



Water Health Open knowLedge (WHOW)

Data pre-processing is finalised

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	Knowledge graph definition
Deliverable number	3.1 [Milestone #4]
Version (date)	1 (2022-05-30)
Due date	2021-05-31
Responsible organisation	National Council of Research (ISTC-CNR)
Editors	ISTC-CNR, Aria SpA, ISPRA
Abstract	This deliverable introduces the data pre-processing that is necessary to be carried out at data providers sides in order to prepare the data to be transformed in the WHOW knowledge graph

Keywords	Data processing, RDF, knowledge graph, data provision, technical design
-----------------	--

Editor(s)

Anna Sofia Lippolis, Giorgia Lodi, Andrea Giovanni Nuzzolese (Institute of Cognitive Sciences and Technologies of the Italian National Research Council - ISTC-CNR).

Gianluca Carletti (ARIA SpA)

Elio Giulianelli, Marco Picone, Giulio Settanta (ISPRA)

Reviewers

Francesco Poggi (Institute of Cognitive Sciences and Technologies of the Italian National Research Council - ISTC-CNR)

Annalisa Minelli (ISPRA)

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).

Disclaimer

The sole responsibility of this publication lies with the author. The European Union is not responsible for any use that may be made of the information contained therein.

Confidentiality



The information in this document is public and can be used according to the following licence CC-BY 4.0.

Change Log

Version	Date	Organisation	Description
0.1	2022-03-31	ISTC-CNR	Creation of the Table of Content
0.2	2022-04-13	ISTC-CNR, ARIA SoA	Draft of section 3 and 2.1
0.3	2022-04-21	ALL	Draft of Executive Summary, Introduction, revisions of section 3 and 2.1, draft of section 2.2
0.4	2022-04-29	ALL	Revisions of all sections
0.5	2022-05-09	ALL	Document ready for internal review
0.6	2022-05-26	Aria SpA, ISPRA	Revisions of section 2.1 and 2.2 based on reviewers' comments
0.9	2022-05-28	ISTC-CNR	Revisions of some points of the documents based on the reviewers' comments
1.0	2022-05-30	ALL	Document ready for delivery

Table of contents

List of Tables	7
List of Figures	8
Executive Summary	9
Introduction	10
Project Overview	10
Objectives	10
Relationships with other activities	11
Structure of the document	12
Data pre-processing at WHOW data providers	13
Data pre-processing at ARIA S.p.A.	13
Open data Pre-processing overview	13
Under the ingestion step, the data owner is in charge of pre-processing activities	14
Under the publication step, pre-processing is in charge of the open data team	14
Pre-processing procedures for datasets used in the WHOW use cases	14
Dataset provided directly by ARPA Lombardia	15
Datasets provided by ARPA Lombardia via the Aria S.p.A. Open Data portal	15
Regione Lombardia Health-related datasets	16
Regione Lombardia Geoportal Datasets	17
Regione Lombardia Environmental risk areas datasets	17
Data pre-processing at ISPRA	18
Privacy preserving processes	Errore. Il segnalibro non è definito.
Data pre-processing in the WHOW Linked Open Data reference architecture	32
WHOW data providers' datasets: pre-processing needs	34
3.1.1 Formats handling	35
3.1.2 Encoding handling	35
3.1.3 Strings handling	35
3.1.4 Dates handling	36
3.1.5 Handling of incoherent data with respect to the expected content	37
3.1.6 Handling of data that conveys more than one semantic concept at a time	38

Pre-processing implementation	40
Conclusions	43
References	44

List of Tables

Table 1: List of datasets provided directly by ARPA Lombardia

Table 2: Datasets provided by ARPA Lombardia via the Aria S.p.A. Open Data portal

Table 3: Regione Lombardia datasets about health

Table 4: Regione Lombardia Geoportal Datasets

Table 5: Regione Lombardia Environmental risk areas datasets

Table 6: Example data structure of ISPRA-SINTAI System measurements aggregated by monitoring points (Stations)

Table 7: Quality controls for Eionet water data

Table 8: Same entity written in three different ways in three datasets

Table 9: Result obtained by harmonising the strings for naming production

Table 10: Example of “Data” column of Dataset 5 before and after pre-processing of dates

Table 11: Comparative result obtained by distinguishing CAS, WISE and other alphanumeric codes

List of Figures

- Figure 1: Relationships with other deliverables (and their due time) and activities
- Figure 2: Overview of open data pre-processing process
- Figure 3: Deployment in a single buoy
- Figure 4: Deployment in the central concentration infrastructure
- Figure 5: Pesticides data collection and storage workflow
- Figure 6: Procedure overview
- Figure 7: WHOW high level linked open data reference architecture. In the architecture the data preparation layer is highlighted by the box with the violet dotted border.
- Figure 8: Data pre-processing layer design
- Figure 9: Example of date format in the “Data” column of Dataset 5
- Figure 10: Example of incoherent data in the “CAS” column in Dataset 1
- Figure 11: Example of multiple semantic concepts in the “Valore” column of Dataset 1
- Figure 12: Result of the pre-processing operation of the “Valore” column in Dataset 1
- Figure 13: Scheme of pre-processing process

Executive Summary

When dealing with an increasing available number of open datasets, possibly coming from a variety of data sources and actors, the probability of encountering anomalies in the data is quite high. It is in fact simply unrealistic to expect that the data is perfect: errors can be introduced in different steps of the data management process, during the data collection phase, during the data processing and even during the data publication phase.

Especially in the open data context, where data providers tend to publish datasets with the almost sole purpose of enabling a public consultation by human-beings (concentrating on a transparency principle, only), quality issues in data structures and content emerge quite frequently. However, these issues become very easily concrete barriers for the re-use in applications and services development scenarios, where machines need to understand and treat the data. As already described in deliverable 5.1 on SDGs and KPIs [5], the FAIR principles (Findability, Accessibility, Interoperability, Reusability) which we base the WHOW project on, guide the openness of the data we consider. This approach is also promoted by recent initiatives of open data communities such as the Italian “datiBeneComune” initiative, where a published online report¹ clearly asks government institutions to open data having in mind these principles, thus strongly limiting the aforementioned barriers.

The need therefore to ensure that:

- data is properly prepared for any machine-based processing;
- data quality characteristics also of the standard ISO/IEC 25012, we also identified in deliverable 5.1 on SDGs and KPIs [5], are met;
- personal data protection, if applicable, is satisfied according to the related GDPR regulation;

is of utmost importance to create a sound and sustainable linked open data production process, and typically it involves more than half of time in data cleansing activities.

This deliverable describes the operations carried out at each data provider of the WHOW project regarding the data pre-processing; that is, a preparatory step in the data management process (i.e., linked open data process) necessary to enable an effective successive machine elaboration of the data.

In the deliverable, we distinguish between two layers of data pre-processing:

- pre-processing done before entering the WHOW-specific linked open data process by WHOW data providers within their existing open data infrastructures;
- pre-processing that is carried out in the WHOW designed linked open data processes. This is enabled to prepare the datasets, identified as relevant for the three use cases, to be transformed into the open knowledge graph (Linked Open Data) of each data provider.

In essence, this deliverable represents a focus on a specific phase foreseen in the overall design of the linked open data process, as described in deliverable D3.2 - “Linked Open Data process design is finalised”.

¹ <https://vorrei.datibenecomune.it/dati-che-vorrei/come-li-vorrei/>

1 Introduction

This is deliverable “3.1 - Data pre-processing is finalised”. It is the result of the activities conducted in the context of task 3.2 of Activity 3 related to “Knowledge Graph 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) [1]. 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 sets 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 describes the operations to be performed on the data in order to prepare it to be transformed into a knowledge graph with a harmonised semantics and uniform standard data format.

In the design of the Linked Open Data Reference Architecture of the project, a specific layer on data preparation is foreseen which addresses a number of functional requirements (e.g. FR-01 and FR-08) such as those related to the possible manipulations to be performed on the data in order to prepare it for the creation of the knowledge graph.

In general, data can be subject to a variety of quality issues and anomalies, especially when it comes from various sources and stakeholders. Examples of these anomalies include the use of different standard formats for the dates, the extensive use of strings rather than codes to identify things of the real world, that can differ from one dataset to another even when referring to the same thing (e.g., “Lago di Como” or “Como - lago”), inconsistencies with respect to the semantics expressed by specific column names in tabular datasets, etc.

Data processed by ARIA and ISPRA data providers, in their respective open data and Linked Open Data infrastructures, typically undergo a number of operations prior to their publication or sharing, defined on the basis of quality checks implemented in the deployed data management processes. This is particularly true in all those public institutions that act as aggregators with respect to a plethora of other administrations, as is the case of ISPRA and ARIA: these quality checks are thus absolutely necessary prior to every data publication.

Notwithstanding these controls, further processing activities on the data are necessary before the transformation of the data into the knowledge graph, based on a standard format such as RDF and a common semantic layer (represented in WHOW by the whow ontology network we are creating). Therefore, even in the case of the specific WHOW processes we are developing at each data provider, defined with respect to the datasets identified for the three use cases of Deliverable 2.1, data pre-processing operations have to be performed.

This deliverable first introduces the data pre-processing activities already performed by ARIA and ISPRA in their open data infrastructures and second discusses additional pre-processing operations to be performed on the already processed data so as to prepare it for the transformation in RDF according to defined common ontologies and standards. The deliverable represents a focus of a specific phase of the linked open data process defined in Deliverable 3.2.

1.3 Relationships with other activities

The present deliverable D3.1 is an another important milestone (#4) of the WHOW project. It covers all the activities of task 3.2 that are linked with other tasks of other activities of the project.

The following Figure 1 shows such relationships.

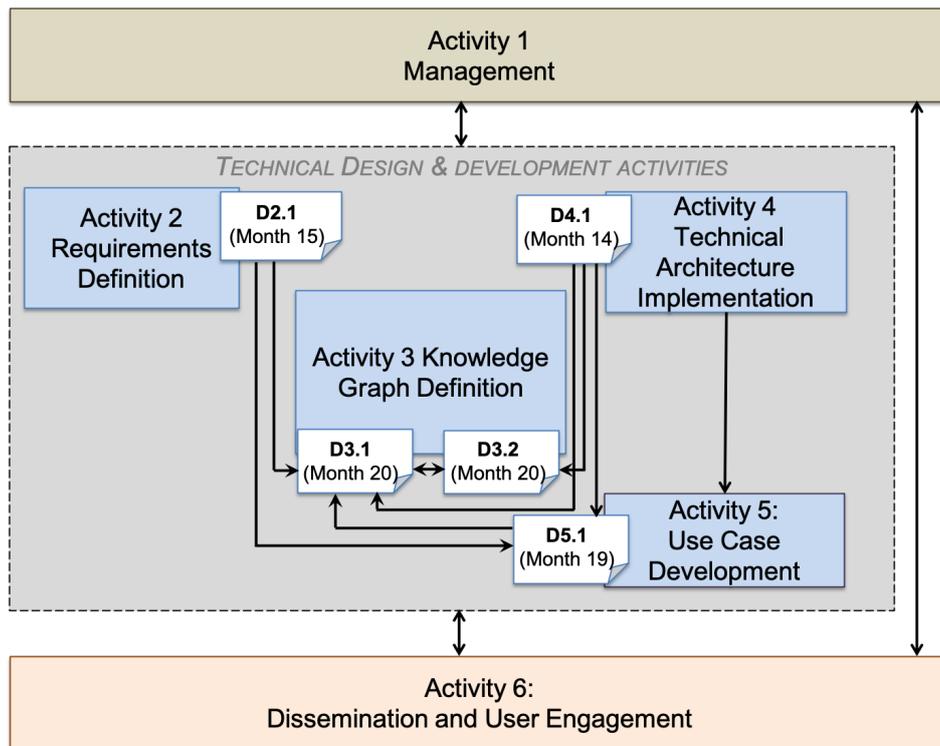


Figure 1: Relationships with other deliverables (and their due time) and activities

Specifically, the content of this deliverable is strictly dependent on Deliverable 3.2 on the linked open data process design. In fact, data pre-processing introduced here is a specific phase of the overall process, also reported in the deliverable on the Linked Open Data Reference Architecture (Deliverable 4.1)..

