

AI4EOSC Data Management Plan

Version 1

Description

The primary objective of the initial Data Management Plan (DMP) for the AI4EOSC project is to ensure compliance with GDPR regulations and to provide a comprehensive overview of how the project is managing the data generated within the scope of AI4EOSC in accordance with the relevant principles of making data findable, accessible, interoperable, and reusable (FAIR). The DMP outlines the type of data that the project is generating, how it will be made accessible for verification and re-use, and how it will facilitate potential re-use of the collected and processed data.

Within the list of objectives, work packages and tasks of the project, some FAIR-related activities can be found. For example, Task 7.2 aims to support the different use cases to ensure that the different digital objects potentially reusable along the workflow, adopt the FAIR principles: data, models, predictions, metadata, publications, software, etc. This means that, apart from the datasets used as input for model training, the complete workflow will be connected to identify the different components, trying to make all the scientific pipeline reproducible.

Funder

European Commission||EC

Grant

Artificial Intelligence for the

Researchers

Ignacio Heredia (orcid:0000-0001-6317-7100), Valentin Kozlov (orcid:0000-0002-8770-3619), Álvaro López García (orcid:0000-0002-0013-4602), Fernando Aguilar Gómez (orcid:0000-0001-9462-4831)

Organizations

Institute of Physics of Cantabria (IFCA), Spanish National Research Council (CSIC), Madrid, Spain, KIT - Steinbuch Centre for Computing

1. Main Info

Title of DMP: [AI4EOSC Data Management Plan](#)

Description:

The primary objective of the initial Data Management Plan (DMP) for the AI4EOSC project is to ensure compliance with GDPR regulations and to provide a comprehensive overview of how the project is managing the data generated within the scope of AI4EOSC in accordance with the relevant principles of making data findable, accessible, interoperable, and reusable (FAIR). The DMP outlines the type of data that the project is generating, how it will be made accessible for verification and re-use, and how it will facilitate potential re-use of the collected and processed data.

Within the list of objectives, work packages and tasks of the project, some FAIR-related activities can be found. For example, Task 7.2 aims to support the different use cases to ensure that the different digital objects potentially reusable along the workflow, adopt the FAIR principles: data, models, predictions, metadata, publications, software, etc. This means that, apart from the datasets used as input for model training, the complete workflow will be connected to identify the different components, trying to make all the scientific pipeline reproducible.

Researchers:

[Ignacio Heredia \(orcid:0000-0001-6317-7100\)](#)

[Valentin Kozlov \(orcid:0000-0002-8770-3619\)](#)

[Álvaro López García \(orcid:0000-0002-0013-4602\)](#)

[Fernando Aguilar Gómez \(orcid:0000-0001-9462-4831\)](#)

Organizations:

[Institute of Physics of Cantabria \(IFCA\)](#)

[Spanish National Research Council \(CSIC\), Madrid, Spain](#)

[KIT - Steinbuch Centre for Computing](#)

Contact: [Fernando Aguilar Gómez](#)

2. Funding

Funding organizations: [European Commission](#)||[EC](#)

Grants: [Artificial Intelligence for the European Open Science Cloud](#)

Project: [Artificial Intelligence for the European Open Science Cloud](#)

3. License

License: [CC-BY-4.0](#)

Access Rights: [Public](#)

4. Templates

Descriptions

UC3 – AUTOMATED THERMOGRAPHY [Input data - TUFSeg]

Use case 3 leverages multispectral unmanned aerial vehicle (UAV) based imaging and AI to identify “hot spots” (thermal anomalies) in urban settings and thus improve the efficiency of energy-related systems. One of the two scenarios being examined is:

Thermal Urban Feature Segmentation (TUFSeg): Detecting thermal hotspots caused by common urban features (cars, manholes, streetlamps, etc.) to aid district heating network operators in their search for pipeline leakages by automatically removing these false alarms from the list of potential suspects.

Template: [Horizon Europe](#)

Type: [Dataset](#)

1.1 Brief description of the described research output

1.1.1 What kind of research output are you describing?

Research Data

1.1.2 Is it physical or digital?

Digital

1.1.3 Are you generating or re-using it?

New

1.1.4 What is the type of the described dataset?

Observational

1.1.6 What is its expected size?

8.4 GB

1.1.7 Why are you collecting/generating or re-using it?

- To obtain information
- To make informed decisions

1.1.8 What is its origin / provenance?

As stated in section 1.1.3, the data is novel and generated by us for research purposes as well as for the AI4EOSC project. The data was acquired using DJI's Matrice 600 Pro and 300 RTK UAVs and a dual camera with a 4k RGB camera and FLIR's XT2 thermal sensor. For more information, view the associated publications.

1.1.9 To whom might it be useful ('data utility')?

- Researchers
- Decision makers

Researchers in the field of urban infrastructure, multispectral imaging, or foundation models; decision makers in infrastructure maintenance such as building owners or network operators to compare with existing and previously monitored systems.

2.1 Publications

2.1.1 Does the described output support any scientific publication?

Yes

2.1.2 Is there a data availability statement provided along with the publication?

No

2.3 Software

2.3.1 Does the described output use or support any software?

Yes

<https://github.com/emvollmer/TUFSeg>

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 What type(s) of persistent identifier(s) are used for the described dataset / output?

Data identifiers

DOI

3.1.1.2 Will you provide metadata for the described dataset / output?

Yes

3.1.1.3 What type(s) of metadata?

- Descriptive
- Administrative

3.1.1.4 Do the metadata use standardised vocabularies?

No

3.1.1.6 Are the metadata searchable?

No

3.1.1.8 Are keywords provided in the metadata?

No

3.1.1.9 Are metadata harvestable?

No

3.2.1 Repository

3.2.1.1 In which repository will the dataset / output be deposited?

Zenodo

3.2.1.2 Is the selected repository a trusted source?

No

3.2.1.5 Does the repository(ies) assign datasets / outputs with persistent identifiers?

Yes

3.2.1.6 Does the repository(ies) resolve the identifiers to a digital object?

Landing page

3.2.1.7 Does the repository support versioning?

Yes

3.2.2 Data

3.2.2.2 How is the dataset / output shared?

Open

3.2.2.5 Are there any methods or tools required to access the dataset / output?

Yes

Couldn't find it? Insert it manually

Tools for decompressing the folders (.tar.std or .zip) are required, though these are standard packages and terminal commands.

3.2.2.7 Please provide information about the tools needed to access the dataset / output.

Couldn't find it? Insert it manually

Tools for decompressing the folders (.tar.std or .zip) are required, though these are standard packages and terminal commands.

3.2.2.8 Is the described dataset / output supported by a data access committee?

No

3.2.2.9 Please specify how the dataset / output will be accessed during and after the project ends

Via NextCloud and Zenodo during the project, after the project via Zenodo.

3.2.2.10 Please specify how long after the project has ended the dataset / output will be made accessible for

As they will all have been published on Zenodo, there shouldn't be subject to an embargo.

