



D3.1

Database of geospatial data

Deliverable for the Horizon Europe Project BirdWatch

Version 1.2



**Funded by
the European Union**

Legal Disclaimer

This document reflects only the views of the author(s). Neither the European Global Navigation Satellite Systems Agency (EUSPA) nor the European Commission is in any way responsible for any use that may be made of the information it contains. The information in this document is provided “as is”, and no guarantee or warranty is given that the information is fit for any particular purpose. The below referenced consortium members shall have no liability for damages of any kind including without limitation direct, special, indirect, or consequential damages that may result from the use of these materials subject to any liability which is mandatory due to applicable law.

This document and the information contained within may not be copied, used, or disclosed, entirely or partially, outside of the BirdWatch consortium without prior permission of the project partners in written form.

© 2023 by BirdWatch Consortium.



**Funded by
the European Union** 2

Document Information

GA Number	101082634	Type of Action	Horizon-IA
Full Title	BirdWatch - a Copernicus-based service for the improvement of habitat suitability of farmland birds via satellite-enabled monitoring, evaluation and optimisation of CAP greening measures		
Project Acronym	BirdWatch		
Start Date	February 1st, 2023	Duration	36 Months
Project URL	https://birdwatch-europe.org/		
Deliverable	D3.1: Database of geospatial data		
Work Package	WP3		
Project Month of Delivery	Contractual	M6	Actual M6
Nature	DATA	Dissemination Level	PU
Lead Beneficiary	EURAC		
Responsible Author			
Contributions from	Nika Oman Kadunc – SINERGISE Nejc Vesel – SINERGISE Basil Tufail – EURAC Bartolomeo Ventura – EURAC Alexander Jacob – EURAC		



History of Changes

Version	Issue Date	Stage	Description	Comments	Contributor
1.0	31.07.2023		Initial version of D3.1		
1.1	03.08.2023		Version 1.1 of D3.1		Nika Oman Kadunc – SINERGISE Basil Tufail – EURAC
1.2	04.08.2023		Version 1.2		Nastasja Scholz - LUP

1	Introduction.....	5
2	Selection of datasets.....	6
2.1	Available relevant datasets.....	7
2.2	Bird Species Requirements.....	10
2.3	EO-parameters relevant for bird habitat assessments.....	11
2.4	Species Distribution Model data requirements.....	13
3	Vector Datasets.....	14
3.1	Land Parcel Identification System (LPIS) datasets.....	14
3.1.1	Exploration of Birdwatch test regions' LPIS datasets.....	15
3.1.2	Harmonization of BirdWatch test regions' LPIS datasets.....	18
3.2	The EuroCrops dataset.....	20
3.3	BirdWatch vector database description.....	21
4	Raster Datasets.....	23
4.1	10m Annual Land Use Land Cover (LULC).....	23
4.2	Mowing intensity and first mowing date datasets.....	24
4.2.1	Calculation of the mowing intensity.....	24
4.3	Access to raster data.....	27
4.3.1	BYOC SH.....	28
4.3.2	openEO framework.....	28
4.4	Spatio-temporal Asset Catalog API.....	30
5	Conclusion.....	32



1 Introduction

This document describes in detail the geospatial datasets that were prepared and structured into geospatial databases for the BirdWatch project until M6 of the project. At this first stage of the BirdWatch project, the initial dataset selection was made based on the preliminary information on the remote sensing parameters needed for the bird habitat assessments. Taking into account Bird Species Requirements (listed and described in Section 2.2). The datasets were prepared for the first iteration of the Species Distribution Model (SDM) and will be expanded and iterated upon as the SDM testing results are gathered and analysed. The basic technologies used for the vector and raster database structure and data access were set according to SDM input data requirements (described in Section 2.4) and in accordance with the consortium partners managing different phases (WPs) of the BirdWatch project that need the data as input for the activities of their respective WPs.

In Section 2, the dataset selection process for the BirdWatch project is described in detail, considering all available information and requirements for the geospatial databases to be used throughout the BirdWatch project. Section 2.1 describes the readily available datasets that were made accessible to support activities of WP2000. Section 2.2 lists the Bird Species Requirements that need to be considered while making the preliminary selection of the parameters and, consequently, bird habitat assessments – these parameters are described in Section 2.3. Section 2.4 lists the requirements for the data needed to serve as an appropriate input for the Species Distribution Model (SDM).

Section 3 describes in detail the vector datasets that are (or will be further) used as datasources for the geospatial vector database of the BirdWatch project. In Section 3.1, we describe the LPIS datasets and the necessary data harmonisation steps of the four BirdWatch project testing regions. Further, in Section 3.2, the EuroCrops dataset is presented. It will be used in areas where LPIS data may not be available or as a complementary dataset. Section 3.3 describes the technical structure and data access for the BirdWatch vector database.

In Section 4, geospatial raster databases required and prepared for the initial testing of the SDM are presented. First, a readily available dataset for land cover is described in Section 4.1. As mowing cycles are already known to be of importance for the bird habitat suitability assessment, initial maps for mowing intensity were created specifically for BirdWatch and are presented in Section 4.2. Section 4.4 describes the BirdWatch raster database access along with its frameworks and standards.



2 Selection of datasets

The process of dataset selection started early in the project (M1) in order to support all activities of the BirdWatch partners in WP2000. The process of selection incorporated gathering all (preliminary) information available regarding potentially relevant parameters for bird habitat assessments for different bird species (Bird Species Requirements) as well as the requirements for the Species Distribution Model (SDM) – these requirements are further described in detail in Sections 2.2 and 2.4. A list of available datasets was prepared in order to enable early exploration and testing. The list is presented in Section 2.1. Section 2.3 presents the initial parameters selected for SDM testing. This list will be iterated upon, adding and / or eliminating parameters, as the results of SDM testing will be gathered.

To prepare the data into a format suitable for the SDM, the information from the relevant datasets described in this document will be rasterized and combined into a raster stack that will be aggregated to EuroStat Grid 200mx200m cells. The raster stack will be hosted on Sentinel Hub BYOD (see section 4.3) from which it will be made available to the species distribution modelling component. The workflow is depicted in Figure 1.

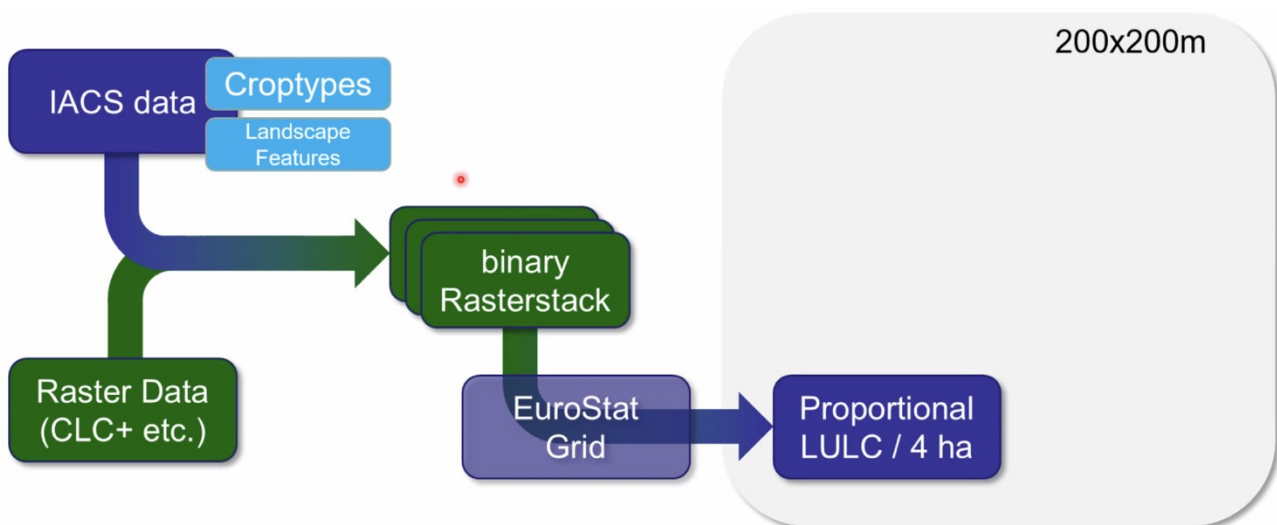


Figure 1: Workflow of data used in the first phases of the BirdWatch project.

2.1 Available relevant datasets

Based on the first preliminary information gathered about bird habitat assessment parameters and taking into account all possible relevant EO-data related to them, a list of available datasets was prepared (in M1 of the BirdWatch project) by Sinergise and Eurac research and shared among consortium partners. The goal of this activity was to support and facilitate activities of other work packages, mostly WP2000, namely to gather all information on what EO-data is readily available for usage. The selection for the datasets considered all relevant parameters listed in the preliminary voting table prepared by partners working on WP 2000 and availability of the datasets from the perspective of regional coverage for all the test regions as well as public availability and license of the datasets. Since the geospatial datasets provided within the BirdWatch project have public dissemination level, open and public datasets were preferred. Multiple datasets were listed where available.

