D4.1 – FAIR Principles Guidelines

WP4 FAIRness assessment in the water industry

Author/s: Barry Evans (UNEXE), Fernando López Aguilar (FIWARE), Benoit Orihuela (EGM)

Date: 31/05/2023

Co-funded by the
European Union

GRANT AGREEMENT NUMBER	101070262		
FULL TITLE / ACRONYM	Water Data Management Ecosystem for Water Data Spaces / WATERVERSE		
START DATE	1 October 2022	DURATION	36 months
END DATE	30 September 2025		
PROJECT URL	https://waterverse.eu/		
DELIVERABLE	D4.1 FAIR Principles Guidelines		
WORK PACKAGE	WP4		
CONTRACTUAL DATE OF DELIVERY	31/05/2023		
ACTUAL DATE OF DELIVERY	31/05/2023		
TYPE	Report	DISSEMINATION LEVEL	PU
LEAD BENEFICIARY	UNEXE		
RESPONSIBLE AUTHOR	Barry Evans (UNEXE), Fernando López Aguilar (FIWARE), Benoit Orihuela (EGM)		
CONTRIBUTIONS FROM	Matteo Basile (ENG), Sergio Baena Miret (CET), Gerasimos Antzoulatos (CERTH), Gael Poujol (EGM), Aitor Corchero (EUT), Siddharth Seshan (KWR), Eloy Hernández Busto (EUT)		
ABSTRACT	Deliverable 4.1. outlines the development of the WATERVERSE FAIR Guidelines to be utilised within the water sector. The development of these guidelines will be based upon the existing principles of FAIR, integrating aspects of MELODA5 and related context for use within the WATERVERSE.		

Disclaimer

Copyright message

REVISION HISTORY

Version	Date	Who	Description
V1	19/05/2023	B. Evans (UNEXE) F. Aguilar (FIWARE) B. Orihuela (EGM)	Version 1 submitted for internal review
V2	31/05/2023	B. Evans (UNEXE) F. Aguilar (FIWARE) B. Orihuela (EGM)	Version 2 submitted with changes suggested from internal review

QUALITY CONTROL

Role	Date	Who	Approved/Comment
Reviewer	27-05-23	Gerasimos Antzoulatos (CERTH)	Comments added. Requires minor revisions
Reviewer	28-05-2023	Lydia Vamvakeridou-Lyroudia (KWR)	Comments added. Requires minor revisions

EXECUTIVE SUMMARY

The WATERVERSE mission is to develop a Water Data Management Ecosystem (WDME); that is programming languages, packages, algorithms, cloud-computing services, and general infrastructure to collect, store, analyse, and leverage water data. More specifically, it will make data management practices and resources in the water sector accessible, affordable, secure, FAIR and easy to use, improving usability of data and the interoperability of data-intensive processes, thus lowering the entry barrier to data spaces, enhancing the resilience of water utilities, and boosting the perceived value of data, and therefore the market opportunities behind it.

One of the key outputs within the WATERVERSE is the development of WATERVERSE FAIR Guidelines to facilitate the sharing and dissemination of data/outputs that are specifically tailored to the water sector.

This deliverable focuses on establishing criteria for achieving FAIR Digital Objects and implementing the necessary components to create a FAIR Ecosystem whilst additionally recognising the importance of going beyond these principles to ensure optimal conditions for digital asset reuse. To address this, the FAIR principles are expanded upon to integrate requirements from the MELODA5 metrics, which encompasses factors related to data publishers and content.

In line with the recommendations of the second European Open Science Cloud High Level Expert Group (EOSC HLEG) report, this deliverable will define a Minimum Viable Ecosystem that adheres to the principles of a research ecosystem that will enable the development of efficient and effective services to meet the requirements of FAIR and MELODA5 principles.

In summary, the aim within this deliverable is to establish clear guidelines, metrics, and an ecosystem that promotes the adoption and measurement of FAIRness within the WATERVERSE context, for the benefit of the water sector.

Related Deliverable: Deliverable 4.2 (M8, M18, and M36)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

API	Application Programming Interface
CA	Consortium Agreement
CSO	Combined Sewer Overflow
CSV	Comma-Separated Values
DCAT	Data Catalogue
ETSI	European Telecommunications Standards Institute
EU	European Union
FAIR	Findable Accessible Interoperable Reusable
GDPR	General Data Protection Regulation
IP	Internet Protocol
IPR	Intellectual Property Rights
JSON	JavaScript Object Notation
LDP	Linked Data Platform
MAC	Media Access Control
MELODA	Metric for the evaLuation of Open Data
NGSI	Next Generation Service Interfaces
PUC	Pilot Use Case
RDF	Resource Description Framework
RSS	Really Simple Syndication
SCADA	Supervisory Control And Data Acquisition
SDM	Smart Data Model
SQL	Structured Query Language
URL	Uniform Resource Locator
URN	Uniform Resource Name
WAN	Wide Area Network
WDME	WATERVERSE Data Management Ecosystem
WP	Work Package

1.0 Introduction

The WATERVERSE project aims to develop a Water Data Management Ecosystem (WDME) for making data management within the water sector more accessible, affordable, secure, fair, and easy to use. To facilitate this the WATERVERSE is adopting FAIR principal guidelines first defined in Wilkinson et al., (2016) [1] and dimensions of MELODA5 outlined in Abella et al. (2020) [2]. The project is divided up into six work packages (WPs) as shown in Figure 1, where WP4 "FAIRness assessment in the water industry" forms an integral part, interconnected with the other WPs.

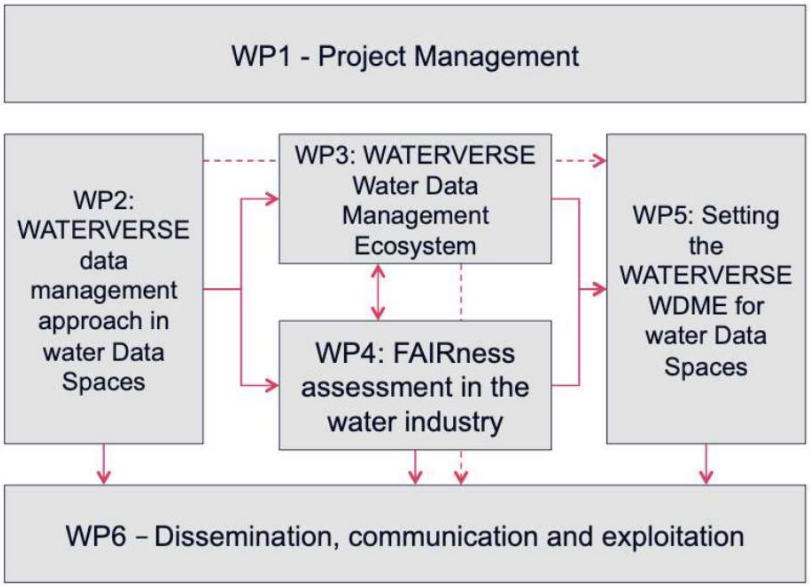

Figure 1: WATERVERSE Work Package Structure

There are six pilot case studies participating within the WATERVERSE project each with their own defined use cases and subsequent data management requirements:

- Netherlands: Prediction of water quality and its impact in the treatment steps
- Germany: Digital Village Twin for flood protection and territorial management
- Cyprus: Leveraging the potential of water digital twin and water analytics tools
- United Kingdom: Addressing the challenges of Combined Sewer Overflow (CSO) performance
- Spain: Management of the integral water cycle and open innovation
- Finland: Smart water tools and risk management

As part of the FAIRness assessment WP4 will assess the FAIRness of (meta)data flows, tools, services and infrastructures that are digitally represented within their workflows.

As the initial output of Work Package 4, Deliverable 4.1 outlines the WATERVERSE FAIR guidelines, which forms a key component of the WATERVERSE Data Management Ecosystem (WDME) laid out in WP3. The WATERVERSE FAIR Guidelines are a combination of the FAIR principles and MELODA5. These guidelines are to be implemented across each of the six pilot case studies as a means of ensuring good data management within the project and to demonstrate as an exemplar as how following said guidelines can enhance the utilisation, comprehensibility, and reusability of data within the water industry and beyond.

The Deliverable is structured in the following way:

- Outline of challenges associated with adoption of FAIR guidelines within a workflow.
- Definitions of FAIR and MELODA5 principles and they are applied.
- How FAIR and MELODA5 principles are to be adopted within the WATERVERSE.
- Details about the tools and resources that will be utilized.
- The European Added Value of research being undertaken.
- Summary with final remarks and conclusion about the document.

2.0 Main Challenges for the implementation of FAIR principles

Implementing FAIR guidelines within an industry sector can be a complex and challenging process, where issues including but not limited to available resources, ethics, and data security need to be considered. This section outlines these challenges and how they're envisioned to be tackled within the scope of the WATERVERSE project.

2.1 Allocation of Resources

The adoption of FAIR guidelines within an organisation may be resource-intensive and time-consuming requiring long-term commitment, including potential investment in resources and personnel. However, the benefits of improved data/information dissemination may indirectly outweigh these challenges through making the data/information being more easily disseminated throughout the organisation and to third parties. To facilitate the adoption/integration of WATERVESE FAIR guidelines into the pilot case studies workflow, each pilot case study has an external technical partner/facilitator that will be acting as a mediator between the case study and the technology/framework provider FIWARE/EGM. In terms of data model creation that is under scope of each of the use cases FIWARE will provide support with the creation of metadata and FAIR data models.

2.2 Data Security

Within industry, especially sectors relating to the provision of critical services, the flow of data and information is restricted and controlled under licence due to commercial, confidentiality or/and security reasons. Due to this, methods to ensure data is secure will be implemented within the framework that will allow for controlled access to data where required.

2.3 Ethical Aspects (EUT)

This section is mainly devoted to the description of the Ethical Aspects adhered to the WATERVERSE project. Under this section, WATERVERSE will consider two main aspects: (i) the Intellectual Property Rights (IPR) and (ii) the assessment and protection of personal data and research activities. This last aspect is in relation to the existent regulations in data privacy as specified in the GDPR and compatible regulations.

2.3.1 General Information & Assessments

In the WATERVERSE project, there are important aspects to be considered in the collaboration with stakeholders and potential users. In this regard, there potentially could exist ethical issues related to ensuring informed consents for participating in stakeholder engagement sessions, anonymity, and confidentiality of information. All of these aspects are related with the rules associated with the voluntary involvement of human participants in the EU. Based on this, data collected in WATERVERSE are user interviews, opinions, reviews and policy-based questions that will be associated with different digital components of the project and also, with the demo-cases. Thus, a non-exhaustive list is provided as follows:

- Close involvement of operators, managers and regional representatives that help in the identification of efficient practices, data and needs/requirements that serves to demonstrate the Open Water Data space envisioned (digital platform).

- The visualisation and human interfaces to collect information from users so that the WATERVERSE WDME and subsequent modules could support and help in the operational and planning decision-making. This will help to put in practice efficient operational and planning practices in the demo-cases.
- A series of interviews/workshops/communities of practices with key stakeholders and decision-makers of the case-studies.
- Planned contacts with representative stakeholders that represent targeted end-users. Interviews should be carried out in two-way mode: online meetings (e.g. Teams, Zoom, etc) or face-to-face. The selection of the mode will be convenient depending on the audience and the periodicity of the meeting. Commonly, interviews should help to define the expected functionalities, identify the digital platform (open data space) needs, requirements, and services to be offered. Also, it will cover validation and demonstration of the platform in the effectiveness of open data spaces inside the water industry. Specifically, on interlinking water value chains across industries and administrations. These meetings will be also useful at business/explicable level in terms of replicability, transferability, and the identification of distribution channels to achieve specific targeted clients.
- The relevant stakeholders (end-users, potential developers, potential partners, etc) will have the opportunity to test and review the latest products and services offered by WATERVERSE (WDME).
- Methodology and procedures for sensitive data processing, manipulation and storing will be specified as a part of the ethics as well. It is important to emphasise special efforts that will be devoted to ensuring data security, integrity, and privacy inside the digital platform. Specifically, in terms of security and privacy modules offered by FIWARE to ensure these aspects. Also, and according to the directives, mechanisms to delete personal data will be provided in an easy and usable manner.

To strengthen further commitment of WATERVERSE partnership research and innovation, a shaped ethical practices and guidelines will ensure fair and equal power relationships between researchers and participants. Specifically, the consortium agrees on complying with the principles laid down in the European Code of Conduct for Research Integrity[1], published by the European Science Foundation. These principles mainly highlight:

- Honesty in the communication of the researchers' goals and intentions, in reporting methods and procedures and in conveying interpretations.
- Reliability in performing research.
- Objectivity, which requires facts capable of proof, and transparency in the handling of information.
- Impartiality and independence.
- Openness and accessibility.
- Duty of care-all researchers have a duty of care for humans, animals, biodiversity, the environment, or the objects that they study.
- Fairness in providing references and giving credit for the work of others.
- Responsibility for the scientists and researchers of the future.
- Care will be taken to minimise the potential collection of personal data (e.g., while taking photos or recording videos in both face-to-face and teleconference events).

[1] https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/european-code-of-conduct-for-research-integrity_horizon_en.pdf

Considering other aspects, WATERVERSE will not involve any potential vulnerable groups or people unable to consent (children, those with a learning disability or cognitive impairment, or individuals in a dependent or unequal relationship). Additionally, it will not involve sensitive topics which may induce psychological stress, anxiety or humiliation, deception or any potential increased danger to participants, or the collection of personal data from participants.

Further, WATERVERSE digital solutions, architectures, processes, and methodologies will not involve the collection, manipulation and processing of the following type of data:

- Research involving sensitive topics - for example participants' sexual behaviour, their illegal or political behaviour, their experience of violence, their abuse or exploitation, their mental health, or their gender or ethnic status.
- Research involving groups where permission of a gatekeeper is normally required for initial access to members - for example, ethnic or cultural groups, native peoples or indigenous communities.
- Research involving deception, or which is conducted without participants' full and informed consent at the time the study is carried out.
- Research involving access to records of personal or confidential information, including genetic or other biological information, concerning identifiable individuals.
- Research which would induce psychological stress, anxiety or humiliation or cause more than minimal pain.
- Research involving intrusive interventions - for example, the administration of drugs or other substances, vigorous physical exercise, or techniques such as hypnotherapy. Participants would not encounter such interventions, which may cause them to reveal information, which causes concern, in the course of their everyday life.
- Research involving the tracking or observation of participants (e.g. surveillance or localization data, and Wide Areas Network -WAN- data, such as IP address, MACs, etc.). However, 'cookies' could potentially be used on the website and the graphic user interfaces to understand and analyse how users behave while interacting with the WATERVERSE WDME;
- A privacy statement will be put on the website regarding the use of external services like Google Analytics (or similar) to track and get statistics from users in the use and interaction with the website. A similar privacy statement will be put on the graphical user interface with similar purposes. At this point, it is important to notice that none of the collected data by WATERVERSE project requires a notification or authorization for the collection and/or processing of the personal data to authorities or other responsible entities.

To ensure all WATERVERSE research and participatory approaches follow good ethical practices and ensure fair and equal power relationships between researchers and participants, the consortium will study to agree, sign, and make public an ethics agreement based on the European Code of Conduct for Research Integrity, published by the European Science Foundation.

Complementary, WATERVERSE consortium also agrees to follow the rules and guidelines of the GDPR EU[2] regulation adhered to the data privacy and security of personal information across their digital and non-digital developments[3]. Consequently, the following conditions enable to elaborate the Data Impact Assessment[4] to ensure correct protection of the users/stakeholders' information:

- Use and elaboration of newer technology.

[2] https://gdpr.eu
[3] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN
[4] https://gdpr.eu/data-protection-impact-assessment-template/

- Track of people's location and behaviour.
- Systematically monitoring a publicly accessible place on a large scale.
- Processing personal data related to "racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation".
- Data processing is used to make automated decisions about people that could have legal (or similarly significant) effects.
- Processing children's data.
- Processing could result in physical harm to the data subjects if it is leaked.

Considering the GDPR guidelines, specific consent should be practised:

- **Consent must be freely given**. No data subject will be cornered into agreeing upon the usage of their data. Consent to data processing will not be a condition of using the data. The one exception is when some piece of data is needed for the data subject to provide them a data related service.
- **Consent must be specific**. The request for consent will be presented in a manner which is clearly distinguishable from the other matters. It will be clear what data processing activities are carried out, granting the subject an opportunity to consent to each activity.
- **Consent must be informed**. The users will be aware of the data processor's identity, the processing activities that will be conducted, the purpose of the data processing, and that they can withdraw their consent at any time. The latter will be described in plain language ("in an intelligible and easily accessible form, using clear and plain language"). That means no technical jargon or legalese. Anyone accessing the digital twin will be able to understand what they are asked to agree to.
- **Consent must be unambiguous**. There will be no question about whether the data subject has consented. Consent must be clear on any circumstances.
- **Consent can be revoked**. In the WATERVERSE digital tools, data from users will have the right to withdraw consent at any time. This process will be foreseen to be made easy for them to do so.