3.2.3 Metadata

3.2.3.1 Will you provide metadata even if the described dataset / output can not be openly shared?

Yes

3.2.3.2 Under which license will metadata be provided?

Creative Commons Zero (CC0)

3.2.3.3 Do metadata provide information about how to access the described dataset / output?

Yes

3.2.3.4 Will metadata remain available after the dataset / output is no longer available?

Yes

3.3 Making data and other outputs interoperable

3.3.1 Does your (meta)data use a controlled vocabulary?

Yes

3.3.3 Have you applied a standard schema for your (meta)data?

Yes

Couldn't find it? Insert it manually

Dublin Core

3.3.7 Does the described dataset / output provide qualified references with other outputs?

Yes

ORCID to data managers, producers, etc.

3.4 Increasing data and other outputs reuse

3.4.2 What reusability and / or reproducibility methods are followed?

- Readme files
- Codebooks

- Other

All datasets and code have supplementary README files and all required explanations to re-use. See the provided code and dataset link at 2.1.1.

3.4.3 Will you provide the described dataset / output in the public domain?

Yes

3.4.4 Do you intend to ensure (re)use by third parties after your project finishes?

Yes

Via Zenodo

3.4.5 Is provenance well documented?

Yes

Zenodo

4.1 Allocation of resources

4.1.1 What will be the cost of making the described output FAIR?

0

Euro

- Storage
- Archiving
- Re-use

Direct cost

4.1.2 How will this cost be covered?

Use of institution infrastructure

4.1.3 Identify the people who will be responsible and their role(s) in the management of the described output

Elena Vollmer (orcid:0000-0002-8805-3726)

6.1 Ethical aspects

6.1.2 Does the described dataset / output contain sensitive information?

No

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

UC3 – AUTOMATED THERMOGRAPHY [Input data - TBBRDet]

Use case 3 leverages multispectral unmanned aerial vehicle (UAV) based imaging and AI to identify “hot spots” (thermal anomalies) in urban settings and thus improve the efficiency of energy-related systems. One of the two scenarios being examined is described in this description.

Thermal Bridges on Building Rooftops Detection (TBBRDet): Detecting thermal hotspots on building rooftops caused by thermal bridges to support urban planners and building owners with retrofitting plans.

The utilized data is given in the form of multispectral imagery, simultaneously acquired standard red green blue (RGB) and thermal infrared (TIR) images. Via custom processing pipelines, this raw data is undistorted, aligned and merged into 4-channel images. Select images are annotated with the VGG Image Annotator, thus creating JSON files. In preparation for AI model training, these are processed according to formats required by the utilized methods and toolboxes.

Template: [Horizon Europe](#)

Type: [Dataset](#)

1.1 Brief description of the described research output

1.1.1 What kind of research output are you describing?

Research Data

1.1.2 Is it physical or digital?

Digital

1.1.3 Are you generating or re-using it?

New

1.1.4 What is the type of the described dataset?

Observational

Primary data are raw RGB and TIR images. These are both comprised of three channels with values between 0 and 255 format. In RGBs, channels are red, green and blue; in TIRs, all three are identical as these are grayscale intensity images. Secondary data are the undistorted, aligned and merged images.

Publishing of this data has been previously discussed. The input data for AI model training consists wholly of the processed, secondary data.

1.1.5 What is its format?

The raw RGB and TIR images are given in JPG format, the merged images are NPYs. On Zenodo and NextCloud, these are stored as zipped folders to be extracted for further use.

1.1.6 What is its expected size?

35MB

1.1.7 Why are you collecting/generating or re-using it?

- To obtain information
- To make informed decisions
- To develop a product

The overarching objective lies in automating the detection of thermographic anomalies to accelerate the implementation of necessary counter-measures and repairs in urban infrastructure. AI can be used to this end as either a stand-alone tool (TBBDet) or through integration into a pre-existing analysis pipeline (TUFSeg) to counteract the current inability to quickly and accurately pinpoint areas of unwanted heat loss leading to reduced energy efficiency. As such datasets are not yet in existence for these specific inspection aims, we created our own.

1.1.8 What is its origin / provenance?

As stated in section 1.1.3, the data is novel and generated by us for research purposes as well as for the AI4EOSC project. The data was acquired using DJI's Matrice 600 Pro and 300 RTK UAVs and a dual camera with a 4k RGB camera and FLIR's XT2 thermal sensor. For more information, view the associated publications.

1.1.9 To whom might it be useful ('data utility')?

- Researchers
- Decision makers
- Education
- Economy

2.1 Publications

2.1.1 Does the described output support any scientific publication?

a. Yes

Remote Sensing

b. Yes

Automation in Construction

2.1.2 Is there a data availability statement provided along with the publication?

No

2.3 Software

2.3.1 Does the described output use or support any software?

Yes

<https://github.com/emvollmer/TBBDet>

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 What type(s) of persistent identifier(s) are used for the described dataset / output?

Data identifiers

DOI

<https://doi.org/10.5281/zenodo.7360996>

<https://doi.org/10.5281/zenodo.7022736>

3.1.1.2 Will you provide metadata for the described dataset / output?

Yes

3.1.1.3 What type(s) of metadata?

- Descriptive
- Administrative

In the case of the TBBD dataset and as mentioned in the Zenodo publication itself, the experimental metadata is structured with the Spatio Temporal Asset Catalog (STAC) specification family. This specification provides a standardized way for describing geospatial assets. It defines related JSON object types of Item, Catalog, and Catalog, extending on Collection as the basis. One STAC Collection JSON object provides information about the recorded images and the environmental conditions during recordings. It also contains information about the overall bounding box of the entire area in which images were recorded. This object links to related STAC Item JSON objects containing information about the recorded city blocks and the cameras. The objects for the city blocks contain the GeoJSON geometry of

the respective block and the corresponding bounding box. The objects containing the camera information are based on an existing STAC extension for camera related metadata.

Metadata of the archived NumPy files for each image was structured using the Data Package schema from the Frictionless Standards. This standard describes a collection of data files. Therefore, metadata about all containerized NumPy files of the six flight paths is provided within a JSON-based file.

All files are represented in a standardized way as FAIR Digital Objects (FAIR DOs) to enable machine actionable decisions on the data in spirit of the FAIR principles.