Vector datasets, namely LPIS data was gathered for each test region and are presented in Table 1. The extensive list of raster datasets available through different platforms together with availability information and links to data access was created and is presented in Table 2. This is the first list of selected datasets further based on relevancy and requirements. The list would be iterated to other datasets as well.

Table 1: List of LPIS vector datasets potentially relevant for bird habitat assessment for different test regions of the BirdWatch project.

BirdWatch test Region	LPIS dataset	Link to download
Flanders	Landscape elements	https://landbouwcijfers.vlaanderen.be/open-geodata-kaart-landschapselementen
	Agricultural parcels	https://landbouwcijfers.vlaanderen.be/open-geodata-landbouwgebruikspercelen
Germany - Brandenburg	Agricultural parcels	https://geobroker.geobasis-bb.de/gbss.php?MODE=GetProductInformation&PRODUCTID=9e95f21f-4ecf-4682-9a44-e5f7609f6fa0
Lithuania	Land cover type layer, EFA Layer	https://www.geoportal.lt/metadata-catalog/catalog/search/resource/details.page?uuid=%7B05B2E0DB-31DC-7382-EBB7-7EC5531858A8%7D
South Tyrol	LISS (Land use layer) LAFIS (Agriculture parcels)	Not open-source data, thus shared internally through database with partners.



Table 2: List of available raster datasets potentially relevant for bird habitat assessment.

EO data type	Dataset	Link to description and download
Crop types	EuroCrops for the available countries	https://github.com/maja601/EuroCrops
	Global croplands data	https://www.usgs.gov/apps/croplands/app/map?lat=0&lng=0&zoom=2
Land cover	Corine Land Cover	https://land.copernicus.eu/pan-european/corine-land-cover/clc2018?tab=download and on SH: https://collections.sentinel-hub.com/corine-land-cover/
	CORINE Land Cover Accounting Layers	https://collections.sentinel-hub.com/corine-land-cover-accounting-layers/
	ESA WorldCover	https://collections.sentinel-hub.com/worldcover/
	10m Annual Land Use Land Cover LULC	https://collections.sentinel-hub.com/impact-observatory-lulc-map/
	Global Land Cover 100m	https://collections.sentinel-hub.com/global-land-cover/
	Dynamic World	https://dynamicworld.app/
Vegetation	Seasonal Trajectories, 10-daily	https://collections.sentinel-hub.com/seasonal-trajectories/
	Vegetation Indices, Daily	https://collections.sentinel-hub.com/vegetation-indices/
	Vegetation Phenology and Productivity Parameters Season 1, Yearly	https://collections.sentinel-hub.com/vegetation-phenology-and-productivity-parameters-season-1/
	Vegetation Phenology and Productivity Parameters Season 2, Yearly	https://collections.sentinel-hub.com/vegetation-phenology-and-productivity-parameters-season-2/
	Global Forest	https://earthenginepartners.appspot.com/science-2013-global-forest
	High Resolution Layers: Forests	https://land.copernicus.eu/pan-european/high-resolution-layers/forests
	High Resolution Layers: Grassland	https://land.copernicus.eu/pan-european/high-resolution-layers/grassland
	High Resolution Layers: Small Woody Features	https://land.copernicus.eu/pan-european/high-resolution-layers/small-woody-features
Elevation	DEM GTOPO30	https://www.eea.europa.eu/data-and-maps/data/digital-elevation-model-of-europe
	Copernicus DEM	https://collections.sentinel-hub.com/copernicus-dem/
	Mapzen DEM composite of elevation data	https://collections.sentinel-hub.com/mapzen-dem/
Infra-structure	Road network - Europe	https://public.opendatasoft.com/explore/dataset/europe-road/export/?refine.icc=BE
	Global Human Settlement Layer - GHS-	https://collections.sentinel-hub.com/global-human-settlement-layer-ghs-built-s2/

	BUILT-S2	
Hydrology	Global Surface Water	https://collections.sentinel-hub.com/global-surface-water/
	Water Bodies	https://collections.sentinel-hub.com/water-bodies/
	High Resolution Layers: Water & Wetness	https://land.copernicus.eu/pan-european/high-resolution-layers/water-wetness
	EU-Hydro	https://land.copernicus.eu/imagery-in-situ/eu-hydro
Structural elements - hedges, bushes, trees	Linear vegetation features around agricultural parcels in Flanders https://landbouwcijfers.vlaanderen.be/open-geodata-kaart-landschapselementen	
Other potentially useful datasets	Water framework directive including surface water bodies and groundwater bodies https://www.eea.europa.eu/data-and-maps/data/wise-wfd-spatial-3	
SAR data : Sentinel-1 Level 2, backscatter and coherence	Sentinel-1 GRD	https://collections.sentinel-hub.com/sentinel-1-grd/
	CARD4L	https://docs.sentinel-hub.com/api/latest/data/sentinel-1-grd/#card4l-data
Landsat / Sentinel 2 data	Level 2 of Landsat-8/9 and Sentinel-2 with their cloud and cloud shadow masks; different spectral bands and texture measurements	https://docs.sentinel-hub.com/api/latest/data/
EO data products derived from S1 and S2 data	Analysis ready data: Sentinel 1 + Sentinel 2	https://custom-scripts.sentinel-hub.com/custom-scripts/sentinel/sentinel-1/ https://custom-scripts.sentinel-hub.com/custom-scripts/sentinel/sentinel-2/
	Spectral indices	https://custom-scripts.sentinel-hub.com/custom-scripts/sentinel/sentinel-2/#remote-sensing-indices
	Texture measures Sentinel 1+ Sentinel 2: Haralick texture	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6386443/ eo-learn code to calculate Haralick features: https://eo-learn.readthedocs.io/en/latest/_modules/eolearn/features/haralick.html
	Phenology - Sentinel 2	https://collections.sentinel-hub.com/seasonal-trajectories/
	Soil moisture - Sentinel 1	Calculated from Sentinel-1, Further details in provided D3.3.
	Coherence - Sentinel 1	Calculated from Sentinel-1, Further details in provided D3.3.
	Grassland mowing	https://area-monitoring.sinergise.com/docs/markers/mowing-marker/

Together with the list of available datasets, access to Sentinel Hub service platform was granted to partners in the scope of the BirdWatch project to explore usability of the raster datasets. Five "Basic" Sentinel Hub user accounts were created for the purposes of BirdWatch project data exploration and preliminary testing.

2.2 Bird Species Requirements

The selection of datasets to prepare into the geospatial databases was based on the bird species requirements gathered and identified in the activities of WP 2000. Table 3 lists the types of habitat requirements collected for each bird species, along with the EO data and EO data-based measures to derive the presence or absence of a habitat requirement.

Table 3: Types of habitat requirements collected for each bird species, along with the EO data and EO data-based measures to derive the presence or absence of a habitat requirement.

Requirement type	Requirement expressions	EO and EO-based data	Examples for EO-based measures
Structural habitat characteristics	Open (e.g., wide, open fields) Semi-open (e.g., open fields with some vertical structures or landscape elements) Dense (dense growth of e.g., shrubs, trees or bushes) Mosaicked (mix of open and dense structures)	Sentinel 1 and -2 Readily available land cover classifications (e.g., Dynamic World)	Percentage land cover type per habitat Proportional extent of land cover type per habitat Texture measures (e.g., grey-level occurrence) Radar backscatter
Landscape elements	Trees, high shrubberies Low shrubberies, bushes, hedges Water bodies and water edge vegetation Man-made structures	Sentinel 1 and -2 Readily available land cover classifications Digital terrain models & digital elevation models	Percentage cover per land cover type Spectral & radar vegetation indices Radar backscatter Difference calculations
Vegetation types	Crop types Grassland types	Sentinel 1 and -2 Readily available land	Percentage land cover type per habitat

		cover and crop type classifications (e.g., EuroCrops)	Spectral & radar vegetation indices Texture measures (e.g., grey-level occurrence) Time series metrics
--	--	---	--

Based on the Bird habitat requirements per species (described in the BirdWatch Deliverable D2.2), summary of the parameters was made and the parameters with most occurrences were chosen for the database preparation for the first iteration of the Species distribution model (SDM) testing. The summary of the habitat requirements is presented in Table 4.

Table 4: Number of times specific habitat requirements are mentioned in the list of bird species habitat requirements; 2nd column is only filled for cases in which different data sources might be required.

