1st WHOW Co-Creation Meeting

Project Name: WHOW (Water Health Open knowledge)

Agreement number: INEA/CEF/ICT/A2019/2063229

Action No: 2019-EU-IA-0089

Co-financed by the Connecting Europe
Facility of the European Union

Table of Contents

1	Introduction	31.1
	Project overview	31.2
	European value	31.3
	Milestones and timeline	41.4
	Project partners	42
	European legal framework	43
	Use cases	73.1
	Water for human consumption & Infectious diseases	73.1.1
	Water for human consumption	73.1.2
	Infectious diseases	83.2
	Bathing waters	93.2.1
	Legal framework	93.2.2
	Opportunities	103.3
	Bioaccumulation	103.3.1
	Legal framework	103.3.2
	Monitoring activities	113.3.3
	Datasets	113.3.4
	Opportunities	113.4
	Ostreopsis ovata	113.4.1
	Context	123.4.2
	Monitoring activities	123.4.3
	Datasets	123.4.4
	Opportunities	133.5
	Extreme events	133.5.1
	Monitoring activities	133.5.2
	Datasets	134
	Technical Architecture	134.1
	Objective	144.2
	Requirements	144.3
	Existing data architectures	144.4
	WHOW Toolkit and pluggable knowledge graphs	154.5
	Challenges	164.6
	Opportunities	175
	Datasets	175.1
	Data Modelling – common patterns	185.2
	Ontologies	195.3
	General observations	19

1 Introduction

On May 10th 2021, the first online meeting of the WHOW co-creation programme took place, with over 80 participants. It was an opportunity to share the project objectives and challenges with the interested stakeholders and engage them in the relevant project activities.

The objective of this document is to provide an overview of the WHOW project and the results of the meeting during which the project uses cases, technical architecture and relevant datasets were discussed.

This document should serve as the basis for continuing the dialogue with our stakeholders and gather their feedback and suggestions to move forward on the various topics elaborated in the following sections.

1.1 Project overview

The WHOW project aims to foster the creation of the first European knowledge graph on water consumption and quality, health parameters and dissemination of diseases to be reused for advanced analysis and development of innovative services.

The project will leverage the use of the Open Linked Data paradigm. Water related datasets from European countries and Copernicus (the European Union's Earth observation programme) will be integrated and made available for re-use on data.europa.eu, the official portal for European data. Health related datasets from Italy will be identified and linked to specific water datasets.

WHOW will target use cases in the creation of the knowledge graph, identifying and integrating the relevant set of indicators for Sustainable Development Goals (SDGs), and will leverage a co-creation programme where interested stakeholders and users will be engaged from the initial phases of the project.

The initiative will support the Public Open Data Digital Service Infrastructure by helping to boost the development of information products and services based on the re-use and combination of environmental data and health data on disease dissemination.

1.2 European value

In Europe, legislative measures aim to protect water and its ecosystems to guarantee sustainable water usage and to reduce risks for human and living beings using a cross-domain approach covering physical, biological, chemical parameters of single water bodies and of the surrounding environment.

Wider availability and higher quality of data on water pollution, water consumption, biodiversity, health parameters can help in better assessing the issues at stake and foster the development of innovative services, allowing decision makers and communities to make more informed decisions on policies and lifestyles.

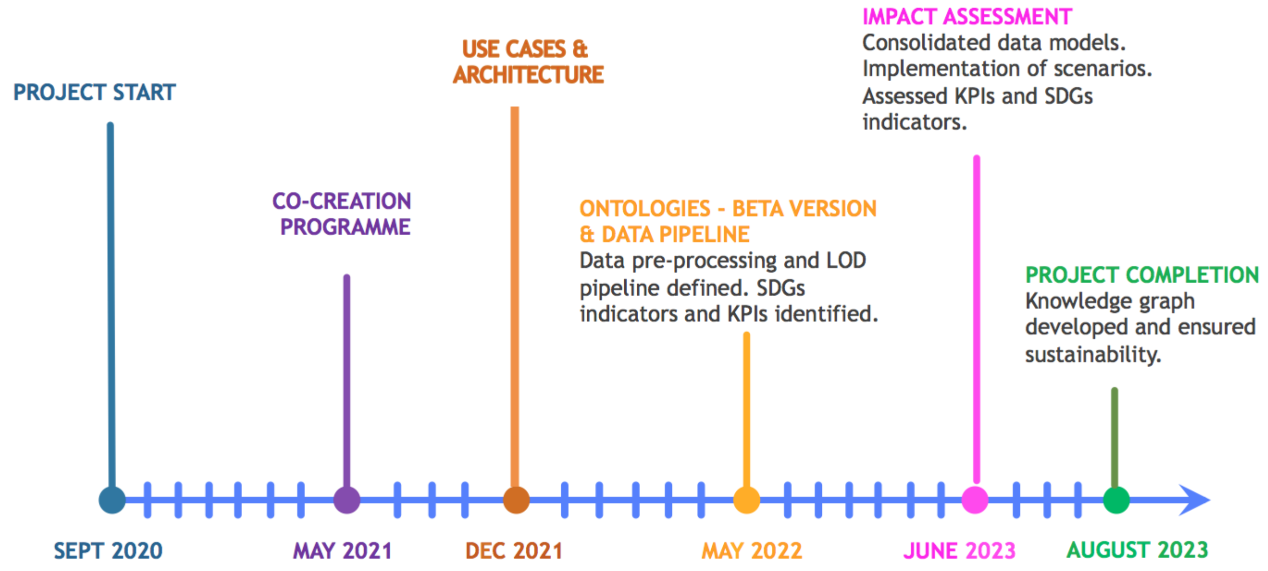
Europe is still struggling in providing a holistic view on these topics with data fully harmonized in terms of semantics, formats and licences. Even at the national level, datasets from different administrative levels show a fragmented context.

WHOW will address these issues by creating of a framework of data, data models and supporting services to provide a consistent view on water and health data to be effectively re-used to foster economic growth and better decision-making.

The framework will be a powerful tool for promoting transparency and attract experts in these areas that will actively participate in the co-creation of the WHOW framework. It should serve as reference model also at the international level.

1.3 Milestones and timeline

Launched in September 2020, the WHOW project will focus on the following milestones and related deliverables that are discussed in the next sections of the document.