3.1.1.4 Do the metadata use standardised vocabularies?

No

3.1.1.6 Are the metadata searchable?

Yes

3.1.1.7 How are searchable metadata provided?

- Registry/Catalogue
- Metadata repository

3.1.1.8 Are keywords provided in the metadata?

Yes

<dc:subject>drone</dc:subject>

<dc:subject>UAV</dc:subject>

<dc:subject>thermal bridges</dc:subject>

<dc:subject>thermal</dc:subject>

<dc:subject>object detection</dc:subject>

<dc:subject>computer vision</dc:subject>

<dc:subject>hyperspectral</dc:subject>

<dc:subject>remote sensing</dc:subject>

3.1.1.9 Are metadata harvestable?

Yes

OAI-PMH at Zenodo

3.2.1 Repository

3.2.1.1 In which repository will the dataset / output be deposited?

Zenodo

3.2.1.2 Is the selected repository a trusted source?

No

3.2.1.5 Does the repository(ies) assign datasets / outputs with persistent identifiers?

Yes

3.2.1.6 Does the repository(ies) resolve the identifiers to a digital object?

Landing page

3.2.1.7 Does the repository support versioning?

Yes

3.2.2 Data

3.2.2.1 What is the described dataset / output title?

Hyperspectral (RGB + Thermal) drone images of Karlsruhe, Germany - Raw images for the Thermal Bridges on Building Rooftops (TBBR) dataset

3.2.2.2 How is the dataset / output shared?

Open

3.2.2.5 Are there any methods or tools required to access the dataset / output?

Yes

Couldn't find it? Insert it manually

Tools for decompressing the folders (.tar.std or .zip) are required, though these are standard packages and terminal commands.

3.2.2.7 Please provide information about the tools needed to access the dataset / output.

Couldn't find it? Insert it manually

Tools for decompressing the folders (.tar.std or .zip) are required, though these are standard packages and terminal commands.

3.2.2.8 Is the described dataset / output supported by a data access committee?

No

3.2.2.9 Please specify how the dataset / output will be accessed during and after the project ends

Via NextCloud and Zenodo during the project, after the project via Zenodo.

3.2.2.10 Please specify how long after the project has ended the dataset / output will be made accessible for

As they will all have been published on Zenodo, there shouldn't be subject to an embargo.

3.2.3 Metadata

3.2.3.1 Will you provide metadata even if the described dataset / output can not be openly shared?

Yes

3.2.3.2 Under which license will metadata be provided?

Creative Commons Zero (CC0)

3.2.3.3 Do metadata provide information about how to access the described dataset / output?

Yes

3.2.3.4 Will metadata remain available after the dataset / output is no longer available?

Yes

3.3 Making data and other outputs interoperable

3.3.1 Does your (meta)data use a controlled vocabulary?

Yes

As mentioned above, the TBBR metadata uses a controlled vocabulary: The Spatio Temporal Asset Catalog (STAC) specification provides a standardized way for describing geospatial assets while the metadata of the archived NumPy files for each image was structured using the Data Package schema from the Frictionless Standards, which also uses a controlled vocabulary.

3.3.3 Have you applied a standard schema for your (meta)data?

Yes

Couldn't find it? Insert it manually

Dublin Core

3.3.5 What is the methodology followed?

The dataset uses the Spatio Temporal Asset Catalog (STAC) specification for structuring the experimental metadata. STAC provides a standardized way for describing geospatial assets. It defines related JSON object types of Item, Catalog, and Collection. For the metadata of the archived NumPy files, the Data Package schema from the Frictionless Standards was used. This standard describes a collection of data files.

3.3.6 What community-endorsed interoperability best practices are followed?

The enforced practices by Zenodo are adhered to. Zenodo follows the FAIR principles, using JSON Schema as the internal representation of metadata and offering exports to other popular formats such as Dublin Core or MARCXML.

3.3.7 Does the described dataset / output provide qualified references with other outputs?

Yes

ORCID to data managers, producers, etc.

3.4 Increasing data and other outputs reuse

3.4.2 What reusability and / or reproducibility methods are followed?

- Readme files
- Other

All datasets and code have supplementary README files and all required explanations to re-use. See the provided code and dataset link at 2.1.1.

3.4.3 Will you provide the described dataset / output in the public domain?

Yes

3.4.4 Do you intend to ensure (re)use by third parties after your project finishes?

Yes

Via Zenodo

3.4.5 Is provenance well documented?

Yes

Zenodo

3.4.6 What documented procedures for quality assurance do you have in place?

Not available

4.1 Allocation of resources

4.1.1 What will be the cost of making the described output FAIR?

0

Euro

- Storage
- Archiving
- Re-use

Direct cost

4.1.2 How will this cost be covered?

Use of institution infrastructure

4.1.3 Identify the people who will be responsible and their role(s) in the management of the described output

James Kahn (orcid:0000-0002-8517-2359)

6.1 Ethical aspects

6.1.2 Does the described dataset / output contain sensitive information?

No

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

UC1 – AGROMETEOROLOGY [Input Data]

Anthropogenic-driven climate change is expected to affect, mainly to increase, frequency and intensity of various types of extreme events. One of them are thunderstorms that generate adverse phenomena for farmers:

- high winds can break and damage buildings, equipment, material and plants;
- hail can cause leaf damage reducing yield or destroying plants, machines, cars, and buildings;
- flash floods can endanger humans, animals, damage machines and equipment, can lead to loss of topsoil as well as damage to crops.

In this use case we process radar imagery - utilising AI ensemble stacking and federated learning - to create added value products for improving farmers activity, namely timely and precise warnings of thunderstorms for farmers. As an enhancement, the farmer wants to know when exactly thunderstorms will hit his area so that he can get everyone to safety. Nowcasting can provide localised warnings around 30 minutes ahead of thunderstorms. We integrated the use case with the platform, which enabled us to use existing AI model templates and state-of-the-art computing resources, perform automated testing and use other platform tools to tackle the problem. The model is a neural network currently, where the number of layers can be set by the user.

Template: [Horizon Europe](#)

Type: [Dataset](#)

1.1 Brief description of the described research output

1.1.1 What kind of research output are you describing?

Research Data

1.1.2 Is it physical or digital?

Digital

1.1.3 Are you generating or re-using it?

New

Data were measured previously by MicroStep-MIS meteorological radar

1.1.4 What is the type of the described dataset?

Observational

Primary data were collected previously by MicroStep-MIS radar. For the purpose of this project we have created a secondary dataset.

1.1.6 What is its expected size?

1TB

1.1.7 Why are you collecting/generating or re-using it?

- To obtain information
- To make informed decisions

We need input data for thunderstorm forecasting.

1.1.8 What is its origin / provenance?

The data is gathered by an instrument called MMR-116 (Mini Meteorological Radar). Product sheet can be found here:

https://www.microstep-mis.com/drupal/web/sites/default/files/datasheets/MMR-116_product%20sheet_0.pdf

1.1.9 To whom might it be useful ('data utility')?

- Researchers
- Research communities
- Decision makers

2.1 Publications

2.1.1 Does the described output support any scientific publication?

Yes

2.1.2 Is there a data availability statement provided along with the publication?

No

2.3 Software

2.3.1 Does the described output use or support any software?

Yes

<https://github.com/MicroStep-MIS/thunderstorm-nowcast-microstep/tree/main>

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 What type(s) of persistent identifier(s) are used for the described dataset / output?

