



MARCLAIMED

Data Management Plan v1

Deliverable D7.1

WP7

Author: Luc GASSER

Date: 25/07/2024



**Funded by
the European Union**

GRANT AGREEMENT NUMBER	101136799		
ACRONYM / FULL TITLE	MARCLAIMED: integrated Decision Support Tool for Reliable and Affordable Application of Managed Aquifer Recharge with Alternative Water Resources in River Basin and Drought Management Plans		
START DATE	01/02/2024	DURATION	36 months
END DATE	31/01/2027		
PROJECT URL	https://marclaimed.eu/		
DELIVERABLE	D7.1		
WORK PACKAGE	WP7		
CONTRACTUAL DATE OF DELIVERY	31/07/2024		
ACTUAL DATE OF DELIVERY	25/07/2024		
NATURE	DMP	DISSEMINATION LEVEL	PU
LEAD BENEFICIARY	EGM		
RESPONSIBLE AUTHOR	Luc GASSER		
CONTRIBUTIONS FROM	Luc GASSER (EGM) Franck LE GALL (EGM) Simone PIO (AGDA) Daniela BENEDICTO VAN DALEN (AW) Sara ESPINOSA MARTINEZ (CETAND)		
ABSTRACT	This document is the first version of the Data Management Plan (DMP) of the MARCLAIMED project. Its goal is to monitor the generated data with respect to their privacy and confidentiality, ensure that the ethical standards for data use, storage and share are applied throughout the project, ensure that the technical standards are applied for and determine which data can be shared in the open data initiative. In relation with those objectives, MARCLAIMED will do its best to apply the Guidelines on FAIR data to its activities. This document will follow the structure proposed by the EC Data Management Plan template for Horizon Europe programme. It presents the main data identified at this stage of the project and highlights how the project will follow the FAIR guidelines in relation to those data.		

Disclaimer

Any dissemination of results reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

© **MARCLAIMED Consortium, 2024**

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both. Reproduction is authorised provided the source is acknowledged.

Copyright message

TABLE OF CONTENTS

EXECUTIVE SUMMARY	6
1. DATA SUMMARY	7
Data re-use	7
Types and format of data	7
Purpose of data generation	7
Expected size	8
Data origin	8
Data utility	8
2. FAIR DATA	9
Making data findable, including provisions for metadata	9
Data identification	9
Data discovery	9
Metadata discovery and indexation	10
Making data accessible	10
Trusted repository	10
Data identification	10
Data public availability	10
Accessibility	11
Restrictions	11
Identity verification for data access	11
Data access committee	12
Metadata public availability	12
Making data interoperable	12
Interoperability	12
Mapping to commonly used ontologies	12
Increase data re-use	12
Data analysis validation	12
Freely available data	13
Third party usability	13
Standards	13
Data quality assurance	13
3. OTHER RESEARCH OUTPUTS	14
4. ALLOCATIONS OF RESOURCES	15
Costs	15

Data management responsible	15
Long term preservation	15
5. DATA SECURITY	16
Provisions	16
Storage	16
6. ETHICS	17
Possible ethical issues	17
Consent for data sharing and long-term preservation	17
7. OTHER ISSUES	18
8. CONCLUSIONS	19

ACRONYMS:

ABM (Agent-based modelling),
AI (Artificial Intelligence),
API (Application Programming Interface),
ASR (Aquifer Storage and Recovery),
ASTR (Aquifer Storage, Transfer and Recovery),
AWARE (Available WATER REMaining per area),
AWR (Alternative Water Resources),
AWRSS (Alternative Water Resources Supply Systems)
BM (Business Model),
CDP (Communication and Dissemination Plan),
CoP (Community of Practice)
DMP (Data Management Plan)
DSS (Decision Support System)
DT (Digital Twin)
DWTP (Drinking Water Treatment Plant)
EU (European Union)
EC (European Commission)
HLEG AI (High Level Expert Group on Artificial Intelligence)
IDST (Integrated Decision Support Tool)
IIP (Intellectual Property)
IPR (Intellectual Property Rights)
KPI (Key Performance Indicator)
MAR (Managed Aquifer Recharge)
ML (Machine Learning)
MLMs (Machine Learning Methods)
PIML (Physics-Informed Machine Learning)
PoMs (Programmes of Measures)
RCP (RepresentativeConcentration Pathways)
RES (RESults)
ROI (Return Of Investment)
SAT (Soil-Aquifer Treatment)
SO (Specific Objectives)
SoA (State of the Art)
SUS (System Usability Scale)
TG (Target Group)
TRL (Technology Readiness Level)
WFD (Water Framework Directive),
WP (Work Package)
WWTP (Waste Water Treatment Plant)

EXECUTIVE SUMMARY

This document is the first version of the Data Management Plan (DMP) of the MARCLAIMED project. This deliverable is produced early in the project (M6, 31/07/2024) and a second and third versions of this Data Management Plan will be delivered on M24 (31/01/2026) and M36 (31/01/2027) respectively. This DMP is being produced as part of task “T7.4: Data requirements planning and management”, which is part of WP7 - Project Coordination and Management.