In addition, the content of the present deliverable depends on the datasets as identified in the definition of the use cases (Deliverable 2.1) and on the data quality KPIs that have been defined in Deliverable 5.1 , as shown in Figure 1.

1.4 Structure of the document

Section 2 describes the state of the art of the pre-processing activities currently in place at ISPRA and ARIA open data infrastructures. A specific focus on the privacy data quality characteristic is provided in this section when applicable.

Section 3 introduces additional pre-processing operations that are performed in the context of WHOW specific data management processes, on the datasets identified as relevant for the three use cases of the project. The section describes the typical operations that are performed, providing examples with the original datasets to be then transformed in RDF for knowledge graph production purposes.

Finally, Section 4 concludes the deliverable.

2 Data pre-processing at WHOW data providers

This section discusses state of the art data pre-processing processes currently deployed at Aria S.p.A. and ISPRA.

2.1 Data pre-processing at ARIA S.p.A.

This section provides a brief description of general pre-processing policies at ARIA S.p.A. (Azienda Regionale per l’Innovazione e gli Acquisti) and the actual pre-processing procedures on the datasets needed to realise the WHOW use cases.

Open data Pre-processing overview

The open data portal is implemented by Socrata Connected Government Cloud (SCGC) SaaS (Software as a Service) solution. The reader can find a detailed description about the open data portal and the open data process in the deliverable 3.2 - “Linked Open Data Process Design is finalised”.

In summary, the Open Data process is performed by the data owner, responsible for the data, and the open data team, responsible for the operations on the open data portal. It consists of 3 steps:

- *Ingestion*: the data owner and the open data team agree on the licence and publication policies. Based on that, the data owner is in charge of pre-processing activities that consist of privacy and data quality checks.
- *Publication*: the data owner is in charge of checking whether the dataset is consistent for publication and of uploading it on the open data portal. To do that, pre-processing activities are initiated in automatic, semi automatic or manual modes: they are again about privacy and data quality checks. Lastly, when the dataset is ready, this is published to the SCGC using an automated procedure.
- *Access*: any end-users can access, visualise and download the dataset.

As shown in Figure 2, pre-processing happens both during the Ingestion and publication steps.

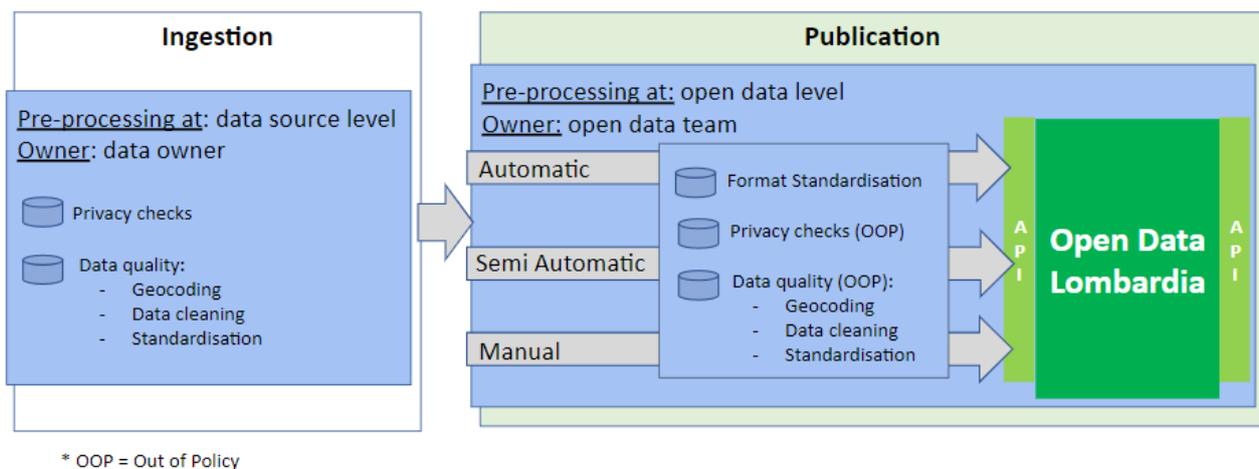


Figure 2: Overview of open data pre-processing process

Under the ingestion step, the data owner is in charge of pre-processing activities

After the data owner and the open data team agree on licence and publishing policies, the data owner performs privacy and data quality checks. They are carried out by the data owner rather than the open data team, so that once the dataset version is given to the open data team it is ready to be published on the open data SCGC Cloud.

As for the privacy checks, the data owner is responsible for assessing that no personal information is included in the dataset: if any, this should be either erased or obscured. Furthermore, when it comes to the compliance with the statistical confidentiality, this is again a data owner's task. Although it does not involve personal data, it could allow unique identification of a person due to the limited number of observations. For instance, this happens when anagraphic data (e.g., age) related to some other data (e.g. a dietary regime) make personal identification possible without using personal data that directly identifies a person.

As far as the quality checks are concerned, the most frequent tasks are the following:

- *geocoding*: automatic recognition of incomplete and / or misspelt addresses, even if in different formats;
- *data cleaning*: filtering anomalous cells, checking for mistakes and fixing them. Common examples are duplicates (duplicated observations), erroneous data joining, inconsistency between cells (e.g. working status is unemployed but job title is not NA);
- *format standardisation*: it principally involves date and address format and type validity, capital lettering, female and male nouns/adjectives rather than neutral.

Under the publication step, pre-processing is executed by the open data team

At this stage, pre-processing is meant to be limited to: small changes as format standardisation (e.g. date and time format), checks related to the correct processing of URL addresses (e.g. the variables included in the dataset are consistent with the ones communicated to the open data teams) and checks related to data updating, based on the updating frequency communicated to the open data team and included in the metadata file.

As explained earlier, the data owner is supposed to provide the metadata file and the dataset with privacy and data quality checks already performed. Nevertheless, sometimes, the data owner provides datasets not ready to be published and the open data team performs "Out of Policy" data manipulation tasks that produce the files ready to be published.

The open data team normally executes a local check. In extreme cases, if inconsistencies or personal data are spotted in datasets, the open data team may operate to fix the dataset or ask for further changes to the data owner.

With regard to the privacy checks for the variables to be included, they consist of a comparison between column names and a list of words, commonly related to personal information (social security number, name, surname, address...). In case of match, then the column is not inserted.

Pre-processing procedures for datasets used in the WHOW use cases

In this paragraph, the reader finds a detailed overview about pre-processing procedures applied to the datasets involved in the WHOW use cases implementation. Datasets can be grouped by homogeneous pre-processing procedures into:

- Datasets provided directly by ARPA Lombardia (Agency for Environment Protection in Lombardy)
- Datasets provided by ARPA Lombardia via the Aria S.p.A. Open Data portal
- Regione Lombardia Health-related datasets
- Regione Lombardia Geoportal datasets
- Regione Lombardia Environmental risk areas datasets

Datasets provided directly by ARPA Lombardia

ARPA Lombardia provides datasets on the environmental domain, about chemical, physical, micro-bacteriological monitoring of surface and ground water. In the following Table 1, the list of datasets is included.

Table 1: List of datasets provided directly by ARPA Lombardia

Dataset name	Format
PFAS data (perfluoroalkyl substances) surface waters - Year 2018	XLSX
PFAS data (perfluoroalkyl substances) groundwater - Year 2018	XLSX
Analytical data of river water bodies	XLSX
Analytical data of lake water bodies	XLSX
Analytical data of groundwater	XLSX
Chemical status of groundwater	XLSX
Environmental Radioactivity Monitoring Network	XLSX
Weekly inflows per basin	XLSX
Weekly outflows per basin	XLSX
Monthly inflows per basin	XLSX
Monthly outflows per basin	XLSX
Height of the lakes	XLSX

ARPA Lombardia publishes these datasets only on its own portal and, as a consequence, these datasets are not available from the Regione Lombardia Open data Portal and they do not undergo pre-processing activities at publication level, as summarised in the previous paragraph.

All datasets listed in Table 1 are not available on the Aria S.p.A. open data catalogue, but rather on the ARPA open data portal. As these ones show inconsistencies both in semantics and syntax (as further detailed in Section 3), the data owner - ARPA - will have to run some pre-processing activities at the data source level to clean the data. After that, in order to be included in the linked open data project, they should be uploaded also to the Aria S.p.A. open data portal.

Datasets provided by ARPA Lombardia via the Aria S.p.A. Open Data portal

This group includes datasets about weather and waterways conditions. In the following Table 2, the list of datasets is included.

Table 2: Datasets provided by ARPA Lombardia via the Aria S.p.A. Open Data portal

Dataset name	Format
Weather sensor data	CSV, JSON, TSV, XML
Meteorological Stations	CSV, JSON, TSV, XML
Interpolation of hourly precipitation observations	ZIP, TXT
Flow rate (or hydrometric height) data relating to the watercourse monitoring network	CSV, JSON, TSV, XML
Surface Waters (lakes) - LTLECO - LTLeco is a descriptor that integrates the values of 3 parameters detected on the lake: total phosphorus, transparency and hypolimnic oxygen. Reference year: 2013	XLSX
Surface Waters (water courses) - LIMECO - LIMeco: descriptor that integrates the values of 4 parameters measured on a watercourse: ammonia nitrogen, nitric nitrogen, total phosphorus and dissolved oxygen (100 -% saturation). Reference year: 2013	XLSX

At the data source level, for all these datasets, ARPA is in charge of the pre-processing tasks.

At the publication level, for all datasets, the open data team does not run pre-processing tasks. Only for the “Flow rate (or hydrometric height) data relating to the watercourse monitoring network”, LTLECO 2013 and LIMECO 2013, the publication mode is manual; for the other ones it is automatic.

Regione Lombardia Health-related datasets

This group includes datasets from the epidemiological observatory of Regione Lombardia, whose information is available in the open data portal specific section. These datasets cover drug consumption, medical hospitalizations, medical services offered by Regional Health System and reported diseases - including infectious ones.

Mostly, they are regional datasets fed through data flows transmitted by the hospitals and healthcare authorities. This data gathering process is standardised and based on a unique regional tool: before being stored to the relevant regional thematic information systems, a dataset has to pass both syntax and semantics checks, implemented specifically for the dataset type.

After the process is finalised, these datasets are uploaded to the regional health data warehouse, and then, through predefined views, they are gathered into the open data system for the subsequent steps. During the publication step, the open data team does not run pre-processing tasks.