None

3.1.1.2 Will you provide metadata for the described dataset / output?

Yes

CF (Climate and Forecast) Metadata Conventions

3.1.1.3 What type(s) of metadata?

- Descriptive
- Administrative

3.1.1.4 Do the metadata use standardised vocabularies?

Yes

3.1.1.5 Please provide URL/Description of used vocabularies

<https://www.ametsoc.org/index.cfm/ams/publications/glossary-of-meteorology/>

3.1.1.6 Are the metadata searchable?

No

3.1.1.8 Are keywords provided in the metadata?

No

3.1.1.9 Are metadata harvestable?

No

3.2.1 Repository

3.2.1.1 In which repository will the dataset / output be deposited?

Project NextCloud

<https://share.services.ai4os.eu/>

3.2.1.2 Is the selected repository a trusted source?

No

3.2.1.5 Does the repository(ies) assign datasets / outputs with persistent identifiers?

No

3.2.1.7 Does the repository support versioning?

Yes

3.2.2 Data

3.2.2.1 What is the described dataset / output title?

MMR-116_radar_data

3.2.2.2 How is the dataset / output shared?

Shared

3.2.2.5 Are there any methods or tools required to access the dataset / output?

No

3.2.2.8 Is the described dataset / output supported by a data access committee?

No

3.2.3 Metadata

3.2.3.1 Will you provide metadata even if the described dataset / output can not be openly shared?

Yes

Will remain shared on NextCloud

3.2.3.2 Under which license will metadata be provided?

Creative Commons Zero (CC0)

3.2.3.3 Do metadata provide information about how to access the described dataset / output?

Yes

3.2.3.4 Will metadata remain available after the dataset / output is no longer available?

Yes

3.3 Making data and other outputs interoperable

3.3.1 Does your (meta)data use a controlled vocabulary?

Yes

3.3.3 Have you applied a standard schema for your (meta)data?

Yes

Couldn't find it? Insert it manually

CF (Climate and Forecast) Metadata Conventions

3.3.7 Does the described dataset / output provide qualified references with other outputs?

No

3.4 Increasing data and other outputs reuse

3.4.2 What reusability and / or reproducibility methods are followed?

- Readme files
- Codebooks
- Analyses

3.4.3 Will you provide the described dataset / output in the public domain?

No

3.4.4 Do you intend to ensure (re)use by third parties after your project finishes?

No

3.4.5 Is provenance well documented?

No

4.1 Allocation of resources

4.1.1 What will be the cost of making the described output FAIR?

0

Euro

Storage

Direct cost

4.1.2 How will this cost be covered?

- Use of national infrastructure
- Collaboration with other Projects

5.1 Data Security

5.1.1 What security measures are followed?

- Passwords
- Physical access control

5.1.2 What conditions do the security measures meet?

- Data access
- Data storage

5.1.3 How will you preserve the described dataset / output in the long term?

Tape Backups, security access.

6.1 Ethical aspects

6.1.2 Does the described dataset / output contain sensitive information?

No

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

UC3 – AUTOMATED THERMOGRAPHY [AI module - TBBRDet]

Use case 3 leverages multispectral unmanned aerial vehicle (UAV) based imaging and AI to identify “hot spots” (thermal anomalies) in urban settings and thus improve the efficiency of energy-related systems. One of the two scenarios being examined is:

Thermal Bridges on Building Rooftops Detection (TBBRDet): Detecting thermal hotspots on building rooftops caused by thermal bridges to support urban planners and building owners with retrofitting plans.

Template: [Horizon Europe](#)

Type: [Dataset](#)

1.1 Brief description of the described research output

1.1.1 What kind of research output are you describing?

Models

1.1.2 Is it physical or digital?

Digital

1.1.3 Are you generating or re-using it?

New

1.1.4 What is the type of the described dataset?

Other

The main output is the AI models itself: A Swin-T. Alternative models are a MaskRCNN R50, TridentNet and FSAF, however these perform worse than the Swin-T, which is why the Swin-T is the main focus.

1.1.6 What is its expected size?

550MB

1.1.7 Why are you collecting/generating or re-using it?

To make informed decisions

The models are created to counteract the current inability to quickly and accurately pinpoint areas of unwanted heat loss leading to reduced energy efficiency in urban infrastructure. The overarching objective lies in automating the detection of thermographic anomalies to accelerate the implementation of necessary counter-measures and repairs. AI can be used to this end as either a stand-alone tool (TBBDet) or through integration into a pre-existing analysis pipeline (TUFSeg).

1.1.8 What is its origin / provenance?

Derived from UC3 input data

1.1.9 To whom might it be useful ('data utility')?

- Researchers
- Decision makers
- Education
- Economy

2.1 Publications

2.1.1 Does the described output support any scientific publication?

Yes

2.1.2 Is there a data availability statement provided along with the publication?

Yes

<https://github.com/emvollmer/TBBDet>

2.2 Datasets

2.2.1 Does the described output use or support any published dataset?

Yes

2.3 Software

2.3.1 Does the described output use or support any software?

No

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 What type(s) of persistent identifier(s) are used for the described dataset / output?

None

DOI

There are plans for submitting the models to Zenodo automatically

3.1.1.2 Will you provide metadata for the described dataset / output?

Yes

Apart from the data that can be found in GitHub repository, the AI4EOSC dashboard has been developed taking into account the FAIR principles. That is why, a list of metadata terms have been added to any single model, including:

- Title, description, summary
- DOI of the data science application
- Documentation, source code, docker image and dataset links
- How the model should be cited
- Creation and update dates
- Libraries involved

- Type of task the model does
- Category and type of data that the model performs

3.1.1.3 What type(s) of metadata?

- Descriptive
- Structural
- Reference

3.1.1.4 Do the metadata use standardised vocabularies?

No

3.1.1.6 Are the metadata searchable?

Yes

3.1.1.7 How are searchable metadata provided?

Registry/Catalogue

AI4EOSC Platform

3.1.1.8 Are keywords provided in the metadata?

Yes

3.1.1.9 Are metadata harvestable?

No

3.2.1 Repository

3.2.1.1 In which repository will the dataset / output be deposited?

Zenodo

<https://zenodo.org/>

3.2.1.2 Is the selected repository a trusted source?

No

3.2.1.5 Does the repository(ies) assign datasets / outputs with persistent identifiers?

Yes

3.2.1.6 Does the repository(ies) resolve the identifiers to a digital object?

Landing page

3.2.1.7 Does the repository support versioning?

Yes

3.2.2 Data

3.2.2.1 What is the described dataset / output title?

Thermal Bridges on Building Rooftops Detection (TBBRDet)

3.2.2.2 How is the dataset / output shared?

Open

3.2.2.5 Are there any methods or tools required to access the dataset / output?

Yes

Couldn't find it? Insert it manually

AI4EOSC platform, Python or Docker

3.2.2.7 Please provide information about the tools needed to access the dataset / output.

AI4EOSC platform, Python or Docker

3.2.2.8 Is the described dataset / output supported by a data access committee?

No

3.2.3 Metadata

3.2.3.1 Will you provide metadata even if the described dataset / output can not be openly shared?

Yes

3.2.3.2 Under which license will metadata be provided?

Creative Commons Zero (CC0)