1.4 Project partners

1. Celeris Advisory Limited (Ireland)
2. Istituto Superiore per la Protezione e la Ricerca Ambientale (ISPRA)
3. Azienda Regionale per l'Innovazione e per gli Acquisti S.p.A. (ARIA S.p.A)
4. Consiglio Nazionale delle Ricerche – Istituto di Scienze e Tecnologie della Cognizione (CNR-ISTC)

2 European legal framework

Water resources protection has been on the agenda of the European Commission since the first wave of European water legislation began with standards for the rivers and lakes used for drinking water in 1975, and culminating in 1980 with the setting of binding quality targets for drinking water. EU legislation included quality objectives for fish waters, shellfish waters, bathing waters and groundwater. Its main instrument for emission control was the Dangerous Substances Directive (1976).

In 1988, a number of potential improvements to this legislation were identified, resulting in a second wave of water legislation in the 1990's, namely, the urban Waste Water Treatment Directive, providing for secondary (biological) wastewater treatment, and even more stringent treatment where necessary, and the Nitrates Directive, addressing water pollution caused by nitrates used in agriculture. Both directives were adopted in 1991. In particular, the Nitrates Directive mandates EU member states to designate as vulnerable zones all land areas draining into waters that are or are likely to become affected by pollution ("nitrate-sensitive" areas). Member states are also called upon to establish codes of good agricultural practices binding on farmers and to implement programmes of measures relevant to the prevention of non-point source pollution. Monitoring obligations are also provided for in this directive.

Other developments were a Directive for Integrated Pollution and Prevention Control (IPPC), adopted in 1996, addressing pollution from large industrial installations, and a Drinking Water Directive adopted in 1998, reviewing quality standards and provisions for tightening them where necessary. This Directive, which was amended in 2003, 2009 and 2015, has been recently revised to provide for better access to safe drinking water to all, as well as better public information. Directive (EU) 2020/2184 on the quality of water intended for human consumption (recast) entered in force on 12 January 2021, and Member States have two years to transpose it into national legislation. The Directive lays down the essential quality standards at EU level. A total of 48 microbiological, chemical and indicator parameters must be monitored and tested regularly.

In the mid-1990's, a more global approach to water resources management was considered in order to address the increasing demand by citizens and environmental organizations for cleaner rivers and lakes, groundwater and coastal beaches. Thus, the commission initiated a process of consultation leading to a widespread consensus that, while considerable progress had been made in tackling individual issues, the current water policy was fragmented, both in terms of objectives and of means. All parties agreed on the need for a comprehensive framework.

In response to this, the commission presented a Proposal for a Water Framework Directive with the following key aims:
(i) expanding the scope of water protection to all waters, including surface water and groundwater, considered within defined "river basin districts";
(ii) achieving "good status" for all waters by fixed deadlines, through the setting and coordination of objectives within each river basin, and the introduction and implementation of programmes of measures;
(iii) establishing a river basin management plan for each river basin district. The plan is essentially a detailed account of how the objectives set for a river basin (ecological status, quantitative status, chemical status and protected area objectives) are to be reached within set deadlines. The plan provides an analysis of the characteristics of the river basin, a review of the impact of human activity on the status of waters in the basin, an estimation of the effects of existing legislation and of what remains to be done to meet the objectives, and an indication of the measures required to fill the gap. An economic analysis of water use within the basin must be carried out in order to facilitate an assessment of the cost-effectiveness of the various measures;
(iv) a "combined approach" to emission limit values (effluent standards) and water quality standards requiring the adoption of the best available technology for point-sources and, on the effect side, the coordination of environmental objectives with a view to achieving the overall status objective;
(v) water prices reflecting the true cost and acting as an incentive for sustainable water use, thus helping to achieve the environmental objectives set under the directive;
(vi) public participation in river basin planning and management, to balance the interests of various groups and ensure plan implementation and enforceability;
(vii) streamlining legislation by repealing seven of the first wave directives.
The Water Framework Directive, which for the first time provides for the management of surface water and groundwater at the EU level in a comprehensive manner, was adopted in 2000. Since then, EU Member States have taken action to align their respective legal systems and institutional arrangements to its requirements and to those of the "second wave" directives which remain in force.

Since the implementation of the Water Framework Directive has resulted to be more complex than originally envisaged, in 2001 the Member States have agreed with the Commission and Norway on a common implementation strategy that aims at developing a common understanding of approaches and informal technical guidance on best practices, thereby reducing the changes of wrong application. The strategy is largely based on the sharing of information and experiences and the production of guidelines on technical issues.

To complete the European legal framework, the European Parliament and the Council adopted the Groundwater Directive (2006/118/EC) and the Floods Directive (2007/60/EC; while Directive 2008/105/EC lays down environmental quality standards for priority substances and other pollutants.

The Groundwater Directive aims at defining criteria for assessing the good chemical status and groundwater quality trends, identifying significant and sustained upward trends in the concentration of pollutants and defining starting points for trend reversal. For groundwater bodies which are considered to be at risk pursuant to the analysis of pressures and impacts to be carried out under the Water Framework Directive, member states must establish threshold values and report on them to the European Commission.

The Water Framework Directive stipulates that the good quantitative and chemical status objective for groundwater is to be arrived at through the prevention or limitation of the input of pollutants and the maintenance of a balance between groundwater abstraction and recharge. EU Member States have the duty to implement all the measures necessary to “reverse any significant and sustained upward trend in the concentration of any pollutant resulting from the impact of human activity in order progressively to reduce pollution of groundwater, and to establish programmes for monitoring the status of groundwater. The Directive prohibits the direct discharge of pollutants into groundwater.

The Floods Directive aims to reduce and manage the risks that floods pose to human health, the environment, cultural heritage and economic activity. It requires member states to carry out preliminary assessments to identify the river basins and associated coastal areas at risk of flooding and, for the areas so identified, to develop flood risk maps and establish flood risk management plans that take into consideration development in the long term, as well as the possible effects of climate change. These plans are to be coordinated with the river basin management plans prepared under the Water Framework Directive.

A “fitness check” of European directives was conducted in order to assess their relevance, effectiveness, efficiency and coherence. In December 2019, the Fitness Check concluded that water legislation is broadly fit for purpose, with room for improvement related to investments, implementation, integrating water into other policies, chemical pollution, administrative simplification and digitalisation. In 23 October 2020, an Inception Impact Assessment was launched to address the findings of the Fitness Check in relation to chemical pollution and the legal obligation to regularly review the lists of pollutants affecting surface and groundwater.

EU Member States have made significant progress towards the transposition of the Water Framework Directive and other EU directives into their legislation. Among them, Belgium, France, England and Wales and Germany have amended their water legislation or adopted new legislation between 2002 and 2004. In a nutshell, this legislation provides for enhanced resource planning within formally identified river basin districts, and introduces mechanisms and procedures for public participation in this exercise. It further provides for the establishment of environmental objectives and programmes for monitoring the water status, and for the adoption and implementation of programmes of measures within the river basin district framework. The basin management plans must include all the information specified in the Water Framework Directive.