Table 3: Regione Lombardia datasets about health

Dataset name	Format
Consumption of drugs in Regione Lombardia for level I ATC	CSV, JSON, TSV, XML
Indicative dataset ESAC consumption of antibiotics	CSV, JSON, TSV, XML
Dataset Hospital Assistance for ACC and ATS	CSV, JSON, TSV, XML
Antibiotic consumption dataset in terms of DDD for level III and IV	CSV, JSON, TSV, XML
Health conditions dataset by province, gender and cause	CSV, JSON, TSV, XML
Infectious diseases Regione Lombardia rates by sex and age	CSV, JSON, TSV, XML
Dataset Health conditions by municipality, age and gender	CSV, JSON, TSV, XML

Dataset name	Format
Delivery of drugs in Regione Lombardia for the first 10 second-level ATCs	CSV, JSON, TSV, XML
Number of drugs dispensed in Regione Lombardia of the first 10 active ingredients	CSV, JSON, TSV, XML
Hospital Assistance for DRG and ATS	CSV, JSON, TSV, XML
Regione Lombardia SDO dataset	CSV, JSON, TSV, XML
Dataset Average Hospitalization And Average Accesses By Discipline Discipline And ATS	CSV, JSON, TSV, XML
Dataset Hospital Assistance for MDC and ATS	CSV, JSON, TSV, XML
Dataset Hospital Assistance for Discipline Discipline and ATS	CSV, JSON, TSV, XML
Dataset Medium Inpatient And Medium Access For MDC And Facility	CSV, JSON, TSV, XML
Dataset Average Hospitalization And Average Accesses By Discipline Discipline And Structure	CSV, JSON, TSV, XML
Dataset Medium Inpatient And Medium Access For ACC And Facility	CSV, JSON, TSV, XML
Number of drug packs dispensed in Regione Lombardia	CSV, JSON, TSV, XML
Dataset Medium Inpatient And Medium Access For DRG AND ATS	CSV, JSON, TSV, XML
Dataset Medium Inpatient And Medium Access For MDC And ATS	CSV, JSON, TSV, XML

Regione Lombardia Geoportal Datasets

Datasets published on the geoportal cover territory characteristics, important to study water quality and extreme events effects.

Table 4: Regione Lombardia Geoportal Datasets

Dataset name	Format
Soil defence works	ZIP, SHP
Large dams	ZIP, SHP
Unified regional hydrographic network	ZIP

As these datasets are already included in the regional geoportal, they do not undergo any pre-processing activities on the open data portal.

Regione Lombardia Environmental risk areas datasets

This group again covers territory characteristics, important to study water quality and extreme events effects. In detail, these are about zones at hydro-weather risk, contaminated areas or recently reclaimed in Regione Lombardia.

Table 5: Regione Lombardia Environmental risk areas datasets

Dataset name	Format
List of reclaimed sites in Lombardy - Year 2020	CSV, JSON, TSV, XML
List of contaminated sites in Lombardy - Year 2020	CSV, JSON, TSV, XML
Homogeneous zones for hydro-weather risk - List of Municipalities	CSV, JSON, TSV, XML
Homogeneous zones for hydro-weather risk - Centroids	CSV, JSON, TSV, XML

For this datasets group, Regione Lombardia is in charge of pre-processing tasks at the data source level.

For the first two datasets (“List of reclaimed sites in Lombardy - Year 2020”, “List of contaminated sites in Lombardy - Year 2020”) the publication mode is semi-automatic and open data pre-processing consists of geo-coding tasks.

For the other two datasets (“Homogeneous zones for hydro-weather risk - List of Municipalities”, “Homogeneous zones for hydro-weather risk - Centroids”) publication mode is manual and no pre-processing tasks are run by the open data team.

2.2 Data pre-processing at ISPRA

ISPRA owns and manages a large number of environmental data structures (more than 150, each containing more than one dataset), deriving from institutional activities, European reporting, and research projects. Some of these descend from monitoring activities implemented by ISPRA, others derived from data collection activity from local and regional institutions, following the establishment of the National Environmental Information Network as imposed by the Italian legislation (Law 28 June 2016, no. 132).

The information is extremely heterogeneous. There are observations, indicators, samplings, censuses, territorial defence interventions. The data is punctual or on grids (regular or not), time series, spatial series, transects, spatial data (areas, perimeters or points). Values can be expressed as numbers, amounts, strings, dates. A lot of information is organised in databases, others in GIS layers. The file formats for the acquisition and dissemination are heterogeneous, i.e. shapefile, raster, text, binary, spreadsheet, xml, but also specific formats such as NetCDF, GRIB.

Given the complexity of the environmental information managed by ISPRA, it is extremely difficult to imagine common pre-processing procedures. Activities of harmonisation, validation, correction of the acquired information can be carried out during the data acquisition, ingestion and publication of the data.

Within ISPRA it is possible to classify the datasets on the basis of the pre-processing procedures into three groups.

- Data produced and published by ISPRA. It includes all the situations in which ISPRA follows the entire information production chain, from the design of the acquisition system to publication. In this case, ISPRA can identify the suitable pre-processing procedures for each specific chain, based on national or international directives or guidelines. These procedures tend to produce validated and harmonised data.
- Data produced by local authorities and published by ISPRA: in this framework local / regional authorities / regional environmental protection agencies are responsible for data acquisition, while ISPRA has to collect and publish the information on a national basis. In this case, ISPRA adopts pre-processing procedures which have the main purpose to harmonise information from different sources.
- Data produced by ISPRA or other entities and disseminated by ISPRA through official supranational channels: often used for the transfer of information for the European reporting. The pre-processing procedures are often codified at European level and provide for the harmonisation of information and formats, the identification of anomalous values, the validation of data.

An example will be given for each of them, taking into account some dataset useful for the WHOW project.

For the datasets that ISPRA proposes for the project, there are no problems relating to the disclosure of sensitive data, and for this reason no information manipulation operations are carried out in accordance with the European GDPR legislation.

2.2.1 Case 1 - Marine Monitoring Networks

ISPRA manages several environmental monitoring networks. In the marine field, together with other institutional and research activities, the institute manages two important marine-weather observational networks. The national wave network (15 buoys located off the Italian coast in open sea, devoted to the physical condition of the sea and meteorological components measurement) and the national tide gauge network (36 stations along the Italian coast for sea levels and meteorological parameters measurement). ISPRA accumulated more than twenty years of experience in the entire data production and management chain, from the infrastructure design to the dissemination of the observed data. Therefore, over the years, it was possible to consolidate measurement operations, formats and protocols useful for the transfer and data dissemination activities, in order to minimise the operations after the marine observation.

Figures 3 and 4 show a standard operational deployment for a single measurement station and the data concentration infrastructure in ISPRA, respectively.

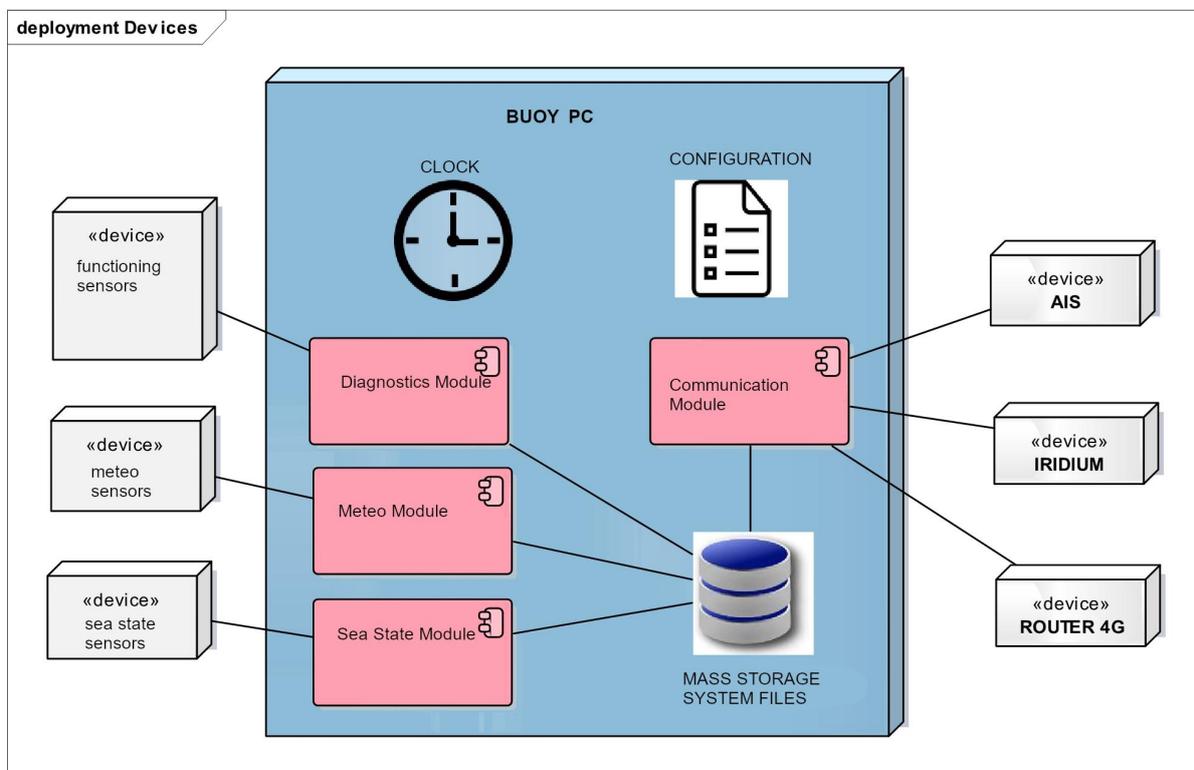


Figure 3: Deployment in a single buoy

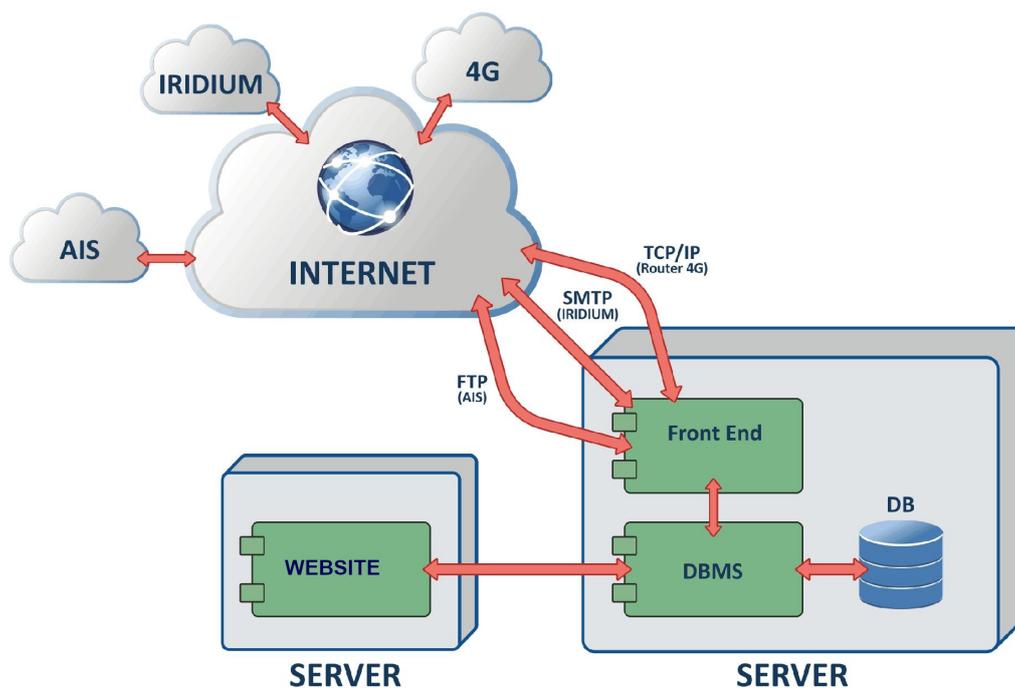


Figure 4: Deployment in the central concentration infrastructure

The pre-processing operations are performed in real time at two separate moments:

- After the observation at the station: data quality checks are performed through the definition of appropriate thresholds (depending on the used sensors and the observed phenomena) and diagnostic checks regarding the instrumentation functioning (control of the battery voltage, absence of interference during the data communication).
- At the concentration infrastructure: checks are performed on the amount of received data, the accuracy and integrity of the delivered messages, the continuity of transmission systems and database synchronisation.

Further operations are executed at a deferred time. They include the validation of the historical series through the comparison of measured values, the identification of repeated measurements, the removal of anomalous values.