Requirement type	Requirement expressions which might need different data input	Number of occurrences
Structural habitat characteristics		10
Landscape Elements		
	Trees	10
	Low shrubberies	10
	Water bodies	7
	Man-made	5
Vegetation types		
	Crop type	6
	Orchards	3
	Grassland type	7
Species diversity		9
Soil Moisture		10
Terrain		3
Mowing (start & end / intensity of mowing)		10

2.3 EO-parameters relevant for bird habitat assessments

Considering the Bird species requirements and the corresponding parameter occurrence summary statistics presented above, a list of parameters that would comprise the geospatial database was created and agreed upon with partners on the live meeting in Potsdam in June 2023. The list is presented in Table 5. The parameters were grouped according to availability in different dataset types (Landscape features, Crop type, Landcover and Remote Sensing (RS) derived data). Some



similar crop types were grouped together into common subgroups according to relevance for the bird habitat requirements to reduce the number of parameters which contributes to efficiency of the SDM as well as facilitates the ingestion of parameters into the database.

The RS-derived datasets were calculated and created specifically for the BirdWatch project as the necessary data for the bird species habitat requirements was not yet available as a ready-to-use dataset.

For each of the parameters, a dataset containing the information was either selected from the list of available datasets described in Chapter 2.1 of this document or calculated from EO-data. The chosen datasets listed in Table 5 are further described in detail in the next Sections of this document.

Table 5: List of parameters identified based on the bird species requirements to be ingested into the vector and raster geospatial databases.

Parameter group	Parameters	Dataset(s)	Dataset type
Landscape Features	Hedgerows / Shrubs	LPIS	Vector
	Trees		
	Field margins		
Crop type	Winter Cereals	LPIS EuroCrops	Vector
	Summer/Spring Cereals		
	Other Cereals		
	Maize		
	Root Crops (sugar beet, potatoes etc.)		
	Vegetables (salad, cabbage etc.)		
	Rapeseed		
	Legumes (soy, peas, beans etc.)		
	Permanent Grassland		
	Cultivated Grassland		
	Fallow		
	Orchards and Berries		
	Grapevine		
	Hops		
	Sunflower		
	Forest Trees/SRF		
	Miscellaneous		
Landcover	Forest	10m Annual Land Use Land Cover (LULC)	Raster
	Water		
	Urban Areas / Streets		
RS-derived data	Soil Moisture	Sentinel -1	Raster

	Mowing intensity	Sentinel-2 L2A	
	First mowing date	Sentinel-2 L2A	

2.4 Species Distribution Model data requirements

As the data prepared for the geospatial databases will serve as input for the species distribution modelling (SDM), we took into consideration the SDM data requirements as provided by Bird-Watch project partners of WP4000. These requirements are listed below.

1. General data requirements:

- Continuous raster data are needed for SDMs. For example, early or late mowing as a category is not useful but the day or month in a year when mowing took place could be useful.
- How many parameters can be included in the model depends on the number of occurrence points per species (as a rule of thumb: 10 occurrence points are needed for 1 parameter). We will also have to test for correlation between the parameters. When parameters are correlated the habitat requirement list can help to identify the one parameter out of the correlated ones that will be included in the model. The other correlated parameters will not be included. Statistical outputs could also be used to identify which parameter out of the correlated ones will be included.

2. Temporal proximity of environmental data (e.g. EO-based data):

- In general, the most recent bird occurrence data are the most relevant
- If one year has not enough bird occurrence data, data from previous years can be included as well
- The environmental data should be derived from the same time period as the bird occurrence data

3. Parameter selection for the final habitat models:

- The first selection of parameters at the moment is only preliminary. The data will go into the model and the model will then inform us on how well each parameter explains the species occurrence data. Thus, which parameters are "useful" and which ones are not.
- For a lot of parameters we do not know yet if they will explain the occurrence data well or not. Here, the model output will help. For example, we could use proportion of high and low grass per grid cell under the category "land cover type" but we could also have a category "proportion of grass per cell and height" as another category. For these cases we will have to run the models with each category setup to see which categories explain the occurrence data better.



3 Vector Datasets

Vector based datasets are a huge input requirement for the Birdwatch project. Datasets for Land parcel information which describes the usage of parcels to crop type are available in vector formats. Detailed information like shrubs hedgerows which have been identified as important features for bird habitats are also present within the LPIS database or sometimes as separate vector format data. These can be in either shapefile to GeoJSON format which are the most used vector data formats. In the following section a brief overview of most important vector datasets for the Birdwatch project is given.

3.1 Land Parcel Identification System (LPIS) datasets

The most accurate and up-to-date and thus very important source of vector-based information relevant for bird habitat assessments are the LPIS datasets provided by EU countries or specific regions. LPIS is a digital database that contains detailed information about land parcels within a specific geographic area. It is typically used by government agencies, agricultural authorities, and farmers to manage agricultural land and monitor compliance with land use regulations and is maintained and updated by government agencies responsible for agriculture or land management.

LPIS data provides a comprehensive inventory of agricultural land, including both arable and non-arable parcels, landscape features and more. In general, LPIS data includes various attributes for each land parcel, such as:

1. Parcel identification number: A unique identifier assigned to each land parcel.
2. Geospatial information: Coordinates and boundaries of the parcel, usually in the form of geographic coordinates (latitude and longitude) or a polygonal shape.
3. Land use category: Classification of the land parcel based on its primary use, such as cropland, grassland, orchard, or fallow land.
4. Area: The size or area of the land parcel, typically measured in hectares or acres.
5. Ownership information: Details about the landowner or landholder.
6. Eligibility and subsidy status: Information related to eligibility criteria for agricultural subsidies or grants, including compliance with specific land use regulations or environmental requirements.
7. Historical data: Records of land parcel changes over time, including past land use, land ownership, and any modifications made to the parcel.

Together with LPIS datasets with geospatial information, usually a text document containing metadata of the dataset and explanation of vector file attributes, classes and their formats is obtained.



Specific attributes and details of LPIS data vary depending on the country or region, as different jurisdictions have their own systems and data structures. Thus, to construct a common and standardised vector database for the purposes of BirdWatch project, harmonisation of LPIS datasets for different test regions is needed.

We explored the contents of BirdWatch test region's LPIS datasets and investigated the harmonisation issues for the purpose of ingestion of EO parameters from different LPIS datasets into a common BirdWatch vector database.

3.1.1 Exploration of Birdwatch test regions' LPIS datasets

The BirdWatch project test regions LPIS datasets were provided by the BirdWatch consortium partners (listed in Table 1 of this document), each from their own country, and data contained within them was explored. The goal of this activity was to identify which parameters, relevant for bird habitat assessments, were available within these datasets and to investigate harmonisation or other potential (i.e. missing data, ...) issues.

LPIS data availability issues

In providing the BirdWatch test regions LPIS data, most of them are publicly available and a direct download link is provided by the authorities. There are, however, availability issues for Germany as there is no single aggregated LPIS dataset for the whole country, but rather respective federal states provide their own LPIS data (16 in total). Additionally, most of the datasets are not openly available for download; some are available per request (directly from the authorities) and this process is sometimes slow and, in some cases, the request may not be granted. Also, in some states, LPIS data is not public and there are issues with publishing results sourcing from these datasets.

For the purposes of preliminary testing of SDM for BirdWatch, LPIS dataset for the Germany's Brandenburg state was used, as it is freely available. For other German states, LPIS will be gathered where possible along the course of the BirdWatch project; in the regions where this will not be possible, the EuroCrops dataset (described below in Section 3.2 of this document) will be used to obtain the relevant parameters in the region.

LPIS data exploration

Investigation and comparison of the structure, attributes and classes of each dataset was conducted. We found that all categories of the LPIS datasets (structure, attributes and classes) vary from region to region. An example of the LPIS complementary documentation from Lithuania and Germany, showcasing the differences in structure and attributes of the two datasets, is depicted in Figure 2.



