Reference Data Module (RDM) Protocol

Document Ref:	ESA World Ecosystem Extent Dynamics (WEED)
ESA Contract No:	4000144495/24/I-NS
Version:	Version 1.0
Creation Date:	Aug 11, 2025
Last Modified:	

Submitted by:IIASA

BC3 BASQUE CENTRE FOR CLIMATE CHANGE Klima Aldaketa Ikergai
EXCELENCIA MARÍA DE MAEZTU
iDiv
IIASA
vito

DOCUMENT RELEASE SHEET

	Name	Signature	Date
Book Captain			
Contributing Authors			
Approval (internal)			
Approval (ESA)			
Distribution			

DOCUMENT CHANGE LOG

Issue	Issue date	Pages affected	Relevant information

Table of Contents

Executive summary **2**
Introduction to the RDM protocol **3**
RDM design prototype **4**
RDM data specification **6**
RDM metadata 8
RDM STAC **9**
Validation **10**
Annex **12**

Executive summary

Availability of high-quality in-situ reference data remains one of the few bottlenecks for training and validating accurate ecosystem type classification models. The World Ecosystem Extent Dynamics (WEED) Reference Data Module (RDM) is an online application which will host global harmonized reference datasets on ecosystem type. Through the RDM and its associated Spatio Temporal Asset Catalog (STAC), users will be able to view, query, download, contribute, share and publish ecosystem reference data. This functionality will be available stepwise in the future through an intuitive user interface and a more advanced API service at https://weed-rdm.iiasa.ac.at/.