3 Use cases

3.1 Water for human consumption & Infectious diseases

The first use case concerns the possible correlations between water intended for human consumption monitoring data and reported infectious diseases. It has been conceived since in recent years several interconnections have been detected. The most recent is connected to Norovirus, detected in local aqueducts next to Garda lake, in 2018.

3.1.1 Water for human consumption

Thanks to EU directive 93/1998, waters for human consumption are defined according to their use into:

1. All **water** either in its original state or after treatment, **intended for drinking, cooking, food preparation or other domestic purposes**, regardless of its origin and whether it is supplied from a distribution network, from a tanker, or in bottles or containers;
2. All **water** used in any food-production **undertaking for the manufacture, processing, preservation or marketing of products or substances intended for human consumption** unless the competent national authorities are satisfied that the quality of the water cannot affect the wholesomeness of the foodstuff in its finished form.

3.1.1.1 Legal framework

This use case has the advantage of being regulated by EU directive 93/1998 and further amendment (up till the most recent directive 2184/2020). Such directive determines in detail:

- definitions mentioned above,
- risks associated with the management of these waters,
- monitoring activities,
- microbiologicals, chemicals and physicals parameters to be monitored
- minimum frequency of sampling and analyses
- specifications for the analysis of parameters

Directive 98/83/EC has been adopted in Italy by Legislative Decrees n. 31/2001, 27/2002 and further amendments. Such regulation has been furtherly adopted at regional level stating guidelines for application of Legislative Decrees mentioned above (Circular 03/16/2004 n.15 D.G. Health 15 / SAN / 2004 and further amendments).

Assuming a general member states adoption of the EU directives above, the use case stands under a semantic umbrella that facilitates the comparison among the measurements collected along the different European regions.

3.1.1.2 Datasets

In Regione Lombardia ATS (Local healthcare authorities) have an important role in monitoring. ATS in compliance with the european, national and regional regulatory framework, are responsible for taking and analysing samples, and collecting reports and related measurements in the territorial

scope of competence. These local datasets are then gathered by Regione Lombardia and stored in a central database.
Such datasets are based on a general concept of sample report:

- Sampling point
- Type of sampling point (such as drinking water, aqueducts, wells),
- Sampling point address and coordinates
- Sample date
- Parameter (code and description),
- Unit of measure,
- Value,
- Acceptability threshold

Such datasets are NOT available in open data.

3.1.2 Infectious diseases

Although infectious diseases must be reported all over Europe, the regulatory framework, unlike the water for human consumption, is more fragmented at European level.

For the analysis of the current use case the following resolutions have been taken in account:

- Regional resolution 26/02/2015 n. X/3190 «Interventions of surveillance, prevention, prophylaxis and control of infectious diseases in view of Expo 2015»
- Regional resolution 17/12/2018 n. XI/1046 Annex H «Methods of reporting and notification of infectious diseases and investigations for epidemiological purposes»

According to regional regulation, in Lombardy infectious diseases must be promptly reported by GPs, pediatricians, hospital/health care facilities doctors through a regional platform at the time of the suspected diagnosis according to clinical / laboratory criteria and in any case as soon as possible.
Regional regulation specifies the list of diseases, including those related to poor water quality (eg. Legionellosis, Hepatitis, Food poisoning), where it has to be applied and how such cases have to be reported.
The file opened into a regional platform is connected to a patient, to a temporal and territorial context, and linked to different information sheets according to the pathology.

All files collected into the regional platform are CLOSED for patient's privacy protection.

An aggregate representation of such data is published as infectious diseases open dataset at Regione Lombardia's Open Data Portal, on the Epidemiological observatory's dedicated site.
The dataset contains the number of infections reported by gender, age cluster, pathology and year of first symptoms for the whole regional territory. No territorial drills are on.

3.1.2.1 Law and privacy protection

Since the use case had positive feedback from the co-creation group, and main datasets needed for implementation are closed, the first actions to be taken are targeted to open a dialog with data providers and to start feasibility studies, where needed.
Regarding water for human consumption, the dataset do not contain personal data so no privacy issues would concern data providers.
Unlike it, a law and privacy study would be needed for infectious diseases dataset in order to understand the feasibility of an extension of actual open datasets and of new datasets opening.

The new EU directive 2020/2184 takes into consideration the risks related to the use of water resources and not, as in the past, mainly the conservation of the resource. The same directive also requires greater transparency towards citizens, so it can foster the opening of new datasets from both public and private bodies in charge of monitoring and managing water networks (e.g. AMAP spa, MM spa).

As an alternative data source for health impact, the ISTAT (Italian Statistic Institute) dataset on the *causes of death* can fit to the purpose. It is open but updated within 24 months from death events.

3.1.2.2 Opportunities

Opportunities are different for different targets:

- Citizen awareness (common People),
- Data driven policy and comparison (Public bodies),
- Research (Science, Journalism),
- Data driven business (enterprises).

3.2 Bathing waters

The second use case is centered on bathing waters and can be developed in a two-fold approach. A first one where, like the first use case, health diseases evidences (e.g. dermatological effects on skin, diarrhoea or stomach problems if swallowed, etc.) can be correlated to bathing waters bad quality reports.
A second opposite approach is to consider what bathing water's bad quality can be triggered by, such as nitrates' extensive use in agriculture or wastewater treatment plant spills and untreated water.

3.2.1 Legal framework

For the risks connected to bathing waters, the EU's Bathing Water Directive (and its revision 2006/7/CE) drives monitoring activities for European countries (parameters, monitoring periods, acceptability thresholds) and facilitates the comparison among the measurements collected in different regions.
The actual directive is scoping the monitoring to types of bacteria that indicate pollution from sewage or livestock (e.g. Escherichia coli and intestinal enterococci). Directive addresses adequate monitoring to cyanobacterial proliferation risk as well.

In Italy, the above directive has been adopted by Legislative Decrees 116/2008 and Health Ministry decrees 30/03/2010 and 19/04/2018.

3.2.1.1 Datasets

In Regione Lombardia monitoring is up to ATS (Local healthcare authorities). ATS are responsible for taking and analysing samples, and collecting all related measurements in the territorial scope of competence. These measurements are taken at least every month during bathing season from about 250 sampling points, then gathered by Regione Lombardia and stored in a central database, the same storing water for human consumption sampling reports. The measurements database is not available in open data. However, the Ministry of Health gathers such datasets from the regions, and provides the publication on national Water Portal with Open license CC-BY 3.0. At the moment, datasets are human readable, but not machine-readable.

