Data Management Plan Information

SWAIN Data Management Plan

This Data Management Plan (MDP) concerns the datasets generated by the Sustainable Watershed Management Through IoT-Driven Artificial Intelligence (SWAIN) project as a basis for scientific publication of the project results. These datasets include data generated by numerical experiments, associated documentation, and metadata. The aim is to create a modular, scalable, and secure platform enabling data collection, storage, and easy access, in a unified and standardized manner.

Funder

Austrian Science Fund (FWF), Swiss National Science Foundation (SNSF), Scientific and Technological Research Council of Turkey (TUBITAK), Academy of Finland (AKA)

Grant

CHIST-ERA

Organisations

Istanbul Technical University, Finnish Environment Institute, University Bogazici, Vienna University of Technology, Università della Svizzera italiana

Researchers

Sabtain Ahmad (orcid:0000-0003-3798-6355), Heidi Ahkola, Slobodan Lukovic (orcid:0000-0003-4079-651X), CESARE ALIPPI (orcid:0000-0003-3819-0025), Ulas Tezel (orcid:0000-0001-8322-1666), Ivona Brandic (orcid:0000-0001-7424-0208), Tolga Ovatman (orcid:0000-0001-5918-3145), Atakan Aral (orcid:0000-0002-2281-8183), Mehmet Tahir SANDIKKAYA (orcid:0000-0002-9756-603X), Vincenzo De Maio (orcid:0000-0002-7352-3895)

Datasets

Title: Environmental data regarding Ergene and Kokemäki rivers

Template: CHIST-ERA

This dataset contains environmental data collected and generated during the SWAIN project. Additionally, it includes the relevant environmental data collected prior to the project.

Dataset Description

1.1.1 How will new data be collected or produced?

1.1.1.1 Explain which methodologies or software will be used if new data are collected or produced

• SeaDataNet European Directory of Marine Environmental Data (EDMED), Open Knowledge Maps

• Environmental sensor data: Environmental sensor data will be collected from Ergene River, Turkey, and Kokemäenjoki River, Finland. Sensors will be deployed by international research partners to collect the data to assess the water quality and detect micropollutants in these two rivers. The collected data includes but is not limited to timestamped, the concentration of various chemicals and other pollutants in the rivers, hydrological data (e.g., streamflow, groundwater levels), metrological data (e.g., precipitation, wind speed, and direction), land use, topography, and industrial sites. Computational architecture data: System logs of the data processing architecture will be collected and analyzed as an output of simulations, emulations, and proof-of-concept implementation. These logs will include timing and scheduling information, energy measurements, hardware and software failure traces, and performance indicators such as accuracy, precision, recall, error rate, etc.

1.1.1.2 Explain how data provenance will be documented

Data creation, transformation, and processing will be recorded through a combination of a W3C Provenance Data Model and a README text file that describes the data collection and processing methods.

1.1.2 What data (for example the kind, formats, and volumes), will be collected or produced?

1.1.2.1 Give details on the kind of data

Observational (e.g., sensor data, data from surveys), Experimental (e.g., gene sequencing data), Simulation (e.g., climate modeling data)

1.1.2.2 Give details on the data format

numeric (databases, spreadsheets)Text files - MS Word docs, .txt files, PDF, RTF, XML (Extensible Markup Language), Models - 3D, statistical, Software - Java, C, Python, Discipline specific formats - Flexible Image Transport System (FITS) in astronomy, Crystallographic Information File (CIF) for crystallography

1.1.2.3 Justify the use of certain formats

SWAIN data is composed of many types of data that will be collected from multiple locations from two rivers. In order to accommodate different types of data with different granularity, a consolidated geodatabase is created to store data harvested from different locations in the watershed. The objective is to store data in a standardized format so that it can be queried and manipulated using a single common interface. A database based on MongoDB is designed since it provides a general and widely-adopted non-relational database and uses a simple query syntax and has many drivers to interact with the database.

1.1.2.4 Give details on the volumes

TB (terabyte)

Comment: The data collected from various locations from two rivers in order to effectively and continuously measure their health condition and detect presence of hazardous chemicals is estimated to be in the order of terabytes, whereas the size of computational data is estimated to be in the order of gigabytes.