Brandenburg, Germany

Landscape Features (LE):			
Field	Data Type	Length	Description
LE_ID	Text	16	Landscape Feature ID
AREA_QM	Numeric	18 - 4 decimal places	area in square meters, corresponds to the subsidizable maximum area for direct payments
PERI_M	Numeric	18 - 4 decimal places	perimeter
TK10_BLATT	Text	14	page of the topographical map
KREIS_NR	Numeric	9	Number from the office of agriculture
L_TYP	Text	2	Type of LE (Standard deduction area: "LT"; freely digitized area: "LE")
L_BEZEICH	Text	2	Type of LE (see list below)
CODE	Numerisch	9	Type of LE as number
GROSSE_P			area
X_COORD	Numeric	18 - 6 decimal places	X-coords centroid
Y_COORD	Numeric	18 - 6 decimal places	Y-coords centroid
GUELTVON_F	Date		validity date start
GUELTBIS_F	Date		validity date end
FGUE_JAHR	Text	4	validity year

Type of Landscape Feature		
Description	Number of instances	Area sum in ha
BR = Tree rows	10506	1414,1499
EB = Individual trees	86	0,3575
FG = Wetland areas	7021	611,2194
FR = Field hedgerow	18058	1153,6167
FR = Field margins	783	138,9844
FS = Rock and stone ridges as well as naturally stoned areas	158	6,1642
HK = Hedges or field boundaries	14224	2772,6056
NT = Natural stone and dry stone walls, stone wall made of loose stones	419	13,0693
Sum		

Lithuania

Reference parcel layer		
FIELD NAME LT	FIELD NAME EN	Description
DATA_NUO	DATE	UPLOAD DATE
BLOKAS_ID	RP_NUMBER	UNIQUE REFERENCE PARCEL NUMBER
PLOTAS_HA	AREA_HA	RP area in ha
GKODAS	GOE_CODE	CODED VALUE FOR ELIGIBILITY

GKODAS values	Description
bl1	Eligible reference parcel – arable land, permanent grassland, permanent crop
bl1b	Eligible reference parcel not declared in previous years
bl2	Non-eligible reference parcel. Built up area in cities
bl3	Non-eligible reference parcel. Forests, groups of trees, shrubs
bl6	Non-eligible reference parcel. Built up area outside cities
bl9	Non-eligible reference parcel. Wetlands, quarries, temporarily soaked areas, etc.
Other objects	
gc11p	Highway
gc12o	Road with pavement

Figure 2: Example of the LPIS data complementary documentation from Brandenburg, Germany and Lithuania, showcasing the differences in structure, classes and attributes of the two datasets.

Despite the observed differences, all LPIS data within this exploration contain polygons, parameters, additional parameters for each attribute, which enables extraction of the vector polygons of selected relevant parameters. An example of visualised vector polygons for Lithuanian LPIS data is shown in Figure 3.

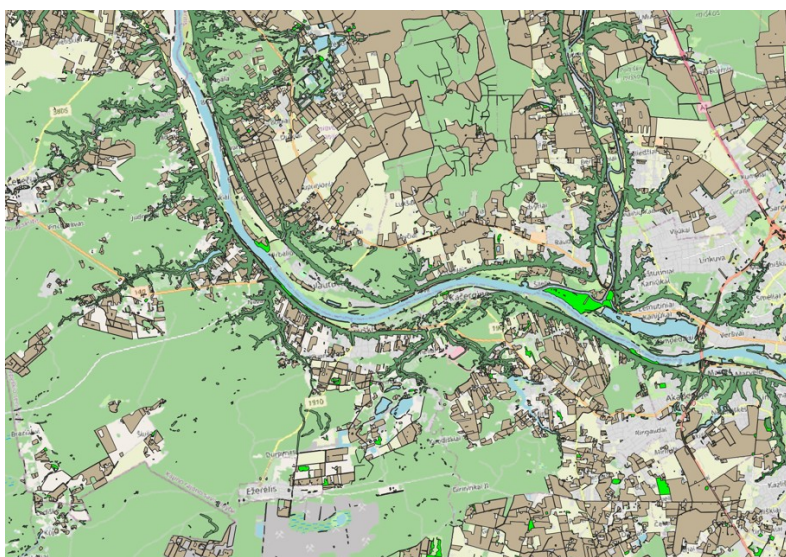


Figure 3: An example of visualized vector polygons for Lithuanian LPIS data.

The dataset available for the South Tyrol area is called LAFIS (sistema informativo agricolo forestale provincia autonoma di Bolzano). It is periodically generated by the agriculture and forest services of the Province of Bolzano. In the absence of any LPIS dataset present for the region it serves as a good substitute.

The recent version of the data was generated in 2020 and before that there is a version from 2015.

The latest version includes 100 classes including various fruits, vegetables and types of pastures and meadows.

Based on the requirement of the birds features can be extracted from the LAFIS layer. For example, Figure 4 shows a subsection of the province of South Tyrol with hedges which are taken from the LAFIS layer.

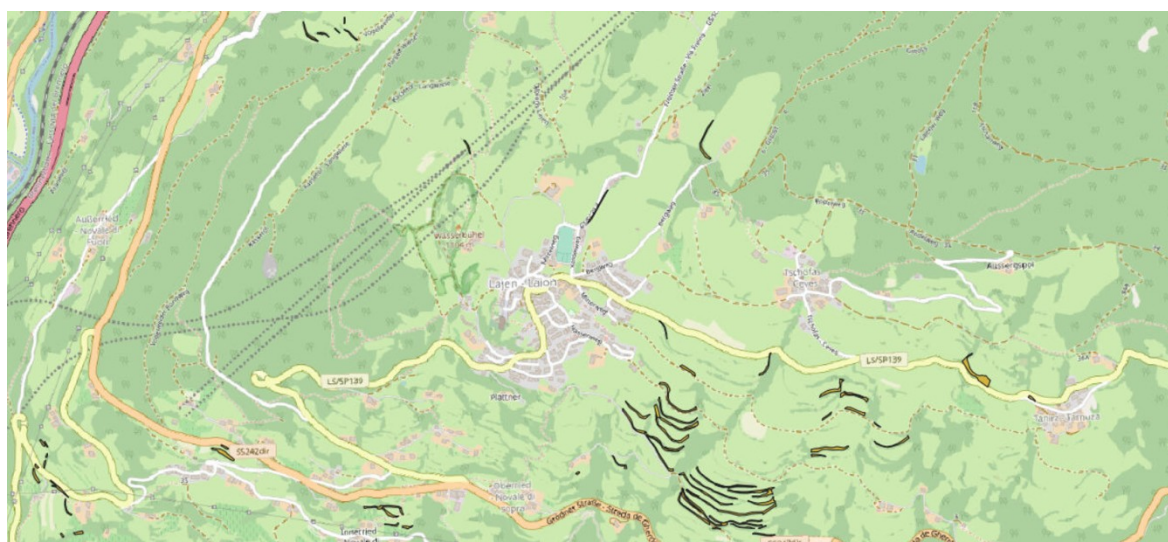


Figure 4: An example of visualized vector file for hedgerows for South Tyrol.

The second data available for the province of South Tyrol is the LISS dataset which is a land cover layer produced by Eurac internally. It is a better option compared to generic land use layers like Corine. The data comprises 16 classes which is shown in Table 6.

Table 6: 16 classes of land cover layer for South Tyrol LISS dataset.

LISS code	LISS legend
120	Artificial
210	Farmland
221	Vineyards
222	Fruit plantations
231	Intensively used meadows and pastures
250	Mixed agriculture
315	Forest
321	Extensively used pastures and natural meadows
322	Krummholz
323	Dwarf shrubs
324	Grasslands with trees
332	Bare rocks
333	Sparse vegetation
430	Glaciers
510	Wetlands
520	Water bodies

3.1.2 Harmonization of BirdWatch test regions' LPIS datasets

Harmonization of LPIS data refers to the process of integrating and standardizing LPIS datasets from different sources or regions to ensure consistency, compatibility, and interoperability.

The harmonization of LPIS data involves several steps:

1. **Data Integration:** The collected LPIS data is brought together into a unified dataset. This involves merging and aligning data from different sources, removing duplicates, and resolving any conflicts or inconsistencies.
2. **Standardization:** The next step is to standardize the LPIS data to ensure uniformity and compatibility. This includes establishing common data formats, coordinate reference systems, attribute schemas, and coding schemes.
3. **Data Transformation:** LPIS data may need to be transformed a consistent coordinate system to enable seamless integration. This step ensures that the data aligns accurately with other geospatial datasets.
4. **Attribute Mapping:** During harmonization, attribute mapping is performed to match and reconcile similar attributes across different datasets. This creates a consistent and comprehensive set of attributes.
5. **Quality Assurance:** It is essential to perform quality checks and validations to identify and correct any errors or inconsistencies in the harmonized LPIS data.



6. Metadata Documentation: Documenting the metadata is crucial to provide information about the source, processing steps, data accuracy, and other relevant details. This documents the characteristics and limitations of the harmonized LPIS dataset.

With the data exploration, the first issue was encountered with the language: attributes and classes and the metadata documentation of the respective LPIS datasets are each provided in the native language of the country of origin. To enable investigations and comparisons, translations of these documents to English language were provided by BirdWatch consortium partners.

To investigate the harmonization issues with the give LPIS datasets, we chose the example of the “Trees” parameter and proposed a harmonized extraction of this parameter. Figure 5 shows all different attributes corresponding to the “Trees” parameter in three of the LPIS datasets depicting the problem of harmonizing between different LPIS datasets.

