PROJECT DELIVERABLE REPORT

NAIADES

Greening the economy in line with the sustainable development goals

D5.7: Predictive AI analytics for consumer confidence – Mid-term

A holistic water ecosystem for digitisation of urban water sector

SC5-11-2018

Digital solutions for water: linking the physical and digital world for water solutions

Document Information

Grant Agreement Number	820985	Acronym	NAIADES
Full Title	A holistic water ecosystem for digitization of urban water sector		
Topic	SC5-11-2018: Digital solutions for water: linking the physical and digital world for water solutions		
Funding scheme	Innovation action		
Start Date	1stJUNE 2019	**Duration**	36 months
Project URL	www.NAIADES-project.eu		
EU Project Officer	Alexandre VACHER		
Project Coordinator	CENTER FOR RESEARCH AND TECHNOLOGY HELLAS - CERTH		
Deliverable	D5.7: Predictive AI analytics for consumer confidence		
Work Package	WP5 – NAIADES Smart Framework: AI analytics and predictive services		
Date of Delivery	**Contractual** M18	**Actual**	M18
Nature	R – Report	**Dissemination Level**	PU-PUBLIC
Lead Beneficiary	JSI		
Responsible Authors	Joao Pita Costa (JSI), Matej Posinković (JSI), Matej Čerin (JSI), Marko Grobelnik (JSI)	Email	joao.pitacosta@ijs.si matej.posinkovic@ijs.si matej.cerin@ijs.si marko.grobelnik@ijs.si
		Phone	+ 386 1 477 3528
Reviewer(s):	Nikos Angelopoulos (KT), Andreea Paunescu (SIMAVI)		
Keywords	Water Observatory, Text Mining, Data Analytics, Predictive AI analytics, digital twin, news monitoring, MEDLINE, SDG		

Revision History

Version	Date	Responsible	Description/Remarks/Reason for changes
0.1	2/9/2020	JSI	First version with ToC
0.2	9/9/2020	JSI	Context and argumentation on change of task description and objectives added
0.3	1/10/2020	JSI	Digital twin approach and related systems added
0.4	12/10/2020	JSI	Pilot 1 described
0.5	01/11/2020	JSI	Data sources and first pilot described
0.6	12/11/2020	JSI	Version ready for internal review
0.7	19/11/2020	KT, SIMAVI	Deliverable review and comments
0.8	20/11/2020	JSI	Corrected version for approval
1.0	30/11/2020	JSI	Final version for submission
1.1	11/5/2021	JSI	Elaborated on the related-work that considers

			GIS (in section 2.3) and how the geolocated data is present in the existing views (pilot 1) and the planned views (pilot 2). Added a couple of paragraphs in Section 4.2 to better describe the relation between the Water Observatory and the overall NAIADES architecture, its independence in functionality, and its complementarity towards the data available in NAIADES and the geolocation of some of the "Local data" (pp 41). Updated subchapter 3.6.3 ("Expected outcomes") of the chapter 3.6 ("Local data") to elaborate on how the specific configurations and priorities of use cases impact the Water Observatory (pp 37).
1.1	9/7/2021	KT, SIMAVI	Deliverable review and comments
1.2	9/7/2021	JSI	Corrected version for approval
1.3	16/7/2021	JSI	Final version for submission

Table of Contents

Summary 6
1 Introduction 7
1.1 Task focus and plan change 7
1.1.1 Problems of the Task 5.4 implementation 7
1.1.2 Alternative work plan 8
1.1.3 Comparison of the outcomes 8
1.1.4 Value for the NAIADES users 9
1.2 Structure of the deliverable 9
2 NAIADES Water Observatory 11
2.1 Background and Motivation 11
2.2 Vision and Scope 12
2.3 Related Work 12
3 Data Sources 15
3.1 Worldwide news 16
3.1.1 Dataset Description 16
3.1.2 Relevance to NAIADES 16
3.1.3 Expected Outcomes 17
3.1.4 Preliminary Results 18
3.2 Google trends & Social Media 19
3.2.1 Dataset Description 19
3.2.2 Relevance to NAIADES 20
3.2.3 Expected Outcomes 20
3.2.4 Preliminary Results 21
3.3 Economic, societal, water and other indicators through time 21
3.3.1 Dataset Description 21
3.3.2 Relevance to NAIADES 27
3.3.3 Expected Outcomes 28
3.3.4 Preliminary Results 28
3.4 BioMedical Research 28
3.4.1 Dataset Description 28
3.4.2 Relevance to NAIADES 30
3.4.3 Expected Outcomes 30
3.4.4 Preliminary Results 31
3.5 Reused EC-funded datasets 32

3.5.1 Dataset Description....32
3.5.2 Relevance to NAIADES....34
3.5.3 Expected Outcomes....35
3.5.4 Preliminary Results....35
3.6 Local Data....35
3.6.1 Dataset Description....35
3.6.2 Relevance to NAIADES....36
3.6.3 Expected Outcomes....36
3.6.4 Preliminary Results....37
4 Water Observatory Framework....38
4.1 Methodology and Implementation....38
4.2 Pilot 1 System Architecture....39
5 Water Observatory Pilot 1....45
5.1 Pilot description....45
5.1.1 GWO News Dashboard....45
5.1.2 GWO Indicator Dashboard....47
5.1.3 GWO Biomedical Dashboard....47
5.2 Extended Exploratory Dashboards....49
5.2.1 External News Dashboard....49
5.2.2 External Indicator Dashboard....51
5.2.3 External Biomedical Dashboard....52
6 Conclusions and Future Work....54
7 Bibliography....56

List of Figures

Figure 1 – Screenshot of the Blue Dot system showcasing the surface water of a selected region in Spain [24]....15
Figure 2 – Main topics appearing in the worldwide news about water scarcity....18
Figure 3 – Main categories relating to water scarcity in the worldwide new....19
Figure 4 – Google Trends comparison results on search queries about water contamination....21
Figure 5 - Freshwater withdrawl as a proportion of available freshwater resources, with highlighted case in Spain [30]....25
Figure 6: Trending topics in water resource management as observed by Microsoft Academic....30
Figure 7: Methodology adopted for the NAIADES Water Observatory....39
Figure 8: The global view of the pilot 1 over usage and data sources....40
Figure 9: Context of integration of the NAIADES Water Observatory in the global system architecture (adaped from D2.3)....40
Figure 10: System architecture for the Global Water Observatory - pilot 1....41
Figure 11: System architecture for the exploration of indicators....42

Figure 12: System architecture for the news monitoring system 43
Figure 13: System architecture for the biomedical research explorer..... 44
Figure 14: News dashboard at the NAIADES Global Water Observatory 46
Figure 15: News dashboard configuration panel, showcasing filter options to fit the news stream to the exact needs of the use cases..... 47
Figure 16: Indicators dashboard at the NAIADES Global Water Observatory 47
Figure 17: Biomedical dashboard at the NAIADES Global Water Observatory..... 49
Figure 18: External news dashboard, showcasing the main aspects of the news and their impact in social media channels..... 50
Figure 19: Timeline of news articles analysis, where the water-related articles can be analysed from the perspective of the events that they integrate..... 50
Figure 20: News categories interactive view, where the water-related articles can be analysed from the perspective of their impact in other areas of interest..... 51
Figure 21: External dashboard for indicators, showing the potential to include D3JS code for more sophisticated visualisation modules..... 52
Figure 22: The further exploration of the details of the biomedical research on the Biomedical Dataset, enhanced by the powerful knowledge-base MeSH Headings..... 52
Figure 23: Biomedical science dashboard describing through interactive visual modules aspects of the published research on water-related diseases, chemical pollutants, etc..... 53
Figure 24: The multitime-series analysis of the weather parameters in the UK, using Markov chains in a complex data visualisation available through the Streamstory technology. 54

List of Tables

Table 1: Related systems and initiatives to the NAIADES Water Observatory..... 13
Table 2: Data Sources feeding the NAIADES Water Observatory..... 15

List of Acronyms

ML	Machine Learning
SDG	Sustainable Development Goal
GWO	Global Water Observatory
GWM	Global Water Monitoring
AMAEM	Naiades partner
IHF Delft	Naiades partner
GIS	Geographic Information System
WHO	World Health Organization
GHO	Global Health Observatory
MeSH	Medical Subject Headings

Summary

The water sector is facing rapid development in the direction of the smart digitalisation of resources, much motivated and supported by the UN global initiative of the Sustainable Development Goal 6. In that context, the efforts to address the specific challenges related to water management data and priorities multiply globally. There are several "digital twin" systems dedicated to water, each of which focusing on the different aspects of the digitalisation of signal to support water management companies, as well as water "observatories", that are usually meant as Geographical Information Systems that showcase the different aspects of water resources through time.

In this deliverable, we propose a slightly different approach that integrates heterogeneous data sources to try and solve common research questions, as well as to support water management companies in their current problems. This NAIADES Global Water Observatory, now released in its pilot 1 at naiades.ijs.si, puts together: (i) real-time information from multilingual world news on water topics; (ii) data visualisation of water-related indicators through time, sourced in the SDG6 and other UN data; and (iii) scientifical knowledge from published biomedical research on water related topics (e.g., water contamination).

A forthcoming pilot 2 will ingest a range of other data sources, as described in this deliverable, where we also explore expected outcomes and preliminary results from early exploration of the available data. In that, we also describe the potential to ingest complementary local data and configure global sources to parameters addressing local priorities, and thus provide a local dimension that shall be explored with the NAIADES use case providers throughout the following months.

The initial workplan defined for the Task 5.4 to which this deliverable reports, was changed in the context of consortium agreement, due to limitations as described in the introduction section, offering valuable tools to the NAIADES user that do not overlap with other tasks in the project or restrict their functionality.

1 Introduction

The NAIADES project has a global initiative to the digitalization of the water sector and as such aims to provide the early adopters with insight from their own data as well as data collected from their users. This data, relative to the local context, is not enough to put the water provider service in a global context.
The early analysis of the workplan of Task 5.4 showed that the unavailability of useful data (as discussed below) will compromise results regarding the assessment of consumer confidence, being initially defined as a rather passive strategy towards it. The new workplan provides the NAIADES user with a Global Water Observatory that will provide value to their business and help solving a range of problems. In this section, we will be discussing the new work plan in the context of the work package structure and project objectives, as well as the projected impact we expect from the output of this work.

1.1 Task focus and plan change

This task aimed to use unsupervised machine learning methods to predict consumer confidence. In this domain, consumer confidence mostly relates to water source frequency and flavour (what relates to salinity). On 14.05.2020 the WP/task lead (JSI), the coordinator (CERTH), and the use case partner AMAEAM held a meeting to discuss and approve an alternative plan as described below. The problem of the original plan of this task, as well as the alternative plan, are also described below.

1.1.1 Problems of the Task 5.4 implementation

The plan analysis of task 5.4 and the available data and resources show evidence of the low potential for value both to the use case partners, as well as for NAIADES and its end users. This is mostly because the initially considered data is too scarce for the proposed approach. Thus, the overall consumer confidence at this level is hardly responsive to this kind of influence in the short term. We would need a larger data sample both in terms of geographic scope and time, which is not possible to be provided at the moment by AMAEM, the only use case with the required infrastructure for data collection and customer interaction required. The following points present in detail the several problems of the Task 5.4 implementation as it is defined in the Grant Agreement.

The level of locality promised to be set to "confidence in residential or commercial buildings", which is only supported by the data at Alicante; though, there is no data from consumers that can support this locality. On the other hand, a time horizon of 6 hours is a very difficult time window to achieve, mostly because we do not have historical data that can support that work. Customer satisfaction is mostly driven by large problems with water source and water quality, and how the company can efficiently solve the problem. And, also, this is the only case when a customer's opinion can rapidly change in a window of 6 hours about the topic of water supply.

In the original plan, the proposed AI service would provide consumer confidence monitoring in residential or commercial buildings, based on (i) the water consumption data per residential and commercial building provided by AMAEM; and (ii) information from online sources such as email, Facebook or Twitter. Though, a preliminary analysis shows that the population is not frequently talking about water or the lack of it in social media (as seen in Twitter). In the follow-up of social media, most complaints are related to water price. Early results show that little is discussed about lack of water in the social media, unless when a large water-related problem happens. The crawling of data from forums/social media might be insufficient. It is expected that, in the case of problems with water sourcing, people will talk about it in the media but, at the moment, we have no data from such events to have our algorithms learning from. Also, the idea of scraping freely available web sources can be done, but its efficiency is doubtful.

As for the planned interactive feedback, we see this as the interaction of AMAEM through the company's website and twitter account that will allow the general public to alert for water scarcity, maybe helped by local authorities to campaign for its awareness (this could be done by water providers or local authorities).

AMAEM can also set it up as an alarm triggered by the amount of input (social media/local news) about, e.g, the lack or taste of water. This could assess the different interactions and could turn them into indicators of the severity of the situation. Regarding data analytics, we can say something about prediction of water scarcity in interaction with the Task 4.4 (AI empowered critical water consumption monitoring) for the case of AMAEM (the only case with local data per neighbourhood). But it is difficult to predict when the alarms triggered by customers are going to happen as we have no historical data for those interactions with the general public.

1.1.2 Alternative work plan

Due to the above reasoning, we proposed an alternative workplan that was accepted with common agreement by the NAIADES consortium. The new task and deliverables description are now defined as follows:

Task 5.4 Predictive AI analytics for consumer (M3-M18 & M20-M30) [JSI(10)] - A computationally intelligent framework will be employed to build a global monitor on water, ingesting a range of open data sources, providing business intelligence capabilities to the NAIADES users, and thus indirectly improving consumer confidence. This Global Water Monitor will ingest: (i) world news data (including blog posts and the impact of news in social media); (ii) economic, societal, water and other indicators through time based on the U.N.'s Sustainable Development Goals on Water (SDG6) or the World Bank Data; (iii) data of published scientific research, that can help better understand, e.g., water contamination; and other complementary data sources. It will help water providers to explore best practices and have an early assessment to global and cross-border problems and their known solutions.

D5.7 Predictive AI analytics for consumer confidence [M18 = Nov2020] - This deliverable redefines the task (in the context of the proposed changes), defines customer satisfaction and the work settings in the context of the project, describes the data sources, provides initial info about the technology to be developed, preliminary results and elaborates on the value for the use case (AMAEA).

D5.8 Predictive AI analytics for consumer confidence - Final [M30 = Nov2021] - This deliverable describes all the final data sources used (that were not described), the data cleaning and pre-processing, the final technology and end results and extends to other use cases.

1.1.3 Comparison of the outcomes

The outcomes of this task, once the change is implemented, follow the development of a global monitoring system over global data. Instead of focusing locally, the task is redefined in order to focus on water issues on a global level. The goal would be to ingest and integrate various data sources related to the water across the world and provide a system (e.g. in a form of a web portal) which we name "NAIADES Water Observatory". The available data sources which we could immediately ingest and analyse are:

- Global news monitoring in all world languages (using https://eventregistry.org/ system) including blogs and impact of news in social media
- Social media data (i.e. Twitter) on the topic of water
- Google trends data using through the limited access API
- Economic, societal, water and other indicators through time - e.g. WorldBank (https://data.worldbank.org/), UN (http://data.un.org/)
- Sustainable Development Goals (with SDG6 on Water and Sanitation - https://www.sdg6data.org/)
- Biomedical science on water contamination from PubMed/MEDLINE - https://www.ncbi.nlm.nih.gov/pubmed/ (with over 26 million records)
- Health data - from WHO (https://www.who.int/gho/database/en/)
- Research/Science publications based on Microsoft Academic Graph (https://academic.microsoft.com/)

- Weather & Environment indicators - access through ECMWF (European Centre for Medium-Range Weather Forecasts) (https://www.ecmwf.int/)
- Reusage of EC funded results such as the outcomes of, e.g., the Global Earth Observation for integrated water resource assessment (in collaboration with the NAIADES partners IHE Delft – ongoing discussion for preparation of work)

These data sources, their expected output, and their preliminary results are described in detail in Section 3.

In order to "localize" this global information, we also analyse and consider the complementary input of local data sources provided by AMAEM such as, e.g., local priorities at Alicante/regional level and data collection related to it. This is further discussed in Section 3.6. From the task description: "Interactive feedback phase where the consumer's confidence level will be assessed using information via social media (news shared on Facebook, and Twitter feed analysis). Though, we expect that these data points of called alarms by the consumers will be low, and their significance, limited.

1.1.4 Value for the NAIADES users

As mentioned in the opening of this section, the new workplan brings much value to the adopters of the NAIADES solution, offering a global system that integrates a heterogeneous variety of data sources to provide useful insight.

We are exploring the usefulness of the Global Water Observatory with the use-case AMAEM, that already at Pilot 1 will be able to:

- Explore the global and local news on water-related topics, with a customized multilingual news source available;
- Understand the impact of some events of interest in social media (Facebook), as well as explore social media itself (Twitter);
- Define alerts to, e.g., water scarcity or water contamination over the news;
- Explore the public image of the company and their decision-making, complementing the information provided in the yearly questionnaire;
- Align with the progress of the country in the global indicators defined by the SDG 6;
- Explore water contamination from the most trusted scientific sources, and the good practices to deal with these problems;
- Reutilize datasets built on the global monitoring of water;
- Align with the input of local data sources, based on the AMAEM provided data.

The NAIADES users will also gain value by being able to:

- Solve an ill-defined task, based on inexistent data, transferring the effort to useful new capabilities.
- Including powerful input from credible sources customized to the needs of the users over open data.

Some of this value can already be accessed at the first pilot of the NAIADES Global Observatory, as described in Section 5.

1.2 Structure of the deliverable

The remainder of this deliverable proceeds as follows:

- Section 2 focuses on the description of the NAIADES Water Observatory, its vision and range, as well as its position in the context of other technologies available. It proceeds with the description of the Digital Twin concept, and the approach we will be taking in NAIADES.
- Section 3 describes the data sources that are ingested in the system, which is the potentially relevant information extracted from each of them, and which are the preliminary results already obtain from exploratory data analysis.

- ♣ Section 4 presents the NAIADES Water Observatory Framework and explains how it addresses the needs and expectations of the use cases, complementing the insight provided by other tasks in the project, enabling interaction and transversal value.
- ♣ Section 5 presents the first version of the NAIADES Water Observatory, and proceeds with a discussion on the results obtained with it.

The deliverable concludes with final remarks and the next steps for the upcoming period.

2 NAIADES Water Observatory

The priorities in European Union are rapidly changing towards sustainability and environmental efficiency, transversally to most domains of action. The European Commission's Green Deal aiming for a climate neutral Europe in 2050, and boosting economy through green technology [1] provides a new framework to understand and position water resource management in the context of the challenges of tomorrow [2].

The NAIADES Global Water Observatory will not only contribute to the improvement of European sustainability in water-related matters, but will also provide the local actors on the water resource management an active role in that. In the following section we will be describing our motivation, our vision and the scope of the work presented in this deliverable and led throughout the Task 5.4. We will also present some of the most prominent work done globally in this direction and discuss how we differ from it.

2.1 Background and Motivation

Water is fundamental to all human activity and ecosystem health and is a topic of rising awareness in the context of the recent discussions on climate change. The water resource management is central to those concerns, with the industry accounting for over 19% of global water withdrawal, and agricultural supply chains are responsible for 70% of water stress [3]. In 2015 the UN established "clean water and sanitation for all" as one of the 17 Sustainable Development Goals, aiming for eight targets to be achieved by 2030 [4]. Some of these focus central topics to the NAIADES priorities such as the increase of water use efficiency (Target 6.4), or the implementation of integrated water resources management at all levels (Target 6.5). Although the several efforts done in this context so far, the UN secretary-general points out in April 2020 that SDG 6 is "badly off track" compromising the progress on the 2030 Agenda [5].

As noted by the Organisation for Economic Co-operation and Development (OECD), the 'water crisis' has often proven to be a crisis of governance [6], where water scarcity is largely caused by mismanagement of available resources, leading to global prioritisation [7]. The intention to globally monitor water resources is not new, and already in the late 1960s [8] the first spatially-distributed water resources model appeared, with first operational uses of satellite observations in water resources developed in the early 1980s [9]. Though, the reliable management of water resources is only possible under condition of availability of adequate qualitative and quantitative information about state of the water body at any moment of time.
Taking advantage of the recent technological progress enabling much innovation that was unthinkable a few years ago, the concept of the *Digital Twin* is increasingly entering the water sector as an innovation driver.