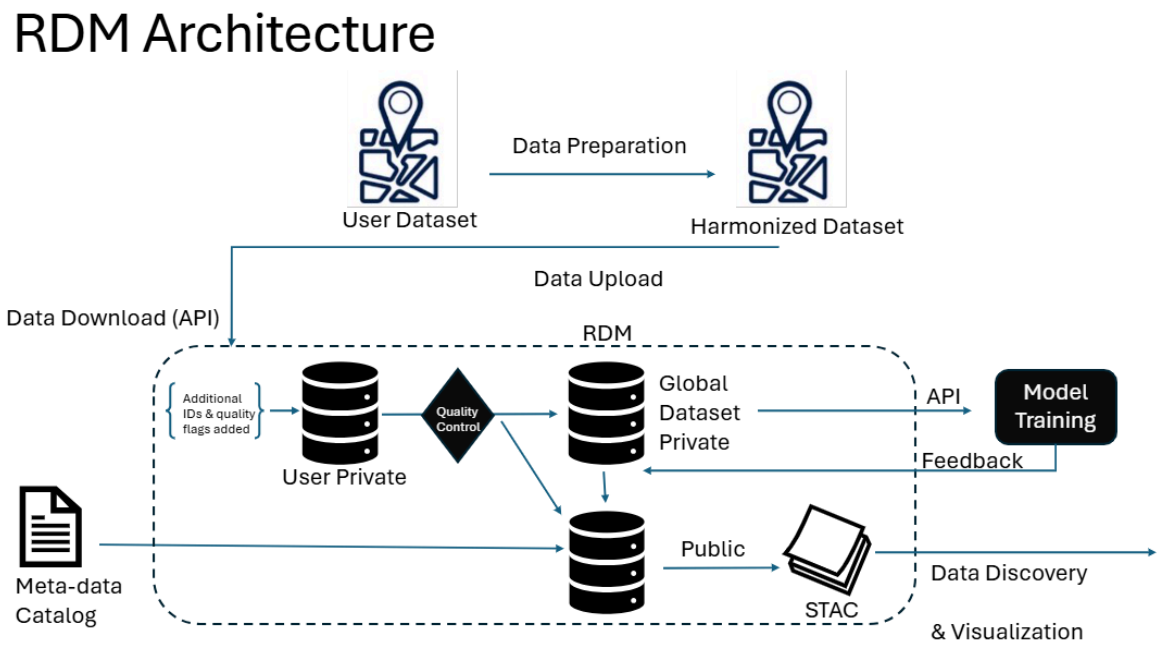
Introduction to the RDM protocol

Reference data underpins the World Ecosystem Extent Dynamics (WEED) solution toolbox. The WEED Reference Data Module (RDM) will host a global collection of harmonized and curated in-situ reference datasets related to ecosystems. First and foremost it will host national reference datasets provided by the WEED Champion Users, complemented with additional global, regional and national datasets. Data providers will specify the associated licence for each dataset. Data access permissions will be controlled such that data providers may specify which data is public and which is restricted. In addition, datasets will be provided with standardized metadata (compatible with the GEO-Atlas, making use of the WEED metadata catalog BExIS2 - Search in Public Datasets) along with attributes mapped to unified typologies (i.e. GET, EUNIS and EU typology for accounting). Built-in automated data quality checks, along with careful curation by WEED data moderators, will ensure high and transparent data quality. Via the RDM, users will be able to view, query, download, contribute to, and share in-situ reference data based on the specific access restrictions associated with any given dataset.

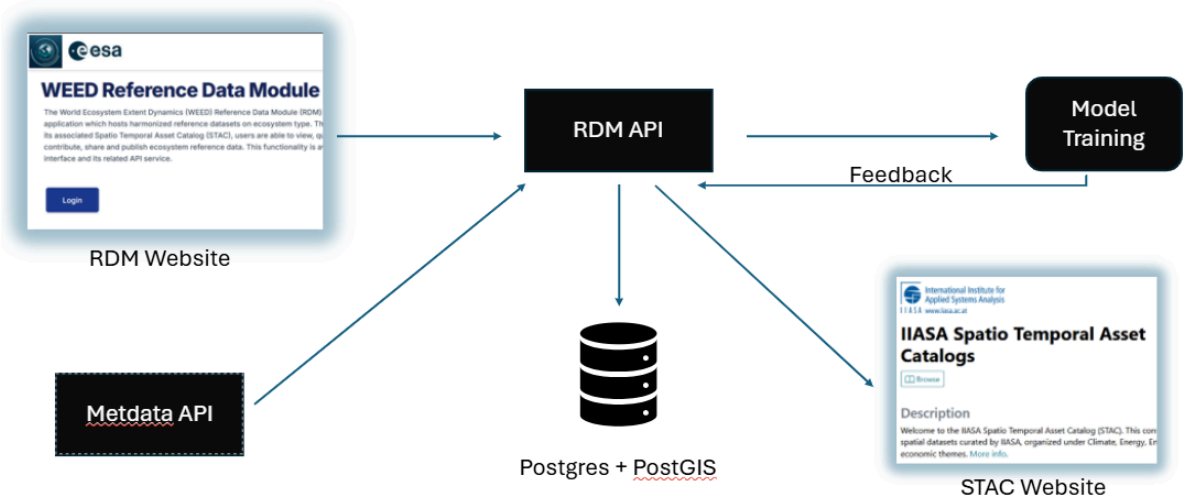
In WEED, reference data will be used for training the Machine Learning (ML) models and for validation of the resulting ecosystem maps (both ML and rule-based). Users will have the ability to upload their own reference data. This data should be in the form of point data (csv format) with geographic coordinates and a column which indicates the ecosystem class of each point (based on the EUNIS or GET typologies - with more typologies to be added in the future). See below for the complete description of field attributes required for upload. In addition to data which WEED Champion Users upload, other reference datasets (local, regional, national, continental, global) may also be added to the RDM. Users can also download all public data. Private data will require registration and access may be restricted.

RDM design prototype

The following briefly describes the WEED RDM workflow and the process to upload data into the RDM. Firstly, a user will assemble their dataset ensuring that basic attributes are present (i.e. geometry, date, etc.) along with the respective meta-data. See sections below for further instructions. In particular, one of the key typology classes must be added to the dataset prior to upload. This can be done by conversion from a national typology or ecological descriptions, or cross-walking from a standard typology (i.e. national typology to IUCN-GET). Support for this step will be provided by the WEED partners (iDIV, UniBonn). Once this is complete, the datasets shall be provided to IIASA for upload. Upon upload, further checks will be performed to ensure consistency and quality. Uploaded data will be made available via the RDM STAC and ultimately the RDM dashboard for data discovery.



Visual representation of how in-situ reference data will travel through the RDM, once implemented.