2.3.2 Intellectual Property Rights

Intellectual Property Rights (IPR) will receive special attention from the beginning. All rules regarding management of knowledge and IPR are governed by the CA. Thus, WATERVERSE will not act in contradiction with the rules laid down in Annex II of the Grant Agreement. The CA will address background and foreground knowledge, ownership, protected third party components of the products, and protection, use and dissemination of results and access rights. In this regard, the following principles are applied:

- **Confidentiality**: During the project duration and beyond, the contractors shall treat any information, which is designated as property by the disclosing contractors, as confidential. They also shall impose the same obligations to their employees and suppliers.
- **Pre-existing know-how**: Each Contractor is and remains the sole owner of its IPR over its pre-existing know-how. The Contractors will identify and list the pre-existing know-how over which they may grant access rights for the project. The Contractors agree that the access

rights to the pre-existing know-how needed for carrying out their own work under the project shall be granted on a royalty-free basis.

- **Ownership and protection of knowledge**: The ownership of the knowledge developed within the project will be governed by an open-source licence.
- **Open data**: Data and results obtained during the project that are based on open public-sector data will be made available free of charge.

The procedures for the dissemination, protection, and exploitation of intellectual property rights (IPR) are in the Consortium Agreement (Section 6: Governance Structure, Sub-section 6.2.4). The intention has been to balance the requirements necessary to protect such intellectual property and the foreseen dissemination objectives. IPR will be applied according to the rules of the employer under the applicable European and national laws and regulations.

2.4 Other issues

One of the initial identified challenges in the adoption and implementation of FAIR guidelines within a workflow is the misconception that FAIR data must be open data. Due to the sensitivity/confidentiality of certain datasets within industry and for security reasons not all data/information can be fully open access.

It is therefore important to clarify that FAIR data does not need to be Open Data, with data still able to be FAIR even if under restricted licence.

3.0 FAIR Principles and MELODA5 Dimensions in practice

The industry sector plays a significant role in generating and utilising vast quantities of data [3] with companies providing products and services now relying on the analysis of large amounts of data to develop new products and services, whilst maintaining day to day operations. As the use and availability of data continues to grow in both research and industry, the importance of adhering to good data standards to ensure data usability has become increasingly evident. In a study by Jahanddideh-Tehrani et al., (2021) [4] five challenges were identified within the water sector relating to data:

- Poor Quality of Water Data
- Lack of Integrated Water Portal Systems
- Limited Access to Data
- Big Data Problems
- Data Processing

It is envisioned that the implementation of good data and metadata standards can go towards addressing these issues through improving both the access and understandability of data both within the water sector and externally to other sectors. The application therefore of metrics that provide insight into the ease of which your data can be accessed and disseminated, along with guidelines as to how the reach of the data can be improved would thus be a valuable resource for industrial partners. The concept of FAIRness of data itself, originally born in an academic domain, was a result of the urgent need to improve the infrastructure supporting the reuse of scholarly data, to allow for the maximum benefits of research outputs, facilitating knowledge discovery and innovation [1]. To aid in improving the access and reusability of data Wilkinson et al., (2016) [1] proposed the FAIR Data Principles, based on the 4 key tenants that both data and metadata should be:

1. **Findable**: Data should be easy to locate and search for using appropriate keywords and metadata.
2. **Accessible**: Data should be available in a format that can be easily accessed and used by others.
3. **Interoperable**: Data should be structured in a way that allows it to be easily integrated with other data sources.
4. **Reusable**: Data should be available for reuse by others and should be accompanied by sufficient documentation and metadata to make this reuse possible.

Expanding upon each of these four tenants a number of criteria for each tenant needs to be satisfied (Figure 2) where data and metadata that follows these criteria are said to be “FAIR”.

Figure 2: FAIR Principles Guidelines (D7.3. FIWARE[5])

An additional aspect and step beyond the FAIR assessment of data sources within the WATERVERSE project are the inclusion of the dimensions outlined in MELODA5 (Figure 3) that are designed to assess the reusability of data sources. The acronym stands for "Metric for the evaLuation of Open Data". MELODA5 is a metric designed to evaluate the reusability of open datasets. The latest version of MELODA5, version 5.0 was released in 2019 where its primary objective is to assist open data publishers in providing sufficient information to locate and reuse data. The metric is based on eight dimensions, some of which pertain to the dataset's metadata, such as access mechanism, licence, and update frequency, while others relate to the data's structure and content, such as standardisation, geographical content, and technical format. Additionally, two dimensions (added in this latest iteration of MELODA5) are associated with the publisher, including reputation and its dissemination and communication activities. Each dimension is structured into several levels ranging from three to five, and a weight is assigned to each level. By combining the results of each level, a comprehensive evaluation is generated, which is referred to as the MELODA5 score. The eight dimensions of MELODA5 are summarised as follows:

- **Legal licensing**: The licence assigned to the dataset. It could be Private use, Non-commercial reuse or Commercial reuse or no restrictions.
- **Access to information**: Access to information or mechanisms by which information is released. it could be classified in Web access or unique URL parameters to dataset, Web Access unique with parameters to single data or API or query language.

[5] https://www.fiware4water.eu/sites/default/files/delivrables/F4W-D7.3-InitialDMP_final.pdf

- **Technical standard:** Technical standards or technical structure in which the data are available from the point of view of openness. It can be classified in Closed standard reusable and open non reusable, Open standard reusable and Open standard with individual metadata.
- **Data model standardisation**: Data Model standardisation from the standpoint of dissemination and adoption. It can be classified in Own data model, Own ad hoc data model standardisation published, Local standardisation and Global standardisation.
- **Geolocation content**: whether the content of the dataset includes fields with geographical information. It can be classified in No geographic information, Simple or complex text fields and Coordinates or full geographical information.
- **Updating frequency of data:** Frequency by which the data are refreshed (independently of changes of the values), it can be Longer than 1 month, Monthly, Updating period ranges from 1 month to 1 day, Daily, Updating period ranges from 1 day to 1 hour, Hourly. Updating period ranges from 1 hour to 1 minute. Seconds. Updating period is less than 1 minute.
- **Reputation:** Account of the prestige of the portal between re-users and other data publishers. It could be No information about the reputation of the data source, Statistics or reports published on users opinions and Indicators or rankings on reputation of the data source.
- **Dissemination:** If the publisher has dedicated campaigns to disseminate and promote reuse of the data assets. It could be Communication / dissemination not systematic, Available resources on updates and Proactive dissemination / push dissemination.

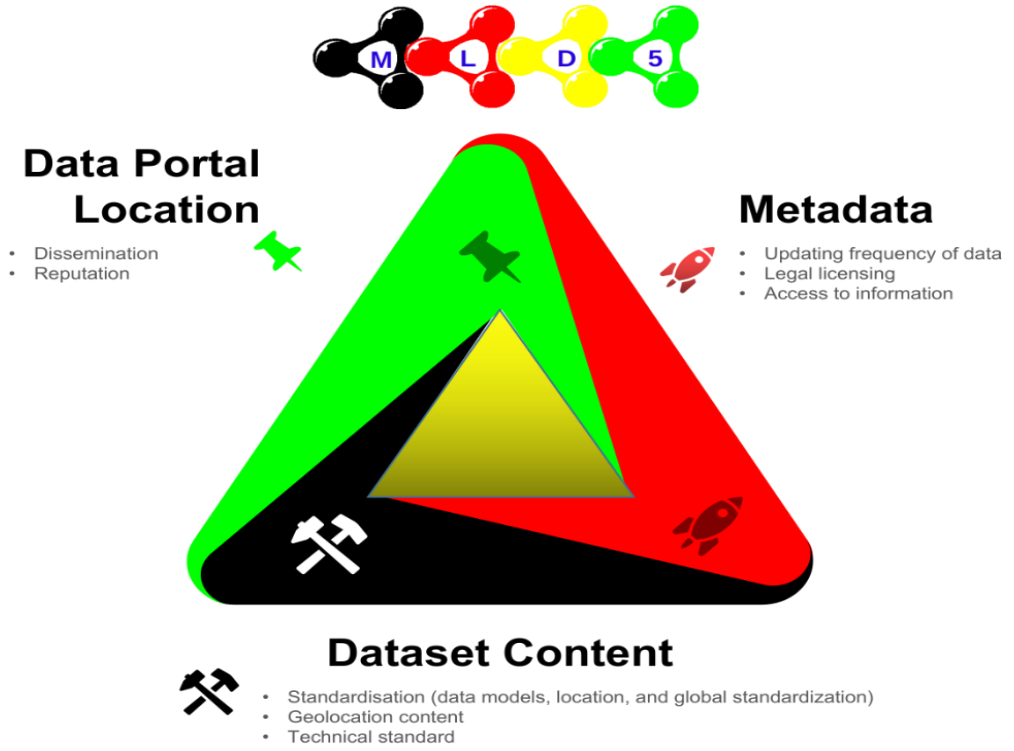

Figure 3: MELODA5 Dimensions

MELODA5 continues to promote the use of FAIR principles for open data and provides a practical framework for implementing these principles in practice. Four of the eight principles of MELODA5 overlap with the requirements of FAIR principles (Table 1) with the exceptions being geolocation, update frequency, dissemination, and reputation.

MELODA5 Dimesion	FAIR Principle ID	Description
Licence	R1.1	(meta)data are released with clear and accessible data usage licence.
Access to data	A1	(meta)data are retrievable by their identifier using a standardised communication protocol.
Technical format	I1	(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
Data model (standardisation)	I1, I2, I3	(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. (meta)data use vocabularies that follow FAIR principles. (meta)data include qualified references to other (meta)data.
Geolocation	-	metadata contains spatial information relating to location and coordinate reference system of the data.
Update frequency	-	metadata contains information relating to time interval between updates of the presented data.
Dissemination	-	metadata providing information as to means of which data is to be accessed/shared.
Reputation	-	Metadata providing insight into the credibility of the data source.

Table 1: MELODA5 Dimensions and their Synergies with FAIR Principles

3.1 Findable: Techniques for making data discoverable and accessible

To facilitate both the use and reuse of data, it is crucial to ensure its discoverability. Therefore, it is important that both data and metadata are easy to locate by both humans and computers respectively. The "Findable" aspect of data from the FAIR guidelines relates as to whether data that is produced and/or used within the project is discoverable with metadata, and locatable by a standardised file naming convention. Previous work within Fiware4Water (reference) adopted a naming convention to facilitate the findability and accessibility of given datasets that was divided up into 3 parts (Figure 4).

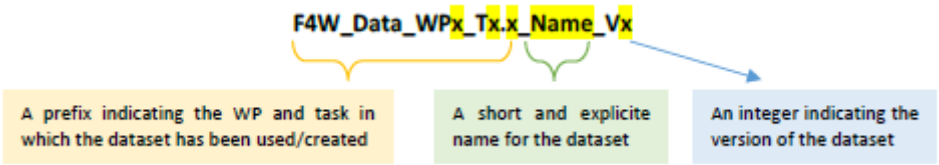

Figure 4: Dataset naming convention employed in FIWARE4Water (D7.3. FIWARE[6])

[6] https://www.fiware4water.eu/sites/default/files/delivrables/F4W-D7.3-InitialDMP_final.pdf

In aqua3S[7], a similar naming convention was applied where filenames consisted of prefix relating to the work package the data was obtained for, a descriptor of data source e.g. Sensor, the location/use case where the data was obtained, and finally a unique suffix identifier.

In WATERVERSE, specific rules have been adopted for generating file names that have a consistent structure, while simultaneously describing the content of files and the relationships to other processes, files etc. Specifically, the following naming conventions have been applied depending on the sources or characteristics from which those datasets be generated, collected, and managed, by the WATERVERSE partners or processes.

a) For the datasets collected by pilot use cases the name convention that will be followed is:

WATERVERSE_<PUC*x*>_<*name of source*>_<*description of source*>_V#.#

where

- *PUCx*: is the number of Pilot Use Case
- *name of source:* indicates a descriptor of data source e.g. Sensor, SCADA, etc.
- *description of source:* indicates any kind of short identification of the source e.g. measurements such as chloride, water flow, conductivity, level, Precipitation forecast etc.
- *V#.#*: implies the version of the respective data

As an example: WATERVERSE_PUC1_ RWS_API_ Chloride_V0.1

b) For the datasets generated by the WATERVERSE data management ecosystem (ROs/tools) the name convention that will be followed is:

WATERVERSE_<WP*x*>_<*name of tool or RO#*>_<PUCx>_V#.#

where

- *WPx:* is the number of Work Package
- *name of tool or RO#:* indicates the name of the tool or the Research Output number, e.g. RO#6_AI_Data_Validation_Tool
- *PUCx:* is the number of Pilot Use Case that original dataset generated
- *V#.#:* implies the version of the respective data

As an example: WATERVERSE_WP3_ RO#6_AI_Data_Validation_Tool_PUC1_V0.1

c) For various usage datasets such as datasets generated for project management, communication, and dissemination purposes etc, the name convention that will be followed is:

WATERVERSE_<WP*x*>_<T*x. x*>_<*description of dataset*>_V#.#

where

- *WPx*: is the number of the Work Package
- *Tx.x*: is the number of the task
- *description of dataset:* indicates a short description of the dataset, e.g. Newsletter, Subscribers, Events_and_Publications, etc.
- V#.#: implies the version of the respective file

[7] https://cordis.europa.eu/project/id/832876

As an example: WATERVERSE_WP2_T2.1_Stakeholder_Group_Contact_List_V0.1

The adoption of such naming conventions can immediately convey meaning directly to the user about the dataset being queried. In addition to the identifiers, it is important to consider what metadata will be created and the metadata standards that will be utilised. The inclusion of metadata with the dataset will facilitate the user in understanding the data that is being presented whilst providing further context to the data such as when the data was generated, how the data was generated i.e. type of sensor used for capturing the data, the accuracy of the sensor, and the location (where geographical information is applicable and used) as to where the data was obtained.

In the context of discoverability, having metadata that clearly describes aspects of the dataset, allows users to search for data without having to know that said data exists, the user can base their custom search on parameters such as information they are seeking and where and when, and they can be presented with list of datasets that align with their search criteria.

3.2 Accessible: Standards and protocols for data access and sharing

Accessibility of generated data within a project is an important aspect both within FAIR and MELODA5 principles, certain criteria needed to be considered. One of the first aspects to consider is how the data will be accessible by individuals. For this, consideration is needed as to where data is stored and what tools are needed to access/fetch the data and whether specialist tools are required to interpret the data. To enable access to data collected/processed by project partners the corresponding metadata that generated with said datasets will need to be accessible from an open data portal via an API. One example of a potential open data portal is that of the European Open Data Portal[8] platform which was developed by the European Commission to provide access to a multitude of open data sets from various EU institutions and member states.

Within certain cases there will be the need to restrict access to datasets dues to security, commercial, and/or licensing constraints. As such infrastructure should be in place to monitor and control (where applicable) access to datasets.

3.3 Interoperable: Methods for ensuring data compatibility and integration with other datasets

Interoperability refers to the ease and ability of different systems, tools, and applications to communicate, exchange, and utilise data. To achieve interoperability, data providers will need to use common data standards, metadata schemas, and ontologies that allow data to be easily shared and understood by others. This includes using common file formats, data models, and terminologies that are widely accepted and recognised within the sector. Additionally, interoperability requires the use of open data protocols and technologies that enable data to be shared and accessed across different systems and platforms. This includes the use of APIs (Application Programming Interfaces), web services, and other data exchange mechanisms that facilitate data sharing and integration across different domains and disciplines. In the scope of the WATERVERSE project, data handling and APIs will be carried out via the use of context brokers utilising the FIWARE framework.

[8] https://data.europa.eu/en

To enable increased interoperability within the WATERVERSE all data flows will be translated into JSON (JavaScript Object Notation) file formats. This file format is a lightweight interchange format meaning that it is designed to use minimal storage space and network bandwidth, whilst still within a structured format that is easy for both humans and machines to interpret. Additionally, JSON file formats have the flexibility to accommodate a wide range of data types including data with spatial properties, which makes it a suitable candidate for the WATERVERSE project. Furthermore, adding to its interoperability, the standardised format of JSON files allows for easy exchange between different systems and platforms providing greater reach of the generated/processed data.