We proposed the following harmonized categories as possible values of the “Trees” parameter to ingest into the vector database:

- Forest
- Individual trees
- Other

Germany:

- BG = groups of trees
- BR = tree rows
- EB = individual trees, if not a point feature
- Wa = Forest
- So = Solitaire

Lithuania:

- BI3 (Non-eligible reference parcel. Forests, groups of trees, shrubs)
- mg0 (Tree group / forest)

Flanders:

- Trees in group
- Rows of trees
- Arboriculture - forest and hedge plants
- Arboriculture - fruit cultivation
- Arable land with more than 100 trees per ha
- Arboriculture - forest and hedge plants
- Forest
- Christmas trees
- Avenue trees
- Pasture with non-harvestable trees (> 100 trees per ha)
- Pasture with harvestable standard trees (> 100 trees/ha)
- Solitary trees

Figure 5: All different attributes corresponding to the “Trees” parameter in three of the LPIS datasets depicting the problem of harmonizing between different LPIS datasets.

The quality checks and validation will be done with manual checks on a sample of the data to validate and verify the harmonization procedure. Also, visual analysis of data will be made to check the overall consistency. In case of detected errors, we will also compare data from different sources to check for inconsistencies, missing data etc.



3.2 The EuroCrops dataset

EuroCrops is a dataset collection that contains geo-referenced polygons of agricultural croplands from 16 countries of the European Union as well as information on the respective crop species grown there. Figure 6 shows participating countries. These semantic annotations are derived from self-declarations by farmers receiving subsidies under the common agricultural policy of the European Commission. Over the last years, the individual national crop datasets have been manually collected, the crop classes have been translated into the English language and transferred into the newly developed hierarchical crop and agriculture taxonomy (HCAT) containing over 300 crop classes. EuroCrops is publicly available under continuous improvement through an active user community 1. An example of EuroCrops polygons is shown in Figure 7. More Information and documentation on the EuroCrops data collection is available at <https://github.com/maja601/EuroCrops>.

From the HCAT classes, we created 17 crop type categories that correspond to the Bird species requirements (see listed crop types in Table 5).

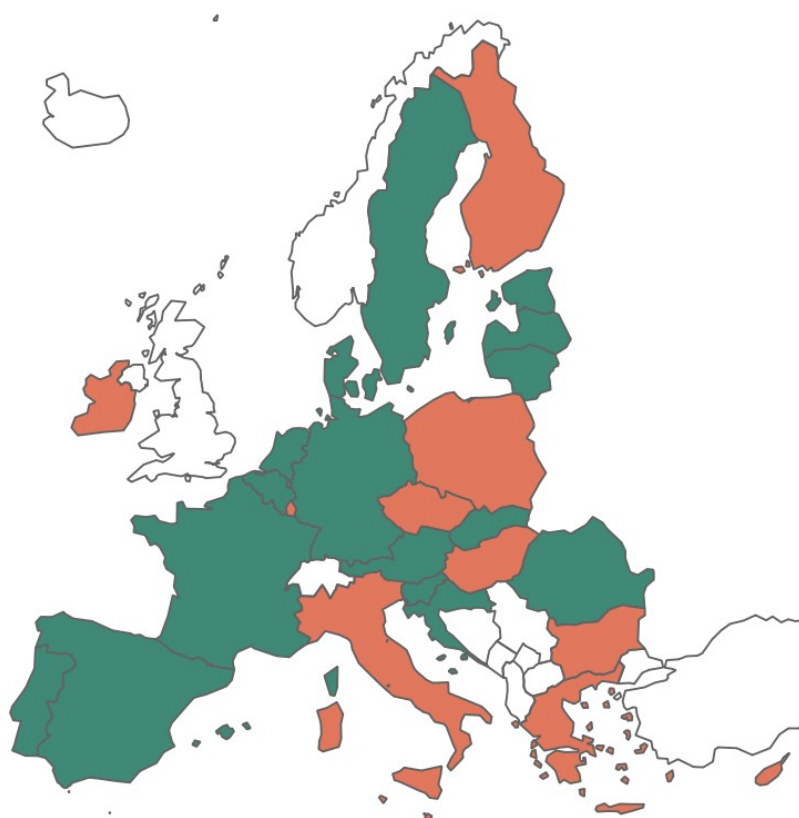


Figure 6: Countries participating in EuroCrops project are marked with green (source: <https://github.com/maja601/EuroCrops>)





Figure 7: An example of EuroCrops polygons (source: <https://www.eurocrops.tum.de/>)

3.3 BirdWatch vector database description

In the framework of the BIRDWATCH project, it is necessary to collect and organize the data provided and/or collected by the partners, creating a common Database (DB). The DB is a POST-GRES open-source object-relational database system. It is an open-source, highly stable database system that provides support to different functions of SQL, like foreign keys, subqueries, triggers, and different user-defined types and functions.

The direct connection to the DB is possible using different clients (R, Python, PgAdmin, etc.) and/or by a basic, but powerful, web interface for data preview and queries. The connection will be possible for all project partners using specific credentials and connection parameters to be shared. An example of the credentials that will be shared.

Host	birdwatch_host
name	birdwatch (schema: hydrology)
Port	XXXX
Username	birdwatch_user
Password	write to <bartolomeo.ventura@eurac.edu>

The datasets to be managed will consist of various information about different parameters such as trees, high shrubberies, waterbodies, etc., and they will belong to different areas (Flanders, South



Tyrol, Germany, Lithuania). Because of these inhomogeneities, it will be necessary to perform a preliminary analysis that will then lead to harmonise all available data, as described in Section 3.1.2 of this document.

In the initial phase, up to five fundamental parameters will be selected by the data providers, with the possibility to add a few optional ones for the sake of completeness. Since they know their data, it is important to collect the main information that will be used in future queries. The goal is to have an overview, as complete as possible, of the entire dataset to be inserted in the DB. Consequently, it will be accordingly created. This will speed up its usage when the users will start using the DB.

Currently, the DB has been created and additional configurations are being performed to enable full accessibility.

With all the permissions to access the DB, a generic request can be formulated as follows: host-name/TABLENAME/OPTIONALFILTERS.

In the meanwhile, a GUI has been created where only a selected user can access and see the data shared among the partners: <https://maps.eurac.edu/maps/1309>.

The credentials for this user have been already shared. In Fig 8, it is possible to have an overview of one of the available layers. It is worthwhile noting that, if you click on one polygon, all the information contained is shown at the right part of the layer.

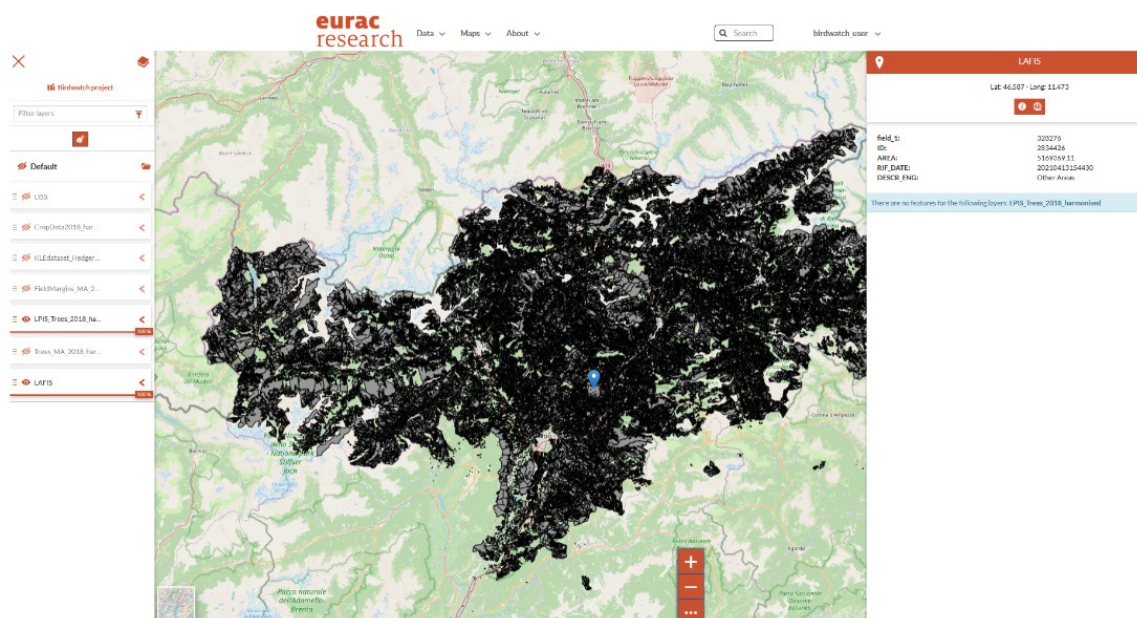


Fig 8: Screenshot of vector data GUI set up to visualize and access the data.

4 Raster Datasets

4.1 10m Annual Land Use Land Cover (LULC)

The 10m Annual Land Use Land Cover (LULC) is a land cover dataset produced by Impact Observatory, Microsoft, and Esri collaboratively. The data collection is derived from ESA Sentinel-2 imagery at 10m resolution using Impact Observatory's state of the art deep learning AI land classification model which is trained by billions of human-labelled image pixels. It is produced on annual basis worldwide, so it was chosen as a suitable dataset for the BirdWatch. There are 9 LULC classes generated by the algorithm: Built Area, Crops, Trees, Water, Rangeland, Flooded Vegetation, Snow/Ice, Bare Ground, and Clouds. An example of an area in Germany mapped by the classes of the LULC is presented in Figure 9.