Current depiction of the RDM software components.

RDM data specification

Prior to upload, reference data needs to be prepared according to the table below, with field names and formats exactly as specified. Mandatory items to be filled are indicated. Unknown items may be assigned NA. Please contact the project partners for support in preparing these files (i.e. iDiv, IIASA). Finally these files should be provided to IIASA for upload into the RDM.

Owing to the close cooperation between the GEO-ATLAS and WEED we have considered the data specifications for the GEO-ATLAS when designing the WEED data specs. As the GEO-ATLAS data specification is designed for the specific task of collecting a new global ecosystem reference dataset from image interpretation, some unique features can be expected in their structure. The WEED data specification on the other hand needs to be able to incorporate reference data from a variety of sources, additional typologies beyond the IUCN-GET etc., and hence will be more generic in design. The WEED RDM will however be able to directly ingest the GEO-ATLAS reference data. Basic items such as IDs, latitude, longitude, timestamps and ecosystem codes are common to both specifications. The WEED specification will however separate descriptive meta-data, providing that in a related metadata file described in the next section.

Prior to uploading any dataset, it is important to assign at least one of the global standard typologies to a dataset (e.g. EUNIS or IUCN-GET are supported at this time). This allows for both training (via the ML pipeline) or validation of the resultant map products via the ARIES tool, assuming the reference data was intended for validation. Support for this step will be provided by the WEED partners (iDive and UniBonn).

Reference dataset template describing the 10 attributes required for upload into the RDM.

Attributes required for WEED RDM data upload *				
Attribute Name	**Item Type**	**Example**	**Description**	**Mandatory**
lat	Float	37.6534	Latitude (y) coordinates: user uploads and RDM converts into WGS84 (EPSG:4326): for polygons/rasters we store centroid	yes
lon	Float	22.3045	Longitude (x) coordinates: user uploads and RDM converts into WGS84 (EPSG:4326): for polygons/rasters we store centroid	yes