3.2.3.3 Do metadata provide information about how to access the described dataset / output?

No

3.2.3.4 Will metadata remain available after the dataset / output is no longer available?

Yes

3.3 Making data and other outputs interoperable

3.3.1 Does your (meta)data use a controlled vocabulary?

Yes

<https://github.com/ai4os>

3.3.2 If you created the vocabulary, where can it be found?

<https://github.com/ai4os>

3.3.3 Have you applied a standard schema for your (meta)data?

No

3.3.4 Will you provide a mapping to more commonly used ontologies?

Yes

We will use an extension of PROV-O to improve the provenance information

3.3.7 Does the described dataset / output provide qualified references with other outputs?

Yes

The model provenance system within AI4EOSC project will track the relationship among the models and any other part of the workflow, like the input data.

3.4 Increasing data and other outputs reuse

3.4.1 What internationally recognised licence will you use for your dataset / output?

BSD 3-Clause "New" or "Revised" License (BSD-3-Clause)

3.4.2 What reusability and / or reproducibility methods are followed?

- Codebooks

- Other

AI4EOSC provenance system

3.4.3 Will you provide the described dataset / output in the public domain?

Yes

3.4.4 Do you intend to ensure (re)use by third parties after your project finishes?

Yes

3.4.5 Is provenance well documented?

Yes

AI4EOSC provenance system

3.4.6 What documented procedures for quality assurance do you have in place?

Use of tools for automatic checks

4.1 Allocation of resources

4.1.1 What will be the cost of making the described output FAIR?

0

Euro

- Storage
- Re-use

Direct cost

4.1.2 How will this cost be covered?

Use of institution infrastructure

5.1 Data Security

5.1.3 How will you preserve the described dataset / output in the long term?

Github based version controls and back-up

6.1 Ethical aspects

6.1.2 Does the described dataset / output contain sensitive information?

No

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

UC3 – AUTOMATED THERMOGRAPHY [AI module - TUFSeg]

Use case 3 leverages multispectral unmanned aerial vehicle (UAV) based imaging and AI to identify “hot spots” (thermal anomalies) in urban settings and thus improve the efficiency of energy-related systems. One of the two scenarios being examined is:

- Thermal Urban Feature Segmentation (TUFSeg): Detecting thermal hotspots caused by common urban features (cars, manholes, streetlamps, etc.) to aid district heating network operators in their search for pipeline leakages by automatically removing these false alarms from the list of potential suspects.

Template: [Horizon Europe](#)

Type: [Dataset](#)

1.1 Brief description of the described research output

1.1.1 What kind of research output are you describing?

Models

1.1.2 Is it physical or digital?

Digital

1.1.3 Are you generating or re-using it?

New

1.1.4 What is the type of the described dataset?

Other

The main output are the AI models itself: U-Net for TUFSeg

1.1.6 What is its expected size?

780MB

1.1.7 Why are you collecting/generating or re-using it?

To make informed decisions

The models are created to counteract the current inability to quickly and accurately pinpoint areas of unwanted heat loss leading to reduced energy efficiency in urban infrastructure. The overarching objective lies in automating the detection of thermographic anomalies to accelerate the implementation of necessary counter-measures and repairs. AI can be used to this end as either a stand-alone tool (TBBRDet) or through integration into a pre-existing analysis pipeline (TUFSeg).

1.1.8 What is its origin / provenance?

The models are created to counteract the current inability to quickly and accurately pinpoint areas of unwanted heat loss leading to reduced energy efficiency in urban infrastructure. The overarching objective lies in automating the detection of thermographic anomalies to accelerate the implementation of necessary counter-measures and repairs. AI can be used to this end as either a stand-alone tool (TBBRDet) or through integration into a pre-existing analysis pipeline (TUFSeg).

1.1.9 To whom might it be useful ('data utility')?

- Researchers
- Research communities
- Decision makers
- Education

2.1 Publications

2.1.1 Does the described output support any scientific publication?

Yes

2.1.2 Is there a data availability statement provided along with the publication?

Yes

<https://github.com/emvollmer/TUFSeg>

2.2 Datasets

2.2.1 Does the described output use or support any published dataset?

a. Yes

[Hyperspectral \(RGB + Thermal\) drone images of Karlsruhe, Germany - Raw images for the Thermal Bridges on Building Rooftops \(TBBR\) dataset](#)

b. Yes

[Thermal Bridges on Building Rooftops - Hyperspectral \(RGB + Thermal + Height\) drone images of Karlsruhe, Germany, with thermal bridge annotations](#)

2.3 Software

2.3.1 Does the described output use or support any software?

Yes

<https://github.com/emvollmer/TUFSeg>

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 What type(s) of persistent identifier(s) are used for the described dataset / output?

None

3.1.1.2 Will you provide metadata for the described dataset / output?

Yes

Apart from the data that can be found in GitHub repository, the AI4EOSC dashboard has been developed taking into account the FAIR principles. That is why, a list of metadata terms have been added to any single model, including:

- Title, description, summary
- DOI of the data science application
- Documentation, source code, docker image and dataset links
- How the model should be cited
- Creation and update dates
- Libraries involved
- Type of task the model does
- Category and type of data that the model performs

3.1.1.3 What type(s) of metadata?

- Descriptive
- Administrative
- Structural

3.1.1.4 Do the metadata use standardised vocabularies?

No

3.1.1.6 Are the metadata searchable?

Yes

3.1.1.7 How are searchable metadata provided?

Registry/Catalogue

AI4EOSC Dashboard

3.1.1.8 Are keywords provided in the metadata?

Yes

3.1.1.9 Are metadata harvestable?

No

3.2.1 Repository

3.2.1.1 In which repository will the dataset / output be deposited?

Zenodo

<https://zenodo.org/>

3.2.1.2 Is the selected repository a trusted source?

No

3.2.1.5 Does the repository(ies) assign datasets / outputs with persistent identifiers?

Yes

3.2.1.6 Does the repository(ies) resolve the identifiers to a digital object?

Landing page

3.2.1.7 Does the repository support versioning?

Yes

3.2.2 Data

3.2.2.1 What is the described dataset / output title?

Thermal Urban Feature Segmentation (TUFSeg)

3.2.2.2 How is the dataset / output shared?

Open

<https://github.com/emvollmer/TUFSeg>

3.2.2.5 Are there any methods or tools required to access the dataset / output?

Yes

Couldn't find it? Insert it manually

git, AI4EOSC Dashboard

3.2.2.6 Please provide information about the method(s) needed to access the dataset / output.

GIT, AI4EOSC dashboard

3.2.2.7 Please provide information about the tools needed to access the dataset / output.

AI4EOSC dashboard, GIT

3.2.2.8 Is the described dataset / output supported by a data access committee?

No

3.2.3 Metadata

3.2.3.1 Will you provide metadata even if the described dataset / output can not be openly shared?

Yes

3.2.3.2 Under which license will metadata be provided?

Creative Commons Zero (CC0)

