Co-financed by the European Union
Connecting Europe Facility

Water Health Open knoWledge (WHOW)

Milestone #9: SDGs and KPIs are defined

Project number	**Agreement number: INEA/CEF/ICT/A2019/2063229** **Action No: 2019-EU-IA-0089**
Project acronym	**WHOW**
Project title	**Water Health Open knoWledge**
Project duration	**36 months (01/09/2020 – 31/08/2023)**
Programme	**Connecting Europe Facility (CEF) Telecom**

Activity title	**Use Cases Development - Task 5.2 Identification of relevant set of indicators for SDGs**
Deliverable number	**5.1 [Milestone #9]**
Version (date)	**1.0 (2022-04-30)**
Due date	**2022-04-30**
Responsible organisation	**National Council of Research (ISTC-CNR)**

Editors	**Celeris, ISTC-CNR, Aria SpA, ISPRA**
Abstract	**This deliverable introduces the so-called WHOW data quality framework. It encompasses: i) a set of indicators distinguished between business and technical indicators; ii) some priorities assigned to the indicators; iii) and an evaluation method that varies from business to technical indicators.** **Among the business indicators, SDGs indicators have been selected after an analysis of the UN Agenda 2030 in addition to thematic indicators strictly dependent of the WHOW use cases and requests emerged during the co-creation programme.** **The technical indicators are mainly based on the FAIRness principles, data quality characteristics as those defined in the ISO/IEC standard 25012, metadata quality and semantic quality principles.**
Keywords	**Water, Health, Sustainable Development Goals, Key Performance Indicator, Metadata Quality**

Editor(s)

Anna Sofia Lippolis (ISTC-CNR), Giorgia Lodi (ISTC-CNR), Andrea G. Nuzzolese (ISTC-CNR), Gianluca Carletti (Aria Spa), Carmen Ciciriello (Celeris), Anna Pasian (Celeris), Patrizia Borrello (ISPRA), Roberta De Angelis (ISPRA), Marco Picone (ISPRA), Emanuela Spada (ISPRA)

Reviewers

Francesco Poggi (ISTC-CNR)

Edmund Gray (Celeris/ TSL)

Contributors

Paola Rita Esposito (Celeris)

Acknowledgement

This work was partially supported by the European Commission (EC) through the Connecting Europe Facility (CEF) programme under the WHOW project WHOW (grant agreement no. INEA/CEF/ICT/A2019/2063229).

Change Log

Version	Date	Organisation	Description
0.1	2022-02-14	ISTC-CNR	Preliminary ToC
0.2	2022-03-10	Celeris	Preliminary content for SGDs section
0.3	2022-03-17	ARIA Spa, ISPRA, ISTC-CNR	Preliminary content for Thematic and Technical indicators
0.4	2022-03-24	ISTC-CNR	Executive Summary, Introduction
0.5	2022-03-31	All	Consolidation of the overall document
0.6	2022-04-04	All	Document ready for internal reviews (quality assessment process)
1.0	2022-04-30	All	Document ready to be delivered after revisions based on comments from external reviewers

Table of contents

List of Tables

List of Figures

Executive Summary

The benefits of making open data available is not immediately apparent to data providers and consumers because they are often intangible.

For instance:

- they may help in creating insights that improve specific research and decision-making activities;
- they may be the means through which new applications, services and websites are developed;
- they may support in improving products and existing processes in order to augment productivity and efficiency;
- they may assist in understanding how to act in order to provide more sustainable actions as a society for the defence of our planet and wellbeing.

However, simply publishing open data is not enough for such advantages to become evident. It is widely acknowledged that for open data to be truly effective, it must: i) be able to answer relevant questions of public interest with respect to the reference application domain; ii) be of high quality over time, supported by sustainable processes.

In a blog post of 2017[1], the Open Knowledge Foundation clearly stated that "*The open data community needs to shift focus from mass data publication towards an understanding of good data quality.*"

In the light of these observations, we believe that an important aspect in the process of publishing open data is to give data providers any tool that can guide them in releasing open data of high quality, helping them to improve what they can obtain even at an early stage of the open data process, and to facilitate the show up of the above mentioned intangible advantages.

To do so, it is also important to understand what 'open data of high quality' means. This deliverable proposes a *data quality assessment framework* for evaluating the quality of open data according to different aspects. In particular, the framework consists of:

- a set of indicators, and associated metrics when applicable, spanning from business, strictly related to the application domain, to more technical indicators that are defined according to quality dimensions available at the state of the art;
- an evaluation mechanism that, based on priorities that can be assigned to indicators, allows data providers to understand the l*evel of achieved quality* for their data.

It is worth noticing that there are a certain number of evaluation frameworks that have been adopted over years in the open data context. Examples include the Open Data Maturity of the data.europa.eu portal [1], the open data certificates from Open Data Institute (ODI) [2], the Open Data Barometer [3], the FAIR Data Maturity Model [4]. However, according to our research, these frameworks do not focus on specific relevant questions for target domains, and some of them do not take into account quality dimensions for all data aspects, from metadata of datasets to data content and data semantics. This is also due to their general nature. In contrast, the framework proposed in this deliverable relies on indicators defined in most of these experiences while targeting the objectives of specific open data processes. This leads to the definition of a

[1] https://blog.okfn.org/2017/05/31/open-data-quality-the-next-shift-in-open-data/.

framework that is able to unify in one tool the minimum set of indicators that are significant for the open data process of a data provider, capable of guiding it in publishing data that may show a real impact on the reuse.

Among the business indicators in the framework proposed in this paper, we include some of those that have been defined in the context of the UN 2030 Agenda for Sustainability, identified as relevant to the WHOW project as well as thematic indicators important for answering the most crucial questions underlying the WHOW use cases.

The technical indicators instead are mainly based on the FAIR principles (Findability, Accessibility, Interoperability, Reusability) that, we claim, guide the openness of all the data in every application domain, not only in open science. This approach is also promoted by recent initiatives of open data communities such as the "datiBeneComune" Italian initiative, where a published online report clearly asks government institutions to open data having in mind these principles[2]. In particular, the deliverable includes technical indicators selected from the: i) FAIR Data Maturity Model; ii) ISO/IEC 25012 data quality model standard; iii) Metadata Quality Assurance model of data.europa.eu; scientific literature on the quality of the semantic of the data.

Thus, the deliverable discusses the definition of the framework, the methodologies we applied for the selection of all its indicators and extends the preliminary analysis on the data quality framework introduced in the "Use Cases Definition" deliverable.

[2] https://vorrei.datibenecomune.it/dati-che-vorrei/come-li-vorrei/

1. Introduction

This is deliverable “5.1 - SDGs and KPIs are identified”. It is the result of the activities conducted in the context of task 5.2 of Activity 5 related to “Use Case Development”. It extends deliverable 2.1 “Use Cases Definition”.

1.1 Project Overview

The WHOW project aims to foster the creation of the first open and distributed 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 leverages the Linked Open Data paradigm. Water related datasets from Italy and other European countries and Copernicus (the European Union's Earth observation programme) will be used to support the construction of WHOW’s knowledge graph, intended as a federation of knowledge graphs deployed at each data provider willing to join the WHOW community. The knowledge graph will be documented on data.europa.eu, the official portal for European data, thanks to the adoption of shared metadata models such as DCAT-AP and its extensions that are relevant for the type of data treated in WHOW (e.g., GeoDCAT-AP). Selected health related datasets mainly from Italy will be linked to specific water datasets.

WHOW targets identified use cases in the creation of the knowledge graph. In order to evaluate such use cases relevant set of indicators for Sustainable Development Goals (SDGs) are identified along with Key Performance Indicators for metadata and data quality. A co-creation programme, where interested stakeholders and users are engaged from the initial phases of the project, is set-up so as to consider real needs of data re-users.

The initiative supports 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 Objectives

This deliverable discusses a set of selected indicators as well as an evaluation methodology for them that together form the so-called **WHOW data quality framework**.

Note that, there is no willingness here to ‘reinvent the wheel’: the WHOW data quality framework can be thought of as an aggregator of the most important indicators and metrics for open data quality that recently emerged and are dispersed in a variety of different evaluation models and tools.

The term data quality, we use to qualify the framework, encompasses not only the technical characteristics of the data, but also its ability to answer important questions that are relevant for possible re-users of the application domain(s). Therefore, the indicators composing the framework are both business indicators, that is, indicators related to the specific water and health application domains investigated in the WHOW projects, and technical indicators; that is, indicators related to different technical aspects of open datasets such as descriptive metadata, semantics of dataset content, accessibility, interoperability. Furthermore, the framework includes indicators for monitoring the Sustainable Development Goals that have been analysed and defined as relevant to support the implementation of the WHOW use cases, already described in Deliverable 2.1.

The importance of considering SDGs in the context of open data emerges from several contexts. Notable is the European Parliament resolution of 14 March 2019 on the Annual strategic report on the implementation and delivery of the Sustainable Development Goals (SDGs) (2018/2279(INI))[3] where a precise call on the Commission is mentioned in order to add data related to the SDGs to the high-value datasets as defined in the directive on open data and public sector information, encouraging the Member States to publish all reports on the SDGs under a free license.

The World Bank Group, in a blog post[4] from as far back as 2015, explicitly highlights that "*Open Data can help achieve the SDGs by providing critical information on natural resources, government operations, public services, and population demographics*". Since then, initiatives have been taken to open up data to support monitoring of each SDGs, and many of these datasets, mostly statistical, can be found in the SDG data hub provided by the United Nations[5].

However, more fine-grained thematic aspects must be considered in order to qualify successfully an open data initiative that takes into account one or more specific application domains. To this end, the WHOW data quality framework embodies so-called thematic indicators that have been identified by the data providers of the WHOW project according to the three use cases and their legislation bases.

From a technical perspective, the indicators included in the WHOW data quality framework are all guided by the FAIR principles, already mentioned in our previous deliverables on the WHOW architecture and use cases. We claim that any open data publication process, also carried out in a context which is different from the open science context, can be improved if those principles are at the forefront. However, these are general and agnostic of technological implementations; they are guidelines that ensure not only that data/metadata are traceable and persistent over time, but also that they are published with high quality international data representation standards in mind. Among these standards, in the WHOW data quality framework we refer to the Metadata Quality Assurance of data.europa.eu, crucial for the metadata quality evaluation, and for enabling future harvesting activities that document the produced WHOW knowledge graph in the European data portal, and the ISO/IEC 25012 and its implementing standard ISO/IEC 25024 on the data quality model. In particular, from these latter standards, we select according to a policy we describe in this deliverable, a set of relevant data quality characteristics and their associated indicators and metrics that we want to assess when releasing open datasets. Finally, since an important role in the WHOW project is played by the harmonised semantic conceptualisation of the datasets forming the knowledge graph, indicators of the semantic soundness are listed, as presented in the reference scientific literature.

To conclude, we state that the proposed WHOW data quality framework may allow data providers and consumers to assess the readiness, implementation and impact of the WHOW knowledge graph.

[3] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52019IP0220&from=EN

[4] https://blogs.worldbank.org/digital-development/sustainable-development-goals-and-open-data

[5] https://unstats-undesa.opendata.arcgis.com/search?collection=Dataset

1.3 Relationships with other activities

The present deliverable D5.1 is another important milestone (#9) of the WHOW project. It covers all the activities of task 5.2 of Activity 5 that have dependencies linked with other tasks of other activities of the project.

The following Figure 1 shows such relationships. In particular, the content of this deliverable; that is, the results of the activities of task 5.2, is both related to the definition of the use cases and the preliminary data quality requirements there identified, and the design of the technical architecture. In addition, relationships with the on-going work of the semantic data model of the knowledge graph of WHOW are identified when semantic indicators are described and included in the framework introduced in this document.