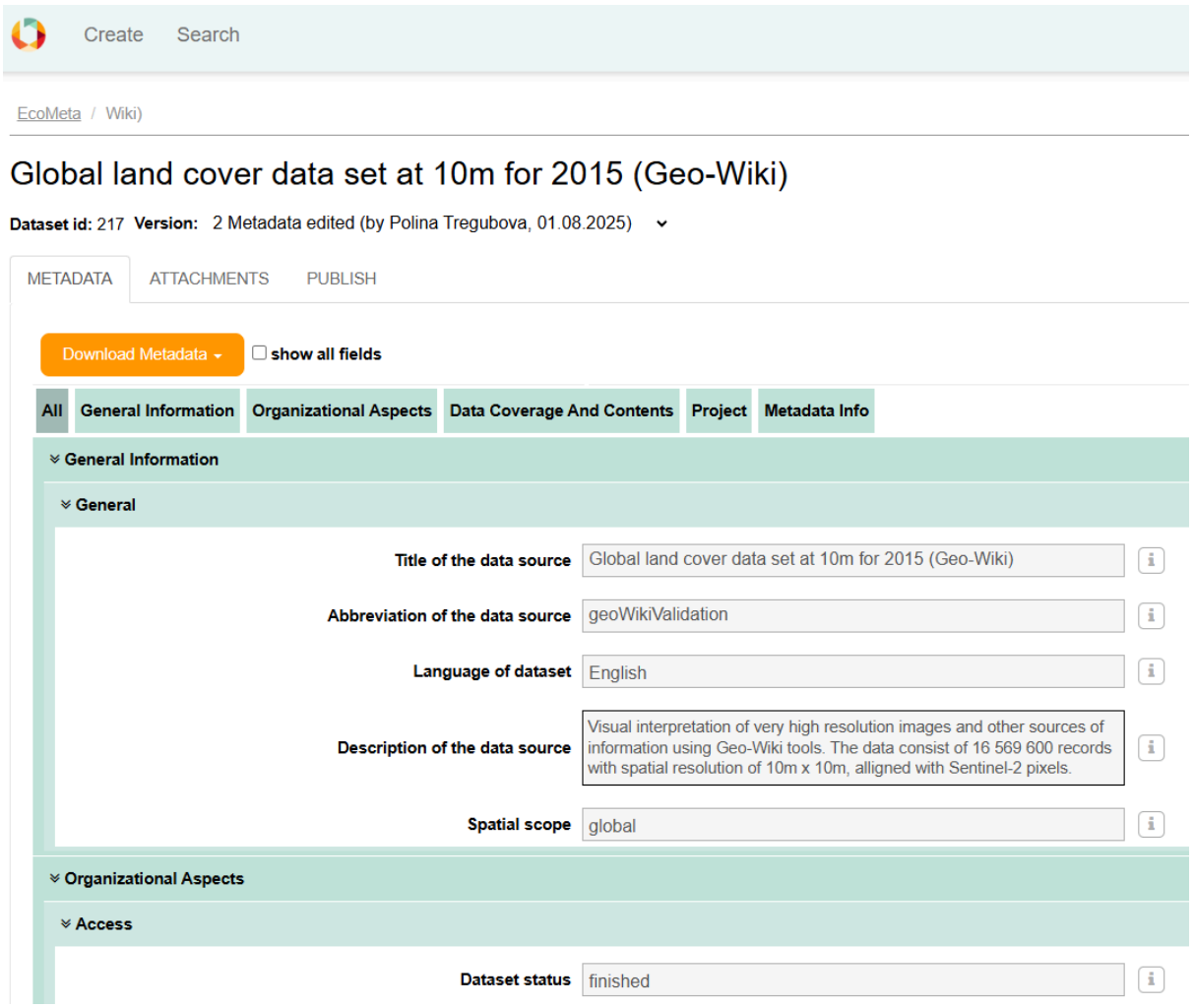
orig_id	Integer	345	Original ID from reference data	no
orig_class	Character	Forest	Original reference data class i.e. crop species, landcover type, ecosystem type, etc.	no
description	Character	Dry coniferous forest	Additional information to describe each location	no
timestamp	Date	1/1/2020	The date at the time of sampling (insitu or image)	yes
eunis2021plus	Character	Mediterranean evergreen Quercus forest	Dominant ecosystem code or class name at level 3 if possible	one of these at a minimum
iucn_get	Character	Annual croplands	Dominant ecosystem code or class name at level 3 if possible	one of these at a minimum
EU	Character	Dense Urban Areas	Dominant ecosystem code or class name at level 3 if possible	one of these at a minimum
type	Character	Training	Specify if data is for training or validation purposes	yes
*** all other data to be entered in WEED metadata catalog**				

RDM metadata

Metadata required for each uploaded file shall be provided via the WEED meta-data catalog (https://ecosystems-metadata.uni-jena.de/). These are items associated with the dataset which do not need to be stored in the datafile itself, and refer to items such as the dataset name, description, language and citation. This will be used to populate the RDM STAC and will be visible to anyone browsing the RDM data, in order to understand the data provenance.

The fields present here are aligned with the GEO-ATLAS while providing a complete description of the file for the RDM STAC. In the case of the GEO-ATLAS, a total of 32 fields are annotated onto every record and contain a mixture of both data and metadata. As explained above, owing to the difference in design of the two initiatives, both the WEED RDM data and the WEED meta-data specification needs to be somewhat broad to fully describe the variety of incoming datasets.

