Guide for filling the template to instantiate the LULC ontology

Martin Cubaud – October 2024

Content

[General principles of the template 2](#_heading=h.gjdgxs)

[1) Paper metadata 3](#_heading=h.30j0zll)

[a. Doi - mandatory information 3](#_heading=h.1fob9te)

[b. Title 3](#_heading=h.3znysh7)

[c. type of publication 3](#_heading=h.2et92p0)

[d. Authors 3](#_heading=h.tyjcwt)

[e. Affiliation Name 3](#_heading=h.3dy6vkm)

[f. Affiliation address 3](#_heading=h.1t3h5sf)

[g. Journal 3](#_heading=h.4d34og8)

[h. Year 4](#_heading=h.2s8eyo1)

[i. Keywords 4](#_heading=h.17dp8vu)

[j. Abstract 4](#_heading=h.3rdcrjn)

[k. Number of citations 4](#_heading=h.26in1rg)

[2) Process type 4](#_heading=h.lnxbz9)

[3) Procedure 4](#_heading=h.35nkun2)

[4) Instruments 4](#_heading=h.1ksv4uv)

[a. Algorithms 4](#_heading=h.44sinio)

[b. Tool used names 5](#_heading=h.2jxsxqh)

[c. Tool used types 5](#_heading=h.z337ya)

[d. Tool used is collaborative 5](#_heading=h.3j2qqm3)

[5) Input data 5](#_heading=h.1y810tw)

[a. Input data names 5](#_heading=h.4i7ojhp)

[b. Input data natures and resolution 5](#_heading=h.2xcytpi)

[c. Input data date 5](#_heading=h.1ci93xb)

[d. Input is VGI 6](#_heading=h.3whwml4)

[e. Input data raster/points/lines/polygon 6](#_heading=h.2bn6wsx)

[f. Input is training, validation, both or neither 6](#_heading=h.qsh70q)

[g. Training dataset size 6](#_heading=h.3as4poj)

[h. Validation dataset size 6](#_heading=h.1pxezwc)

[6) Output data 7](#_heading=h.49x2ik5)

[a. Output data names 7](#_heading=h.2p2csry)

[b. Output data natures and resolution 7](#_heading=h.147n2zr)

[c. Output data raster/points/lines/polygon 7](#_heading=h.3o7alnk)

[7) Operator 7](#_heading=h.23ckvvd)

[a. Operator type 7](#_heading=h.ihv636)

[b. Operator description 7](#_heading=h.32hioqz)

[8) Study area 7](#_heading=h.1hmsyys)

[a. Study area name 7](#_heading=h.41mghml)

[b. Geographic extent type 8](#_heading=h.2grqrue)

[c. Belongs to country 8](#_heading=h.vx1227)

[9) Classification 8](#_heading=h.3fwokq0)

[a. If classification, nomenclature name 8](#_heading=h.1v1yuxt)

[b. If classification, nomenclature level 8](#_heading=h.4f1mdlm)

[c. If classification, nomenclature classes 8](#_heading=h.2u6wntf)

[d. Number of classes 8](#_heading=h.19c6y18)

[10) Validation metrics 9](#_heading=h.3tbugp1)

[a. OA 9](#_heading=h.28h4qwu)

[b. mF1 9](#_heading=h.nmf14n)

[c. mIoU 9](#_heading=h.37m2jsg)

[d. kappa 9](#_heading=h.1mrcu09)

[e. global recall (producer accuracy) 9](#_heading=h.46r0co2)

[f. global precision (user accuracy) 10](#_heading=h.2lwamvv)

[g. per class binary accuracy 10](#_heading=h.111kx3o)

[h. per class recall (producer accuracy) 10](#_heading=h.3l18frh)

[i. per class precision (user accuracy) 10](#_heading=h.206ipza)

[j. per class F1 score 10](#_heading=h.4k668n3)

[k. per class IoU 10](#_heading=h.2zbgiuw)

[l. user defined algorithm quality assessment metrics 11](#_heading=h.1egqt2p)

[11) Criterion of contributions of the paper 11](#_heading=h.3ygebqi)

[a. CodeAvailability 11](#_heading=h.2dlolyb)

[b. DataAvailability 11](#_heading=h.sqyw64)

[c. Challenge 11](#_heading=h.3cqmetx)

[d. Strength 11](#_heading=h.1rvwp1q)

[e. Weakness 11](#_heading=h.4bvk7pj)

[**Questions and advanced use of the template 11**](#_heading=)

# General principles of the template

The template is an Excel file created to help the instantiation of the LULC ontology. The template does not contain every aspect that can be described in the ontology. All the definitions are within the ontology and reminded here. Once the template is filled, a python code is used to automatically instantiate the ontology in .owl format. The ontology can then for instance be visualized with tools like WebProtégé.