The parameters identified to be relevant for the bird habitat assessment that will be part of the species distribution modelling are identified by these LULC dataset classes:

- **Forest – „Trees“ dataset class:** Any significant clustering of tall (~5 m or higher) dense vegetation, typically with a closed or dense canopy; examples: wooded vegetation, clusters of dense tall vegetation within savannas, plantations, swamp or mangroves (dense/tall vegetation with ephemeral water or canopy too thick to detect water underneath).
- **Water – „Water“ dataset class:** Areas where water was predominantly present throughout the year; may not cover areas with sporadic or ephemeral water; contains little to no sparse vegetation, no rock outcrop nor built up features like docks; examples: rivers, ponds, lakes, oceans, flooded salt plains.
- **Urban Areas / Streets – „Built area“ dataset class:** Human made structures; major road and rail networks; large homogenous impervious surfaces including parking structures, office buildings and residential housing; examples: houses, dense villages / towns / cities, paved roads, asphalt.



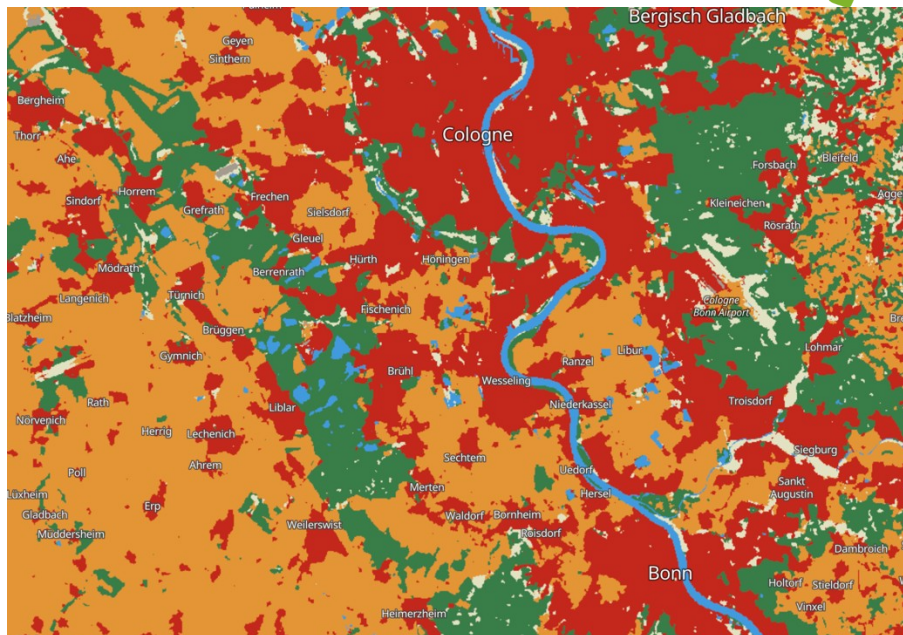


Figure 9: An example of an area in Germany mapped by the classes of the 10m Annual Land Use Land Cover (LULC).

More detailed description of the data collection is available at:

<https://collections.sentinel-hub.com/impact-observatory-lulc-map/>

4.2 Mowing intensity and first mowing date datasets

The mowing intensity and the first mowing date parameters are mapped by calculating the Siner-gise's pixel-based mowing marker.

The mowing marker identifies all observations where mowing events lead to a drop in the vegetation index values (NDVI). The mowing event itself cannot be detected via satellite imagery, but its consequences are observable. The marker identifies mowing events by identifying drops in the NDVI profile of features-of-interest (FOIs) claiming to be permanent meadows or growing grass, clover or grass/clover mixtures.

4.2.1 Calculation of the mowing intensity

The algorithm used by the mowing marker is based on an algorithm that searches for variations in NDVI to identify candidate mowing events. The algorithm doesn't depend on the neighboring data and solely uses the NDVI time-series of the target FOI.

The detection is performed in the following steps:

1. Perform cloudy observation filtering using the s2cloudless cloud detector.



2. Perform other outlier-like filtering using observation outlier detector.
3. Search for pairs of local maxima and minima from the remaining valid observations.
4. Keep only pairs where:
NDVI drops below threshold value and,
NDVI recovers above threshold value.
5. Extract information like duration, number of observations, etc. and construct the marker.

An example of a detected mowing event in the time interval during the period from April 15th to Oct 15th is shown in Figure 10.

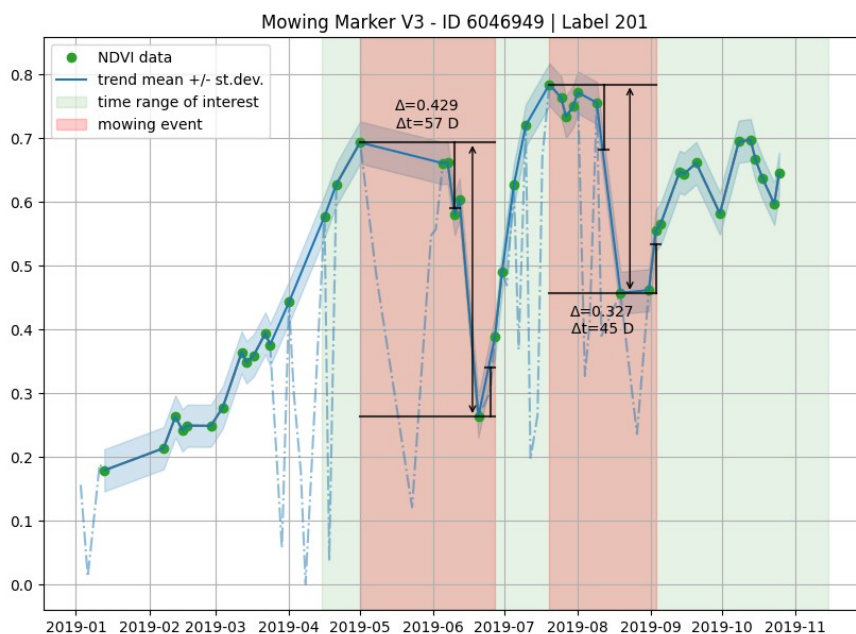


Figure 10: An example of a detected mowing event in the time interval during the period from April 15th to Oct 15th. The dashed line represents the whole downloaded NDVI timeseries before filtering steps.

The thresholds for NDVI drops and recovery are different, which is achieved by applying different scaling factors to the estimated value of fluctuation in the time-series (blue band on the image). The rate of fluctuation in the NDVI time-series is estimated via the differences between the consecutive values in the time-series and averaged over a large sample of FOIs beforehand. A fluctuation estimate in the form of a single number is obtained and used for all the FOIs in the area so that the sensitivity of the detector does not depend on the number of valid observations in the time-series of the FOI. An example of the mowing marker mowing intensity raster data for a region in Brandenburg is shown in Figure 11.

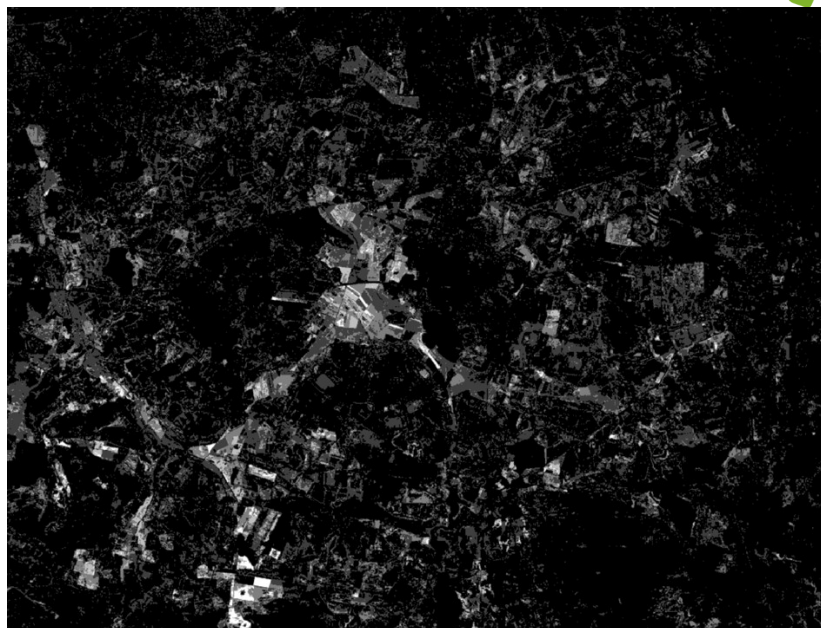


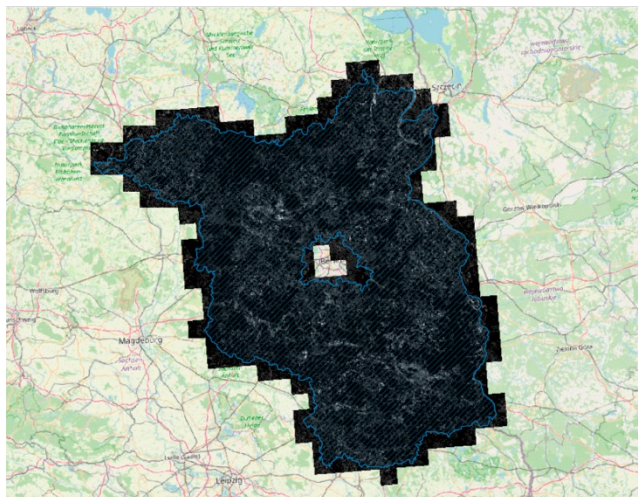
Figure 11: An example of the mowing marker mowing intensity raster data for a region in Brandenburg. The brighter pixels represent higher mowing intensity detected.