To facilitate the interoperability of data within the WATERVERSE project, the project is implementing the use of established Smart Data Models. These models define a format that allows for data interchange between organisations through the adoption of open standards. Previous work in FIWARE4Water developed Smart Water data models to be utilised within water supply networks that allows for the connection between NGSI platforms to EPANET. If the project partner was collecting data relating to water quality, they could utilise the Water Quality model within Smart Environment (Figure 5)


```
{
    "id": "urn:ngsi-ld:WaterQualityObserved:waterqualityobserved:Sevilla:D1",
    "type": "WaterQualityObserved",
    "NO3": 0.01,
    "conductivity": 0.005,
    "dateObserved": "2017-01-31T06:45:00Z",
    "location": {
        "coordinates": [
            -5.993307,
            37.362882
        ],
        "type": "Point"
    },
    "measurand": [
        "NO3, 0.01, M1, Concentration of Nitrates"
    ],
    "pH": 7.4,
    "temperature": 24.4,
     "flow": 127.53,
    "@context": [
        "https://uri.etsi.org/ngsi-ld/v1/ngsi-ld-core-context.jsonld",
        "https://raw.githubusercontent.com/smart-data-models/dataModel.WaterQuality/master/context.jsonld"
    ]
}
```


Figure 5: Example of Smart Data Model representing Water Quality

3.4 Reusable: Best practices for data documentation and curation

The reusability aspect of data falls under both FAIR and MELODA5 dimensions. Data that is produced/processed within the project should be reusable in the sense that it can be used in different contexts and purposes, by providing machine-readable information on how to reuse the data, including the limitations/restrictions in place upon said data due to its licence.

For the data to be machine-readable it needs to be formatted in a way that can be easily read and processed by a computer. To facilitate this, data should be structured using a standardised format or language that can be recognised by a computer such as CSV, XML, JSON, GeoJSON, WKT etc. These

standardised data formats are both human and machine-readable and as such do not require specialised proprietary software to read.

3.5 Geolocation: Assigning geolocation properties to data

The geolocation information is a key aspect to facilitate the reuse of the data and help to search the information based on the proper location of interest. Both data/datasets, and processes can provide a location information to specify where the data is obtained, the datasets are generated to the location in which this information is maintained, and which is the location of the responsible to maintain these processes.

Geolocation data refers to any data that provides with reasonable precision the location of the information (any objects, entity, or element) on our planet. Typically, this information is obtained from a signal generated from an electronic device (e.g., GPS receivers, mobile phones, smart watches, sensors). The most common way to provide the location is with longitude and latitude coordinates (Figure 6). Longitude is a location measurement east or west of the first meridian at Greenwich, the north-south (imaginary) line that joins both geographic poles and Greenwich, London. Latitude is a measurement of the distance both north or south from the Equator. Both are measured in degrees, minutes, and seconds (e.g., 51° 37′ 40″ N, 8° 45′ 45″ E).

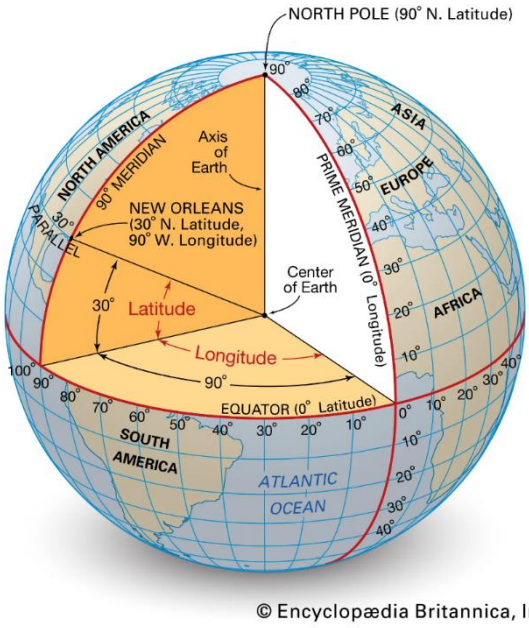

Figure 6: Representation of geographical coordinates system

It is possible that the location also reflects some more complex structures like a Polygon, Multipolygon, or areas, though these can also be basically described as a set of points (latitude and longitude coordinates) that defines the geometry of them.

In terms of data representation, there are several ways to represent the location in terms of data structure currently in use:

- Shape or shapefiles, a vector data storage binary format.
- GeoJSON, is a geospatial data interchange format based on JSON objects.
- TopoJSON, is an expansion of GeoJSON for encoding topologies, eliminating redundancy, and allowing related geometries to be stored in the same file.

Shapefiles are complex, and are commonly used, however, they do not compress very well. GeoJSON (raw) files (Figure 7) are text-based files that in addition to the non-geographical attributes, describe the geolocation of each point/vertex (longitude and latitude (WGS84 coordinate system in decimal degrees), or easting and northing for projected coordinates) within the spatial dataset and as such can result in large file sizes. In contrast to shapefiles however, as these files are in essence text files represented within a JSON style format, they compress well and are relatively fast to encode and decode. TopoJSON are minimal in space consumption but require long times to process the encode and decode and this process consumes huge amounts of memory.

```
"location": {
    "type": "GeoProperty",
    "value": {
        "type": "Point",
        "coordinates": [-8.5, 41.2]
    }
}
```

Figure 7: Example of location (point) representation in GeoJSON format (value)

Taking into consideration this analysis, the adoption of GeoJSON is preferred to facilitate the management of data as well as the compression. Moreover, this is also aligned with the location format adopted in FIWARE Technology and FIWARE NGSIv2 and ETSI Next Generation Service Interfaces NGSI-LD APIs[9].

3.6 Update Frequency

In a data model environment, the update frequency principle shows how often information is expected to be updated inside a dataset. Therefore, it reflects the frequency in which it is expected to update the content of your datasets. We should not misunderstand this with the value of how often we check the data of a dataset. The expected values for update frequency may be:

- Every day, the updated frequency of the dataset is every day.
- Every year, the updated frequency of the dataset is every year.
- Live, when the dataset updates are continuous and ongoing.
- As needed, when we do not know really the prediction of these updates.
- Never, datasets that never change.

The last option is not the best choice, though it is kept for the case that some datasets may have no update frequency, nevertheless the score of this MELODA5 dimension will be 0. The WATERVERSE FAIR Guidelines recommend the nearest less frequent regular value and dismiss the use of “As needed” or “Never”. This selection helps to monitor the datasets freshness.

[9] https://www.etsi.org/deliver/etsi_gs/CIM/001_099/009/01.06.01_60/gs_cim009v010601p.pdf

Regarding the representation of those values, we adopt the ISO 8601, especially the duration representation of date represented by the following format:

P[n]Y[n]M[n]DT[n]H[n]M[n]S or P[n]W

Where [n] is replaced by the value of each of the date and time (leading zeroes are not needed). The capital letters have the following meaning:

- P is the identification tag to express a period time.
    - Y is the “year tag” that follows the number of calendar years.
    - M is the “month tag” that follows by the number of calendar months. It is recommended not to put more than eleven (11) months, if it is the case, increase the number of years.
    - W is the “week tag” that follows the number of weeks designator that follows the value for the number of weeks.
    - D is the “day tag” that follows the number of days.
- T is the “time tag” that shows that the next characters make reference to a time information.
    - H is the “hour tag” that follows the value for the number of hours, the recommendation is not specify more than twenty-three (23) hours.
    - M is the “minute tag” that follows the value for the minutes, the recommendation is not specify more than fifty-nine (59) minutes.
    - S is the “second tag” that follows the value for the seconds, the recommendation is not specify more than fifty-nine (59) seconds.

It is possible to represent a fraction of values represented with either comma or point (e.g., “P0.5D” or “P0,5D” for half a day updated frequency).

Finally, if we wanted to represent some update frequencies, for example, “P1D” indicates that the update frequency will be every day, "PT0S" or "P0D” will indicate a live update frequency. For consideration, empty string value “” will be considered as a dataset that will not be updated.

3.7 Dissemination

According to the MELODA5 dimensions, the dissemination dimension is mainly about letting people know about any update of the data (new data in the dataset or enrichment of the dataset).

To obtain a good level of dissemination, proactive ways of communication must be set up:

- All updates are precisely documented and promoted in different places.
- Push notifications mechanisms are configured and active. It can be a text based RSS (Really Simple Syndication) feed, a Twitter feed, or any other mechanism that allows users to be notified instead of having to go to a website to see if something has happened since their last visit.

To increase the dissemination, other means can also be considered. One common way to do it is to use the data federation principles. By referencing a dataset on many data portals, using the harvesting mechanisms, this allows any update to be automatically promoted on these satellite data portals.

The MELODA5 dimensions define three levels for this dimension:

- Communication / dissemination not systematic.

- Available resources on updates (i.e., RSS feed).
- Proactive dissemination / push dissemination (information automatic and timely).

Based on these defined levels, a metric on the dissemination level of a dataset or data process can be manually evaluated by measuring or checking:

- If any update is thoroughly documented and communicated in any way (level 1).
- The number of channels where any update is notified.
- The number of other data providers platform where the data is referenced.

Achieving level 2 or 3 then mainly depends on the sum of the two counters based on the number of channels and federated data providers:

- A sum less than or equal to 2 maps to level 2.
- A sum greater than 2 maps to level 3.

3.8 Reputation: Methods for determining reputation of data sources

According to the MELODA5 dimensions, the reputation dimension is about giving users insights and feedback on the usages of the dataset.

To provide this, some tools have to be made available to users so that they can share their experience with the dataset (which can also help the data producer to improve it). It is also important to display some statistics about the usages of the dataset: count of downloads / API calls, count (and links) of reuses of the data, etc.

The MELODA5 dimensions define three levels for this dimension:

1. No information about the reputation of the data source.
2. Statistics or reports published on user's opinions.
3. Indicators or rankings on reputation of the data source.

As for the dissemination, the reputation can be evaluated only with a manual checking process:

- If there is no information about the usages of the data.
- Number of usage statistics published about the data.
- Number of reuses of the data.

Achieving level 2 or 3 then mainly depends on the sum of the two counters based on the number of channels and federated data providers:

- A sum between to 2 and 5 maps to level 2.
- A sum greater than 5 maps to level 3.

4.0 Defining a FAIR Ecosystem for the WATERVERSE

This section outlines how the FAIR and MELODA5 dimensions will be implemented within the WATERVERSE utilising FIWARE infrastructure.

4.1 Selection and Mapping of FAIR principles and MELODA5 dimensions into a Minimum Viable Ecosystem

WATERVERSE project is developing two steps beyond the current State of the Art in the application of FAIR principles. The first one consists of the adoption of MELODA5 dimensions for improving the Open Data quality of datasets together with the application of FAIR principles. Tables 2- 5 within this section outline each of the FAIR principles guidelines and how they are implemented within the FIWARE framework. In these tables, we also specify how we will implement those FAIR Principles and MELODA5 Dimensions and how we can apply these principles and dimensions into the corresponding data models.

The second consist in the adoption of FIWARE technology and the Smart Data Models (SDM) program for definition of the Data Models involve the adoption of the standards used in this community for data representation, JSON/JSON-LD for data format representation and JSON Schema for data model definition. Once we have defined a JSON Schema, the Smart Data Models program automatically generates several examples of representation of an entity data in different formats (e.g., JSON, JSON-LD, CSV, DTDL, etc…). In the case of using FIWARE Technology the Entity Data representation format used is either JSON or JSON-LD. The adoption of the ETSI standard NGSI-LD API for informational representation of something that is supposed to exist in the real world (physically or conceptually) also facilitates the use of JSON-LD data format representation. Additionally, the use of the FIWARE NGSIv2 also facilitates the use of JSON data format representation. Both use the same data representation defined in the SDM program for data modelling representation. Figure 8 shows the NGSI-LD information model utilised in order to explain some of the links between the FAIR principles and MELODA5 dimensions.

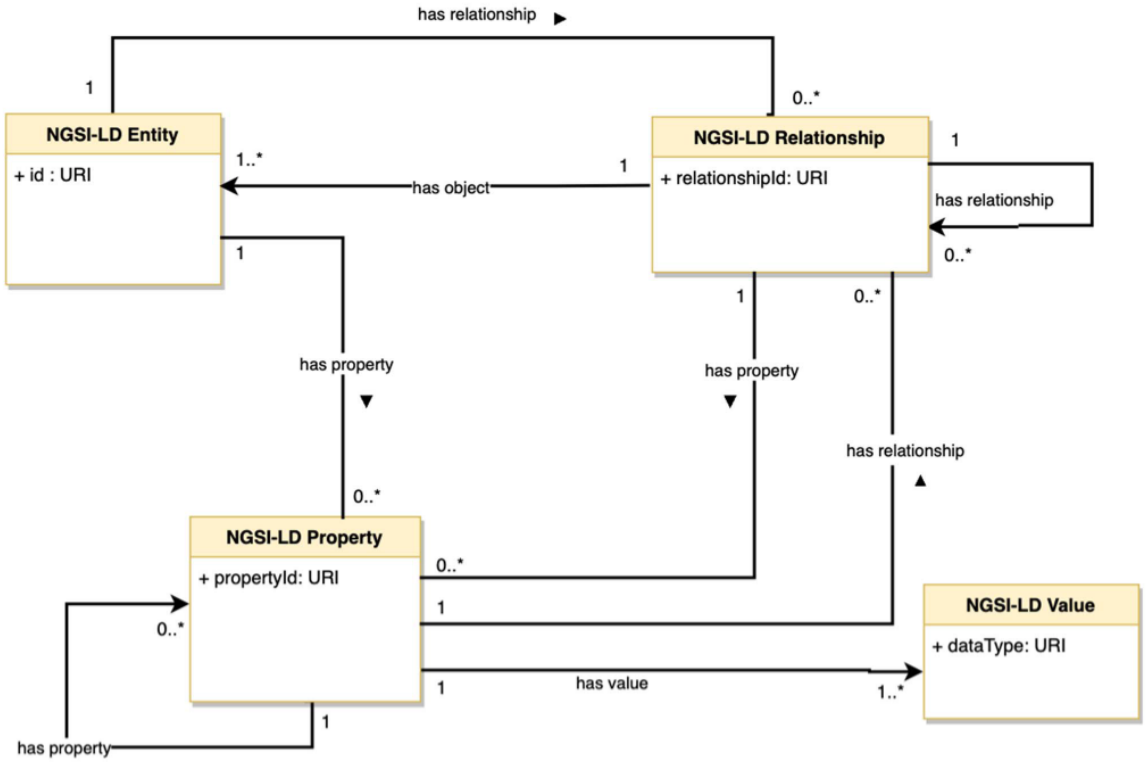

Figure 8: NGSI-LD information model

Moreover, the WATERVERSE project is going another step beyond the current application of FAIR principles with the adoption of this principles and dimensions not only on the datasets but also in normal data representation and any data process that we are defining in the project. It is clear that not all the principles or dimensions can be applied to them, but several can be applied. This activity is very innovative and therefore its definition and application may evolve throughout the execution of the project based on the results we obtain in the different Pilots with the intention of refining and improving the metrics in use. Therefore, for convenience, the terminology that we use are the following:

- **Metadata for entity data** makes reference to the underlying model that it is defined to represent any domain specific data (e.g., wastewater tank, pipeline, etc.).
- **Metadata for datasets** makes reference to the information that facilitates the data exchange of datasets (e.g., license of datasets, URL to access data, format of the data, compression algorithm, etc.). A dataset is a collection of data, published or curated by a single source, and available for access or download in one or more formats.
- **Metadata for data services** makes reference to the set of data processing functions used over datasets (e.g., anomaly detections, anonymization, data cleaning and filtering, etc.).

Besides, we adopt the use of DCAT Application Profile (DCAT-AP) for the definition of datasets and data services. DCAT-AP[10] is a DCAT profile for sharing information about Catalogues containing Datasets and Data Services descriptions in Europe developed in the context of Action 1.1 – Improving semantic interoperability in European eGovernment systems[11] of the European Commission's Interoperability Solutions for European Public Administrations (ISA) programme[12].

The next subsections describe the adoption of the FAIR principles and MELODA5 dimensions into the corresponding entity data, datasets, and data services metadata and which solution we have adopted to fulfil them. For understanding, each subsection ends with a table with the corresponding implementation option adopted. The font colour in the column Implementation, especially red font, indicates that this option has to be implemented in the task 4.2 during the execution of the project. The normal colour means that it is not needed to implement anything else apart the suggested thing to cover the principle or dimension.

4.1.1 Findable principle

Analyse the Findable principles of FAIR, the first point is the management of **Unique Identifier (F1)**. This is the most important FAIR principle because if we have no global unique and persistent identifiers it is not possible to achieve other aspects of FAIR. We should not confuse with the meaning of these identifiers that are related to the FAIR principles I1-I3. The purpose is to assign a global unique persistent identifier to the data and metadata. This is something that we can achieve basically adopting FIWARE NGSIv2 or ETSI NGSI-LD API. Figure 8 shows in detail the Information Model of an Entity and each entity is identified with a unique identifier provided by a URI in the **id** property as shown in Figure 9. This is something that will be applied both entity data, metadata for entity data, metadata for datasets, and metadata for data services.

```
{
  "id": "urn:ngsi-ld:Pump:74azsty-70d4l-4da9-b7d0-3340ef655nnb",
  "type": "Pump",
  ...
}
```

Figure 9: Defining a Unique identifier for an asset

Additionally, related to the metadata for data service, there is a specific property, endpointURL, which is the root location or primary endpoint of the data service (an IRI) and therefore can be considered as unique identifier of the data service as well.