The goal of this task and this deliverable is to monitor the generated data with respect to their privacy and confidentiality, ensure that the ethical standards for data use, storage and share are applied throughout the project, ensure that the technical standards are applied for and determine which data can be shared in the open data initiative. In relation with those objectives, MARCLAIMED will do its best to apply the Guidelines on FAIR data to its activities. This document will follow the structure proposed by the EC Data Management Plan template for Horizon Europe programme.

At this stage of the project, all of the data that have been identified to be generated within the project are either data belonging to the demo site that won't all be shareable or data involving personal information that must be kept confidential. However, for now, few data have been identified to be publicly shareable and are described in this DMP. Moreover, all the data processed and generated within the project will use the NGSI-LD standard, which provides a strong interoperability, as described later in the document.

1. Data Summary

This section will provide an initial description of the data, their types, formats, purposes, expected size, origin and utility.

A large amount of the data collected during the project will be from the demo site which are the followings:

- Demo site 1 – Mediterranean coastal aquifer: combined scheme (ASTR+ASR) of MAR with reclaimed water by infiltration ponds and mix of surface water along with reclaimed water by a deep injection well; Location: Barcelona, NW Spain. Owner: AIGUES and ACA, Tech Partner: CETBCN. Data management for this DS will be handled by AIGUES.
- Demo site 2 – Atlantic coastal aquifer: ASTR scheme of SAT-MAR with secondary treated wastewater using four infiltration basins. Location: Comporta, Portugal. Owner: AgdA, Tech Partner: LNEC. Data management for this DS will be handled by AgdA and LNEC.
- Demo site 3 – North Atlantic coastal aquifer: ASTR scheme of MAR with agricultural tile drainage water by deep injection well. Location: Breezand, NW Netherlands. Owner: AW. Data management for this DS will be handled by AW.

Data re-use

Each of the 3 demo sites have well established infrastructures that already includes multiples sensors, thus already existing data from the demo site will be re-used for the project, such as:

- Historical data from existing sensors/piezometer
- Infrastructure location and plan of the demo sites
- Weather historical data
- Historical data for water volume in reservoirs, water volume from well extraction and water volume from WWTP
- Water quality historical data from WWTP and water quality historical data from the aquifers
- Socio-economic historical data

The project will also make use of publicly available data (open weather API for example)

Types and format of data

MARCLAIMED will use and generate different types of data, for which it will use the most commonly used formats/files types, for instance:

- For text documents: Microsoft Word (.docx/.doc) or PDF (.pdf).

- For tabular data: Microsoft Excel (.xlsx/.xls) or Comma Separated Values (.csv) or JSON (.json) files.
- For images: TIFF (.tif/.tiff), JPEG (.jpeg) or PNG (.png).
- For audio: Free Lossless Audio Codec (FLAC) (.flac) or WAV (.wav).
- For video: MPEG-4 (.mp4).
- For geospatial data: Shapefile (.shp), raster (.tiff), GeoPackage (.GPKG) or GeoJSON (.geojson/.json) files

Purpose of data generation

The main goal of MARCLAIMED is to demonstrate the feasibility of using Alternative Water Resources through the means of Managed Aquifer Recharge. In that regard, MARCLAIMED will gather sensor measurements and chemical lab analysis from the demo sites and use those data to generate new ones with the MARCLAIMED tools. Each tool was developed to simulate the response of the demo site systems to infrastructure changes (i.e. addition of an MAR method, for example) or to different scenarios (climatic, demographic, economic, ...). The newly generated data from the tools will thus be used to demonstrate the positive (or negative) impact of MAR on the demo sites.

The second main goal of MARCLAIMED is to raise awareness on the existence of MAR and especially to increase its social acceptance. To assess the current state of this social acceptance as well as its evolution during the project, MARCLAIMED will carry out interviews and focus groups involving stakeholders related to water management as well as other people that might be impacted by the activities of MARCLAIMED.

