Data Management Plan GEO.INFORMED

Project Name Spatio-temporal deep learning workflows for transforming remote sensing data into geo-indicators for environmental policy support - Data Management Plan GEO.INFORMED

Project Identifier S006421N

Grant Title S006421N

Principal Investigator / Researcher Ben Somers

Project Data Contact ben.somers@kuleuven.be; +32 16 37 91 01

Description The general objective of this project is to develop deep learning workflows for deriving environmental indicators (parameters) from freely available remote sensing data sources. The main end-users are the policy supporting organizations in the Flemish Government's Environmental Policy Domain. We will combine freely available remote sensing data with in-situ data, either obtained from existing monitoring efforts by the end-users or obtained within the framework of the project (= input data). We will create indicator maps and the deep learning scripts needed to derive these maps (= output data).

Institution KU Leuven

1. General Information

Name applicant

Stien Heremans (INBO), Matthew Blashko (KU Leuven), Ils Reusen (VITO) and Ben Somers (KU Leuven)

FWO Project Number & Title

SBO project S006421N - Spatio-temporal deep learning workflows for transforming remote sensing data into geo-indicators for environmental policy support

Affiliation

- KU Leuven
- Other

2. Data description

Will you generate/collect new data and/or make use of existing data?

- Generate new data
- Reuse existing data

Describe the origin, type and format of the data (per dataset) and its (estimated) volume, ideally per objective or WP of the project. You might consider using the table in the guidance.

WP 1: System analysis of policy needs

Input data: (1) use case descriptions (Origin: Google Form, workshops

; Type: qualitative observational; Format: Excel; Estimated volume: 50 KB);

Output data: (1) Description of core geo-indicators (Type: qualitative; Format: Excel and Word, Estimated Volume: 50 KB)

WP 2: Feasibility analysis and core indicator selection

Input data: (1) Description of core geo-indicators (Origin: WP1; Type: qualitative; Format: Excel and Word, Estimated Volume: 50 KB); (2) Deep learning and remote sensing literature (Origin: Google Scholar; Type: reference; Format: Pdfs in Zotero database; Estimated Volume: 100 MB)

WP3: Data collection

Input data: (1) Existing reference data for use cases (Origin: Geopunt.be and end-users; Type: observational, experimental, derived or combined; Format: Shapefile, Excel, Access database(s), SQL database(s); Estimated volume: 5 GB); (2) New reference data for use cases (Origin: Fieldwork and orthophoto/satellite interpretation; Type: observational, experimental; Format: Shapefile, Excel, Access database(s), SQL database(s); Estimated volume: 1 GB) (3) Satellite imagery (Origin: Terrascope; Type: observational; Format: safe format, cloud-optimized geotiffs(COG) or netcfd ; Estimated volume: 2194 GB/year)

Output data: (1) Database with deep learning reference data (Type: quantitative observational; Format: geodatabases; Estimated volume: 5 GB); (2) Spectro-temporal fused satellite image stacks (Type: compiled; Format: cloud-optimized geotiffs(COG) or netcfd; Estimated volume: 100 s of GB); (3) Labelled image patches (Type: compiled; Format: shapefiles, JSON, GeoJSON, geodatabase/XML/character delimited files (e.g. csv); Estimated volume: order of GB)

WP4: Machine learning algorithm design

Input data: (1) Labelled image patches (Origin: WP3; Type: compiled; Format: shapefiles, JSON, GeoJSON, geodatabase/XML/character delimited files (e.g. csv); Estimated volume: order of GB), (2) existing Pytorch code base (Origin: Pytorch open source machine learning library; Type: canonical ; Format: software; Estimated volume: several megabytes)

Output data: (1) new Pytorch code blocks and scripts (Type: experimental; ; Format: software; Estimated volume: several megabytes).

WP5: Sampling design for cal/val

Input data: (1) Labelled image patches (Origin: WP3; Type: compiled; Format: shapefiles, JSON, GeoJSON, geodatabase/XML/character delimited files (e.g. csv); Estimated volume: order of GB), (2) existing Pytorch code base (Origin: Pytorch open source machine learning library; Type: canonical ; Format: software; Estimated volume: several megabytes)

Output data: (1) Pytorch scripts (Type: experimental; ; Format: software; Estimated volume: several megabytes); (2) Ground sampling design for use cases (Type: simulated; Format: shapefiles; Estimated volume: 500 MB)

WP6: Workflow assimilation and user feedback

Input data: (1) Terrascope raw and/or derived products (Land and Water) Origin:

Terrascope; Type: observational ; Format: safe format, cloud-optimized geotiffs(COG) or netcfd ; Estimated volume: 2194 GB/year); (2) new Pytorch code blocks and scripts (Type: experimental; ; Format: software; Estimated volume: several megabytes). Source code is a very small component of data storage, which is dominated by databases and annotations above;

(3) New in situ information (from WP5/WP3) (Origin: Geopunt.be and end-users; Type: observational, experimental, derived or combined; Format: standardized format like GeoJSON; Estimated volume: order of GB)

Output data: (1) Workflow (pre-processing, post-processing, and pre-trained model integration) code (Type: pre-operational ; Format: software; Estimated volume: several megabytes). (2) Geo-indicatoren (Type: combined; Format: cloud-optimized geotiffs(COG) or netcfd; Estimated volume: 10 s of GB); (3) Viewer (Type: combined; Format: geodatabase and code; volume: 36,5GB/year and 1,2GB)

The answers to this section were checked by:

- Other

DMP responsible at INBO

3. Legal & ethical issues

Will you use personal data? If so, shortly describe the kind of personal data you will use (add the reference to your file in your host institution's privacy register - not relevant yet)

- No

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, add the reference to the formal approval by the relevant ethical review committee(s)

- No

Does your work possibly result in research data with potential for tech transfer and valorisation? Will IP restrictions be claimed for the data you created? If so, for what data and which restrictions will be asserted?

- Yes

In the ‘samenwerkingsovereenkomst’, it is indicated that joint results can be exploited through a joint ownership agreement in case of a patent application. For other types of valorization, a written agreement is needed from all partners involved in obtaining the result. Once the project has been completed for more than 24 months, joint results can be valorized by a single partner without the consent from the other involved partners.

Do existing 3rd party agreements restrict dissemination or exploitation of the data you (re)use? If so, to what data do they relate and what restrictions are in

place?

- Yes

The reference data obtained from the end-users in WP3 will remain their property, unless already made publicly available. We can only use these data, but not further disseminate them.

The answers to this section were checked by:

- Others

4. Documentation & metadata

What documentation will be provided to enable reuse of the data collected/generated in this project?

For each use case, a metadata file with a description of the reference data will be created.

The Pytorch scripts will be documented on Github.

Terrascope documentation and metadata available on https://terrascope.be/nl

Will a metadata standard be used? If so, describe in detail which standard will be used. If no, state in detail which metadata will be created to make the data easy/easier to find and reuse.

- Yes

For the following data we will not use a metadata standard:

For reference data collected from the end-users in WP3, a text file will be made with a reference to the existing metadata files at the end-user organization. In case this does not exist, a text file will be made containing the date that the data was collected, the coordinates and/or study area, the people that collected the data, a short explanation of each variable and its units, the file format, and user rights and acknowledgements.

For reference data collected on the field in WP3, a text file will be made containing the date that the data was collected, the coordinates and/or study area, the people that collected the data, a short explanation of each variable and its units, the file format, and user rights and acknowledgements.

For scripts generated in WP4-5, metadata will be added in the scripts either as comments or separate text sections (Jupyter Notebooks).

For the following data we will use a metadata standard:

For satellite derived products metadata are included in GeoJSON file format. The GeoJSON implementation is based on the following standards: https://www.ogc.org/standards/eo-geojson. OGC-17-003r1 – OGC EO Dataset Metadata GeoJSON(-LD) Encoding Standard OGC-17-084-v0.9.0-D2 - GeoJSON EO Collection Metadata Encoding Standard

For the spectro-temporal fused satellite image stacks created in WP3: standardized openEO formats https://openeo.org/platform/

For the labelled image patches created in WP3; labelling via shapefile of other labelling tool not decided yet; many standards available: JSON (VGG,COCO), pascal VOC XML, ….

For the workflows generated in WP6, standardized OpenEO; https://openeo.org/platform/

The answers to this section were checked by:

- Other

5. Data storage & back up during the FWO project

Where will the data be stored?

Literature database: stored in Zotero, pdfs in KU Leuven Box account (backed up in the cloud) Reference data: stored on AWS (backed up) and backed up also in Google Drive (INBO enterprise license with unlimited storage) Metadata on the use cases: stored on KULeuven OneDrive (backed up in the cloud) and on the Microsoft Teams pages of this project (backed up in the cloud) Scripts and code blocks: stored in Github and on the PSI server: RAID system with multiple physical storage locations, and tape backup for selected data checkpoints

How is back up of the data provided?

All storage options mentioned above have an integrated back-up option

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely. If no or insufficient storage or backup capacities are available then explain how this will be taken care of.