The second principle, **data are described with rich metadata (F2)**, means that the definition of the data should include as much as possible metadata information to facilitate machines to develop searching and sorting processes based on this metadata. This helps other people to locate this data and therefore increase the reuse and citations of it. In the WATERVERSE project, it is achieved with the use of the context definition in the JSON-LD files (Figure 10). This context is used to map the terms

[10] https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe/release/211

[11] European Commission. Interoperability Solutions for European Public Administrations (ISA). Improving semantic interoperability in European eGovernment systems. http://ec.europa.eu/isa/actions/01-trusted-information-exchange/1-1action_en.htm

[12] European Commission. Interoperability Solutions for European Public Administrations (ISA). http://ec.europa.eu/isa/index_en.htm

in the JSON-LD payloads to IRIs (Internationalized Resource Identifier). We can see an IRI as a sequence of characters. A mapping from terms to IRIs is defined which allows using a term (short, or compacted form) instead of an IRI (long or expanded form) in order to identify resources. This allows us to describe (semantically) any term described through an IRI and use the term in JSON-LD (smaller payloads).

```
{
  "@context": {
    "accuracy":
"https://smartdatamodels.org/dataModel.WaterDistributionManagementEPANET/accuracy",
    "address": "https://smartdatamodels.org/address",
    "alternateName": "https://smartdatamodels.org/alternateName",
  ...
}
```

Figure 10: Data description in metadata

Therefore, the adoption of JSON-LD involves the use of context definition in the data and it is associated to all kind of entity data, including the metadata representation of datasets and data services using DCAT-AP.

Following, **metadata include clearly a unique identifier (F3)**. Usually, the content of the datasets and data services and their metadata definitions in DCAT-AP are different information. DCAT-AP classes used to define the metadata information form where datasets and data services are Datasets and DataServices classes. These classes need to have also their corresponding data models to define what it expected to be defined in these classes and therefore both of them need to have a unique identifier. It is also valid for the Entity Data and its own Data Model definition where a unique identifier is needed to identify the corresponding data model. Therefore, it is needed to make a relationship between them through a globally unique identifier defined in the data models. This is achieved through the use of JSON Schema in order to define the metadata used in a data model. JSON Schema defines a best practice to include an $id property (Figure 11) as a unique identifier for each schema.

```
{
  "$id":
"https://smart-data-models.github.io/dataModel.WaterDistributionManagementEPANET/Pipe/schema.json",
  ...
}
```

Figure 11: Metadata identifier

Afterwards, the entity data uses a specific Uniform Resource Name (URN) for its own unique entity "id" (Figure 12), which includes the reference to the JSON Schema Id, to be precise the entity type (in the previous case, Pipe), together with a global unique identifier for the data.

```
"id": "urn:ngsi-ld:Pipe:74azsty-70d4l-4da9-b7d0-3340ef655nnb",
```

Figure 12: Entity id example

Last but not least, **(meta)data have to be findable (F4)**, deals with the way in which the data can be found or searchable through internet. These data and datasets should be discoverable, including indexing. The first step consists of the adoption of the same solution explained in F2, the use of @context in JSON-LD and the use of URIs to define properly the terms used on JSON-LD. It is mandatory that these URIs be publicly accessible. Smart Data Models program automatically checks that the URIs used to define the terms (properties) are accessible.

Additionally, in order to link the corresponding data with the definition of the data model, a new property relating to the validation schema (Figure 13) is defined to provide the link to the data model. It is applied to the metadata for entity data, metadata for datasets, and metadata for data services. In this way, the client solutions can implement a validation tool to check that the structure of the data is following the data model defined. This property is also related to the Data Model MELODA5 dimension.

```
{
 "validationSchema": {
    "type": "property",
    "value":
"https://github.com/smart-data-models/dataModel.WaterDistributionManagementEPANET/blob/master/Pipe/schema.json"
 }
}
```

Figure 13: Validation schema

Regarding automatic discoverable and indexing, all data models and terms or properties used in the Smart Data Models program are automatically indexed by Google Engine. Every term is data models has an associated page (https://smartdatamodels.org/<Subject>/<term>), e.g. https://smartdatamodels.org/dataModel.WaterQuality/measurand. Currently (11th May, 2023) there are 20,939 terms available for searching and more than 960 pages indexed by Google Search Engine.

Table 2 summarises all the choices selected to provide the implementation of the Findable principles.

FAIR principles	FAIR	Implementation
Unique identifier	F1	• NGSI-LD includes id • JSON Schema includes $id • DCAT-AP:DataService:endpointURL property

FAIR principles	FAIR	Implementation
Data includes metadata	F2	• @context information includes the metadata information about the data, published in the Smart Data Model program
Metadata include identifier	F3	• JSON Schema includes $id
Metadata is searchable	F4	• @context information includes the metadata information about the data, this data is represented through a URI that it is publicly accessible • Smart Data Models automatically allows the use of Google Search Engine to find terms • Data includes a validationSchema property linking to the Data Model definition (metadata for entity data, metadata for datasets and metadata for data services).

Table 2: FAIR Principles – Findable (red font indicates to be implemented in task 4.2)

4.1.2 Accessible principles

Regarding Accessible principles, the first point to analyse is the **retrieval of data and metadata using standard communications protocol (A1)**. This principle outlines how data and metadata can be obtained based on their identifiers. This access should be implemented without the use of proprietary tools or communication methods such as via the adoption of FIWARE NGSIv2 or ETSI NGSI-LD as APIs to access the Entity information. To be more precise, the resource entities/{entityId} represents an entity defined by an Id (URI) as it is defined in F1 in the previous section. Therefore, we can use the resource method GET to access the corresponding information associated with the URI (entity identifier). With the adoption of these protocols, there is no additional need to make any other development or adoption.

Regarding the access to the metadata definition of a data model through the JSON Schema id, this is something that is currently not implemented but will be added to the list of services available to be implemented within the scope of the WATERVERSE project.

The second principle, the **use of an open, free, and universally implementable protocol (A1.1)**, means that in order to maximise the data resume, the protocol that we selected to be used to access the data should be free and open source. These properties allow the implementation of any solution to facilitate the data access. These objectives are obtained as well with the use of these APIs. They are fully available and there are several implementations of Brokers using these protocols.

- FIWARE NGSIv2, this is the last version of the FIWARE API still in use in several FIWARE implementations and also used in several FIWARE Components in the FIWARE Catalogue. You can access to the detailed information in the following link https://github.com/telefonicaid/fiware-orion/blob/master/doc/manuals/orion-api.md
- ETSI NGSI-LD, the standardized version of the API based on the previous version started in 2017 and the last released version (v1.6.1) is from August 2022. You have the public access to the specification in the following link

https://www.etsi.org/deliver/etsi_gs/CIM/001_099/009/01.06.01_60/gs_cim009v010601p.pdf

Regarding the third principle**, authentication and authorization procedure of the protocol (A1.2)**. This is important to clarify that the Accessibility (A) does not necessarily mean open or free data, but one should provide the exact conditions under which this data can be accessible. This must be defined in a way that machines can understand how to access the data and therefore it affects the decision about the localisation of the repository in which the data will be shared. We need to make a differentiation between data access and metadata access.

In the first case, the adoption of FIWARE NGSIv2 and ETSI NGSI-LD are compatible with the use of OAuth2 flows. In fact, this is the approach that is followed in the FIWARE Ecosystem to provide authentication and authorization access to the data based on token. OAuth 2.0 is the industry-standard protocol for authorisation and it is focused on client development simplicity while providing specific authorization flows for web applications, desktop applications, and so on. It is possible also to define fine grained authorization access to the data based on XACML[13] standard which allows defining a core schema and corresponding namespace for the expression of authorization policies in XML format against objects that are themselves identified in XML.

The second case is related to the way in which we define the authentication and authorization access to a datasets and data services. FIWARE Community has adopted the use of DCAT-AP[14] (Data Catalogue Vocabulary - Application Profile) which is an RDF data vocabulary to facilitate the interoperability of data catalogues published on the web. This application profile easily allows the definition of the metadata associated to a dataset. Specifically, DCAT-AP has a single property that allows the link between an Agent (typically, an organisation) to a Dataset through the property dct:publisher[15] defined as “An entity responsible for making the dataset available”. In some specific use cases in which there is a data exchange between domain-specific portals, maybe should be needed to express other more precise agent roles. During the execution of the project, we will evaluate the necessity to develop extensions of the base profile in order to define additional properties with more specific roles and translate the suggestion to the DCAT-AP joinup initiative (Figure 14) through the creation of new issues of participating in meetings.

Additionally, it is possible to define the Rights Statement of your dataset through the dct:RightsStatement[16]. Which specify the intellectual property rights over the dataset as well as indicate if it is open data, has access restrictions or is not public. Associated with a Catalogue. Additionally, associate to the Data Service may provide information about access or restrictions based on privacy, security, or other policies.

In case of data services, the DCAT-AP DataService class also define the property accessRights, also dct:RightsStatement which may include information regarding access or restrictions based on privacy, security, or other policies for the corresponding service.

[13] http://xml.coverpages.org/XACMLv20CD-CoreSpec.pdf
[14] https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe/release/211
[15] http://purl.org/dc/terms/publisher
[16] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/2012-06-14/#terms-RightsStatement

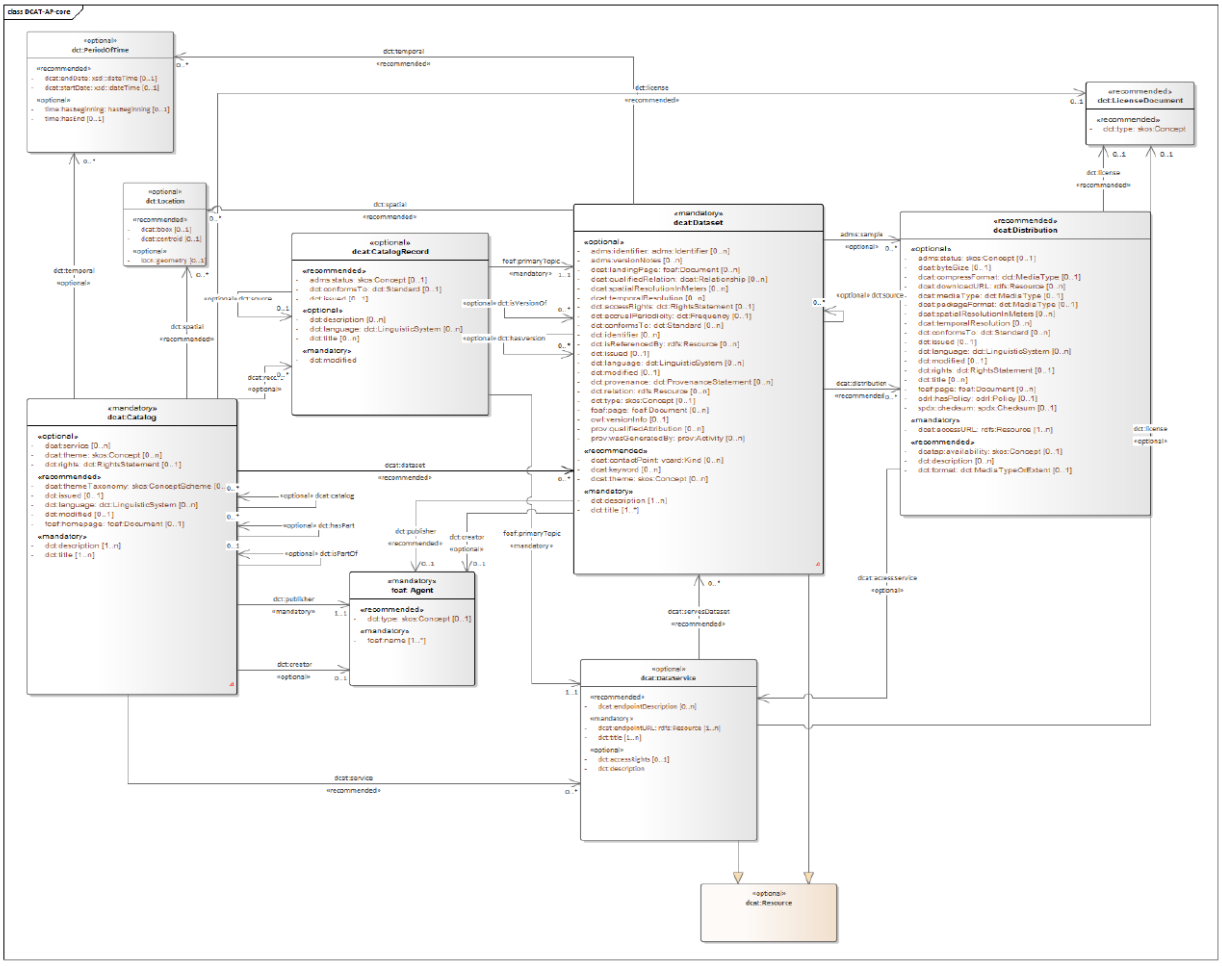

Figure 14: DCAT Application Profile UML Class Diagram[17]

Finally, the fourth principle**, metadata are accessible (A2)**, means that the metadata information, related to an entity data, a dataset, or a data service, is maintained even though the corresponding datasets are not available any more. This principle is related to the registration and indexing issues defined in F4. We differentiate between metadata associated with entity data, which is represented in a data model, and the metadata associated with a dataset, which is represented in DCAT-AP Classes. In the first case, the access to the metadata is guaranteed through the Smart Data Models program in which all the JSON Schemas are publicly accessible through GitHub under specific domains (https://github.com/smart-data-models) (Figure 15). In the second case, DCAT-AP dataset class defines a property landing page, which refers to a web page that provides access to the Dataset, its Distributions and/or additional information like the JSON Schema definition of the data models. This information is pointing to the original data provider and not any kind of aggregator of datasets or (meta)data information. In case of metadata for data services, the DCAT-AP DataService class also includes the property endpoint description, which contains a description of the service available via this endpoint, given specific details about the actual endpoint instance.

[17] https://github.com/SEMICeu/DCAT-AP

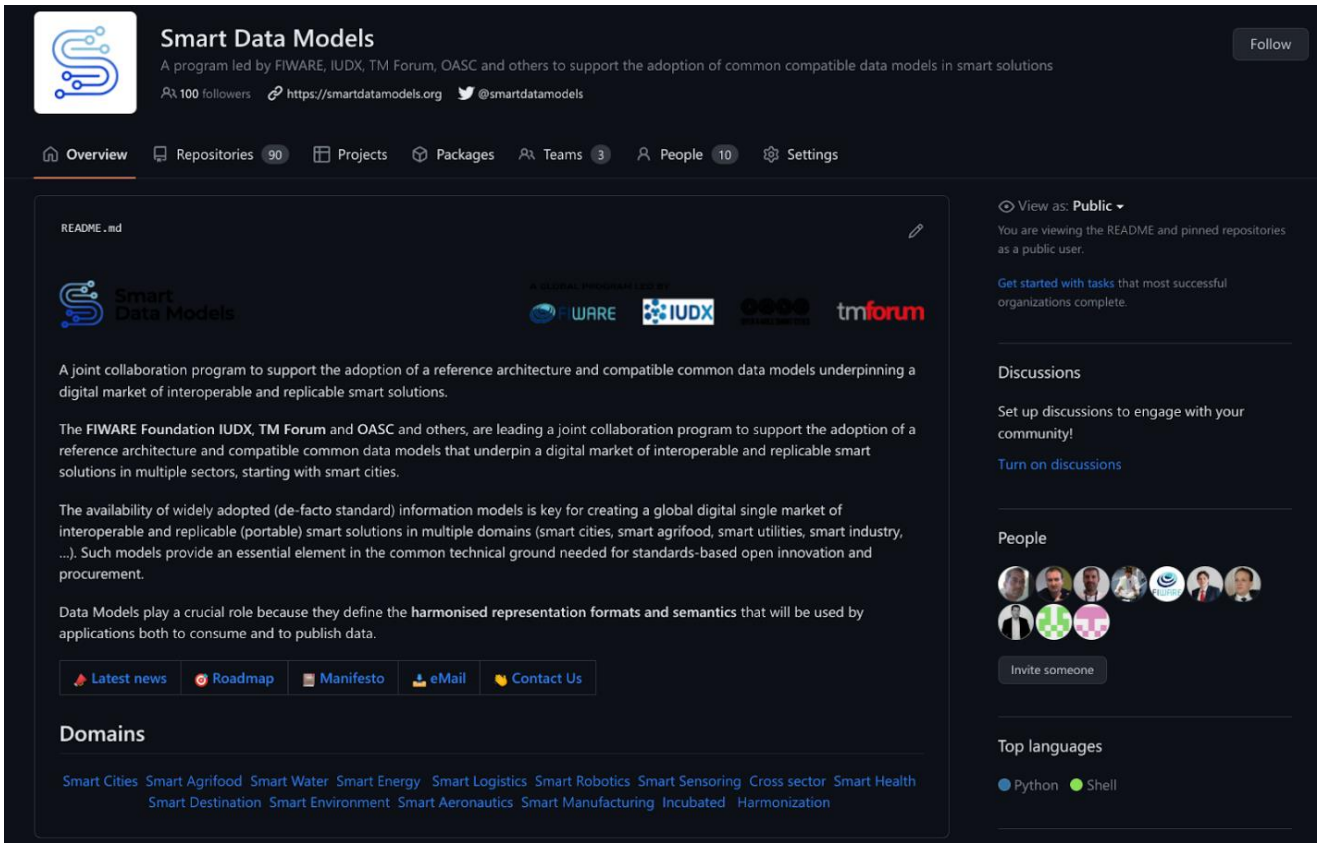

Figure 15: Smart Data Models program in GitHub[18]

Table 3 summarises all the choices selected to provide the implementation of the Accessible principles.