Expected size

At this early stage of the MARCLAIMED project, the expected size of the data that will be generated can't be assessed precisely. A broad estimation of the number of sensors and sampling both deployed and to be deployed can still be provided:

- Expected number of sensors: ~30
- Expected number of sampling points: ~20

The number of different variables that is measured by each type sensor as well as the measurement can vary a lot and is still not fully documented, meaning that trying to make an more precise estimation of the total number of measurements would hold too much uncertainty and will thus be left for a later version of the DMP.

Data origin

The data generated during the MARCLAIMED project will come from the demo sites, they will mainly be:

- Measurements from sensors
- Chemical lab analysis results from water retrieved at the sampling points

- Tools outputs (new variable prediction, forecasts, ...)
- Users and stakeholders' viewpoints collected through interviews and focus group

Data utility

Beside their use within the MARCLAIMED project, data generated within MARCLAIMED will include multiple data related to water quality as well as data related to operational management of wastewater treatment plants. Such data might be useful to multiple types of data consumers such as:

- Public authorities, who might benefit from new insights to conduct management activities and risk mitigation.
- Start-ups or businesses willing to provide new services based on the processed data offered by MARCLAIMED.
- Utility providers, customers, research and university audience, internal & external sales network, energy advisors in the water ecosystem.
- Researchers who are working in the areas of interest of MARCLAIMED. Researchers can use a variety of the data produced within MARCLAIMED using:
 - Datasets for building their own AI/ML models and applications
 - AI Models for testing and comparing their own models
 - MARCLAIMED tools for developing their own modules

2. FAIR data

The **FAIR (Findability, Accessibility, Interoperability, and Reusability)** data principles emphasise the capacity of computational systems to find, access, interoperate, and reuse data with no or minimal human intervention because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.

MARCLAIMED considers these principles particularly relevant to build the basis for a solid solution. The following sections explain specifically how MARCLAIMED will ensure that the diverse datasets handled within the project will comply as much as possible with those FAIR data principles.

Making data **findable (F)**

Data is considered findable if they comply to the following:

- F1. (Meta)data is assigned a globally unique and eternally persistent identifier.
- F2. Data (including datasets and AI models) are described with rich metadata.
- F3. (Meta)data are registered or indexed in a searchable resource.
- F4. Metadata specifies the data identifier.

Data identification

MARCLAIMED plans to establish whether the data produced and/or used in the project are identifiable and traceable through a standard identification mechanism (e.g., persistent and unique identifiers such as Digital Object Identifiers, DOI). This will include naming conventions and identifying search keywords.

Data discovery

Metadata will be provided as much as possible for all the data that will be collected during the project.

- For models, especially the MARCLAIMED tools, this includes aspects relating to its general application, methods applied and nature of data it consumes.
- For services, this will include aspects in relation to what data in processes and operations it can perform.
- For datasets, this will include owner's data, ways in which the data are produced, ways in which the data are sent to the IDST.

What specific metadata will be created will be further considered and, in the event that metadata standards do not exist, MARCLAIMED will outline the type of metadata created and how it will be created. As an initial plan, MARCLAIMED foresees the use of DCAT-AP for encoding datasets metadata.

Metadata discovery and indexation

All data within the MARCLAIMED project will be stored within the IDST datacore, which uses the NGSI-LD API to query and retrieve data. This API allows to perform discovery based on the name of any variable name in the database, meaning that any variables within the IDST datacore can be used as keywords. Specific endpoints within the NGSI-LD allows for the retrieval of all metadata names in the databases, allowing them to be later used as key words.

Making data accessible (A)

Data is considered as accessible if they comply to the following:

- A.1. (meta)data are retrievable by their identifier using a standardised communications protocol.
 - A1.1. The protocol is open, free, and universally implementable.
 - A1.2. The protocol allows for an authentication and authorisation procedure, where necessary.
- A2. Metadata are accessible, even when the data are no longer available.

The data generated by the demo sites will be made accessible through the MARCLAIMED IDST, assuming that the users have proper authorization to access them.

Trusted repository

All data produced and used within MARCLAIMED will be stored in the IDST datacore, which is a NGSI-LD context broker. At this stage, the consortium has not identified the need to deposit the data in another kind of trusted repository. Access to the stored data will be restricted to authenticated and authorised users using fine grained (at NGSI-LD entity level) access control mechanisms provided by the Stellio context broker.

The duration of storage beyond project lifetime remains to be evaluated based on actual produced datasets. Possibility of transfer to a static data repository such as [Zenodo](https://zenodo.org/) will be evaluated.

Data identification

All data in the IDST datacore are stored in entities that each possess a unique identifier (NGSI-LD URI). Each measurement within said entities also possesses an “instanceId” that can be used as a unique identifier for each single measurement.