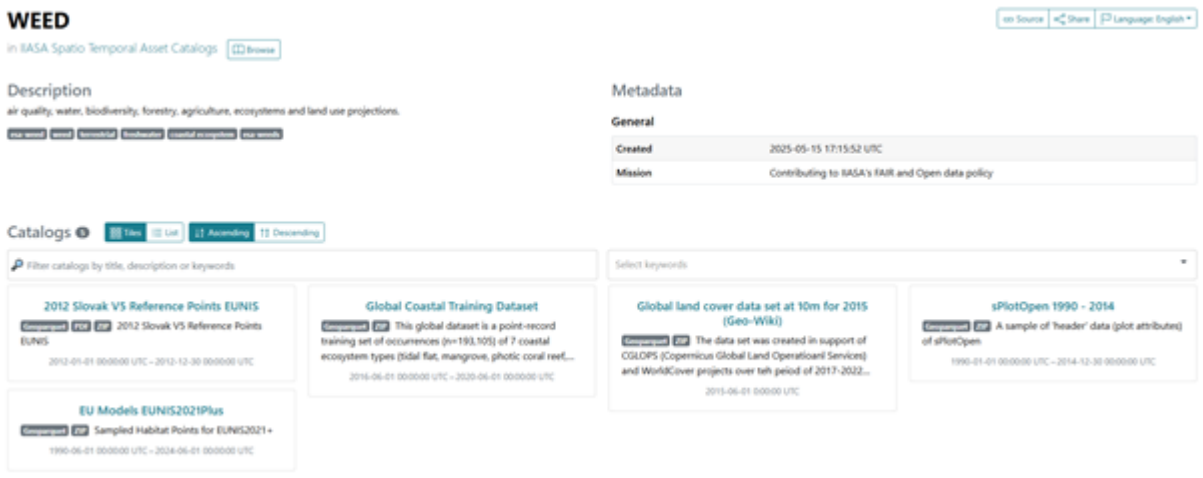
Metadata template example for a reference dataset to be uploaded into the RDM.



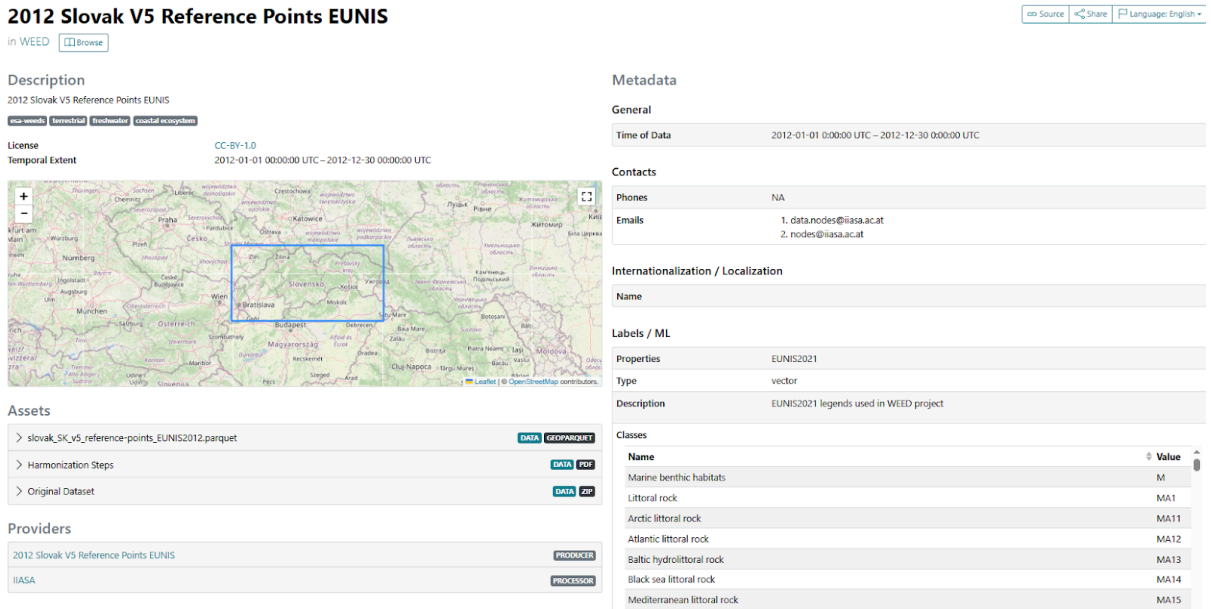
RDM STAC

Once ingested into the RDM, public reference data will be placed into the WEED RDM STAC (i.e. if the user who uploads the data indicates that this data is public, it will appear in the RDM STAC). Access to the RDM STAC will be supported via the https://dataspace.copernicus.eu/ and further instructions will be provided once this feature is in place. Several example datasets are currently available in the RDM STAC, as shown below.

The RDM STAC contains several examples of reference data uploaded via the RDM.



Selecting an individual dataset, a page is provided which shows the dataset meta-data and provides options for downloading the data files (eg. Slovakia 2012 EUNIS). Further instructions will be provided for accessing the datasets once the RDM workflow is complete.