FAIR principles	FAIR	Implementation
(Meta)data are retrievable by their identifier using a standardised communications protocol	A1	• ETSI NGSI-LD allows retrieving data using the Entity Id • The access to Metadata based on $id needs to be implemented.
The protocol is open, free, and universally implementable	A1.1	• ETSI NGSI-LD
Protocol allows authorization	A1.2	• oAuth2 flows in ETSI NGSI-LD API • DCAT-AP:Dataset:publisher Property. Model:'foaf:Agent'. • DCAT-AP:Dataset:accessRights Property. Model:'dct:RightsStatement' • DCAT-AP:DataService:accessRights Property. Model:'dct:RightsStatement'
Metadata are accessible	A2	• Accessible through Smart Data Models program

[18] https://github.com/smart-data-models

FAIR principles	FAIR	Implementation
		• DCAT-AP:Dataset:landingPage Property. Model:'dcat:landingPage'. • DCAT-AP:DataService:endpointDescription Property. Model:'rdfs:Resource'

Table 3: FAIR Principles – Accessible (red font indicates to be implemented in task 4.2)

4.1.3 Interoperable principles

About Interoperable principles, the first one to analyse is the **(meta)data use a format, accessible, shared, and broadly applicable language for knowledge representation (I1)**, means that we wanted to be able to exchange and understand data and metadata (both humans and machines). Therefore, it is needed the definition of a common understanding of the digital objects through a language of knowledge representation to describe these objects. WATERVERSE follows the current State of the Art and use JSON/JSON-LD data format to represent both the entity data and metadata information associated to the entity data, datasets, and data services. Additionally, we adopt the JSON Schema in order to define the data models to be used in order to represent the datasets and data services information (using DCAT-AP) as well as the entity data. It helps us to align with the FIWARE NGSIv2 and ETSI NGSI-LD APIs.

Although, this is the main data format that we use inside the project for data manipulation and management, the Smart Data Models program allows the representation of the data in other formats like CSV or SQL Schema as well RDF Turtle how it was suggested to be available in the services (see section 4.2).

Concerning the second principle, **(meta)data use vocabularies that follow FAIR principles (I2)**, means that the vocabulary that we use in order to define our entity data and metadata, associated to the entity data, datasets and data services, should follow the FAIR principles, in order that others, humans and machines, can find, access, interoperate, and reuse them. Therefore, this vocabulary needs to be documented and resolvable using global unique identifiers. The objectives of the WATERVERSE project are to translate the WATERVERSE FAIR Guidelines inside the Smart Data Models program with the intention to provide FAIR principles and MELODA5 dimensions as a characteristic in the definition of new data models. Additionally, the definition of JSON Schemas for each data model include the WATERVERSE recommendations in order to apply these principles and dimensions.

Finally, **(meta)data include qualified references to other (meta)data (I3)**, the objective is to create as many links as possible to enrich the contextual information about the data and metadata. This is resolved in our case through the adoption of JSON Schema and JSON-LD. JSON Schema includes some keywords for including schemas (metadata) together.

- **allOf** means that the content defined in the JSON Schema must be valid against all of the subschemas (example in Figure 16).
- **anyOf** means that the content defined in the JSON Schema must be valid against at least one of the subschemas.
- **oneOf**, means that the content defined in the JSON Schema must be valid against exactly one of the subschemas.

```
"allOf": [
{
      "$ref":
      "https://smart-data-models.github.io/data-models/common-schema.json#/definitions/GSMA-Commons"
                },
                {
                   "$ref":
      "https://smart-data-models.github.io/data-models/common-schema.json#/definitions/Location-Commons"
  }
```

Figure 16: References to other metadata within metadata

Additionally, the properties are defined in the Smart Data Models program following a specific pattern:

[Property|Relationship]. Model:[URI]. <description>

It is not mandatory, but it is really recommended to include the "Model" part in the description. This content makes reference to a qualified metadata definition of the property expressed through an URI and it is used usually to redirect to the definition of this property (e.g., https://schema.org/Text) as shown in Figure 17.

```
"power": {
  "type": "number",
  "description": "Property. Model:'https://schema.org/Number'. Units:'KiloWatt'. The power
supplied by the pump. All units are accepted in [CEFACT](https://www.unece.org/cefact.html) code."
}
```

Figure 17: Defining smart data model within metadata

Finally, it is possible to add the qualified references to other data through the use of the context (@context in Figure 18). This shows the value of a property definition using a URL where you can find more information about the corresponding property.

```
"@context": [
"https://raw.githubusercontent.com/smart-data-models/dataModel.WaterDistributionManagementE
PANET/master/context.jsonld"
]
```

Figure 18: Qualified references to other data within metadata

Table 4 outlines all the options selected to provide the implementation of the Interoperable principles.

FAIR principles	FAIR	Implementation
(Meta)data uses a formal, accessible, shared, and broadly applicable language for knowledge representation.	I1	• Use JSON/JSON-LD data format • Use of JSON Schema for data model definition
(Meta)data use vocabularies that follow FAIR principles	I2	• Smart Data Models should follow the FAIR principles and MELODA5 dimensions
(Meta)data include qualified references to other (meta)data	I3	• Data Models defined in JSON format (JSON Schema) allow the reference to other Data Models ("allOf", "anyOf", "oneOf") • Data Models definition in JSON Schema includes the reference to the semantic definition of the properties. Entity Data includes qualified references to other metadata in the @context

Table 4: FAIR Principles – Interoperable (red font indicates to be implemented in task 4.2)

4.1.4 Reusable principle

About Reusable principles, the first one to analyse is **(meta)data are richly described with a plurality of accurate and relevant attributes (R1)**. The purpose is that not only provide information allowing discovery of data but also information about the content of it. The contribution to the Smart Data Models program and definition of the Data Models following its recommendations, facilitates the description of accurate and relevant attributes. The Smart Data Models program helps to enable actual data interoperability between diverse systems based on open-licensed data models. The Smart Data Models community is focussed on building and maintaining any data model focusing on building and maintaining a data model through:

- flexible and extensible, allowing it to be adapted to a wide range of use cases. Few required attributes and free to use whatever attributes are defined in them;
- based on best practices (actual use cases) and open and adopted standards, ensuring that they are reliable and trustworthy;
- supported by a strong and active community of contributors and users, ensuring that it continues to evolve and improve over time.

Additionally, the entity data and metadata for entity data, datasets, and data services are defined based on required attributes and will be aligned with well-known semantic definitions of the Entities and DCAT-AP vocabulary defined to represent the corresponding attributes for the Datasets and DataServices classes (e.g., schema.org, standard specification, etc.).

Regarding **(meta)data released with a clear and accessible data usage license (R1.1)**, it is about legal interoperability in terms of which usage rights are attached to the entity data, datasets and data services. This is something that should be clearly described both for humans and machines. Additionally, the more the clarity of the license the more important for automated searches.

We differentiate between the metadata of datasets and data services and the entity data. The first one makes reference to the proper definition of the LICENSE to be used in the dataset and data service. The adoption of DCAT-AP, facilitates the definition of the Class Distribution with the property license, which refers to the licence under which the Distribution is made available. It will be defined properly by the owner of the metadata and facilitates the definition of the data usage policy. In case of Data Service class in DCAT-AP, there is a property, license, that contains the licence under which the Data service is made available. Additionally, we decided to include the license property into the dataset class in order to provide the license according to the id of the SPDX standard applied to the overall datasets.

The second one, license of metadata for entity data, is achieved with the rules defined in the Smart Data Model program. It is mandatory that each new defined data model incorporates the corresponding LICENSE file. The preferred license is Creative Commons 4.0 but also will be valid others like Apache 2.0 or other open licenses if they cover the following points:

- Recognise contributions.
- Allow free use and modification of the data models.
- Allow sharing the modifications.
- Do not impose other restrictions to use and adoption.

It is something that it is applied as well to the proper definition of the DCAT-AP classes into the SDM program.

Concerning, **(meta)data associated with detailed provenance (R1.2)**, it is expected that anyone that wanted to reuse a dataset, data service or entity data model knows where they are coming from, how to be acknowledged, who generated or collected them, how has it been processed, if it was published before, if is this data containing data from someone else that it was transformed of completed, etc.

In case of datasets, the detailed provenance is detailed through the use of the Dataset Class and specially through the corresponding property provenance (dct:provenanceStatement). This property is a statement of any changes in ownership and custody of a resource since its creation that are significant for its authenticity, integrity, and interpretation. This is defined in Dublin Core[19]. It may include a description of any changes successive owners made to the resource.

In case of data services, we have decided to adopt the same property, provenance (dct:provenanceStatement) like the list of libraries and tools used in the development and execution of the service expressed as URIs list to the corresponding version of these libraries in order to provide the provenance information about the service.

In case of entity data models definition, currently there is a property in JSON Schema, "$schemaVersion", that provides details about the different versions of the data model. Additionally, there is a file for subject, **CONTRIBUTORS.yaml**, in which we can define the person details of the contributor to the subject in which the data model is included. This information is not enough to provide a detailed provenance of the data model. WATERVERSE will develop a service to manage different versions of data models and the information about the provenance of them with the purpose that we can request this information if it is needed. It is also applied to the data model definition of the DCAT-AP Dataset and DataService classes.

[19] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/#http://purl.org/dc/terms/provenance

Finally, **(meta)data meet domain-relevant community standards (R1.3)** makes reference to the facility to reuse datasets if they are similar, same data type, data organised in standardised way, well-established and sustainable file formats, and documentation following a common template and using a common vocabulary.

With respect to datasets and data services, the adoption of The DCAT Application Profile for data portals in Europe (DCAT-AP) based on the Data Catalogue Vocabulary (DCAT) developed in the W3C defines a specification for metadata records to meet the application needs of Open Data portals in Europe keeping semantic interoperability with other applications based on reusing established controlled vocabularies (e.g. EuroVoc) and mappings to existing metadata vocabularies (e.g., Dublin Core, SDMX, INSPIRE metadata, etc.). DCAT-AP becomes currently a de facto standard in Europe with the side effect to increase the quality of the metadata associated with a dataset and therefore making data more accessible. The adoption of DCAT-AP Data Service class, applied to a dataset, give us the possibility to increase the metadata quality associated to the services that are used to manage any datasets.

Regarding entity data models, the Smart Data Models program has defined guidelines for defining new data models (Smart Data Models guidelines[20]). Additionally, Smart Data Models program defines the structure of the GitHub repository (Figure 19) in order to find any subject (e.g., Smart Water, DCAT-AP, etc.) and any data models (e.g., Pump, Dataset, etc.).

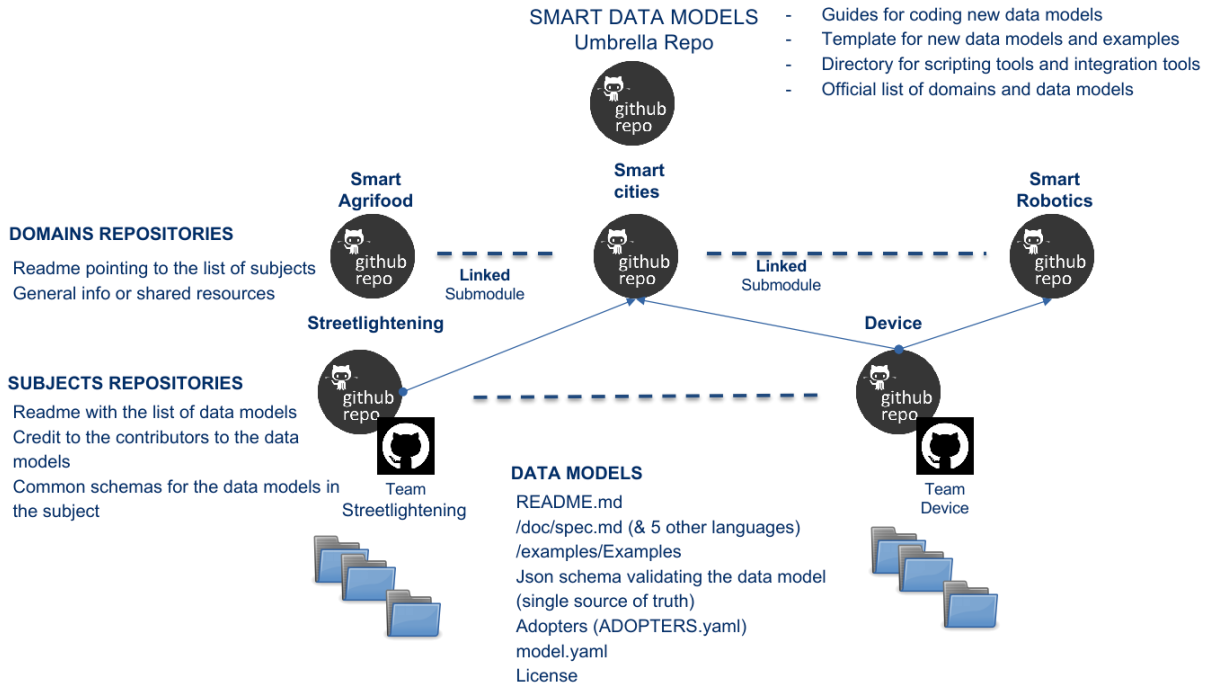

Figure 19: Smart Data Models structure in GitHub

Moreover, it is also defined the contribution workflow (Figure 20) in order that all the steps in the creation of a data models is clearly defined.

[20] https://github.com/smart-data-models/data-models/blob/master/guidelines.md

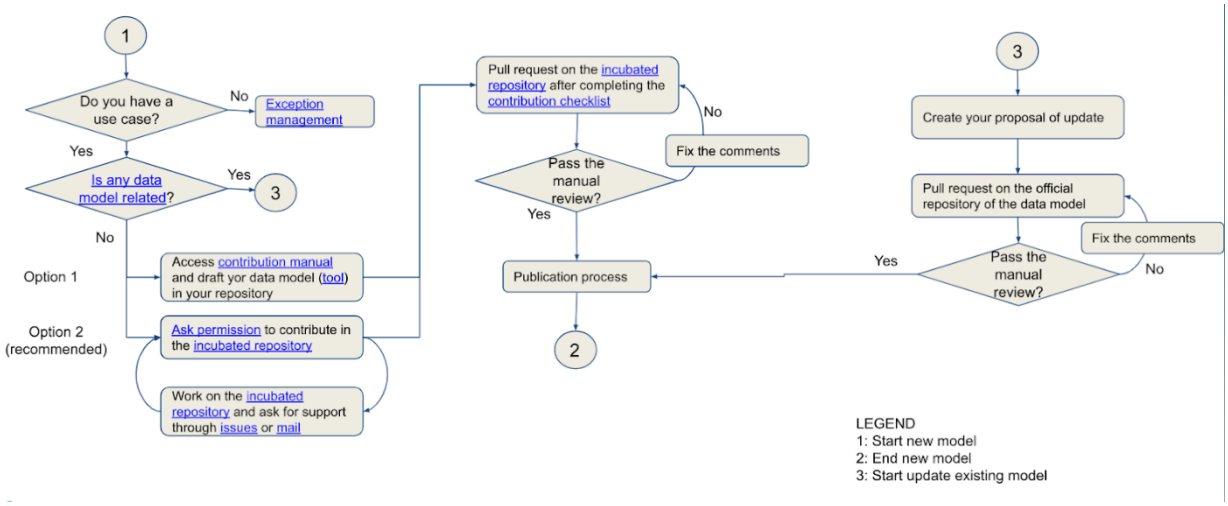

Figure 20: Smart Data Models contribution workflow

Table 5Table 5 outlines all the choices selected to provide the implementation of the Reusable principles.