Data public availability

As stated previously, most of the data identified to be produced at this stage of the project will either be measurements from the demo sites or results from users and stakeholders interviews and focus groups.

Regarding the data from the demo sites:

- DS2 (Comporta, Portugal) has declared that, at this stage, their intention is to make the data on the demo site publicly available. However, as the project develops, they reserve the right, in accordance with their internal policy, to prevent the dissemination of data that could be considered confidential or sensitive to the utility's activities.
- DS3 (Breezand, The Netherlands) has declared that they don't plan on making their data publicly available.
- DS1 (Barcelona, Spain) has not made a decision on whether they expect to publicly share data or not.

All statements made by the demo are subjects to change and will be elaborated in later versions of the DMP.

Regarding the results from the users and the stakeholders interviews, those data will most likely contain personal data and will thus not be made publicly available. The possibility of applying an anonymization process to those data to possibly make shareable will be evaluated during the project.

Accessibility

All the data will be stored in the IDST datacore, which is a NGSI-LD context broker that can provide a free and standardised access by the means of the NGSI-LD API. The NGSI-LD API is meant to facilitate the open exchange and sharing of structured information between different stakeholders. NGSI-LD has been standardised by ETSI (European Telecommunications Standardization Institute).

Restrictions

The IDST will integrate Keycloak as its SSO (single sign-on) service. Keycloak implements full support of the OpenId Connect and OAuth2 standards, it offers a very precise rights configuration, allowing to set up any access restrictions needed in the project.

Identity verification for data access

There is a need to ascertain the identity of any person trying to access the data. Any access will require an account on the SSO of the project, which will involve an email verification to ascertain the identity of the person trying to access the data.

If needed, for the beginning of the project, the account creation on the SSO can be disabled, so that only the admin can create an account and make sure that only the adequate people can access the data.

Access to data will be restricted to project partners, until the dataset owner decides to make it public or to provide its access to a specific set of users.

Data access committee

The need for a data access committee has not been identified at this stage of the project. The vast majority of the data gathered and produced during the project will belong to the demo sites or will be generated from data belonging to the demo sites. The decision to share or not the data will be made by each demo site directly.

Metadata public availability

MARCLAIMED initially plans to apply the same confidentiality level to a dataset and to its metadata, meaning that metadata regarding publicly available data will also be publicly available while metadata regarding confidential data will also be confidential. This aforementioned statement is only a general guideline to be followed that can be bypassed if needed.

Making data interoperable (I)

Interoperability

Data produced within MARCLAIMED will be structured using the NGSI-LD information model to provide a standardised representation of data and metadata across different domains. MARCLAIMED will assess and select the appropriate data models to enrich and accurately represent the data from its use cases. If new data models are being developed, they will be published.

Mapping to commonly used ontologies

MARCLAIMED's approach is to first attempt to extend ontology concepts that are adopted by prominent MAR initiatives. It will then attempt to map to other information models where concepts match or are similar.

Increase data re-use (R)

Data is considered reusable if it complies with the following:

- R1. Meta (data) has a plurality of accurate and relevant attributes.
 - R1.1. (meta)data are released with a clear and accessible data usage licence.
 - R1.2. (meta)data are associated with their provenance.
 - R1.3. (meta)data meet domain-relevant community standards.

Data analysis validation

When needed, data produced in MARCLAIMED will be provided alongside "readme" files that facilitate data reuse in the long term. These documents will include information on the methodology followed, analyses conducted, definitions of variables employed or units of

measurement. Data models will be described using schemas for structured data such as the one provided by schema.org.

Freely available data

As stated previously, at this stage of the project, one of the demo sites has stated its plan to make its data open in the public domain. The type of licence that would be used to do so has not been decided and will be identified in later versions of the DMP.

Third party usability

One of the demo sites of the project that does not plan to share its data publicly has stated that their data will be usable by a university outside of the consortium, to produce scientific publication.

Standards

Data generated in MARCLAIMED will have clear licence and provenance, being thoroughly documented using the standards considered most appropriate at the time of the project execution.

Data quality assurance

The Data Quality Assurance Process is the responsibility of the whole consortium and will be done according to the guidelines specified in this report. This process will consist of the following steps:

- Establishment of the way in which the collection and processing of data is managed. This should be done by distinguishing the different types of data managed in the project: data collected or generated in the different use cases, data reused from other EU project initiatives or data from project communication and dissemination activities.
- Definition of the responsibilities of each of the roles involved in the process.
- Determination of the quality parameters according to the six main dimensions that are commonly used to mark the quality of the same: accuracy, completeness, consistency, validity, uniqueness, and timeliness.
- Implementation of the set of technical tools and Key Performance Indicators (KPIs) that will be used to measure and ensure the quality of the data. The type of tools to be employed will be defined throughout the project according to its needs.