2.2.2 Case 2: the Pesticide dataset

The SINTAI Information System (*ISPRA system for the dissemination and consultation of national water data*)² manages the information collected annually from the Italian regions for the Eionet network, according to the Water Framework Directive 2000/60/EC and Legislative Decree 152/2006 and its implementation decrees. The time schedule is instead on a six years base for the *Water Information System*

² <https://www.sintai.isprambiente.it/>.

for Europe (WISE). The collected data regard the environmental status of Italian inland waters, and contain water monitoring data.

The SIMP Information System (*ISPRA Information System for Pesticide Monitoring*) is the part of SINTAI that annually collects and process pesticide measurements made in inland waters, plus information on monitoring points from the Italian Regions/Autonomous Provinces³. The workflow is reported in Fig. 3.

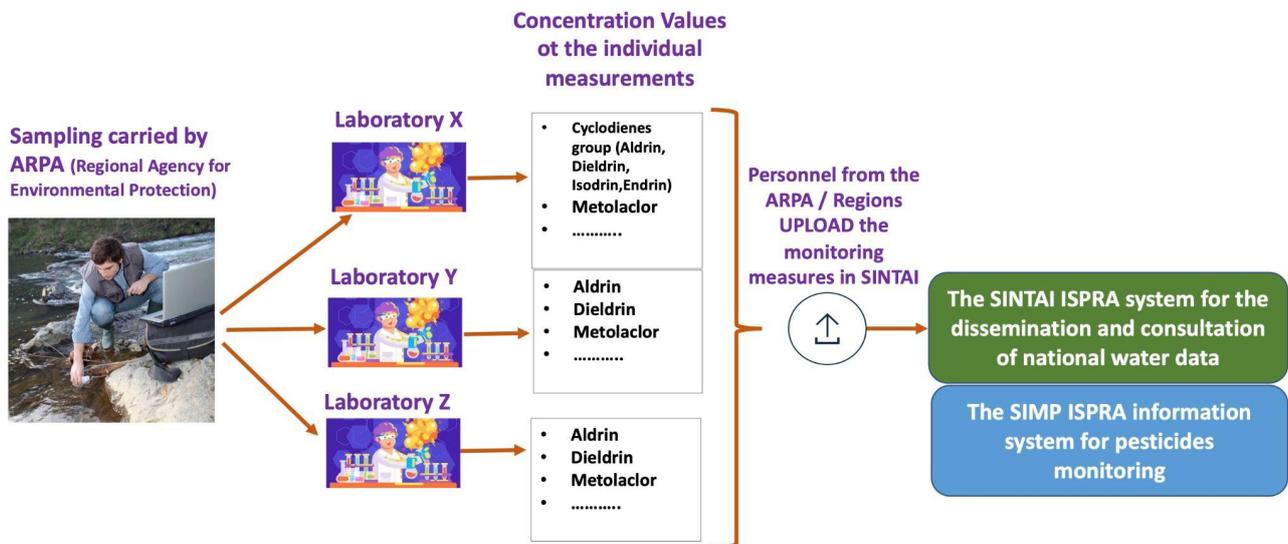


Figure 5: Pesticides data collection and storage workflow

The data transmitted to ISPRA by the Italian regions have the format of **regional tables** as Excel sheets, where data are reported in terms of:

- **Stations**, which contain the information on the monitoring points;
- **Monitoring**, which contains the individual measurements of the substances.

The tables are checked in advance and a data cleansing procedure is performed, before they are stored into a database.

2.2.3 Description of procedures

Regional agencies *upload* the Excel sheets populated with monitoring information in SINTAI; the SIMP accesses the SINTAI, checks if there are new boards (**regional tables**) and the cleansing procedure is carried out.

³ Measurement data are provided through the Regional (ARPA) and Provincial (APPA) Agencies for Environmental Protection, which carry out surveys on the territory and laboratory analyses and transmit the data collected to ISPRA.

After the check/cleansing process in the ISPRA-SINTAI system, the data are aggregated by monitoring point (station) to assess the compliance with EQS (Environmental Quality Standards). An example is reported in Tab. 1. Finally the data will be published according to the Linked Open Data paradigm.

Table 6: Example data structure of ISPRA-SINTAI System measurements aggregated by monitoring points (Stations)

REGION CODE	STATION CODE	YEAR	LONG	LAT	CAS	SUBSTANCE	CONC MAX	CONC AVG	...
01	1095	2020	7.699934	45.065070	118-74-1	ESACLOROBENZENE	0.091	0.01266667	...
01	87010	2020	8.670582	44.680673	118-74-1	ESACLOROBENZENE	0.004	0.0016	...
01	19020	2020	8.540616	45.173581	122-34-9	SIMAZINA	0,03	0.01333333	...

The control process is performed in both an automated and a semi-automated way and consists of the following steps:

1. *Data completeness*: ISPRA experts check if the expected table format has been met by the regions, otherwise they either modify it to meet the requirements or ask the regional contact person (ARPA / APPA) to send the tables again in the established format.
2. *Data cleansing*: The SIMP performs automatic checks / data cleansing on the data corresponding to measurements and monitoring points, enters the monitoring data in the SIMP database and ends the data validation phase. In case of errors found, results are returned in an online report and in log files. The automatic controls are performed by inserting the Excel sheets **Stations** and **Monitoring** in temporary tables, and by executing appropriate stored procedures.

The procedure described above is outlined in Figure 4.

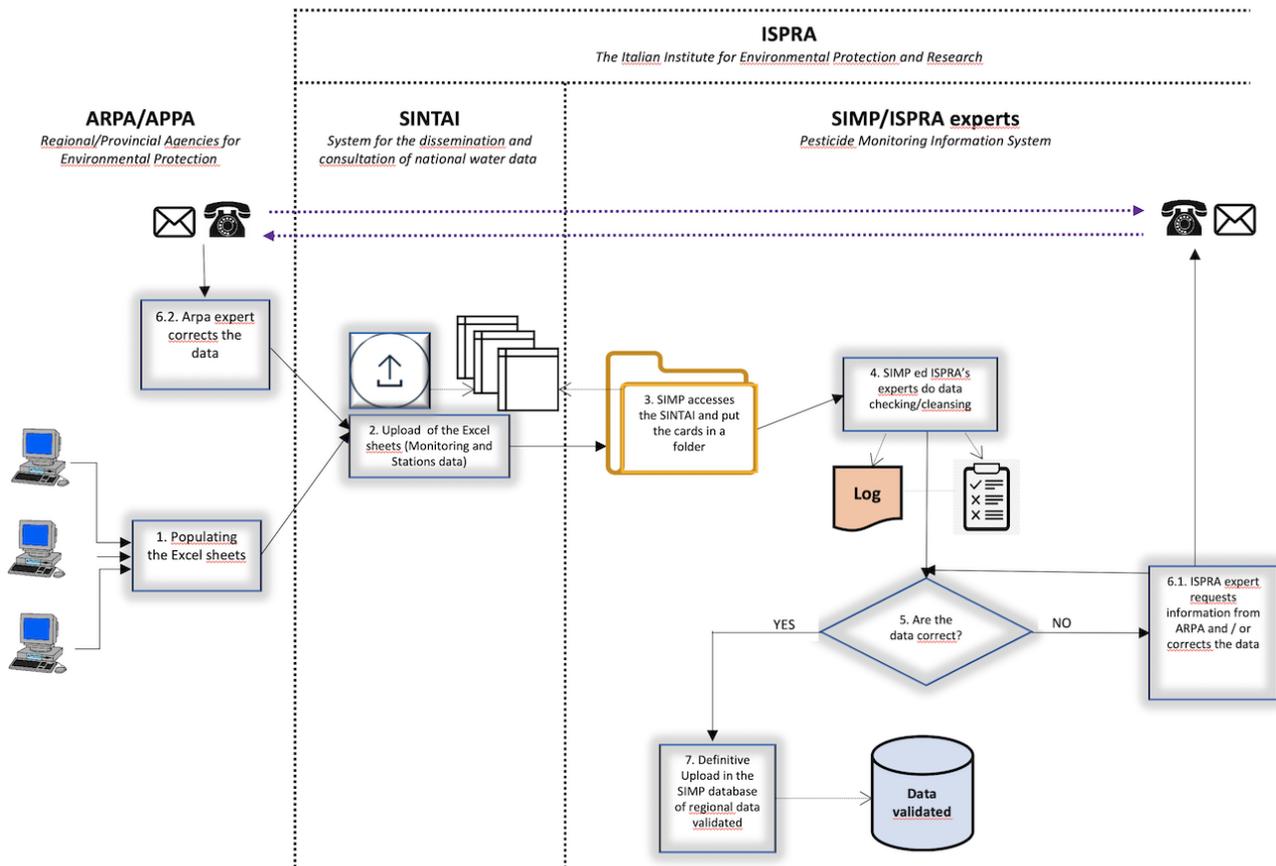


Figure 6: Procedure overview

Data completeness. A number of controls are performed for checking data completeness. First, the presence of all mandatory fields is verified.

Mandatory fields for **Stations** are:

- year;
- region code;
- station type (i.e. surface/groundwater station SW/GW);
- station code;
- information required for georeferencing (if EPSG can't be computed and associated with the cartographic or metric coordinates transmitted, a blocking error is reported).

Mandatory fields for **Monitoring** (apart the station referenced, which must exist in **Stations**) are:

- substance code (*cas*);
- date [DD/MM/YYYY];
- concentration [$\mu\text{g/L}$].

Then a syntactic check of mandatory fields is applied. In particular, a blocking error is raised if:

-
- "region code"⁴ has more than 2 digits or is incorrect;
 - "municipality code"⁵ has more than 6 digits or is incorrect;
 - the geographical coordinates "station long" or "station lat" are missing or are not numeric;
 - the "station datum" is not present;
 - the "station projection" or "station zone" is not present and the coordinates are not geographic;
 - the coordinates of the individual monitoring point do not fall within the administrative limits (ISTAT) of the Region or Autonomous Province it belongs to;
 - monitoring has a negative value for "concentration", or a value too high, or with instrument limit of quantification (LQ) not congruent;
 - substance code (cas) does not exist in the code list database;
 - duplication of measures with different concentration values.

Data cleansing. The first operation is to verify that no marine-coastal station ('station type') is present in the dataset. If it happens that marine-coastal stations are present, then these stations and their associated data (e.g. measurements) are deleted accordingly. The field "station type" must necessarily be SW or GW. Values for river water (RW), lake water (LW), and TW (river, lake and transition) are automatically converted to SW. Then additional data cleansing includes the following operations:

- all "station codes" are converted to uppercase;
- the "region code" is changed to 2 digits by placing "0" in case the value consists of 1 digit only (i.e. the value is in the [0,9] range);
- the "municipality code" is forced to be 6 digits long by placing "0" in front of the number when the digits are less than 6;
- in "station long" and "station lat" all commas (i.e. character ",") are replaced by points (i.e. character ".");
- the georeferencing of the monitoring points (**Stations**) is done by a transformation of coordinates into a single reference system;
- all measurements of non-pesticide substances in the **Monitoring** are removed;
- all duplicate measurements are removed.

2.2.4 Case 3 - Eionet dataset

The European Environment Agency (EEA) is an agency of the European Union that provides independent and qualified information on the environment to policy makers and the public with timely, targeted, relevant and reliable data. The EEA's main collaborators are the European Commission, the European Parliament, the Council of the EU, other EU institutions, the governments of the participating countries as well as the scientific, academic and business communities, NGOs.

To support the EEA mission, in 1994 the partnership network Eionet, European Environment Information and Observation Network, was born. Eionet is made up of the EEA, 27 member states of the European Union plus Iceland, Liechtenstein, Norway, Switzerland and Turkey, and six Western Balkan countries. The main purposes of the Eionet network are the collection and exchange of data, information, indicators and analyses as well as infrastructures and standards.

⁴ Code list from regional administrative unit plus autonomous provinces of Bozen/Trento from ISTAT.