FAIR principles	FAIR	Implementation
(Meta)data are richly described with a plurality of accurate and relevant attributes	R1	• Both data and metadata are described in the corresponding data models in the Smart Data Model program.
(Meta)data are released with a clear and accessible data usage licence	R1.1	• DCAT-AP:Distribution:license Property. Model:'dct:LicenseDocument' • DCAT-AP:dataset:license Property. Model:'dct:LicenseDocument' • DCAT-AP:dataservice:license Property. Model:'dct:LicenseDocument' • Smart Data Models program requests open LICENSE in Data Model definitions.
(Meta)data provenance	R1.2	• DCAT-AP:Dataset:provenance Property. Model:'dct:ProvenanceStatement'. • DCAT-AP:DataService:provenance Property. Model:'dct:ProvenanceStatement'. • Smart Data Models program will implement a versioning system of the data models.
(Meta)data meet domain-relevant community standards	R1.3	• DCAT-AP adoption • Smart Data Models guidelines and workflow.

Table 5: FAIR Principles – Reusable (red font indicates to be implemented in task 4.2)

4.1.5 MELODA5 dimensions

Table 6 shows the 8 dimensions of MELODA5, whilst highlighting synergies with corresponding components from the FAIR principles as defined previously in Figure 2Figure 2. Here we observe that four (4) of the MELODA5 Dimensions (see FAIR column) are already encompassed within the FAIR Principles guidelines. Nevertheless, in some cases, like Access to data and Data Model Standardization, we need to extend them with some properties to be aligned with MELODA5 metrics. However, other four (4) (Geolocation, Update frequency, Dissemination, and Reputation) are not covered or not fully covered.

It is important to mention that MELODA5 makes reference to the metrics associated with Open Data, therefore not all the entity data, datasets, or data services that will be published in the WATERVERSE project will have a valid metric or information about it. In such cases, the properties are not filled in. We decided to include them into the guidelines just to provide even better quality open data in case that we wanted to generate it.

Let's analyse one by one each of the dimensions. The first one makes reference to (meta)data are released with a **clear and accessible data usage licence** (**M1**). It is related to the R1.1 FAIR principle but in the case of MELODA5 we wanted to make reference to the explicit license associated to the information. It can be managed both in metadata (for entity data, datasets, and data services) and entity data specification. Regarding metadata for entity data, the license information is associated to the JSON Schema, each specification by default required the definition of a proper LICENCE file together with the schema. Regarding metadata associated with datasets and data services, the adoption of DCAT-AP facilitates the use of license information. By default, DCAT-AP include license information inside the class Distribution but with no details about which kind of license. It is mentioned that license is a license document giving permission to do something with a resource. We propose to define a new License property inside the class Dataset to explicitly detail which is the license that it is associated with this dataset. Additionally, we extend the property License inside the Data Service class in order to provide the same information about the different type licenses that can be adopted.

The next dimension, **access to information** (**M2**), is related to the A1 FAIR principle but in this case, we wanted to provide details about which kind of access is available to recover the information of the data associated to the metadata for datasets and metadata for data services. At the moment, there is no property in DCAT-AP that cover this information therefore, we suggest the creation of a specific one (Access Mechanism) inside the Dataset class to provide the detailed information about this type of access. Regarding metadata for data services, the DataService class provides the property endpointURL that provide access to the service. Therefore we will use it for this purpose.

Following with the next dimension, **technical format** (**M3**), it is related to the I1 FAIR principle and basically is covered with the adoption of JSON/JSON-LD for representing the entity data information and through the use of JSON Schema to define the metadata information of this data.

The next one, **data model** (**M4**), is related to standardization and is connected also with the I1, I2, and I3 FAIR principles. However in this case, MELODA5 requires to have a clear detail about which kind of information is provided about the standard used in the data and if that information is publicly available to get details afterward. Additionally, it is required to have details about where is the JSON Schema that it is used in order to validate the content of the data. That information currently is not possible to provide into DCAT-AP in the case of a dataset. Hence, we need to extend the Dataset Class with a set of properties, standardization, standardization source, and validation schema to keep that

information. Standardization property will keep the level of Standardization as it is registered in MELODA5 and Standardization Source will give the URL to the corresponding source where we can get details of the defined attributes of the data. Regarding Validation Schema, it will contain the URL of the corresponding JSON Schema used. Moreover, the validation of the schema will be applied to all data models that we generate in the project, which facilitates a direct link between the data and the metadata (related to the I3 FAIR principle).

Additionally, it is possible to define the corresponding meta schema[21] of the data model schemas associated to the definition of the metadata for datasets, metadata for data services and metadata for entity data. This meta schema is a schema against other schemas can be validated and it is self-descriptive due to JSON Schema meta-schema validates itself. Our activity will consist in the modification of the corresponding vocabulary to create a new keyword $conformsTo in order to provide the corresponding link to the meta schema. This meta schema will use to validate the JSON Schemas in the SDM program, both DCAT-AP JSON Schemas associated to the classes Dataset and DataService and JSON Schemas associated to any other Entity Data.

Next one, **geolocation** content (**M5**) is not related to any other FAIR principles. Geoproperties representation using GeoJSON format in the corresponding location property of an entity. In case of DCAT-AP, the property that can provide this information is the property spatial inside the class Dataset. Nevertheless, spatial refers to a geographic region or named place that is covered by the Dataset. We decided to differentiate between geo location and postal address through the properties Geo Location and Address Available. This information helps us to calculate afterwards the MELODA5 metrics in case of geolocation content. In the case of metadata for data services, we do not consider necessary the use of geolocation.

Continuing with the next dimensions, updating frequency of data (M6), is the rate at which the publication of a dataset or data service recurs. DCAT-AP already defines a property Frequency inside the Dataset class that provide that information. Additionally, the cross-domain NGSI-LD information model defines the temporal property modifiedAt with the information about the last modification executed in an NGSI-LD system. This property can be used to calculate the updating frequency of the data.

Next, **dissemination** (**M7**), tries to detail if the corresponding organization responsible of the dataset provide enough information about the dissemination process, if there are resources on updates or even if there is some kind of communications mechanism about the dissemination of the dataset. For MELODA5 metrics, we need to know which is the organization that it is responsible of the dissemination process (Dissemination Organization) and which kind of dissemination activities are developed in order to inform about a dataset (Dissemination). At present, there is no property in DCAT-AP that covers this requirement therefore, we will create them in the corresponding Dataset class in DCAT-AP. In the case of data service, although this is a property very interesting it is not currently clear how we can evaluate the dissemination activities of the organization that published or released a data service. We will need more activity to resolve how we express the dissemination of it.

Finally, **reputation** (**M8**), is a complex metric in MELODA5. The idea is that towards some forms of the Open Data portal managers, we can obtain a reputation score over the published open data in the open data portal. In WATERVERSE project, this questionary will be created for each of the Use Case that want to publish open data in order to get a cross evaluation of the Open Data information. Afterwards, we need two information, from one side which is the Organization that it is responsible

[21] https://json-schema.org/specification.html#meta-schemas

of the publication of the datasets (Reputation Organization) and what should be the value of this reputation (Reputation). At the moment, there is no property in DCAT-AP that covers this requirement therefore, we will create them in the corresponding Dataset class in DCAT-AP. In the case of data service, although this is also very interesting property, it is not clear how we measure the reputation of a company. Therefore, at the moment, we do not apply it. During the execution of WATERVERSE we will analyse this property to see how we can determinate the reputation of an organization in an objective way.

ID	MELODA5 Dimensions	FAIR	Implementation
M1	Licence	R1.1	• Included in schema • Included in DCAT-AP:dataset:license, DCAT-AP:distribution:license, DCAT-AP:dataservice:license
M2	Access to data	A1	• DCAT-AP:dataset:accessMechanism Property, DCAT-AP:DataService:endpointURL Property
M3	Technical format	I1	• JSON/JSON-LD • JSON Schema
M4	Data model (standardisation)	I1, I2, I3	• DCAT-AP:dataset:standardization Property. • DCAT-AP:dataset:standardizationSource Property. • DCAT-AP:dataset:validationSchema Property. • DCAT-AP:dataservice:validationSchema Property. • DCAT-AP:dataset:conformsTo Property. • DCAT-AP:dataservice:conformsTo Property. • $conformsTo keyword in JSON Schema
M5	Geolocation		• DCAT-AP:dataset:geolocation Property. • DCAT-AP:dataset:addressAvailable Property.
M6	Update frequency		• DCAT-AP:Dataset:frequency. Model:'dct:Frequency'
M7	Dissemination		• DCAT-AP:dataset:disseminationOrganization:Property • DCAT-AP:dataset:dissemination:Property
M8	Reputation		• DCAT-AP:dataset:reputationOrganization:Property. • DCAT-AP:dataset:reputation:Property.

Table 6: MELODA5 Dimensions (red font indicates to be implemented in task 4.2).

4.2 Defining WATERVERSE FAIR Services

4.2.1 Metadata 4 Assets

Water utilities and relevant stakeholders of the water sector are responsible for a variety of assets which are crucial in the operation and maintenance of processes that result in the delivery of safe and clean water to the public. In the water sector, an asset could range from a natural source where a utility extracts water, to machinery, such as pumps and valves, and constructed structures, such as tanks, used within the treatment and distribution parts of the urban water cycle. The level of influence

a water utility has also varies among assets. For example, for a natural source of water such as a lake, a water utility has minimal influence on the water levels and quality. Whereas for an asset such as a pump, a water utility has much more control in its conditions and operational protocols.

Irrespective of the influence, effective asset management is essential to ensure that infrastructure is maintained in good condition and risk of failures or breakdowns is reduced. This can involve a range of activities, from routine maintenance and inspections to more advanced techniques such as condition monitoring and predictive maintenance. Furthermore, system performances can be optimised to achieve more sustainable operations. Such activities are only possible when specific data on the assets are available. This is made possible through the deployment of a network of devices that can also be considered as crucial assets within the water sector such as data loggers, sensors, and remote monitoring systems. Such assets provide valuable information about the performance and condition of natural and man-made infrastructures.

With the Smart Data Models program, many data models have been developed within the Smart Water domain that adequately describe various water sector specific assets as entities, along with relevant properties that provide information on the associated metadata. Within these models, ETSI NGSI-LD or NGSIv2 descriptions are utilised. Table 7 illustrates a non-exhaustive list of assets along with the associated data models hyperlinked.

Asset	Relevant domain of Water Sector	Description
Pipe	Treatment, distribution networks	A generic pipe transporting water within a distribution network or between assets in a system (such as a treatment plant).
Valve	Treatment, distribution networks	An asset to regulate or control the flow of water through a pipe or tank.
Pump	Treatment, distribution networks	• An asset to transport water or other fluids through a distribution network or between assets in a system. • An asset to regulate and control the flow of water or other fluids within a system.
Tank	Treatment, distribution networks	• An asset to store water or other fluids in a water supply or treatment system. • An asset to regulate water pressure and improve the stability of the distribution network.
Junction	Treatment, distribution networks	A point where two or more pipes come together, and water can flow in different directions aimed to enable water to be routed to different parts of the network and help to maintain adequate flow rates and pressures.
Sluice Gate	Open channel management	A type of valve used in water management systems, such as irrigation, flood control, and wastewater treatment plants, to control the flow of water.
Reservoirs	Source, distribution networks	A large storage facility to store and supply water for human consumption, agricultural use, and industrial purposes.

Asset	Relevant domain of Water Sector	Description
Blower	Wastewater treatment	An asset to move air or gas through a system by creating a pressure difference (to provide air to the biological treatment processes in wastewater treatment plants).
WasteWaterTank	Wastewater treatment	An asset to hold and treat wastewater before it is discharged into the environment or sent for further treatment.
Device	All domains	A piece of equipment or a tool that is designed to perform a specific function or task: • In water treatment plants: chemical feeders, pumps, valves, filters, and disinfection equipment, which are all used to remove contaminants and ensure the safety and quality of the water. • In water distribution systems: pressure regulators, flow metres, and leak detection equipment, which are used to monitor and control the flow and pressure of water in the network. • Electronic or computerised equipment used in water treatment and distribution, such as sensors, control systems, and remote monitoring and control devices.

Table 7: Common Water Sector assets including their descriptions and defined smart data models (hyperlinked).

It must be noted that in the Smart Data Models program, not all water related assets that can be conceived have been described. The models were developed based on the workflow described in Figure 20 which relies on use cases through projects and contributions of members of the community. It is anticipated that more data models covering more identified assets will be described through projects such as WATERVERSE and other initiatives.

To ensure that the previous assets follow FAIR principles, the following metadata should be included:

1. **Findable**: metadata of each asset is provided that ensures its global and unique identification. As described in Section 4.1.1, this is achieved through the provision of a unique identifier for each asset. Furthermore, rich metadata is provided through the use of the "@context" property.
2. **Accessible**: metadata of each asset must use an open, free, and universally implementable protocol that allows authorization, being accessible by the unique identification and accessible through Smart Data Models program. As described in Section 4.1.2, this is achieved through the provision of a unique identifier for each asset, the use of ETSI NGSI-LD protocol and the oAuth2 authorization.
3. **Interoperable**: metadata of each asset must obey the same data structure, as well as any standards or conventions that are used to describe the asset. As described in Section 4.1.3, this is achieved through the use of the same data format (JSON/JSON-LD) and by following MELODA5 dimensions.
4. **Reusable**: metadata of each asset must include rich descriptions, be accessible and have a clear provenance. As described in Section 4.1.4, this is achieved through the use of the "@context" property, an open license, and a versioning system (a version history).

In addition, metadata for water assets can also be ensured to be MELODA5 compliant. Some of the MELODA5 dimensions are incorporated by ensuring the FAIR principles guidelines are maintained, as discussed in Section 4.1.5. For the MELODA5 dimensions not yet covered, this can be achieved by providing certain optional properties to the entity data model that provide this relevant information. The following properties can be considered:

- **Geolocation**: A property to provide the location of a given water asset or source of the data. This includes the provision of the coordinates. This is relevant in the operation and maintenance of water assets. For example, within a water distribution network, the location of pipes, junctions and valves are highly relevant. Within ETSI NGSI-LD, GeoJSON formatting standards are used.
- **Update frequency**: A property to describe how often a given information is updated, as described in Section 3.6. From a water asset point of view, this typically would refer to the resolution of the data of a given property or parameter measured for the asset. For example, for a pump, the power generated is measured. The values of the data could be measured every minute by a device, which would then be the corresponding update frequency for this given dataset.
- **Date Modified**: A property that will indicate the date and time when the metadata provided for a given water asset was last modified. This is crucial in ensuring version controlling and monitor updates of the entity models defined for a given water asset. For example, a specific device that previously measured one property (such as temperature of the water in a tank) has subsequently been calibrated to also measure the pH. This will require the addition of a new property pH to the data model. Providing meta information on the date of this modification increases transparency and efficiency. Another example could be the update of property that describes the last time it was calibrated (dateLastCalibration).

Provided below, are concrete examples of two assets - a pump (Table 8), and a device measuring the flow created by a pump (Table 9**Σφάλμα! Το αρχείο προέλευσης της αναφοράς δεν βρέθηκε.**). Some common properties of such entity models have been described. The examples illustrate how information defined within the properties of the entity models lead to the compliance with the FAIR principles and MELODA5 dimensions which have been listed, for each of the properties that either provide information on the dataset or the supporting metadata, for a given asset.

Property/ Relationship Key	Property/Relationship Value	FAIR	MELODA5 Dimension	Mandatory / Optional
id	urn:ngsi-ld:Pump:Pump001	F1, F3, A1	-	Mandatory
type	Pump	I1, I2	M3	Mandatory
endsAt	urn:ngsi-ld:Reservoir:Reservoir001	-	-	Optional
power	{value: 100, unitcode: KWT, observedBy: }	I3	-	Optional
flow	{value: 20, unitcode: G51, observedBy: urn:ngsi-ld:Device:Device056}	I3	-	Optional
location	{value: { type: Point, coordinate: [-4.6, 37.5]}}	-	M5	Optional

Property/ Relationship Key	Property/Relationship Value	FAIR	MELODA5 Dimension	Mandatory / Optional
updateFrequency	P1M	-	M6	Optional
modifiedAt	2023-03-25T13:05:00Z	-	M6	Optional
@context	["https://raw.githubusercontent.com/smart-data-models/dataModel.WaterDistributionManagementEPANET/master/context.jsonld"]	F2, F4		Mandatory

Table 8: Example of a Pump asset entity model including common properties and links with FAIR principles and MELODA5 dimensions.

