measures for the Stations in the Campina de Faro Aquifer, providing interactivity and attractiveness to the community.

The first one depicts the geo-positioned stations where quality measures were taken, with dynamic filters related to timespan, council and station are available to build evolutional graphs of levels of Nitrates and Chlorides, with the possibility of drilling down or up in time, besides a descriptive table with numeric values:

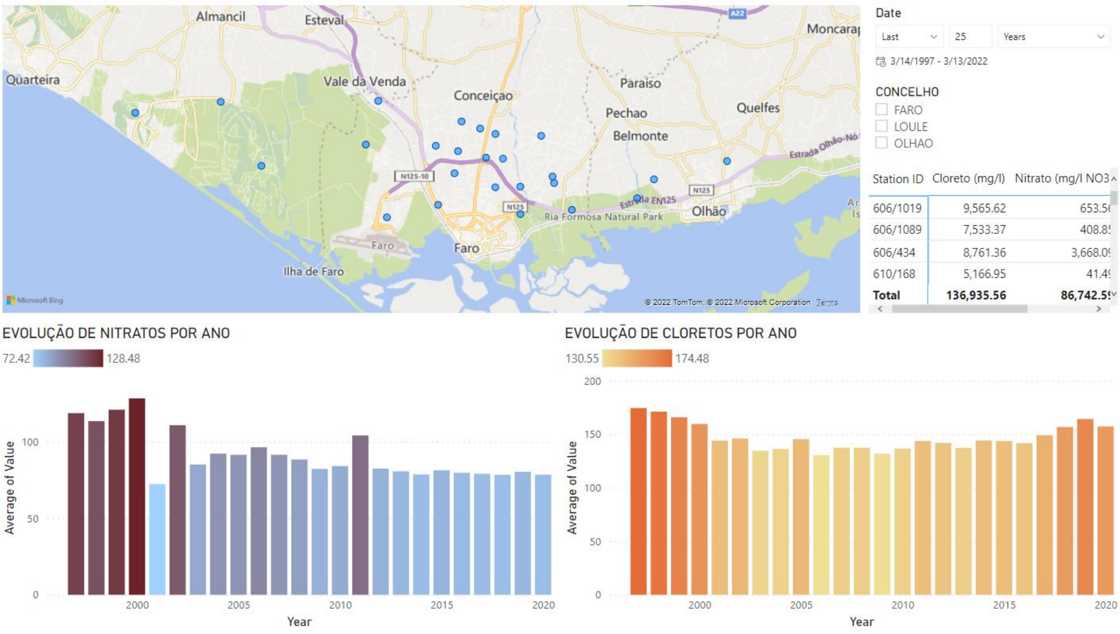


Figure 7 - Water Quality Sample Interactive Dashboard

The second brings to the user the ability to navigate through piezometric values for each borehole available as well as the possibility to explore the Aquifer as a whole, also enhanced with dynamic filters related to timespan and council, and drilling down/up capabilities in time. Although the most accurate hydrogeological calculations are not in place, is possible to see the continuous degradation since 2019. This is the desirable insight output expected to raise awareness to the community, strengthening the governance to better coordinate the water extraction, in quantity and in time-distribution:

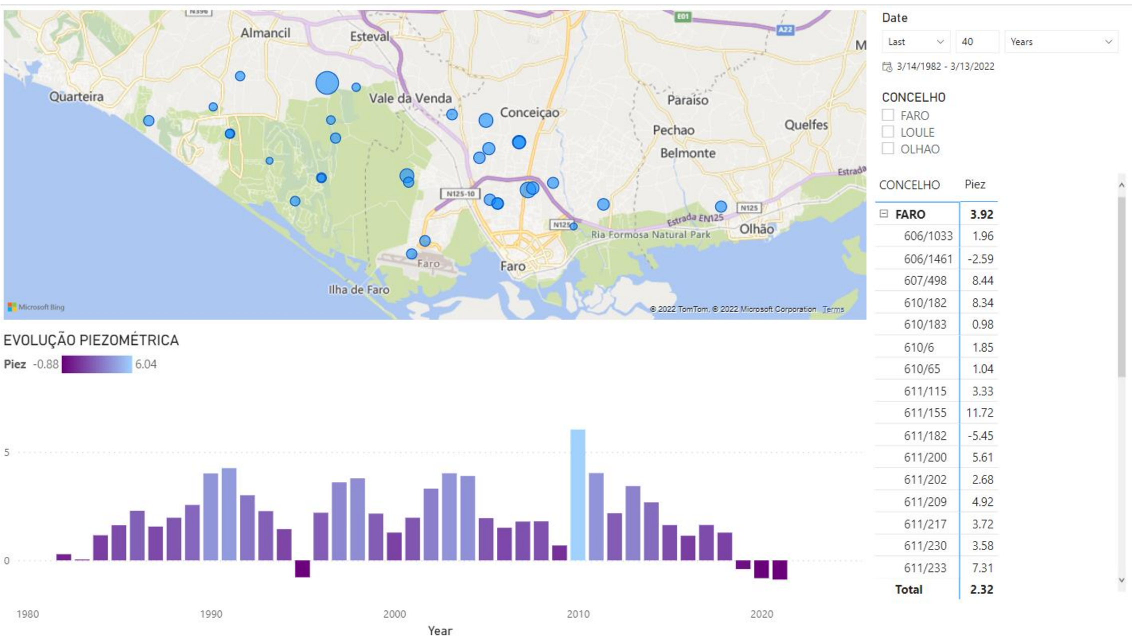


Figure 8 - Piezometric Evolution Sample Interactive Dashboard

Another interesting visual to be built to raise awareness is the year-to-year comparison regarding the piezometric values of the aquifer, where the audience can check the gaps from a period to another. In this view, all the above features are present, but the time-series graphs are overlapped:

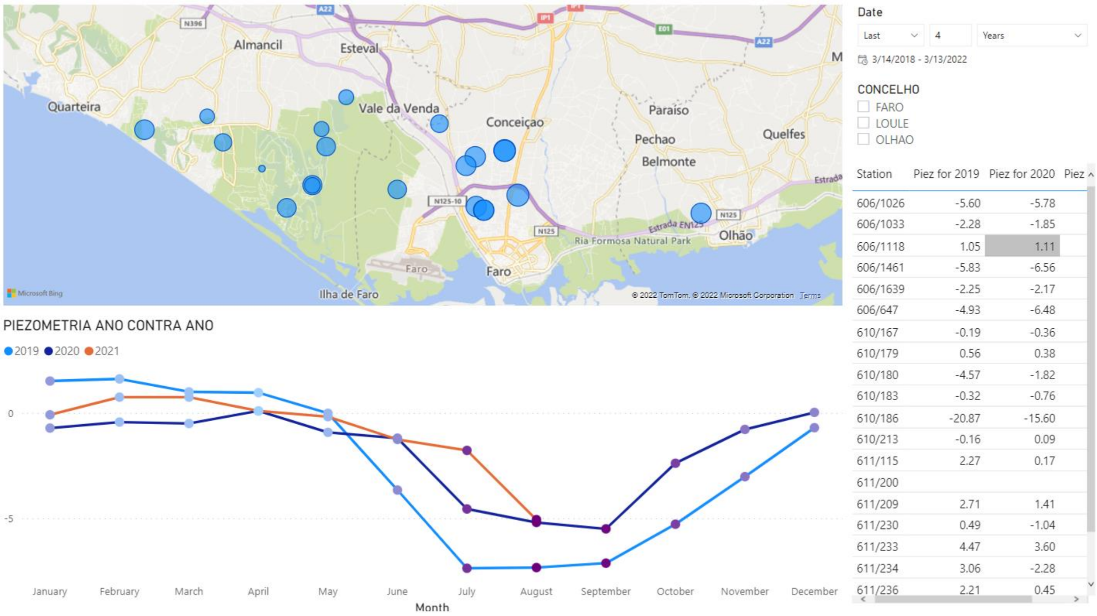


Figure 9 - Interactive Sample Dashboard with year-to-year piezometric comparison

3.6 Third-Party Cloud Services

To make the insights and visuals available to the community, other third-party cloud services are needed. The Microsoft Power BI service (app.powerbi.com) is the SaaS (Software as a Service) part of Power BI, and to where the visuals discussed in the previous chapters in the local environment were published, so end users in the Power BI service and mobile devices can view and interact with them.

