Sentinel-2 Agriculture

Design Justification File

Milestone	Milestone 2
Authors	CESBIO – Olivier Hagolle, Jordi Inglada, Silvia Valero, Marcela Arias, David Morin CS – Mickael Savinaud, Thierry Rabaute UCL – Sophie Bontemps, Guadalupe Sepulcre, Nicolas Matton
Distribution	ESA - Benjamin Koetz

This page is intentionally blank

Table of recorded changes

Issue Record Sheet

ISSUE	DATE	REASON
0.1	16 October 2014	Initial version
0.2	26 January 2015	First version with the design part
1.0	29 January 2015	1st version for ESA review

Table of contents

List of figures

List of tables

1 Introduction

1.1 Purpose and scope

This document is the Design Justification File (DJF) of the Sentinel-2 for Agriculture (Sen2-Agri) project funded by the European Space Agency (ESA).

The overall objective for the Sen2-Agri project is to provide the international user community with (i) validated algorithms to derive Earth Observation (EO) products relevant for crop monitoring, (ii) open source software and (iii) best practices to process Sentinel-2 (S2) data in an operational manner for major worldwide representative agriculture systems. The project outputs are of different natures. It will deliver (i) a core of processing strategies and (ii) an open source and portable solution to convert the S2 Level 1c (L1C) data into relevant EO products. This system will rely on the Orfeo Toolbox (OTB) framework, making use of existing algorithms when relevant and integrating new ones identified through a benchmarking exercise.

The DJF presents the result of all significant design choices, trade-offs, technical analyses, and benchmarking assessments justifying the design of algorithms and processing chains.

It records all relevant information showing that the proposed solutions meet the requirements.

This DJF document is a major input of the other activities of the task 3 (Figure 1-1) as it defines the main choices made by the consortium for the final system. The DJF is also one of the key inputs of the Task 4 to implement the system.

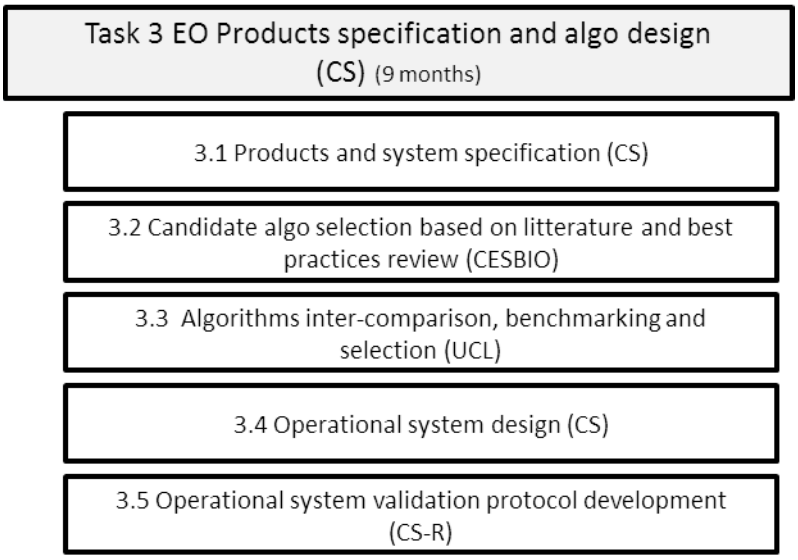

Figure 1-1. Organization of the Task 3 activities (from [AD.1])

1.2 Structure of the document

After this introduction, this document contains 3 main sections that describe:

- The justification report, which will present the benchmarking issues in the forms of individual chapters, as well as a discussion of additional implementation issues;
- The decisions taken in terms of system design;
- The potential gaps between initial and implemented requirements.

1.3 References

1.3.1 Applicable documents

ID	Title	Code	Issue	Date
AD.1	Statement of Work for Sentinel-2 Agriculture Project	EOEP-DUEP-EOPS-SW-13-0004	1.0	26/03/2013
AD.2	Technical proposal of the Sentinel-2 Agriculture Project		1.0	
AD.3	Sentinel-2 for Agriculture User Requirements Document	Sen2Agri_URD_1.2	1.2	25/07/2014
AD.4	Sentinel-2 for Agriculture Technical Specification	Sen2Agri_TS_1.1	1.1	15/10/2014
AD.5	Sentinel-2 for Agriculture Product Specification Document	Sen2Agri_PSD_1.1	1.1	15/10/2014
AD.6	Sentinel-2 for Agriculture Software Development Plan	Sen2Agri_SDP_1.0	1.0	15/10/2014
AD.7	Sentinel-2 for Agriculture Test Data Set documentation	Sen2Agri_TDS_1.1	1.1	08/12/2014