⁵ Code list from municipality administrative unit from ISTAT.

The nodes of the network are the National Focal Points (NFP), which are responsible for the coordination of the networks of national experts for the various environmental issues, and the National Reference Centres (NRC). In addition, to support the collection, management and analysis of data, the EEA has set up consortia between the organisations of the countries belonging to the Eionet network, divided according to the main environmental issues, called European Topic Centres (ETC).

The mandate to represent Italy was entrusted to ISPRA, and as for the NFP, a team of experts has been identified at the DG-SINA Service. Currently, the National Reference Centres (NRCs) bring together ISPRA experts supported by SNPA (National System for Environmental Protection) experts.

The data and information collected are used by the EEA for the creation of indicator data and thematic publications, and made available on the Agency's website.

The type of data collected responds either to specific European regulations or to agreements made among the member countries of the EEA as they are deemed necessary for the activity of the EEA itself.

Concerning the water framework, data refer to the condition and quality of rivers, lakes, groundwater, marine, coastal and transitional waters, release of pollutants into water and quantitative aspects of water resources. There are two channels for data flow, resulting from the monitoring activities at national level:

- The WISE-SoE (Water Information System for Europe-State of Environment) data flow, powered by SNPA monitoring system and agreed within EEA, aimed at producing the SoE Report (SOER).
- The data flow derived from the implementation of EU directives (91/276 / EEC - Nitrates, 91/271 / EEC - Urban wastewater, 2000/60 / EC - Waters, 2007/60 / EC - Floods, 2008/56 / EC - Marine strategy).

ISPRA represents, through the national system SINTAI (Information System for the Protection of Waters in Italy) and the SIC (Centralised Information System for Marine Strategy), the Italian node of the WISE system (Water Information System for Europe). All the data produced by the system of regional (ARPA) and provincial (APPA) agencies are available in SINTAI and SIC. In particular, the SIC collects, manages and shares the MSFD marine-coastal monitoring data at community level, including data deriving from the monitoring campaigns carried out by ISPRA and by other third parties appointed by ISPRA.

ISPRA has the role of reviewing regional data and the national guidelines for Community Reporting, in line with the guidance documents of the European Commission (EC). It is also in charge of making them available, together with the updated Information Standards, on the national SINTAI and SIC systems, as well as coordinating the collection and the release of data on the ReportNet platform of the EEA. The data collected through the ReportNet platform are processed on a regional, national and European scale by the European ETC thematic centres and then flow into the WISE system, to allow information access.

As reported on Eionet guidelines⁶, a set of validation procedures are applied to data collected annually from ARPAs and APPAs. In the water framework, the general applied quality controls are summarised in Table 7.

Each test is performed by automatic procedures and could return an output codified as follow:

⁶ <https://www.eionet.europa.eu/>

- **BLOCKER:** they indicate that the detected error will prevent data submission (data release is not possible).
- **ERROR:** these messages indicate issues that clearly need corrective action by the data reporter.
- **WARNING:** such messages indicate issues that may be an error. Data reporters are expected to double-check relevant records.
- **INFO:** Informative message. Neutral or statistical feedback about the delivery, e.g. number of species reported.

Table 7: Quality controls for Eionet water data

Test Label	Test Description	Status
0.a Data type test	Tests whether the reported values follow the data type defined in the dataset specifications (http://dd.eionet.europa.eu/tables/11122). Tested data types are all numeric, date, date-time and boolean. The records that fail this test must be fixed before they can be checked by the other tests.	BLOCKER
0.b Data constraints test	Tests whether there are any other records which could not be imported into the database for testing for any other reason. One of the possible reasons is that the values are too long, but it can be a different one. The records that fail this test must be fixed before they can be checked by the other tests.	BLOCKER
1 Mandatory values test	Tests the presence of the mandatory values: <ul style="list-style-type: none"> • [monitoringSiteIdentifier] • [monitoringSiteIdentifierScheme] • [parameterWaterBodyCategory] • [observedPropertyDeterminandCode] • [procedureAnalysedMatrix] • [phenomenonTimeReferenceYear] • [parameterSampleDepth] 	BLOCKER
2.a Mandatory values test - conditional - missing result values unjustified	Tests records for the missing result values, which are not justified by using an appropriate flag in the [resultObservationStatus] (http://dd.eionet.europa.eu/dataelements/95711) field. The result values are: <ul style="list-style-type: none"> • [resultUom] • [resultNumberOfSamples] • [resultQualityMinimumBelowLOQ] • [resultMinimumValue] • [resultQualityMeanBelowLOQ] • [resultMeanValue] • [resultQualityMaximumBelowLOQ] • [resultMaximumValue] • [resultQualityMedianBelowLOQ] • [resultMedianValue] 	BLOCKER

2.b Mandatory values test - conditional - missing result values justified	List of records with missing result values, which are justified by the appropriate flag in the [resultObservationStatus] field.	INFO
3 Mandatory values test - conditional - [procedureLOQValue], [resultQualityNumberOfSamplesBelowLOQ]	Tests whether the [procedureLOQValue] and [resultQualityNumberOfSamplesBelowLOQ] is reported for selected determinands. The WISE6 ObservedProperty QC reference (http://cdr.eionet.europa.eu/help/WISE_SoE/wise6/WISE6_ObservedProperty_QC_reference.xlsx) file, in the WISE6 CDR Help (http://cdr.eionet.europa.eu/help/WISE_SoE/wise6), provides reference for the rules used in this test.	ERROR
4 Conflicting values test - 'missing' result values	Tests whether the following values are empty if it is justified by using an appropriate flag in the [resultObservationStatus] (http://dd.eionet.europa.eu/dataelements/95711) field: <ul style="list-style-type: none"> • [resultUom] • [resultNumberOfSamples] • [resultQualityMinimumBelowLOQ] • [resultMinimumValue] • [resultQualityMeanBelowLOQ] • [resultMeanValue] • [resultQualityMaximumBelowLOQ] • [resultMaximumValue] • [resultQualityMedianBelowLOQ] • [resultMedianValue] 	BLOCKER
5 Record uniqueness test	Tests the uniqueness of the records. The following combination of values must be unique with no duplicate records existing: <ul style="list-style-type: none"> • [monitoringSiteIdentifier] • [monitoringSiteIdentifierScheme] • [observedPropertyDeterminandCode] • [procedureAnalysedMatrix] • [phenomenonTimeReferenceYear] • [parameterSampleDepth] 	BLOCKER
6 Valid codes test	Tests the validity of the values against the respective code lists: <ul style="list-style-type: none"> • [monitoringSiteIdentifierScheme] (http://dd.eionet.europa.eu/fixedvalues/elem/75870) • [parameterWaterBodyCategory] (http://dd.eionet.europa.eu/vocabulary/wise/WFDWaterBodyCategory) • [observedPropertyDeterminandCode] (http://dd.eionet.europa.eu/vocabulary/wise/ObservedProperty) • [procedureAnalysedMatrix] (http://dd.eionet.europa.eu/vocabulary/wise/Matrix) • [resultUom] (http://dd.eionet.europa.eu/vocabulary/wise/Uom) • [resultObservationStatus] 	BLOCKER

	(http://dd.eionet.europa.eu/fixedvalues/elem/95711)	
7.a Monitoring site identifier test - format	<p>Tests the validity of the [monitoringSiteIdentifier] value format:</p> <p>1) The country code part of the identifier value must match the one of the reporting country ('UK' for the United Kingdom and 'EL' for Greece).</p> <p>2) The identifier value can't contain punctuation marks, white space or other special characters, including accented characters, except for "-" or "_". Presence of two or more consecutive "-" or "_" characters ("--" or "__"), or their combination ("_-" or "-_"), is however not allowed. The identifier value must use only upper case letters. The third character, following the 2-letter country code, and the last character can't be "-" or "_". The total length of the identifier can't exceed 42 characters.</p> <p>(Regular expressions: <code>^[A-Z]{2}[0-9A-Z]{1}([0-9A-Z_\-]{0,38}[0-9A-Z]{1}){0,1}\$</code> and <code>^[A-Z0-9](\- _)?+\$</code>)</p>	BLOCKER
7.b Monitoring site identifier test - reference	<p>Tests the presence of the [monitoringSiteIdentifier] and its respective [monitoringSiteIdentifierScheme] in the official reference list (http://dd.eionet.europa.eu/vocabulary/wise/MonitoringSite). Only the valid, retired and deprecated identifiers are accepted. Superseded or non-existing are not.</p> <p>The list has been created from previously reported data on monitoring sites. New monitoring sites must be reported via WISE-5 reporting, well before the time series reporting. The time is needed for processing of the delivery and update of the reference list.</p>	BLOCKER
7.c Monitoring site identifier test - retired and deprecated identifiers	List of records with monitoring site identifiers that are retired or deprecated.	WARNING
7.d Monitoring site identifier test - water body category	Tests whether the reported [parameterWaterBodyCategory] matches the category ([specialisedZoneType]) of the water body, to which the respective monitoring site is officially assigned, as reported in the WFD or WISE-5 reporting.	ERROR
8 The [observedPropertyDeterminandCode] test - unexpected	<p>Tests whether the [observedPropertyDeterminandCode] values are expected to be reported in this table.</p> <p>The WISE6 ObservedProperty QC reference (http://cdr.eionet.europa.eu/help/WISE_SoE/wise6/WISE6_ObservedProperty_QC_reference.xlsx) file, in the WISE6 CDR Help (http://cdr.eionet.europa.eu/help/WISE_SoE/wise6), provides reference for the rules used in this test.</p>	WARNING
9 The	Tests whether correct [resultUom] values have been used for the	BLOCKER

[observedPropertyDeterminandCode] and [resultUom] coherence test	observed determinands. The WISE6 ObservedProperty QC reference (http://cdr.eionet.europa.eu/help/WISE_SoE/wise6/WISE6_ObservedProperty_QC_reference.xlsx) file, in the WISE6 CDR Help (http://cdr.eionet.europa.eu/help/WISE_SoE/wise6), provides reference for the rules used in this test.	
10 The [procedureAnalysed Matrix] test	Tests whether the table contains only Water data.	BLOCKER
11.a Value constraints test - numeric parameter and result values	Tests whether the numeric result and parameter values follow the constraints set in the dataset specifications (http://dd.eionet.europa.eu/tables/11500): 1) [resultNumberOfSamples] >=1 2) [resultQualityNumberOfSamplesBelowLOQ], [resultStandardDeviationValue] and [parameterSampleDepth] >= 0 3) [parameterSampleDepth] < 11000	BLOCKER
11.b Value constraints test - [phenomenonTimeReferenceYear]	Tests whether the [phenomenonTimeReferenceYear] values are within the expected range.	WARNING
11.c Value constraints test - [parameterSamplingPeriod]	Tests whether the [parameterSamplingPeriod] value 1) is provided in the requested format (YYYY-MM-DD--YYYY-MM-DD or YYYY-MM--YYYY-MM); 2) starting date is not higher than ending date; 3) represents a period of maximum one year; 4) matches with the value provided in the [phenomenonTimeReferenceYear] field	WARNING
12.a The result value limit test - acceptable limits	Tests whether the following result values follow the acceptable limits for the respective observed determinands: <ul style="list-style-type: none"> • resultMinimumValue] • [resultMeanValue] • [resultMaximumValue] • [resultMedianValue] The WISE6 ObservedProperty QC reference (http://cdr.eionet.europa.eu/help/WISE_SoE/wise6/WISE6_ObservedProperty_QC_reference.xlsx) file, in the WISE6 CDR Help (http://cdr.eionet.europa.eu/help/WISE_SoE/wise6), provides reference for the rules used in this test.	BLOCKER
12.b The result value limit test - expected range	Tests whether the following result values are within the commonly expected range for the respective observed determinands: <ul style="list-style-type: none"> • resultMinimumValue] 	WARNING