The free-tier version of Power BI allows to publish a public web page for general access, however, to embed the visuals in another portal or website, which is the desirable solution, a licensed version is needed. For demonstration purposes, this Work was published to an interactive public webpage that can accessed through this link (https://app.powerbi.com/view?r=eyJrIjoiMGE2ZWE1YzEtOTM1MS00OWVlLTg3NmMtMjQ0M2RiYmY1ZmFmIiwidCI6IjYyYTZkNTAxLTU1MGEtNGE5Mi04ZTRlLWYxNGFiMGFlNGJiZCJ9&pageName=ReportSection31b32625988c546fed31).

Details on how to configure an additional Data Gateway, emulating a Windows Virtual Machine, to securely transfer data through Azure Service Bus and assure that credentials provided by gateway administrators are encrypted to help protect the information in the cloud and only decrypted on the gateway machine, are not going to be described. However, it is important to state that this is an essential step to provide real-time data refresh in Power BI Online Services, and, therefore, in any application that the visuals are replicated.

3.7 Future Integrations

Given the flexibility and openness of the designed solution, continuously improvements of the current connections as well as the possibility to expand to new integrations are desired and planned.

In an ideal scenario, all stakeholders that are already collecting data in the region, besides the Environmental Agency, in their private boreholes, and are committed to disclosure the information to enhance the overall visualization as well as contribute to more rich models for the aquifer, could be connected to the solution as additional data sources. Being themselves also benefitted from a more coordinated governance if more precise communication and forecasts are made, this seems to be a plausible pledge, though is acknowledged not to be an easy task to approach potential leads and convince them to contribute, where sectoral and political frictions are possibly the main reasons for private stakeholders keep their data in private. Nevertheless, if new data providers are available in the region, it is estimated a two-weeks work window to design and deploy new connectors in the existing workflow, considering one dedicated full-time equivalent (FTE), considering that the provider counterpart also counts with an accessible human resource.

Considering that the aquifer recharge occurs as precipitation falls on the land surface, infiltrates into soils, and moves through pore spaces down to the water table, another desirable integration is related to Weather Forecast. It is useful for the local community to forecast, plan and mitigate future droughts, relating to the current aquifer water levels, as it is useful for the research community to enrich the models in the region, being able to make the solution proposed by this Work a go-to place for every information that surrounds Campina de Faro Aquifer. With high-availability and well documented API for current weather and forecasts, OpenWeather provides a free plan to execute up to 60 calls/minute, or 1,000,000 calls/month, to obtain minute forecast for one hour ahead, hourly forecast for two days, and daily forecast for seven days. Additionally, for scholars, the platform offers a six-month trial to request daily weather data for the next 30 days, which makes it a strong and recommended option to be implemented in the project. A three-week work window to design and deploy new connectors with OpenWeather API in the existing workflow is estimated, considering one dedicated full-time equivalent (FTE). For more granular and precise information, e.g., specific areas smaller than city-level measures, local weather stations need to be provisioned, and integrate those, despite possible, is not foreseen in the project.

Ultimately, the project aims to have the integration with eGroundwater final product database built and deployed, aligned with the ambition to be a perennial solution that can abide further and beyond its initial source of inspiration. As discussed in the introductory chapter of this Work, eGroundwater ICT-based tool aims to have enhanced information acquisition through citizen science techniques. Therefore, the database that underlays the solution is exceptionally rich and utterly powerful to engage stakeholders in the co-development of sustainable groundwater governance, as the citizens can visually understand their efforts in contributing with data being transformed into actions. eGroundwater ICT-based tool is being developed by a third-party private company, and the first prototype is planned to be released after this Work submission, therefore is not yet possible to estimate the time to be spent in the integration. However, assuming that the company provides integrative APIs, and is flexible to adjust any firewall or security rules to make the data reachable, a four-week work window is a reasonable deadline to design and deploy the integration with the existing system.

4 Operations

Albeit this Work focus on building an experimental tool to throw light into new possibilities of working with data for enhanced governance and engagement, additional operations aspects are going to be described in the sections below but were not integrated in the existing system. Furthermore, as it will be discussed in Chapter 5, it is recommended that a resource (fraction of an FTE) is allocated to deal with operational aspects of the tool. Additionally, to cope with the extra infrastructure resources demanded by the following sections, an AWS optimized cost estimation is provided in the Annex.