The mowing marker for the purpose of running SDM tests was calculated for the BirdWatch test region of Flanders and for the Germany's Brandenburg region. Also, first mowing dates were mapped for these two regions. The time period chosen was 2018 as bird occurrence data is available for that period for both regions – as per SDM requirements listed and described in Section 2.4.

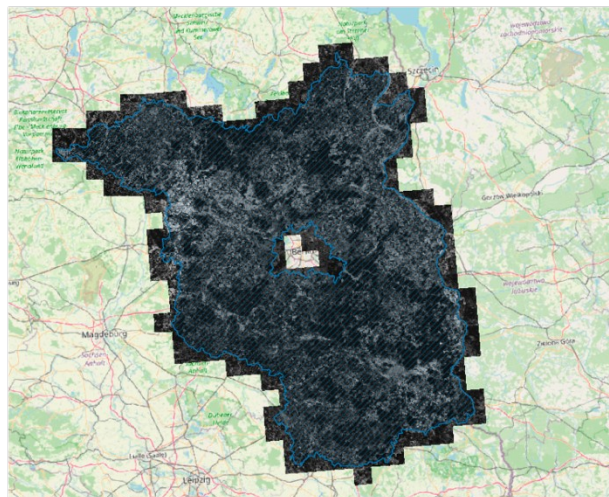
Figure 12 shows the mowing intensity and first mowing date maps respectively for the two regions. Please note that the lower number of detected mowing events in the middle of Flanders (visible as a large NE/SW-strip in the lower two images) is an artefact which is due to overlapping satellite swaths. The middle of the AOI has no swath overlap and therefore, fewer mowing events were detected.

Brandenburg region, Germany

Mowing intensity map

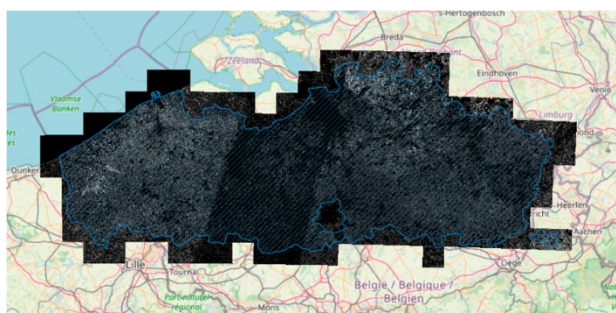


First mowing date map



Flanders region, Belgium

Mowing intensity map



First mowing date map

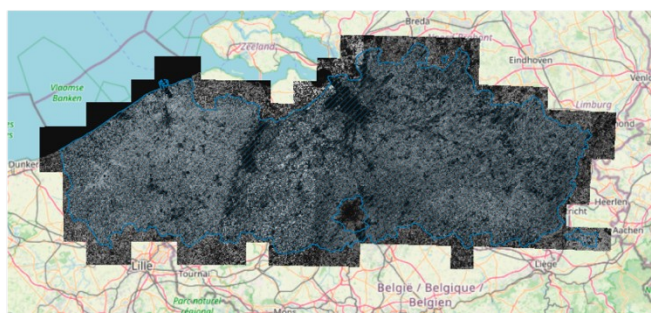


Figure 12: Mowing intensity and first mowing date maps for regions of Brandenburg and Flanders.

4.3 Access to raster data

Sentinel Hub is a cloud-based platform developed by Sinergise that provides access to raster EO data. The raster data are available through the Sentinel Hub platform via the “Bring your own Data” functionality. The openEO framework is a platform for accessing EO datasets that will be used throughout the BirdWatch project.

4.3.1 BYOC SH

The Bring Your Own Data (BYOD) functionality of Sentinel Hub is a feature that allows users to integrate and process their own satellite or geospatial data within the Sentinel Hub environment.



**Funded by
the European Union** 27

With the BYOD functionality, users can leverage the capabilities of Sentinel Hub to analyse their own custom data alongside the existing data available in the platform. This feature enables seamless integration of user-provided data with the extensive range of satellite imagery and derived products already accessible in Sentinel Hub.

To utilize the BYOD functionality, we uploaded the GeoTIFF mowing marker data to the Sentinel Hub environment. The supported data formats for BYOD include various other raster and vector formats commonly used in the geospatial domain, such as NetCDF, Shapefile, and others. To ingest data into Sentinel Hub platform as BYOD, the following steps must be taken:

- Storing the raster data in the cloud optimized geotiff (COG) format on AWS S3 bucket.
- Configuring the bucket's permissions so that Sentinel Hub can read them.
- Importing tiles using the Dashboard or API.

The BYOD functionality of Sentinel Hub is described in more detail in the deliverable D3.2 of the BirdWatch project. More information on the use of the Sentinel Hub BYOD API can also be found in the Sentinel Hub documentation: <https://docs.sentinel-hub.com/api/latest/api/byoc/>.

With the BYOD functionality of Sentinel Hub, different raster data can be uploaded and visualized and will be used further in the BirdWatch project for farmland mapping or could be potentially also used for ingesting SDM outputs and creating different partial maps.

4.3.2 openEO framework

The OpenEO framework is an open and standardized approach to enable interoperability and access to Earth observation (EO) data and processing services. It aims to simplify and streamline the use of EO data by providing a common interface and set of APIs that developers can use to access and process data from multiple sources.

OpenEO defines a set of common data formats, metadata standards, and API specifications that allow users to interact with a variety of EO data providers and process backends consistently. This standardization helps to overcome the fragmentation and incompatibility issues that often exist in the EO data ecosystem.

Figure 13 summarizes the use of openeo restful API and the implementation of pre-defined processes in R, python and Javascript clients.



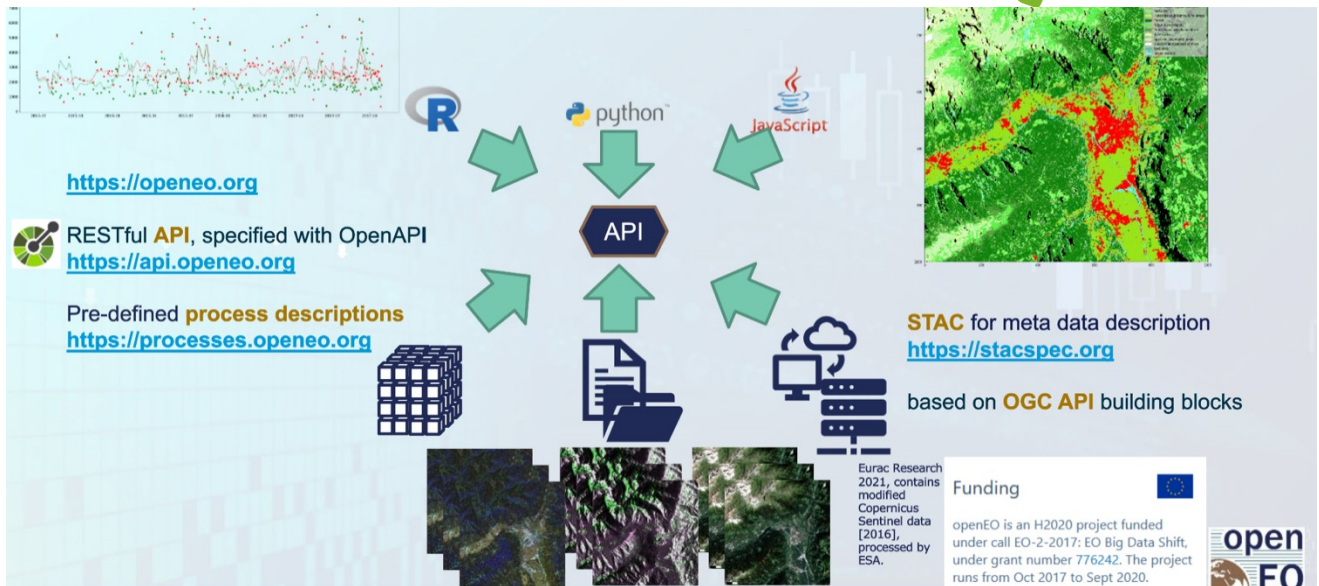


Figure 13: Overview of OpenEO client sides, API and pre-defined workflows.