The technical term *digital twin* appears after 2010 and got widely adopted after 2017, especially in the context of IoT technologies. Functionally, a digital twin is a "digital mirror" of a certain observed physical reality. Structurally, a digital twin is a dynamical model which, given a current state of an observed system, is capable of a digital partial reconstruction of such a system. Depending on the technology used to build digital twins, it can offer various analytical operators which allow monitoring, controlling, predicting or other reactive or proactive queries. In its simplest form, a digital twin can be a data collection of recorded sensors – a global observatory – while in more sophisticated scenarios, it can perform operations such as anomaly detection, optimization of parameters, prediction of future evolution, (re)scheduling of tasks, causality modelling, decision making and other more or less complex operations to manage the observed physical system.

Data driven digital twins are still not in a general industrial practice due to the lack of AI know-how and possibly lack of relevant IoT data to reconstruct the underlying physical processes. The current marketplace for generic digital twin construction technology is not yet very mature, with key providers positioning between Bosh, IBM, Siemens and General Electric. The majority of products are centred

around internal businesses of the corresponding companies related to mostly IoT or manufacturing. The engineering paradigm of the digital twin arrives to water resource management with: (i) online sensors becoming cheaper and ubiquitous; (ii) improved usefulness of Big Data analytics, processed for patterns and monitored for signals; (iii) the vast remote computing resources in the Cloud making them inexpensive and more accessible [10]. This allows businesses to reorganize organisational processes and workflows towards an improved cost-effectiveness deriving from the eminent digitalisation of the industry.

2.2 Vision and Scope

Taking into account the early stage of modernisation of processes in the water sector, and having in mind the specific challenges of the data collected, most of the existing approaches focus on physics-based or data-driven models (as later discussed in Section 2.3), often spatially related to a Geographic Information System (GIS), or focusing on decision-making workflows, often connected to the enterprise resource planning (ERP) software of the organisation [10]. In NAIADES we take a different approach and focus on multiple layers of heterogeneous data that provide us with several aspects of water-related priorities, ingesting data from open data sources and already existing systems (see Section 3). These are aligned with the NAIADES goals, focusing on the following main objectives:

Objective GWO 1. Provide a global perspective of the monitored water-related priorities (over news, social media, various indicators, scientific research, etc.) allowing for evidence-based decision-making
Objective GWO 2. Allow for alerts automatically generated from real-time data streams, improving the risk management and time to react.
Objective GWO 3. Ease the exploration and visualisation of the ingested data, allowing insight-driven improvements in business intelligence.

The Global Water Observatory has an underlying engine based on real-time monitoring and distributed data management system, that can be used to connect the physical and digital worlds through the smart sensing technologies and methods developed in other tasks throughout the project. In the context of NAIADES, the value of this approach is much related to specific water-related technologies and methods that improve the business of the use cases and the satisfaction of their customers. In this context, water companies can identify trends in technologies related to water and obtain useful early warning signs on research and solutions to their benefit in the future. Variations of this methodology are to be considered throughout the project highlighting aspects of the use cases priorities in regards to their services and in the light of the environmental changes that Europe is facing.
Additionally, we are also able to extract highly important information from the historical repository of more than ten years of worldwide news, including several water-related events covered in media such as, e.g., draught, floods, or political problems in access to water sources. Based on that data we can create stories on the evolution of events, including the storyline of news in a particular event. Our research is at the moment in the direction of building event templates (similar to Markov chains) that can be useful in identifying the usual happenings in such events. Having such a template for a water-related event in media might allow us to say something about upcoming events 4 and 5 probabilistically, from events 1, 2, and 3 that happened, based on projections of how things may evolve from what we know of collected news articles in the last 10 years. It is early to discuss the effectiveness of this result in the light of the available data, but indicators will serve us as milestones to build and validate these news stories. This achievement can affect the full range of NAIADES target audiences, from the general public to their water-providing companies, and policy-makers that can use the monitoring and event prediction information to their interest.

2.3 Related Work

Due to the rapidly growing awareness of the sustainability challenges that we are facing in Europe and worldwide in the context of the water resource management, there has been much work done to develop systems that are able to collect information about the available water and even simulate and forecast that in the near future. These are usually geolocation-based systems ingesting water-related data to enable real-time monitoring of resources and usage. In the following sections, we will discuss some of these as listed

and shortly described in the following Table 1. We will get into detail in some of these and contrast them to the proposed scope of the NAIADES Global Water Observatory.

Table 1: Related systems and initiatives to the NAIADES Water Observatory

ID	Type	Initiative	Focus	Comparison to NAIADES
S1	Commercial Solution	GoAigua system [11]	Improvement of water management workflow	Digitalisation of the water management at NAIADES use-cases in WP5
S2	GIS open system	Blue Dot [[12]	Data visualisation of bodies of water over time, globally.	Not considered GIS data ingestion but complemented in the Water Observatory (as discussed below and in 4.1)
S3	Open system and dataset	JRC Global Surface Water [13]	Data visualisation and analysis of surface water dynamics	Partial data in analysis to be used and ingested within T5.4
S4	Open dataset	SDG 6 Data Portal [14]	Data availability related to the SDG6 monitoring	Partial data is already ingested in the pilot 1 of the GWO
S5	Commercial solution	DIH Water Digital Twin [15]	ICT system based on forecasting and monitoring services	Digitalisation of the water management at NAIADES use-cases in WP5
S6	Commercial solution	Anglia Water's Digital Twin [16]	Digital Twin technology developed by Black & Veatch for Anglia Water	Current and predictive analysis in near real-time, also provided at NAIADES WP5 outputs.
S7	Regional solution	Águas do Porto Digital Twin [17]	Sustainable and integrated management of the entire urban water cycle of the city.	Real-time heterogeneous data integration, including GIS, real-time network sensors, and household meters.
S8	Regional solution	Companhia Águas de Joinville [18]	Local reduction of water shortcomings and losses, using hydraulic simulations.	In NAIADES we can also do some simulation and optimize water resource planning.
S9	Regional solution	San Diego's North City Pure Water Facility [19]	Model the periodic nature of backwashes, f lushes, drains, and recycles.	NAIADES can improve the planning of water sourcing and problem solving.
S10	Commercial solution	Atkins digital twin survey platform [20]	Access, analyse and further develop high-resolution 3D digital models water assets.	NAIADES is not developing any GIS solution but provides predictions from real-time data on water assets within a defined region ingested from ECMWF (see Section 3.5).

In the following paragraphs we will be discussing the highlights of some of the examples of water observatories listed in Table 1. These systems were identified through a research survey over existing methods and technologies that relate to water in the context of the SDG6 and NAIADES.

The **GoAigua** system [11] is a digital twin technology, developed by Idrica and implemented in the water utility Global Omnium in Spain. It allows, e.g., the city of Valencia to optimize its water management at the network level, improving efficiency in daily operations, plan real-time scenarios, and make some prediction on its future behaviour [21]. It provides major benefits for those organizations aiming to improve their system operation based on an already existing digital management of the water cycle and extract value from their already collected distributed data. It is focused on workflows in the water sector, including the management of water distribution networks, hydraulic efficiency or leak/fraud detection (in similar way than several of the WP5 outputs at NAIADES). It is better suited to those companies that already have their infrastructure in place and know well what they want to monitor.

Other regional and commercial solutions focus on the seamless integration of real-time heterogeneous data, including GIS, real-time network sensors, and household meters. This is the case of, e.g., the **Águas do Porto**'s Oporto Water Utility – AdP – deployed in the city by Bentley, being responsible for the sustainable and integrated management of the entire urban water cycle of the portuguese city [17], integrating heterogeneous real-time data sources (including GIS, real-time network sensors, household meters, SCADA, etc.) in the form of indicators and dashboards to enhance business intelligence. On the other hand, Welsh Water's **Atkin Digital Twin** opts for high-resolution 2D and 3D digital models of water assets, based on data collected from drones and laser scans [20].

The **Blue Dot** [12] is a water observatory based on a geographic information system GIS, exhibiting information on water levels of lakes, reservoirs, wetlands and similar water bodies over a timeline. This technology is based on the remote sensing of Copernicus satellite imagery and Sentinel Hub services. The key benefit of the service is the accumulation of global current and historical water level data in one place. Due to its cost-effective approach, anyone is able to access water level information freely; not only authorities, but also citizens can now better understand the state of their local and global environment. This overall general audience water education potential is addressed in the NAIADES Water Observatory over the public instances of the system configured to the priorities that are of most value to the technology adopters (the NAIADES use cases in Alicante, Braila and Carouge).

Similarly, the **JRC's Global Surface Water** [13] [22] is a system to display water resource availability through time, mapping the location and temporal distribution of water surfaces at the global scale. It is collecting data for 36 years and provides statistics over the collected data. It was designed to support better informed water-management decision-making. It identifies some of the different facets of surface water dynamics such as, e.g., occurrence, recurrence, seasonality and change of intensity. We are ingesting data related to natural resources from ECMWF in the NAIADES Global Water Observatory, that can complement the data made available by the JRC as an open dataset [23]. We are also ingesting geolocated data to configure the data sources to the NAIADES use case priorities (including news and social media, local indicators and published research). It should be noted that – at least for the moment – the Global Observatory is not sharing any data with the NAIADES data manager, not taking any data from the platform (NAIADES architecture and its interoperability are described in D2.9 and D3.1).

Finally, we also want to highlight the **SDG 6 Data Portal** [14] which is the system maintained by the UN to monitor the worldwide progress of the water-centred sustainable development goal. It provides data on all the SDG 6 global indicators and other key social, economic and environmental parameters. Moreover, this portal offers tailored options for visualization and analysis of the data, through maps, charts and tables. This portal is a flagship product of UN's Water Integrated Monitoring Initiative focusing SDG 6, offering in-depth information across all SDG 6 indicators, as well as tailor-made analytical and visualization tools. It makes available a rather comprehensive dataset focusing the SDG 6 indicators, part of which is ingested by the NAIADES Global Water Observatory (see Section 3.3).

In the review we have conducted in the context of Task 5.4, we haven't identified any system which includes news and social media or any text mining-related work in the context of monitoring water centred topics. We haven't also identified tools that aim the ease of exploration of the published research on what concerns, e.g., water contamination and can provide easy access to further information on best practices. Furthermore, with the exception of GoAigua, the systems identified have a rather global granularity and don't seem to be fit to any customization to local priorities, as is envisioned by NAIADES (see Section 3.6).

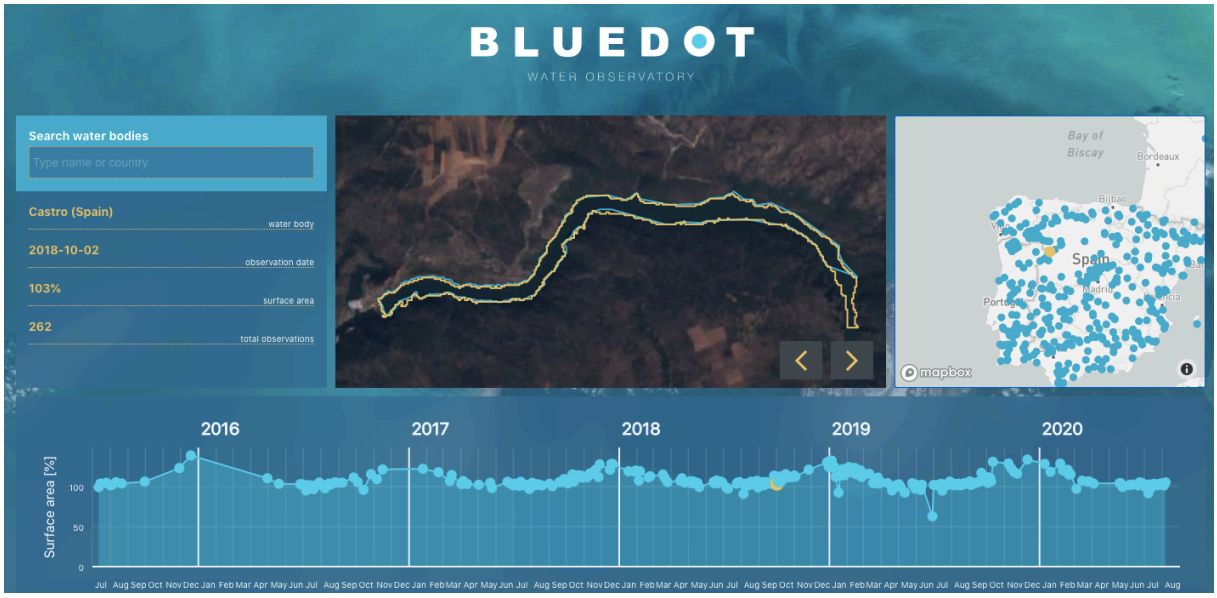

Figure 1 – Screenshot of the Blue Dot system showcasing the surface water of a selected region in Spain [24]

3 Data Sources

The Global Water Monitor (GWM) ingests a range of heterogeneous data sources, each of which with their specific challenges. In this section we list the ingested data sources (see Table 2) and (i) describe them in terms of content and data types; (ii) discuss their ingestion and integration in the overall system; (iii) describe the expected output from it; and (iv) discuss the early results generated from each data source. In this summary table presented below we list and describe the data sources, naming their source, proposing a potential usage in NAIADES, and assigning each one of them a unique identifier for later reference.

Table 2: Data Sources feeding the NAIADES Water Observatory

ID	Name	Source	Usage in NAIADES	Short description
N1	Worldwide News	Newsfeed.ijs.si	Explore news on water-related topics	News articles with timestamp, title and URL
N 2	Facebook Impact	Event Registry	Assess the impact of news in social media	Number of posts in Facebook mentioning a certain article
N 3	UN Open Dataset	UN Portal	Get access to global indicators	Global indicators on water resource management
N 4	Worldbank Open Data	Worldbank	Get access to global indicators	Global indicators on the access to clean and drinkable water
N 5	SDGs	SDG 6 web portal	Get access to global indicators	Global indicators of progress on the achievement of SDG 6
N 6	JRC	Global Surface Water	Assess available water sources	The volume of available water by GIS coordinates over time
N 7	Statistical Water Data	Eurostat	Assess water-related statistics	Different statistical time series on water-related aspects
N 8	European Water Data	EU Open Data Portal	Assess European water priorities	A range of water-related topics described by time series
N 7	Google Trends	Google	Get alerts on recent search patterns	The Google Trends data on any query over the restricted API
N 10	Twitter Feed	Twitter	Get alerts on recent	The historical data of a part of the global Twitter feed
N 11	Weather	ECMWF	Obtain forecast of	The current data and forecasts on

	data		extreme weather conditions	the weather conditions
N 12	MEDLINE	USA National Library of Medicine	Explore the research and best practices	The 26+ million biomedical research article abstracts and titles
N 13	Microsoft Academic Graph	Microsoft Academic	Explore the research and patents	The published research articles and patents (including MEDLINE)
N 14	WHO data	World Health Organization	Access health indicators	The global health indicators over WHO priority domains

* the listed datasets are planned to be ingested according to availability at source in the lifetime of the project.

Within the ingested data sources we consider primary data sources, in which we base our research on, and secondary data sources, that will be confirming or reinforcing results obtained from primary data sources. Typically, primary data sources are relatively clean, abundant and structured enough to derive representative results. On the other hand, the secondary data sources are typically scarce, with missing data values, or noisy (e.g. Twitter has a lot of noise) granularity. In the table above, we consider primary the data sources numbered 1, 4, 5, 6, 11, 12 and 13, while all others are secondary data sources.

3.1 Worldwide news

The news monitoring is an interesting and useful feature of NAIADES' GWM, allowing the user to assess a real-time news stream of global or local multilingual news (according to the choice of the user).

3.1.1 Dataset Description

The current language technologies can deal only with words and sentences, generating bottlenecks when moving from simple textual representation to semantic representation. In the latter we aim to observe the semantic/conceptual aspect of the textual content, not just lexical (words and phrases) and syntactical (sentences). The global media monitor is the system Event Registry (http://eventregistry.org/) can track over 100 thousand global media sources across 100 languages in near-real-time (having a few minutes delay).

This system aggregates global media content in a semantically meaningful way, collecting over 300 thousand news articles daily and arranging them into events. These are news stories talking about the same topic and clustered into events that further connected into story-lines, enabling tracking of evolving topics. This system can be used to track water-related topics on different levels of resolution: from local issues to country-level issues and global trends.

3.1.2 Relevance to NAIADES

We can look into the local and national or cross border news about water scarcity, using the Wikipedia term "water scarcity" (https://en.wikipedia.org/wiki/Water_scarcity) in our worldwide news engine Event Registry, and source this data stream and related properties of the event (in the last 12 months there were 491 such news on this topic, which already brings some signal). Also, we count in our global media coverage, more than 600 news on water AND quality (with multilingual capability based on Wikipedia concepts and using water quality. Other topics like "water supply" can be used when appropriate. We can also learn from related events in their sequence of related news stories. Additionally, the sentiment will be determined from news, social media and emails, which will be a base for assessing customer confidence. The sentiment will be matched to available data (water quality params), which is how we will establish a correlation between available data and customer confidence. The following are water-related Wikipedia terms that can be used to monitor cross-lingual news at a global or local scale:

- /Fresh_water (67 languages) - any naturally occurring water except seawater and brackish water
- /Drinking_water (76 languages) - also known as potable water, is water that is safe to drink or to use for food preparation.
- /Drought (96 languages) - an event of prolonged shortages in the water supply
- /Water_scarcity (25 languages) - the lack of freshwater resources to meet the standard water demand.
- /Water_pollution (62 languages) - the contamination of water bodies, usually as a result of human activities.
- /Waterborne_diseases (11 languages) - conditions caused by pathogenic micro-organisms that are transmitted in water.
- /Trihalomethane (16 languages) - were the subject of the first drinking water regulations.
- /Protozoa (85 languages) - a group of single-celled eukaryotes, either free-living or parasitic, which feed on organic matter (Waterborne_diseases)
- /Chloroform (61 languages) - or trichloromethane, is an organic compound existing as a colourless, strong-smelling, and dense liquid that is a common water contaminant.
- /Drought_in_Spain (english only) – briefly describes the mentioned situations, and the term can be used to search news annotated with this topic, but is not useful for the multilingual feature of the news engine due to the limited availability of the languages for this term.
- /Coliform_bacteria (19 languages) – bacteria that can be found in water
- /Organochlorine (22 languages) – organic compound containing chlorine
- /Pesticides (82 languages) – substances that are chemically produced to control pests
- /Nitrate (63 languages) - common components of fertilizers.

Several other water contamination terms can be considered to generate alarms, e.g., Bromoform, Bromodichloromethane, and Dibromochloromethane, that are all Wikipedia terms and thus can be used as multilingual search terms. This list of terms will be further developed throughout the project, covering the priorities addressed at NAIADES in which news monitoring can be useful. The exploration tool described in Section 5.2.1 will also help the NAIADES use-cases to contribute to the improvement of data collection.

3.1.3 Expected Outcomes

The news monitoring and exploration can provide much insight to the global and local awareness of water-related issues, and that awareness can lead us to solutions and best practices, but also to the relative dimension of some problems as, e.g., water contamination, hinting the NAIADES user to valuable information. Moreover, the measurement of the impact of the news in social media (the number of shares in Facebook) allows to better understand the public dimension of the topic, which can be useful in decision-making, assessing "voices" that are otherwise harder to reach.

In the NAIADES Global Water Observatory we make available a real-time news engine for news in water at global and local level, with potential to customize the news stream to own priorities (already available in the pilot 1 of the observatory, see Section 5.1.1). But we also make available a news data exploratory dashboard (see Section 5.2.1) that further allows the user to query the global news data and utilize the data visualization modules to:

- Access to news over time to explore the media on any water-related topic;
- Analyse news reflecting public opinion on water companies and their activities and decision-making;

- Identify key topics related to a selection of news to explore the topic of interest;
- Explore the relation of the news and the social media;
- Analyse the sentiment in news, for a particular query.

The descriptions of the user stories related to these explorations will be explored with the NAIADES use-case partners and described in the upcoming deliverable D5.8 to be released in M30.

3.1.4 Preliminary Results

Even though challenges were encountered in the context of this topic in social media (mentioned in Section 1.1.1) we have already obtained promising results in the exploration of pilot 1 that show some evidence of the usefulness of the technology.

A preliminary exploration shows that, since the beginning of 2020, 21.2762 news articles were published worldwide about 22.600 events covering topics that include fresh and drinking water, drought and water scarcity. This was based on the query »Fresh water OR Drinking water OR Drought OR Water scarcity«. On the other hand, the Wikipedia terms-base query »Water pollution OR Waterborne diseases OR Trihalomethane OR Protozoa OR Chloroform OR Bromoform OR Bromodichloromethane OR Dibromochloromethane« provides us with 3898 news across 449 events.