- Yes

Yes. The satellite imagery and derived products can be accommodated on the Terrascope platform at VITO for the duration of the BELSPO funded Terrascope contract (currently up to mid-2023). The reference data take only 11 GB, while the storage capacity on INBO’s AWS server is several TB. Scripts can be freely stored on Github. For metadata, documentation and literature, three cloud platforms are available: KULeuven’s OneDrive/Box (2TB), VITO’s Microsoft Teams (25TB) and INBO’s Google Drive (unlimited).

What are the expected costs for data storage and back up during the project? How will these costs be covered?

For additional shared resources (storage and processing) at Terrascope during the project €5000 will be covered from the project by VITO.

Data security: how will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

OneDrive, Teams and Google Drive secure, enterprise cloud storage service with centralized security controls and multi-layered encryption, authorized by KULeuven, VITO or INBO.

The answers to this section were checked by:

- Others

6. Data preservation after the FWO project

Which data will be retained for the expected 5 year period after the end of the project? In case only a selection of the data can/will be preserved, clearly state the reasons for this (legal or contractual restrictions, physical preservation issues, ...).

The satellite data and derived products (WP3) and geo-indicator rasters created in WP6 (for further exploitation) up to 20TB will be stored on Terrascope (on the condition that Terrascope receives further BELSPO funding) and remain there after the project. The reference data collected in this project (WP3) will remain on INBO's Google Drive, and published in open repositories like GBIF, Zenodo and Environmental Data. The reference data obtained from the end-users will only be published (if not done already) with their explicit consent. Otherwise, the data will be securely stored on INBO's Google Drive only.

The materials for dissemination and communication will remain available for the broad public via the project website and through ResearchGate.

Where will the data be archived (= stored for the longer term)?

Same as described in the previous question. Only the project website will be discontinued 5 years after the end of the project, but all project outputs will remain available on ResearchGate. Satellite data and derived geo-indicator rasters will be stored for the longer term at Terrascope at VITO (on the condition that Terrascope receives further BELSPO funding)

What are the expected costs for data preservation during the retention period of 5 years? How will the costs be covered?

No costs. Terrascope already accommodates the satellite imagery, which require the largest storage capacity. INBO's Google Drive can easily accommodate the remainder of the data.

The answers to this section were checked by:

- Others

7. Data sharing and reuse

Are there any factors restricting or preventing the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)?

- Yes. Specify:

Source code of WP6 workflows cannot be shared.

Sharing of geo-indicator raster data will be subject of discussion with the particular end-user in order not to jeopardize further post-project exploitation together with the end-user. If needed separate logins can be created for the different end-users.

Which data will be made available after the end of the project?

The reference data collected by the project (Zenodo, GBIF, Environmental Data Initiative), the scripts and code blocks (GitHub)

Where/how will the data be made available for reuse?

- In an Open Access repository

The reference data collected by the project (Zenodo, GBIF, Environmental Data Initiative), the scripts and code blocks (GitHub)

When will the data be made available?

- Upon publication of the research results

Who will be able to access the data and under what conditions?

Everyone, with free and open access.

What are the expected costs for data sharing? How will the costs be covered?

None, as Zenodo, GBIF, Environmental Data Initiative and Github are free and open.

The answers to this section were checked by:

- Others

8. Responsibilities

Who will be responsible for data documentation & metadata?

For reference data (WP3): Stien Heremans (INBO)

For scripts and code (WP 4-5): Matthew Blaschko (KULeuven) and Ben Somers (KU Leuven)

For satellite imagery and derived products (WP3): Ils Reusen (VITO)

For workflows (WP6): Tanja Vanachteren (VITO)

For geo-indicator rasters (WP6): Ils Reusen (VITO)

Who will be responsible for data storage & back up during the project?

For reference data (WP3): Stien Heremans (INBO)

For scripts and code (WP 4-5): Matthew Blaschko (KULeuven) and Ben Somers (KU Leuven)

For satellite imagery and derived products (WP3): Ils Reusen (VITO)

For workflows (WP6): Tanja Vanachteren (VITO)

For geo-indicator rasters (WP6): Ils Reusen (VITO)

Who will be responsible for ensuring data preservation and reuse ?

For reference data (WP3): Stien Heremans (INBO)

For scripts and code (WP 4-5): Matthew Blaschko (KULeuven) and Ben Somers (KU Leuven)

For satellite imagery and derived products (WP3): Ils Reusen (VITO)

For workflows (WP6): Tanja Vanachteren (VITO)

For geo-indicator rasters (WP6): Ils Reusen (VITO)

Who bears the end responsibility for updating & implementing this DMP?

The PI bears the end responsibility of updating & implementing this DMP.