Some general guidelines on how to fill the template:

* If an information is not known or not given in the article, you can leave the cell empty.
* For some sections, you can enter a list of semicolon (‘’;’’) separated values. If you do so, it is most of the time expected that the other related cells of the section contain a list of the same length. Spaces around the semicolon are possible but not mandatory.
* If in such a list, one of the values is not known or given by the article, you should leave a blank in such a way that the number of semicolons stays the same.
* A row in the template corresponds to a process of an article (cf 2) process type). To identify the rows as the same article, you should at least fill the doi cell for each row.
* Sometimes you are asked to choose a value in a hierarchy of classes of the ontology. You must choose the one that corresponds the most precisely, at any level of the hierarchy.

# Paper metadata

This part contains the metadata of the paper such as DOI, title, authors, keywords, abstracts, etc. All the elements are listed below.

You can just fill the doi and leave the remaining empty since all the metadata will be automatically extracted from the doi of the article. The doi should be repeated on each row of the template concerning the same article.

## Doi - mandatory information

Defines the Digital Object Identifier (DOI)

## Title

Defines the title of the publication. The sub-title needs to be part of the title. The title can be automatically generated from the paper metadata

## type of publication

It defines the type of the publication paper. It can be a book, chapter, conference, journal or report. It is automatically generated from the paper metadata.

## Authors

The list of the authors of the scientific publication. It is automatically generated from the paper metadata. Here the list should be separated with ‘ and ‘ instead of ‘ ; ‘. Each author should be written in the format « last name, first name ». If only the initials are known, write only the initials.

## Affiliation Name

It describes the name of the affiliation (e.g. laboratory, university, etc.) of each author. It is automatically generated from the paper metadata. It should be a semicolon separated list, with the same length as the number of authors, even if some shares the same affiliation, and in the same order. If the affiliation of one author is not known, leave a blank between two semicolons. If the affiliation of all the authors is unknown, leave the cell blank.

## Affiliation address

It describes the address of the affiliation (e.g. laboratory, university, etc.) of each author. It is automatically generated from the paper metadata. It should be a semicolon separated list, with the same length as the number of authors, even if some share the same affiliation, and in the same order. If the affiliation of one author is not known, leave a blank between two semicolons. If the affiliation of all the authors is unknown, leave the cell blank.

## Journal

Name of the journal or the conference or the book in which the paper is published. It is automatically generated from the paper metadata.

## Year

Year of which the paper was published. It is automatically generated from the paper metadata.

## Keywords

List of the keywords of the paper, separated by semicolons. It is automatically generated from the paper metadata.

## Abstract

Represents the abstract of the publication. It is automatically generated from the paper metadata.

## Number of citations

Defines the number of citations of the paper. It is automatically generated from the paper metadata at the time of the instantiation.

# Process type

A process is a step needed in the lifecycle of LULC data (e.g. classification). A process is defined by input data, output data and resources needed to make the process.

In the context of a review paper, we consider a process to be a research work or a step in meeting the needs of a research goal.

We consider that unless its design is an important contribution of the paper, the validation of each process can be considered as a part of the steps of the process, i.e. the procedure. This also includes the required steps to make this validation (e.g. sampling design…). This allows in case of multiple processes described in a paper to clearly identify how each process is validated.

* Complete: one process per row. If the paper describes multiple processes, i.e. it has several research goals, each one should be in a different row of the Excel template. *Example: change\_detection*
* Use the Excel file prefilled values corresponding to the respective processes defined in the ontology.

# Procedure

The procedure describes how the research goal (the process) is achieved.