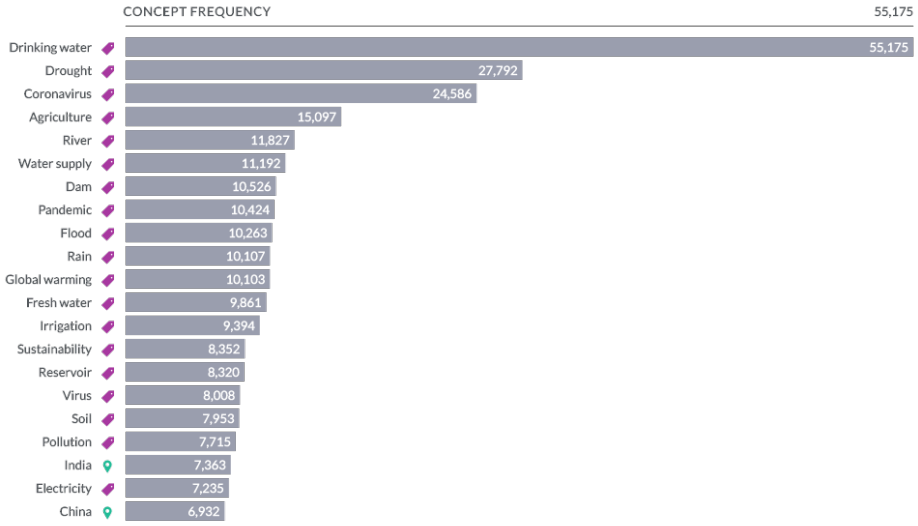

Figure 2 – Main topics appearing in the worldwide news about water scarcity

The social media impact of some of the news is evident with, e.g., the news »Edinburgh faces high alert warning for drought as Scotland dries up«[1] being shared 162 times in Facebook, giving evidence of the attention focus in European sourced news about water. On the other hand, when looking into the main concepts present in the news over these topics, we can see that the pandemic is taking an important role, considering news as »Coronavirus advice highlights Asia's terrible water scarcity«[2], but also topics like agriculture, or global warming (and others related to climate change, as in recent events like the Australian bush fires) are pointed out (see Figure 2). It is particularly interesting to notice that 5.2% of the news sample generated by this query points out the relation to energy utilities (see Figure 3). Surprisingly 40% of the news sentiment at this news articles sample is positive with over 34% neutral.

[1] https://www.edinburghlive.co.uk/news/edinburgh-news/edinburgh-weather-water-shortage-scotland-18414945

[2] https://asia.nikkei.com/Opinion/Coronavirus-advice-highlights-Asia-s-terrible-water-scarcity

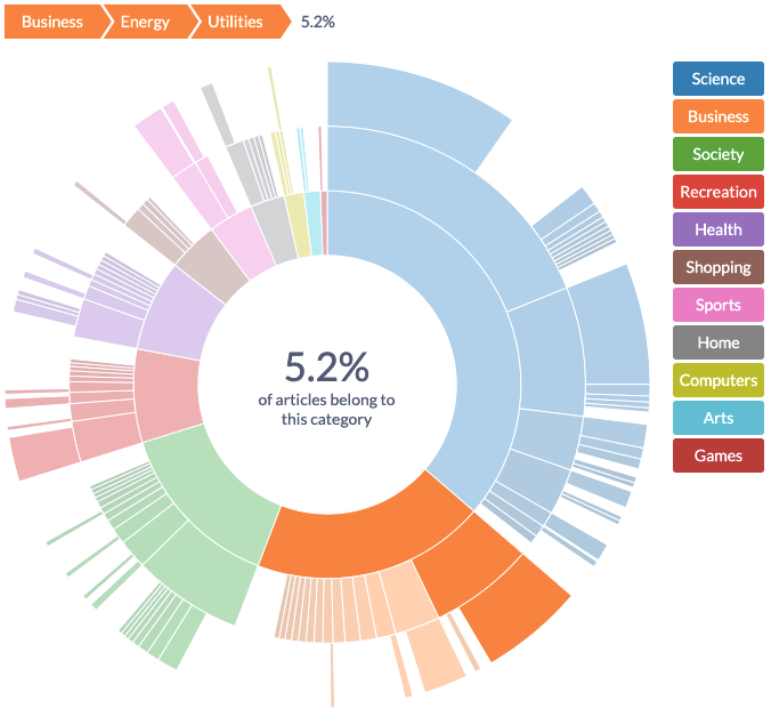

Figure 3 – Main categories relating to water scarcity in the worldwide new

3.2 Google trends & Social Media

The real-time monitoring of trends can be useful for the NAIADES user, allowing for the most current perspective on the water-related landscape. We will be providing this signal over two different sources that can complement each other – Google Trends and Twitter – that can also be complemented locally by customer data provided by the NAIADES use cases as described in Section 3.6. Typically, the nature of this data is noisy and will only serve to complement the information validated over other data sources.

3.2.1 Dataset Description

Google Trends

As their slogan is expressing, the value of knowing what are people searching over the internet (considering that most internet users perform their queries through Google) is to provide insight in times of uncertainty. This tool is frequently used by marketing agencies to identify trends and top queries across several regions and languages worldwide. It is made available by Google through a web portal[3] where the user can compare search terms, and through an API available to a limited number of institutions (including JSI).

This dataset and real-time tool can thus help NAIADES with the awareness of customer trends and concerns in what relates to water. Usually there are two types of queries:

- topic query which might be cross lingual (this is still to be investigated)
- keyword (search term) query which is exactly what people type in search (typically less results)
- entity type, usually representing a company, location, etc.

Due to the nature of this kind of data, it is mostly useful to trigger alarms than must then be verified, rather than to provide us information to ingest. The case of Google Flu Trends in 2012 [25] is an example of the eminent risk in the usage of this data source when looking for information. Nevertheless, the same case shows success when using this data and its insight in appropriate ways through nowcasting [26]. In

[3] https://trends.google.com/trends/

NAIADES we will be exploring Google Trends in what respects to water quality, water scarcity and water contamination in global and local contexts to better understand what alarm triggers can be useful.

Twitter

The role that the social media channel Twitter[4] is occupying in worldwide research is unseen, providing insight through the access to their API for academic research purposes (to which also JSI has access) [27]. The nature of this short information instances shared by online users all over the world, together with hashtags that serve as text hand-annotation of categories labelling the data, allows the researcher to have a collective overview over a certain topic. Twitter also allows the researcher to define location and other parameters that can be useful in the context of NAIADES.

Though, much like in the case of Google Trends, the Twitter data can lead to inaccuracies due to the great amount of noise that derives from the own nature of the data often based on personal opinion. Nevertheless, it has shown to be useful in defining triggers for topic-based alarms or identifying trends [28]. The aims we propose in the usage of this data source reflect the nowcasting potential associated with it [29].

3.2.2 Relevance to NAIADES

Although the noise can negatively impact the results in both the data sources, the nowcasting role that they occupy in the Water Observatory is relevant for the assessment of the general audiences. While in the case of Google Trends the terms of search that can be relevant to NAIADES range between key phrases and entities, the Twitter terms of search are key phrases and hashtags.

In Google Trends we can find the following entities that can be used in a focused search:

- Water – chemical compound
- Water scarcity – topic
- Water pollution – topic (no topic existing for "water contamination", that still can be used as key phrase)
- Drought – topic

In Twitter we can find the following hashtags that can be useful in a focused search:

- #water
- #watercontamination
- #waterdamage
- #EndTheWaterCrisis
- #waterpollution
- #LackofWater
- #drought

The choice of location can be an additional input in both of the data sources, with granularity at country level. Moreover, Twitter allows for advanced parameters in the search (e.g. searching specific accounts) although its usefulness is uncertain.

3.2.3 Expected Outcomes

Following the impact assessment of news in social media (by shares in Facebook, as discussed in Section 3.1, the further investigation of the usefulness of these data sources is to be developed in the duration of this project (as it was planned in the original task description). It has been shown to be very difficult to assess customer satisfaction over social media (as discussed in Section 1.1.1). Though, its usefulness lies on the power of nowcasting global audiences. As mentioned earlier regarding both of the data sources

[4] https://developer.twitter.com/

highlighted in this section, their potential can be leveraged in the real-time information that they can provide over what people are searching and what they are talking about.

3.2.4 Preliminary Results

A preliminary analysis of the Twitter data shows that the population is not frequently talking about water or the lack of it in social media (as in the query <agua calidad Alicante>). In fact, most complaints are related to water price. It is expected that, in the case of lack of water sourcing, people will talk about it in the social media but, at the moment we have no data from such events in substantial amounts. The monitoring of specific topics to water problems in Alicante in the social media is very limited, where most water-related content is about the beautiful seaside in Alicante pointing most data to sea water. Examples of such search queries in Twitter are, e.g., <"falta de agua" AND "alicante">, or, similarly with the query <“Escasez de agua”>.

Also, Google Trends shows little interaction on water quality in Spain (with peaks of less than 100 searches being achieved over the past 12 months). Nevertheless, the queries on water contamination and "Water Pollution" topics give us good results across the world (see Figure 4) but show evidence of the low popularity of the topic in what refers to search queries. It is also evident from this search that the usage of Topic will allow for a cross-lingual query. These query optimizations will be further investigated in the lifetime of the project towards the improvement and accuracy of obtained results.

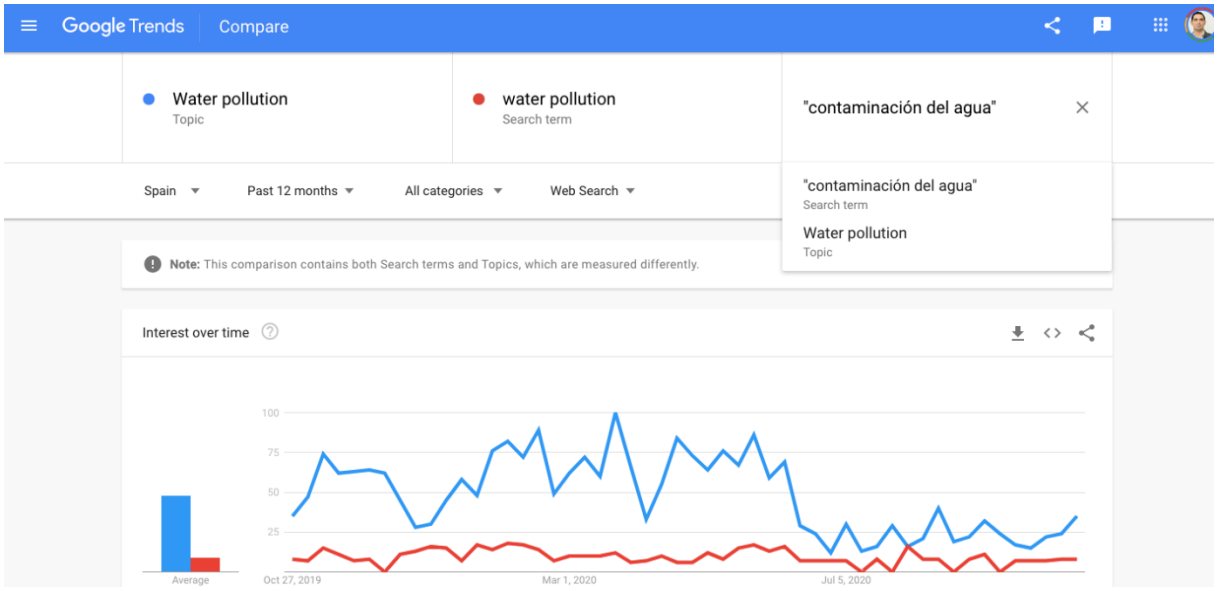

Figure 4 – Google Trends comparison results on search queries about water contamination

3.3 Economic, societal, water and other indicators through time

The quantitative aspects of water-related topics are being collected over datasets and indicators of major institutions, e.g., the UN, the WHO or the World Bank. The following paragraphs present the reader with their potential relevance, how they are being used in NAIADES, and what are the first obtained results.

3.3.1 Dataset Description

In this section we will be describing datasets and indicators that relate to water topics and can be useful to the GWO user. These are sourced over (i) the Sustainable Development Goals, where SDG 6 is on Water; (ii) the UN Open Dataset, that span a range of water-related topics; and (iii) the World Bank, that

concerns mostly datasets and indicators that can be foundational regarding major societal pillars worldwide, but also at local level.

World Bank Open Data

The World Bank Group is a family of five international organizations, with 183 member countries and 130 offices worldwide. In its extended function, it collects economic and societal data and other indicators through time (freely accessible at the World Bank Open Data portal[5]), in which water is a relevant topic.

The data is protected under a Creative Commons Attribution 4.0 International License, free to copy, distribute, adapt and display, or include the data in other products for commercial or non-commercial purposes, unless indicated otherwise in the data or indicator metadata.

This World Bank Open Data (data.worldbank.org) is a well-established organisation providing data by indicator/country, to which NAIADES relates to over climate change, fresh water, etc. It is easily accessible over the World Bank Open Data Catalogue[6] where, e.g., the dataset for *fresh surface water abstracted* includes 1366 records.
The Data Catalogue includes 48 datasets that include Spain based time series (available for download) of which 12 of them relate to water. Though, many of which are not relevant as, e.g., "Global - Electrical Conductivity In Surface Water". An example of a useful dataset is the **Access To Water** 2004-2019 consisting of a set of indicators (with values 0 and 1) and coverage including EU.

From the available, the relevant hand-picked datasets are as follows:

- **Health Nutrition And Population Statistics**
  https://datacatalog.worldbank.org/dataset/health-nutrition-and-population-statistics
  - Data Type: Time Series
  - Year: 1960 - 2019
  - Periodicity: Quarter
  - Last Updated: Sep 18, 2020
  - Granularity: National
  - Coverage: include the geolocation of the NAIADES use-cases and other EU countries
  - Relevance: complementary data to correlate
- **Global Financial Inclusion And Consumer Protection**
  https://datacatalog.worldbank.org/dataset/global-financial-inclusion-and-consumer-protection
  - Data Type: Time Series
  - Year: 2017 - 2017
  - Periodicity: Other
  - Last Updated: Jun 27, 2019
  - Granularity: National
  - Coverage: include the geolocation of the NAIADES use-cases and other EU countries
  - Relevance: complementary data to correlate
- **World Development Indicators**
  https://datacatalog.worldbank.org/dataset/world-development-indicators
  - Data Type: Time Series
  - Year: 1960 - 2020
  - Periodicity: Annual
  - Last Updated: Sep 16, 2020
  - Granularity: National
  - Coverage: include the geolocation of the NAIADES use-cases and other EU countries
  - Relevance: Environment and Natural Resources

[5] https://data.worldbank.org/

[6] https://datacatalog.worldbank.org/

- **Sustainable Development Goals**
  https://datacatalog.worldbank.org/dataset/sustainable-development-goals
  - Data Type: Time Series
  - Year: 1990 - 2019
  - Periodicity: Annual
  - Last Updated: Jul 02, 2020
  - Granularity: National
  - Coverage: include the geolocation of the NAIADES use-cases and other EU countries
  - Relevance: wb_waterdata; water_deliver_services; water_sustain_water_resources; water_build_resilience
- **Millennium Development Goals**
  https://datacatalog.worldbank.org/dataset/millennium-development-goals
  - Data Type: Time Series
  - Year: 1990 - 2015
  - Periodicity: Annual
  - Last Updated: Sep 19, 2018
  - Granularity: National
  - Coverage: include the geolocation of the NAIADES use-cases and other EU countries
  - Relevance: Agriculture and Food Security; Climate Change; Environment and Natural Resources
- **World Integrated Trade Solution Trade Stats**
  https://datacatalog.worldbank.org/dataset/world-integrated-trade-solution-trade-stats
  - Data Type: Time Series
  - Year: 1988 - 2017
  - Periodicity: Annual
  - Last Updated: Dec 28, 2018
  - Granularity: National
  - Coverage: include the geolocation of the NAIADES use-cases and other EU countries
  - Relevance: Macroeconomics, Trade & Investment to find correlations
- **Atlas Of The Sustainable Development Goals 2018**: From The World Development Indicators
  https://datacatalog.worldbank.org/dataset/atlas-sustainable-development-goals-2018-world-development-indicators
  - Data Type: Time Series
  - Periodicity: Periodicity not specified
  - Last Updated: Dec 28, 2018
  - Granularity: National
  - Coverage: include the geolocation of the NAIADES use-cases and other EU countries
  - Relevance: Agriculture and Food Security; Climate Change; Environment and Natural Resources

All of the above datasets include the geolocation of the NAIADES use-cases and other European countries.

The World Bank Data catalogue also features other datasets that might be relevant for NAIADES including:

- **Earth Observation for Sustainable Development** [105 datasets] - A collection of datasets produced through partnership between the World Bank and the European Space Agency (ESA) Earth Observation for Sustainable Development (EO4SD) initiative.
- **World Bank Water Data** [2 datasets] - From Data to a Water-Secure World for All. World Bank Water Data surrounding the three pillars of Sustain Water Resources, Deliver Services and Build Resilience.

SDGs

Another reliable and important source of global and national wide data is that extracted from the SDGs set up by the UN (where the indicator 6 is on "clean water", directly related to the project), and the datasets on progress per country that can be accessed and used for the global study.

The indicators data, that serves as evidence of national progress collected per indicator, is available through the *Global SDG indicators database* (https://unstats.un.org/sdgs/indicators/database/). In this portal the raw data can be easily downloaded in two procedures: (i) if the data requested is not complex (e.g., referring to only one subindex) then the download is direct; and (ii) if the complexity is big and the requested dataset includes several subindexes, the data is generated and then the download link is sent to the requester.

Additional information about the SDG 6 that directly relates to water topics can be accessed through the SDG6 portal (https://www.sdg6data.org/). This is a data portal that provides several highlights on the SDG6 data, including status summaries and data visualisation modules (based on templates).

The data has a yearly frequency, granularity set at country level and covers 2001-2017. The following are specifics of the SDGs on the 6th indicator:

- SDG 6 Clean water and sanitation Ensure access to water and sanitation for all
- Desired metrics
    - Water embedded in trade adjusted for environmental impact
    - Quality of drinking water and surface waters
- Goals
    - 6.1 By 2030, achieve universal and equitable access to safe and affordable drinking water for all
        - 6.1.1 Proportion of population using safely managed drinking water services - C060101
        - 6.4 By 2030, substantially increase water-use efficiency across all sectors and ensure sustainable withdrawals and supply of freshwater to address water scarcity and substantially reduce the number of people suffering from water scarcity
        - 6.4.1 Change in water-use efficiency over time C060401
        - 6.4.2 Level of water stress: freshwater withdrawal as a proportion of available freshwater resources C060402
    - 6.5 By 2030, implement integrated water resources management at all levels, including through transboundary cooperation as appropriate
        - 6.5.1 Degree of integrated water resources management implementation (0–100) C060501
        - 6.5.2 Proportion of transboundary basin area with an operational arrangement for water cooperation C060502
    - 6.a By 2030, expand international cooperation and capacity-building support to developing countries in water- and sanitation-related activities and programmes, including water harvesting, desalination, water efficiency, wastewater treatment, recycling and reuse technologies
        - 6.a.1 Amount of water- and sanitation-related official development assistance that is part of a government-coordinated spending plan C060a01
    - 6.b Support and strengthen the participation of local communities in improving water and sanitation management
        - 6.b.1 Proportion of local administrative units with established and operational policies and procedures for participation of local communities in water and sanitation management C060b01
- Indicators / Obs Mean Std.Dev. Min Max
    - For high-income & OECD countries: population using safely managed water services (%) 42 96.1 4.9 81.5 100.0
    - For all other countries: Population using at least basic drinking water services (%) 93 76.8 19.4 19.3 99.9

- For high-income & OECD countries: population using safely managed sanitation services (%) 47 86.1 12.1 60.1 100.0
- For all other countries: Population using at least basic sanitation services (%) 107 57.9 28.9 7.1 100.0
- Freshwater withdrawal as % total renewable water resources 180 65.4 287.3 0.0 2603.5
- Imported groundwater depletion (m3/year/capita) 170 10.4 18.3 0.1 148.2

The chart in Figure 5 shows the potential usefulness of the above mentioned SDG6-related datasets when compared in the attempt to identify patterns.