	<ul style="list-style-type: none"> • [resultMeanValue] • [resultMaximumValue] • [resultMedianValue] <p>The WISE6 ObservedProperty QC reference (http://cdr.eionet.europa.eu/help/WISE_SoE/wise6/WISE6_ObservedProperty_QC_reference.xlsx) file, in the WISE6 CDR Help (http://cdr.eionet.europa.eu/help/WISE_SoE/wise6), provides reference for the rules used in this test.</p>	
12.c The result value limit test - confirmed outliers	<p>List of records with confirmed result values outside the commonly expected range for the respective observed determinands.</p> <p>The WISE6 ObservedProperty QC reference (http://cdr.eionet.europa.eu/help/WISE_SoE/wise6/WISE6_ObservedProperty_QC_reference.xlsx) file, in the WISE6 CDR Help (http://cdr.eionet.europa.eu/help/WISE_SoE/wise6), provides reference for the rules used in this test.</p>	INFO
13.a Logical coherency rule test - result values	<p>Tests the following logical coherence rules regarding the result values:</p> <ol style="list-style-type: none"> 1) [resultMeanValue] >= [resultMinimumValue] 2) [resultMaximumValue] >= [resultMeanValue] 3) [resultMedianValue] >= [resultMinimumValue] 4) [resultMaximumValue] >= [resultMedianValue] 5) [resultMaximumValue] >= [resultMinimumValue] 6) [resultStandardDeviationValue] <= ([resultMaximumValue] - [resultMinimumValue]) 7) IF [resultMinimumValue] < [resultMaximumValue] THEN [resultStandardDeviationValue] > 0 8) IF [resultNumberOfSamples] = 1 THEN [resultMinimumValue] = [resultMeanValue] = [resultMaximumValue] = [resultMedianValue] 9) IF [resultNumberOfSamples] = 1 THEN [resultStandardDeviationValue] = 0 10) [resultQualityNumberOfSamplesBelowLOQ] <= [resultNumberOfSamples] 11) IF [resultQualityNumberOfSamplesBelowLOQ] = 0 THEN [resultQualityMinimumBelowLOQ] = [resultQualityMeanBelowLOQ] = [resultQualityMaximumBelowLOQ] = [resultQualityMedianBelowLOQ] = False 12) IF [resultNumberOfSamples] = 1 THEN [resultQualityMinimumBelowLOQ] = [resultQualityMeanBelowLOQ] = [resultQualityMaximumBelowLOQ] = [resultQualityMedianBelowLOQ] 13) IF [resultQualityNumberOfSamplesBelowLOQ] = [resultNumberOfSamples] THEN [resultQualityMinimumBelowLOQ] = 	BLOCKER

	[resultQualityMeanBelowLOQ] = [resultQualityMaximumBelowLOQ] = [resultQualityMedianBelowLOQ] = True	
13.b Logical coherency rule test - result statistics values and [procedureLOQValue]	Tests the following logical coherence rules: 1) [resultMinimumValue] >= [procedureLOQValue] 2) [resultMeanValue] >= [procedureLOQValue] 3) [resultMaximumValue] >= [procedureLOQValue] 4) [resultMedianValue] >= [procedureLOQValue]	BLOCKER
13.c Logical coherency rule test - result below LOQ = True, result statistics values and [procedureLOQValue]	Tests the following logical coherence rules: 1) IF [resultQualityMinimumBelowLOQ] = True THEN [resultMinimumValue] = [procedureLOQValue] 2) IF [resultQualityMeanBelowLOQ] = True THEN [resultMeanValue] = [procedureLOQValue] 3) IF [resultQualityMaximumBelowLOQ] = True THEN [resultMaximumValue] = [procedureLOQValue] 4) IF [resultQualityMedianBelowLOQ] = True THEN [resultMedianValue] = [procedureLOQValue]	ERROR
13.d Logical coherency rule test - result below LOQ = False, result statistics values and [procedureLOQValue]	Tests the following logical coherence rules: 1) IF [resultQualityMinimumBelowLOQ] = False THEN [resultMinimumValue] > [procedureLOQValue] 2) IF [resultQualityMeanBelowLOQ] = False THEN [resultMeanValue] > [procedureLOQValue] 3) IF [resultQualityMaximumBelowLOQ] = False THEN [resultMaximumValue] > [procedureLOQValue] 4) IF [resultQualityMedianBelowLOQ] = False THEN [resultMedianValue] > [procedureLOQValue]	WARNING
1 Record uniqueness test - AggregatedData	Tests the uniqueness of the records across all delivered files. The following combination of values must be unique with no duplicate records existing: <ul style="list-style-type: none"> • [monitoringSiteIdentifier] • [monitoringSiteIdentifierScheme] • [observedPropertyDeterminandCode] • [procedureAnalysedMatrix] • [phenomenonTimeReferenceYear] • [parameterSampleDepth] The records creating duplicates only within the same file, are not shown in this test result. They are shown in the result of the respective 'Record uniqueness test' of that specific file.	BLOCKER

Additional specific tests can then be implemented for particular datasets.

3 Data pre-processing in the WHOW Linked Open Data reference architecture

Notwithstanding the activities that data providers perform to process the data in such a way as to guarantee a certain level of quality, further pre-processing of the datasets that feed into the WHOW process (see deliverable 3.2 for more details) is necessary to prepare them for their transformation into Linked Open Data, according to a common semantic model.

The architectural layer devoted to data pre-processing is the layer named **data preparation**. For the sake of readability and self-containment of this deliverable, we include Figure 7 that illustrates the high level architecture we described in deliverable 4.1, with a focus on the involved layer. The latter is highlighted by the box with the violet dashed border.

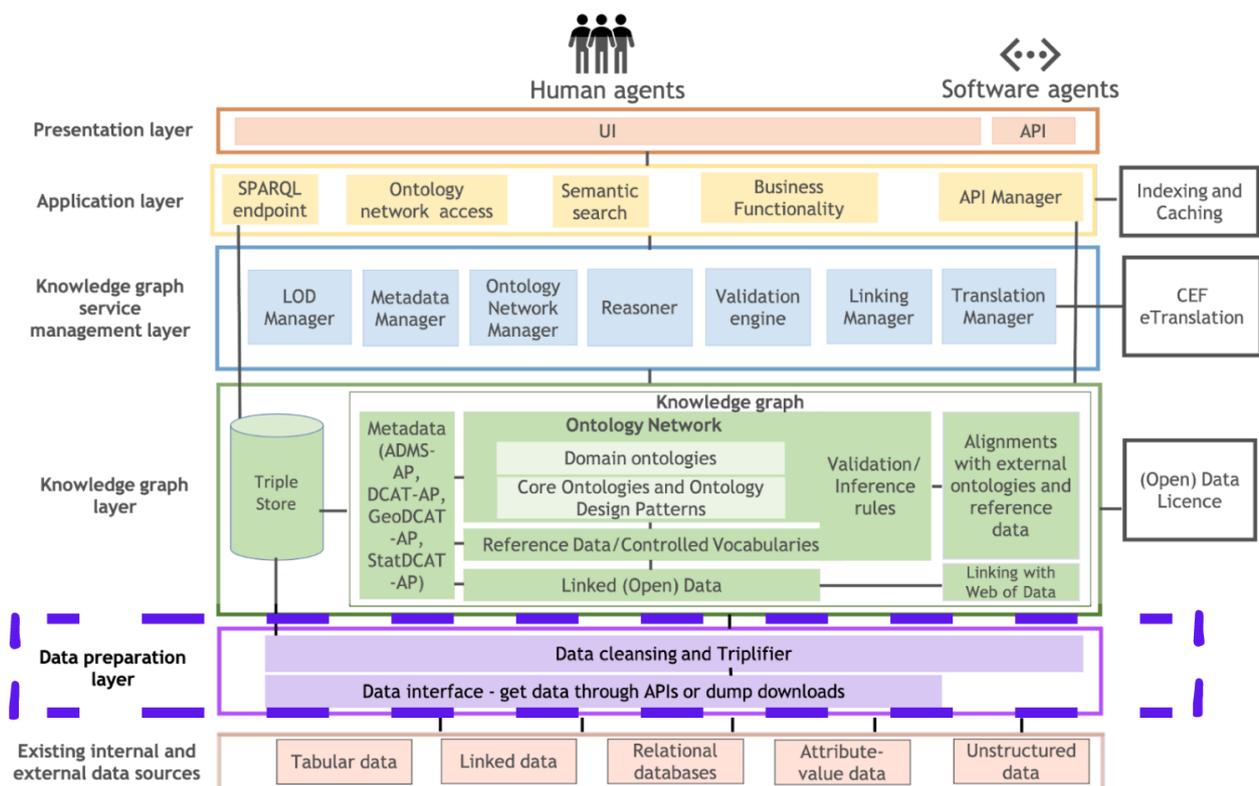


Figure 7: WHOW high level linked open data reference architecture. In the architecture the data preparation layer is highlighted by the box with the violet dashed border.

Specifically, there are two components of this layer dedicated to data pre-processing operations; namely, the data cleanser and the triplifier itself. Figure 8 below illustrates in an UML component diagram all the components of the data preparation layer and the interaction between the data cleanser and the transformer component, in turn instantiated by the triplifier. We remind here that the data cleanser provides services for cleansing data. Those services are focused on detecting and correcting/removing corrupt, inaccurate, incoherent facts from data. The DataCleanser provides access to the data cleansing

services to other components through the IDataCleansing interface. In the layer the transformer is the only other component that interacts with the DataCleanser. The Transformer is meant as the component of the layer aimed at performing the generation of RDF data from original sources. It is instantiated through the triplifier that performs a physical transformation of the original open data to RDF. For an extensive description of the overall layer and all the interactions between the various components shown in Figure 8, the interested readers can refer to the deliverable 4.1 [6].