4.1 Quality Assurance & Quality Control

Acknowledging that setting this system was only the initial step, advancing with new integrations requires a certain level of quality assurance and management, for the information to remains accurate, reliable, and readily available to all stakeholders. Data that are inaccurate, incomplete, or inconsistent, coming from an uncalibrated sensor or human errors, can create serious complications, because they may create inaccuracies in decisions about the actions that should be taken by the community (Fitz-Gerald, 2004).

Few aspects are paramount to be verified when ingesting the data, hence rigorous and careful data pipeline in every new integration must be designed to assure that format and data patterns are consistency on each record, there are no values out of an expected range, and that enforcement of data integrity is in place. The ideal solution resides in building external scripts to be constantly running in the background focusing on filtering out spurious data while generating alerts to the operational responsible. It is also desirable that an interface to let further users customize how rigorous the filtering thresholds are configured is built, enhancing the usability of the tool.

Furthermore, the introduction of data lineage and data traceability is also key and needs to be further developed. While data lineage accounts for the documentation of the data life cycle, data traceability will guarantee that the data is following the designed life cycle as expected.

4.2 High-Availability

High-availability indicates that the system is optimally designed for durability, redundancy, and automatic failover such that the applications supported by the system can operate continuously and without downtime for a long period of time. Cloud hosted systems usually face great challenges to meet high-availability standards, and many services can be used to reduce, or eliminate, the downtime of a service, such as checkpointing, load balancing, and redundancy (Endo et al., 2016). Considering that MongoDB set up is the most essential asset in the solution presented by this Work, this section will attain to describe the recommended ways to pursue always-on service.

MongoDB features a replica set mechanism that comprises a group of *mongod* processes that maintain the same data set. Replica sets helps providing redundancy and, therefore, high-availability, through multiple copies of data on different database servers, replication provides a level of fault tolerance against the loss of a single database server. Furthermore, replication could lead to increase in read capacity as inputs can be sent to different servers. Maintaining copies of data in isolated AWS instances can also increase data locality and availability. Maintaining additional copies for dedicated purposes, such as disaster recovery, reporting, or backup, is a long-term goal, but it is recommended to be kept on track.

The structure of replication considers that the primary node receives all write operations. A replica set can have only one primary capable of confirming writes write concern; although in some circumstances, another *mongod* instance may transiently believe itself to also be primary. The primary records all changes to its data sets in its operation log. MongoDB Replica Set deployment architectures, and steps to configure asynchronous replication, automatic failover, read operations, transactions and change streams, can be extensively found in Mongo's documentation repository, under the address https://docs.mongodb.com/manual/replication/.

4.3 Security

AWS is responsible for protecting the infrastructure of the computing instance, where third-party auditors regularly test and verify its effectiveness as part of the AWS Compliance Programs. The chosen region to deploy the instance and AWS network architecture is built to meet the most security-sensitive requirements. However, operational security measures need to be implemented to reduce vulnerabilities and assure integrity.

Those measures are mainly related to controlling network access to the instances, configuring virtual point codes and security groups; Managing the credentials used to connect to the instances: Managing Linux operating system and software deployed to the guest operating system, including updates and security patches; Configuring Identity Access Management (IAM) roles that are attached to the instance and the permissions associated with those roles.

Additionally, in what concerns MongDB database, some key features are also desirable to be implemented, such as authentication, access control and encryption. Special attention to authentication needs to be paid, and that can be enforced through enabling access control. With access control enabled, users are required to identify themselves and can only perform actions that adhere to the permissions granted by the roles assigned to their user. Use of Secure Sockets Layer (SSL) and Transport Layer Security (TSL) for all incoming and outgoing connections is also recommended to encrypt communication between components of the MongoDB deployment, as well as between all applications and MongoDB. A comprehensive guide on how to further implement, audit and periodically check security measures for the database can be found in Mongo's documentation available in the address: https://docs.mongodb.com/manual/administration/security-checklist/.

Local environment or other cloud service providers' security measures are not going to be discussed, since the administration of those systems falls out of the scope of this Work.

5 CONCLUSION

This Work presented a practical approach, inspired by literature references, to build a technological framework to extract, process and store data related to the Campina de Faro Aquifer from different, existent, and proposed future, data sources, as well as offered meaningful and interactive online sample visuals to engage stakeholders in participatory governance. Furthermore, the solution presented copes with the need for programmatic and non-programmatic access of the underlaying data for university students and researchers to advance technical analysis that depends on groundwater measures in the region.