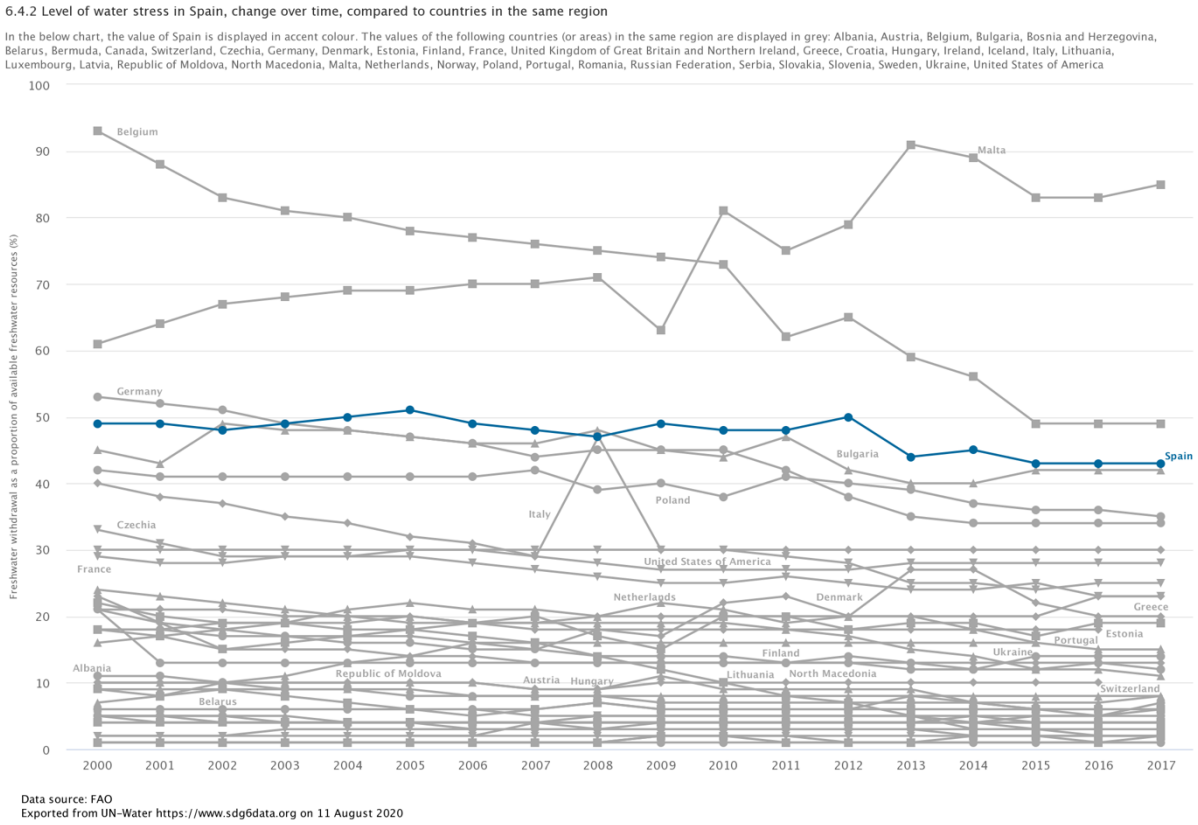

Figure 5 - Freshwater withdrawal as a proportion of available freshwater resources, with highlighted case in Spain [30]

UN Open Dataset

To complement the specific information and associated data provided to monitor progress on the SDGs, the UN provides other open datasets through their data portal accessible at http://data.un.org/. The available data is in most cases of statistical nature. Some of these are already ingested by the pilot 1 of the NAIADES Global Water Observatory, as described in Section 5.1.2.

It includes several datasets that relate to water and are not repeating from the SDG6 datasets mentioned above, complementing what the latter provide. Within these, the following are relevant to our work:

- Fresh surface water abstracted
  - 1366 records
  - Source: Environment Statistics Database | United Nations Statistics Division
- Net freshwater supplied by water supply industry
  - 1264 records

    - Source: Environment Statistics Database | United Nations Statistics Division
- Net freshwater supplied by water supply industry to: Households
    - 1117 records
    - Source: Environment Statistics Database | United Nations Statistics Division
- Falling Water
    - 964 records
      Source: Energy Statistics Database | United Nations Statistics Division
- Inland waters
    - 10428 records
      Source: FAOSTAT | Food and Agriculture Organization
- Coastal waters
    - 69 records
      Source: FAOSTAT | Food and Agriculture Organization
- Total population supplied by water supply industry
    - 880 records
    - Source: Environment Statistics Database | United Nations Statistics Division
- Population using improved drinking-water sources (%)
    - 3299 records
      Source: WHO Data | World Health Organization
- Investment in water and sanitation with private participation (current US$)
    - 360 records
      Source: World Development Indicators | The World Bank
- Public private partnerships investment in water and sanitation (current US$)
    - 369 records
      Source: World Development Indicators | The World Bank
- Proportion of population with sustainable access to an improved water source
    - 549 records
      Source: The State of the World's Children | United Nations Children's Fund

The source of the above-mentioned datasets shows that some of the information might overlap with other selected data sources to be ingested by the system. We will take this into consideration when providing the final pilot of the NAIADES Global Water Observatory.

World Health Organisation

The World Health Organization (WHO) is a specialized agency of the United Nations focusing on public health, topic which includes water-related matters such as, e.g., water contamination, water-born diseases, and access to clean water and basic hygiene. The WHO provides international health standards and guidelines, and collects data on global health issues through the World Health Survey.

It provides a publicly available Global Health Observatory (GHO)[7] based on a data repository, maintained by the WHO, serving as a gateway to the health-related statistics for its 194 member states. The GHO system allows the access to over 1000 indicators on priority health topics, including mortality/burden of diseases, environmental health, or the Millennium Development Goals (which include water and sanitation), among others. It is to be noticed that the available datasets are best estimates of the WHO, using methodologies for specific indicators that aim for comparability across countries and time. These datasets are updated when the used methodology changes or when more recent or revised data become available.

[7] https://www.who.int/data/gho/

The WHO's Global Health Observatory data repository[8] collects data per WHO region, world bank status class, etc. It provides access to over 1000 indicators on priority health topics, offering a Query API[9] and metadata information[10].

An example of a useful dataset provided by the WHO is the water sanitation and hygiene dataset of indicators, that shows that, in 2017, 71% of the world's population used safely managed drinking water services, while two billion people worldwide are still lacking basic sanitation services, and three billion people lack basic handwashing facilities with soap and water in their households. This directly relates to exposure and burden of disease, which are datasets that can also be ingested by our NAIADES system.

Moreover, the WHO also makes available the International Classification of Diseases (ICD)[11] which is a diagnostic tool used globally including epidemiology, health management and clinical purposes. The ICD includes 26 +2 roots (supplements, extensions) only for disease types (unlike MeSH). This system of diagnostic codes can be used for classifying diseases, and includes a wide variety of symptoms, social circumstances, and external causes of injury or disease, available at https://icd.who.int/dev11/l-m/en.

EuroStat Data

Eurostat (European Statistical Office) is a Directorate-General of the European Commission that provides statistical information to the institutions of the EU and provides statistical methods across its member states. It makes available online statistical data over its open datasets, hierarchically ordered in a navigation tree, with tables distinguished from multi-dimensional datasets. The statistics are extracted via an interactive tool over the Eurostat portal.

Eurostat acts in several main topics, some of which appropriate for the ingestion in the NAIADES Global Water Observatory include: Quality of life indicators (Population and social conditions); Environment (Environment and energy); and Climate change (Environment and energy). Currently there are 2870 results related to water topics. Relevant datasets are:

- **Water exploitation index** - The indicator presents i) the annual total fresh water abstraction in a country as a percentage of its long-term annual average (LTAA) available water from renewable fresh water resources; ii) the annual groundwater abstraction as a percentage of the country's long-term annual average groundwater.
- **Water use balance - Matches found in dimension**: "Hydrological parameters - flows - sources - water supply - treatment of water", position: "Hydrological parameters - flows - sources - water supply - treatment of water > Reused water"
- **Water resources – long-term annual average**: The minimum period taken into account for the calculation of long-term annual averages is 20 years. - Actual evapotranspiration is the volume of **water** transported from the ground (including inland **water** surfaces) into the atmosphere by evaporation and by transpiration of plants. - Internal flow

Thus, the usefulness of this dataset is evident in what regards topics of interest in NAIADES, and a selection of datasets will be ingested after further analysis.

3.3.2 Relevance to NAIADES

It is shown from the preliminary analysis of available data sources above, that there is a wide range of topics that can be taken into consideration and be further explored to help predict impactful behaviours of water stress levels, water supply usage, water sourced underground, etc. The frequency and granularity

[8] https://apps.who.int/gho/data/view.main

[9] https://apps.who.int/gho/data/node.resources.examples?lang=en

[10] https://apps.who.int/gho/data/node.metadata

[11] https://icd.who.int/en

of the datasets identified to be ingested by the system depend on the data structure and data collection from the entities that build and make available such resources. In most cases, these are compatible for comparison analysis as already shown in the pilot 1, discussed in Section 5.1.2. This information complements the one obtained from GIS tools as discussed in Section 3.5.

Moreover, the exploration of SDG 6 subindices and related indicators is in line with the initial work done in WP2 and reported in D2.3 on the "Gap Analysis of the Existing SDG and EU Framework for Smart Water Management" over the methodology linking smart water technologies to SDG compliance at Task 2.2. This work contributes to the results achieved in that effort and will allow the NAIADES user to further explore aspects of the national compliance over time. In that note, we hope to contribute with this work to the assessment of SDG impact reported in M36 on the deliverable D9.12.

3.3.3 Expected Outcomes

The ingestion of a wide range of indicators from well-established sources will promote evidence-based decision-making within NAIADES users and will also allow alignment with common global goals. Moreover, the timeline exploration of several of these indicators can help the user to better understand the evolution of environmental scenarios that impact water-related priorities. It should also help the preparedness and risk management of water management stakeholders observing global scenarios.

3.3.4 Preliminary Results

The analysis of data to be ingested shows that the NAIADES user will be able to analyse the progression of population access to drinking water and water-use efficiency over time, and how it relates to water stress level from freshwater withdrawal as a proportion of available freshwater resources.

Also, in pilot 1, the NAIADES user can explore the availability of water from freshwater surface, falling water and inland waters over time, and see the already evident impact of climate change in some European regions, showcasing the potential of using different indicators through time to address complex questions (as those drafted in the Conclusions section of this deliverable).

3.4 BioMedical Research

In the context of biomedical information to ingest to the system, we will be leveraging of existing open datasets that contemplate this type of research – MEDLINE (in pilot 1) – and extend it to general research and patenting – Microsoft Academic Graph (in pilot 2) – that will allow us to have a wider perspective over water-related risks and best practices (e.g., on water contamination) but also on new technologies and methodologies that can improve customer confidence by improving the capabilities of the NAIADES use cases. In this context we will also be ingestion data from the WHO to establish focus points around their well-established priorities.

3.4.1 Dataset Description

MEDLINE

The biomedical search engine PubMed, freely available since 1997, provides access to references and abstracts on life sciences and biomedical topics. The biomedical dataset MEDLINE is the underlying open database, maintained by the United States National Library of Medicine (NLM) at the National Institutes of Health (NIH). Currently, MEDLINE is based of structured information sourced on more than 27 million records dating from 1946 to the present day. It is complemented with a comprehensive controlled vocabulary – the MeSH Headings – composed of 16 major categories (covering anatomical terms, diseases, drugs, etc.) that further subdivide from the most general to the most specific in up to 13 hierarchical depth levels. These biomedical classes that aim to index publications in the life sciences are

hand-annotated by NIH human experts, including descriptors and classifiers. These biomedical classes allow the NAIADES user to explore a certain biomedical related topic, e.g., related to water contamination, which relying on curated information.

Having useful and validated technology being produced in Europe is highly relevant, in particular when the tools made available are dealing with sensitive data as in the case of the Health domain. In that, the gain of automated knowledge discovery from MEDLINE/MeSH is transversal in medical research and can highly impact the biomedical research progress. In the context of the meaningful integration and usage of data, the Horizon 2020 project NAIADES will develop the assessment to scientific knowledge in pilot 1 using MEDLINE (see Section 5) to extract information on water contamination events and best practices, using MeSH heading annotations and substance annotations.

Microsoft Academic Graph

The knowledge accessible from the Microsoft Academic Graph (MAG) provides us with a heterogeneous relational structure containing scientific publication records, with citation relationships between those publications, as well as authors filiations, and venues, throughout a variety of fields of study. This graph is updated on a weekly basis, and used to feed well established search engines as, e.g., Bing, Cortana, or Microsoft Academic.

This dataset includes over 230 million articles, with MEDLINE being part of that set, but not being restricted to the medical domain. This allows the user to, e.g., explore technologies associated to water management or identify innovation in water sustainability or withdrawal from recent research. The main topics associated with NAIADES (as explored through the citation matrix in [31]) include the path "All > Environmental science > Water resource management" where the following subtopics are considered:

- UN World Water Development Report
- Reservoir storage
- Water management system
- Water research
- Water resources
- Climate change impact assessment
- Water vulnerability
- Water sustainability
- Water withdrawal
- Freshwater resources

In the chart of Figure 6 we show the trending topics in water resource management, based on citation growth in the past 5 years.

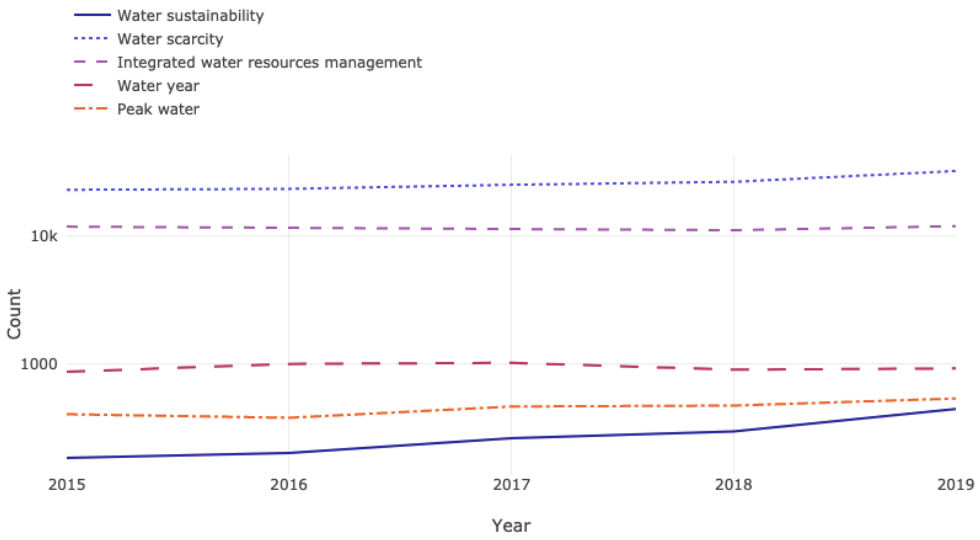

Figure 6: Trending topics in water resource management as observed by Microsoft Academic

3.4.2 Relevance to NAIADES

The results from scientific research provide insightful expert information that can extend the know-how of the water management stakeholders in priority topics such as water quality or water contamination. It is often the case that the problem lies in identify the appropriate information from the ever-rising amount of science being published every day. In NAIADES, we will also focus on this problem and in how can we find value from research with tools that ease the identification of appropriate information.

In the pilot 1 we are making available a set of tools allowing for the exploration of biomedical topics where some information relates to water at a local and global level. This is sourced from the MEDLINE dataset throughout over more than 26 million scientific articles published worldwide in many languages since the 60's (see dataset description above). The scientific research assessment tool already available in pilot 1 allows us to explore the best practices on, e.g., water contamination. Using the metadata that associates to each scientific article the hand-annotation by experts, we can identify over 250 thousand articles where the specific substance "Water" is a focus topic.

In the forthcoming pilot 2 we will be extending this knowledge to a much larger exploration of scientific knowledge by ingesting the Microsoft Academic Graph (as described above) that includes also other topics not limited to biomedical research and will provide us access to insight on new technology and methodologies as well as specific insight from water research. It also provides access to information on worldwide patents that can provide perspectives on the evolution of the business landscape.

3.4.3 Expected Outcomes

Following the value creation described above, the NAIADES Global Water Observatory is expected to enhance the competence of the user in solving water-related problems and enrich its business intelligence with insight from best practices and success stories. In detail, we expect the following milestones to be achieved with the available technology:

- Easier access to the published science on water and related topics, allowing the NAIADES user to explore the cases with focus in water contamination, water quality, etc.
- Design problem-answer strategies based on scientific evidence sourced in well-established work.
- Anticipate contamination events based on known cases and improve its level of preparedness.

3.4.4 Preliminary Results

Already in the biomedical dashboard of the GWO pilot 1 described below in Section 5.1.3, the NAIADES user can obtain rich information by using simple queries on water-related topics, e.g., happening in Spain. It allows for specific queries based on key-phrases, medical categories (the hand-annotations by MEDLINE experts, i.e., MeSH headings) and even substances (another category available in the metadata), trough powerful queries over Lucene syntax. Besides accessing MeSH through "MeshHeadingList.desc" we can also utilize the class "ChemicalList.NameOfSubstance", which are just two items in the metadata that can be useful, aside from dates, etc. Examples of such queries show good results:

- The query <ChemicalList.NameOfSubstance: "Water"> – provides 253033 scientific articles on water as substance and one of the focus points in research;
- The query <ChemicalList.NameOfSubstance: "Water" AND spain AND MeshHeadingList.desc: "Chloroform"> – provides 6 scientific articles specific on Spanish water contamination by Chloroform and is utilizing the two mentioned classes simultaneously – biomedical categories and substances – as well as Boolean operators and the key phrase "Spain" for location mention.

This is just to mention a few useful explorations of the already available data. The biomedical exploration tool described in Section 5.2.3 also allows the user to retract the article references in which these results are based on. Other examples of terms of interest in MEDLINE are as follows:

Chloroform (Chemicals and Drugs Category) [6276 results]
https://www.ncbi.nlm.nih.gov/mesh/68002725

Bromoform (Supplementary Concept) [376 results]
https://www.ncbi.nlm.nih.gov/mesh/67015044

Bromodichloromethane (Supplementary Concept) [291 results]
https://www.ncbi.nlm.nih.gov/mesh/?term=Bromodichloromethane

Chlorodibromomethane (Supplementary Concept) = Dibromochloromethane [113 results]
https://www.ncbi.nlm.nih.gov/mesh/?term=Dibromochloromethane

Trihalomethane (Chemicals and Drugs Category) - THM [1032 results]
https://www.ncbi.nlm.nih.gov/mesh/?term=Trihalomethane

Boron (Inorganic Chemicals) [10271 results]
https://meshb.nlm.nih.gov/record/ui?ui=D001895

Bacteria (Bacteria) [174744 results]
https://meshb.nlm.nih.gov/record/ui?ui=D001419

Legionella (Bacteria) [5998 results]
https://meshb.nlm.nih.gov/record/ui?ui=D007875

Escherichia Coli [288730 results]
https://meshb.nlm.nih.gov/record/ui?ui=D004926

Further discussions will take place in the following months to better understand the useful terms to consider in queries using these text-mining-based technologies.

3.5 Reused EC-funded datasets

In order to leverage the existing open data on water-related topics made available by several European Institutions, we also consider the ingestion of certain data sets that derive from EC-funded projects. This will be the focus topic of this section, and the forthcoming data exploration to be exposed in the pilot 2.

3.5.1 Dataset Description

ECMWF (European Centre for Medium-Range Weather Forecasts)

The ECMWF (https://www.ecmwf.int/) has one of the largest supercomputer facilities and meteorological data archives in the world, producing global numerical weather predictions and other data for EU member states and the broader community. It operates two services from the EU's Copernicus Earth observation programme: (i) the Copernicus Atmosphere Monitoring Service (CAMS); and (ii) the Copernicus Climate Change Service (C3S). Forecasts are produced four times per day, using advanced computer modelling techniques to analyse observations and predict future weather. The ECMWF offers data sourced over 90 satellite data sources in a daily basis, summing to 40 million observations processed and used daily (the majority of these are satellite measurements).

This dataset is based on more than 150 weather indicators with spatial resolution of 1km worldwide. It is used in several EC-funded projects that relate this data to their research problems. An example of that is the EW-SHOPP project[12] where the JSI has developed a REST API for automated data extraction[13]. This API enables access to two sources of weather data: (i) the MARS (Meteorological Archival and Retrieval System) weather data archive of the ECMWF; and (ii) the OpenWeatherMap platform[14] (OWM).