3.2.3.3 Do metadata provide information about how to access the described dataset / output?

No

3.2.3.4 Will metadata remain available after the dataset / output is no longer available?

Yes

3.3 Making data and other outputs interoperable

3.3.1 Does your (meta)data use a controlled vocabulary?

Yes

<https://github.com/ai4os>

3.3.2 If you created the vocabulary, where can it be found?

<https://github.com/ai4os>

3.3.3 Have you applied a standard schema for your (meta)data?

No

3.3.4 Will you provide a mapping to more commonly used ontologies?

Yes

We will use an extension of PROV-O to improve the provenance information

3.3.7 Does the described dataset / output provide qualified references with other outputs?

Yes

The model provenance system within AI4EOSC project will track the relationship among the models and any other part of the workflow, like the input data.

3.4 Increasing data and other outputs reuse

3.4.1 What internationally recognised licence will you use for your dataset / output?

BSD 3-Clause "New" or "Revised" License (BSD-3-Clause)

3.4.2 What reusability and / or reproducibility methods are followed?

- Codebooks
- Other

AI4EOSC Provenance System

3.4.3 Will you provide the described dataset / output in the public domain?

Yes

3.4.4 Do you intend to ensure (re)use by third parties after your project finishes?

Yes

3.4.5 Is provenance well documented?

Yes

AI4EOSC provenance system

3.4.6 What documented procedures for quality assurance do you have in place?

Use of tools for automatic checks

4.1 Allocation of resources

4.1.1 What will be the cost of making the described output FAIR?

0

Euro

- Storage
- Re-use

Direct cost

4.1.2 How will this cost be covered?

Use of institution infrastructure

4.1.3 Identify the people who will be responsible and their role(s) in the management of the described output

Elena Vollmer (orcid:0000-0002-8805-3726)

5.1 Data Security

5.1.3 How will you preserve the described dataset / output in the long term?

Github based version controls and back-up

6.1 Ethical aspects

6.1.2 Does the described dataset / output contain sensitive information?

No

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

UC2 – INTEGRATED PLANT PROTECTION [AI Module]

This use case aims to determine the risk of disease in crops and determine the phases of plant growth and the condition of crops. The developed AI models are going to be integrated into existing national advisory platforms, operated by WODR and PSNC.

Currently, WODR and PSNC operate a national advisory platform for farmers (eDWIN), which includes a network of meteorological ground stations, the Farm Management System, and ground observations of the occurrence of diseases and pests. The current solutions are based on predictive mathematical models.

With AI4EOSC, the plan is to add to the current mathematical prediction models the ML/DL-based models used for the early detection of plant diseases and add new sources of data. The initial focus is on detection of the fungal disease occurrences in rye and sugar beet.

Template: [Horizon Europe](#)

Type: [Dataset](#)

1.1 Brief description of the described research output

1.1.1 What kind of research output are you describing?

Models

1.1.2 Is it physical or digital?

Digital

1.1.3 Are you generating or re-using it?

New

1.1.4 What is the type of the described dataset?

Other

AI model

1.1.6 What is its expected size?

500MB

1.1.7 Why are you collecting/generating or re-using it?

- To make informed decisions
- To develop a product

1.1.9 To whom might it be useful ('data utility')?

- Researchers
- Decision makers

2.1 Publications

2.1.1 Does the described output support any scientific publication?

No

2.1.2 Is there a data availability statement provided along with the publication?

No

2.2 Datasets

2.2.1 Does the described output use or support any published dataset?

No

2.3 Software

2.3.1 Does the described output use or support any software?

Yes

<https://github.com/ai4os-hub/integrated-plant-protection>

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 What type(s) of persistent identifier(s) are used for the described dataset / output?

None

DOI

There are plans for submitting the models to Zenodo automatically

3.1.1.2 Will you provide metadata for the described dataset / output?

Yes

Apart from the data that can be found in GitHub repository, the AI4EOSC dashboard has been developed taking into account the FAIR principles. That is why, a list of metadata terms have been added to any single model, including:

- Title, description, summary
- DOI of the data science application
- Documentation, source code, docker image and dataset links
- How the model should be cited
- Creation and update dates
- Libraries involved
- Type of task the model does
- Category and type of data that the model performs

3.1.1.3 What type(s) of metadata?

- Descriptive
- Structural
- Reference

3.1.1.4 Do the metadata use standardised vocabularies?

No

3.1.1.6 Are the metadata searchable?

Yes

3.1.1.7 How are searchable metadata provided?

Registry/Catalogue

AI4EOSC Platform

3.1.1.8 Are keywords provided in the metadata?

Yes

3.1.1.9 Are metadata harvestable?

No

3.2.1 Repository

3.2.1.1 In which repository will the dataset / output be deposited?

Zenodo

<https://zenodo.org/>

3.2.1.2 Is the selected repository a trusted source?

No

3.2.1.5 Does the repository(ies) assign datasets / outputs with persistent identifiers?

Yes

3.2.1.6 Does the repository(ies) resolve the identifiers to a digital object?

Landing page

3.2.1.7 Does the repository support versioning?

Yes

3.2.2 Data

3.2.2.1 What is the described dataset / output title?

Integrated Plant Protection

3.2.2.2 How is the dataset / output shared?

Open

3.2.2.5 Are there any methods or tools required to access the dataset / output?

Yes

Couldn't find it? Insert it manually

AI4EOSC platform, Python or Docker

3.2.2.7 Please provide information about the tools needed to access the dataset / output.

AI4EOSC platform, Python or Docker

3.2.2.8 Is the described dataset / output supported by a data access committee?

No

3.2.3 Metadata

3.2.3.1 Will you provide metadata even if the described dataset / output can not be openly shared?

Yes

3.2.3.2 Under which license will metadata be provided?

Creative Commons Zero (CC0)

3.2.3.3 Do metadata provide information about how to access the described dataset / output?

No

3.2.3.4 Will metadata remain available after the dataset / output is no longer available?

Yes

3.3 Making data and other outputs interoperable

3.3.1 Does your (meta)data use a controlled vocabulary?

Yes

<https://github.com/ai4os>

3.3.2 If you created the vocabulary, where can it be found?

<https://github.com/ai4os>

3.3.3 Have you applied a standard schema for your (meta)data?

No

3.3.4 Will you provide a mapping to more commonly used ontologies?

Yes

We will use an extension of PROV-O to improve the provenance information

3.3.7 Does the described dataset / output provide qualified references with other outputs?

Yes

The model provenance system within AI4EOSC project will track the relationship among the models and any other part of the workflow, like the input data.