Complete the cell by adding the procedure; If the procedure is composed of different steps, you need to separate them with semicolons. *Example : Preprocessing of the Sentinel-2 Imagery ; Change Detection Algorithm applied to construction sites ; photo interpretation.*

# Instruments

## Algorithms

The semicolon separated list of the names of the algorithms developed or applied in the procedure. *Example: Sen2Cor ; linear co-registration ; Normalized Difference Vegetation Index (NDVI) ; change vector analysis (CVA).*

## Tool used names

The semicolon separated list of the names of the tools developed or applied in the procedure. A tool is a human-computer interface that allows a task to be performed (e.g. a software). *Example : Laco-wiki; Paysages*

## Tool used types

Complete for each of the tools, how they were used. Four values are possible: annotation, storage, validation or other.

You have to create a semicolon separated list of the same length as the **column “tool used names”**. *Example: validation; validation;*

## Tool used is collaborative

For each of the tools, you have to precise if the tool is collaborative, i.e. if several people can use the tool collaboratively.

You have to create a semicolon separated list of the same length as the **column “tool used names”**.

Write: **yes** if the tool is collaborative, otherwise **no;**

If you don’t know, you can leave a blank between two semicolons, so that the number of semicolons stays the same.

*Example: yes; yes*

# Input data

## Input data names

The list of names of the data used as input of the process, separated by semicolons. The names can be for instance the name of a sensor, of a database, of a product…

*Example: BD Ortho; lu\_changes\_gers\_HP\_HG\_2019*

## Input data natures and resolution

This column is a semicolon separated list of the same length as the **column “input data names”**, and describes two aspects.

Firstly, it describes the nature or the thematic of the input data. You can choose what is for you the closest nature in the tab “examples\_of\_data\_nature” in the excel file. If you cannot find any corresponding nature, you can write user\_defined\_data(), and write inside the parenthesis what should be in your opinion the nature of the data.

Secondly, for each input data, you have to write its resolution or minimal mapping unit in the format data\_nature:resolution. You have to write the unit of the resolution. If you don’t know the resolution, or if this type of data doesn’t have a resolution, you can write only the data nature.

Example: satellite:20m; addresses; digital\_surface\_model:50cm

## Input data date

This column is a semicolon separated list of the same length as the **column “input data names”**, and describes the date of the data in three possible formats.

* If the data is used at a single date: write only the year of the data; *Example: 2019.*
* If the data is used at a few dates: write the years in which it is used in the format [year1, …, yearN]. *Example: [2016, 2019].*
* If the data is sampled during a period of time: write in the format period (starting\_year, ending\_year). *Example: period(2015-2019).*

## Input is VGI

This column is a semicolon separated list of the same length as the **column “input data names”**.

For each data, write:

* **yes** if the input data is a Volunteered Geographic Information, i.e. if the data is produced by volunteers in a context of citizen science or crowdsourcing initiatives;
* otherwise write **no**.

## Input data raster/points/lines/polygon

This column is a semicolon separated list of the same length as the **column “input data names”**. It describes the geometry of the input data.

For each input data:

* choose between raster, points, lines or polygon.

If an input data doesn’t have any geometry, leave a blank in such a way that the list keeps the same length.

## Input is training, validation, both or neither

This column is a semicolon separated list of the same length as the **column “input data names”**.

For each input data write:

* *training :* if the input data is used to train a machine learning algorithm,
* *validation:* , it the dataset is used to assess the quality of the output of the process,
* *both :* if the input data is used for both the training and the validation;
* *neither* : if none of these cases apply.

## Training dataset size

This column is a comma separated list of the same length as the **number of “input data”** used as training dataset (including the data that are used for both training and validation).

For each training dataset, you have to indicate its size in number of pixels, surface area or number of objects. You should write the unit. *Example: 467 polygons*

If it is not known, you have to leave a blank in such a way that the list keeps the same length.

## Validation dataset size

This column is a comma separated list of the same length as the number of input data used as validation dataset (including the data that are used for both training and validation).

For each validation dataset; indicate its size in number of pixels, surface area or number of objects. You should write the unit. *Example: 689 polygons*

If it is not known, you have to leave a blank in such a way that the list keeps the same length.

# Output data

## Output data names

The semicolon separated list of the names of the data created as output of the process. If they don’t have any given name, you have to create one. The names should be explicit to allow identifying the outputs. *Example: LU\_classification\_2019; LU\_classification\_2021; change\_detection*

## Output data natures and resolution

This column is a semicolon separated list of the same length as the **column “output data names”**, and describes two aspects: the nature of the input data and its resolution or minimal mapping unit.