Property/ Relationship Key	Property/Relationship Value	FAIR	MELODA5 Dimension	Mandatory / Optional
id	urn:ngsi-ld:Device:Device056	F1, F3, A1	-	Mandatory
type	Device	I1, I2	M3	Mandatory
controlledAsset	urn:ngsi-ld:Pump:Pump001	F2	-	Optional
controlledProperty	value: [power]	I3	-	Optional
deviceState	value: ok	I3	-	Optional
location	{value: { type": "Point, coordinate": [-4.6, 37.5]}}	-	M5	Optional
updateFrequency	P1M	-	M6	Optional
modifiedAt	2023-05-15T15:33:00Z	-	M6	Optional
@context	["https://uri.etsi.org/ngsi-ld/v1/ngsi-ld-core-context.jsonld", "https://raw.githubusercontent.com/smart-data-models/dataModel.Device/master/context.jsonld"]	F2, F4	-	Mandatory

Table 9: Example of a Device asset entity model including common properties and links with FAIR principles and MELODA5 dimensions.

Over and above the mentioned FAIR+MELODA5 compliance as listed in Table 8 and Table 9, the use of JSON/JSON-LD format within the Smart Data Models programme, lead to the following with the components of the Accessible and Reusable principle (described in Section 4.1.2 and 4.1.4) and other MELODA5 dimensions M1-M4 (described in Section 4.1.5).

4.2.2 WATERVERSE FAIR Implementation Profiles

The WATERVERSE Implementation Profiles arise to accelerate broad community convergence on MELODA5 and FAIR Implementation principles. In fact, the implementation profiles are related to a collection of FAIR and MELODA5 implementation choices that take into account the community of practice related to each FAIR and MELODA5 Dimensions. These principles concern implementation strategies that can be used as a basis for optimising the reuse of existing FAIR and MELODA5 enabling resources and interoperability within and between (water) domains.

In order to facilitate the collection of feedback from users and to allow an analysis of whether and to what extent the FAIR and MELODA5 dimensions are implemented, a questionnaire depicted in the table was created (Table 10).

Code	FAIR Principles	MELODA5 Dimensions	Questions	FAIR enabling resources types
WIP01	F1		What globally unique, persistent, resolvable identifiers do you use for metadata records?	Identifier type
WIP02	F1		What globally unique, persistent, resolvable identifiers do you use for datasets?	Identifier type
WIP03	F2		Which metadata schemas do you use for findability?	Metadata schema
WIP04	F3		What is the technology that links the persistent identifiers of your data to the metadata description?	Metadata-Data linking mechanism
WIP05	F4		In which search engines are your metadata records indexed?	Search engines
WIP06	F4		In which search engines are your datasets indexed?	Search engines
WIP07	A1.1	Access to data	Which standardised communication protocol do you use for metadata records?	Communication protocol
WIP08	A1.1	Access to data	Which standardised communication protocol do you use for datasets?	Communication protocol
WIP09	A1.2	Access to data	Which authentication & authorisation technique do you use for metadata records?	Authentication & authorisation technique
WIP10	A1.2	Access to data	Which authentication & authorisation technique do you use for datasets?	Authentication & authorisation technique
WIP11	A2		Which metadata longevity plan do you use?	Metadata longevity

Code	FAIR Principles	MELODA5 Dimensions	Questions	FAIR enabling resources types
WIP12	I1	Data Model (standardization)	Which knowledge representation languages (allowing machine interoperation) do you use for metadata records?	Knowledge representation language
WIP13	I1	Data Model (standardization)	What globally unique, persistent, resolvable identifiers do you use for metadata records?	Knowledge representation language
WIP14	I2	Data Model (standardization)	What globally unique, persistent, resolvable identifiers do you use for datasets?	Structured vocabularies
WIP15	I2	Data Model (standardization)	Which metadata schemas do you use for findability?	Structured vocabularies
WIP16	I3	Data Model (standardization)	What is the technology that links the persistent identifiers of your data to the metadata description?	Metadata schema
WIP17	I3	Data Model (standardization)	In which search engines are your metadata records indexed?	Data schema
WIP18	R1.1	License	In which search engines are your datasets indexed?	Data usage licence
WIP19	R1.1	License	Which standardised communication protocol do you use for metadata records?	Data usage licence
WIP20	R1.2		Which standardised communication protocol do you use for datasets?	Provenance model
WIP21	R1.2		Which authentication & authorisation technique do you use for metadata records?	Provenance model
WIP22		Technical standard/format	Which authentication & authorisation technique do you use for datasets?	-
WIP23		Technical standard/format	Which metadata longevity plan do you use?	-
WIP24		Geolocation	Which knowledge representation languages (allowing machine interoperation) do you use for metadata records?	-
WIP25		Update frequency	Which knowledge representation languages (allowing machine interoperation) do you use for datasets?	-
WIP26		Update frequency	Which structured vocabularies do you use to annotate your metadata records?	-

Code	FAIR Principles	MELODA5 Dimensions	Questions	FAIR enabling resources types
WIP27		Update frequency	Which structured vocabularies do you use to encode your datasets?	-
WIP28		Dissemination	Which models, schema(s) do you use for your metadata records?	-
WIP29		Data model	Which models, schema(s) do you use for your datasets?	-

Table 10: Questions relating to WATERVERSE Implementation Profiles and FAIR and MELODA5 principles.

The questionnaire is composed by 28 questions and it look as a table that has five columns:

- "Code" that is a unique identifier of question;
- "FAIR principles". The FAIR principle identification code (if the question is related to FAIR and not MELODA5). It is represented with a letter F, A, I or R that corresponds to the FAIR principle with which it is associated and a number that determines its sequence.
- "MELODA5 dimensions". The MELODA5 dimension to which the question refers.
- "Questions", which represents the question addressed to the user.
- "FAIR enabling resource types" that represents the resources' types to which the question is referred (e.g, Metadata, Search Engine, …).

The answers to the questionnaire can be of various types. To make the questions more accessible to users, in some cases, sample answers have been provided. Specifically:

- WIP01 → This is referred to as an identifier type used by pilot to identify metadata records. It could be for example PURL or DOI and so on…
- WIP02 → Also here it is referred to as an identifier type that the pilot uses to uniquely identify the dataset.
- WIP03 → This is a question related to the Metadata schemas used by the pilot for findability of the dataset.
- WIP04 → The question refers to the Metadata-Data linking mechanism adopted by the pilots to link dataset in the metadata file.
- WIP05 → This is a technical question related to the search engines that pilots use to federate their dat (e.g, CKAN, Google Scholar and so on…).
- WIP06 → Same as above but for metadata.
- WIP07 → This is related to the communication protocol that pilots use to access the metadata (e.g. Web access or unique URL parameters to dataset, Web Access unique with parameters to single data, API or query language, …).
- WIP08 → Same as above but for dataset.
- WIP09 → Related to authentication & authorisation technique used by the pilots to access metadata (e.g. oauth2, saml, …).
- WIP10→ same as above but for datasets.
- WIP11-Which metadata longevity plan do you use?
- WIP12 → Related to the knowledge representation language that pilots use for metadata records (e.g. OWL, RDF, SKOS, JSON-LD, …).
- WIP13 → same as above but for datasets.

- WIP14 → Related to the structured vocabularies used by the pilots to annotate metadata records (e.g., DCAT-AP);
- WIP15 → Same as above but for datasets.
- WIP16 → Related to the schema that pilots use to standardise metadata (e.g., own data model standardisation; Own ad hoc data model standardisation published (harmonisation); Local standardisation; Global standardisation, …).
- WIP17 → same as above but for data.
- WIP18 → referred to the metadata usage licence, e.g., private use (copyright); non-commercial use (GPL, …); commercial reuse or no restrictions (MIT, Apache, ...), etc.…
- WIP19 → same as above but for datasets.
- WIP20 → This is related to the provenance model used for metadata.
- WIP21 → Same as above but for datasets.
- WIP22 → Related to the format used to store data, e.g. JSON, XML, CSV, PDF, etc.
- WIP23 → Same as above but for metadata.
- WIP24 → Related to the geolocation information (eventually) included in the dataset, e.g. no geographic information; simple or complex text field; coordinates or full geographical information; one text field or several text fields; etc.
- WIP25→ Related to the frequency updates of metadata, e.g. Longer than 1 month; Monthly. Updating period ranges from 1 month to 1 day; Daily. Updating period ranges from 1 day to 1 hour; Hour. Updating period ranges from 1 hour to 1 minute; Seconds. Updating period is lower than 1 minute.
- WIP26→ same as above but for dataset.
- WIP27 → If the frequency update of dataset information is included in the metadata, yes or not.
- WIP28 → Related to the dissemination approach for the dataset, e.g. dissemination/communication not systematic; available resources on updates (i.e., RSS feed); Proactive dissemination / push dissemination (information automatic and timely).

The questionnaire just described will be administered to the six WATERVERSE pilots and results will be reported in D4.2.

4.2.3 FAIR Digital Objects

The FAIR Digital Objects (or WATERVERSE Data Point) is a concept related to the metadata of a dataset. In fact, it stores information about the dataset and aims to give anyone the power of putting their own data on the web. In order to do that, a set of definitions and components should be defined. This chapter presents the Fair Digital Objects and describes the Fair Digital Object Frameworks (FDOF) that is a framework which defines a model to represent objects in a digital environment and a mechanism to create, maintain and (re)use these objects.

In general, a Digital Object is a sequence of bits that represents an informational unit such as a document, a file, a video, etc. A FAIR Digital Object is when the digital object is represented according to FAIR principles: identified by a globally-unique, persistent and resolvable identifier with a predictable resolution behaviour, described by related metadata records.

In order to implement the described FAIR principles and thus realise a FAIR system, an infrastructure is needed that can support the handling of digital objects according to the requirements defined by the FAIR principles: The FAIR Digital Object Framework (FDOF).

So, the FDOF is the basis on which a FAIR system can be developed. In particular, it aims at tackling core issues raised by the FAIR principles regarding the optimal reuse of digital objects. Starting from this framework, the obtained system and data can better interoperate with the existing systems. It is important to specify that FDOF does not aim to replace existing protocols and standards (such as the WEB) but aims to be complementary to them in order to integrate them with FAIR principles).

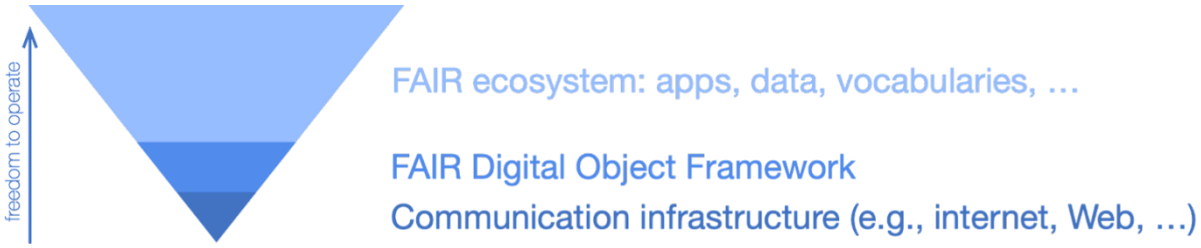

Figure 21: Defining the FAIR Digital Object Framework

For two or more systems to interoperate and talk to each other, they must follow rules and standards that enable them to do so. If a system wanted to communicate with a black box, it could not do so because it would not know how to interpret the data or interrogate the component. Unless both follow guidelines and standards. Interoperability between systems is in fact only guaranteed when all resources respect the defined standards and guidelines. The FDOF defines some of these rules and guidelines required by the FAIR principles and, to the extent that the objects involved conform to its specifications, a higher level of FAIRness and, thanks on that the interoperability is achieved. The set of guidelines defined by the FDOF mainly cover three areas:

- Predictable identifier resolution behaviour.
- Method to access more information about the object given its identifier.
- Object typing system.

In general, a FDO is a Digital Object that is identified by a globally unique, persistent, and resolvable identifier, characterised by the FDOF typing system and described by metadata records (Figure 22).

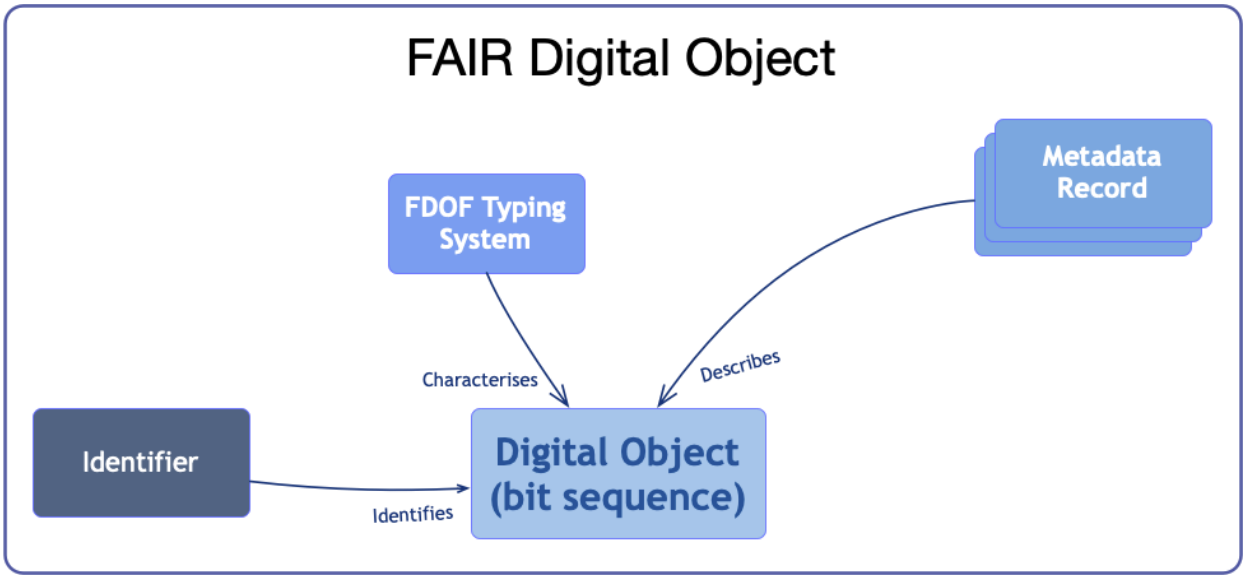

Figure 22: FAIR Digital Objects Attributes

'Predictable Identifier Resolution Behaviour' refers to the fact that identifiers of digital objects behave predictably, so that artificial agents can know what to expect when an identifier is resolved. Currently, that is not always the case. For example, it is possible to use a DOI, used by IEEE to identify published documents, to construct a URL that points to the HTML page of the identified document. But on that HTML page there are many other URLs (authors of the article, advertisements, and so on) of which only one corresponds to the PDF of the document being searched for. An intelligent client would not be able to find it except after many attempts and using non-trivial logic. To tackle this issue, the FDOF defines a predictable resolution behaviour based on an identifier named FDOF's Identifier Record (FDOF-IR).

A FDOF is a specific type of metadata, containing information about:

- The object's type
- The object's metadata record(s) and
- The object's location(s).

These are the minimal information required by the infrastructures/applications in order to interoperate with it. In fact, knowledge of these three pieces of information means that the client can identify the type of object (through the reference to the type of object), operate directly on the object (through the reference to the location of the object) and obtain further information on the object (through the references to the object's metadata records): important information when the client has just discovered an object identifier and knows nothing about the object.

However, the FAIR-IR not only defines the information that is to be included in the ID but also defines the format in which it is to be represented in order to guarantee predictability of the metadata presentation format. The FAIR-IR defines that it must be presented as RDF (that explicitly provides semantic annotations together with the data content), minimally using the Turtle and JSON-LD syntaxes. In this way, the client not only knows the needed information of the data (needed metadata) but also knows how to extract this needed information starting from the FAIR-IR.

Since FDOF does not aim to replace existing systems, as mentioned earlier, it has defined a set of mechanisms and protocols useful for integration with existing systems. Not all the resources are already in place using the FAIR-IR identification. For this reason, the FDOF resolution behaviour can also offer methods to directly request the information contained in the FDOF-IR (object type, object metadata record and object location). In order to do that FDOF defines three different ways:

- FDOF resolution protocol, that defines a sets of methods that can be used to retrieve the information (GET, GETIR, GETMETADATA and GETTYPE) related to the object, the FDOF-IR, the metadata and the object types.
- FDOF-P using HTTP accept headers that defines a sets of header parameters that can be used to access the object, the FDOF-IR, the metadata and the object types.
- FDOF-P using HTTP Signposting.

The object's metadata record(s) are very important parts of FDOF. In order to have FAIR systems and data it is required that we have rich metadata to describe them. As mentioned before, the default representation of the metadata record must be RDF. Furthermore, the FDOF also defines that the response to the METADATA method must be a Linked Data Platform (LDP) container representing the maximal collection of metadata records for the object.