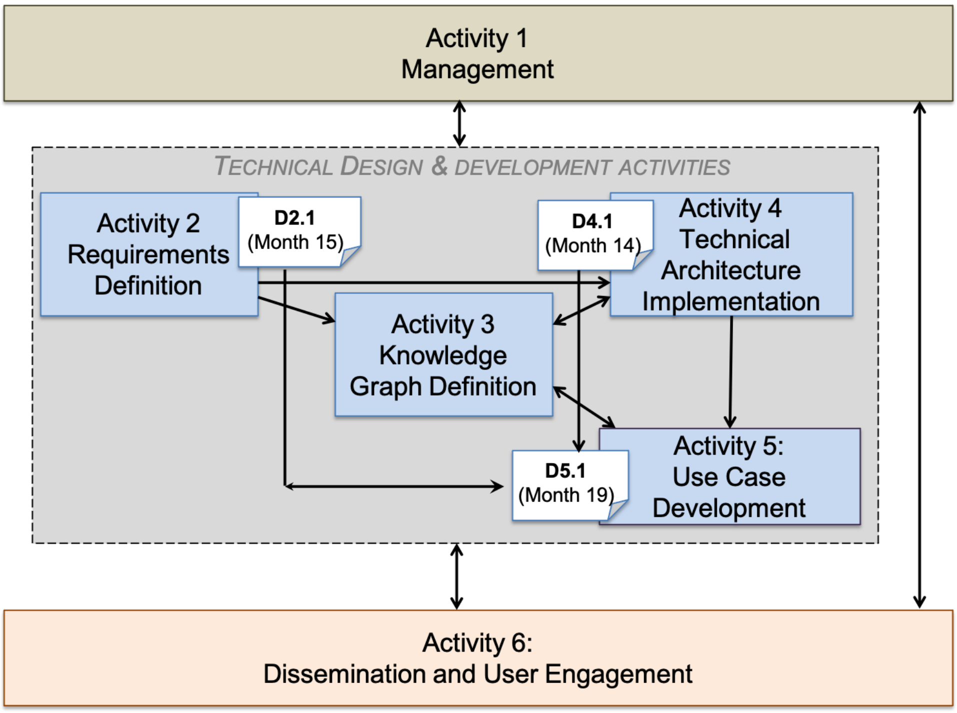

Figure 1: Relationships with other deliverables and activities

1.4 Structure of the document

The rest of this document is structured as follows. Section 2 describes the WHOW data quality framework with the set of Key Performance Indicators (KPIs) that have been identified. Section 3 introduces the business KPIs distinguishing between the selected Sustainable Development Goals (SDGs) of Agenda ONU 2020 from the thematic KPIs we defined for supporting the three use cases of the project. Section 4 describes the technical KPIs for the assessment of the quality of the produced WHOW knowledge graph. Technical KPIs are discussed from different perspectives according to the FAIR principles. Finally, Section 5 concludes the deliverable.

2. The WHOW data quality framework

This section introduces the data quality framework of the WHOW project. The first fundamental component of the framework is represented by the set of indicators and related metrics (this latter atomic and simple composition measures that contribute to the calculation of indicators) that have been identified.

In general, the design of the framework reflects the meaning we ascribe to the term **data quality**, which is not merely a technical requirement to be achieved, but also includes the ability of open datasets to generate interest in possible data consumers and thus actual impacts in one or more specific domains.

The framework classifies the indicators into two main macro-categorises, as shown in Figure 2:

- **Business KPIs**, consisting of SDGs indicators, relevant for the project use cases, and additional thematic KPIs, also identified through the interactions with the co-creators;
- **Technical KPIs**, where different technical aspects about data quality (e.g., quality of the metadata, quality of the data, quality of the data semantics) are considered in the identification of the indicators and their metrics. The indicators we included in this part are the result of the discussion about the technical architecture described in Deliverable 4.1, also done in the context of the co-creation programme.

We claim that such a framework can be effectively reused in other data scenarios: business KPIs can be customised on a case-by-case basis according to the different needs of the target application domains and their users, while technical KPIs can be universally reused in all (open) data contexts since they are, to the best of our knowledge, the minimum set of indicators of well-established standards and scientific works needed to qualify a piece of data as high quality.

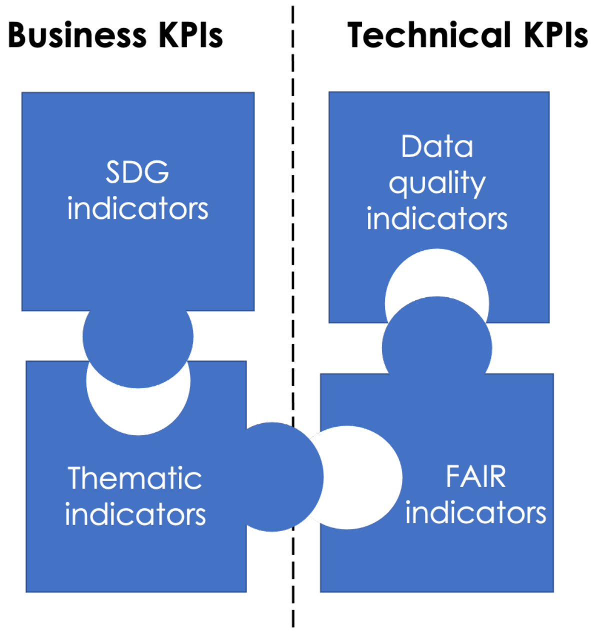

Figure 2: The WHOW data quality framework - Business and Technical KPIs

Along with the set of indicators, the framework provides:

- **priorities** associated with thematic and technical indicators and their metrics. Inspired by the FAIR Data Maturity Model, we introduced, for all technical KPIs, a priority level that assumes the following values: Essential, Important, Useful. This level can be effectively used in the evaluation method of the framework to assess the level of achieved data quality. As also explained by the FAIR model, according to the context, some indicators can take different levels of priority. This makes the framework flexible enough for being adopted in various data scenarios;
- **evaluation method.** As reported in [4] there might exist different evaluation methods to be used. In our framework, we adopt a hybrid approach: for some indicators we use the **pass-or-fail method,** in other cases we calculate a number/percentage that we then verify against predefined thresholds. Please note that we distinguish between the evaluation method for business KPIs and that for technical KPIs. This is due to the different nature of the evaluation in the two cases: when evaluating technical KPIs, one measures the degree of technical quality (e.g., completeness, accuracy, compliance to standard, presence of certain metadata, etc.) of the data included in the linked open datasets. In some cases (e.g., the presence of certain metadata), this is independent of the specific application domains to which the data refer. On the contrary, when assessing business KPIs, one shows a state of a phenomenon in the relevant application domain that, among other things, is also influenced by the quality of the datasets used to describe facts about that phenomenon.

3. Business Key Performance Indicators

In general, Key Performance Indicators (KPIs) are critical indicators used to measure progress toward an intended goal. Thus, KPIs provide a well-defined blueprint for strategic and operational improvement of even certain processes, creating an analytical basis for decision support as they help focus attention on what matters most. As Peter Drucker said, "*What gets measured gets done*".

If we think about the specific open data context of the WHOW project, KPIs can be very helpful to measure the progress toward the production of a semantic knowledge graph that is capable of providing connected data of high quality, ready to be successfully re-used by any final user for any purpose.

In this section we focus our discussion on the business KPIs of the WHOW data quality framework, dividing them in indicators coming from the UN Agenda 2030 for sustainability that are relevant for the goals of our project, and thematic business KPIs related to the use cases we consider.

3.1 SDGs indicators in the UN Agenda 2030

Since the Seventies to the recent agreements such as the 2030 Agenda for Sustainable Development, the 2015-2030 Sendai Framework for Disaster Risk Reduction, the 2015 Addis Ababa Action Agenda on Financing for Development, *sustainability* has been identified as a global goal that anyone, within public or private sectors will have to strive for.

At the heart of the 2030 Agenda there are five critical dimensions: *people, prosperity, planet, partnership and peace*. Addressing the critical issues around these 5 pillars, the UN 2030 Agenda envisages "a world of universal respect for human rights and human dignity, the rule of law, justice, equality and non-discrimination".

Paragraph 55 of the 2030 Agenda for Sustainable Development adopted by the United Nations states: "The Sustainable Development Goals and targets are integrated and indivisible, global in nature and universally applicable, taking into account different national realities, capacities and levels of development and respecting national policies and priorities. Targets are defined as aspirational and global, with each government setting its own national targets guided by the global level of ambition but taking into account national circumstances. Each government will also decide how these aspirational and global targets should be incorporated in national planning processes, policies and strategies. It is important to recognize the link between sustainable development and other relevant ongoing processes in the economic, social and environmental fields".

Signed on 25 September 2015 by the governments of the 193 Member Countries of the United Nations, and approved by the UN General Assembly, the Agenda sets out 17 Sustainable Development Goals, SDGs, which are part of a broader programme of action consisting of 169 associated targets. All Countries are called upon to define their own sustainable development strategies to achieve the objectives, and report the results within a process coordinated by the UN.

Every country on the planet is expected to contribute to addressing these major challenges towards a sustainable path by developing its own National Strategy for Sustainable Development. Each signatory State is evaluated annually at the UN through the work of the High-level Political Forum (HLPF), responsible for assessing progress, results and challenges for all countries, and by national and international public opinion. Every four years there is also a debate on the implementation of the 2030 Agenda at the UN General Assembly before Heads of States and/or Governments. The first such review was carried out in September 2019.

During the opening speech of the Plenary Session of the European Parliament chaired by Ursula von der Leyen (July 2019), the European Commission presented a rich action programme to be implemented over the next five years. The programme clearly reveals the Union's willingness to achieve sustainable development objectives, also in relation to the Paris Agreement on climate change, and sets the ground for a global EU strategy for the years 2019-2024. In this context, the European Union has also committed to transposing and defining the principles of the 2030 Agenda for sustainable development. The SDGs are transposed at EU level to guide the Member States in the final definition of their strategic objectives.

3.1.1 UN Sustainable Development Goals

The Sustainable Development Goals (SDGs) were adopted by the United Nations in 2015 as a universal call to action to end poverty, protect the planet, and ensure that by 2030 all people enjoy peace and prosperity. The 17 SDGs (Sustainable Development Goals), with their 169 targets, form the core of the 2030 Agenda. Progress towards these Targets is agreed to be tracked by 232 unique Indicators. For example, the SDG 6 - is Clean Water and Sanitation; it defines 8 targets (e.g., 6.3 target - *by 2030, improve water quality by reducing pollution, eliminating dumping and minimising release of hazardous chemicals and materials, halving the proportion of untreated wastewater and substantially increasing recycling and safe reuse globally)* and 11 indicators to measure the 8 targets (e.g., 6.3.2 indicator - the *proportion of bodies of water with good ambient water quality*).

The UN SDGs are a blueprint to a better future for all. They provide a focus for how businesses, governments and civil society can tackle these challenges in order to promote a more sustainable future for all. They balance the economic, social and ecological dimensions of sustainable development, and place the fight against poverty and sustainable development on the same agenda for the first time. In order to measure and record progress on the SDGs will first require the answer to the complex question of how the Sustainable Development Goals should be correctly implemented. The Agenda 2030 defines two key features as fundamental to their implementation: the fact that they have been created as universal objectives, applying to all countries; and that they have been created as an indivisible package of goals and objectives all of which need to be pursued in an integrated way.

In the light of the annual SDG Progress Report, the UNECE Executive Secretary Olga Algayerova stated: *"On the basis of past trends, the region would achieve only 23 targets by 2030. Progress in 57 targets should accelerate. For 80 targets, almost half of the total, there is no sufficient national data to track change over time. Much progress is necessary therefore not only to meet the targets but also to improve data availability, including in countries with well-developed statistical systems.".* Figure 3 reports which targets are on track with respect to the deadline of 2030 [14].