3.2.2 Opportunities

Same as before, opportunities are different for different targets:

- Citizen awareness (common People),
- Data driven policy and comparison (Public bodies),
- Research (Science, Journalism),
- Data driven business (enterprises).

3.3 Bioaccumulation

The bioaccumulation is a process that occurs when a chemical substance is accumulated in the tissues of an organism through any possible pathways of absorption.
This storage is not balanced by any losing process and it is quantified by the bioaccumulation factor, defined as the ratio between the concentration of the chemical element in the organism and the concentration in the surrounding environment. The bioaccumulation of pollutants can occur directly from the environment in which the organism lives, through the respiratory surfaces and / or the skin. For aquatic organisms, the bioconcentration factor is typically referred to the absorption of the pollutant from water and can be easily estimated by numerical models using pollutant concentrations as input.

Human health could be affected by bioaccumulation in case of direct contact with contaminated water resources or by biomagnification. Biomagnification occurs larger organisms (as predators or humans) feed upon the already contaminated organism lives and in turn absorb elements into their own tissues at a higher concentration. The more contaminants are eaten, the more pollutants it will have in the new organism. Because the amounts of contaminants become more and more concentrated at each trophic level, apex predators (and also humans) could risk to gain fatal levels of accumulation in their bodies. As a result, the main effects on human health regards the interaction between chemical substances and tissues, from gastrointestinal or respiratory disease, poisoning, tumours. But there are also secondary effects related to the inability to use the fishery resource for human consumption in case of highly level of contamination.
A large number of chemical substances can activate a bioaccumulation process, from industrial waste, to pesticides, from plastics in water to drugs or antibiotics for human or animal consumption.

3.3.1 Legal framework

A set of recent European directives looks to the water and in particular to the sea with an holistic approach. In this sense, the water bodies are not separately evaluated by chemical, biological and physical aspects.

The European Directive 2000/60/CE (Water Framework Directive, WFD) defines waters as “not a commercial product like any other but, rather, a heritage which must be protected, defended and treated as such”. The main objectives of the Water Framework Directive are to protect all high status waters, prevent further deterioration of all waters and to restore degraded surface and ground waters to good status, according to the new concepts of Good ecological status and Good chemical status, by establishing the environmental quality standards for a set of chemical, biological, hydromorphological elements. In this sense, the Water Framework Directive is complementary to other EU regulations such as the REACH regulation on chemical elements, the directive on industrial emissions and regulations on pesticides.

According to the European Directive 2008/56/CE (Marine Strategy Framework Directive, MSFD) “the marine environment is a precious heritage that must be protected, preserved and, where practicable, restored with the ultimate aim of maintaining biodiversity and providing diverse and dynamic oceans and seas which are clean, healthy and productive"

This Directive defines eleven qualitative descriptors which describe what the environment will look like when the Good Environmental Status has been achieved and to do this, each Member State must implement a marine strategy for its marine waters, in cooperation with other Member States sharing the same marine region.

3.3.2 Monitoring activities

Following the above-mentioned directives, each EU Member State carries out a monitoring program in relevant points of each water body to observe the concentration of identified and alternative substances. National and European data portals collect these data usually every year.

In Italy, Regional Agencies provide these data that are collected by ISPRA and forwarded to the European dedicated portal (such as EIONET).

Some research institutions are involved in the observation of the presence of chemical substances in the tissues of fishes.

3.3.3 Datasets

SIC www.db-strategiamarina.isprambiente.it/app/#/
CDR-Eionet https://cdr.eionet.europa.eu
WISE https://water.europa.eu/marine
REACH https://echa.europa.eu/it/information-on-chemicals/registered-substances

3.3.4 Opportunities

The application of this use case could provide several opportunities. First of all, it could push for the opening of some useful datasets, in particular data on concentration of substances in organisms and the number of infection/diseases due to bioaccumulation (in terms of cases, accesses to medical care, health expenditure). Many are the opportunities for stakeholders.

Citizens can have clear information about environmental status and risks for human health, while the business world related to fishery and tourism can have data to identify areas for aquaculture or for human leisure. Public authorities can better monitor critical status, while researchers can define accurate models for the transmission mechanism between water end organism, or for clinical studies.

3.4 Ostreopsis ovata

Ostreopsis ovata is a microalgae belonging to the dinoflagellate group, hailing from tropical and subtropical regions and found in recent years also in temperate zones and in many Mediterranean countries, in particular in areas characterized by poor hydrodynamic conditions and shallow waters (eg closed bays). In optimal environmental conditions and with high value of sea temperature, the number of cells can rapidly increase until reaching giving origin to blooms.

The toxic microalga Ostreopsis ovata has been detected for the first time in Italian waters in 1994 along the Tyrrhenian coast of Lazio, while in previous years was observed in other Mediterranean locations such as France, Lebanon, Spain, since 1979. Two potentially toxic species of Ostreopsis are known to be present in the Mediterranean Sea: O. ovata and O. fattorussoi. Another non toxic species is O. siamensis

Along the Italian coast, intense and recurrent blooms of O. ovata occurred in several locations of the Tyrrhenian Sea (Sicily, Lazio, Tuscany, Liguria), Ionian Sea and Adriatic Sea (Puglia, Marche, Gulf of Trieste) with toxic effects on humans and benthic organisms such as limpets, mussels, olothurians, sea urchins (suffering or mortality).

The blooms have been associated with phenomena of human intoxication whose symptoms (cough, irritation of the upper airways, muscle or joint pain, conjunctivitis, rhinorrhea, fever due to the production of ovatoxins - OVTXs) could disappear spontaneously in the following 24-72 hours.
In tropical regions, lethal case following the ingestion of contaminated fish products have been registered to the production of palytoxin (PLTX).
In the Summer 2005 along the Ligurian coast, hundreds of people required medical attention because of respiratory problems and fever which apparently was caused by the exposure to the algal toxins through the marine aerosol.

3.4.1 Context

The Bathing water Directive (2006/7/CE) consider, the surveillance of algae bloom in the "bathing water profiles"; the mentioned water framework directive and marine strategy framework directive consider the surveillance of harmful algae blooms too. In particular, the marine strategy takes into account the algae blooms in descriptors 2 and 5 (Non-indigenous species, Eutrophication)

3.4.2 Monitoring activities

In Italy a microalgae surveillance is active since 2007 (Toxic Algae Program Directive of the Minister of the Environment no. GAB / 2006/6741 / B01). The monitoring activities are carried out by the Regional Agencies (ARPA) along the coasts of the 15 coastal regions, generally from June to September / October. Each Agency reports the concentration in terms of cells per litre on the bottom and in water column in a number of monitoring points. The observing activities are carried out during the summer season every two weeks or monthly. There is a threshold value for the protection of human health. According to the results of these monitoring activities, authorities could restrict permissions for the use of some locations.