The format to describe the two aspects is: *data\_nature:resolution*.

* for *data\_nature*: choose what is for you the closest nature in the tab “examples\_of\_data\_nature” in the excel file. If you cannot find any corresponding nature, you can choose user\_defined\_data(), and write inside the parenthesis what should be in your opinion the nature of the data.
* for the *resolution*: write the unit of the resolution. If you don’t know the resolution, or if this type of data doesn’t have a resolution, you can write only the data nature.
* *Example 1: land\_use:10m; Example 2: land\_use*

## Output data raster/points/lines/polygon

This column is a semicolon separated list of the same length as the **column “output data names”**. It describes the geometry of the output data.

For each output data you have to choose between raster, points, lines or polygon. If an output data doesn’t have any geometry, you have to leave a blank in such a way that the list keeps the same length.

# Operator

It describes who did the process. It can be a person or a computer.

## Operator type

It is a semicolon separated list of the operator type: choose between person or computer.

## Operator description

It is a semicolon separated list of the same length as the previous column (operator type).

For each operator: write any relevant information about them.

# Study area

## Study area name

The name of the study area.

If several areas are studied independently, you should separate them with semicolons. Don’t use semicolons if they are studied jointly. For example, if the study area is the union of two regions and they are being analyzed as a single unit, they should not be separated by a semicolon but rather written as "region 1 and region 2."

## Geographic extent type

It is a semicolon separated list of the same length as the previous column (study area name).

For each study area indicate the scale among: ;

* local (e.g. a municipality, a department, or a small area in a department)
* national (a whole country)
* regional (e.g. the Mediterranean Region covering portions of three continents Europe, Africa, and Asia) or
* global (the whole world).

## Belongs to country

It is a semicolon separated list of the same length as the previous column.

* If the study area is local or national: write the name of the country.
* If it is regional: list the countries [country1, country2…].
* If it is global or if you don’t know, you can leave a blank in such a way that the list keeps the same length.

# Classification

Complete this section only if the process involves a (land use and/or land cover) classification; otherwise leave it empty. A nomenclature refers to the set of classes used for the classification.

## If classification, nomenclature name

This column designs the name of the nomenclature.

Write the name of classification. If the article contains several nomenclatures, you can separate them with a semicolon.

If the nomenclature doesn’t have a name, leave the cell empty.

## If classification, nomenclature level

Complete:

* level of the nomenclature 1,2 or 3 if the nomenclature is hierarchical
* 1, if the nomenclature is not hierarchical.

You can separate them with a semicolon, if there are several nomenclatures.

## If classification, nomenclature classes

The semicolon separated list of the classes used. Example: residential ; industrial ; infrastructure ; other ; construction in progress ; destruction ; no change ; unknown.

If there are several nomenclatures, use a pipe “|” to separate the nomenclatures.

## Number of classes

Complete the number of classes of the nomenclature. If there are several nomenclatures, you can separate them with a semicolon.

# Validation metrics

If the process is validated, quantitative metrics are computed to evaluate the quality of the results.

This part is optional. If you consider that it is important for you or for the article, then complete as follows:

* computed : If a metric is computed but you cannot enter a value, you can write computed in the appropriate column.
* values: if the values are available. If they are only included in a figure, you should estimate them as accurately as possible and enter them into the relevant fields of the ontology.
* If the process involves binary classification (e.g. change/ no change), pay attention if the metrics used are the global metrics as defined below, or the metrics for the positive class.

If a classification is evaluated, the confusion matrix M is defined as is the number of elements that belong to the class i and were predicted to the class j; n is the number of classes and N is the total number of predicted elements.

**Global metrics: these metrics are defined for all the classes at the same time.**

## OA

The overall accuracy is defined from the confusion matrix as .

## mF1

The mean F1 score is defined as the macro average of the F1 scores of each class i:

, where F1(i) is the F1 score of the class i (cf per class F1 score)

## mIoU

The mean intersection over union is defined as , where IoU(i) is the intersection over union of the class i (cf per class IoU)

## kappa

Cohen’s kappa is defined as , where Po = OA and Pe =

## global recall (producer accuracy)

The global recall (also known as global producer accuracy, or global sensitivity) is defined as the macro average of the recall values of each class:

## global precision (user accuracy)

The global precision (also known as global user accuracy) is defined as the macro average of the recall values of each class:

**Per class metrics: these metrics are defined for each of the classes separately.**

If the paper provides the results for all the classes, you can list the values for each class in the same order as in the column *“ah. If classification, nomenclature classes”*, separated with semicolons.