There are flagrant limitations that reduces the reach of this Work, especially regarding the time needed to build, adapt, and keep improving via a constant feedback loop with the local stakeholders. As seen from the chapters hereby presented, an interactive tool that aims to create impact and engagement among its users, should have a continuous and collaborative learning curve. Thus, the full potential, real effect, and complete capabilities can only be observed if the project continues to be developed and further stages expanded. Therefore, showing ways forward with glimpses of hope and understanding of the opportunity to deploy a long-lasting, low-cost, and impactful solution, which can be used either by universities that have a strong connection to the Aquifer and to the local community, by governmental entities, or by civil society organizations, is likewise a major intent of the project.

Thus, despite not implementing all optimal recommended requirements, several sections unveiled proposed solutions to continuously improve the tool, giving the work the needed sense of continuity and movement, with ambitious goals to be achieved in, confidently and most sought, upcoming steps.

6 BIBLIOGRAPHY

Aboody, D., & Baruch, L. (2000). Information asymmetry, R&D, and insider gains. *Journal of Finance*, *55*(6), 2747–2766. https://doi.org/10.1111/0022-1082.00305

Al-Kodmany, K. (1999). Using visualization techniques for enhancing public participation in planning and design: process, implementation, and evaluation. *Landscape and Urban Planning*, *45*(1), 37–45. https://doi.org/10.1016/S0169-2046(99)00024-9

Campina de Faro aquifer | eGROUNDWATER. (n.d.). Retrieved February 19, 2022, from https://egroundwater.com/case-studies/campina-de-faro-aquifer/

Chul, D., Tae, S. &, & Eom, H. (2008). E-Government and Anti-Corruption: Empirical Analysis of International Data. *Intl Journal of Public Administration*, *31*(3), 298–316. https://doi.org/10.1080/01900690701590553

Dienel, P. C. (1989). Contributing to Social Decision Methodology: Citizen Reports on Technological Projects. *Social Decision Methodology for Technological Projects*, 133–151. https://doi.org/10.1007/978-94-009-2425-3_8

Endo, P. T., Rodrigues, M., Gonçalves, G. E., Kelner, J., Sadok, D. H., & Curescu, C. (2016). High availability in clouds: systematic review and research challenges. *Journal of Cloud Computing*, *5*(1), 1–15. https://doi.org/10.1186/S13677-016-0066-8/FIGURES/13

Fitz-Gerald, S. (2004). Management information systems: managing the digital firm, 8th Edition. In *International Journal of Information Management* (Vol. 24, Issue 2). https://doi.org/10.1016/j.ijinfomgt.2003.12.006

Ganapati, S. (2011). *Use of Dashboards in Government*. www.businessofgovernment.org

Goodchild, M. F., Parks, B. O., & Steyaert, L. T. (Louis T. . (1993). *Environmental modeling with GIS*. 488.

Greenwald, B. C., & Stiglitz, J. E. (1990). *Asymmetric Information and the New Theory of the Firm: Financial Constraints and Risk Behavior*. https://doi.org/10.3386/W3359

Iwanaga, T., El Sawah, S., & Jakeman, A. (2013). Design and implementation of a web-based groundwater data management system. *Mathematics and Computers in Simulation*, *93*, 164–174. https://doi.org/10.1016/J.MATCOM.2012.11.009

Jankowski, P. (2009). Towards participatory geographic information systems for community-based environmental decision making. *Journal of Environmental Management*, *90*(6), 1966–1971. https://doi.org/10.1016/j.jenvman.2007.08.028

Marques, M., Quadrado, F., & Rodrigues, R. (1999). *HISTORIAL DO SITE SNIRH*. www.inag.pt/snirh

Marwala, T., & Hurwitz, E. (2015). *Artificial Intelligence and Asymmetric Information Theory*. https://doi.org/10.48550/arxiv.1510.02867

Marzouki, A., Lafrance, F., Daniel, S., & Mellouli, S. (2017). The relevance of geovisualization in Citizen Participation processes. *ACM International Conference Proceeding Series*, *Part F128275*, 397–406. https://doi.org/10.1145/3085228.3085240