Table 1-1: Applicable documents

1.3.2 Reference documents

ID	Title
RD.1	Earth Explorer Ground Segment File Format Standard, PE-TN-ESA-GS-0001, Issue 1.4, 13 June 2003
RD.2	Sentinel-2 MSI – Level-2A Input Output Data Definition, S2PAD-VEGA-IODD-0001, Issue 1.0, 11 March 2014
RD.3	User, installation and operating manual maccs chains, LAIG-MU-MAC-010-CS, Issue 4.3, 17 Avril 2013
RD.4	The Data Hub Application Programming Interfaces, https://scihub.esa.int/userguide/BatchScripting, last access 27 January 2015

Table 1-2: Reference documents

1.3.3 Acronyms and abbreviations

Acronym	Definition
AD	Applicable Document
TDS	Test Data Set
S4-T5	Spot4-Take5
L8	Landsat 8
CPU	Central Processing Unit
RAM	Random Access Memory
CESBIO	Centre d'Etudes Spatiales de la Biosphere
URD	User Requirement Document
VM	Virtual Machine
TBD	To Be Defined
TBC	To Be Confirmed
CNES	Centre National d'Etudes Spatiales
MACCS	Multi Mission Atmospheric Correction and Cloud Screening
CS-SI	Communications et Systèmes d'Informations
CS-R	Communications et Systèmes d'Informations - Romania
API	Application Programming Interface
DUE	Data User Element
UTM	Universal Transverse Mercator
WRS2	Word Reference System 2
ESA	European Space Agency
JECAM	Joint Experiment for Crop Assessment and Monitoring
S2	Sentinel 2
S2AGRI	Sentinel 2 Agriculture
Sen2Cor	Sentinel-2 L2A processor
LUT	Lookup Table
Sen2-Agri	Sentinel2 - Agriculture
UCL	Université Catholique de Louvain
CFI	Customer Furnished Item
DEM	Digital Elevation Model
ATBD	Algorithm Theoretical Basis Documents
LAI	Leaf Area Indicator

NDVI	Normalized Difference Vegetation Indicator
MGRS	Military Grid Reference System

Table 1-3: Acronyms

2 Justification reports

2.1 Benchmarking

2.1.1 Benchmarking activity

A key activity of the project consists in a benchmarking exercise that aims at selecting the “best” algorithms for fulfilling to a maximum extent to the user requirements and the products specifications [AD.1]. The benchmarking should contribute to improve the understanding of algorithms performance.

For each product, a minimum of 5 concurrent algorithms have to be benchmarked – i.e. applied, assessed and inter-compared in considering the user-focus with critical attention – on a set of pre-defined test sites. A list of 12 sites was defined in the earlier stage of the project [AD.7] and EO and in-situ data were acquired for each of them, constituting the Test Data Set (TDS). This TDS is made of high spatial resolution time series from the year 2013 (images mainly acquired by Spot 4-Take 5 (S4-T5) and complemented with Landsat 8 (L8) and RapidEye (RE) sensors) and of field data from the same year.

The algorithms tested for each product were selected based on literature review and on an exploratory phase carried out to select the most interesting ones out of all possibilities offered by the state of the art.

In order to make this exercise as transparent as possible, the following information was made available in advance:

- The input dataset – namely the TDS – with detailed specifications;
- The specifications of the expected output products [AD.5];
- The detailed characterization of each test site [AD.3 and AD.7];
- The analytical approach for products inter-comparison: for each product, a list of criteria was agreed before the exercise and guided the products assessment and inter-comparison.

2.1.2 Benchmarking reports

The report of the benchmarking exercise is presented as 4 separate chapters, one for each product:

- Benchmarking for L3 monthly composite product;
- Benchmarking for L4 dynamic crop mask product;
- Benchmarking for L4 crop type product;
- Benchmarking for L3 biophysical product.

Each report contains, at a minimum, the following content:

- The rationale behind the selection of the algorithms to benchmark;
- The definition of the benchmarked algorithms;
- The list of evaluation criteria;

- The presentation and discussion of the results;
- The conclusions and algorithms selection.