If the paper doesn’t use or doesn't show the results for all the classes of the column *“ah. If classification, nomenclature classes”*, you can write the results in the format:

*class\_name1:value1; class\_name2:value2…*

## per class binary accuracy

The binary accuracy for the class i is defined as

## per class recall (producer accuracy)

The recall per class, also known as producer accuracy per class, true positive rate per class, or sensitivity per class, is defined for the class i as

## per class precision (user accuracy)

The precision per class, also known as user accuracy per class or positive predictive value per class, is defined for the class i as

## per class F1 score

The F1 score per class, also known as F-score or F-measure, is defined for the class i as the harmonic mean of the producer and user accuracies of the class i:

## per class IoU

The intersection over union (IoU) per class, also known as Jaccard Index per class, threat score or critical success index, is computed as

**Other metrics**

The metrics previously mentioned are the most often found in LULC articles.

Nevertheless, the ontology has the flexibility to add other user-defined metrics. For example, it is possible to compute **weighted means** for metrics like PA, UA, F1, and IoU, which adjust for class imbalance. These weighted metrics should **not** be entered as global PA, UA, F1, or IoU. Instead, users are encouraged to include these weighted averages under this "user-defined" section. Other variants of the metrics should also be considered as “user-defined” in order to ensure fair comparisons and avoid confusions.

## user defined algorithm quality assessment metrics

This column can contain semicolon separated values.

* use the format *metric\_name:value* for a global user-defined metric
* use the format *metric\_name(class\_name):value* for a per class user-defined metric. Example: specificity(forest):0.95; specificity(water):0.87; mean specificity:0.91; roc curve:computed

# Criterion of contributions of the paper

## CodeAvailability

Write **yes, the link or the doi** if the code of the paper is available; otherwise write no.

## DataAvailability

Write **yes** or a **link**, or a **doi,** if the data of the paper is available; If several links are required, you have to separate them with spaces “ “. Otherwise write **no**.

## Challenge

This column reflects the opinion of the reader on the difficulties of the process. This opinion can be influenced by reading the discussion section of the article.

## Strength

This column reflects the opinion of the reader on the strengths of the methodology used to answer the research question. This opinion can be influenced by reading the discussion section of the article.

## Weakness

This column reflects the opinion of the reader on the weaknesses of the methodology used to answer the research question. This opinion can be influenced by reading the discussion section of the article.

# Questions and advanced use of the template

1. **What should I do if I want to describe a hierarchical land use or land cover nomenclature?**

To describe a hierarchical nomenclature, you should write in the classification section the nomenclatures from low level to high level, and group together the classes from that have the same parent using parenthesis “( )“. The order of the groups should match the previous level of the nomenclature: thus, if a class don’t have children in higher level, you should repeat it or leave a blank between two semicolons.

E.g.: this two level nomenclature :

| Primary production | Agriculture |
| --- | --- |
| Forestry |
| Mining |
| Secondary production |  |
| transport networks, logistics and utilities | Transport network |
| Public utility network |

Should be described in the template as:

| If classification, nomenclature name | If classification, nomenclature level | If classification, nomenclature classes | Number of classes |
| --- | --- | --- | --- |
| OCSGE\_LU\_level\_1; OCSGE\_LU\_level\_2 | 1;2 | Primary production ; secondary production ; transport networks, logistics and utilities | (agriculture ; forestry ; mining ) ; ; (transport network; public utility networks) | 3 ; 5 |

1. **What should I do if the paper is evaluated on several validation datasets, or separately in different study areas?**

If the paper is evaluated on several validation datasets or in different study areas, you can create a new row in the template for each validation dataset or study area. Fill the same information for the paper, the process, the output data, and the operator.

If the paper is evaluated on several validation datasets, you have to only mark one input used as validation dataset for each row. If it is evaluated on several study areas, fill in the study area section the information specific to each study area.

For the validation metrics, you can fill the metrics for each validation dataset or study area.