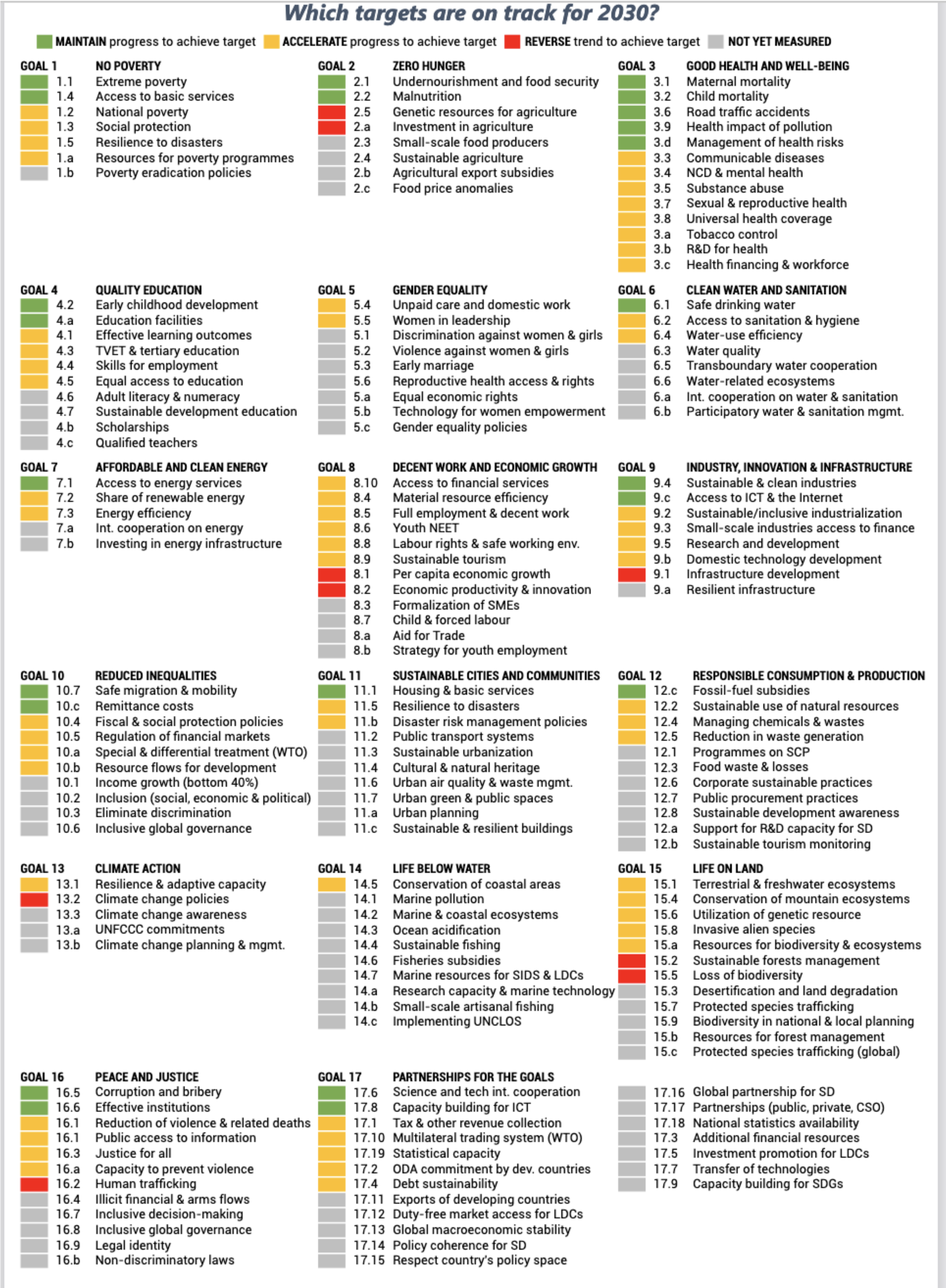

Figure 3: Which SDGs' targets are on track for 2030

At the EU level, Eurostat plays a central role in providing statistical reporting on the SDGs. Eurostat[6] shows the progress towards the SDGs as of 2019. It is worth noting that for SDG 6, SDG 14 (these both are related to water and are thus linked to the identified WHOW datasets) and SDG 16 trends cannot be calculated due to the lack of time series for more than 25% of the related indicators, as shown in Figure 4.

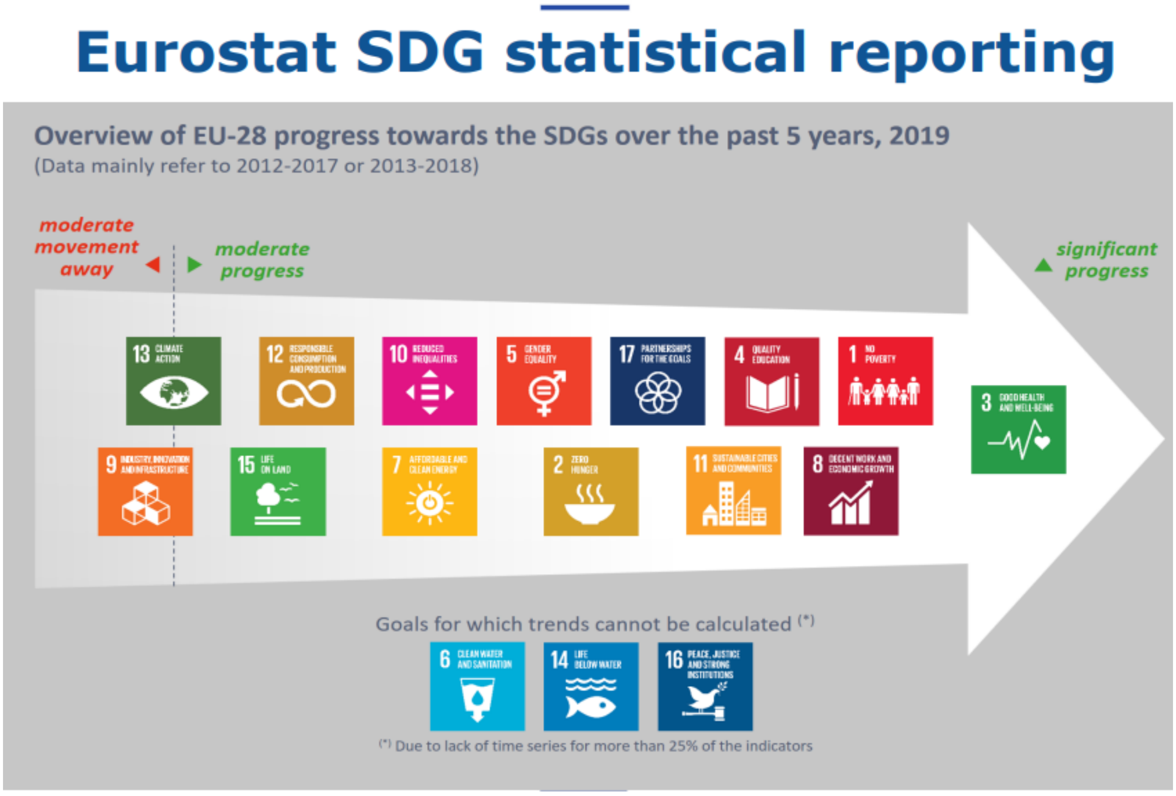

Figure 4: Eurostat SDG statistical reporting

For each indicator, if targets, baselines and historical data is available, it calculates “target paths”, otherwise it makes an assessment about growth rates.

3.1.2 Water and Health SDGs

Access to safe drinking water is essential for good health, welfare and productivity and is widely recognized as a human right. Drinking water may be contaminated with pathogens or with chemical and physical contaminants, leading to harmful effects on health. While improving water quality is critical to prevent the transmission of many diseases (such as diarrhoea which exacerbates malnutrition and remains a leading global cause of child deaths), improving the accessibility and availability of drinking water is equally important for health and welfare.

Sea water, atmospheric water, ice and snow cover, surface water, groundwater, soil water, water in the human water use chain and wastewater are all interconnected within the hydrological cycle. To promote

[6] https://ec.europa.eu/regional_policy/rest/cms/upload/07102019_042252_EWRC%20presentation%20SDG.pdf

effective and sustainable choices it became essential to gather and manage updated, interoperable and re-usable datasets the public and private sector may refer to while dealing with the complexity of water cycle components and interactions. Achieving transformative and timely impacts requires breaking silos across different disciplines and working on an integrated approach.

Figure 5: The source-to-sea concept

The source-to-sea[7] concept, shown in Figure 5, identifies six key flows that connect the source-to-sea system from land systems to open oceans: water, sediment, pollutants, materials and ecosystem services. It describes six steps to guide analysis and planning; and presents a framework for elaborating a theory of change, all with an aim of designing initiatives that support healthy ecosystems and sustainable green and blue economies.

The relevant SDGs to be taken into account to reduce pollutants to improve water quality are as follows:

SDG 6: WATER AND SANITATION

SDG 3: GOOD HEALTH AND WELL-BEING

SDG 12: RESPONSIBLE CONSUMPTION AND PRODUCTION

SDG 14: LIFE BELOW WATER

SDG 15: LIFE ON LAND

Systematic data acquisition is done in the form of:

[7] https://siwi.org/why-water/source-to-sea/

- assessment studies (for time-independent data and a for ‘snapshot’ at a single moment of time)
- monitoring activities (for time-dependent data), making use of both classical field work methods and innovative remote sensing and automatic monitoring techniques.

As stated in [18]

> “*Once “sufficient information is available”, water accounting, numerical modelling and other types of analysis may contribute to gaining in-depth knowledge on the water systems’ performance and on the water services. Such knowledge is indispensable for identifying water resources management needs, issues and opportunities.* T*ime series of groundwater and surface water quantity and quality are usually collected through monitoring networks. Their design and operation should preferably be coherent, also with those of meteorological and other related monitoring networks“.*

The collected data form the building block for developing, evaluating and adjusting water resources management strategies, but also for real-time management decisions (e.g., on water allocation or on water level control actions in the field). It needs to be emphasised that the collected monitoring data should be made easily accessible or disseminated as soon as possible among all stakeholders and agencies that have a role to play in the area concerned. This includes monitoring data sharing between neighbouring countries, in case of transboundary water bodies.

Numerical simulation models play an important role in the development of the required knowledge. On the one hand, they allow a deeper understanding of the local hydrological processes and to calculate water fluxes and transport of solutes and suspended matter.

3.1.3 Interlinkages between Water and Health SDGs

The SDGs are founded on the principle that they are “integrated and indivisible” – progress in one area is dependent upon progress in many others. Translating this idea into practical action is going to be one of the key challenges for the agenda. The clear UN recommendation is that the SDGs should be considered through an integrated approach, “*noting that sustainable development interventions cannot be put in an economic, social and/or environmental box*”.

This concept becomes crystal clear when referring to water (SDG 6): to ensure the well-being of ecosystems in the whole *source to sea stream*, a robust understanding of the systemic linkages between the different segments of this stream is vital. This understanding maps the path and highlights the relationships between upstream pressures and downstream effects.

Recognizing the “integrated and indivisible” quality of the SDGs territories is therefore the key to define the interdependencies with other relevant goals and make use of them to embed highly effective, coordinated actions at regional and national level.

That is even more the case with regard to health. There are several arenas to interconnect and monitor in order to achieve and preserve the health-sector goals. The health of people is not solely a health-sector

responsibility; it is also highly impacted by issues covered by the SDG 6, such as safe management of drinking water and faecal waste, and the attainment of hygiene. Beyond SDG 6, water is explicitly mentioned in relation to SDG 3 (health impacts from water-borne diseases and contaminated water), SDG 11 (water-related disasters), SDG 12 (water pollution), and SDG 15 (conservation of freshwater ecosystems). This complex of linkages to SDG 6 speaks to the interlinkage of all SDGs that provides an integrated set of priorities with interdependent goals and targets [19].
The goal for health and well-being is directly influenced by efforts to reduce poverty (SDG 1), knowledge of health behaviours (SDG 4) and their gendered differences (SDG 5), clean water, sanitation and hygiene (SDG 6), safe and healthy living environments (SDG 11), and agricultural products (SDG 15).

Health is affected by a multitude of factors, inherent to each individual but also dependent on environmental and economic circumstances. The massive number of connections between SDG 3 and other SDGs here suggests identifying those more directly related to the purposes of the WHOW project and the use cases adopted.