From an input set of weather parameters and a bounding box representing area of interest and timesteps for the forecast we access a collection of weather measurement forecasts where each measurement forecast is represented with:

1. Location: latitude and longitude of the measurement point
2. Timestamp: time and date of the forecasted measurement
3. Days offset: representing the number of days before timestamp when the forecast was made
4. Weather parameter: the parameter of the measurement (i.e. temperature, humidity, wind speed, air pressure, precipitation, etc.)
5. Value: the forecasted value of given weather parameter

We plan to use this data in the context of the second pilot of the Global Water Observatory, representing this data over the stream story technology[15], developed at JSI, and capable of providing complex data visualisation through interactive Markov chain-based representations accompanied by charts that help the interpretability of the obtained results. These improve prediction models based on knowing which state of the Markov chain we are in.

JRC Global Surface Water

The Joint Research Centre (JRC) is the European Commission's science and knowledge service carrying out research aiming to provide independent scientific advice and support to EU policymaking. Within its areas of activity, it includes *Environment and climate change*. In that context the JRC developed a water dataset in the framework of the Copernicus Programme, mapping the location and temporal distribution of water

[12] https://www.ew-shopp.eu/

[13] https://github.com/JozefStefanInstitute/ew-shopp-public/tree/master/analytics/weather

[14] https://openweathermap.org/

[15] http://streamstory.ijs.si/

surfaces worldwide throughout 1984-2019. It provides statistics on the extent and change of those water surfaces.

This dataset [22] is provided free of charge and without restriction of use, being based of satellite imagery from Landsat and produced under the Copernicus Programme [16]. It supports applications including water resource management, climate modelling, biodiversity conservation and food security. Moreover, it feeds a GIS-based system to visualise the bodies of water worldwide [17], and observe water occurrence and change of intensity over time in the period of 1984 - 2019. It also includes water transitions from permanency to seasonality, as well as maximum water extent.

The JRC water resource assessment platform is complemented by the eartH2Observe Water Cycle Integrator [18], providing data visualisation on worldwide water-related indicators, including drought, meteorology, optical water quality, surface water, rainfall, among others. It was developed in the context of eartH2Observe FP7 project [19], focusing global earth observations for integrated water resource assessment, integrating available global earth observations (EO). The data portal is an European contributor to the GEOSS water cycle platforms and communities [20]. The usage of these datasets is still in discussion.

EU Open Data Portal

The European Commission makes available a useful collection of open datasets over a web portal – the European Data Portal (EDP) [21] – over 25,000 datasets related to water [32]. It was set up in 2012, following European Commission Decision 2011/833/EU on the reuse of EC-funded results. The data is made available free of charge and without any copyright restrictions. Examples of these datasets are as follows:

- Item: Hydrological data water stations
  Source: data.gov.gr
  Description: a dataset that states the daily water data in different hydrological stations, such as water temperature and conductivity, in Greece.
- Item: Surface water chemistry
  Source: edgi.geology.cz
  Description: a database of chemical analysis of the surface water in the Czech Republic.
- 'Water - Evolution of price water
  Source: data.gouv.fr
  Description: a dataset that shows the evolution of the price of water in France.
- Water supply plan
  Source: opendata.dk
  Description: a database on the Municipality of Denmark's water supply plan.

Although the specificity of the datasets regarding their coverage, they can be considered useful in some local scenarios (as discussed in Section 3.6. Nevertheless, some of the datasets identified have a global coverage and can be ingested in the more general version of the system. These are some examples:

- Item: Waterbase - Water Quantity
  Source: European Environment Agency
  Description: Waterbase is the generic name given to the EEA's databases on the status and

[16] https://global-surface-water.appspot.com/download
[17] https://global-surface-water.appspot.com/map
[18] https://wci.earth2observe.eu/portal/
[19] https://cordis.europa.eu/project/id/603608/de
[20] http://www.earthobservations.org/gci.php
[21] https://data.europa.eu/euodp/en/data/dataset

quality of Europe's rivers, lakes, groundwater bodies and transitional, coastal and marine waters, and on the quantity of Europe's water resources.

- Item: Report on the quality of drinking water
  Source: Directorate-General for Environment
  Description: Member States reporting on the microbiological, chemical and indicator parameters quality under Council Directive 98/83/EC on the quality of water intended for human consumption (Drinking Water Directive)
- Item: Copernicus Land Monitoring Service - High Resolution Layers - Water and Wetness
  Source: European Environment Agency
  Description: The combined Water and Wetness product is a thematic product showing the occurrence of water and wet surfaces over the period from 2009 to 2015. Two products are available: (A) The main Water and Wetness (WAW) product with defined classes of (1) permanent water, (2) temporary water, (3) permanent wetness and (4) temporary wetness; and (B) The additional expert product: Water & Wetness Probability Index (WWPI). The products show the occurrence of water and indicate the degree of wetness in a physical sense, assessed independently of the actual vegetation cover and are thus not limited to a specific land cover class and their relative frequencies.

Zenodo

The open-access repository Zenodo was developed under the European OpenAIRE program and is currently operated by CERN. It allows researchers to deposit data sets and any other research related to digital artifacts, assigning for each submission a persistent digital object identifier (DOI), which makes the stored items easily citeable.

This open data repository includes currently 1288 open datasets in the following formats: zip (1010), xlsx (504), txt (451), xml (443) and csv (323). Though much of these relate to aspects of biology such as, e.g., biodiversity, taxonomy or insect, that are out of scope of this work. Specifically, in the context of *water contamination* we identify 1327 datasets.

The predicted problem foreseen in the usage of these datasets are their specificity, many times due to their relation with a particular published research. Examples of these that might still be useful (if, at least, their coverage includes Europe and relates to know problems and priorities) in the context of *water contamination* are the following:

- Data_Current_Pollution_Reports-20200528 [DOI 10.5281/zenodo.3871677] - Data used in the work "An overview of snow albedo sensitivity to black carbon contamination and snow grain properties based on experimental datasets across the northern Hemisphere" https://zenodo.org/record/3871677#.X3QzIJMzbLd
- Global Reservoir Geometry Database [DOI 10.5281/zenodo.1322884] - This is a global-scale reservoir storage-area-depth dataset including 6,824 major reservoirs. https://zenodo.org/record/1322884#.X3Q0cpMzbLc

The usefulness of the datasets provided in Zenodo is often relative to the topic of research that these support. Hence, the ingestion of any of these datasets will only be done if relevant data can be identified in the duration of the project.

3.5.2 Relevance to NAIADES

The nature of the weather data, as well as the water resources data, makes it extremely useful due to its frequency and granularity, when exploring the potential impacts of climate change, planning resources upfront, or defining risk mitigation strategies. The reusage of EC-funded data and global systems for this aim is of great value, also because it is representing European interests.

The access to information on the available water resources is one of the topics of interest in this research, given that the AMAEM use case relies on water resource transport over 100 km away from their location. Thus, the information on water resource availability can be particularly useful, when accessed at a regional level. Moreover, the weather information sourced at the ECMWF can also be of great value to the NAIADES user, complementing this perspective given the granularity and frequency of the ingested data.

The data sourced at the EU Open Data Portal and at the Zenodo research and data sharing platform are often more specific to the projects and research problems that drove their collection but, as pointed out above, can be useful in several contexts of the NAIADES work. Their granularity and detail can bring complementary information that unveils aspects of water-related problems not evidenced by other available datasets at the Global Water Observatory.

3.5.3 Expected Outcomes

With the availability of data of this type and structure, we expect the NAIADES user to be capable to explore certain aspects of the water-related problems in more detail, use the outcomes for simulation, and to be able to use the evidence identified in their decision-making. In particular, the use case partners will be able to explore the weather dynamic in their region, based on historical data, through the complex data visualization made available by the Streamstory technology [33] as discussed in Section 6. Moreover, the NAIADES user will also be able to explore aspects of available water resources based on interactive data visualization sourced at the JRC's Global Surface Water Explorer [22].

3.5.4 Preliminary Results

Although the brief exploration of the data in analysis, we have already been able to identify data sources that can provide exploratory capabilities to the NAIADES users, focusing on weather, water resources, quality of drinking water, etc. The availability of this data will complement the ingested indicators (discussed in Section 3.3) and the data exploration tool, of which a first version is already available in the first pilot of the Global Water Observatory (as discussed in Sections 5.1.2 and 5.2.2).

3.6 Local Data

In the following section we will be discussing the localization of the Global Water Observatory, which is based on the customization of parameters (as discussed in Sections 3.1 and 3.4), but also in the ingestion of available local data that has some compatibility to other datasets in the system. This follows the discussions with the owners of NAIADES use cases, particularly AMAEM in the context of pilot 1, to be scaled-out to the other two use case interests and priorities in the following months.

3.6.1 Dataset Description

A preliminary analysis of the data availability of NAIADES users, based on the reality of its use cases, shows that some data can be ingested to provide information that is closer to local problems and priorities. This follows the developments in WP2 as reported in the deliverable D2.2, and the discussions with the use-case partner AMAEM. Based on the latter, we expect to be able to ingest the following datasets if available:

i. Local government priorities – we will explore with the NAIADES use case owners the main points of interest at a regional level, and the data collection related to it, to align with policymaking at a local level;
ii. Company priorities – we will also be discussing the specific interests of the use case owners, and the related data available, that can impact their business and, consequently, the satisfaction of their customers;

iii. Local Public Health priorities – we will explore topics related with the main concerns of regional public health institutes, particularly in the context of water contamination.

Most of the above will help us, through the use case owners, to tune the parameters for, e.g., the news monitoring or the biomedical exploration (as explained in the above sections). In the context of (ii) the AMAEM has confirmed their access to online sensors with basic parameters and lab analysis (thousands per year, also available on their website), and the interaction with consumers through customer services and twitter (not so frequent).

As for the planned interactive feedback, we see this as interaction of AMAEM through the company's website and twitter account that will allow the general public to alert for water leaks and contaminations, maybe helped by local authorities to campaign for its awareness (this could be done by water providers or local authorities). Early results show that little is discussed about lack of water in the social media, unless when it happens massively. The consumer confidence is usually setup by the local authorities and their awareness campaigns. AMAEM can also set it up as an alarm triggered by the amount of input (social media and local news) about lack of water. This could assess the different interactions and could turn them into indicators of the severity of the situation.

AMAEM is outsourcing yearly surveys to a random sample of Alicante's population, mostly based on water quality, water price, and company's innovation public image, with several years of data for internal consumption. The usefulness of this data would be related to the water quality, but AMAEM is not sure if this data can be sourced to the project. Also, the usefulness of the data is not clear, depending on more info on what is asked/answered. Also, there is a recorded complaint follow-up system, with category and location for about 10 or 20 per day in all topics, where info could be extracted. Some info could be extracted when doing in parallel the action to user awareness at schools, but the data collection has not happened yet and the timing when that data summary and anonymization will be available would severely impact the efficiency of the task.

There is also a relationship between customer satisfaction and water taste according to the water source, but there is no detection of water taste, and it is known that the user's water taste gets used to a source and takes time to get used to another. Thus, this can introduce more noise than a solution. Nevertheless, we can only learn from this data after a substantial amount of it is collected. Another option is to have this interaction directly at the electronic receipt, but AMAEM can't provide JSI with the aggregated and anonymised data. But there is a big probability that the opinion is rather based on the price. These approaches are taken into consideration in the exploration of these local scenarios as sources of the complementary signal.

3.6.2 Relevance to NAIADES

The ingestion of data of this local nature can highlight specific aspects of the NAIADES user's business positioning it in the global and national perspectives provided by the rather wide scope of the Global Water Observatory. It will also promote the further integration of the observatory with the overall NAIADES platform and its local aims. This will highlight the value of the observatory in the exploration of solutions for local problems, but also in identifying points of improvement in the business of use case owners and in the relation with their customers.

3.6.3 Expected Outcomes

With the availability of local data, the NAIADES user will be able to position its own goals, achievements and priorities in the global context provided by those other datasets ingested. It is expected that the compatibilities on the local and global datasets ingested will represent problems in the usage and interpretability of results. This should also be the case regarding the further investigation on the optimisation of the system to identify meaningful insights, transversal to these datasets.

To this aim, we made available the pilot 1 to be explored by the NAIADES use-case partners and to be used in further discussions on the configuration of the observatory to the local priorities and interests. We will be exploring the context of the local priorities described above in the following:

- Water resource assessment, as described in Section 3.5.1;
- Weather data at regional level, also described in Section 3.5.1;
- News monitoring customization to local priorities, partially available in pilot 1;
- Water-related events analysis as, e.g., floods, based on historical news data;

3.6.4 Preliminary Results

Regarding the own priorities expressed by the AMAEM, we are exploring: (1) water quality and contamination, (2) perception water taste (based on Spanish keywords “agua”, “sabor”, “color”, “desagradable”, “cloro”, “cal”), (3) water leaks and pressure, (4) overall perception of the company to customer service. The current challenges relating to the COVID19 effect (4), as there have been some complains, directly communicated and expressed online (e.g., in the valuation in Google Maps). Also, regarding (4) we will investigate discussions in the media and social media about discharges to the sea (“contaminación marina”, i.e., sea pollution; and “vertidos”, i.e., discharges) and city construction works (“obras”, i.e., work in progress), as well as “odores” (odour) related with wastewater treatment stations. Moreover, we confirm results on tuples: (1) “odor” (odour) & “depuradora” (waste water treatment plant); and “odor” & “alcantarillado” (waste water network).

The news query <“agua” AND (“sabor” OR “color” OR “desagradable” OR “cloro” OR "cal")> identifies 55248 news articles in the context of 4694 events since the beginning of 2020, of which 528 news are associated with Spain. In the latter, 77 news articles relate to clorine and 62 about food, but there is an abundance of COVID19 related topics that include noise in the obtained signal. A preliminary analysis using the same query over Twitter shows evidence of noisy signal and a very small amount of useful results. We will continue to explore ways to optimize queries together with use case partners in order to improve the methods of exploration of the data sources in the local context.

Regarding customer satisfaction assessment, there are no results yet due to the fact that no data was provided (which relates to the problem exposed in Section 1.1.1). That data availability is still in study by the use case owners, and will need to comply with privacy policies (e.g. GDPR) ensured before the data is handled for analytics. The bottlenecks encountered in this aim and the secondary relevance of these outcomes, as expressed by use case owners, limit this assessment to the nowcasting in the context of social media (Twitter) and search results (Google Trends) discussed in Section 3.2.

4 Water Observatory Framework

In the following section we will discuss the overall framework of the NAIADES Global Water Observatory, in the context of the objectives and expected impact described in Section 1. This observatory will draft the NAIADES Water Digital Twin, an original concept and approach earlier described in Section 2.2 and deployed in the context of Task 5.4.

4.1 Methodology and Implementation

Following the earlier discussions on the vision and scope of the NAIADES Global Water Observatory in Section 2.2, we will be presenting in this section the methodology adopted to implement that concept in the context of NAIADES. We will also discuss what are we already making available in the pilot 1, and what we expect to release as pilot 2 in M30, as reported by the deliverable D5.8.

Taking into account the schema in Figure 7, with this work we consider the construction of the NAIADES Global Water Observatory into phases going from lower to higher complexity. We start by putting together data sources that are meaningful to a range of stakeholders that are targeted by NAIADES, from the water providers and their customers to decision makers that can leverage the information provided to established evidence-based policies.

At the data collection phase, we are concerned with addressing properly the challenges in the heterogeneous nature of the data, their different frequency and size, as well as the levels of access to it established by data providers. These parameters taken into consideration were carefully described in Section 3 to ensure the appropriate data ingestion into the system. The selection of data sources is done manually, but their ingestion is automated, and their frequency of update depends solely on the data provider. At this stage we are collecting data from many different data sources such as the Worldwide Media, the Microsoft Academic Graph, the Word Bank, the United Nations SDGs, etc.

A forthcoming stage is in the data cleaning, data processing and data integration prior ingestion. This step is highly important to allow the data quality that is needed in order to obtain useful insights from it. In this step we include the data curation, where the most meaningful datasets are selected and parsed. We also include the exploratory data analysis and some data visualisation for the purpose of prototyping what is then available at the Water Observatory. A version of this is already available at the pilot 1, easy to use by non-technical users, as described in Section 5.2.

The Observatory phase is then possible when the curated data streams of a selection of dynamic data sources are live in the system and can be used to obtain insight on particular topics of interest, monitor KPIs associated with business priorities, and allow a global and local perspective on water-related topics. These include interactive data visualisations of indicators and statistical data, the dynamic view of the news sources over priorities, or the survey over scientific research. This allows for insight on topics such as water scarcity and water contamination that will be running examples in pilot 1 sourced from the shared interest of use case partners in Alicante.

In the context of the planned work of this project we are exploring the concept of a m\eaningful Water Digital Twin that builds over the Global Water Observatory to higher the complexity to data interoperability. This is usually difficult to achieve in full due to the heterogeneity of the data, the different characteristics of the data sourced (frequency, data types, etc) and the domain knowledge needed to identify new challenges covering a wide range of business intelligence priorities. Nevertheless, useful aspects of it can be achieved, some of which are already evident from the released pilot 1. An example of this is to track a source of water contamination in the news, its impact in the social media, and explore the range of the problem in the published scientific research, as well as extract good practices to deal with this problem.

We had a final stage to this diagram that is usually forgotten in a theoretical framework, which is the adaptation of the system to the needs and priorities in the NAIADES user side. Here we consider the

ingestion of local data, the customization of news streams, the availability of exploratory dashboards, the shareable instances for policy makers, and the APIs for 3rd party integration. This is all considered by design (see Section 4.2), most of it already available in pilot 1 (see Section 5).

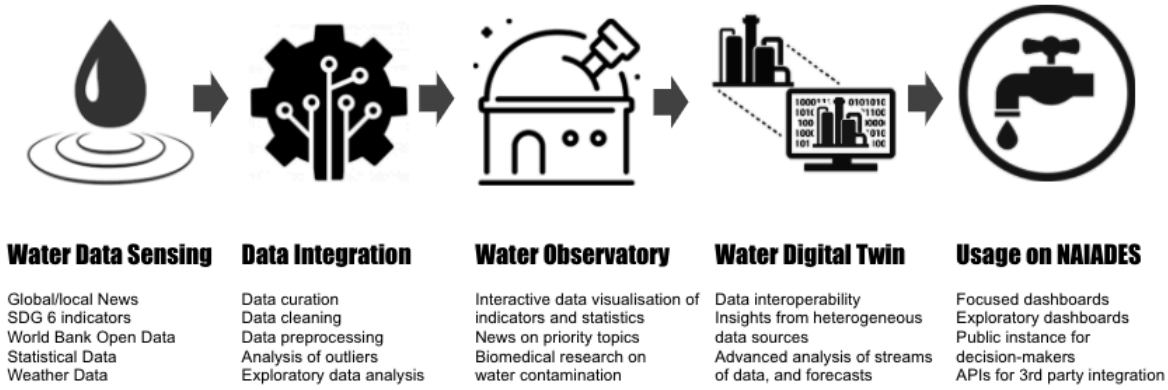

Figure 7: Methodology adopted for the NAIADES Water Observatory

A typical example of the aimed outcome of such an intelligent system is the following sequence of events:

1. a new technology is identified in **scientific research**
2. the **patents** around it multiply, alerting for its importance
3. new mentions of its usage arise in the **market** landscape
4. companies relating to it are now able to guarantee new **investment**
5. **media** is more and more aware of the importance of the technology (unknown in step 1)
6. **github** mentions show the growing technology communities in contact with the trend
7. the **job** market also reacts to the trend

The system that is able to access the data sources that relate to the bold items above, is also able to track the term throughout the several phases of popularization. It is also able to show the current status of a particular topic of interest, and optimally alert for potentially trendy topics in the future.

4.2 Pilot 1 System Architecture

To enable an appropriate sense of usability of the NAIADES Global Water Observatory from its early version in the proof-of-concept of pilot 1, the NAIADES user can leverage insight from data sources of different nature. This first version is, thus, composed of three dashboards that serve three different aims:

1. News: the customizable monitoring of news mainly on water scarcity and water contamination
2. Indicators: the information fed from open data sources and integrated to provide insight on the status of priority water-related topics
3. Medical: the published research on biomedical aspects of water contamination, and lessons learnt to promote problem-solving.