Validation

Ecosystem maps generated via the WEED solution (i.e. k.Lab) will be provided in various formats (e.g. geoTIFF) and available for download. The maps will be in one or more of the accepted typologies (e.g. IUCN-GET, EUNIS, etc.), depending on which typology the user selects for mapping and which typologies are available. The resulting ecosystem class names or codes in the map must be directly comparable to the class names or codes in the RDM reference data. A variety of validation workflows will then be available to the user, including the following described below. All of these approaches rely on IIASA's Laco-Wiki Map Accuracy Tools (https://tool.laco-wiki.net/).

Validation using a single RDM reference dataset (currently implemented)

If all conditions are met (i.e. the map and the validation data are directly comparable) a quantified validation may be performed. Selecting an uploaded validation dataset in the RDM, select the validation button. This brings the user to a page which allows for selection of the respective typology, and an option to upload a geotiff map, along with the associated QML file (Once the QML files are standardized they will not need to be uploaded and this option would disappear). Selecting submit, a confusion matrix will automatically be generated and downloaded as an excel table for analysis. This approach relies on the user ensuring that the validation points are both accurate and representative of the region.

Validation using a map (forthcoming)

This option will allow for the upload and subsequent validation of an ecosystem map. Based on the map coordinates (bounding box), all available reference data in the RDM will be used to validate the map. Selecting submit, a confusion matrix will automatically be generated and downloaded as an excel table for analysis. In the future, a sampling scheme could also be incorporated to select validation points from the RDM in a robust manner.

Once we have completed testing of the workflow, further instructions will be provided on how to perform this validation.

RDM Maintenance and Sustainability

IIASA will continue to host and maintain the WEED RDM beyond the lifetime of the WEED project. This will be accomplished in coordination with the World Cereal RDM which continues to be funded for the next several years. Should no further funding be provided for the WEED RDM, basic maintenance and support will be provided by IIASA. Users may continue to contact IIASA for support in the future, as currently provided.

Annex

Data specification (including metadata) of the Global Ecosystems ATLAS training dataset, along with the equivalent field in the WEED data specification (or metadata).

Condensed Field Names (SHP)	Field	Required	Format	Collected in Earth System Studio?	Description	WEED Equivalent Field
date	date	Y	Date	Automatically populated	Date indicating the date of data collection (MM-DD-YYYY)	Timestamp (in metadata)
sample_id	sample_id	Y	Integer	Sample: pre-populated before ingestion Interactive: automatically populated	Individual identifier of each sample point.	ID
grid_id	reference_grid_id	Y	Text	Automatically populated	Unique identifier that refers to Global Ecosystems Atlas reference grid	Mgrsid10, grid20id
latitude	latitude	Y	Numeric	Automatically populated	Latitudinal coordinates in decimal degrees	center_lat
longitude	longitude	Y	Numeric	Automatically populated	Longitudinal coordinates in decimal degrees	center_lon
method	method	Y	Text	Automatically populated	Principal method of data acquisition in the primary source project (usually image interpretation).	Origin (also in metadata)
int_min	interpretation_scale_min	Y	Integer	10m	Indicator of minimum pixel scale (meters) at which the point should be used.	NA
int_max	interpretation_scale_max	Y	Integer	100m	Indicator of maximum pixel scale (meters) at which the point should be used.	NA
iucn_realm	iucn_realm	Y	Text	Automatically populated from value in iucn_efg_code_10m	Realm as defined by the IUCN Global Ecosystem Typology.	NA
iucn_biome	iucn_biome	Y	Text	Automatically populated from value in iucn_efg_code_10m	Biome as defined by the IUCN Global Ecosystem Typology	NA
iucn_efg	iucn_ecosystem_functional_group	Y	Text	Automatically populated from value in iucn_efg_code_10m	Ecosystem Function Group as defined by the IUCN Global Ecosystem Typology	NA
efg_10m	iucn_efg_code_dominant_10m	Y	Text	Specified during training data collection/annotation	Alphanumeric code representing an Ecosystem Function Group as defined by the IUCN Global Ecosystem Typology at 10 meter pixel scale	Iucn_get Eunis2012 Eunis2021plus Ramsar (dominant ecosystem code at level 3)