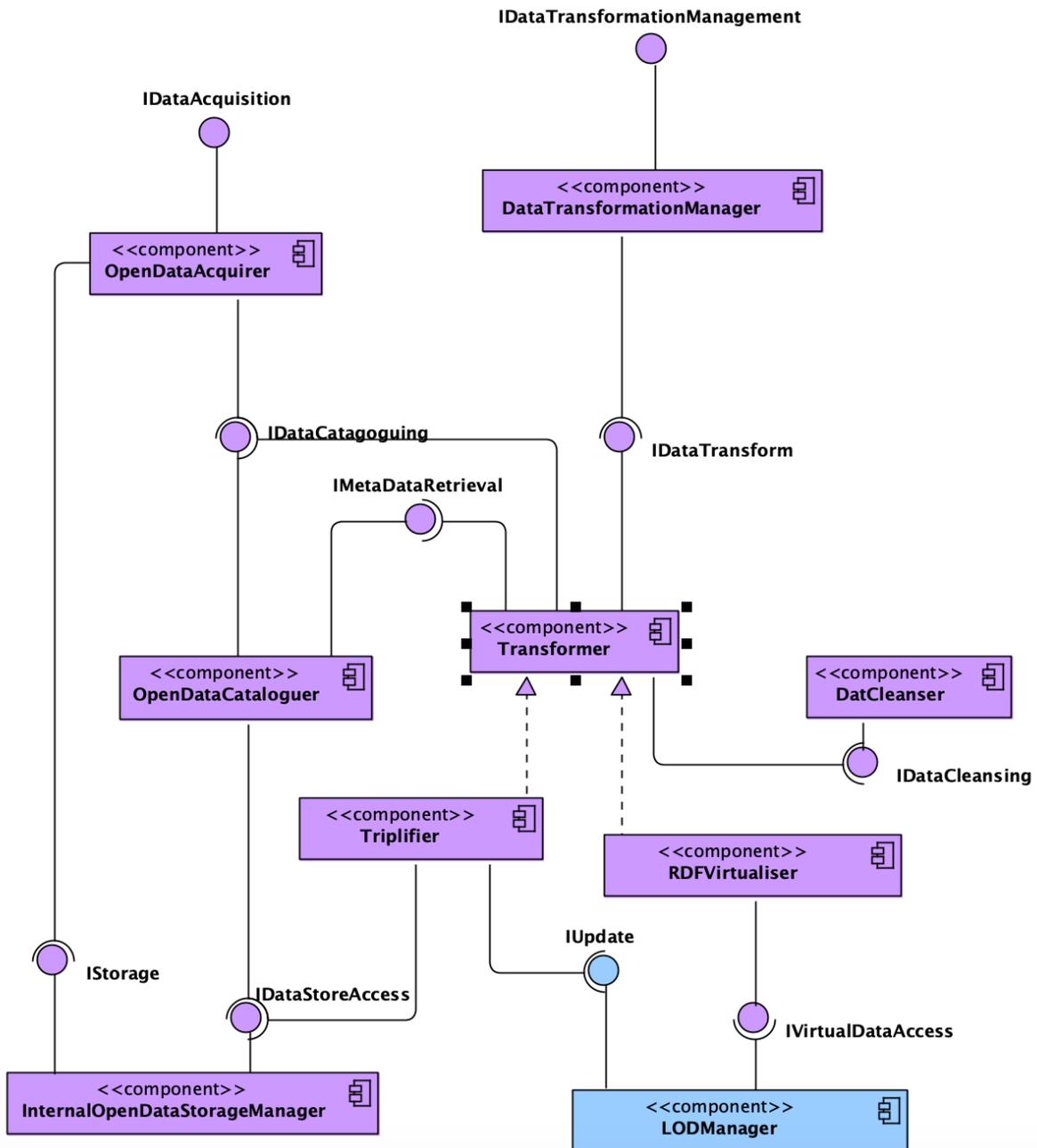


Figure 8: UML Diagram of the data preparation layer design

As described in the following subsections, where specific types of pre-processing tasks are explained with their current implementation, the operations are mainly executed by the data cleanser but also during the triplification process, which is done by exploiting ETL functions (e.g., string replace, normalised dates, etc.) offered by the RML mapping language we have chosen.

3.1 WHOW data providers' datasets: pre-processing needs

The need for a pre-processing carried out by the data cleanser and triplifiers components of our architecture also aims to further improve the quality of the data on the basis of the quality principles described in deliverable 5.1 on SDGs and KPIs. WHOW processes in fact mandate the provision of high quality datasets with respect to metrics such as syntactic and semantic accuracy, completeness, but also persistent and neutral URIs for the elements of the WHOW knowledge graph (i.e., descriptions of entities and their relationships and facts related to those entities). This latter aspect is dealt with within the WHOW process following the rules defined in [1] and using the w3id.org service, also recommended by the Italian national guidelines for public sector information valorisation [3].

We have identified so far six types of pre-processing activities that are part of the WHOW linked open data process and must be carried out on the WHOW datasets of all three use cases of deliverable 2.1:

1. Formats handling
2. Encoding handling;
3. Strings handling;
4. Dates handling;
5. Handling of incoherent data with respect to the expected content;
6. Handling of data that conveys more than one semantic concept at a time with respect to the defined data model (or ontology).

Note that pre-processing operations usually depend on the data, its format and structure. The operations are therefore likely to be rather customised for specific datasets. However, there are other types of pre-processing operations that are standard and can be applied to a variety of datasets from different application domains..

The following subsections examine in further detail such pre-processing needs, along with their resolution. The datasets used as examples present different data on surface water bodies, and have been named for the sake of simplicity as Dataset 1 (Analytical data on lake water bodies),⁷ Dataset 2 (Height of the lakes),⁸ Dataset 3 (PFAS data in surface waters)⁹, Dataset 4 (Analytical data on river water bodies)¹⁰ and Dataset 5 (Ostreopsis Ovata concentration)¹¹.

⁷ Analytical data on lake water bodies, Arpa Lombardia (2019), <https://www.arpalombardia.it/Pages/Dati/2019/Acque/Dati-analitici-corpi-idrici-lacustri-2019.aspx?tipodati=1&tema=Acque&sottotema=Sottotema%20Ambientale&ordine=1>.

⁸ Height of the lakes, Arpa Lombardia (2018), <https://www.arpalombardia.it/Pages/Dati/2018/Idrometeorologia/Altezza-laghi-2018.aspx?tipodati=1&tema=Idrometeorologia&sottotema=Sottotema%20Ambientale&ordine=1>.

⁹ PFAS data (perfluoroalkyl substances) surface waters, Arpa Lombardia (2018), https://www.arpalombardia.it/sites/DocumentCenter/Documents/PFAS/Allegato_1_DATI_PFAS_2018_ACQUE_SUP.xlsx.

3.1.1 Formats handling

Dataset ingested in the WHOW linked open data process usually come in different formats (e.g., XLS, CSV, JSON, etc.). In particular, some datasets are available only in proprietary formats (e.g., XLS, XLSX). In this case, we have carried out a pre-processing operation in order to transform the format in an open tabular format like CSV (Comma Separated Value) to be successively elaborated by well-consolidated mapping languages like RML we decided to adopt according to what is described in deliverable 3.2.

3.1.2 Encoding handling

Another problem common to many open datasets, especially tabular ones, is the character encoding used. Usually, many open datasets make use of different encodings such as ASCII, Windows, etc.; however, this may cause issues when dealing with specific characters. For example, in Italian, some words contain accented letters, e.g. names of cities. The use of certain encoding schemes means that these words are not correctly represented in the file. In this respect, also following the recommendations of the national guidelines on the valorisation of public sector information [3], we have carried out pre-processing operations to transform all the encodings of the input files into UTF-8, which solves the above-mentioned problems.

3.1.3 Strings handling

Often different datasets share common elements. When we first started the data processing work in WHOW, strings were the only fields of the datasets to be used in order to uniquely identify some elements of the water domain such as rivers and lakes. However, these strings are not uniformly used across datasets for the same entity. An example can be the case of the datasets for surface water. Taking Dataset 1, 2 and 3 as references, we can notice that the same lake is written in three different ways. Table 8 shows this issue for ISEO lake:

Table 8: Same entity written in three different ways in three datasets

Dataset 1	Dataset 2 (column name)	Dataset 3
ISEO	LAGO D'ISEO	LAGO D'ISEO (SEBINO)

The need therefore arises to standardise these strings. In fact, this heterogeneity is not only an issue when using strings to generate codes for persistent URIs, since they may produce three different URIs for the same real object, but it is also a problem when presenting the name of the same entity in the knowledge graph. In this latter case, if we took the three strings of Table 1 to generate the name of Iseo lake we would have three different names associated with the same entity.

¹⁰ Analytical data on river water bodies, Arpa Lombardia (2019), <https://www.arpalombardia.it/Pages/Dati/2019/Acque/Dati-analitici-corpi-idrici-fluviali-2019.aspx?tipodati=1&tema=Acque&sottotema=Sottotema%20Ambientale&ordine=1>.

¹¹ *Ostreopsis Ovata* concentration, ISPRA Ambiente (2020), https://annuario.isprambiente.it/sys_ind/847.

Therefore manipulations of these strings were performed according to these rules, for instance:

- Lowercase all the letters of the strings with the exception of the first one;
- In addition to the preferred name of the object, add the object type (e.g., “Lago” - lake);
- Eliminate any other alternative name included among brackets.

The result of this manipulation is reported in Table 3.

Table 9: Result obtained by harmonising the strings for naming production

Dataset 1	Dataset 2 (column name)	Dataset 3
Lago d’iseo	Lago d’iseo	Lago d’iseo

It is worth highlighting that, during our work in WHOW, we immediately reported to data owners these problems, especially those related to the lack of reference codes for water bodies. In general we argue that this practice undermines interoperability. In the case of Lombardy Region datasets, the data owner is ARPA (Regional Environmental Protection Agency) contacted through ARIA SpA. They understood the issues and provided us with their internal codes they are going to add in the original datasets. This makes it possible to avoid string manipulation to generate URIs for some entity: the provided codes can be instead used. At the same time, it was nevertheless necessary to manipulate the strings in order to better harmonise the text used in the knowledge graph, in particular for the same entities described differently across datasets. This has been done following the rules as earlier reported so that the names for the same entity are the same across the RDF datasets.

3.1.4 Dates handling

According to the standard ISO 8601¹², a field of type date should be expressed in the form YYYY-MM-DD. This format date is also widely adopted in the context of XML Schema Description (xsd), in turn used in OWL and RDFs standards in order to represent primitive types such as dates, integers, decimals, etc. However, not every dataset that includes dates meets this requirement. Fig. 5 provides an example of this problem we encountered for Dataset 5.

Data
06-08-20
06-09-20
07-06-20
07-07-20

Figure 9: Example of date format in the “Data” column of Dataset 5

¹² <https://www.w3.org/TR/NOTE-datetime>.

In this case, a pre-processing activity that normalises the date according to the above-mentioned standard has been carried out, leading to the result shown in Table 10.

Table 10: Example of “Data” column of Dataset 5 before and after pre-processing of dates

Dataset 5 “Data” column before pre-processing	Dataset 5 “Data” column after pre-processing
06-08-20	2020-08-06
06-09-20	2020-09-06
07-06-20	2020-06-07
07-07-20	2020-07-07

3.1.5 Handling of incoherent data with respect to the expected content

Most input datasets are available in tabular format. Some of their columns indicate contents that sometimes do not correspond to the instance values included in the dataset. For example, this is the case for the Chemical Abstract Service (hereafter CAS) column for Datasets 1 and 4. Most of its content relates to CAS codes. However, Water Information System For Europe (WISE) codes are also present along with others that do not conform to either CAS or WISE specifications. Fig. 10 provides a real example of this problem.

CAS
479-61-8
WISE 3133-05-9
71-52-3
72-54-8-SUM

Figure 10: Example of incoherent data in the “CAS” column in Dataset 1

To deal with this, we defined a pre-processing operation that allows us to create three different datasets that can be successively used in an easy manner for RDF transformation purposes. One dataset includes, among the other columns, only the CAS one, and CAS codes as instance values of that column; a second

dataset that includes, among the other columns, only the WISE one and WISE codes as instance values of that column; and a final dataset that includes only alphanumeric codes referring to chemical substances but not referring to any known standard (for instance, the values BIO-ESC-COL, or 72-54-8-SUM).

To differentiate CAS codes from alphanumeric ones as previously mentioned, an official CAS registry numbers dataset¹³ has been used against the values available in the WHOW datasets. In this way, it was possible to distinguish substances that could have a Unique Identifier in the CAS system and those that did not.

The three brand-new generated datasets can be found on the WHOW Github repository.¹⁴

In general, this solution allows us to create the correct instances of the entity `ChemicalSubstance` of the Water Monitoring Ontology we have developed so far.

The result of this manipulation is reported in Table 11.

Table 11: Comparative result obtained by distinguishing CAS, WISE and other alphanumeric codes

CAS dataset ¹⁵	WISE dataset ¹⁶	Unmatched dataset ¹⁷
479-61-8	3133-05-9	72-54-8-SUM

Please note that, also this issue has been reported to data owners who recognised the confusing semantics conveyed with a column named only CAS, as in Fig. 6. In the upcoming months we will evaluate whether to deprecate this pre-processing operation on the basis of the actions that possibly data owners decide to take for clarifying the semantics of their original datasets.

3.1.6 Handling of data that conveys more than one semantic concept at a time

In some other cases, data conveys more than a semantic concept within the same column. For instance, in Datasets 1, 3 and 4, columns identifying values also include the 'limit of quantification' for them, as illustrated in Figure 11:

¹³ Retrieved from <https://www.epa.gov/tsca-inventory/how-access-tsca-inventory#download>.