2.2 Additional implementation issues

2.2.1 Implementation issue related to the high data volume processing

This analysis should be done in considering the two types of demonstration case: national and local. The local case handles a large data volume (less than 1TiB) but which can be managed with classic approach. This is not the case for the data volume of the national case, which is estimated to 4.2 TiB without any compression over a 6-month season.

A large part of this volume (80 %) is linked to the L2A production which is needed along the season to produce the Sen2-Agri outputs products. This data volume imposes a constraint on the hardware because the users need to have large disk storage. This issue is even important if we consider the redundancies needed to avoid data loss.

Another issue raised by the Sen2-Agri system (especially over national cases) is the temporal distribution of the data generation. Indeed, if we consider a national case, the L2A data volume will increase regularly due to the multiple orbits coverage. For example, France is covered by 7 different orbits during a 10-days cycle. Therefore the data volume will increase regularly during the entire cycle and over the season (see Figure 2-1).Using the same reasoning, it should be noticed that the L1C input data flow will follow the same behavior even if we will not store this data.

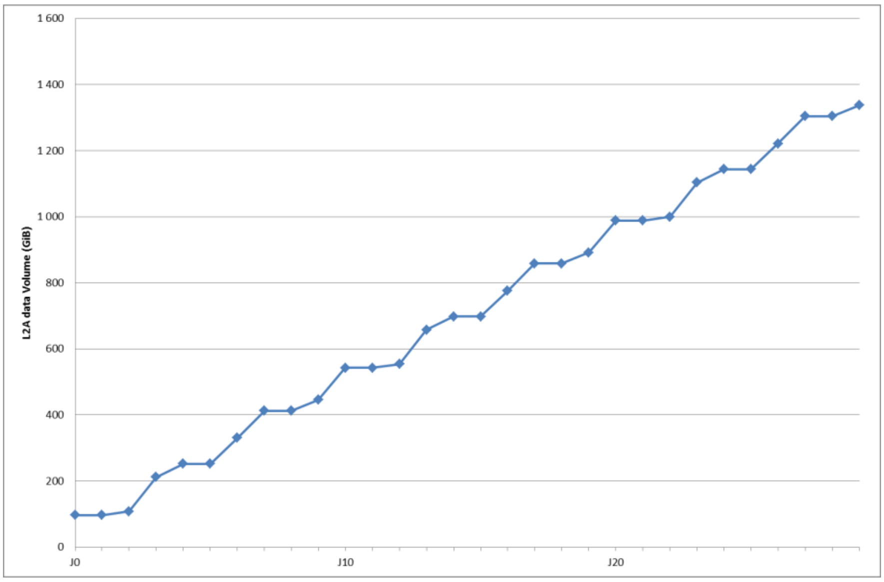

Figure 2-1. L2A data volume over France during one month with one satellite (189 tiles or 2.28M km² are covered during one cycle).

Concerning the Sen2-Agri products, the data volume is lower. Moreover (L4A, L4B and L3A) are delivered monthly. For example, the volume of a monthly composite product over France is estimated at 445GiB, i.e. around 30% of the total L2A volume over the period. The composite products will be stored for the user during a TBD period but not re-used by the system.

Accordingly, the high data volume associated with the national case need to be carefully handled. This is why it is proposed to manage the data storage into two categories:

- Warm storage for data which are needed for immediate processing;
- Cold storage for data which are needed for further processing or dissemination/archiving.

The warm storage could be used for example to store:

- L1C product for the L2A processing
- L2A product for the L3B (biophysical indicator) processing
- L2A product for the L3A (composite) processing
- L2A products for the L4A (crop mask) and L4B (crop type) processing

After processing, input (except L1C) and output data are moved to the cold storage if they are not needed for further immediate processing. For example L2A product generated by the L2A processing would be kept on the warm storage until the L3B processing and L3A processing are completed. After that they would be moved to the cold storage.

The cold storage could be used to store:

- All the L2A products;
- The outputs products delivered to the user.

The warm storage could correspond, for instance, to internal hard disk of a computer. It is quite expensive but it has low latency. The cold storage could correspond, for instance, to external drives connected to a computer. It has a low cost but a high latency. The use of two types of storage allows to limit the cost of the data storage and to reduce the data latency.