3.4.3 Datasets

Ostreopsis ovata concentration (https://annuario.isprambiente.it/sys_ind/search)
Environmental data yearbook (ISPRA - https://www.isprambiente.gov.it/ /pdf/environmental-data-yearbook)
Report on the Quality of the Urban Environment - Bathing waters: Ostreopsis ovata indicator www.areeurbane.isprambiente.it/it/pubblicazioni
ARPA bulletins/report:
https://www.artaabruzzo.it/acque-mc.php?id_page=2
http://www.arpacal.it/index.php/24-tematiche-ambientali/balneazione/706-ostreopsis-ovata
https://www.arpacampania.it/ostreopsis-ovata
https://www.arpae.it/it/temi-ambientali/mare/dati-e-indicatori/dati-ostreopsis-ovata
http://www.arpa.fvg.it/cms/tema/acqua/balneazione/ostreopsis/
https://www.arpalazio.it/web/guest/ambiente/acqua/acque-di-balneazione
https://www.arpal.liguria.it/homepage/acqua/acque-marino-costiere/ostreopsis-ovata.html
https://www.arpa.marche.it/balneazione-nuovo/ostreopsis-cf-ovata
https:// www.arpa.puglia.it/pagina3092_report-alga-tossica-ostreopsis-ovata.html
https://www.sardegnaambiente.it/index.php?xsl=612&s=411800&v=2&c=4581&idsito=21
https://www.arpa.sicilia.it/temi-ambientali/mare/monitoraggio-ostreopsis-ovata/
https://www.arpat.toscana.it/temi-ambientali/acqua/balneazione/ostreopsis-ovata/dati-monitoraggio/ostreopsis-ovata
https://www.arpa.veneto.it/temi-ambientali/acqua/monitoraggio-algale-ostreopsis-ovata
https://www.isprambiente.gov.it/it/pubblicazioni/rapporti
Eionet, Marine Strategy informative systems (https://www.eionet.europa.eu; https://water.europa.eu/marine; https://www.isprambiente.gov.it)
Physical marine parameters form National and Regional Environmental Agencies and International Programs (https://www.isprambiente.gov.it, https://marine.copernicus.eu)

3.4.4 Opportunities

The application of this use case could provide several opportunities, in particular for stakeholders. Citizens can have clear information about bathing waters and risks for human health, while the business world related to tourism can have data to identify areas for human leisure. Public authorities can better monitor critical status and can promptly adopt appropriate measures to guarantee the citizen safety, while researchers can define accurate models for the prediction of the localization of the blooms, or for clinical studies.

3.5 Extreme events

Waters, in all the possible physical status (solid, liquid, gas), is strictly related with human health. So, any extreme events involving snow and ice, river and sea, water vapor, (e.g. storms, avalanches, floods, shortage of water, sea storms, high tides, strong currents, heat waves) could generate accidents, drowning, breakdown in communication and connections, hydrogeological instability, pollutant diffusion among large areas, loss of crops, loss of purification capacity of surface water. Risks for human beings are increasing with the climate change that exacerbate the intensity of the events, but also with the increasing use of the coastal zone and flooding areas, the soil consumption, the variation of hydrographic networks.

It is also well known that an extreme event could generate effects (and even on human health) far from the location of the event. A river flood flushes a great amount of sediments and pollutant into the sea; an extreme rain precipitation could overload the treatment plants and drain water waste directly to the sea, sea storms could push pollutant miles and miles far from the origin.

3.5.1 Monitoring activities

Hydrological parameters are collected continuously in every region of the world. All the European Member States have meteorological offices, hydrological services, marine observing system, so the physical parameters are generally well organised with long time series. During the last years, several products have been developed using in situ measurements, satellite observation, forecasting models (e.g. Copernicus)

3.5.2 Datasets

All meteo-marine parameters monitored by regional authorities and national institutions

4 Technical Architecture

In the context of the WHOW project, a technical architecture will be designed and implemented. All the software libraries will be made available as open-source software, published on GitHub[1]. This will allow us to interact with possible developers, hence keeping the entire development process as much open as possible.

4.1 Objective

Our main objective is to contribute to the Digital Service Infrastructure (DSI). Thus, laying the foundations for the construction of the first European Open Knowledge Graph (OKG) linking: (i) environmental data on water consumption and pollution, and (ii) health parameters, such as the diffusion of specific diseases.

[1] https://github.com/

With the term knowledge graph, we mean a collection of interlinked descriptions of entities – real-world objects and events, or abstract concepts – where:

- descriptions have formal semantics that allow both people and computers to process them in an efficient and unambiguous manner;
- entity descriptions contribute to one another, forming a network, where each entity represents part of the description of the entities, related to it, and provides context for their interpretation.

In essence, in the Open Knowledge Graph, the data has a formal semantics provided by ontologies organised in a network. Additionally, the OKG is represented as Linked Data, under the form of triples `subject-predicate-object` interconnected with each other. Accordingly, the OKG forms a large-scale distributed knowledge base. The latter can be queried for any purpose by anyone, where automatic reasoning can be enabled on it.

4.2 Requirements

There are different requirements to be taken into account in the design and implementation of the WHOW's reference technical architecture. The most important one is to guarantee **sustainability** of the overall approach over time: the knowledge graph should be functioning even at the end of the project and incrementally fed by additional data that may become available in the future.
To this end, we argue that the following additional requirements have to be met:

- A **modular approach is to be favoured** in the creation, management and access of the OKG;
- A **fully decentralized paradigm should be adopted** so that to maintain rightsholdership on the data. This means no need to centralise the data. In fact, the data (i) remain under the full control of the data providers; (ii) are queried on demand by other data sources in order to respond to stakeholders' needs.
- The **compliance with Semantic Web standards** as well as European directives (e.g., INSPIRE[2]), core vocabularies (e.g., Core Location Vocabulary[3]) and existing ontologies related to the domains considered in WHOW must be ensured.
- The **FAIR principles[4]**; that is, data must be Findable, Accessible, Interoperable, and Reusable in order to enable knowledge discovery, integration, sharing and reuse.