¹⁴ <https://github.com/whow-project/ontologies/tree/main/controlled-vocabularies/chemical-substances/datasets>.

¹⁵ <https://github.com/whow-project/ontologies/blob/main/controlled-vocabularies/chemical-substances/datasets/chemical-substances-lakes-rivers-dataset.csv>.

¹⁶ <https://github.com/whow-project/ontologies/blob/main/controlled-vocabularies/observable-properties/WISE/wise-codes-dataset.csv>.

¹⁷ <https://github.com/whow-project/ontologies/blob/main/controlled-vocabularies/chemical-substances/datasets/unmatched-chemical-substances-dataset.csv>.

VALORE
3.9
<0,5
2.3
<0,5
1.3
<0,5
9.5
2
21.7

Figure 11: Example of multiple semantic concepts in the “Valore” column of Dataset 1

In light of this situation, in the still unstable Water Monitoring Ontology, we envisage a concept of observed value of the water quality observation as well as properties of the observed value that model the limits of quantification if available and represented in the original datasets by the mathematical operators < or > (Figure 11).

Even in this case, a pre-processing operation is required to deal with these semantic concepts of the ontology. The pre-processing transforms the mathematical operator ‘<’ into the string ‘lt-’ and the operator ‘>’ into the string ‘gt-’. (cf. Figure 12).

VALORE
3.9
lt-0.5
2.3
lt-0.5
1.3
lt-0.5
9.5
2.0
21.7

Figure 12: Result of the pre-processing operation of the “Valore” column in Dataset 1

This is necessary for creating persistent URIs of the observed values. In fact, the URI is obtained by concatenating the transformed limit of quantification, the value and the measurement unit.

Further manipulations have been also done to urlify the measurement unit column of the dataset, in particular we replace slashes and white spaces with dashes. The resulting URI for an observed value of a water quality observation is then the following: <https://w3id.org/environmental-data/data/value/lt-0.5-mg-l>.

Finally, the pre-processing of Figure 12 also aims at populating the respective ontology elements, as previously introduced, during the mapping process (see Section 3.2.2) so as to instantiate the properties of maximum and minimum lower bound with respect to a certain value.

3.2 Pre-processing implementation

As shown in Figure 13, the pre-processing implementation is based on a workflow that consists of two phases: (i) a phase implemented through Python scripts that we have coded for the specific cases encountered in the analysed datasets; (ii) a second phase implemented during the RDF transformation through the RML mapping language that includes ETL functions used for this purpose.

Python scripts take as input the various original datasets in tabular format producing pre-processed CSV files that are then fed into the triplifier component (consisting of an RML processing engine and a set of RML mapping scripts) where further manipulations of the data are performed to transform it into RDF.

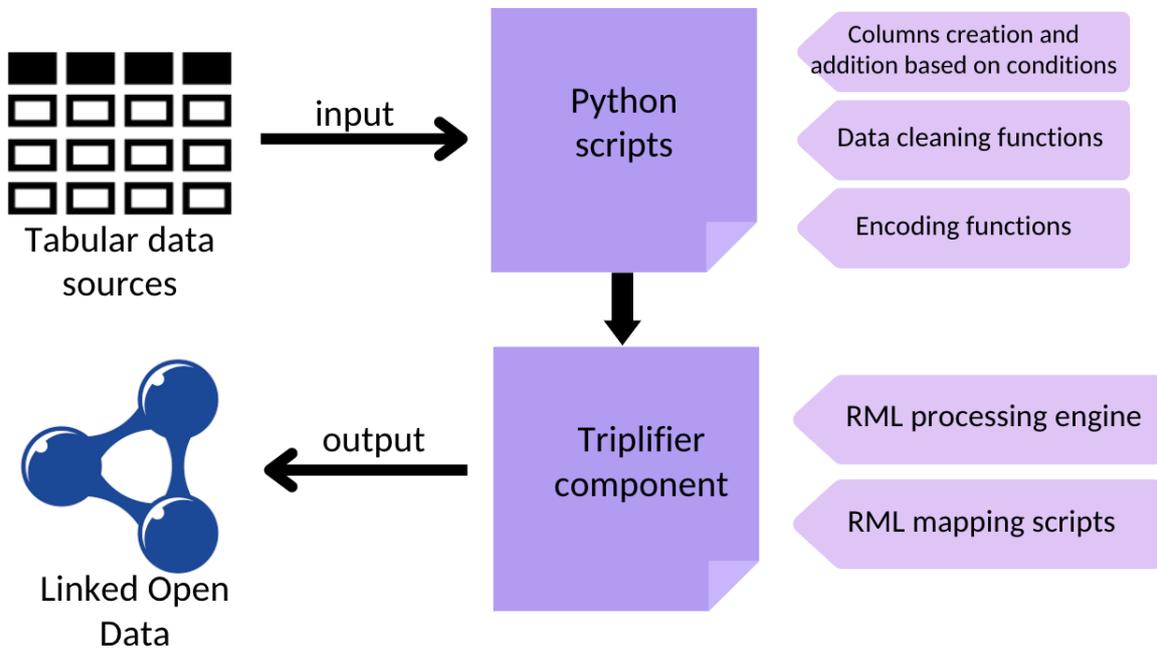


Figure 13: Scheme of pre-processing process

3.2.1 Pre-processing Python scripts

Python scripts have been implemented and considered useful to generate tidier datasets with respect to the original ones. This included functions for the pre-processing of input formats¹⁸, encoding functions¹⁹ and other data cleansing functions, such as:

- converting decimal commas into points, as per ISO 80000-1;²⁰
- converting coordinates from strings to integers.

Moreover, ad-hoc python scripts, strictly dependent on the input datasets, have been implemented for string handling purposes. For instance, in the case of Dataset 1 (Analytical data on lake water bodies), this included:

- Deletion of “CASID” from the “CAS” column;
- Identification of “WISE” in the “CAS” column and creation of a separate dataset for WISE codes;
- Deletion of “WISE” substring;
- Identification of alphanumeric codes by matching the CAS ones using an official CAS registry dataset.

3.2.2 RML Mapping

Based on the analysis reported in deliverable 3.2, RML is the well-known and widely used mapping script language that has been chosen as preferred means for RDF transformation. As earlier described, additional data manipulations have been carried out by exploiting the ETL functionalities offered by RML through its functions framework.

Using this framework is in fact possible to manipulate data during the triplification process.

Specifically, some strings handling, handling of incoherent data with respect to the expected content and handling of data including more than a semantic element have been managed through RML using a set of functions. The RMLMapper engine has been used to read and interpret those mapping rules and functions. RMLMapper is an open source Java library which executes RML rules to generate Linked Data.²¹ and it is the reference implementation for RML-based mapping tools.

RMLmapper relies on declarative rules to define how knowledge graphs are generated, and on ETL functions for specific data manipulation requirements. For instance, in the case of string handling, examples of functions include: the transformation of strings into upper or lower case, the replacement of specific characters, the joining of strings (via `array_join` function) and so on. For example, functions that we have used in some cases are::

- To Lowercase,²²

¹⁸ See the conversion from Excel to CSV: https://github.com/whow-project/architecture/blob/main/ispra-lod-infrastructure/soilc_excel_to_csv.py.

¹⁹ See the script `utf8_converter`: https://github.com/whow-project/architecture/blob/main/ispra-lod-infrastructure/utf8_converter.py.

²⁰ http://store.uni.com/catalogo/uni-cei-en-iso-80000-1-2013?__store=en&josso_back_to=http%3A%2F%2Fstore.uni.com%2Fjosso-security-check.php&josso_cmd=login_optional&josso_partnerapp_host=store.uni.com&__from_store=it.

²¹ See <https://github.com/RMLio/rmlmapper-java>.

²² <https://rml.io/docs/rmlmapper/default-functions/#tolowercase>.

-
- String replace²³.

In addition, in those common cases of manipulation of dates the functions we used are:

- Normalise date²⁴;
- Normalise DateTime²⁵.

Both require identifying the input date format and they transform that format according to the reference XSD standard for the date (YYYY-MM-DD) or the dateTime (YYYY-MM-DDThh:mm:ss), respectively.

A more complex manipulation, that involves both the python scripts and RML functions, is the case in which some CSV columns in input convey more than one semantic concept at a time. In particular, this applies for the observed values that include an indication of the quantification limit. To deal with this, all the following steps and RML functions have been applied:

- In the re-processed CSV file produced by the python scripts, identify the brand-new substrings we introduce to indicate mathematical operators. The substrings are lt- (lower than) or gt- (greater than) and the RML function used to locate the substring is "Contains"²⁶;
- Take only the substring that starts with "-" followed by a value. This substring is used to map the `ObservationValue` class of the water monitoring ontology we have defined so far;
- If lt- is present (use of `trueCondition` function of RML) the datatype property `hasMaxValue` will be populated, otherwise the datatype property `hasMinValue` will be populated;
- If the dataset includes mathematical operators like `<=` or `>=`, the properties of the ontology (`isMaxIncluded` and `isMinIncluded`) will be mapped.

²³ <https://rml.io/docs/rmlmapper/default-functions/#replace>.

²⁴ <https://rml.io/docs/rmlmapper/default-functions/#normalizedate>.

²⁵ <https://rml.io/docs/rmlmapper/default-functions/#normalizedatetime>.

²⁶ <https://rml.io/docs/rmlmapper/default-functions/#contains>.

4 Conclusions

In this deliverable, we described the data pre-processing operations that are applied to the datasets considered in the context of the WHOW project. In particular, the deliverable first introduces the current state of the art of such operations performed in the data management infrastructures of WHOW data providers. Secondly, it focuses on the data pre-processing that is nonetheless necessary within the specific linked open data creation processes that we envisage in the project and that we describe in Deliverable 3.2.

In general, these operations are core activities in data management, especially when datasets are formed on the basis of content provided by heterogeneous data sources. Probably more than 80% of the efforts in data management are devoted to the data cleaning and preparation phases. Both Lombardy Region and ISPRA act as data aggregators with respect to other local data owners and this inevitably leads to putting in place data validation checks before publication for anyone.

Despite these efforts, whose goals we recall are to improve the quality of ISPRA and ARIA datasets overall and to ensure that data protection requirements are met, further pre-processing tasks need to be performed in the linked open data specific processes of WHOW. While the pre-processing mentioned above is mainly focused on data quality assurance, the latter in WHOW is also instrumental in preparing the data for transformation into a knowledge graph, where persistent URIs must be created and a semantic model enforced. The categories of interventions on the data described in this deliverable are quite common and can be necessary for a variety of open datasets.

As for the external datasets coming from sources like EIONet or Copernicus, at the time of this writing we have not observed specific manipulations different from those already described in this deliverable. In this respect, one final remark is worth highlighting. In this deliverable we have presented the main data pre-processing tasks that are performed by data providers and in the WHOW project. As additional datasets may be processed and added during the lifetime of the project, it may be possible that other types of data pre-processing tasks than those described will be performed. In this case, we are planning to report them in future deliverables.

References

1. 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>.
2. Data Catalog Vocabulary (DCAT) - Version 2, W3C Recommendation (2020), <https://www.w3.org/TR/vocab-dcat-2/>.
3. 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>.
4. Requisiti per la pubblicazione di dati di livello 4 e 5 (2017), AGID + Team Digitale, <https://docs.italia.it/italia/daf/ig-patrimonio-pubblico/it/stabile/publdatigov.html#requisiti-per-la-pubblicazione-di-dati-di-livello-4-e-5>.
5. WHOW project, “Milestone #9: SDGs and KPIs are defined” - April 2022.
6. WHOW project. “Design of the technical services for knowledge graph management” https://github.com/whow-project/deliverables/blob/main/ArchitectureDeliverable_final.pdf