2.1.1 What metadata and documentation will accompany the data?

2.1.1.1 Indicate which metadata will be provided to help others identify and discover the data

Structural, Administrative, Descriptive

2.1.1.2 Indicate which metadata standards will be used

- EML (Ecological Metadata Language)

- Metadata standards to be used also include the European Information Environment and Observation Network (Eionet). Metadata is stored in research metadata profile and is stored in .xml format. The metadata can also be harvested from standard OGC WCS web-interface (http://www.opengeospatial.org/standards/wcs) by Etsin service

2.1.1.3 Indicate how the data will be organised during the project

The use of naming conventions and version numbers that indicate the date data has been created will be adopted. Database schemas describing the entities and relationships between them have been made available on GitHub for everyone to use.

2.1.1.4 Consider what other documentation is needed to enable re-use

Data will be released according to the European information environment data standard. Future updates to the data will introduce the use of unique identifiers, including different measurement types and machine-readable JSON schemas and sample documents from MongoDB collections.

2.1.1.5 Consider how this information will be captured and where it will be recorded

- Open Knowledge Maps

- The metadata information will be recorded in machine-readable JSON format with the description stored on Github in the form of a README file.

2.1.2 What data quality control measures will be used?

2.1.2.1 Explain how the consistency and quality of data collection will be controlled and documented

• Calibration, Standardised data capture, Data entry validation

• Data standards developed by the European Environmental Information and Observation Network (Eionet) will be employed. Automatic calibration based on AI techniques will be implemented to ensure consistency and handle missing data.

• Strategies for near-real time data quality assurance for the environmental data delivered in the project will be developed.

• Several automatic data quality control tests which identify and label/remove erroneous measurements will be implemented. Based on pre-determined parameter thresholds, noise and missing data will be calculated using data-driven methods such as machine learning models.

3.1 Reused Data

3.1.1 How will existing data be re-used?

• To reproduce and validate findings, To compare and combine with other data

• https://github.com/steve3nto/swain-data

3.1.2 Where can re-used data be found?

• other

• https://github.com/steve3nto/swain-data

3.1.3 Which data will be re-used?

• other

• Chemical and microbial data measured during the CONPAT project, observed and modelled data from the FMI and SKYE's Hertta database.

3.1.4 State any constraints on re-use of existing data if there are any

No constraints for using the existing data.

3.1.5 Briefly state the reasons if the re-use of any existing data sources has been considered but discarded

The re-use of existing data sources has not been discarded. It is considered useful.

4.1.1 How will data and metadata be stored and backed up during the research?

4.1.1.1 Describe where the data will be stored and backed up during research activities and how often the backup will be performed

• HPC | National HPC Infrastructure

• every 3 Months

• u:Cloud at University of Vienna, TUowncloud at Vienna University of Technology, Cloud Services at the Università della Svizzera italiana

4.1.2 How will data security and protection of sensitive data be taken care of during the research?

4.1.2.1 Explain how the data will be recovered in the event of an incident

Data will be stored at least in two different locations in order to avoid a single point of failure and be able to provide continued access in case of failure at one of the locations.

4.1.2.2 Explain who will have access to the data during the research and how access to data is controlled, especially in collaborative partnerships

The data is hosted at one of the international partners (Università della Svizzera italiana) and every partner can gain access to the database, manipulate and contribute to the development of the database. To coordinate and share resources between the work packages, data is temporarily stored on Github to provide before finally transporting it to the database.

4.1.2.3 Describe the main risks and how these will be managed

Since no personal, human-related, or sensitive data will be collected during the project, the only risks involved are consistency and correctness of the data which we aim to handle through data-driven quality assurance methods.

4.1.2.4 Explain which institutional data protection policies are in place

Data protection policies are described by each partner institution: https://search.usi.ch/en/organisational-units/392/ethics-and-communication-law-center https://dsba.univie.ac.at/fileadmin/user_upload/p_dsba/datenschutzerklaerung_websites_V04_26062020_EN.pdf https://www.tuwien.at/en/tu-wien/organisation/central-divisions/data-protection-and-document-management/data-protection-at-tu-wien

5.1.1 Personal data

5.1.1.1 Are there any personal data to be formulated?

No

5.1.1.2 Explain whether there is a managed access procedure in place for authorised users of personal data

No persona data to be formutlated.

5.1.2.1 Data ownership and accessibility

5.1.2.1.1 Who will be the owner of the data?

Participating institutes where data will be stored for at least 3 years after the completion of the project. The data will be open to the public.

5.1.2.1.2 Explain what access will apply to the data?

- Open

- SWAIN aims to make openly available all datasets produced as an outcome of the project. Datasets will be published after the completion of the project.