efg2_10m	iucn_efg_code_secondary_10m	Y	Text	Specified during training data collection/annotation	Alphanumeric code representing a second majority Ecosystem Function Group as defined by the IUCN Global Ecosystem Typology at 10 meter pixel scale	NA
efg_100m	iucn_efg_code_dominant_100m	Y	Text	Specified during training data collection/annotation	Alphanumeric code representing an Ecosystem Function Group as defined by the IUCN Global Ecosystem Typology at 100 meter pixel scale	NA
efg2_100m	iucn_efg_code_secondary_100m	Y	Text	Specified during training data collection/annotation	Alphanumeric code representing a second majority Ecosystem Function Group as defined by the IUCN Global Ecosystem Typology at 100 meter pixel scale	NA
source	source	Y	Text	Automatically populated	Organisation/Project source of the training data	In metadata
producer	data_producer	Y	Text	Automatically populated	Group responsible for producing the training	In metadata
int_name	interpreter_name	Y	Text	Automatically populated or specified during training data collection and/or annotation	Name of the analyst who collected the training data point	NA
int_conf	interpreter_confidence_level	Y	Integer	Specified during training data collection/annotation	Confidence levels expressed using five qualifiers that relate to a numeric value of between 1 and 5: "very low","low","medium","high","very high". See Mastrandrea et al. (2010) Guidance note for lead authors of the IPCC fifth assessment report on consistent treatment of uncertainties.	Quality_flag (NA, pass or fail, indicates if correct, or requires checking)
rev_name	reviewer_name	Y	Text	Specified during training data review	Name of person reviewing the training data point	NA
rev_conf	reviewer_confidence_level	Y	Integer	Specified during training data collection/annotation	Confidence levels expressed using five qualifiers that relate to a numeric value of between 1 and 5: "very low","low","medium","high","very high". See Mastrandrea et al. (2010) Guidance note for lead authors of the IPCC fifth assessment report on consistent treatment of uncertainties.	Validated (NA, true, false, confirms if point has been reviewed or not)
samp_type	sample_type	Y	Text	Automatically populated	The sampling type (stratified or interactive) used to identify a training	In metadata

					data point.	
year_start	valid_year_start	Y	Integer	Specified during training data collection/annotation	The start year for which the training data point has been confirmed against a reference imagery	Timestamp (see metadata)
year_end	valid_year_end	Y	Integer	Specified during training data collection/annotation	The end year for which the training data point has been confirmed against a reference imagery	Timestamp (see metadata)
hom_10m	homogeneity_estimate_dominant_10m	Y	Integer	Specified during training data collection/annotation	The percentage covered by the dominant Ecosystem Function Group within a 5 meter radius of the training data point (or by an equivalent grid cell)	NA
hom2_10m	homogeneity_estimate_secondary_10m	Y	Integer	Specified during training data collection/annotation	The percentage covered by the second most dominant Ecosystem Function Group within a 5 meter radius of the training data point (or by an equivalent grid cell)	NA
hom_100m	homogeneity_estimate_dominant_100m	Y	Integer	Specified during training data collection/annotation	The percentage covered by the dominant Ecosystem Function Group within a 50 meter radius of the training data point (or by an equivalent grid cell)	NA
hom2_100m	homogeneity_estimate_secondary_100m	Y	Integer	Specified during training data collection/annotation	The percentage covered by the second most dominant Ecosystem Function Group within a 50 meter radius of the training data point (or by an equivalent grid cell)	NA
int_com	interpreter_comment	Y	Text	Specified during training data collection/annotation	Other relevant interpreter comments providing additional context	NA
rev_com	reviewer_comment	Y	Text	Specified during training data review	Reviewer feedback on training data point	NA
task_id	ess_task_id	Y	Text	Automatically populated within ESS	ESS generated ID for each task (includes both sample points and interactive points)	NA
annot_id	ess_annotation_id	Y	Text	Automatically populated within ESS	ESS generated ID for each annotation (sample point)	NA
NA						Orig_id (ID from original dataset for maintaining link to original data)
NA						Orig_class Original class (e.g. land cover, crop species, etc.)
NA						Description (Additional description

						about the plot location - for example a general description of overall site in which the plot is located)