These dashboards come together to provide the user with a global perspective in real-time, where five different tiers of usability are made available (see Figure 8). The tiers allow for the extended usability of the Global Water Observatory, transversally to the data sources available. In that, we offer the NAIADES user exploratory dashboards for the further investigation over news, to get deeper into the indicators ingested, and to explore the biomedical research in detail. Moreover, each of the three dashboards have versions built to be exposed by, e.g., iframe through a publicly available channel that can be used for integration in high management KPI-monitoring dashboards. Furthermore, we also offer a part of the information in these through APIs easily integrable with own systems.

Figure 8: The global view of the pilot 1 over usage and data sources

In the context of the overall NAIADES architecture, as described in deliverable D2.3, the Global Water Observatory occupies an independent position (see Figure 9, extracted from D2.3), much like the platform for *Behavioural Change Support*. Indeed, the Water Observatory ingests mostly data provided by open sources (as described in Section 3) at pilot 1. Nevertheless, we are investigating the ingestion of local data and consumer interaction data that will complement the global perspective proposed. The Global Water Observatory is an independent module in the NAIADES overall architecture, based on open data sources as news, social media, published research, etc. More information on the architectural context of the Global Water Repository within NAIADES is found in the deliverable D2.9.

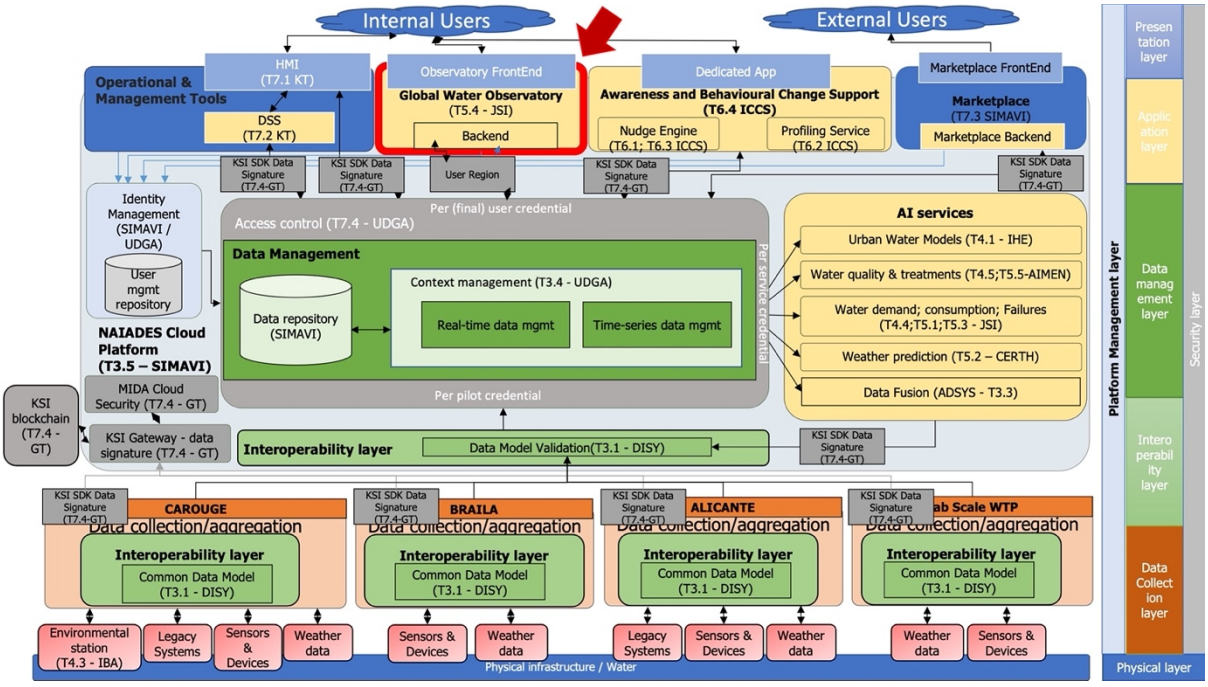

Figure 9: Context of integration of the NAIADES Water Observatory in the global system architecture (adapted from D2.3)

The architecture of the pilot 1 will be extended to pilot 2 based on the construction of new tiers of data workflows that put together data sources of similar nature to which some compatibility of data types can

be identified and related to which the outcomes also provide complementary answers. It will include a new dashboard exploring the water-related event types of the interest of NAIADES use-cases, exploring events of different nature as e.g., floods or water taste, in the context of worldwide multilingual news and social media through Twitter, in an ongoing collaboration with the NAIADES partners AMAEM and IHE Delft. It will also include a resource-focused dashboard that will reflect geo-located weather conditions that can impact those resources. This will expose local weather data ingested at the global European initiative ECMWF. It will expose medium-term predictions and provide an interactive data visualisation module based on Markov chains technology to explore impacts. This will be complementing other GIS data available, which is not the focus of the NAIADES Water Observatory but can be integrated in the NAIADES platform taking in consideration its versatile software architecture and data interoperability (see D2.9 and D3.1).

The Water Observatory is using "spatial data" (in the sense of geolocated data but not as GIS coordinates) in most of its existing (pilot 1) and planned (pilot 2) tools as in:

- natural resources - we are collecting weather data from the ECMWF (European Centre for Medium-Range Weather Forecasts), specifically: rainfall, temperature and humidity;
- news - in most news articles the location is extracted and can be used to explore the similar scenarios in specific world regions;
- social media - in the tweets the location is extracted to be used in filters and specify to use case areas;
- indicators: all the datasets that we load as indicators are geolocated to indicate country - region - city as geolocation granularity;
- research: some of the published research ingested has geolocation associated.

In the following we shall discuss the architecture of the currently available pilot 1, having in perspective the planned pilot 2, that will be available later in M30.

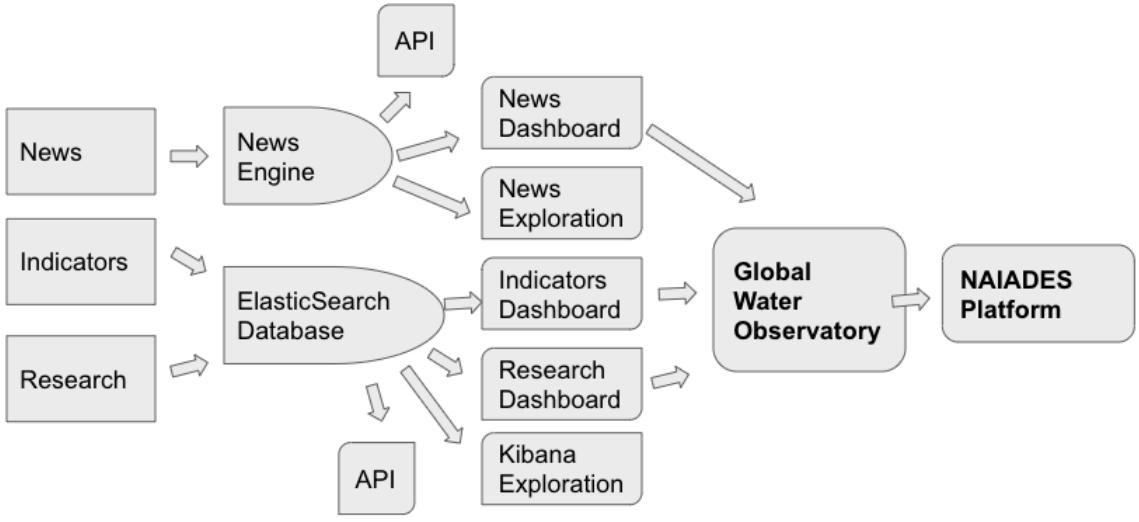

Figure 10: System architecture for the Global Water Observatory - pilot 1.

In the following we will be discussing the specific architectures of the three available tiers (news – indicators – biomedical). These architecture descriptions will be extended in deliverable D5.8 to include the system architecture of the upcoming tiers. The scheme in Figure 10 (also mentioned in the deliverable D2.9) describes the overall architecture of this pilot 1, based on two main core technologies: (i) the Event Registry news engine, and (ii) the versatile Elasticsearch engine. In both cases we can access powerful queries to the ingested data, easy to understand data visualisation modules, shareable dashboard instances and API access for 3rd party integration.

Let us now describe the dashboards, one-by-one. We shall begin with the *indicators' dashboard*, that provides the NAIADES user with interactive data exploration tools that allow for the KPI-monitoring over several water-related topics that include the SDG 6, the World Bank Open Data, the UN data, etc. The

description of functionality of this module is then described in Section 5.1.2, while its corresponding data source is provided in Section 3.3.

In this module (see Figure 11) we ingest different data sources that include relevant indicators, within an automated data collection. Considering their well-established data types, the data integration is possible and, whenever limitations appear due to lack or poor quality of the data, the dataset is pre-processed to allow for data completion (whenever possible), or at least the improvement of data quality.

After the appropriate data integration, the dataset enters the data management component that is fed over a HTTP API, based on a settings file. This data management component is based on the Elasticsearch technology [34], fundamental for both the interactive data visualisations (created with D3JS), and the Indicators Explorer view. The latter allows the NAIADES user to explore the raw data through the Kibana [35], using a Lucene-based query that can leverage the loaded metadata, and to easily build visualisation modules that can define a new thematic dashboard of data visualisation modules.

While the *Indicators Dashboard* is integrated in the overall NAIADES platform, as discussed earlier in this section, the *Indicators Explorer* is an external dashboard that allows for further exploration of the loaded datasets in their most appropriate upload frequencies. The latter then allows for a public instance provided by Kibana that can be used through iframe to be integrated in, e.g., a high-management dashboard. It also allows for an API, native to Elasticsearch. See Figure 11 for further details.

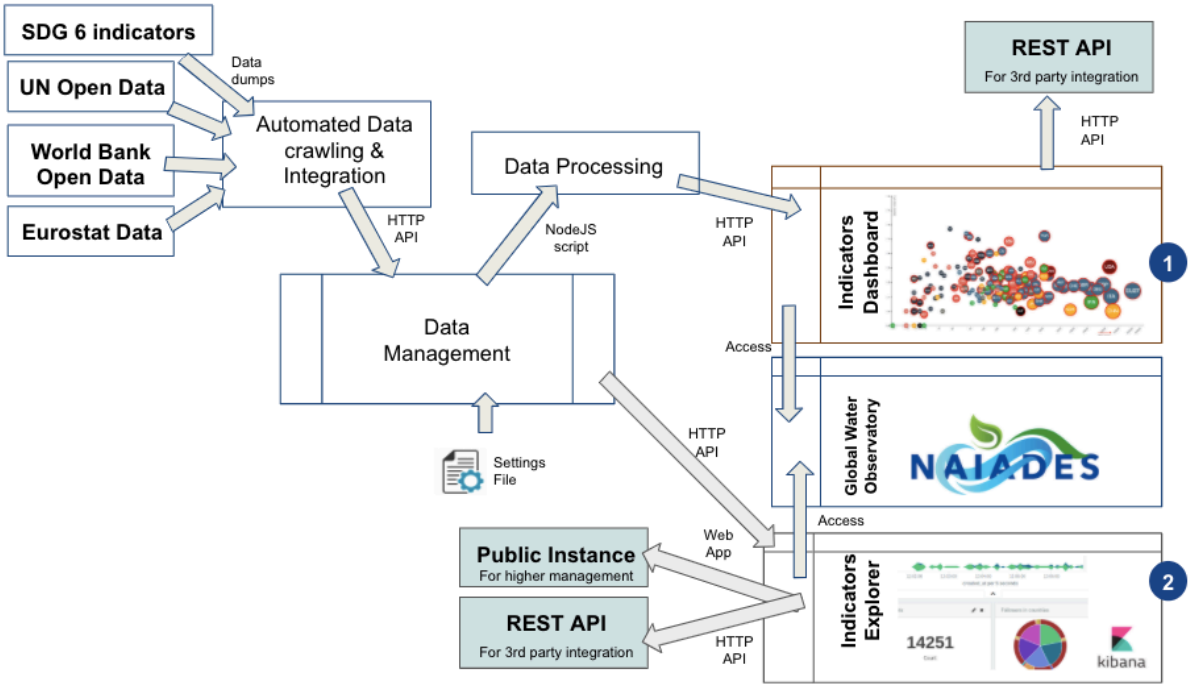

Figure 11: System architecture for the exploration of indicators

We continue with the *news dashboard* that provides the NAIADES user with the real-time news monitoring over water-related topics such as *Water Scarcity* and *Water Contamination*. The description of functionality of this module is then described in Section 5.1.1, while its corresponding data source is provided in Section 3.1. In this module we crawl the news over the internet using the *Newsfeed* [22] technology, that provides the system with a continuous stream of news articles, real-time aggregated and semantically enriched (through the *Wikifier* [23] technology), sourced from RSS-enabled sites across the world. The news articles are also

[22] http://newsfeed.ijs.si/

[23] http://wikifier.org/

automatically annotated with a classifier built to assign DMOZ classes [24] to it, adding does to the metadata of the item. The dataset and corresponding metadata are then stored and managed in the QMiner engine[25], a nodeJS-based technology built for data management. All the latter technologies were built by JSI in the context of other EC-funded projects and are being refocused and customized to address the specific challenges within NAIADES.

From the data management module. the real-time news data is accessed by the news dashboard that can be configured by the NAIADES user to tune the topics of interest in the configuration web app (this functionality is further described in Section 5.1.1. To further explore a topic in the news, the NAIADES user can access the *News Explorer* that offers a series of data visualisation modules allowing for the analysis of the topic over a timeline of published news, the relation between aspects of the subsampled related news, or even the DMOZ categories that relate with it (read Section 5.1.1 for more detail). The architecture is further described in the illustrative schema of Figure 12.

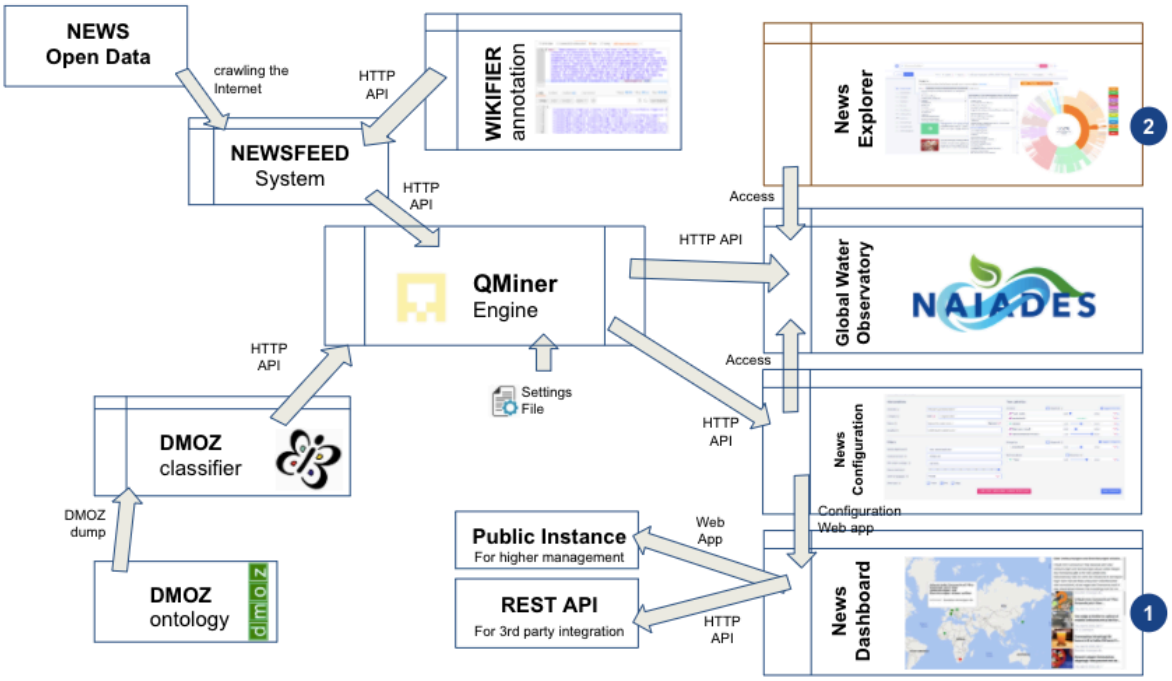

Figure 12: System architecture for the news monitoring system

Finally, in the following paragraphs, we will be discussing the architecture of the biomedical module that is focusing on the exhaustive exploration of water contamination information from scientifical research articles published worldwide and available through the MEDLINE biomedical open dataset (see Section 3.4 for more detail). The dataset is collected from the official FTP source made available by the North American National Library of Medicine (NLM) over an XML dump and uploaded to the Elasticsearch data management system through a python script.

Then the dataset is called over and HTTP API by the *SearchPoint* technology[26] to load the dataset and respective metadata, thus allowing for powerful Lucene-based queries and further interaction over a movable pointer (see Section 5.2.3 for more details on functionality). This will lead to the refinement of the search of information that can then be extended over the Biomedical Explorer, which feeds over the same dataset through *Kibana,* but also allows for the analysis of raw data, or the easy construction of data

[24] https://dmoz-odp.org/

[25] https://qminer.github.io/

[26] http://searchpoint.ijs.si/

visualisation modules from templates, and for an interactive data visualisation dashboard. The latter can be made publicly available through, e.g., iframe to be integrated in high-management KPI monitors.

Moreover, the *Biomedical Explorer* is available through an HTTP API for easy integration with 3rd party solutions. This architecture is illustrated by the comprehensive diagram in Figure 13.

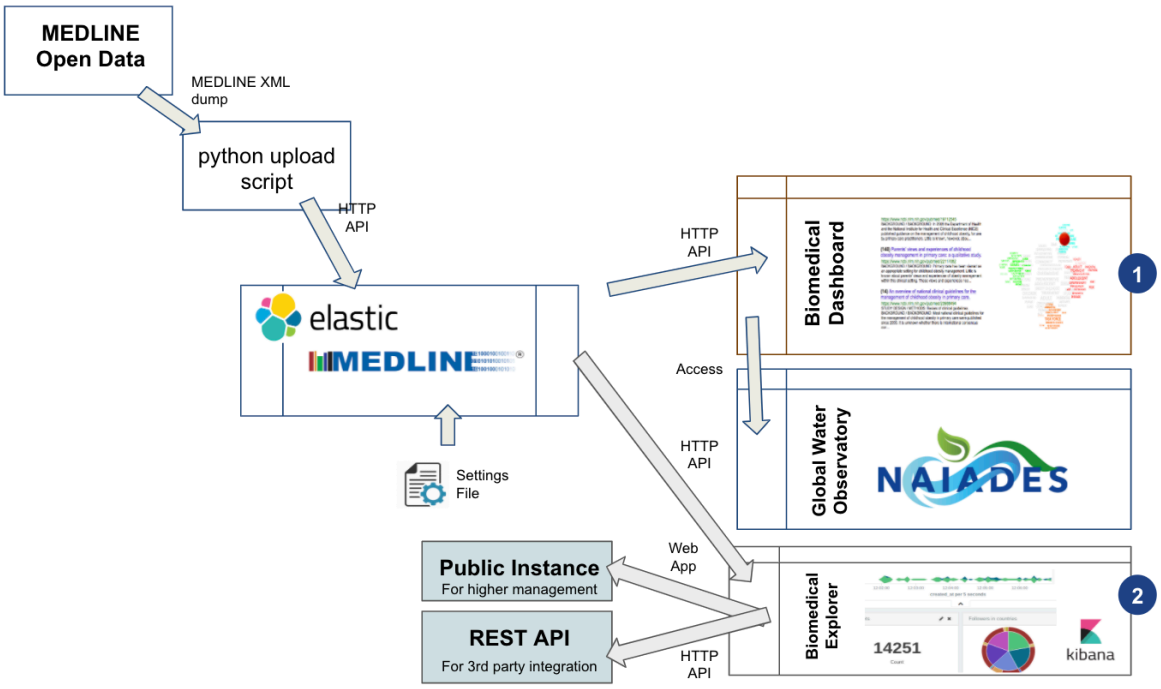

Figure 13: System architecture for the biomedical research explorer

5 Water Observatory Pilot 1

In the following section we will be presenting the first version of the pilot of the NAIADES Global Water Observatory, already available at naiades.ijs.si. It was developed during the first phase of the project, according to the change of work plan agreed within the consortium and described in Section 1. It will be updated by a second and final version in M30 and reported in D5.8.

5.1 Pilot description

Following the methodology and architecture described in Section 4, the NAIADES Global Water Observatory pilot 1 provides the user of the NAIADES platform, as earlier extensively discussed, with the global and local insight that can be transformed into business intelligence, and help companies to steer their strategies towards customer satisfaction. We will be describing the first pilot of this observatory through the already available verticals (or views) News – Indicators – Biomedical, first at the level of the specific dashboards that constitute the tabs in the online instance, and then by the extended exploratory instances, including public instances and APIs, for each of the three verticals.