5.1.2.1.3 Will the data be openly accessible, or will there be access restrictions?

Openly accessible

5.1.2.2 Intellectual property rights

5.1.2.2.1 Are intellectual property rights affected?

No

5.1.2.3 Third-party data restrictions

5.1.2.3.1 Indicate whether there are any restrictions on the re-use of third-party data

No restrictions exist.

5.1.3 Ethical issues

5.1.3.1 What ethical issues and codes of conduct are there, and how will they be taken into account?

- Other

- Legal permission regarding the collection of environmental data will be handled by corresponding international research partners. Creative Commons CC4.0-BY is the standard open data license that will be used. Therefore, there will not be any restrictions regarding the sharing and re-use of data.

6.1.1 How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

6.1.1.1 Explain how the data will be discoverable and shared

Deposit in a trustworthy data repository, Use of a secure data service

6.1.1.2 Outline the plan for data preservation and give information on how long the data will be retained

Data will be stored in participating institutes' open data infrastructure for at least 3 years after the completion of the project. Another potential storage that could be used is IDA (https://www.fairdata.fi/en/ida). The service is produced by CSC and offered free of charge to researchers and universities working with funding from the Academy of Finland.

6.1.1.3 Explain when the data will be made available

SWAIN aims to publish all datasets after the completion of the project.

6.1.1.4 Indicate the expected timely release

2024-06-01

6.1.1.5 Will exclusive use of the data be claimed?

No

6.1.1.7 Indicate whether data sharing will be postponed or restricted

- Postponed

- Postponed until the publications that use the data are available.

6.1.1.8 Indicate who will be able to use the data

Researchers, Research communities, Decision makers

6.1.1.9 Is it necessary to restrict access to certain communities or to apply a data sharing agreement?

No

6.1.2 How will data for preservation be selected, and where data will be preserved long-term?

6.1.2.1 Indicate what data must be retained or destroyed for contractual, legal, or regulatory purposes

- none
- Retained
- No data has to be destroyed.

6.1.2.2 Indicate how it will be decided what data to keep

As per PI's knowledge, there are very few publicly available datasets for environmental (e.g., river, watershed) monitoring. We intend to publish all our datasets with the aim to facilitate researchers not just to be able to reproduce research produced in the SWAIN but also to enable decision-makers and researchers to contribute towards building solutions for effective environmental monitoring.

6.1.2.3 Describe the data to be preserved long-term

Observational (e.g., sensor data, data from surveys), Experimental (e.g., gene sequencing data), Simulation (e.g., climate modeling data)

6.1.2.4 Explain the foreseeable research uses (and/ or users) for the data

The data might be useful to researchers, research communities, decision-makers, the public, and the economy. The datasets produced within SWAIN could be useful to anyone interested in learning continuous monitoring of watersheds and what are different types and sources of chemicals affecting the water quality.

6.1.2.5 Indicate where the data will be deposited

International Journal of Environment, Engineering & Education

6.1.2.6 Demonstrate that the data can be curated effectively beyond the lifetime of the grant

The university cloud repositories allow indefinite data storage without cost.

6.1.3 What methods or software tools are needed to access and use data?

6.1.3.1 Indicate how the data will be shared

Repository

6.1.3.2 Indicate whether potential users need specific tools to access and (re-)use the ata.

ArcGIS Map Viewer, MongoDB Compass, Git

6.1.4 How will the application of a unique and persistent identifier to each data set be ensured?

6.1.4.1 Explain how the data might be re-used in other contexts

To obtain information, To share information, To make informed decisions, To develop a product, To improve a product, To combine with other data

6.1.4.2 Indicate whether a persistent identifier for the data will be pursued

- Yes

- DOI, URI

7.1.1 Who will be responsible for data management?

7.1.1.1 Outline the roles and responsibilities for data management/stewardship activities

- Atakan Aral (orcid:0000-0002-2281-8183)

- Overview and management

7.1.1.2 Is it a collaborative project?

Yes

7.1.1.3 Explain the co-ordination of data management responsibilities across partners

Each partner has a PI and co-PI(s) who are responsible for data management.

7.1.2 What resources will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

7.1.2.1 Explain how the necessary resources to prepare the data for sharing/preservation have been costed in

Use of institution infrastructure

7.1.2.2 Indicate whether additional resources will be needed to prepare data for deposit or to meet any charges from data repositories

No