The key components of the OpenEO framework include the following:

- **API Specification:** OpenEO defines a unified API specification that enables users to access and manipulate EO data and processing services. This specification outlines the endpoints, parameters, and responses that developers can use to interact with the system.
- **Data Providers:** OpenEO supports multiple data providers, which can include satellite imagery providers, data archives, and other EO data sources. These providers expose their data through the OpenEO API, allowing users to seamlessly access data from different sources.
- **Processing Backends:** OpenEO is designed to work with various processing backends, which can include cloud-based platforms, high-performance computing clusters, and local processing environments. The framework abstracts the underlying processing infrastructure, allowing users to switch between backends without modifying their applications.
- **User Applications:** Developers can build EO applications on top of the OpenEO framework using their preferred programming language. The framework provides libraries and client SDKs (Software Development Kits) to facilitate the development process.
- **Scalability and Parallelism:** OpenEO supports scalable processing by leveraging parallel computing capabilities of the underlying processing backends. This allows users to efficiently process large volumes of EO data and perform complex analyses.
- **Geospatial Operations:** The framework offers a wide range of geospatial operations and functions that can be applied to EO data, such as spatial filtering, temporal aggregation, data fusion, and machine learning algorithms. These operations can be combined to create custom workflows and analysis pipelines.

The goal of OpenEO is to promote openness, transparency, and collaboration in the EO community by providing a standardized framework for accessing and processing EO data. It enables developers to create innovative applications and services by leveraging the resources available in the EO domain.

Now with a federation set up that splits the processes based on available resources, this provides more efficient processing on large scale datasets. Figure 14 shows how the backend operates as a federation with an aggregator in the middle where jobs are split if needed.

<https://openeo.cloud>

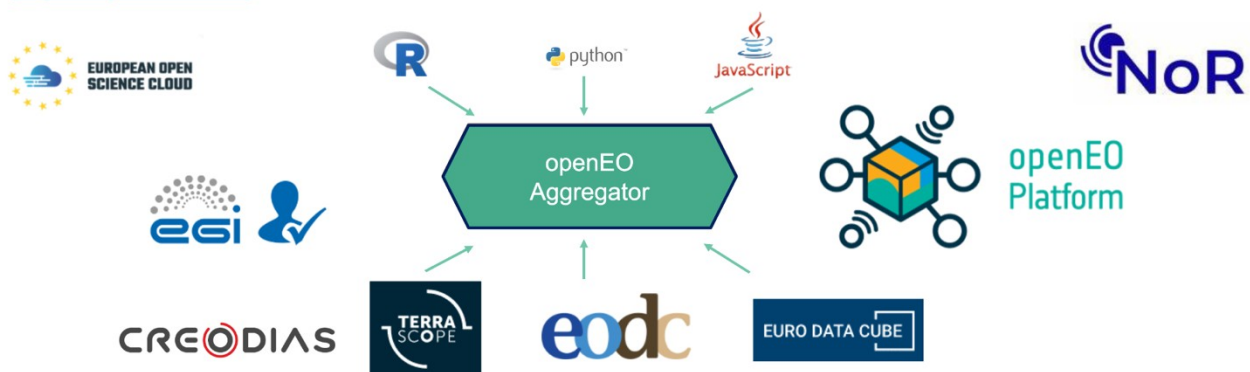


Figure 14: Description of the OpenEO federated backend and the common aggregator.

4.4 Spatio-temporal Asset Catalog API

SpatioTemporal Asset Catalog API (STAC API) is an open standard and specification that defines a web API for searching and accessing geospatial data. It is designed to provide a common interface for working with spatiotemporal data assets, such as satellite imagery, aerial photographs, weather data, and other geospatial information.

The STAC API is built upon the principles of the STAC specification, which aims to standardize the way geospatial data is organized and described. The specification defines a core set of metadata fields and extensions that can be used to describe the properties and characteristics of geospatial assets.

SpatioTemporal Asset Catalog (STAC) is a way to describe data that can be described in terms of time and space. In its simplest form, a STAC catalog describes a set of STAC collections. Each STAC collection contains a set of STAC items. These entities correspond to our notion of providers, collections/datasets, and granules/files.

It provides the specification of a STAC catalog and a STAC API. The STAC catalog can be a static description of data or, when used in conjunction with STAC API, the result of a search of one or more STAC catalogs. STAC allows data providers a great way to provide relay standardized information about their dataset to a potential user without handing over the data itself.

It is a metadata and discovery standard that is becoming widely adopted in the earth science data domain. It helps create workflows and pipelines that can keep data analysis completely in the cloud computing environment.

The STAC API:

- Defines OpenAPI schemas for searching and discovering metadata
- Aligns with and extends the OGC API- Features specification
- Includes a search endpoint for spatiotemporal and attribute queries within or across Collections

The API provides a RESTful interface that allows users to search for data based on spatial and temporal filters. It supports various query parameters that enable users to narrow down their search based on location, time range, data type, and other criteria. The API response typically returns a list of STAC Item objects, which contain metadata and links to the actual data assets.

One of the key features of the STAC API is its ability to support large-scale and distributed geospatial data. It allows data providers to publish their catalogs and assets using a decentralized approach, where each item can be hosted and accessed independently. This decentralized nature enables efficient discovery and access to geospatial data from multiple sources without the need for a centralized catalog.

The STAC API is widely adopted by organizations and platforms in the geospatial community, including satellite imagery providers, Earth observation agencies, geospatial data infrastructure providers, and data science communities. It promotes interoperability, data discovery, and easier integration of geospatial data into various applications, including GIS software, data analysis tools, and machine learning workflows.

The STAC API provides a standardized and efficient way to search, discover, and access spatiotemporal data assets, making it easier for users to work with geospatial information in a scalable and interoperable manner.

Metadata management with STAC API is schematically presented in Figure 15.



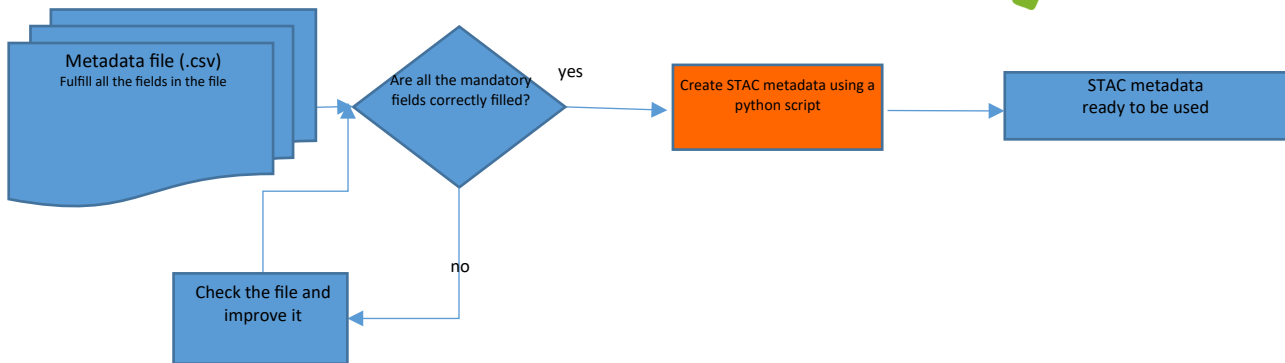


Figure 15: Metadata management with STAC API.

Sentinel Hub is providing a STAC API v1.0 conformant endpoint for discoverability and access to raster metadata information. To see the usage, check the documentation:

<https://docs.sentinel-hub.com/api/latest/api/catalog/>

5 Conclusion

In this document, we have presented the initial BirdWatch project geospatial databases and the technical aspects of data structures and data accessibility that will be used throughout the project.

We explained the dataset selection process, provided detailed descriptions of vector datasets used in the BirdWatch project, including LPIS datasets and the EuroCrops dataset. We also described the the BirdWatch vector database technical structure. We presented the raster geospatial databases, including land cover and mowing intensity maps relevant for bird habitat suitability assessment. The BirdWatch raster database accessibility, frameworks, and standards used were also described.

The work done for the initial preparation of the geospatial databases has served as support of WP2000 and the EO-data will be used for initial testing of the SDM. As the project progresses, the presented geospatial database structures and the datasets used will be expanded and iterated upon, based on the requirements and needs of the SDM and other components of the Birdwatch project.