4.3 Existing data architectures

In order to start designing the architecture, a preliminary assessment of the state-of-the-art architectures has been conducted in order to devise the required elements. In such an analysis, the Linked Data Platform[5], currently a Web standard, has been considered along with the infrastructures that are currently deployed, or considered for future deployment, by WHOW's data providers.
In particular, ISPRA already provides a Linked Open Data platform that offers data transformation and linked open data publication functionalities. ARIA does not currently deploy a specific linked open data infrastructure; however, in the context of the Interreg project GIOCONDA[6], ARIA is evaluating the possibility to use the linked open data platform that has been designed and that allows them to keep the data up-to-date according to the changes occurring in the original datasets.
Additionally, we take into account existing and largely adopted off-the-shelf solutions, such as Open Link Virtuoso[7] or GraphDB[8], that provide core functionalities and service for data management, representation, serving, and querying. With regards to the adoption of the aforementioned solutions,

[2] https://inspire.ec.europa.eu/
[3] https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/core-location-vocabulary
[4] https://www.go-fair.org/fair-principles/
[5] https://www.w3.org/TR/ldp/
[6] https://progetti.interreg-italiasvizzera.eu/it/b/78/gestioneintegrataeolisticadelciclodivitadegliopendata
[7] https://virtuoso.openlinksw.com/
[8] https://graphdb.ontotext.com/

we remind that both Open Link Virtuoso or GraphDB are often used as triplestores in Linked Open Data projects.

4.4 WHOW Toolkit and pluggable knowledge graphs

Based on the previous analysis, we propose to create a so-called WHOW toolkit illustrated in Figure 1.

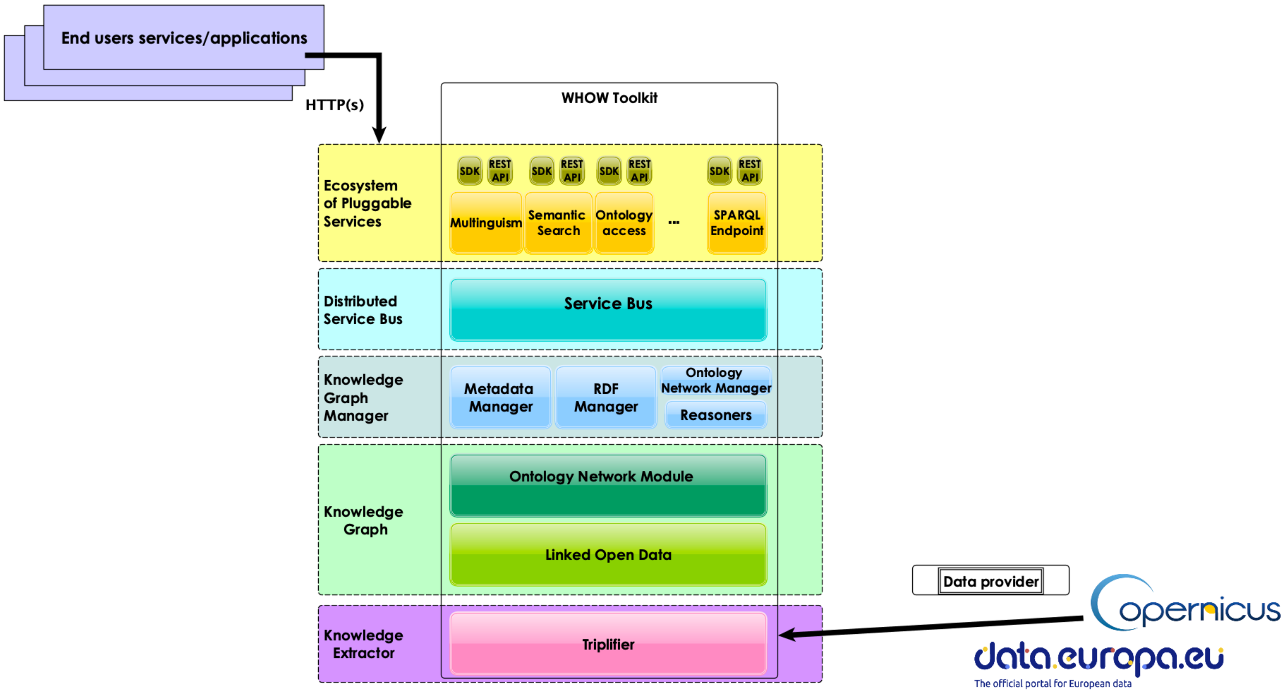


Figure 1: WHOW toolkit

The toolkit can be used by data providers in order to create, publish and make queryable the open knowledge graph on water consumption and pollution and health parameters. The toolkit is a layered reference architecture that starting from data providers' datasets and any other external resources of interest (e.g., data coming from Copernicus), it extracts knowledge through a component named *triplifier*. This is a software element aimed at transforming data available in any structured format (e.g. CSV, JSON, XML, relational database, etc.) into triples, based on `subject-predicate-object`. We remind that triples constitute the abstract model of data representation offered by the RDF[9] (Resource Description Framework) Web standard, to be used when creating linked (open) data. The data will be linked together and with other datasets available in the Web of Data, thus enriching the knowledge. The semantics, that can be understood by computer and human beings, is given by a network of ontologies that will be defined for such a purpose. Overall, the ontology network and the linked open data will form the so-called WHOW knowledge graph that can be successively managed through a set of services pluggable into the toolkit according to the various needs. A service bus layer is considered for the management of the federation of pluggable services to be offered to the application layer that, in turn is then enabled to access the data via APIs.

More in detail, the data produced with the triplification process will be stored in triple stores. Triple stores provide all CRUD (create, read, update and delete) functions at the level of single entities, single triples, sets of entities, sets of triples, single graphs (i.e. named graph) and sets of graphs. All

[9] https://www.w3.org/TR/rdf-concepts/

these elements will be queryable via SPARQL endpoints. WHOW will leverage existing endpoints of data providers and will strengthen their capabilities to support a wide range of queries, if required. An RDF Manager service will be implemented as a higher-level abstraction on the physical triple store to make transparent the way in which Linked Open Data expressed in RDF are managed by applications compared to the way in which they are physically stored. In this way, if triple stores must be changed or new ones are made available WHOW knowledge graph ecosystem, there would not be any repercussions on the levels above that would continue to operate
A Metadata Manager will provide all the necessary functionalities to guarantee that data is documented according to the DCAT-AP metadata profile and its national extensions, and to GeoDCAT-AP when geospatial data is treated.
Searching in, accessing to and reasoning on the knowledge graph will be enabled through the development of a set of services, pluggable in the framework through the service bus. For the implementation of the Reasoners component, WHOW will leverage state-of-the-art open-source solutions such as HermiT[10] or Pellet[11]. Finally, WHOW will evaluate multilingual resources and the eTranslation CEF building block in order to break any possible language barrier and increase the reuse by anyone, including other Member States.