Matheus, R., Janssen, M., & Maheshwari, D. (2020). Data science empowering the public: Data-driven dashboards for transparent and accountable decision-making in smart cities. *Government Information Quarterly*, *37*(3), 101284. https://doi.org/10.1016/J.GIQ.2018.01.006

Meitner, M. J., Sheppard, S. R. J., Cavens, D., Gandy, R., Picard, P., Harshaw, H., & Harrison, D. (2005). The multiple roles of environmental data visualization in evaluating alternative forest management strategies. *Computers and Electronics in Agriculture*, *49*(1), 192–205. https://doi.org/10.1016/J.COMPAG.2005.03.002

Sheppard, S. R. J. (2001). Guidance for crystal ball gazers: developing a code of ethics for landscape visualization. *Landscape and Urban Planning*, *54*, 183–189.

Sieber, R. (2008). Public Participation Geographic Information Systems: A Literature Review and Framework. *Http://Dx.Doi.Org/10.1111/j.1467-8306.2006.00702.X*, *96*(3), 491–507. https://doi.org/10.1111/J.1467-8306.2006.00702.X

Stoler, A. B., Walker, B. M., Hintz, W. D., Jones, D. K., Lind, L., Mattes, B. M., Schuler, M. S., & Relyea, R. A. (2017). Combined effects of road salt and an insecticide on wetland communities. *Environmental Toxicology and Chemistry*, *36*(3), 771–779. https://doi.org/10.1002/ETC.3639

The Jefferson Project: Building The Future of Freshwater Protection | Lake George Association. (n.d.). Retrieved March 8, 2022, from https://lakegeorgeassociation.org/science-protection/jefferson-project

The Jefferson Project Data Dashboard at Lake George. (n.d.). Retrieved March 8, 2022, from https://jeffersonproject.live/sensors

Tian, Y., Zheng, Y., & Zheng, C. (2016). Development of a visualization tool for integrated surface water–groundwater modeling. *Computers & Geosciences*, *86*, 1–14. https://doi.org/10.1016/J.CAGEO.2015.09.019

Winslow, L. A., Hahn, T. H., DeVaul Princiotta, S., Leach, T. H., & Rose, K. C. (2018). Characterizing hydrologic networks: Developing a tool to enable research of macroscale aquatic networks. *Environmental Modelling & Software*, *104*, 94–101. https://doi.org/10.1016/J.ENVSOFT.2018.03.012

Wolhuter, A., Vink, S., Gebers, A., Pambudi, F., Hunter, J., & Underschultz, J. (2020). The 3D Water Atlas: a tool to facilitate and communicate new understanding of groundwater systems. *Hydrogeology Journal*, *28*(1), 361–373. https://doi.org/10.1007/S10040-019-02032-W/FIGURES/14

7 ANNEXES

7.1 Annex I – List of Water Quality Parameters and Correspondent ID from SNIRH

Parametro	Param ID	Parametro	Param ID
2,4,5-T (ug/l)	679756938	Coliformes Fecais	100003206
2,4-D (ug/l)	679759986	Coliformes Fecais UFC	667943058
Alacloro (ug/l)	100002985	Coliformes Totais	100003211
Alcalinidade (mg/l CaCO3)	100002986	Coliformes Totais UFC	667943264
Alcalinidade total (mg/l CaCO3)	100002988	Condutividade (uS/cm)	100003231
Aluminio (mg/l)	100002991	Condutividade de campo a 20C (uS/cm)	100003222
Arsenio dissolvido (mg/l)	100002993	Condutividade de laboratorio a 20C (uS/cm)	100003225
Arsenio total (mg/l)	100002994	Cor (mg/L Pt-Co)	100003234
Atrazina (ug/l)	100002995	Cromio total (mg/l)	100003237
Azoto amoniacal (mg/l NH4)	1210	Cadmio dissolvido (ug/l)	100003050
Benomil (ug/l)	830447476	Cadmio total (mg/l)	100003736
Bentazona (ug/l)	648625294	Calcio (mg/l)	100003099
Benzeno (ug/l)	1214	Desetilatrazina (ug/l)	100733299
Bicarbonato (mg/l)	100003040	Desetilsimazina (ug/l)	679761532
Carbono Organico Total (mg/l C)	100003104	Desetilterbutilazina (ug/l)	679761538
Chumbo dissolvido (ug/l)	100003139	Diazinao (ug/l)	648625306
Chumbo total (mg/l)	100003741	Dicamba (ug/l)	830447458
Cloreto (mg/l)	100003164	Dicloroprope (ug/l)	648160532
Clorfenvinfos (E+Z) (ug/l)	440410418	Dimetoato (ug/l)	648625288
Cloroxurao (ug/l)	830447470	Diurao (ug/l)	679761492
Clorpirifos (ug/l)	648797996	Dureza (mg/l)	100003262
Clortolurao (ug/l)	679760640	Dureza total (CaCO3)	2045
Cobre dissolvido (mg/l)	100003189	Dureza total (mg/l CaCO3)	100003276
Cobre total (mg/l)	100003746	E Coli (/100ml)	908915712
		E Coli (UFC/100ml)	1283