In term of CPU and RAM configuration, we estimate that the processing of local case can be handled by personal computer. The local case corresponds to 7 S2 tiles and all the processing planned for the moment can be parallelized by tiles (or by spectral bands). With a computer with 4 available threads, it is estimate that the local case can be dealt with a time frame compatible with requirements.

For the national case which corresponds to 50 S2 tiles, the volume of data to process is more complex to handle by a classical computer. However, we think that, it is not necessary to use a grid of computers (local or into the cloud) to meet the time requirement. According to the first benchmarking results and to the fact that the L1C products will be delivered from different orbits, we estimate that a high quality computer (Xeon Processor with a minimum of 8 threads and 32 GB of RAM) could support the processing of a national case.

To our opinion, cloud computing is not the good solution for national cases. Indeed, the main advantage of cloud computing is the scalability of the available resource. Due to the multiple orbits coverage and the future S2B, the CPU and RAM resources will be busy during the full 10-day (or 5-day with S2B) cycle. Depending on the study area, the CPU and RAM configuration will be constantly used during the season. This is the reason why we think the scalability of the cloud computing is not useful and could be too expensive.

The large input data flow from ESA facilities could also be an issue. For a 6 month season, the maximum volume of L1C data for a national case is estimated at 709 GiB (19.5 GiB by cycle). It could be problematic if the user facility has low internet bandwidth. In the demonstration phase, the system will be operated from a hosting facility with a high input and output internet bandwidth. That will not necessarily be the case for all users. This is why, for users with limited bandwidth resources, it is recommended to move the system to a hosting facility. With this configuration, the user should retrieve only the output products which represent a smaller data volume. If they want, users could create a Virtual Machine (VM) with the Sen2-Agri system and host this VM into the infrastructure of a cloud provider with a dedicated configuration.

In our demonstration configuration, we will use a dedicated server rented in a hosting facility connected to large data storage. We will evaluate the different configurations between rent fixed disk storage and cloud disk storage for the cold storage. We think that the cloud storage could be useful here due to its scalability.

2.2.2 Implementation issues related to the potential synergy with complementary satellites sensor

2.2.2.1 Variety of sensors

As required by the URD [AD.3], The Sen2-Agri system shall be designed to take into account the availability of L8 data into the operational phase. Moreover during the pre-demonstration and the test phases the system shall support the data of the TDS [AD.7]. Into this TDS, the EO data were acquired by different sensors, each one with its own spectral and spatial characteristics. These characteristics are summarized in Figure 2-2.

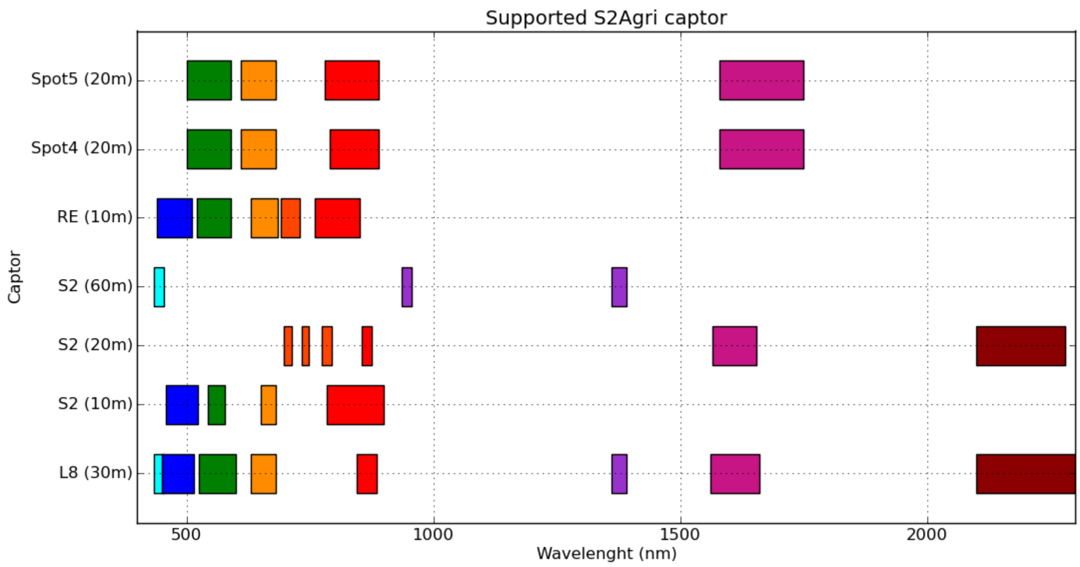

Figure 2-2 : Spectral and spatial characteristics of the sensor supported by the system.