4.5 Challenges

There are many challenges to face in the construction of WHOW's technical architecture. The first challenge consists in the harmonisation and integration of data that is heterogeneous in formats (e.g., CSV, linked open data, geospatial formats, etc.) and in its semantics. This will be done by using semantic web open standards: defining possibly shared ontologies and controlled vocabularies that are compliant with existing European and Italian standard data models will allow users of WHOW's digital artefacts to 'speak the same language' in the usage and exchange of water and health data, meeting semantic and technical interoperability requirements.
A second technical challenge is about combining and harmonising APIs, implemented according to different approaches, to be used to access the data that must be transformed in Linked Open Data. This challenge is very important because it constitutes the main entry point for any future new data provider that wants to join the knowledge graph by leveraging the whole toolkit made available. It is in fact WHOW's capability to accommodate a variety of APIs for original data access that will allow WHOW's knowledge graph to expand itself, including all possible stakeholders, in Italy and in the rest of Europe.
A third technical challenge is represented by the ability to link and use datasets that are not under the control of WHOW's data providers and can be very big in size. It is the case of the Copernicus data that may be used at different level of granularities. With this regard, WHOW will explore very recent research topics such as virtual graphs; that is, the ability to construct at run-time a graph according to the ontologies being used so that to allow end-users to query the data using the virtual graph as well. The data is not replicated in the triples stores of the WHOW's data providers but it is virtually created and queried when needed, in a fully decentalised environment. Finally, the fourth technical challenge is the ability to provide a flexible architecture that can accommodate different ways to offer knowledge discovery and access in compliance with the FAIR principles.

During the kickoff meeting of the co-creation programme, possible co-creators also highlighted additional challenges that are worth being considered during the lifespan of the project. One of these is the ability to link together data that is possibly collected according to different methods and procedures. Provenance of the data, quality of the data and of the links created among the data are important metadata to be defined and provided when publishing the entire knowledge graph.

---

[10] http://www.hermit-reasoner.com/
[11] https://github.com/stardog-union/pellet

4.6 Opportunities

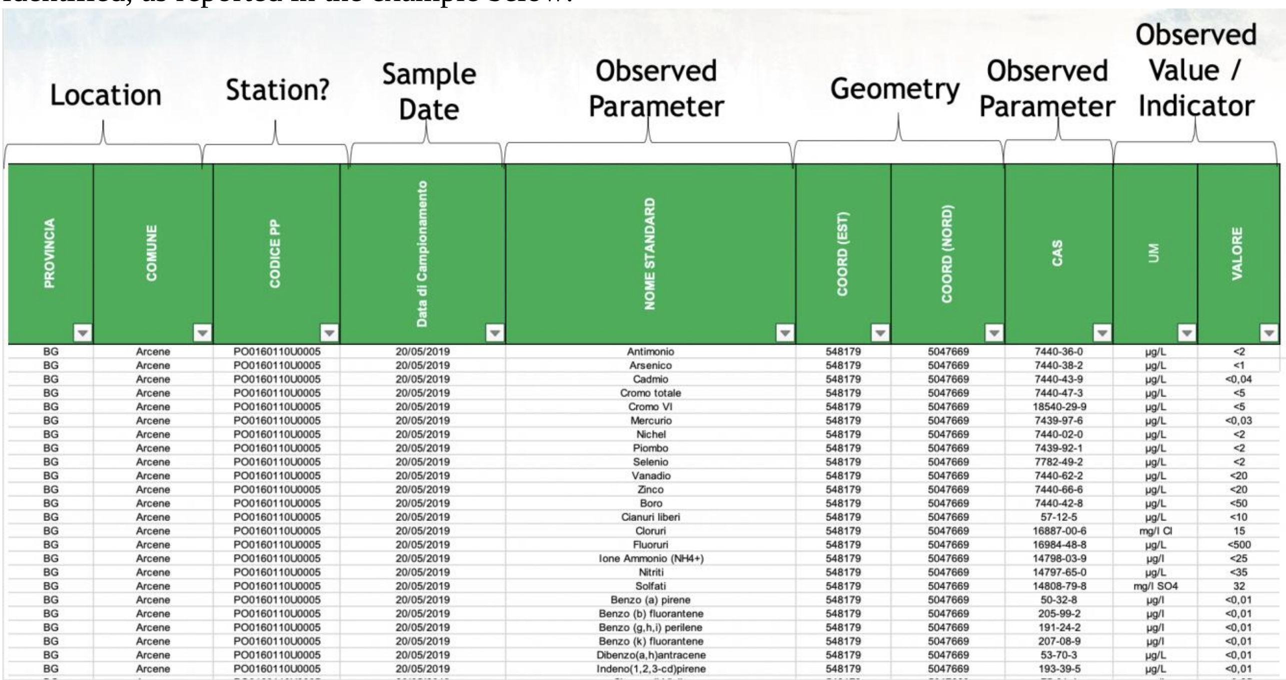
The opportunities that the design and implementation of WHOW technical architecture can offer are many. To begin with, any data provider can join WHOW knowledge graph with almost no code required: the toolkit will contain the necessary software libraries that can be used to i) interface data, supporting the previously mentioned use cases, and ii) transform the data according to the linked open data paradigm. The data will be interconnected with the portion of knowledge graph already produced and published, thus enriching the data itself and breaking possible silos. Through automatic reasoning new paths of knowledge can be discovered, thanks to the definition of a common semantic layer that will also facilitate the data exchange, preserving the meaning of the data in the interactions.

Finally, any end-users will be able to access the data when needed, in order to build new applications/services and any other interesting (research) studies.

5 Datasets

In order to face the semantic harmonisation and interoperability challenges earlier described, we started an analysis of a first set of datasets that have been considered relevant by WHOW's data providers in supporting the use cases previously introduced. In particular, we considered the datasets on the monitoring of water quality and other parameters for inland, marine and ground waters. We preliminary analysed five datasets: two datasets from ARIA, and specifically from ARPA Lombardia, on the quality of inland and ground waters, one dataset already published as linked open data from ISPRA on the national tidal network, and two datasets on inland waters from Umbria and Friuli Venezia Giulia Italian regions.

In the analysis we tried to highlight the main concepts behind each column of the tabular data that were found. During the kick-off meeting of the co-creation programme, these concepts have been identified, as reported in the example below.