3. Other research outputs

Other main research outputs in MARCLAIMED are the followings:

- Output related to the operational efficiency and economic sustainability tools:
 - Deep learning models trained for (MARVIS)
 - Physics informed ML models (PREDMAR)
 - Decision trees and system optimization (MARAISSER)
 - Water availability forecast models (OASIS FLOW)
 - AWR performance simulation models (MARINSURE)
 - Cost recovery models (RECOVER)
- Output related to the IDST:
 - Data models used by the IDST
 - Tool management API
 - IDST graphical interface
- Output related to the co-creation of AWR strategies
 - AWR strategies definition
 - AWR strategies performance assessment and ranking

MARCLAIMED being still in an early stage of the project, none of the above-mentioned output is advanced enough to assess whether or not the consortium plan on sharing and making it available for reuse. This will be discussed in the second version of the data management plan.

4. Allocations of resources

Costs

At this stage of the project, since no data have been identified as publicly shareable, the costs of making data FAIR cannot be estimated and will be addressed in the next version of this deliverable if any applicable data are identified later.

Data management responsible

Data management in MARCLAIMED will be done as part of T7.4 Data requirements planning and management, of which EGM is the task leader. EGM will thus be responsible for data management in MARCLAIMED.

Long term preservation

Resources for long-term preservation, associated costs and potential value, as well as how data will be kept beyond the project and for how long, will be discussed in the next version of the following plan.

5. Data security

Provisions

Specific data security provisions that will be applied for the project will be detailed in Deliverable D7.5 - Research Ethics and Data Protection Monitoring Report v1.

Storage

For the duration of the project, datasets will be stored either:

- In the IDST datacore
- In the consortium's shared Google Drive, provided by the project coordinator (CETAND)
- On the local storage of the demo site that produced the data (when applicable)

Most of the time, data is most likely to be duplicated on at least 2 of the above-mentioned storage, ideally, in the IDST and the demo site local storage. The shared Google Drive will be mostly used as an intermediary step to transfer the data from the local storage to the IDST, as well as the main sharing point for text documents such as deliverables.

6. Ethics

Possible ethical issues

The questions regarding ethics or legal issues will be presented and answered in detail in Deliverable D7.5 - Research Ethics and Data Protection Monitoring Report v1.

In addition, as part of T2.1 Co-design of the MARCLAIMED IDST, the IDST and specifically the AI models that it will orchestrate will undergo the “Trustworthy AI assessment list” evaluation of the HLEG AI (section 1.2.8 of the DoA). This assessment will be planned as soon as the development of the AI models allows it.

Consent for data sharing and long-term preservation

Data collection, storage, protection, retention and destruction will be carried out through the IDST datacore. Interviewees/beneficiaries/recipients will be informed about data security, anonymity and use of data as well as asked for in accordance. Participation happens on a voluntary basis and via the completion of informed consent prepared accordingly.

7. Other issues

At this stage of the project, the need for other national/funder/sectorial/departamental procedures for data management has not been not identified. The need for such procedures will be continuously monitored during the course project.

8. CONCLUSIONS

This first version of a Data Management Plan for the MARCLAIMED provides a description of the types of data that will be used and generated within MARCLAIMED. Most of this data will be sensor measurements and chemical lab analysis results from the demo site, as well as results of interviews of stakeholders regarding the social acceptance of MAR. The nature of those data being possibly confidential regarding the operational safety of their infrastructure, the decision to share those data to third parties and/or to make it openly available will be made exclusively by the demo sites and is subject to change during the course of the project. At this stage of the project, only one of the demo sites has stated that it is planning to make its data publicly available. The consortium will continue to dynamically assess these sharing policies during the whole project and will try as much as possible to comply with the FAIR guidelines while still accounting for the confidentiality needs of the demo sites.

MARCLAIMED will support decision makers to integrate Managed Aquifer Recharge (MAR) with Alternative Water Resources (AWR) in River Basin and Drought Management Plans supporting the adaptation and resilience of structural supply systems in the context of climate change. The main outcome will be the MARCLAIMED Integrated Decision Support Tool (IDST) that will provide solutions to tackle existing and emerging threats and support decision making and adaptation policies.



MARCLAIMED



**Funded by
the European Union**