2.2.2.2 Spectral resolution

We can notice that each sensor has its spectral bands included into the set of bands provided by S2. However some special cases need to be clarified.

1) For the operational L8 input data, the L2A processor will deal with all bands[1] but will not provide corrected band 9 (dedicated to cirrus detection) as for S2. All the other processors will therefore not provide or used these bands.

2) S2 provided two spectral bands with central wavelength close to the band 5 of L8. Into the composite processing and for the temporal resampling, we will consider that the band 5 of L8 is associated with the band 8a of S2 (TBC).

For the pre-operational phase, the input data are more heterogeneous and several issues appear when we want to mix different data:

- L8 provide more spectral bands than S4-T5, RE and S5-T5:
  - L8 band 7 is not available into all these sensors;
  - L8 blue band is not available in the Spot sensor;
- RE data has spectral specificities:
  - a Red-Edge band which is not available with the other sensors
  - The central wavelength of the IR band is quite different from the other sensors.

For L8-Spot missing bands, their weights will be put to 0 in the compositing. For the temporal resampling, it will not be a problem: the missing data will be interpolated with the data around the acquisition.

For the RE data, the Red-Edge band is discarded for the compositing and for the temporal resampling when RE data are in minority in the EO series. As for the IR band, it is assumed that the band is similar to other.

2.2.2.3 Spatial resolution

During the operational phase, only S2 and L8 data will cohabit into the system, but L8 will be considered as additional data. Therefore, the L8 data will be resampled after L2A processing to the S2 spatial resolution as follows:

- L8 band 2 from 30m to 10m;
- L8 band 3 from 30m to 10m;
- L8 band 4 from 30m to 10m;
- L8 band 5 from 30m to 20m;
- L8 band 6 from 30m to 20m;

---

[1] except thermal bands which are not considered in the framework of this project.

- L8 band 7 from 30m to 20m.

In the pre-operational phase, the TDS EO data are aligned to the Spot4-T5 data spatial resolution (20m).

2.2.2.4 Data format

Each EO product has its own product format but the system will consider a single product format. This format is the Earth Explorer Ground Segment File Format [RD.1] which is used by MACCS software and which is the one of the EO TDS data. As a result, each internal data has to be aligned on this format. The use of a common internal format is justified by the fact that it enables simplifying the handling of mixed data time series. As for the final products, they can have a different format, the conversion being done in the final steps, just before the delivery.

2.2.2.5 Spatial and temporal coverage

In the operational phase, the input times series will be made of S22 and L8 data. These data do not have the same time revisit and for each sensor, the time revisit is not always constant. For L8, the data are not acquired when a 100 % cloud cover is forecast. Moreover, L8 images are acquired but not processed to Level 1T when the cloud cover prevents from using a sufficient number of ground control points. On our side, in terms of L8 processing, we do not plan to produce L2A data if the cloud cover is more than 90 % (TBC). The same rules will be applied for S2 (TBC).

For both sensors, some area could be covered by two different orbits. For instance, an area could be covered with an acquisition of the days 1 and 4 of a 10-day cycle in case of one satellite. We could therefore obtain an irregular time sampling for one area and another irregular time sampling for another area (covered by different orbits). To solve this issue, it is proposed to apply a temporal resampling on a regular time step. It will allow to apply our classification strategy for each tile independently from the orbits of each sensor.

2.2.2.6 Data access

The access to each product is another key issue when dealing with different sensors.

For L8 there is currently no public API to retrieve data automatically from the USGS Data Center. We could use the Google Earth API to retrieve L8 data according to the partnership made by USGS with Google but it should be confirmed and tested.

For Sentinel-2, the project ngEO should provide a public API to retrieve data (TBC). Currently no specific documentation is available on this new tool. Therefore we will design the system according to the API offered by the Sentinel-1 Scientific Data Hub to retrieve data [RD.4].

2.2.3 Projection Issue

Currently, each input data use the same cartographic projection system, which is UTM. In order to avoid losing time in re-projection, the UTM system will also be proposed as the default one for the whole processing and for the different outputs.

Yet, this projection might not be appropriate for all users. Some countries have a different official cartographic projection system as for instance in France with the Lambert 93 system.