PROVINCIA	COMUNE	CODICE PP	Data di Campionamento	NOME STANDARD	COORD (EST)	COORD (NORD)	CAS	UM	VALORE
BG	Arcene	PO0160110U0005	20/05/2019	Antimonio	548179	5047669	7440-36-0	µg/L	<2
BG	Arcene	PO0160110U0005	20/05/2019	Arsenico	548179	5047669	7440-38-2	µg/L	<1
BG	Arcene	PO0160110U0005	20/05/2019	Cadmio	548179	5047669	7440-43-9	µg/L	<0,04
BG	Arcene	PO0160110U0005	20/05/2019	Cromo totale	548179	5047669	7440-47-3	µg/L	<5
BG	Arcene	PO0160110U0005	20/05/2019	Cromo VI	548179	5047669	18540-29-9	µg/L	<5
BG	Arcene	PO0160110U0005	20/05/2019	Mercurio	548179	5047669	7439-97-6	µg/L	<0,03
BG	Arcene	PO0160110U0005	20/05/2019	Nichel	548179	5047669	7440-02-0	µg/L	<2
BG	Arcene	PO0160110U0005	20/05/2019	Piombo	548179	5047669	7439-92-1	µg/L	<2
BG	Arcene	PO0160110U0005	20/05/2019	Selenio	548179	5047669	7782-49-2	µg/L	<2
BG	Arcene	PO0160110U0005	20/05/2019	Vanadio	548179	5047669	7440-62-2	µg/L	<20
BG	Arcene	PO0160110U0005	20/05/2019	Zinco	548179	5047669	7440-66-6	µg/L	<20
BG	Arcene	PO0160110U0005	20/05/2019	Boro	548179	5047669	7440-42-8	µg/L	<50
BG	Arcene	PO0160110U0005	20/05/2019	Cianuri liberi	548179	5047669	57-12-5	µg/L	<10
BG	Arcene	PO0160110U0005	20/05/2019	Cloruri	548179	5047669	16887-00-6	mg/l Cl	15
BG	Arcene	PO0160110U0005	20/05/2019	Fluoruri	548179	5047669	16984-48-8	µg/L	<500
BG	Arcene	PO0160110U0005	20/05/2019	Ione Ammonio (NH4+)	548179	5047669	14798-03-9	µg/l	<25
BG	Arcene	PO0160110U0005	20/05/2019	Nitriti	548179	5047669	14797-65-0	µg/L	<35
BG	Arcene	PO0160110U0005	20/05/2019	Solfati	548179	5047669	14808-79-8	mg/l SO4	32
BG	Arcene	PO0160110U0005	20/05/2019	Benzo (a) pirene	548179	5047669	50-32-8	µg/l	<0,01
BG	Arcene	PO0160110U0005	20/05/2019	Benzo (b) fluorantene	548179	5047669	205-99-2	µg/l	<0,01
BG	Arcene	PO0160110U0005	20/05/2019	Benzo (g,h,i) perilene	548179	5047669	191-24-2	µg/l	<0,01
BG	Arcene	PO0160110U0005	20/05/2019	Benzo (k) fluorantene	548179	5047669	207-08-9	µg/l	<0,01
BG	Arcene	PO0160110U0005	20/05/2019	Dibenzo(a,h)antracene	548179	5047669	53-70-3	µg/L	<0,01
BG	Arcene	PO0160110U0005	20/05/2019	Indeno(1,2,3-cd)pirene	548179	5047669	193-39-5	µg/L	<0,01

It is worth pointing out that although some common elements and common guidelines on the use of specific codes, the datasets used in the analysis present different heterogeneities: the names of the columns differ one another even in the case the represented data is apparently the same (e.g., “sostanza”, “parametro”, “nome standard”), the format of the datasets can be different according to the various data providers (e.g., CSV, XLS, RDF/XML, geospatial data format), the structure of the datasets can vary from long to wide dataset format; that is, in some cases multiple columns are used to hold the values of the attributes (as in the case of the different chemical elements being monitored in the water), in other cases each row contains the value of a particular attribute for a given data point.

5.1 Data Modelling – common patterns

Despite the heterogeneities mentioned above, common concepts were found; this allowed us to preliminarily identify so-called modelling patterns (or ontology design patterns). These patterns are already existing modelling solutions for specific modelling issues that may be encountered. The modelling solutions are typically designed by ontology experts and their usage is proved to be very helpful for producing data of high quality and for meeting semantic interoperability requirements.
In particular, during the kick-off meeting of the co-creation programme, two modelling patterns have been presented: **Location-Geometry** and **Observation-Measure**, this latter as defined in INSPIRE. These two patterns are very useful to model respectively, specific spatial objects with a geometry (as it can be a watercourse) and observations of some property/parameter (as it can be the observation of the values of specific chemical elements in the water), for some feature of interest (as for instance a ground water) in a time period and that produce specific results.
A third pattern, not presented during the meeting for the sake of conciseness, has been already identified: **Indicator-Metric-Parameter**. This pattern, linked to the observation-measure pattern is very useful when any type of indicators, based on metrics, are to be modelled.

5.2 Ontologies

The modelling patterns presented above are already represented using semantic Web standards in the Italian network of ontologies named OntoPiA[12]. This entails that those ontologies can be already used for the semantic representation of the datasets that have been analysed above.
In particular, the modelling pattern Location-Geometry is the main pattern included in the Italian ontology Core Location Vocabulary – Italian Application Profile[13] (CLV-AP_IT). The ontology is fully aligned with the EU Core Location Vocabulary and to the GeoSPARQL[14] vocabulary defined by the OGC standardization body.
The modelling pattern Observation-Measure is the main pattern included in the Italian ontology IoT - Italian Application Profile[15] (IoT-AP_IT). This ontology is meant to represent all the possible observations and measures that are captured by sensors and monitoring facilities/platforms (e.g., water quality, air quality, traffic, etc.). It is fully aligned with the Semantic Sensor Network web standard.
Finally, the pattern Indicator-Metric-Parameter is included in the Italian ontology Indicator - Italian Application Profile (Indicator-AP_IT). This ontology is linked to the IoT ontology only for the observation part, and it is used to model any type of indicator, based on specific metrics, computed on data coming from data sources that have been built through observations.

5.3 General observations

The presence of already existing ontologies allows us not to face the semantic harmonisation challenge by scratch: the re-use of national open standards already aligned with European standards is an added value for the WHOW knowledge graph creation. However, from a first assessment, additional modelling elements could be introduced. For instance, during the meeting it has been

[12] https://github.com/italia/daf-ontologie-vocabolari-controllati
[13] https://w3id.org/italia/onto/CLV
[14] https://www.ogc.org/standards/geosparql
[15] https://w3id.org/italia/onto/IoT

discussed the need to model different types of watercourses with some physical elements that are reported in a set of analysed datasets.
To this end, the WHOW consortium is currently exploring some guidelines for hydrography modelling of INSPIRE and existing ontologies available in the literature. Controlled vocabularies such as GEMET[16] and those included in the Italian registry system will be investigated in order to face the representation of the different types of watercourses.

-

[16] https://www.eionet.europa.eu/gemet/en/about/