There were several interactions with the policy-board that contributed to the identified common needs to clarify or facilitate difficulties. In that:

- the first interactions happened with the Alicante use-case to understand better how the dashboard could be more useful, and will write papers together about the obtained results
- also, we plan to scale to the other two use-cases, based on the success stories with Alicante
- this was also a topic of discussion at the plenary meeting in June 2020 where the consortium approved the general plan based on preliminary results and is planned to integrate a forthcoming hands-on workshop.

The further development of the NAIADES Global Water Observatory will be done in the context of discussions with the consortium and use cases to optimize usability, and the new version will include the results of those discussions.

5.1.1 GWO News Dashboard

The backend of the NAIADES News Monitoring Dashboard is powered by the Event Registry news monitoring system [36] that collects and annotates a stream of real-time news articles sourced over 100 thousand news publishers worldwide in 10+ languages (including Spanish, French and Romanian, covering the languages of the NAIADES use-cases). The details of this dataset are discussed in Section 3.1, but the usefulness of this view/vertical is already evident by what is available in pilot 1 as described in this section.

The news dashboard included in the NAIADES Global Water Observatory version 1 is a data stream of real-time news over a narrow topic, specific to the preselected research topics (see Figure 14). This specification is done over a set of filters that are accessed from the external news dashboard (see Figure 15). That selection of criteria for the news sample is shaping the news stream, deployed on the system by a change in the API call to the external dashboard that is managing the data to be exposed.

Besides the list of news in a real-time fashion, it also provides a visual representation of the main topics in the filtered news stream to provide fast access to the information before scrolling down the news Figure 14 shows an example of the News Dashboard, where the filtered news stream for the Alicante use case is exhibited. The listed news items are in Spanish language due to the fact that this restriction was chosen. Otherwise, the news stream would capture news in other languages and from other locations in the context of water scarcity and water contamination topics.

The News Dashboard comprises:

- A drop-down menu to select a country to which the news stream restricts;
- An ordered list of news articles composed by title (linked to the news source), a snippet of the news body and the URL to the publisher;

- A map view to where the news is originating, zooming when a country is selected;
- Three options to the speed of the exposed news stream – slow, normal, fast (including an option to pause);
- Choice over two topics – water scarcity and water contamination – that can include other topics.

Each use-case can choose its own live news source location and explore in the external dashboard the available sliders and filters in the configuration of news monitoring to set input to the news stream as fits best to their interests.

Figure 14: News dashboard at the NAIADES Global Water Observatory

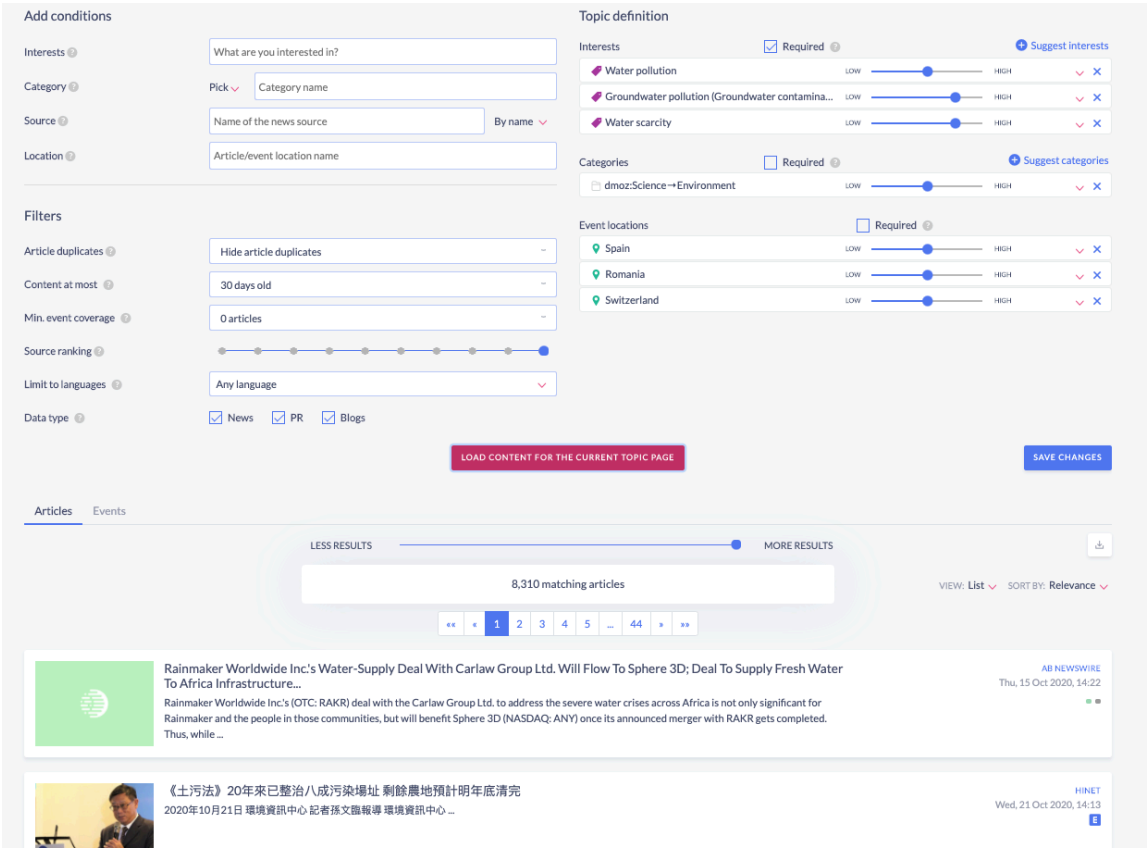

Figure 15: News dashboard configuration panel, showcasing filter options to fit the news stream to the exact needs of the use cases

5.1.2 GWO Indicator Dashboard

In the context of the analysis of global indicators, in the first version of this pilot we put together indicators that are sourced in the UN SDGs and respective supporting data sources, but also at the open dataset of UN. These datasets are fully described in Section 3.3, along with their expected outcomes and preliminary results.

For this pilot 1 we have chosen to exhibit water-related data sources that can be useful in the context of the work done at NAIADES. From SDG 6 indexes and UN data we prepared a dataset over years and the 3-letter country codes to be exposed on the Indicators Dashboard through D3JS-powered data visualisations. To allow the user to identify relations of interest between the indicators, we provide a drop-down menu that changes the Y axis according to the available data sources. The chosen datasets are:

- SDG 6.1.1 [SDG611] - Proportion of population using safely managed drinking water services [label: *population access to drinking water*]
- SDG 6.4.1 [SDG641] - Change in water-use efficiency over time [label: *water-use efficiency over time*]
- SDG 6.4.2 [SDG642] - Level of water stress: freshwater withdrawal as a proportion of available freshwater resources [label: *water stress level*]
- UN dataset 1 [UN1] - Fresh surface water abstracted [label: *Fresh surface water*]
- UN dataset 2 [UN2] - Falling Water [label: *Falling water]*
- UN dataset 3 [UN3] - Inland waters [label: *Inland waters*]

In a similar fashion to what happens in the News Dashboard, here we provide a dropdown menu to restrict information to one of the countries. As mentioned earlier, we allow the user to select what is presented in the axis of the visualisation and represented by the size of the balls, providing three dimensions of data representation. The visualisation is presented in an animation style, that can be controlled and reset over a slider and button, respectively.

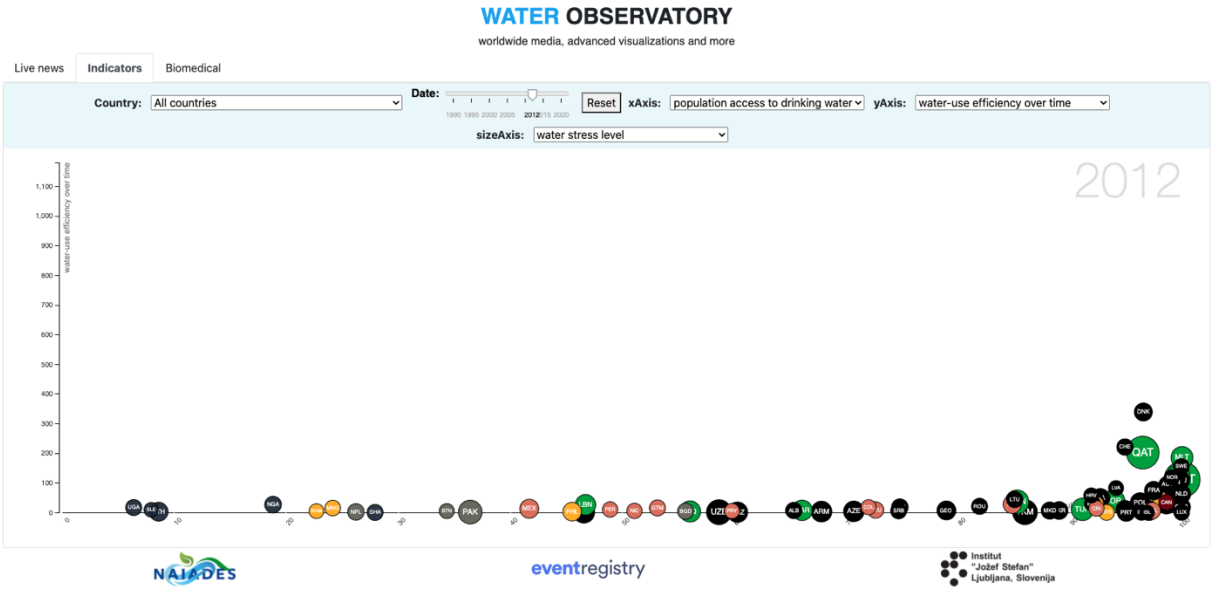

Figure 16: Indicators dashboard at the NAIADES Global Water Observatory

5.1.3 GWO Biomedical Dashboard

When using indexing services to search for information across a huge amount of text documents - MEDLINE being an example - we usually receive the answer as a list sorted by a relevance criteria defined by the search engine, and the answer we get is biased by definition. The Biomedical Dashboard is an interactive visual tool that helps highlight information we are looking for, by reindexing the results of the search based on further input from the user, selecting precomputed clusters/areas of interest. For example, when we enter a search term 'water contamination', the system performs an Elasticsearch-based search over the biomedical dataset, extracting groups of keywords describing different subgroups of results which are the most relevant, and not the most frequent terms. It provides the NAIADES user with a summary of the sampled documents content (e.g. we see groups of results about activity, control, drinking water, assess risks, etc).

Moreover, we can also use the Lucene language syntax to perform more assertive search as, e.g., to search for all the articles that were annotated with the MeSH Headings term "Trihalomethanes", by writing in the search box MeshHeadingList.desc:" Trihalomethanes". The movement of the mouse cursor over word-groups provides us with the relevance criteria of the search result, and exhibits as top results the articles we are interested in. For example, the article "Public drinking water contamination and birthweight, prematurity, fatal deaths, and birth defects." that occupied the position 106 is now in the 7th position. From the provided results we can read the title and the first lines of abstract and, when clicking on it, we are directed to the article in the browser at its PubMed URL location. Example from the Biomedical Dashboard is visible in Figure 17. We can also use the metadata related to hand annotated substances to the biomedical articles.

The Biomedical Dashboard comprises:

- A search box to introduce the main research topic;
- An ordered list of research paper mentions structured by title (linked to the research paper appearance in PubMed), a snippet of the abstract and the URL to PubMed;
- A tag cloud with the subtopics clustered by their proximity (based on frequency of being together in the overall Biomedical Dataset);
- A target pointer (the "searchpoint") that allows the user to go over the tag cloud and change the order of appearance of the articles;
- A pop-up of keywords appearing when it is moved by the user;
- A button to reset the position of the pointer;

An early version of this dashboard is already planned since early June 2020, to which NAIADES users and policymakers are providing feedback. Also, this technology was a point of discussion and potential further collaboration with the Arizona State University. The improvements of this dashboard to its following version in Pilot 2, will address the requests by the use cases.

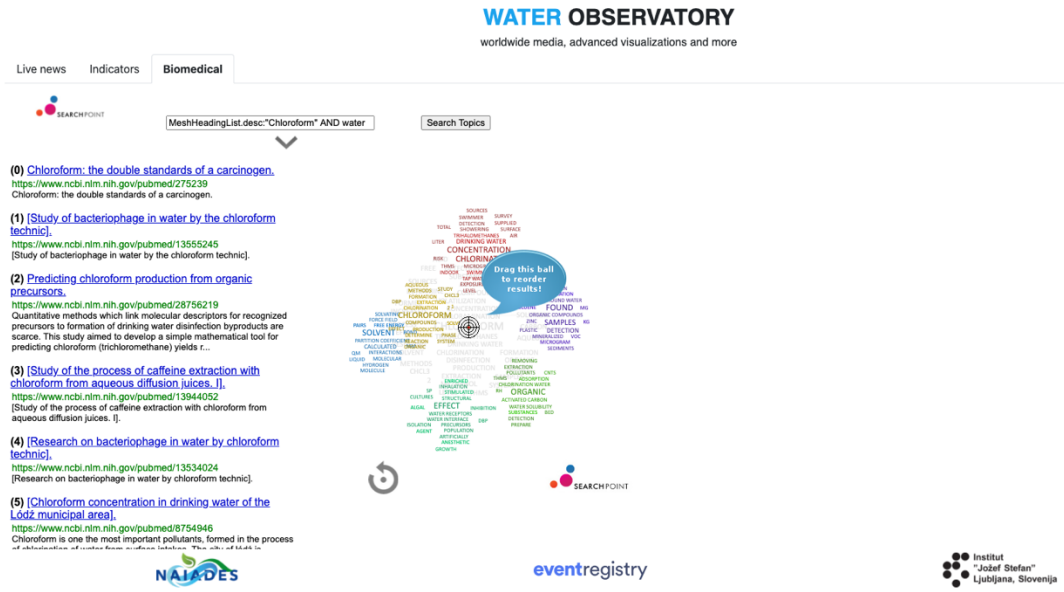

Figure 17: Biomedical dashboard at the NAIADES Global Water Observatory

5.2 Extended Exploratory Dashboards

The NAIADES Global Water Observatory is complemented with external dashboards to each of the verticals that can be used to further explore the data and get insight on a specific event or find evidence to support a certain KPI. In all the three cases we also make available public instances that can be used to either make publicly available certain aspects to customers and enhancing customer satisfaction. Moreover, we make available APIs that can be used to further leverage the available data in 3rd party applications on the NAIADES user side, enhancing the usability of the obtained key exploitable results.

5.2.1 External News Dashboard

This external news dashboard explores in full the data visualisation capabilities of the news engine system described in Section 5.1.1. The system is also able to group the news in events mentioned in these articles, allowing us to explore what is currently being covered in the media worldwide.

For each event, the NAIADES News Monitoring Dashboard provides extensive information, distinguishing between subtopics and perspectives in the stories relating to the news. Each event is based on a list of cross-lingual articles that describe the several aspects of it, including date, location, and impact on social media (i.e. Facebook). The system allows for the NAIADES user to visualize a real-time dynamic world map of news, and to explore a static set of these over a certain time window.

The NAIADES News Monitoring tool can find articles and events related to a particular entity, topic, etc. It enables the user to define topic pages where the parameters can be regulated in order to set up a monitoring system that fits the problem at hand. It can then be provided as a public instance based on the URI code that tracks that topic page where the monitoring of the news collection was defined. Example of a News monitoring dashboard is presented in Figure 18.

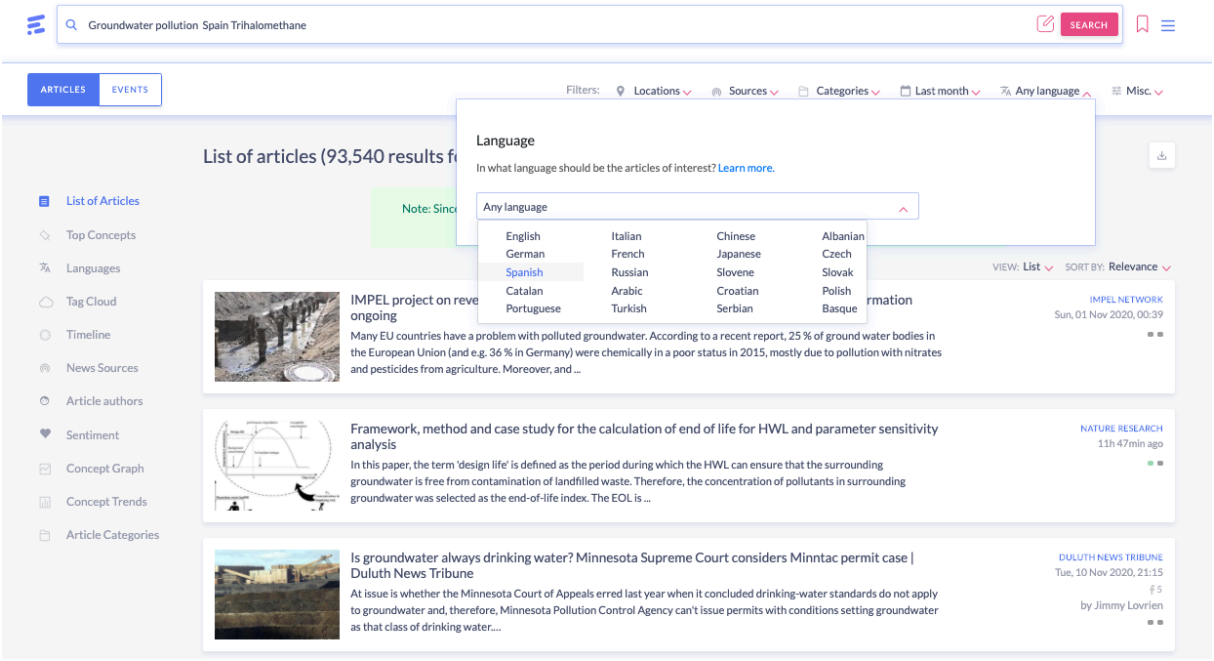

Figure 18: External news dashboard, showcasing the main aspects of the news and their impact in social media channels

The NAIADES user can also explore the evolution of the news (and with it, the media awareness) over a certain water-related topic in a timeline by looking at the sampled news articles as they were identified or updated during a selected period of time. In that context, the user can already identify when a certain topic became a global concern within the public, as are often cases of water contamination (see Figure 19). The charts can be downloaded as images for later use and the zoom option can be activated to explore the news in a particular time frame. Also, when right clicking over the chart, the user can retrieve the subset of news over the queried topic that day or add that to the search definitions.

Figure 19: Timeline of news articles analysis, where the water-related articles can be analysed from the perspective of the events that they integrate

Another particularly useful data visualisation module of this external news dashboard is the *Article Categories* that allow the user to explore the impact, of the sampled news over the topic of interest, in other domains with the quantification of that impact (see Figure 20).

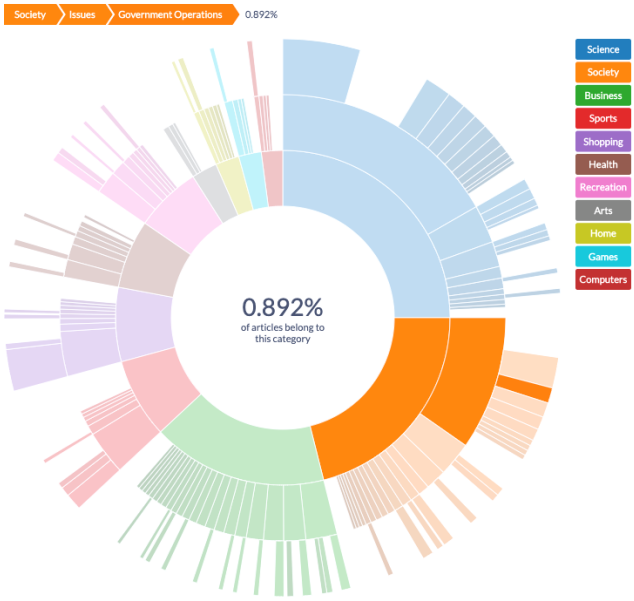

Figure 20: News categories interactive view, where the water-related articles can be analysed from the perspective of their impact in other areas of interest

5.2.2 External Indicator Dashboard

Although still in development, the NAIADES user can already access the external dashboard focused on indicators. As earlier mentioned in Section 4.2, this external dashboard is based on the *Elasticsearch* technology, where the data sourced from the open datasets described in Section 3.3 are stored and managed. This technology allows for powerful queries using Lucene language, and is accessed by the user through the add-on *Kibana* offering the user:

- an interface to explore the raw data and metadata loaded;
- templates to easily build visualisation modules from the loaded indicators, allowing the less technical user to explore the data loaded;
- the possibility to also use D3JS to build visualisations from code (see Figure 21);
- a dashboard construction tool where the user can configure the most useful data visualisation modules to, e.g., monitor a specific KPI;
- the *Timelion* time series data visualiser where the user can combine independent data sources and get further insights from the indicator data.