3.4 Increasing data and other outputs reuse

3.4.1 What internationally recognised licence will you use for your dataset / output?

MIT License

3.4.2 What reusability and / or reproducibility methods are followed?

- Readme files
- Codebooks
- Other

AI4EOSC provenance system, Docker

3.4.3 Will you provide the described dataset / output in the public domain?

Yes

3.4.4 Do you intend to ensure (re)use by third parties after your project finishes?

Yes

3.4.5 Is provenance well documented?

Yes

AI4EOSC provenance system

3.4.6 What documented procedures for quality assurance do you have in place?

Use of tools for automatic checks

4.1 Allocation of resources

4.1.1 What will be the cost of making the described output FAIR?

0

Euro

- Storage
- Re-use

Direct cost

4.1.2 How will this cost be covered?

Use of institution infrastructure

5.1 Data Security

5.1.3 How will you preserve the described dataset / output in the long term?

Github based version controls and back-up

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on sharing the described dataset / output?

no

6.1.2 Does the described dataset / output contain sensitive information?

No

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

UC1 – AGROMETEOROLOGY [AI Module]

Anthropogenic-driven climate change is expected to affect, mainly to increase, frequency and intensity of various types of extreme events. One of them are thunderstorms that generate adverse phenomena for farmers:

- high winds can break and damage buildings, equipment, material and plants;
- hail can cause leaf damage reducing yield or destroying plants, machines, cars, and buildings;
- flash floods can endanger humans, animals, damage machines and equipment, can lead to loss of topsoil as well as damage to crops.

In this use case we process radar imagery - utilising AI ensemble stacking and federated learning - to create added value products for improving farmers activity, namely timely and precise warnings of thunderstorms for farmers. As an enhancement, the farmer wants to know when exactly thunderstorms will hit his area so that he can get everyone to safety. Nowcasting can provide localised warnings around 30 minutes ahead of thunderstorms. We integrated the use case with the platform, which enabled us to use existing AI model templates and state-of-the-art computing resources, perform automated testing and use other platform tools to tackle the problem. The model is a neural network currently, where the number of layers can be set by the user.

Inputs: structured data: images and/or csv

Outputs: structured data: images and/or csv

Template: [Horizon Europe](#)

Type: [Dataset](#)

1.1 Brief description of the described research output

1.1.1 What kind of research output are you describing?

Models

1.1.2 Is it physical or digital?

Digital

1.1.3 Are you generating or re-using it?

New

1.1.4 What is the type of the described dataset?

Derived or compiled

Primary data were collected previously by MicroStep-MIS radar. For the purpose of this project we have created a secondary dataset.

1.1.6 What is its expected size?

Size of code + model is about 1 MB

1.1.7 Why are you collecting/generating or re-using it?

To make informed decisions

Models for thunderstorm forecasting

1.1.8 What is its origin / provenance?

The model is developed within the AI4EOSC Project and it is trained using UC1 input data

1.1.9 To whom might it be useful ('data utility')?

- Researchers
- Decision makers

2.1 Publications

2.1.1 Does the described output support any scientific publication?

No

2.1.2 Is there a data availability statement provided along with the publication?

Yes

UC1 input data

2.2 Datasets

2.2.1 Does the described output use or support any published dataset?

Yes

2.3 Software

2.3.1 Does the described output use or support any software?

No

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 What type(s) of persistent identifier(s) are used for the described dataset / output?

None

DOI

There are plans for publishing automatically model code at Zenodo and mint a DOI

3.1.1.2 Will you provide metadata for the described dataset / output?

Yes

Apart from the data that can be found in GitHub repository, the AI4EOSC dashboard has been developed taking into account the FAIR principles. That is why, a list of metadata terms have been added to any single model, including:

- Title, description, summary
- DOI of the data science application
- Documentation, source code, docker image and dataset links
- How the model should be cited
- Creation and update dates

- Libraries involved
- Type of task the model does
- Category and type of data that the model performs

3.1.1.3 What type(s) of metadata?

- Descriptive
- Administrative
- Reference

3.1.1.4 Do the metadata use standardised vocabularies?

No

3.1.1.6 Are the metadata searchable?

Yes

3.1.1.7 How are searchable metadata provided?

Registry/Catalogue

AI4EOSC Dashboard

3.1.1.8 Are keywords provided in the metadata?

Yes

3.1.1.9 Are metadata harvestable?

No

3.2.1 Repository

3.2.1.1 In which repository will the dataset / output be deposited?

Zenodo

<https://zenodo.org/>

3.2.1.2 Is the selected repository a trusted source?

No

3.2.1.5 Does the repository(ies) assign datasets / outputs with persistent identifiers?

Yes

3.2.1.6 Does the repository(ies) resolve the identifiers to a digital object?

Landing page

3.2.1.7 Does the repository support versioning?

Yes

3.2.2 Data

3.2.2.2 How is the dataset / output shared?

Shared

3.2.2.5 Are there any methods or tools required to access the dataset / output?

Yes

Couldn't find it? Insert it manually

Git, AI4EOSC dashboard

3.2.2.7 Please provide information about the tools needed to access the dataset / output.

GIT, AI4EOSC Dashboard

3.2.2.8 Is the described dataset / output supported by a data access committee?

No

3.2.3 Metadata

3.2.3.1 Will you provide metadata even if the described dataset / output can not be openly shared?

Yes

3.2.3.2 Under which license will metadata be provided?

Creative Commons Zero (CC0)

3.2.3.3 Do metadata provide information about how to access the described dataset / output?

No

3.2.3.4 Will metadata remain available after the dataset / output is no longer available?

Yes

3.3 Making data and other outputs interoperable

3.3.1 Does your (meta)data use a controlled vocabulary?

Yes

<https://github.com/ai4os>

3.3.2 If you created the vocabulary, where can it be found?

<https://github.com/ai4os>

3.3.3 Have you applied a standard schema for your (meta)data?

No

3.3.4 Will you provide a mapping to more commonly used ontologies?

Yes

We will use an extension of PROV-O to improve the provenance information

3.3.7 Does the described dataset / output provide qualified references with other outputs?

Yes

The model provenance system within AI4EOSC project will track the relationship among the models and any other part of the workflow, like the input data.

3.4 Increasing data and other outputs reuse

3.4.1 What internationally recognised licence will you use for your dataset / output?

MIT License

3.4.2 What reusability and / or reproducibility methods are followed?

- Readme files
- Codebooks

- Other

AI4EOSC provenance system

3.4.3 Will you provide the described dataset / output in the public domain?

Yes

3.4.4 Do you intend to ensure (re)use by third parties after your project finishes?

Yes

3.4.5 Is provenance well documented?

Yes

AI4EOSC provenance system

3.4.6 What documented procedures for quality assurance do you have in place?

Use of tools for automatic checks

We use bandit and flake8 for quality control of the code.