Identifying the SDG 3 and SDG 6 interlinkages, and the restoring or preventing applicable actions may have a key role "*in terms of decreased water-borne infections (e.g., acute diarrheal infections, viral hepatitis) and improved nutrition; improving water quality and sanitation also leads to long-term developmental gains. The interaction between these goals is strongest in parts of the developing world where water-borne infectious disease is still prevalent, but water quality and environmental pollution issues are also widespread in many high-income contexts*"[8]

In the 2021 edition of the Sustainable Development Report the Data Table section (Table 4.3, p.67) declares that "*despite our best effort to identify data on SDGs, several sectors and data gaps persist at the international level. Government and international communities must increase investments in SDG data and monitoring systems and build strong data partnerships to support informed SDG decisions and strategies*." [17].

3.1.4 Further interlinkages

All the SDGs must be analysed all together since the actions are systemic and involve all of them. However, in this section, we report a number of SDGs that we believe have clearer connections with the SDGs more linked to our objectives (see sections below).

SDG 2 Zero Hunger - End hunger, achieve food security and improved nutrition and promote sustainable agriculture

Target 2.4: Sustainable food production and resilient agricultural practices Sustainable food production. The UN explains this SDG as[9]:

> *"It is time to rethink how we grow, share and consume our food. If done right, agriculture, forestry and fisheries can provide nutritious food for all and generate decent incomes, while supporting people-centred rural development and protecting the environment.*

[8] https://www.unwater.org/publications/summary-progress-update-2021-sdg-6-water-and-sanitation-for-all/.
[9] https://sdg-tracker.org/zero-hunger

Right now, our soils, freshwater, oceans, forests and biodiversity are being rapidly degraded. Climate change is putting even more pressure on the resources we depend on, increasing risks associated with disasters such as droughts and floods. Many rural women and men can no longer make ends meet on their land, forcing them to migrate to cities in search of opportunities.
A profound change of the global food and agriculture system is needed if we are to nourish today's 815 million hungry and the additional 2 billion people expected by 2050.
The food and agriculture sector offers key solutions for development, and is central for hunger and poverty eradication."

Clearly, clean water available to all can be critical to agricultural production just as protecting crops from extreme events is important to ensure that they are abundant for the benefit of all.

SDG 9 Industry, innovation, infrastructure - Build resilient infrastructure, promote sustainable industrialization and foster innovation

Investments in infrastructure – transport, irrigation, energy and information and communication technology – are crucial to achieving sustainable development and empowering communities in many countries. It has long been recognized that growth in productivity and incomes, and improvements in health and education outcomes require investment in infrastructure.
The UN has defined 8 *Targets* and 12 *Indicators* for SDG 9. Targets specify the goals and Indicators represent the metrics by which the world aims to track whether these Targets are achieved. Innovations in the information architecture could provide *an environmentally grounded understanding* of water resources. Data infrastructure can be used by cities to integrate surface and groundwater resources more wisely.

SDG 12 Responsible Consumption and Production - **Ensure sustainable consumption and production patterns**

Sustainable consumption and production is about promoting resource and energy efficiency, sustainable infrastructure, and providing access to basic services, green and decent jobs and a better quality of life for all. Its implementation helps to achieve overall development plans, reduce future economic, environmental and social costs, strengthen economic competitiveness and reduce poverty.
The UN states that:

> "*By 2020, achieve the environmentally sound management of chemicals and all wastes throughout their life cycle, in accordance with agreed international frameworks, and significantly reduce their release to air, water and soil in order to minimise their adverse impacts on human health and the environment.*"

SDG 15 Life on land

The SDG 15 aims to protect our planet's ecosystems through a sustainable use of them, while preserving biodiversity. Climate change, human beings' interventions in fact have caused over years severe damage to our ecosystems through deforestation, loss of natural habitats and land degradation. In this context, taking into account the WHOW application domains, data regarding water-related ecosystems, including rivers, aquifers and lakes and the recycle and reuse of wastewater for example can be of utmost importance to evaluate the degree of preservation of some of our planet's ecosystems.

3.1.5 Selected SDGs

This section describes the methodology that the WHOW project has followed to identify the relevant SDGs and related indicators. This is necessary in order to create a shortlist of indicators on which the project will focus and assess the potential contribution to close data gaps and/or increase the quality of existing data gathered in the EU.

3.1.5.1 Methodology

The methodology is composed by three layers:

1. Domain
    a. WHOW activities focus on two main domains: Water and Health. Initially SDGs related to these domains will be selected.

2. WHOW Use Cases
    a. WHOW has identified three use cases, as follows: Contaminants in marine environments; Water for human consumption; and Extreme events.
    b. A second selection of relevant SDGs for each use case will be performed to ensure *completeness*, while at the same time assess whether specific targets and indicators of the overall SDGs identified can be considered *in scope or not* for further analysis.
    c. Targets should be considered appropriate for observation and monitoring when suitable for the R.A.C.E.R. Criteria: Relevant, Accepted, Credible, Easy to monitor, Robust.
    d. Targets are identified as most *relevant* when there is a direct correlation between water conditions and impact on human health.

3. WHOW datasets
    a. WHOW has identified a list of datasets for the three use cases. These include data available from WHOW data providers as well as data from external data providers in the EU. WHOW is also aiming at influencing data providers to make available currently closed data, as open data.
    b. Each SDG indicator is calculated using one or more datasets provided by so-called 'Custodians', based on the UN methodology. In Europe, a key role is played by Eurostat in gathering statistics. WHOW will map its datasets to the shortlisted SDG indicators and perform a gap analysis to identify to what extent WHOW datasets can close current data gaps and/or contribute to the harmonisation and quality.

3.1.6 WHOW Use Cases and SDGs indicators

In the light of the above discussion, based on the three use cases we identified in Deliverable 2.1, we analyse in the following which SDGs indicators can be relevant for the proposed WHOW data quality framework.

USE CASE 1 - CONTAMINANTS IN MARINE ENVIRONMENTS

All animals, including humans, can be exposed to and absorb natural and chemical pollutants in their natural surroundings by eating, breathing, and drinking. These pollutants are substances which are not easily broken down in the environment. Some persistent organic pollutants occur naturally, but others are man-made, such as the chemicals used in pharmaceutical, pesticide, industrial, and solvent manufacturing. Additionally, some anthropogenic activities are responsible for releasing heavy metals and toxins into the seawater. One of the reasons why these pollutants cannot be degraded in the environment is that plants and animals have not yet been exposed to them for a long period of time. This means they have not had enough time to evolve the appropriate biological methods for detoxification and elimination.

When substances, such as chemicals or heavy metals, concentrate within the internal organs and tissues of living beings, biomagnification can occur. Biomagnification refers to the condition where the chemical concentration in an organism exceeds the concentration of its food when the major exposure route occurs from the organism's diet.

Because humans are at the top of the food chain, biomagnification is of serious concern. The toxins responsible for health problems include: mercury, lead, chromium, cobalt, cadmium and natural toxins. Humans who are affected by biomagnification tend to have a higher risk of developing certain cancers, liver failure, birth defects, brain damage, and heart disease. Natural toxins in food can cause both acute and chronic health effects with a range of clinical symptoms. Acute symptoms range from mild gastrointestinal upset, neurological symptoms, respiratory paralysis to fatality. This is more likely among the susceptible groups of the population such as children and the elderly. Within hours if not shorter, acute symptoms are seen following ingestion of various marine toxins in shellfish and other seafood. Poisoning from ingested marine toxins is an underrecognized hazard for travellers, particularly in the tropics and subtropics.

USE CASE 2 - WATER FOR HUMAN CONSUMPTION

The concept of this use case derives from the observation of a series of evidence depicting the correlation between the use of contaminated water and negative health effects that have been documented recently in the Lombardy Region (Italy). This use case is therefore born on a regional scale starting from the data available by Lombardy Region, but has the aim of being extended to other Italian regions and possibly also to the European context.

Water can be used for different purposes, such as drinking water for human consumption (domestic, drinking fountains, "Water houses"), while a higher volume of water is used, for example, in the agricultural, manufacturing, etc. Behind the use of water there is a strict regulation that requires the competent structures (eg: ARPA and ATS) periodic checks on the quality of the water so that they respect the precise microbiological, chemical and physical parameters. A decrease in the quality of drinking water, even temporary, can have a direct consequence on the health of its consumers. In fact, some pathologies can be directly related to water as a means of transmission of pathogens.

USE CASE 3 - EXTREME EVENTS

The Extreme Events use case gathers hydro-meteorological and tidal data that provide the knowledge base for analyses related to the impacts of extreme events, more specifically floods and droughts, on the environment and human health.

Floods are natural phenomena which cannot be prevented. Floods have the potential to cause fatalities, displacement of people and damage to the environment, to severely compromise economic development and to undermine the economic activities of communities. However, the increase in the likelihood of their occurrence and the aggravation of their impact are determined by certain human activities (such as increasing human settlements and economic activities in floodplains and the reduction of the natural water

retention by land use) and by climate change (rapid melting of glaciers, precipitation variability and intensification).
Drought is a natural phenomenon. It is a temporary, negative and severe deviation along a significant time period and over a large region from average precipitation values (a rainfall deficit), which might lead to meteorological, agricultural, hydrological and socioeconomic drought, depending on its severity and duration[10].
Linked to the notion of scarcity is the notion of water stress. UNWater[11] defines water stress as: '*Level of water stress: freshwater withdrawal as a proportion of available freshwater resources*'.
For water management purposes, it is important to consider that water scarcity conditions are to be linked to a sharp reduction in water body levels together with water abstraction from those same water bodies. At the same time, it is important to consider the need to meet water demand for various uses. In other words, the above conditions generally occur as a result of the combination of climate factors (drought) and anthropogenic factors (surface and groundwater pressures).
It is therefore necessary to have real-time monitoring data of the main hydrological magnitudes, such as precipitation, temperatures, flow rates of watercourses and springs, groundwater levels, and those relating to the uses of water resources, i.e., abstractions from watercourses, aquifers and springs, including also the water requirements necessary to protect ecosystem services.

Table 1 shows the SDGs that, based on WHOW use cases and datasets already open data and to be open in the lifespan of the project, have been selected as highly relevant. Please note that WHOW datasets labelled as "expected to be open" have been identified as highly relevant during the co-creation programme meetings with data providers. Those datasets are currently closed at each data provider. WHOW project is working to get data owner's authorisation to publish them as open datasets by the end of the project.
For each SDG indicator, the UN metadata are available at: https://unstats.un.org/sdgs/metadata/

Table 1: Selected SDGs and related indicators mapped to WHOW use cases and datasets

SDG	Target	Indicator	WHOW Use Case	WHOW open datasets	WHOW datasets expected to be open	Note
6 - Clean water and sanitation	6.1 - Safe drinking water	6.1.1 Proportion of population using safely managed drinking water services	2	1. Water for human consumption sampling (Lombardy Region) 2. Population census per each municipality (Statistical open data of Lombardy Region)	1. Water for human consumption sampling (Water utilities of Lombardy region)	With these datasets, WHOW can calculate this indicator

[10] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52012SC0382&from=fr
[11] https://www.unwater.org/

SDG	Target	Indicator	WHOW Use Case	WHOW open datasets	WHOW datasets expected to be open	Note
	6.2 Sanitation and hygiene	6.2.1 Proportion of population using a safely managed sanitation services, including a hand-washing facility with soap and water	2	-	1. Wastewater network (Water Utilities of Lombardy Region)	This dataset does not directly allow us to compute the indicator but it contributes to it
	6.3 Water quality and wastewater	6.3.1 Proportion of wastewater safely treated	2	-	1. Wastewater network (Water Utilities of Lombardy Region) 2. Data flow rates (Water Utilities and ARPA of Lombardy Region)	These datasets do not directly allow us to compute the indicator but they contribute to it
		6.3.2 Proportion of bodies of water with good ambient water quality	1,2,3	1. Analytical data of river water bodies (ARPA Lombardy) 2. Analytical data of lake water bodies (ARPA Lombardy) 3. Analytical data of groundwater (ARPA Lombardy) 4. PFAS data (perfluoroalkyl substances) surface waters - Year 2018 (ARPA Lombardy) 5. PFAS data (perfluoroalkyl substances) groundwater - Year 2018 (ARPA Lombardy) 6. Bathing water quality (all years)	-	With these datasets, WHOW can calculate this indicator
	6.4 Water use and	6.4.2 Level of water stress:	2		Water consumption	These datasets do