Parametro	Param ID	Parametro	Param ID
EPTC (ug/l)	680503396	Metalaxil (ug/l)	679761518
Enterococos Intestinais (NMP/100ml)	925417592	Metobromurao (ug/l)	830447464
Enterococos intestinais (UFC/100ml)	775007046	Metolacloro (ug/l)	100291001
Estreptococos Fecais	100003315	Nitrato (mg/l NO3)	100003461
Etilbenzeno (ug/l)	648628698	Nitrito (mg/l NO2)	100003466
Ferro dissolvido (mg/l)	100003325	Niquel total (mg/l)	100003456
Ferro total (mg/l)	100003330	Ometoato (ug/l)	648625280
Fluoreto (mg/l)	100003353	Oxidabilidade ao Permanganato (mg/l)	100003756
Fosfato (mg/l P)	444769264	Oxigenio dissolvido - campo (%)	100003496
Fosfato (mg/l P2O5)	100003363	Oxigenio dissolvido - campo (mg/l O2)	100003502
Fosfato (mg/l PO4)	100733293	Oxigenio dissolvido - lab (%)	100003512
Fosforo total (mg/l P)	100003382	Paratiao (ug/l)	679982580
Hidrocarbonetos dissolvidos (mg/l)	672270472	Pesticidas Totais (ug/l)	100729418
Hidrocarbonetos totais petroleo (C10-C40) (ug/l)	6908038236	Potassio (mg/l)	100003569

Imidaclopride (ug/l)	775007088	Salmonela (pres /aus)	100003581
Isoproturao (ug/l)	679761506	Salmonela 1 000 ml (pres /aus)	411881430
Linurao (ug/l)	679757480	Salmonelas (UFC /1L)	2002373968
MCPA (ug/l)	679756930	Simazina (ug/l)	262281736
MCPP (Mecoprope) (ug/l)	648625272	Sulfato (mg/l)	100003615
Magnesio (mg/l)	100003418	Sodio (mg/l)	100003595
Malatiao (ug/l)	714126118	Tebuconazol (ug/l)	3360129978
Manganes dissolvido (mg/l)	100003423	Temperatura Amostra (C)	100003625
Manganes total (mg/l)	100003428	Terbutilazina (ug/l)	100733296
Mercurio dissolvido (ug/l)	444636054	Terbutrina (ug/l)	775371394
Mercurio total (mg/l)	672270490	Tetracloroetileno ou Percloroetileno ou Tetracloroeteno (ug/l)	672270502
		Tolueno (ug/l)	648620404
		Tricloroetileno ou Tricloroeteno (ug/l)	672270496
		Trifluralina (ug/l)	1393
		Xilenos - mistura de isomeros (ug/l)	648227964
		Zinco dissolvido (mg/l)	100003638
		Zinco total (mg/l)	100003643
		pH - campo	100003566
		pH - lab	100003563

7.2 Annex II – AWS Cost Estimator for enhanced Availability and Security

Service	Upfront	Monthly	First 12 months total	Currency	Configuration summary
Amazon EC2	0	107.31	1287.72	USD	Operating system (Linux), Quantity (3), Pricing strategy (EC2 Instance Savings Plans 1 Year No Upfront), Storage amount (8 GB), Instance type (t3.medium)
Application Load Balancer	0	19.35	232.2	USD	Number of Application Load Balancers (1)
Amazon EC2	0	49.2	590.4	USD	Operating system (Windows Server), Quantity (1), Pricing strategy (EC2 Instance Savings Plans 1 Year No Upfront), Storage amount (8 GB), Instance type (t3.medium)