If users need the Sen2-Agri outputs in a different projection than the UTM one, the system will offer them the possibility to select their own output projection.

2.2.4 Mosaicking issue

The S2 products are tiled according to a scheme based on the UTM MGRS system. Each tile is a 100.000-meter grid square identified by its Grid Zone Designator and pairs of letters. For L8, the system is based on the WRS2 Path/ Row system. These two systems should be aligned to simplify the processing and allow to parallelize the processing by tiles. To this end, the S2 tiling scheme will be used as reference.

However, the S2 tiling scheme and the UTM projection lead to specific cases at the UTM zone borders. Indeed in these areas, a geographic point could be covered by one tile in the projection of the West UTM tile and by another tile in the projection of the East UTM tile. The Sen2-Agri software will consider all the tiles in the processing so that some geographical points could appear with two output values.

2.2.5 MACCS multi-temporal issue

MACCS software use a multi-temporal approach to generate L2A product, i.e. to generate L2A at *t*, we need to have generated previously L2A from *t-1* to *t-N*. N is equal to 8 for the Landsat case. At the start of the period, we should wait N images before producing the first one. Similarly, if a long series of L1C product is not available due to the cloud cover, MACCS will reset its history and wait N images before generating the first after the cloudy period. For Landsat the size of the cloudy period is set to 19 days. This MACCS behavior has an impact on the delivery time of biophysical products which need to be delivered as soon as possible after the L1C delivery.

To solve this issue we have explored two strategies:

- Process *N* L1C products before the user defined period to enable the generation of the first L2A product which corresponds to the beginning of the user defined period.
- Use the single date mode of MACCS to process all data during the initialization period and provide “live” biophysical variables during this period but with a lower quality.

A solution to decrease the impact of the single date mode of MACCS is to reprocess the L2A product with the re-processing mode of MACCS and to deliver high quality biophysical data of this period in a second version. However, this solution dramatically increases the complexity of the system. That is why the first one was selected to enable the production of high quality data at the beginning of the user period. Moreover, CESBIO will try to optimize the initialization period during the first step of MACCS with Sentinel-2 and confirm the size of the cloudy period which resets the temporal approach.

2.2.6 Issue related to the use of two different L2A processors

The L2A processor could be based on MACCS or Sen2Cor software. However MACCS and Sen2Cor software have their own interfaces [RD.2 and RD.3]. In this section we will summarize the impact of these different interfaces on the designed system.

Currently MACCS uses the S2-CNES product format while Sen2Cor use a previous version of the final official product format. These two CFIs tools need to be updated outside the project to support the last version of the official S2 product format. We consider that it will be done for the beginning of the Sentinel-2 data availability. After comparison of the input data used by the two software, some differences appear:

- Both need as input an ozone value: for Sen2Cor it should be provided by the ECMWF data embedded in the L1C products (not the case in the current version but it is planned) while for MACCS it should be provided currently by a specific auxiliary data.
- Both need as auxiliary input a DEM data: only the altitude for Sen2Cor and the altitude, the slope and the aspect at different resolutions and a DEM water body mask for MACCS.
- The tools use different types of LUTs for the atmospheric radiative transfer.

These differences of input or auxiliary data could be managed internally by the Sen2-Agri system for the two first points. For example, by embedding a set of global DEM data into the system and by retrieving a user value for ozone. The management of worldwide LUTs and their update will be managed by each CFI tools.

Another point is the fact that Sen2Cor will be able to deal with L1C product with a number of tiles greater or equal to one while MACCS only works with L1C products on one tile. This difference will have no impact on the system behavior because we will use a single tile by processing to control the multi-processing scheme.

As already defined, we consider that the MACCS output format is our internal format for L2A product. Concerning the outputs, Sen2Cor provides less information than MACCS. Here is the list of data not provided by Sen2Cor, while provided by MACCS:

- the Aerosol Optical Thickness (AOT) and WaterVapour (WV) at 10 m
- the Cloud and Cloud shadow masks at 10 m
- the Geophysical mask at 10m
- the Quality mask at 10m
- the sun too low flag, tangent sun flag and the hidden surface masks (at 10 and 20m)
- the AOT pixel mask

For some masks we need to merge some information from classification map and probability masks of Sen2Core to retrieve masks provided by MACCS:

- Cloud and Cloud shadow at 20 m.
- Geophysical mask at 20 m.