SDG	Target	Indicator	WHOW Use Case	WHOW open datasets	WHOW datasets expected to be open	Note
	scarcity	freshwater withdrawal as a proportion of available freshwater resources			dataset from water service providers of Lombardy Region	not directly allow us to compute the indicator but they contribute to it
3 - Heath and well being	3.9 Reduce illnesses and deaths from hazardous chemicals and pollution	3.9.2 Mortality rate attributed to unsafe water, unsafe sanitation and lack of hygiene (exposure to unsafe Water, Sanitation and Hygiene for All	2	1. Mortality, cause of death (ISTAT, published on OD portal Regione Lombardia) 2. Infectious diseases rates by sex and age (Open Data Lombardy Region)	-	These datasets do not directly allow us to compute the indicator but they contribute to it
14 - Life Below Water	14.1 Reduce marine pollution	14.1.1 Index of coastal eutrophication and floating plastic debris density	1	Through the use of external datasets coming from EIONET-	-	The eutrophication is measurable through the nutrients reported in EIONET database
11-Sustainable Cities and Communities (water related disasters)	11.5 Reduce the adverse effects of natural disasters	11.5.1 Number of deaths, missing persons and directly affected persons attributed to disasters per 100,000 population	3	1. Avalanche accidents 2. Mortality, cause of death (ISTAT, published on OD portal of Lombardy Region)	-	These datasets do not directly allow us to compute the indicator but they contribute to it
13 - Climate Action	13.1 Strengthen resilience and adaptive capacity to	13.1.1 Number of deaths, missing persons and directly affected persons	3	Mortality, cause of death (ISTAT, published on OD portal of Lombardy Region)	-	This dataset does not directly allow us to compute the indicator but it

SDG	Target	Indicator	WHOW Use Case	WHOW open datasets	WHOW datasets expected to be open	Note
	climate-related disasters	attributed to disasters per 100,000 population".				contributes to it

3.1.6.1 Contribution of the WHOW project to the relevant SDGs indicators

Two SDG indicators can be calculated using WHOW datasets that are either currently available and open or expected to be open before the project ends:

- **6.1.1** Proportion of population using safely managed drinking water services;
- **6.3.2** Proportion of bodies of water with good ambient water quality.

In contrast, the WHOW project can contribute to the calculation of six SDG indicators. In particular for the following indicators, the related WHOW datasets are already available:

- **3.9.2** Mortality rate attributed to unsafe water, unsafe sanitation and lack of hygiene (exposure to unsafe Water, Sanitation and Hygiene for all;
- **11.5.1** Number of deaths, missing persons and directly affected persons attributed to disasters per 100,000 population;
- **13.1.1** Number of deaths, missing persons and directly affected persons attributed to disasters per 100,000 population.

In addition, following indicators, that WHOW expects to make the related datasets available before the project ends:

- **6.2.1** Proportion of population using a safely managed sanitation services, including a hand-washing facility with soap and water;
- **6.3.1** Proportion of wastewater safely treated;
- **6.4.2** Level of water stress: freshwater withdrawal as a proportion of available freshwater resources.

Finally, the following indicator can be computed using datasets coming from external data sources such as EIONET, linked to the datasets we are going to open during the lifespan of the project:

- **14.1.1** index of coastal eutrophication and floating plastic debris density.

3.2 Thematic indicators

This section introduces the so-called *Thematic indicators* based on the defined use cases and feedback received from co-creators.

3.2.1 Thematic Indicators for WHOW Use Case 1

Thematic KPI 1.1: Trend measures of the concentration and spread of *Ostreopsis ovata* along the Italian marine-coastal areas

Ostreopsis ovata is a potentially toxic benthic microalgae, currently present in most of the Italian coastal regions with blooms that can give rise to human intoxication phenomena and toxic effects on marine benthic organisms. The continuous expansion along the Italian coasts of *Ostreopsis ovata*, its blooms and associated health, environmental and economic issues, led to the establishment of a surveillance monitoring program, to clarify the occurrence and trend of blooms at a national and regional level.
In Table 2 we introduce the indicators that have been identified for supporting the scenario we foresee for use case 1 of the project (see related Deliverable 2.1).

Table 2: Thematic KPI 1.1 - WHOW Use Case 1

Indicator ID	Indicator	Metrics	Source and Notes	Priority
IND-THEM-UC1-1	Number of sites with the presence of *Ostreopsis ovata*	National and regional percentage of sites with positive presence	ISPRA Environmental Data Yearbook ISPRA annual reports on *Ostreopsis ovata*	Essential
IND-THEM-UC1-2	Number of sites above *Ostreopsis ovata* health threshold	National and regional percentage of sites above health threshold	ISPRA Environmental Data Yearbook ISPRA annual reports on *Ostreopsis ovata*	Essential

3.2.2 Thematic Indicators for WHOW Use Case 2

Thematic KPI 2.1: Quality of water intended for human consumption open datasets

Use case 2 is inspired to Council Directive 93/98/EC and following updates till CD 2184/2020, whose art. 17 and annexe IV push Member States to ensure that adequate, up-to-date information on water intended for human consumption, including the indicator parameters, is available while complying with applicable data protection rules.

WHOW project aims to open water and health related datasets and harmonise semantics. The indicators, listed in Table 3, specify, at Italian and European level, where water intended for human consumption monitoring datasets are available in open data according to the parameters provided for by the directive (ref. Council Directive 93/98/EC annexes I, II and III).

Indicators are set on a yearly scale and let the impact of the WHOW project be measurable.

Table 3: Thematic indicators for water for human consumption

Indicator ID	Indicator	Metrics	Source and Notes	Priority
IND-THEM-UC2-1.1	Available open datasets	Number of available datasets by territory (municipalities, provinces, regions, countries) and by time (yearly)	Datasets published on national and European open data portal	Important
IND-THEM-UC2-1.2	Available open datasets with parameters	Number of available datasets with parameter type (chemical, physical and microbiological) by territory (countries) and by time (yearly)	Datasets published on national and European open data portal	Important
IND-THEM-UC2-1.3	Available open datasets without of threshold outcomes	Number of datasets without of threshold outcomes availability by territory (countries) and by time (yearly)	Datasets published on national and European open data portal	Useful

Thematic KPI 2.2: Ministry of Health monitoring on water intended for human consumption quality

The indicators reported in this section and listed in Table 4 are the indicators calculated by Lombardy Region and sent to the Italian Ministry of Health for water intended for human consumption quality monitoring purposes, on yearly basis. Indicators consider the trend of the medians of parameters values, listed in the reference legislation (98/83/EC), and related out of threshold outcomes.

The scope of all Indicators is confined at regional level.

Table 4: Thematic indicators for quality of water for human consumption

Indicator ID	Indicator	Metrics	Source and Notes	Priority
IND-THEM-UC2-2.1	Parameters out of threshold outcomes (%)	Percentage of out of threshold outcomes on total samples (yearly, at regional level)	Dataset Water for Human Consumption Sampling (ATS)	Essential

IND-THEM-UC2-2.2	Parameters trend on yearly basis (mean values)	Median of parameter mean values (for every parameter, yearly, at regional level)	Dataset Water for Human Consumption Sampling (ATS)	Useful
IND-THEM-UC2-2.3	Parameters trend on yearly basis (max values)	Median of parameter maximum values (for every parameter, yearly, at regional level)	Dataset Water for Human Consumption Sampling (ATS)	Useful

Thematic KPI 2.3: Lombardy Region Open Data Portal water for human consumption quality indicators

Thanks to the WHOW project, water for human consumption quality monitoring dataset will be soon published with detailed parameters values for Lombardy's territory in the regional Open Data portal. The datasets will be presented introducing the monitoring activity performed through the ATS (local healthcare authorities) according to 98/83/EC Directive and to Italian national decree 31/2001. The page will also present the indicators shown in this section (Table 5) in order to give readers a measure of sampling points distribution, sampling frequency and out of threshold outcomes impact.

The scope of all Indicators is confined at regional level.

Table 5: Thematic indicators for water for human consumption quality measures

Indicator ID	Indicator	Metrics	Source and Notes	Priority
IND-THEM-UC2-3.1	Sampling frequency	Number of samples compared by ATS competence territory (local healthcare authorities) and by Chemical, Physical and Microbiological parameters (yearly)	Dataset Water for Human Consumption Sampling (ATS)	Essential
IND-THEM-UC2-3.2	Active sampling points	Number of sampling points compared by ATS competence territory (yearly)	Dataset Water for Human Consumption Sampling (ATS)	Essential
IND-THEM-UC2-3.3	Parameters out of threshold outcomes (absolute values)	Number of out of threshold outcomes by parameter type (chemical, physical, and microbiological), by ATS competence territory (yearly)	Dataset Water for Human Consumption Sampling (ATS)	Essential

Thematic KPI 2.4: WHOW Co-creation programme indicators

Public and private bodies participating in the WHOW CCP (Co-Creation Programme) confirmed their interest in UC2 and suggested how to address it. Therefore, the indicators in this section respond to the needs gathered during co-creation meetings.

The main interest is related to making linkable datasets on water for human consumption quality and infectious diseases, on a time and territorial basis. CCP members also suggested linking analytical monitoring groundwater dataset, as that can impact water for human consumption quality.

Known cases of diseases caused by ingestion of polluted water will also be reported among the indicators.

The scope of all Indicators (Table 6) is confined at regional level. Indicator IND-THEM-UC2-3.3 is inherited from Thematic KPI 2.3.

Table 6: Thematic indicators based on co-creators needs

Indicator ID	Indicator	Metrics	Source and Notes	Priority
IND-THEM-UC2-3.3 (inherited from Thematic KPI 2.3)	Out of threshold outcomes (water for human consumption parameters)	Number of out of threshold outcomes by parameter type (chemical, physical, and microbiological), by ATS competence territory (yearly)	Dataset Water for Human Consumption Sampling (ATS)	Essential
IND-THEM-UC2-4.1	Out of threshold outcomes (underground water parameters)	Number of groundwater out of threshold by parameter type (chemical, physical, and microbiological), by province (yearly)	Analytical data of groundwater (ARPA Lombardy)	Useful
IND-THEM-UC2-4.2	Reported diseases potentially related to water quality	Number of reported diseases by ATS competence territory (yearly)	Infectious diseases Lombardy Region rates by sex and age (open data Lombardy Region)	Essential
IND-THEM-UC2-4.3	Known cases of correlation between infectious diseases and polluted water	Number of known cases of correlation between infectious diseases and polluted water	Known cases monitoring (Lombardy Region)	Important

3.2.3 Thematic Indicators for WHOW Use Case 3

Thematic KPI 3.1: Economic loss in relation to number of days of bathing prohibition due to the short-term water pollution

Visiting bathing sites located around lakes, rivers, estuaries, and coastal areas is a major summertime recreational activity and provides a range of physical and psychological health benefits.

It has been estimated that every year there are about 120 million tourists on Italian seashores and each tourist spends on average 40 euros a day, contributing 13% to the national GDP.

Massive rainfall events are causing flooding of rivers and streams with severe consequences on the environment. The consequent bacterial contamination poses bathing waters risks, besides damaging tourism and economy.

Therefore, the economic impact of closures to bathing is an important aspect to be considered (Table 7).

Table 7: Thematic indicator for Use Case 3 - Economic loss due to short-term pollution events

Indicator ID	Indicator	Metrics	Source and Notes	Priority
IND-THEM-UC3.1	Economic loss due to short-term pollution events	Average daily financial loss due to closures during the bathing season	Prohibitions: Ministry of Health Costs: Italian National Institute of Statistics (ISTAT), “Mondo balneare”	Important

Thematic KPI 3.2: Sea Surface Temperature trend

The Sea Surface Temperature (SST) is one of the main parameters for the evaluation of climate change effects on marine waters (indicator in Table 8). Chemical and biological parameters and therefore water quality and human health are affected by SST variation (especially the increase). As a direct consequence of the SST variation, the salinity and current flow regimes could change, the eutrophication and acidity level could increase, algal blooms could become more frequent, sea conditions could facilitate the settlement of alien species.