The client application can directly request the metadata record(s) of the object from its FDOF-IR identifier using one of the methods defined by FDOF protocols (FDOF-P). This method is named GETMETADATA and returned, as mentioned, a Linked Data Platform (LDP) container representing the maximal collection of metadata records for the object (Figure 23).

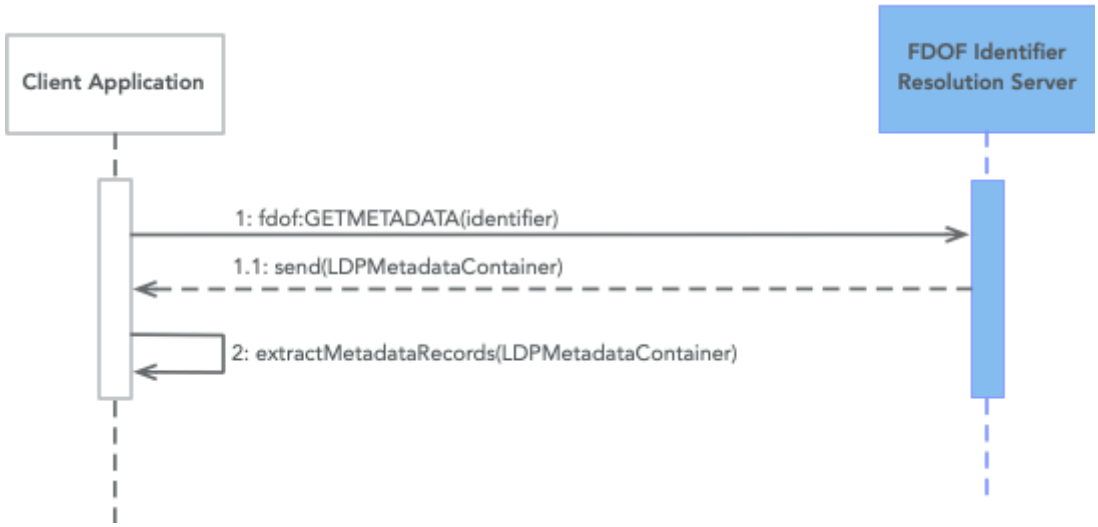

Figure 23: Diagram Flow for request Metadata

There is also another method to get metadata by using a Metadata Source Server. In particular, by using another method defined by FDOF-P, the GETIR method (Figure 24), a client may request the representation of FDOF-IR, which must be RDF, given an identifier of the Digital Object. This request was introduced in order not to disrupt what currently exists on the Internet and to smoothen the integration of FAIR principles into what currently already exists. Once the FAIR-IR is obtained the clients can search in the IR for the references of the metadata records.

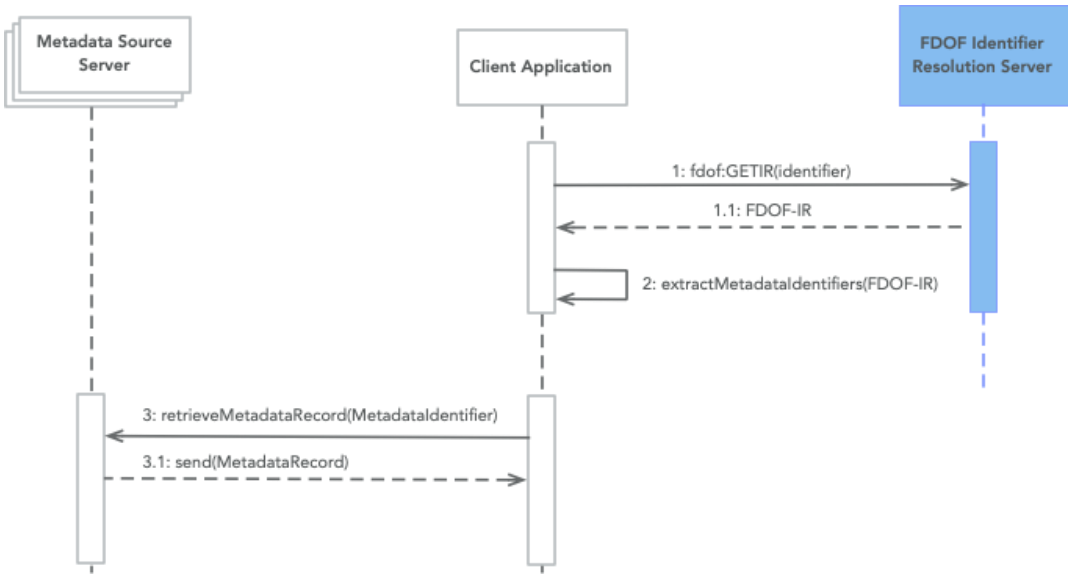

Figure 24: Alternative Diagram Flow for request Metadata

The last point is the object typing system. Understanding the type of digital object is very important for the client in order to understand what kind of operations it can do. The question of what can be done with a given object refers to the types of operations that can be applied to the object and is related to the type of object in question but also to the permissions one has to perform these operations. All these considerations are considered by the FDOF which defines:

- The data format (encoding format, such us JSON, XML, …).
- The type of the digital object with respect to its informational function.
- The entities represented by the DO.
- The operation applicable to the DO.
- The operation allowed on the DO.

In conclusion, the FDOF typing system can determine specific metadata schemas for specific FDO types. This schema must include the most important information needed for the FDO types, so, a minimal set of properties. This guarantees a level of predictability so that a client can expect these properties to be available as descriptors of a FDO of a given type.

5.0 Tools and Resources

5.1 Data Summary

There is a wide variety of data being collected across the six pilot case studies coming from various sources and in a range of differing formats. Table 11 shows a summary overview of some of the data being gathered across these six case studies.

Case Study	Unique IDs	Data Collected/Processed
Netherlands	Fn001, Fn002, Fn003, Fn004, Fn005, Fn006, Fn007, Fn008, Fn048, Fn050, Fn052, Nfn001, Nfn002, Fn059, Dfn001, Dfn002, Fn060, Fn063, Fn064, Fn065, Fn067, Fn068, Fn069, Nfn004, Fn070, Fn076, Nfn005, Nfn006, Nfn007, Fn085, Fn086, Fn087	Wind direction, wind speed, chloride, conductivity, level, value
Germany	Fn009, Fn010, Fn011, Fn012, Fn013, Fn014, Fn015, Fn016, Fn050, Fn051, Nfn001, Fn053, Fn058, Fn059, Dfn001, Dfn002, Fn063, Fn064, Fn065, Fn066, Fn067, Fn068, Nfn004, Fn071, Fn076, Nfn005, Fn086, Fn087	Water level, precipitation, precipitation forecast, water level forecast, soil moisture, soil moisture forecast, ground water sensor, water flow volume sensor, temperature sensor
Cyprus	Fn017, Fn018, Fn019, Fn020, Fn021, Fn022, Fn023, Fn048, Fn049, Fn050, Fn051, Fn052, Nfn001, Fn054, Fn055, Fn058, Dfn001, Dfn002, Fn063, Fn064, Fn065, Fn067, Fn068, Fn069, Nfn004, Fn072, Fn076, Fn077, Fn078, Fn079, Fn080, Fn081, Fn082, Fn083, Fn084, Nfn005, Nfn006, Nfn007, Fn086, Fn087	Flow, pressure, level, chlorine
United Kingdom	Fn024, Fn025, Fn026, Fn027, Fn028, Fn029, Fn030, Fn031, Fn032, Fn047, Fn048, Fn049, Fn052, Nfn001, Fn056, Dfn001, Dfn002, Fn063, Fn064, Fn065, Fn067, Fn068, Fn069, Nfn004, Fn073, Fn074, Fn076, Nfn005, Nfn007, Fn086, Fn087	Storm overflow analog sensors, storm overflow spill start stops, other SCADA data, Environment Agency river / tide gauge data for gauges local to study area, Environment Agency rain gauge data for rain gauges local to study area, Environment Agency Water Quality API, WQ sonde data (Temperature, Conductivity, pH, Ammonium, Turbidity and Dissolved Oxygen, Chlorophyll), DTN rain radar / weather API (subject to review of current contract to determine if we can release for this project)
Spain	Fn033, Fn034, Fn035, Fn036, Fn037, Fn038, Fn039, Fn040, Fn041, Fn047, Fn049, Fn050, Fn051, Nfn001, Dfn001, Dfn002, Fn061, Fn062, Fn063, Fn064, Fn065, Fn066, Fn067, Fn068, Fn069,	Volume of water supplied, registered water volume, clients, investment in social initiatives, volume of groundwater extracted, volume of surface water extracted, volume of WTP water, volume

Case Study	Unique IDs	Data Collected/Processed
	Nfn004, Fn076, Nfn005, Nfn006, Nfn007, Fn086, Fn087	of groundwater purchased, volume of surface water purchased, volume of desalinated water purchased, energy consumption in WTP, energy production in WTP, energy consumption in WWTP, energy production in WWTP, energy consumption in distribution networks, energy production in distribution networks, energy consumption in sewage networks
Finland	Fn042, Fn043, Fn044, Fn045, Fn046, Fn048, Fn050, Nfn001, Fn057, Dfn001, Dfn002, Fn063, Fn064, Fn065, Fn067, Fn068, Nfn004, Fn075, Fn076, Nfn005, Nfn007, Fn086, Fn087	Area ID, Area location, FMI WFS, FMI radar data, update_type, tenant_indentification, synch_status, last_synch_by, last_synch_date, Tenant_url, log_text, data_source, status, precipitation_data_realtime, precipitation_data_daily, Meter reading, Remote metre reading, Graph data, Report data, Result data, Known condition factor, Critical condition factor, Estimated condition factor, Sea water level, Leak area, Signal data, Condition data, Risk alert

Table 11: Summary sample of data being collected over the six pilot case studies derived from D2.1 - WATERVERSE WDME design.

Pathways from the respective case studies are outlined in more detail within D2.1. “WATERVERSE WDME design” as part of the WDME.

5.2 FAIR Services identified

This section presents a list of tools that WATERVERSE wants to provide for the FAIR principles and MELODA5 dimensions. They correspond to the first iteration of services identified by the partners that will be defined in Task 4.2 and implemented in Task 4.3. The purposes of these services are:

- Facilitate the creation of Data Models taking into account the WATERVERSE FAIR principles including FAIR principles and MELODA5 dimensions.
- Calculate the Metadata Quality Assurance (MQA) score over the datasets developed in the WATERVERSE project in order to improve the quality of the metadata generated in the project.
- Manage the data models versions and keep a trace of them to facilitate the checking of the metadata structure of the data models to be used.
- Facilitate the export of the information to the CKAN Open data portals.
- Facilitate the generation of example file formats of the data models.

During the initial phase of identification of services, some of them have already been selected by some partners to be implemented as soon as possible in ongoing task (Task 4.3) and some of them have already started to be developed as shown in Table 12.

Service Identification	Rationale	Partner Involved
S01	Automatic report of FAIR Metrics for FAIR Data defined in T4.1 taking into account the research community for FAIR Digital Objects.	
S02	Automatic report of FAIR Metrics for FAIR Services, it needs more thought. Existing frameworks like CoreTrustSeal (CTS) for repository certification or FAIR Evaluation Services should be used and adapted rather than develop from scratch new schemes based solely on FAIR.	
S03	The WATERVERSE FAIRness Tools will be used in T4.2 to calculate the FAIR DMP and compare it in three moments during the execution of the project.	
S04	Tool to check the extended version of DCAT-AP with FAIRness principles in published datasets in WATERVERSE.	EGM
S05	Certification of the validity of a payload against a specific version of a SDM data model which include the FAIRness principles define in the T4.1	ENG
S06	SDM service to export data model description (using DCAT-AP) into data spaces or other data portals/resources.	FIWARE / EGM
S07	Creation of a SDM database of data models' versions with information of the current version of the data model and the corresponding hash code in GitHub repository.	FIWARE
S08	Develop a Visualization Tool to Facilitate the visualization of the Data Model structure of the data model with the corresponding WATERVERSE FAIR Guidelines.	ENG
S09	Automatic reporting of status and updates of the Data Models to check if they were fully compliant with the WATERVERSE FAIR Guidelines.	CERTH
S10	Export a Data Model into SQL format to facilitate the automatic creation of SQL DB to attract companies in the water sector that are not interested in adopting ETSI NGSI-LD but want to adopt the Data Models in their developments.	FIWARE
S11	Automation and extension of Quality Testing Data Models to facilitate the creation of the new Data Models including the WATERVERSE FAIR Guidelines created in the project.	
S12	Possibility to import/export the modified data models into CKAN Portals through the use of CKAN API.	EGM
S13	All services should be integrated in GitHub through GitHub approach through GitHub CI/CD Actions or Pipelines to be	

Service Identification	Rationale	Partner Involved
	deployed inside the Smart Data Models program following the Manifesto for Agile Standardization (MAS).	
S14	Evaluation of automatic improvement of FAIR data based on the results obtained by the WATERVERSE FAIRness Tools	
S15	Creation of a service to validate automatically the entity data and datasets information based on FAIR principles	EGM
S16	(Optional) Report to contributors when new models are created/updated in their subject.	FIWARE
S17	WATERVERSE Digital Object service based on FAIR Digital Object plus MELODA5 dimensions.	EGM
S18	Tool to calculate the metric value of the MQA based on the FAIR+MELODA5 dimensions.	
S19	Definition of a service to allow accessing the JSON Schemas associated to a Data Model based on the id.	
S20	Automatic generation of RDF Turtle file format examples of the data models.	
S21	Create of simple data models using a template to facilitate the process for companies or persons not familiar with JSON/JSON Schema	FIWARE

Table 12: WATERVERSE FAIR services – 1st iteration

6.0 European Added Value

We can understand the EU added value for the project as the value resulting from an EU project which is additional to the value that would have been created by individual states members alone. This means that there are several areas of interest to cover this added value:

- Enhancing and extending the data models provided by the Smart Data Models initiative with FAIR and MELODA5-related metadata improves their adoption by European Industry and SMEs that can then improve their efficiency by leveraging well-qualified data models.
- Providing datasets enriched with FAIR and MELODA5-related metadata greatly improves their overall quality and makes them easier and more secure to integrate and use by European Industry SMEs, that can then deliver more competitive offerings based on the added value of FAIR data.
- Developing FAIR services tools allows to ensure the quality and robustness of FAIR and MELODA5 metrics over time. It is indeed often difficult to maintain such indicators over the lifetime of a data model or dataset and these tools will greatly ease the process and provide some extra tooling will give European Industry and SMEs the confidence to use the provided data models and datasets.
- More generally, the use of the NGSI-LD specification defined by ETSI as well as the use of the Smart Data Models facilitate the excellence and operational efficiency of the European partners involved in the project.
- The collaboration of the FIWARE Foundation and specially the FIWARE Community gives the opportunity to increase the visibility of the work achieved in the project. Additionally, this visibility beyond EU countries increases the geographic scope of the partners involved in the design and development of the FAIR services given the advantage offered by the FIWARE Marketplace to develop business beyond EU countries.

7.0 Conclusion and Perspectives

As the first deliverable of WP4 “FAIRness assessment in the water industry” this document outlined the work covered in Task 4.1 “Definition of FAIR Digital Objects and FAIR Ecosystem concepts” showing the importance of adopting FAIR principal guidelines coupled with MELODA5. The six pilot case studies participating within the WATERVERSE project will be following these guidelines in the assessment of the FAIRness of their data at the beginning stages of the project, following said guidelines. The FAIRness will be evaluated towards the end of the project to evaluate how the FAIRness has improved. To achieve this evaluation the guidelines, outline herein will be outlined in means of their technical implementation in the Task 4.2 “Implementation of the recommended FAIR principles” with details relating to this task being summarised within the complementary Deliverable 4.2 “FAIR Data Management Plan” that is an evolving document of the course of the project with versions being submitted at months 8, 18, and 36.

8.0 REFERENCE

[1] M. D. Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Sci Data*, vol. 3, no. 1, p. 160018, 2016, doi: 10.1038/sdata.2016.18.
[2] A. Abella, M. Ortiz-de-Urbina-Criado, and C. De Pablos-Heredero, "Meloda 5: A metric to assess open data reusability," *El Profesional de la Información*, vol. 28, Jan. 2020, doi: 10.3145/epi.2019.nov.20.
[3] H. van Vlijmen *et al.*, "The Need of Industry to Go FAIR," *Data Intell*, vol. 2, no. 1–2, pp. 276–284, May 2020, doi: 10.1162/dint_a_00050.
[4] M. Jahanddideh-Tehrani, O. Bozorg-Haddad, and I. N. Daliakopoulos, "The Role of Water Information and Data Bases in Water Resources Management," in *Essential Tools for Water Resources Analysis, Planning, and Management*, O. Bozorg-Haddad, Ed., Singapore: Springer Singapore, 2021, pp. 59–83. doi: 10.1007/978-981-33-4295-8_3

Co-funded by
the European Union