This tool also permits the administrator to set different levels of access to different types of users, that can come handy when we load sensitive data to be explored, as is the case of local data from the use cases. This is a scenario that we are still exploring, permitting a closer connection of the Global Water Observatory with the local priorities.

Figure 21: External dashboard for indicators, showing the potential to include D3JS code for more sophisticated visualisation modules

5.2.3 External Biomedical Dashboard

The freely available medical/scientific research dataset MEDLINE provides a large coverage of that worldwide research (over 26 million citations). It is recognized as an important source of information in the daily life of Healthcare professionals and can be very useful to explore details over, e.g., water contamination. NAIADES provides an interactive text-mining tool generating several visualisation modules enabling the user to extract meaningful information from that Biomedical Dataset.

To extract meaningful information from MEDLINE, NAIADES is using the underlying MeSH (Medical Subject Headings) ontology-like structure. Most of the articles in the MEDLINE biomedical dataset are hand-annotated by humans that assign to each a set of MeSH Heading descriptors. These allow us the exploration of a certain biomedical related topic, relying on curated information made available. The Biomedical Dataset, together with the MeSH annotation, is indexed with ElasticSearch and made available to analytics and visualisation tools.

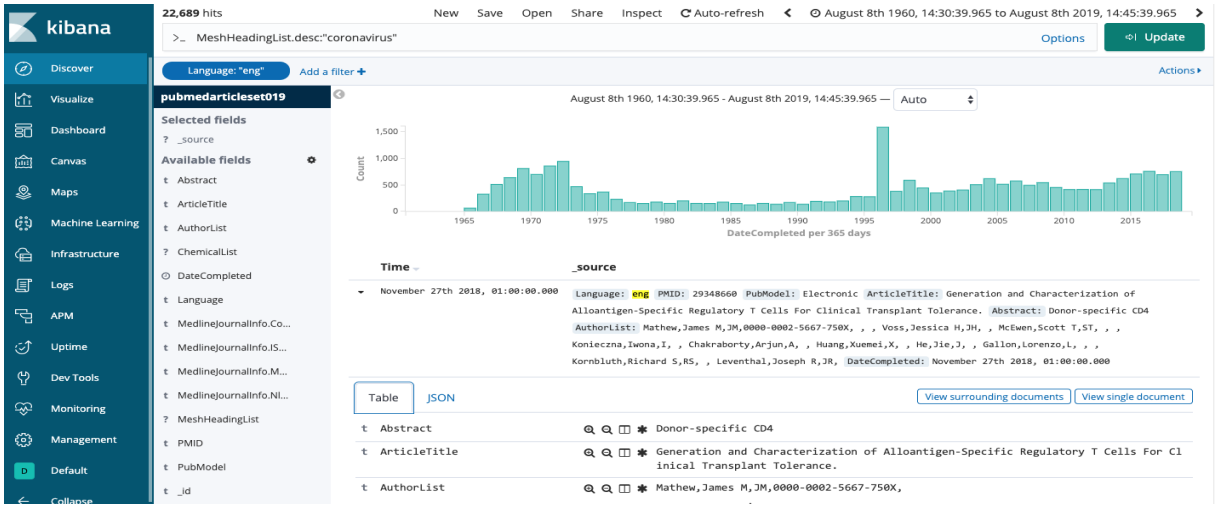

Figure 22: The further exploration of the details of the biomedical research on the Biomedical Dataset, enhanced by the powerful knowledge-base MeSH Headings

The Biomedical dashboard permits the user to profit from data visualisation modules that feed on an instance of the Biomedical Dataset built in Elasticsearch (again with Kibana being used for prototyping the needed visualisations). Also, it allows the NAIADES user to query the dataset and produce data visualisation modules on-the-fly, based on Kibana templates, that can later be included in a customised dashboard, designed to support the end-user workflow. This live dashboard can easily be set up as a public instance without the customization settings. Examples from the Biomedical Dashboard are presented in Figure 23.

The dashboard includes the Kibana menus *Discover* (for data exploration), *Visualize* (to construct data visualisation modules), and *Dashboard* (to visualise, manage and build dashboards based on the visualisation modules in *Visualize* built over the Biomedical Dataset as in *Discover*). Here we will also consider in pilot 2 the levels of data accessibility of different user levels, that also include the different access to available menus.

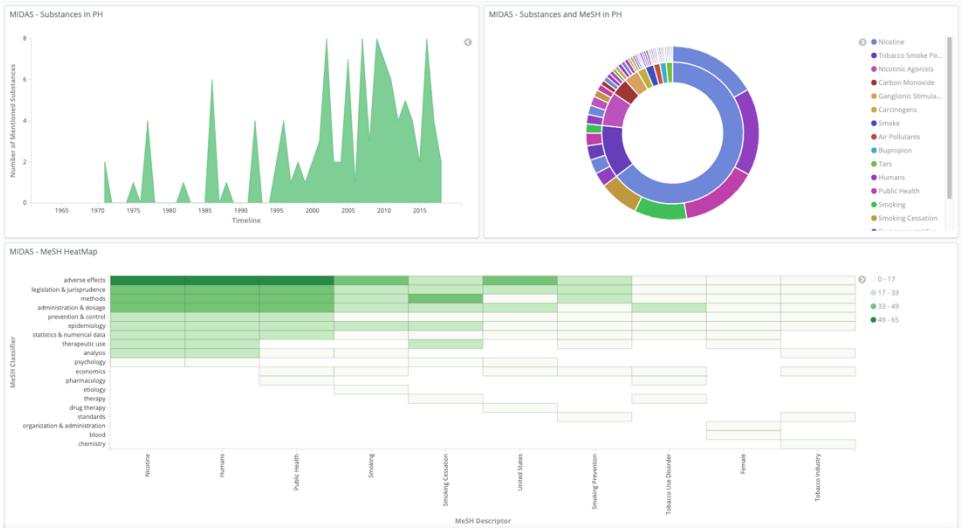

Figure 23: Biomedical science dashboard describing through interactive visual modules aspects of the published research on water-related diseases, chemical pollutants, etc.

6 Conclusions and Future Work

In this deliverable, reporting on the first half of the duration of the project, we present the comprehensive change of workplan, together with the main concepts in which the Task 5.4 will establish its Global Water Observatory. In that context, we have described related work and motivation, as well as the methodology adopted. We have described the first pilot, already available to the NAIADES users, and its architecture. In the data sources section, we have described the data ingested in pilot 1, but also the data we expect to ingest in the final pilot, together with a projection of expected outcomes and preliminary results achieved. That will help us better understand the value of each of the data sources and technological components to the NAIADES users and discuss with the latter on further development in a lean process. This first pilot takes into consideration the use case requirements released in D2.7, relating to general aspects of the platform.

The final pilot of the NAIADES Global Water Observatory, to be available in M30, will be providing the users with an integrated system capable to monitor in real-time the worldwide news and social media, as well as topics searched over the internet provided by Google Trends, regarding issues related to water in the context of NAIADES focus. The user will be able to explore a wide range of indicators and compare trends in a global and local level throughout a meaningful timeline. We will also enable the user to explore the impact of the weather (see Figure 24), as well as the available bodies of water, reusing EC-funded open datasets and initiatives. Moreover, the user will be able to explore in research and patents (not restricted to health-related topics) all matters on water contamination and best practices to solve common problems, but also the technologies that are becoming available and their impact in the market landscape. This final pilot will take into consideration the updated use case requirements to be released in D2.8 at M30. Furthermore, we aim to "localize" this Global Water Observatory, integrating some of the local data that can be provided by the NAIADES use-cases, and customizing news sources to their own priorities, as well as making available data exploration dashboards that allow for further insight and evidence-based policy.

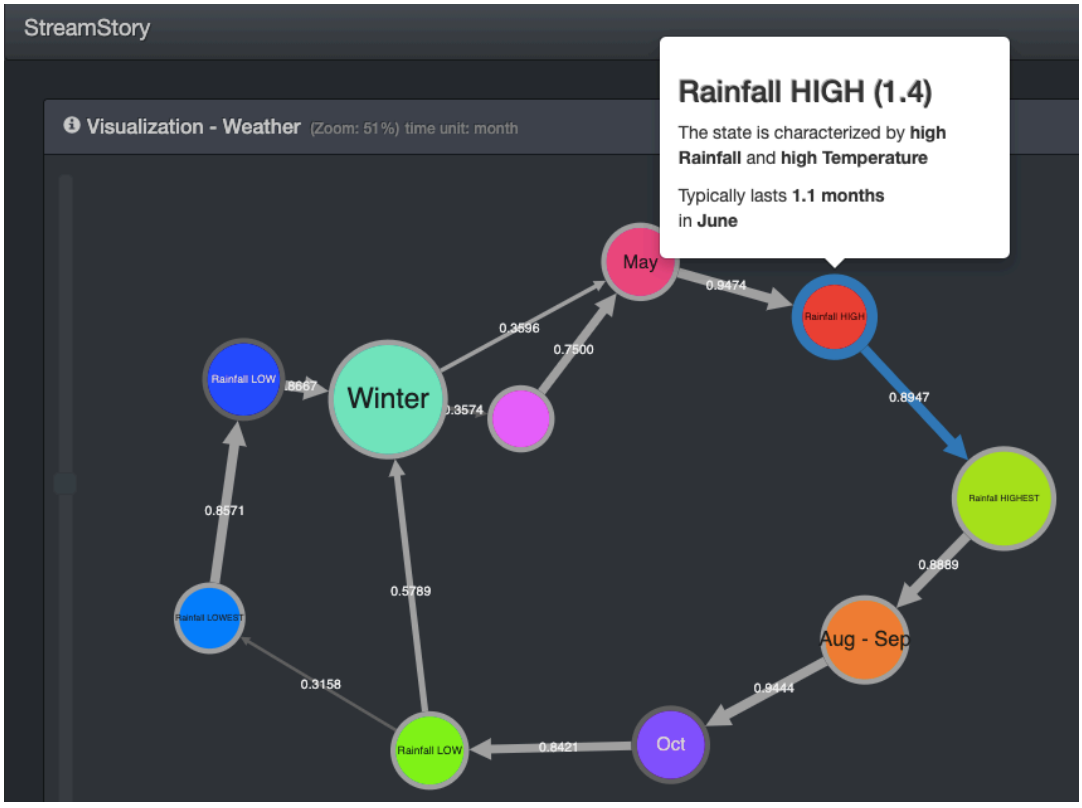

Figure 24: The multitime-series analysis of the weather parameters in the UK, using Markov chains in a complex data visualisation available through the Streamstory technology.

With the development of the ongoing research, the exploration of the potential of Pilot 1 with the NAIADES partners, and the preparation of the forthcoming Pilot 2, we aim to the following research questions:

1. Can we predict water shortages in regions of Spain? Can climate change be forecasted in what respects water availability?

In this we will explore the access to appropriate data, weather conditions in a timeline of events, water shortages and water contamination. We will also analyse the usefulness of news reports, and social media. Moreover, we will try to understand how local we can get in the sense of accessing meaningful insight from the available data granularity.

2. Can we identify water-related problems (e.g., shortage, contamination, salinity) from discussions in social media, news outlets or search engine queries (provided by Google Trends)?

In the context of these livestream data sources noise is a bottleneck and can lead to unreliable alarms. This fairly known problem can sometimes be solved with some smart engineering of the queries that are used to collect the data based on the domain knowledge. We will leverage the experience of the NAIADES partners to better fit our technology to this challenge allowing the users to, e.g., explore filters in the news monitoring system. One of the key aspects of this problem is how to define the trigger(s) that will sound meaningful alarms and how that can positively impact the business of NAIADES adopters.

3. Can biomedical research improve preparedness of providers to water contamination?

A natural answer to this question would be positive, having in mind that many of these problems have common ground and their solutions can derive from published best practices. Sometimes the hard part is to know the right questions to ask, leading to a fruitful exploration. Its success also depends on knowing priorities, ie, the local context, and steer the answers achieved on that direction. We will also investigate the trend identification automation from the publish research, an unsupervised problem that is itself an open research question at global text mining level.

4. Can local data contribute in a meaningful way?

Although the already observed lack of data access from use-case partners (that led us to substantially change the workplan of this task), and the amount of data needed to achieve useful insight, we consider that the local data available can help us partially customize the output of the observatory. We will be exploring the requirements for that positive impact, as well as the appropriate data that can be ingested, the optimal frequency limited to availability, aiming for the "localization" of the global observatory.

5. Can SDG 6 indicators be useful to UC providers?

Having in mind the granularity of most of these indicators, it is well understood already from the early results of Pilot 1 that some insight can be achieved. These insights are identified from the timeline of events that can be analysed simultaneously, even though the large granularity of the data. The new data sources that are being handpicked and explored already show from their description in Section 3.3 the aimed usefulness in understanding the larger picture over water-related topics in time We will be working further to understand in depth that usefulness, within ongoing and planned interactions with use-cases.

7 Bibliography

[1] European Commission, "European Green Deal," 2019. [Online]. Available: https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal_en. [Accessed 1 9 2020].

[2] European Commission, "Water Scarcity & Droughts in the European Union," 2019. [Online]. Available: https://ec.europa.eu/environment/water/quantity/scarcity_en.htm. [Accessed 2020 9 1].

[3] Our World in Data, "Water Use Stress," 2018. [Online]. Available: https://ourworldindata.org/water-use-stress. [Accessed 2020 9 1].

[4] United Nations Development Programme, "Goal6: clean water and sanitation," 2020. [Online]. Available: https://www.undp.org/content/undp/en/home/sustainable-development-goals/goal-6-clean-water-and-sanitation.html. [Accessed 2020 9 1].

[5] V. Blazhevska, "United Nations launches framework to speed up progress on water and sanitation goal," United Nations Sustainable Development, 2020.

[6] OECD, "Water governance in OECD countries: A multi-level approach," OECD, 2011.

[7] A. Akhmouch, C. Delphine and P. G. Delphine Clavreul, "Introducing the OECD principles on water governance," *Water International,* vol. 43, pp. 5-12, 2018.

[8] R. A. Freeze and R. L. Harlan, "Blueprint for a physically-based, digitally-simulated hydrologic response model," *J. Hydrol.,* vol. 9, p. 237–258, 1969.

[9] A. Ramamoorthi, "Snow-melt run-off studies using remote sensing data," *Sadhana,* vol. 6, p. 279–286, 1983.

[10] DHI, "The Digital Twin: What is it and how can it benefit the Water Sector?," 2019. [Online]. Available: https://blog.dhigroup.com/2019/06/06/the-digital-twin-what-is-it-and-how-can-it-benefit-the-water-sector/. [Accessed 2020 9 1].

[11] Idrica, "GoAigua: Smart Water for a Better World," 2020. [Online]. Available: https://www.idrica.com/goaigua/. [Accessed 1 9 2020].

[12] Blue Dot, "Blue Dot Observatory," 2018. [Online]. Available: https://www.blue-dot-observatory.com/. [Accessed 1 9 2020].

[13] Joint Research Center, "Global Surface Water Explorer," 2020. [Online]. Available: https://global-surface-water.appspot.com/#features. [Accessed 1 9 2020].

[14] UN Water, "SDG 6 Data Portal," 2020. [Online]. Available: https://www.sdg6monitoring.org/2020-data-drive/. [Accessed 1 9 2020].

[15] DHI Group, "Urban Water," 2020. [Online]. Available: https://www.dhigroup.com/areas-of-expertise/urban-water. [Accessed 1 9 2020].

[16] Smart Energy International, "The world's first water utility digital twins," 2019. [Online]. Available: https://www.smart-energy.com/industry-sectors/smart-water/the-worlds-first-water-utility-digital-twins/. [Accessed 1 9 2020].

[17] Bentley, "Oporto Water Utility Develops Technology Platform for Integrated Management of Urban Water Cycle," 2019. [Online]. Available: https://prod-bentleycdn.azureedge.net/-/media/files/documents/case-studies/cs_h2porto_ltr_en_lr.pdf. [Accessed 1 9 2020].

[18] Water World, "Digital Twins for Managing Water Infrastructure," 2020. [Online]. Available: https://www.waterworld.com/water-utility-management/smart-water-utility/article/14173219/digital-twins-for-managing-water-infrastructure. [Accessed 1 9 2020].

[19] J. M. Curl, T. Nading, K. Hegger, A. Barhoumi and M. Smoczynski, "Digital Twins: The Next Generation of Water Treatment Technology," *AWWA,* vol. 111, no. 12, pp. 44-50, 2019.

[20] Atkins Global, "Atkins launches digital twin survey platform," 2020. [Online]. Available: https://www.atkinsglobal.com/en-gb/media-centre/news-releases/2020/june/2020-06-22. [Accessed 1 9 2020].

[21] Idrica, "Digital Twin: implementation and benefits for the water sector," 19 2 2020. [Online]. Available: https://www.idrica.com/blog/digital-twin-implementation-benefits-water-sector/.

[Accessed 1 9 2020].

[22] J. Pekel, A. Cottam, N. Gorelick and A. Belward, "High-resolution mapping of global surface water and its long-term changes," *Nature,* vol. 540, no. 7633, pp. 418-422., 2016.

[23] Joint Research Centre, "The new dataset 1984-2019 is available to download," 2020. [Online]. Available: https://global-surface-water.appspot.com/download. [Accessed 1 9 2020].

[24] Blue Dot, "Blue Dot Water Observatory Platform," 2018. [Online]. Available: https://water.blue-dot-observatory.com/2496/2018-10-02. [Accessed 1 9 2020].

[25] D. Buttler, "When Google got flu wrong," *Nature,* vol. 494, p. 155–156, 2013.

[26] V. Lampos, A. Miller, S. Crossan and C. Stefansen, "Advances in nowcasting influenza-like illness rates using search query logs," *Scientific Reports,* vol. 5, p. 12760, 2015.

[27] Twitter, "Twitter for Academic Research," 2020. [Online]. Available: https://developer.twitter.com/en/solutions/academic-research. [Accessed 1 9 2020].

[28] TwitterDev, "Do More with Twitter Data," 2020. [Online]. Available: https://twitterdev.github.io/do_more_with_twitter_data/timeseries.html. [Accessed 1 9 2020].

[29] M. P. Inna Novalija and D. Mladenic, "Towards Social Media Mining: Twitterobservatory," *SIKDD,* 2014.

[30] World Bank, "Level of water stress: freshwater withdrawal as a proportion of available freshwater resources," 2019. [Online]. Available: https://data.worldbank.org/indicator/ER.H2O.FWST.ZS. [Accessed 1 9 2020].

[31] Microsoft Academic Graph, "MAC Citation Matrix," 2020. [Online]. Available: https://academic.microsoft.com/topics/39432304,524765639. [Accessed 1 9 2020].

[32] European Data Portal, "Open water data on the European Data Portal," 2020. [Online]. Available: https://www.europeandataportal.eu/en/highlights/open-water-data-european-data-portal. [Accessed 1 9 2020].

[33] L. Stopar, P. Skraba, M. Grobelnik and D. Mladenic, "Streamstory: exploring multivariate time series on multiple scales," *IEEE transactions on visualization and computer graphics,* vol. 25, no. 4, pp. 1788-1802, 2018.

[34] Elasticsearch, "Elasticsearch," 2020. [Online]. Available: https://www.elastic.co/elasticsearch/. [Accessed 1 9 2020].

[35] Elasticsearch, "Kibana," 2020. [Online]. Available: https://www.elastic.co/kibana. [Accessed 1 9 2020].

[36] G. Leban, B. Fortuna, J. Brank and M. Grobelnik, "Event registry: learning about world events from news," *Proceedings of the 23rd International Conference on World Wide Web,* pp. 107-110, 2014.

[37] J.-F. Pekel, A. Cottam, N. Gorelick and A. S. Belward, "High-resolution mapping of global surface water and its long-term changes. Nature 540„" vol. 540, pp. 418-422, 2016.