4.1 Allocation of resources

4.1.1 What will be the cost of making the described output FAIR?

0

Euro

- Storage
- Re-use

Direct cost

4.1.2 How will this cost be covered?

Use of institution infrastructure

5.1 Data Security

5.1.1 What security measures are followed?

Passwords

5.1.2 What conditions do the security measures meet?

- Data storage
- Data recovery

Github based version controls and back-up

5.1.3 How will you preserve the described dataset / output in the long term?

Github based version controls and back-up

6.1 Ethical aspects

6.1.2 Does the described dataset / output contain sensitive information?

No

6.1.3 Does the described dataset / output contain personal data?

No

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

UC2 – INTEGRATED PLANT PROTECTION [Input Data]

The aim is to determine the risk of disease in agricultural crops and determine the phases of plant growth and the condition of crops. The developed AI models are going to be integrated into existing national advisory platforms, operated by WODR and PSNC.

Currently WODR and PSNC operate a national advisory platform for farmers (eDWIN), which includes a network of meteorological ground stations, the Farm Management System, and ground observations of the occurrence of diseases and pests. The current solutions are based on predictive mathematical models.

With AI4EOSC, the plan is to add to the current mathematical prediction models the ML/DL-based models used for early detection of the plant diseases and add new sources of the data. The initial focus is on detection of the fungal disease occurrences on rye and sugar beet.

Widely collected images of both plants are used as an input for the training. There are different sources of the images.. Some come from other projects, some are being actively collected by our partners and some are collected by us. We are also near the end of getting the synthetic data which will supplement our training datasets. The expected result is the development of a model that can determine whether the expected disease has shown symptoms.

Template: [Horizon Europe](#)

Type: [Dataset](#)

1.1 Brief description of the described research output

1.1.1 What kind of research output are you describing?

Research Data

1.1.2 Is it physical or digital?

Digital

1.1.3 Are you generating or re-using it?

New

We are mostly generating the data but some of the images come from the open source Edwin project.

1.1.4 What is the type of the described dataset?

Observational

The described dataset is of both types. Most of it is the primary data (images being actively collected), but we also have data from other project which has already gone through processing/analysis.

1.1.6 What is its expected size?

5GB

1.1.7 Why are you collecting/generating or re-using it?

- To obtain information
- To make informed decisions

The main objective is to develop a model that will help in the determination of fungal diseases. To ensure the model quality a high quality input data is needed.

As part of our activities in the AI4EOSC project, we are focusing on maximizing our data preparation activities. We are actively acquiring plant images from our

partners, who often work in the fields and provide us with datasets composed of images from the latest season.

Apart from that we are trying to reuse the already existing data by searching the open source datasets from the Edwin project. There are many photos of many plant species, and we managed to get photos of rye and sugar beet.

1.1.8 What is its origin / provenance?

Edwin project, partners in the AI4EOSC project and synthetic data.

1.1.9 To whom might it be useful ('data utility')?

- Researchers
- Industry
- Other

Main goal of the data is to be useful to farmers in making decisions regarding the crop infestations.

2.1 Publications

2.1.1 Does the described output support any scientific publication?

No

2.1.2 Is there a data availability statement provided along with the publication?

No

2.3 Software

2.3.1 Does the described output use or support any software?

Yes

<https://github.com/ai4os-hub/integrated-plant-protection>

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 What type(s) of persistent identifier(s) are used for the described dataset / output?

None

DOI

There are plans to publish data at Zenodo

3.1.1.2 Will you provide metadata for the described dataset / output?

Yes

Plant name	
<input type="checkbox"/> Beet	8
<input type="checkbox"/> Rye	5
Stage	
<input type="checkbox"/> Early	6
<input type="checkbox"/> Mid	3
<input type="checkbox"/> Late	1
Disease	
<input type="checkbox"/> Cercospora Leaf Spot Disease	6
<input type="checkbox"/> Brown Rust Of Rye	4
<input type="checkbox"/> Unaffected	3
Source	
<input type="checkbox"/> Wodr	5
<input type="checkbox"/> Edwin	4
<input type="checkbox"/> Psnc	3
<input type="checkbox"/> Ior	1
Project Name	
<input type="checkbox"/> Ai4Eosc	9
<input type="checkbox"/> Edwin	4

Image 1

3.1.1.3 What type(s) of metadata?

Descriptive

3.1.1.4 Do the metadata use standardised vocabularies?

Yes

3.1.1.5 Please provide URL/Description of used vocabularies

<https://agrovoc.fao.org/browse/agrovoc/en/>

Although the use of standardised vocabularies are not yet applied, there are plans to use :

<https://agrovoc.fao.org/browse/agrovoc/en/>

https://www.eppo.int/RESOURCES/eppo_databases/eppo_codes

3.1.1.6 Are the metadata searchable?

Yes

3.1.1.7 How are searchable metadata provided?

- Registry/Catalogue
- Metadata repository

It will be on Zenodo

3.1.1.8 Are keywords provided in the metadata?

Yes

3.1.1.9 Are metadata harvestable?

Yes

It will be on Zenodo

3.2.1 Repository

3.2.1.1 In which repository will the dataset / output be deposited?

Zenodo

zenodo.org

3.2.1.2 Is the selected repository a trusted source?

No

3.2.1.5 Does the repository(ies) assign datasets / outputs with persistent identifiers?

Yes

3.2.1.6 Does the repository(ies) resolve the identifiers to a digital object?

Landing page

3.2.1.7 Does the repository support versioning?

Yes

3.2.2 Data

3.2.2.2 How is the dataset / output shared?

Open

3.2.2.5 Are there any methods or tools required to access the dataset / output?

No

3.2.2.8 Is the described dataset / output supported by a data access committee?

No

3.2.2.9 Please specify how the dataset / output will be accessed during and after the project ends

Zenodo

3.2.3 Metadata

3.2.3.2 Under which license will metadata be provided?

Creative Commons Zero (CC0)

3.3 Making data and other outputs interoperable

3.3.1 Does your (meta)data use a controlled vocabulary?

No

Although the use of standardised vocabularies are not yet applied, there are plans to use :

<https://agrovoc.fao.org/browse/agrovoc/en/>

https://www.eppo.int/RESOURCES/eppo_databases/eppo_codes

3.3.3 Have you applied a standard schema for your (meta)data?

No

3.3.4 Will you provide a mapping to more commonly used ontologies?

No

3.3.7 Does the described dataset / output provide qualified references with other outputs?

No

3.4 Increasing data and other outputs reuse

3.4.4 Do you intend to ensure (re)use by third parties after your project finishes?

Yes

3.4.5 Is provenance well documented?

Yes

4.1 Allocation of resources

4.1.1 What will be the cost of making the described output FAIR?

0

Euro

Storage

Direct cost

4.1.2 How will this cost be covered?

Use of institution infrastructure

6.1 Ethical aspects

6.1.2 Does the described dataset / output contain sensitive information?

No

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Powered by