Table 8: Thematic indicator for Use Case 3 - sea surface temperature trend

Indicator ID	Indicator	Metrics	Source and Notes	Priority

IND-THEM-UC3.2	Sea Surface Temperature (SST) Trend	Trend and anomalies of SST data	Marine Monitoring networks (tide gauges)	Essential

Thematic KPI 3.3: Frequency of rough and high waves

Particular conditions of the physical state of the sea increase the risk of accidents (drowning, navigation security) or economic loss due to swells and surges. Rough and high waves identify the marine climate with a higher probability of risks for human health (Table 9).

Table 9: Thematic indicator for Use Case 3 - frequency of rough and high waves

Indicator ID	Indicator	Metrics	Source and Notes	Priority
IND-THEM-UC3.3	Frequency of rough and high waves	Ratio between number of wave heights > 2.5 m and total number of waves, on a per year basis	Marine Monitoring networks (buoys)	Important

3.3 Evaluation Method

The evaluation method for the business KPIs, including the selected SDGs and thematic KPIs related to the WHOW use cases differs from the evaluation method we present in section 4.4 for the technical KPIs. As introduced above, the scope of the two evaluations is different, although technical KPIs have a strong impact on the results of business KPIs: the more accurate and complete the data is in open datasets, for example, the more accurate a status on a certain phenomenon can be read from the datasets.

Specifically, for the SDGs, the United Nations offer a well-documented framework that, for each target and each indicator of the targets, provides all the necessary information to compute and assess the related indicators. Interested readers can refer to the E-Handbook for SDG indicators [15] for more information on the evaluation methods adopted.

As far as the thematic indicators are concerned, all the proposed metrics are based on percentages or values to be computed (in specific cases there is a trend to show). Those indicators are addressed to a wide range of consumers, spanning - for example - from the need to communicate information of public interest (such as the water quality), to the need to monitor the effectiveness of some policies or specific actions. Excluding some cases where an indicator tolerability threshold is defined by the legislation (i.e., water quality parameter), a pass/fail methodology for their evaluation is not appropriate. In general, for thematic indicators, the evaluation method is not properly defined since they are used to show a status of a phenomenon in the reference application domain. Hence, the assessment of thematic indicators requires a more peculiar methodologies, which are an objective for the implementation of our evaluation method. In fact, such methodologies aim at contextualising the indicator values in reference evaluation frames. Those frames rely on different methodologies, such as statistical significance, regression analysis, and probabilistic reasoning that take into account historical data, trends and values in order to get meaningful interpretation to scattered indicator values otherwise meaningless. For example, in the WHOW Use Case 3 the value of the IND-THEM-UC3.2 indicator cannot be fully judged as positive or negative. It simply reports a fact that can be read also looking at the data of the previous periods and other elements. Another example is related to the WHOW use case 1: in a specific focus on the *Ostreopsis ovata* of the documentation available in [16][12], it is reported the quantification of the indicator as a trend over time.

Finally, it is worth noting that regardless of the absolute values assumed, the trend analysis of some of these indicators (for example those of the Ministry of Health) also leads to measuring policy application results over time or to evaluating the effectiveness of corrective actions on specific critical issues.

[12] https://annuario.isprambiente.it/sys_ind/847.

4. Technical Key Performance Indicators

This section introduces the technical indicators component of the overall WHOW data quality framework. These KPIs are independent of the specific application domain and most of them, with the exception of the semantic quality indicators, are guided by the FAIR principles: Findability, Accessibility, Interoperability, Reusability. This means that all indicators are associated with the reference FAIR principle(s) among the above four.

The technical component of the WHOW data quality framework consists of the following macro-sections:

- **FAIR principles:** this part of the framework aims at selecting indicators as defined in the well-known FAIR Data Maturity Model;
- **Metadata quality:** this part of the framework concentrates on the metadata associated with open datasets. Metadata is data that describe other data and is a fundamental building block in every (open) data management process. Metadata enables, in fact, easy discoverability and access to the open datasets. In addition, in the context of the WHOW project, metadata of high quality, compliant with European and national standard metadata application profiles, allows data providers to document their datasets in relevant data catalogues; namely, data.europa.eu, the Italian dati.gov.it and the Italian geodati.gov.it;
- **Data quality**: this part of the framework aims at evaluating the quality of the content of the datasets; that is, the quality of the data that is published. In order to achieve this objective, some quality characteristics from the well-known ISO/IEC standard 25012 and their metrics defined in the standard ISO/IEC 25024 are considered.
- **Semantic quality**: since a significant contribution of WHOW is to provide data harmonised from a semantic perspective, we decided to introduce the semantic quality evaluation section in our overall data quality framework. This section includes a set of indicators aiming at evaluating specific aspects (e.g., cognitive ergonomics, usability) of the ontologies we are developing.

For all these sections, we selected a subset of indicators as defined by the related reference standards. The following subsections describe and motivate the methodology for this selection.

It is worth noticing that the metadata quality section represents an exception of this general policy. In fact, for metadata we decided to include all those foreseen by data.europa.eu with some additional ones we included based on national reference metadata application profile standard: since a crucial part is to be harvested by the European data catalogue, the need to respect all the requirements defined in their Metadata Quality Assurance tool is of utmost importance for the success of the project.

Finally, for all the following technical indicators, we indicate the related functional and/or non-functional requirements that the indicator supports and that we defined in the context of Deliverable 4.1 "Design of the technical services for knowledge graph management".

4.1 The FAIR principles indicators

The FAIR principles were defined in 2016 with the aim to specify "*a minimal set of related but independent and separable guiding principles and practices that enable both machines and humans to find, access,* *interoperate and re-use data and metadata*" [4]. In order to avoid the proliferation of different evaluation methods based on various interpretations of the four principles, the RDA working group in 2020 published a FAIR Data Maturity Model. The model consists of indicators that can be used to evaluate the adherence to the FAIR principles. These indicators are designed for re-use in evaluation approaches.

In particular, the FAIR Data Maturity Model consists of three main parts:

- a list of **indicators** that are the different aspects of the FAIR principles to be assessed;
- the importance of the indicators represented through **priorities**. They span from Essential to Useful;
- different **evaluation methods**, that allows the different evaluators to assign a value to each indicator.

Please note that, since the FAIR principles guide most of the proposed technical KPIs of the WHOW data quality framework, we decided to follow the same methodology as the one proposed by the FAIR model.

In this respect, with the use of the FAIR Data maturity Model it is possible to evaluate both the original quality of the datasets being used in WHOW, and the quality to be guaranteed as the outcome of the project activities to be carried out.

In Table 10, we list a subset of the indicators of the FAIR Data Maturity Model that we decided to include in the WHOW data quality framework. As reported in other sections of this document (see below), we decided to not select all of them but only those indicators that are marked as essential and important as priority. This choice is motivated by the idea of providing a data quality framework that is easy to use and implement by data producers and consumers and, at the same time, that is able to offer a robust framework for publishing high quality open datasets.

Table 10: The selected FAIR principles indicators

Indicator ID	FAIR principle	Architectural requirement	Indicator	Method of quantification	Priority
RDA-F1-01M	Findability	FR-06	Metadata is identified by a persistent identifier	Yes/no	Essential
RDA-F1-01D	Findability	FR-06	Data is identified by a persistent identifier	Yes/no	Essential
RDA-F1-02M	Findability	FR-06	Metadata is identified by a globally unique identifier	Yes/no	Essential
RDA-F1-02D	Findability	FR-06, FR-21	Data is identified by a globally unique identifier	Yes/no	Essential
RDA-F2-01M	Findability	FR-14, FR-15, FR-16	Rich metadata is provided to allow discovery	Yes/no	Essential
RDA-F3-01M	Findability	FR-06, FR-16	Metadata includes the identifier for the data	Yes/no	Essential

Indicator ID	FAIR principle	Architectural requirement	Indicator	Method of quantification	Priority
RDA-F4-01M	Findability	FR-14, FR-15	Metadata is offered in such a way that it can be harvested and indexed	Yes/no	Essential
RDA-A1-02M	Accessibility	FR-15	Metadata can be accessed manually (i.e. with human intervention)	Yes/no	Essential
RDA-A1-02D	Accessibility	FR-38, FR-41, FR-42	Data can be accessed manually (i.e. with human intervention)	Yes/no	Essential
RDA-A1-01M	Accessibility	FR-14	Metadata contains information to enable the user to get access to the data	Yes/no	Important
RDA-A1-03M	Accessibility	FR-36	Metadata identifier resolves to a metadata record	Yes/no	Essential
RDA-A1-03D	Accessibility	FR-36	Data identifier resolves to a digital object	Yes/no	Essential
RDA-A1-04M	Accessibility	NFR-02, NR-04, FR-3701	Metadata is accessed through standardised protocol	Yes/no	Essential
RDA-A1-04D	Accessibility	NFR-02, NFR-04, FR-3701	Data is accessible through standardised protocol	Yes/no	Essential
RDA-A1-05D	Accessibility	FR-37, FR-39, FR-40, FR-43	Data can be accessed automatically (i.e. by a computer program)	Yes/no	Important
RDA-A1.1-01D	Accessibility	FR-31, FR-3701	Metadata is accessible through a free access protocol	Yes/no	Essential
RDA-A2-01M	Accessibility	FR-12, FR-13	Metadata is guaranteed to remain available after data is no longer available	Yes/no	Essential
RDA-I1-01M	Interoperability	NFR-04	Metadata uses knowledge representation expressed in standardised format	Yes/no	Important
RDA-I1-01D	Interoperability	NFR-04	Data uses knowledge representation expressed in standardised format	Yes/no	Important
RDA-I2-01M	Interoperability	FR-1401, FR-1402, FR-1403	Metadata uses FAIR-compliant vocabularies	Yes/no	Important
RDA-R1-01M	Reusability	NFR-09	Plurality of accurate and relevant attributes are provided to allow reuse	Yes/no	Essential

Indicator ID	FAIR principle	Architectural requirement	Indicator	Method of quantification	Priority
RDA-R1.1-01M	Reusability	FR-1401	Metadata includes information about the license under which the data can be reused	Yes/no	Essential
RDA-R1.3-01M	Reusability	FR-14, FR-15	Metadata complies with a community standard	Yes/no	Essential
RDA-R1.3-01D	Reusability	NFR-04	Data complies with a community standard	Yes/no	Essential
RDA-R1.3-02D	Reusability	NFR-03, NFR-04	Data is expressed in compliance with a machine-understandable community standard	Yes/no	Important
RDA-R1.3-02M	Reusability	FR-14	Metadata is expressed in compliance with a machine-understandable community standard	Yes/no	Essential

4.2 Metadata Quality indicators

As far as the metadata quality is concerned, the WHOW data quality framework includes the indicators defined in the Metadata Quality Assurance of the data.europa.eu[13]. Since these indicators are crucial for future harvesting activities with the data.europa.eu portal, we decided to replicate them in Table 11. However, we add to each indicator a priority, as it is done in the case of the FAIR indicators. This ensures that a harmonised WHOW data quality framework can be provided. The priorities that have been assigned follow a precise policy: all EU Member States are invited to comply with a metadata application profile named DCAT-AP [5] based on the DCAT W3C Web Standard [6]. DCAT-AP defines precise obligations in terms of availability of specific descriptive metadata of datasets. All mandatory metadata are then "Essential" as priority. In addition, EU Member States can extend the DCAT-AP profile with additional metadata and constraints. All the data providers in WHOW are Italian; therefore, the available Italian extension of the DCAT_AP named DCAT-AP_IT [7], still based on DCAT-AP version 1.1, is taken into account. In the Italian application profile, some metadata is to be mandatorily included. For instance, the Italian application profile introduced an important concept related to data governance that is the organisation that holds the rights on the dataset (so-called rightsHolder). To this end, in Table 11 below we introduced an indicator in addition to the ones of the MQA of the data.europa.eu (the identifier of the indicator is marked with WHOW to distinguish it from the rest) that is meant to verify the presence of this important information. Moreover, the priority is marked "Essential" for all these mandatory metadata. In contrast, the so-called recommended metadata are marked as "Important". For instance, keywords are important to comply with the findability principle of FAIR but they

[13] https://data.europa.eu/mqa/?locale=en.

are not mandatory in the Italian DCAT-AP_IT application profile and in DCAT-AP and thus their priority is important but not essential.

To conclude, it is worth highlighting that, according to European and national rules, in the presence of geospatial datasets some additional metadata is to be taken into account, guided by the GeoDCAT-AP extension [13], born in the context of the INSPIRE framework. With this regard, in Table 11 we marked "Essential" the indicator related to Geo search and introduced a further indicator about the geospatial reference system.

Table 11: European Metadata Quality Assurance indicators

Indicator ID	FAIR principle of reference	Architectural requirement	Indicator	Method of quantification	Priority
IND-MQA-1	Findability	FR-1401, FR-1403	Keyword usage	The system checks whether keywords are defined. The number of keywords has no impact on the score.	Important
IND-MQA-2	Findability	FR-1401, FR-1402, FR-1403	Categories	It is checked whether one or more categories are assigned to the dataset. The number of assigned categories has no impact on the score.	Essential
IND-MQA-3	Findability	FR-1402	Geo search	It checks whether the property is set or not.	Essential
IND-MQA-4	Findability	FR-1401, FR-1403	Time based search	It checks whether the property is set or not.	Useful
IND-MQA-5	Accessibility	FR-1401 FR-1403	AccessURL accessibility	The specified URL is checked for accessibility via a HTTP HEAD request. If the response status code is in the 200 or 300 range, the accessibility of the resource is evaluated positively.	Essential
IND-MQA-6	Accessibility	FR-1401, FR-1403	DownloadURL	It checks whether the property is set or not.	Important
IND-MQA-7	Accessibility	NFR-03	DownloadURL	The specified URL is	Important

Indicator ID	FAIR principle of reference	Architectural requirement	Indicator	Method of quantification	Priority
			accessibility	checked for accessibility via a HTTP HEAD request. If the response status code is in the 200 or 300 range, the accessibility of the resource is evaluated positively.	
IND-MQA-8	Interoperability	FR-1403	Format	It checks whether the property is set or not.	Essential
IND-MQA-9	Interoperability	FR-1401	Media type	It checks whether the property is set or not.	Useful
IND-MQA-10	Interoperability	FR-1401, FR-1403	Format / Media type from vocabulary	The format vocabulary can be found in the data.europa.eu GitLab repository. The media type is check against the IANA list	Useful
IND-MQA-11	Interoperability	NFR-04, FR-44, FR-45. FR-46	Non-proprietary	The distribution is considered as non-proprietary if the specified format is contained in the corresponding data.europa.eu GitLab repository vocabulary.	Essential
IND-MQA-12	Interoperability	NFR-04, FR-44, FR-45. FR-46	Machine readable	The distribution is considered as machine-readable if the specified format is contained in the corresponding data.europa.eu GitLab repository vocabulary.	Essential
IND-MQA-13	Interoperability	FR-1401	DCAT-AP compliance	The metadata is validated against a set of SHACL shapes. The metadata is not compliant, if the SHACL validation reports	Essential

Indicator ID	FAIR principle of reference	Architectural requirement	Indicator	Method of quantification	Priority
				at least one issue. The MQA uses data.europa.eu's DCAT-AP SHACL validation service.	
IND-MQA-14	Reusability	FR-1401, FR-1403	License information	It checks whether the property is set or not.	Essential
IND-MQA-15	Reusability	FR-1401, FR-1403	License vocabulary	This section describes all dimensions that the MQA examines in order to determine the quality. The dimensions are derived from the FAIR principles. The MQA recommends and credits the usage of controlled vocabularies. The data.europa.eu portal publishes its controlled vocabularies in GitLab. The vocabularies are derived from the EU Vocabularies.	Essential
IND-MQA-16	Reusability	FR-1401	Access restrictions	It checks whether the property is set or not.	Useful
IND-MQA-17	Reusability	FR-1401	Access restrictions vocabulary	It is checked whether the controlled vocabulary for access rights is used.	Useful
IND-MQA-18	Reusability	FR-1401, FR-1403	Contact point	It checks whether the property is set or not.	Important
IND-MQA-19	Findability	FR-1401, FR-1403	Publisher	It checks whether the property is set or not.	Important
IND-MQA-20	Reusability	FR-1401	Rights	It checks whether the property is set or not.	Useful
IND-MQA-21	Accessibility, reusability	FR-1401, FR-1403	File size	It checks whether the	Useful

Indicator ID	FAIR principle of reference	Architectural requirement	Indicator	Method of quantification	Priority
				property is set or not.	
IND-MQA-22	Findability	FR-1401, FR-1403	Date of issue	It checks whether the property is set or not.	Essential
IND-MQA-23	Findability	FR-1401, FR-1403	Modification date	It checks whether the property is set or not.	Essential
IND-WHOW-24	Findability	FR-1403	Rights holder	It checks whether the property is set or not	Essential
IND-WHOW-25	Accessibility, Interoperability	FR-1402	Geospatial reference system	It checks whether the property is set or not	Important

4.3 ISO/IEC 25012 indicators

Improving data quality depends on various factors, including guaranteeing the adherence to shared standard quality models. The ISO/IEC 25012:2008 standard defines a set of specific characteristics for defining data quality. These are divided into three main categories:

- **Inherent data quality**: this category includes such data quality characteristics as accuracy, completeness, consistency, currentness (or timeliness) and credibility;
- **System-dependent data quality**: this category includes data quality characteristics that are more related to the infrastructure that provides users with the data. In this category, the standard includes characteristics like availability, portability and recoverability;
- **Inherent and system-dependent data quality:** this category embraces all those characteristics that are accessibility, compliance, precision, understandability, confidentiality, traceability and efficiency.

Once identified, the next step is to quantify them in terms of metrics identifying, through priorities, how to discriminate the goodness or not of the data with respect to the characteristic under consideration. To this end, there exists the ISO/IEC 25024 standard that extends the 2008 ISO/IEC 25012 "Data quality model" to the field of measurements, defining 63 quality measures applicable to the 15 data quality characteristics previously listed with associated calculation functions.

In the context of the WHOW project, we decided not to reuse all the 15 quality characteristics. The motivations for this choice are twofold: on one hand, when we refer to the quality of the produced datasets from a technical perspective, we mean to assess dimensions that are mostly inherent to the data. On the other hand, we wish to facilitate the applicability of the WHOW data quality framework, offering just the minimum set of quality characteristics that can be viewed as building blocks for the definition of the data quality concept.

Owing to these observations, we selected five characteristics, four of which are also recommended in the Italian national guidelines for open data[14] to be followed by the WHOW partners in the national e-government context. These characteristics are all inherent to the data with the exception of the compliance that is also system-dependent. However, compliance is one of the key objectives of the WHOW project: being compliant with semantic standards, shared ontologies and reference data (controlled vocabularies) is an intrinsic feature of the knowledge graph; therefore, it is essential to include it in the proposed data quality framework.

In Table 12, we list the indicators, the quality characteristics of the standard, the possible metrics for computing the indicator and the priority that we assign to each of them. In this work, we also associate the indicator to the FAIR principles that guide the definition of most of the technical KPIs of the WHOW data quality framework.

Please note that for the same characteristic more than one indicator (and then metric) can be specified. For instance, in the case of completeness, we define it with respect to the schema, record or entire population.

Table 12: Data quality indicators from ISO/IEC 25012 and 25024

Indicator ID	**FAIR principle of reference**	**Architectural requirement**	**Quality characteristic**	**Indicator**	**Metric and Method of quantification**	**Priority**
IND-ISO25012-1D	Reusability	FR-11	Accuracy	Syntactic accuracy	Ratio (normalised between 0 and 1) between data attributes with syntactically accurate values on the number of data attributes for which syntactic accuracy is required **Method of quantification:** If ratio <= 0.2 ivery bad, if ratio > 0.2 and <=0.4 insufficient, if > 0.4 and <= 0.6 sufficient, if ratio > 0.6 and <=0.8 good, if ratio > 0.8 and <= 1 excellent	Essential
IND-ISO25012-2D	Interoperability, Reusability	FR-10, FR-17	Accuracy	Semantic accuracy	Ratio (normalized between 0 and 1) between data attributes with semantically accurate	Essential

[14] (ITA) https://docs.italia.it/italia/daf/lg-patrimonio-pubblico/it/stabile/aspettiorg.html#caratteristiche-di-qualita.

Indicator ID	FAIR principle of reference	Architectural requirement	Quality characteristic	Indicator	Metric and Method of quantification	Priority
					values on the number of data attributes for which semantic accuracy is required **Method of quantification**: If ratio <= 0.2 very bad, if ratio > 0.2 and <=0.4 insufficient, if > 0.4 and <= 0.6 sufficient, if ratio > 0.6 and <=0.8 good, if ratio > 0.8 and <= 1 excellent	
IND-ISO25012-3DM	Findability, Accessibility, Reusability	FR-10	Completeness	Schema completeness	Percentage of null values for concepts and properties with respect to the total number of expected values **Method of quantification**: If ratio <= 0.2 very bad, if ratio > 0.2 and <=0.4 insufficient, if > 0.4 and <= 0.6 sufficient, if ratio > 0.6 and <=0.8 good, if ratio > 0.8 and <= 1 excellent	Important
IND-ISO25012-4D	Findability, Accessibility, Reusability	FR-10	Completeness	Record completeness	Number of essential data in a record associated to a non-null value, with respect to the number of essential data in a record which it is possible to measure completeness **Method of quantification**: If ratio <= 0.2 very bad, if ratio > 0.2 and <=0.4 insufficient, if > 0.4 and <= 0.6 sufficient, if ratio > 0.6 and <=0.8	Important

Indicator ID	FAIR principle of reference	Architectural requirement	Quality characteristic	Indicator	Metric and Method of quantification	Priority
					good, if ratio > 0.8 and <= 1 excellent	
IND-ISO25012-5D	Findability, Interoperability, Accessibility,, Reusability	FR-09	Completeness	Population completeness	Percentage of null values with respect to a reference population **Method of quantification**: If ratio <= 0.2 very bad, if ratio > 0.2 and <=0.4 insufficient, if > 0.4 and <= 0.6 sufficient, if ratio > 0.6 and <=0.8 good, if ratio > 0.8 and <= 1 excellent	Important
IND-ISO25012-6D	Findability, interoperability, Reusability	FR-09, FR-11	Consistency	Data Consistency - The degree to which the attributes of the data are not in contradiction with other data in a specific specific context	Ratio of data attributes whose values are semantically correct in the dataset on the number of data attributes semantic rules have been defined for. Ratio of the number of duplicate values for each attribute of the knowledge base on the total number of elements in the knowledge base. **Method of quantification**: If ratio <= 0.2 very bad, if ratio > 0.2 and <=0.4 insufficient, if > 0.4 and <= 0.6 sufficient, if ratio > 0.6 and <=0.8 good, if ratio > 0.8 and <= 1 excellent	Important
IND-ISO25012-7D	Accessibility Reusability	FR12, FR13	Currentness	Data timeliness - Data is up-to-	Based on the use of metadata that indicates	Essential

Indicator ID	FAIR principle of reference	Architectural requirement	Quality characteristic	Indicator	Metric and Method of quantification	Priority
				date with respect to a data frequency of update	when the data was last updated. We distinguish: - **data with a known update periodicity**: data of last update is within the frequency of update with respect to a measurement time; **Method of quantification:** Yes/No - **data with an average update periodicity:** in this case it is possible to calculate the average timeliness with a percentage error. **Method of quantification**: If ratio <= 0.2 ivery bad, if ratio > 0.2 and <=0.4 insufficient, if > 0.4 and <= 0.6 sufficient, if ratio > 0.6 and <=0.8 good, if ratio > 0.8 and <= 1 excellent	
IND-ISO25012-8D	Interoperability	FR-10, FR-14	Compliance	Standard conformity - The degree to which data attributes adhere to existing standards, conventions or regulations	The metric is calculated as the ratio between: (a) Number of data attributes that have values and/or formats compliant with reference standards or regulations (b) Total number of data attributes that have to comply with reference standards or regulations **Method of**	Essential

Indicator ID	FAIR principle of reference	Architectural requirement	Quality characteristic	Indicator	Metric and Method of quantification	Priority
					quantification: If ratio <= 0.2 very bad, if ratio > 0.2 and <=0.4 insufficient, if > 0.4 and <= 0.6 sufficient, if ratio > 0.6 and <=0.8 good, if ratio > 0.8 and <= 1 excellent	

4.4 Semantic quality indicators

Semantic quality evaluation is also a fundamental step to assess both the data quality offered by the WHOW knowledge graph and the efficiency of the WHOW toolkit, as described in the WHOW reference Linked Open Data Architecture deliverable. In this section we list a set of indicators that we selected among the many possible ones cited in [8] and [9] and developed in [10]. Selected parameters of interest have been taken into account and their measures compared to [11] and [12]. As per [11], the framework of reference *OQuaRE* for evaluating the quality of ontologies has been included for the evaluation of the method of quantification.

The indicators we selected and propose in Table 13 are:

- indicators related to the technical content of the ontologies being developed. These types of indicators are useful to assess the *cognitive ergonomics* of the semantic modules. For cognitive ergonomics we mean a principle based on which an ontology can be easily understood, manipulated and exploited by users;
- indicators regarding assessment methods on the *ontology usability.* These indicators are useful to evaluate the transparency and the degree of understandability the ontology can offer to users so as to facilitate its (re)usage in various contexts;
- indicators about the *coverage of the requirements* the ontology can satisfy. This is particularly relevant also for understanding the ability of the semantic model(s) of the WHOW knowledge graph to answer important questions that can be posed by users on the open datasets;
- indicators on the *logical consistency* the ontology can guarantee, meaning the principle based on which an ontology can be easily processed by a reasoner.

Table 13: Semantic quality indicators

Indicator ID	Architectural Requirement	Topic	Indicator	Metrics and method of quantification	Priority
IND-SEM-1	NFR-04	Cognitive ergonomics	Depth	$DITOnto = \max_{1 \leq i \leq N_{leafs}} depth(C_i)$ The metric is calculated as the	Essential

Indicator ID	Architectural Requirement	Topic	Indicator	Metrics and method of quantification	Priority
				maximum depth of a leaf class where Ci is the ith class in the ontology and N is the total number of classes in the ontology. Method of quantification: If ratio <= 8 not acceptable, if ratio >= 6 and <=8 insufficient, if > 4 and <= 6 minimally acceptable, if ratio > 2 and < 4 good, if ratio > 1 and < 2 exceeds the expectations	
IND-SEM-2	NFR-04	Cognitive ergonomics	Tangledness	$TMOnto = \frac{\Sigma_{i=1}^{N} DP(C_i)}{N}$ The metric is calculated as the ratio of multiple parent classes to the total numbers of classes in an ontology Method of quantification: If ratio <= 8 not acceptable, if ratio >= 6 and <=8 insufficient, if > 4 and <= 6 minimally acceptable, if ratio > 2 and < 4 good, if ratio > 1 and < 2 exceeds the expectations	Essential
IND-SEM-4	NFR-04	Cognitive ergonomics	Number of properties per class	The metric is calculated as the number of properties (NOMOnto): Number of properties per class. It is calculated as follows: NOMOnto=∑| PCi| ⁄∑|Ci| Method of quantification: If ratio <= 8 not acceptable, if ratio >= 6 and <=8 insufficient, if > 4 and <= 6 minimally acceptable, if ratio > 2 and < 4 good, if ratio <= 2 exceeds the expectations	Essential
IND-SEM-3	NFR-04, NFR-07, NFR-08	Usability	Class richness	Class Richness (CROnto): its metrics is calculated as the mean number of instances per class. It is calculated as follows: CROnto=∑| ICi| / ∑|Ci|; where ICi, is the set of individuals of the Ciclass.	Essential

Indicator ID	Architectural Requirement	Topic	Indicator	Metrics and method of quantification	Priority
				Method of quantification: If ratio <= 20% not acceptable, if ratio >= 20% and <=40% insufficient, if > 40% and <= 60% minimally acceptable, if ratio > 60% and < 80% good, if ratio < 80% exceeds the expectations	
IND-SEM-4	FR-34	Usability	Numerousness of annotations	Presence of at least one multilingual label and at least one multilingual comment for each entity Method of quantification: Yes/No	Important
IND-SEM-5	NFR-04, NFR-07, NFR-08	Usability	Documentation	Presence of documentation and examples of usage Method of quantification: Yes/No	Important
IND-SEM-6	NFR-04	Usability	Use of naming conventions	Verification of correct naming conventions for ontology entities IDs Method of quantification: Yes/No	Important
IND-SEM-7	FR-47, NFR-04, NFR-09	Logical consistency	Inference verification	It checks the presence of logical inconsistencies by means of a DL reasoner. Method of quantification: Yes/No	Essential
IND-SEM-8	NFR-10	Requirements coverage	Conversion of competency questions into SPARQL queries	Ratio between number of CQ that can be converted into SPARQL queries and total number of CQ Method of quantification: Yes/No	Essential

4.4 Evaluation Method

The method used to evaluate the proposed technical indicators and the related metrics is sometimes referred to as the "Method of quantification" in the tables reported in the previous sections.

For some technical indicators, the method of quantification is the pass-or-fail that allows us to verify whether a resource/dataset meets the requirements expressed in the indicators on a binary basis (Yes/No). There are however cases in which the method of quantification of the indicators and related metrics indicates a percentage or a number (e.g., the number of observed entities and blank nodes in the knowledge graph).

However, both types of quantification methods, together with the assigned priorities can be used for the successive phase in the assessment that attributes a final judgement on the **degree of technical data quality achieved** by the published linked open datasets.

In deliverable 2.1, we already proposed a possible methodology for the final stage of the evaluation. In particular, we revise it and envisage different levels of technical data quality that can be progressively obtained:

- **Very bad data quality level** → when none of the technical indicators are satisfied (i.e., the pass-or-fail-based indicators assumes value No and percentage-based indicators assumes values <= 0.2);
- **Insufficient data quality level** → when only one third of the essential technical indicators are satisfied (i.e., the pass-or-fail-based indicators assumes value Yes and percentage/number based indicators assumes values in the range (0.2-0.4]);
- **Sufficient data quality level** → when only half of the essential technical indicators are satisfied (i.e., the pass-or-fail-based indicators assumes value Yes, and percentage-based indicators assumes values in the range (0.4-0.6]);
- **Good data quality level** → when all the essential technical indicators are satisfied (i.e., the pass-or-fail-based indicators assumes value Yes, and percentage- based indicators assumes values in the range (0.6-0.8]);
- **Very Good data quality level** → when the essential and important technical indicators are satisfied (i.e., the pass-or-fail-based indicators assumes value Yes, and percentage/number based indicators assumes values in the range (0.6-0.8]);
- **Excellent data quality level** → when all technical indicators are satisfied (i.e., the pass-or-fail-based indicators assume value Yes, and percentage-based indicators assume values in the range (0.8-1]).

5 Conclusions

In this deliverable we presented the WHOW data quality framework that consists of a set of business and technical indicators for data quality, and an evaluation methodology that can be applied when releasing the WHOW knowledge graph.

The framework aggregates already existing indicators that are currently scattered across a variety of evaluation models and tools and adds some indicators related to the specific application domain(s) of the project.

In the framework, relevant SDGs monitoring indicators coming from SDG 3, SDG 6, SDG 11, SDG 13 and SDG 14 are also considered to assess how the WHOW knowledge graph can contribute to their achievement.

References

[1] European Commission, data.europa.eu, “Open Data Maturity Report” (2021) - available at https://data.europa.eu/sites/default/files/landscaping_insight_report_n7_2021_0.pdf. DOI: 10.2830/394148.

[2] Open Data Institute (ODI), “Open Data Certificates - the mark of quality and trust of open data” (2022) - available at https://certificates.theodi.org/en/.

[3] World Wide Web Foundation, “Open Data Barometer Global Report- Third Edition” (2016) - available at https://opendatabarometer.org/3rdedition/report/.

[4] FAIR Data Maturity Model Working Group. (2020). FAIR Data Maturity Model. Specification and Guidelines (1.0). https://doi.org/10.15497/rda00050.

[5] DCAT application profile for data portals in Europe (2021) EU ISA Programme (ISA) https://joinup.ec.europa.eu/solution/dcat-application-profile-data-portals-europe.

[6] Data Catalog Vocabulary (DCAT) - Version 2, W3C Recommendation (2020), https://www.w3.org/TR/vocab-dcat-2/.

[7] DCAT-AP_IT v1.1 – Italian Application profile of DCAT-AP (2016), available in ITA at https://www.dati.gov.it/content/dcat-ap-it-v10-profilo-italiano-dcat-ap-0.

[8] Aldo Gangemi et al. “Modelling Ontology Evaluation and Validation”. In: *Proc. of ESWC 2006*. (Budva, Montenegro).
Springer, 2006, pp. 140–154.

[9] Luigi Asprino et al. “Observing LOD using Equivalent Set Graphs: it is mostly flat and sparsely linked”. In: *Proc. of ISWC 2019*, 2019, pp.57–74.

[10] Valentina Carriero et al., “ArCo: the Italian cultural heritage knowledge Graph”. In: 18th International Semantic Web Conference. (Auckland, New Zealand). 2019, pp. 36–52.

[11] S. M. Alghamdi *et al.* “Quantitative evaluation of ontology design patterns for combining pathology and anatomy ontologies”. In: *Sci Rep* 9**,** 4025 (2019).

[12] Yao et. al., “Benchmarking Ontologies: Bigger or Better?”. In: *PLOS Computational Biology*, Online. DOI: https://doi.org/10.1371/journal.pcbi.1001055.

[13] European Commission, “GeoDCAT-AP” available at https://inspire.ec.europa.eu/good-practice/geodcat-ap.

[14] United Nations, “Is the UNECE region on track for 2030? Assessment, stories and insights” (2021) available at https://unece.org/sites/default/files/2021-03/SDG-Report%202021-withcovers.pdf.

[15] United Nations, “E-handbook on SDG indicators”. Available at https://unstats.un.org/wiki/display/SDGeHandbook/Home

[16] ISPRA, “Annual report for environmental data” - available in italian at https://annuario.isprambiente.it/sys_ind/macro

[17] J.D Sachs et al, “Sustainable Development Report 2021 - The Decade of Action for the Sustainable Development Goals”, Cambridge University Press (2021).

[18] Unesco, “Conjunction Water Management - A powerful contribution to achieving Sustainable Development Goals”, 2020, available at: https://books.google.it/books?id=Sf4bEAAAQBAJ&pg=PA17&lpg=PA17&dq=Once+%E2%80%9Csufficient+information+is+available%E2%80%9D+(FAO,+2018),+water+accounting,+numerical+modelling+and+other+types+of+analysis+may+contribute+to+gaining+in-depth+knowledge+on+the+water+systems%E2%80%99+performance+and+on+the+water+services.&source=bl&ots=fCEe8ym9CL&sig=ACfU3U1O4vzrnBDybEn-cDsr3kiOlJbpnQ&hl=en&sa=X&ved=2ahUKEwi69rWE9Lj3AhX0gf0HHV_rBagQ6AF6BAgCEAM

[19] Hall, N.L., Creamer, S., Anders, W. *et al.* “Water and health interlinkages of the sustainable development goals in remote Indigenous Australia”. *npj Clean Water* **3,** 10 (2020). https://doi.org/10.1038/s41545-020-0060-z