The Saturated and Defective pixel masks are mixed at 20m in Sen2Cor while there are available independently from MACCS.

All these differences have a strong impact on the composite generation which needs all the data produced by MACCS to compute the weights used by date. Using Sen2Cor implies to decrease the quality of the composite due to the missing inputs.

3 Design decisions

3.1 General

3.1.1 Overview

The Sen2-Agri System is a standalone processing chain which generates a set of products for agriculture monitoring from Sentinel-2 time series. This chain is composed of a set of independent processing modules orchestrated by a data driven approach. These modules compose a set of tools which can be re-used into other system. This is why these set of modules can be considered as a toolbox.

The logical data flow of the system is described in Figure 3-1. The Figure 3-1describes the data flow as it will be implemented for the operational production system. One idea behind the design is that the final users can re-use the system as it is or only re-use some modules by integrating them in their own processing framework. One of this processing framework is the Sentinel-2 Toolbox. Into this toolbox, the user can run each module independently.

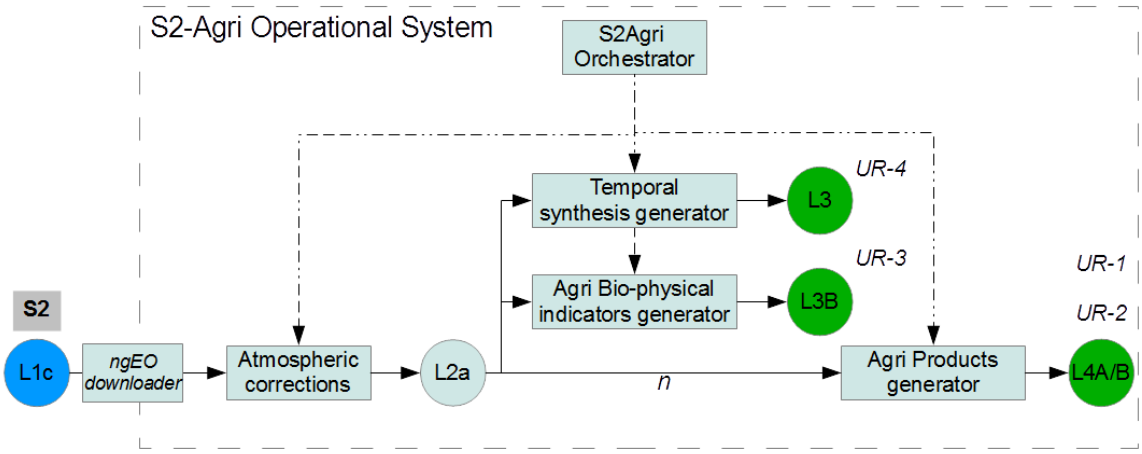

Figure 3-1 : Logical data flow of the Sen2-Agri system

Another idea behind the system is to reduce the human interaction once the operational system is fed with user parameters. The operational system should automatically download data, process them until the end of the season and deliver in time the output products. For that purpose, we will consider a data-driven approach: the processing chain will work as long as input data are available. The availability of input data (L1C products or other mandatory inputs) can be controlled by the user by setting the end of the monitoring time period.

The scalability of the system is a also a critical point due to the high data volume required to provide accurate crop masks or crop type maps. This is why we propose to design an operational system with parallel processing capability and with optimized and low cost data storage. This system will be also compatible with hosting facility or cloud virtualization to benefit from high data bandwidth offered by these services.

To enable multi-processing two strategies are possible:

- Use the native multi-threading capability of Orfeo ToolBox which is the main component of the processing module
- Or use a job scheduler tool to manage a pending queue of processing jobs related to atomic parts of data.

We propose to use the second one which allows to control precisely the data flow. This solution is possible because we can easily restrict OTB applications to run only on a unique thread with a limited RAM consumption.

All this design allows to have a modular system with high processing capability and scalability. Moreover, the final end-user can exploit the Sen2-Agri system at different levels.

3.1.2 Interface of the Sen2-Agri System

The main inputs of the operational system are:

- A time series of Sentinel-2 L1C products automatically downloaded from ESA facility,
- Mandatory user parameters stored into an XML file:
    - Area of Interest,
    - Definition of the monitoring period,
    - Various parameters: time resampling step, L2A processor choice for example
- In-situ data for supervised classification and product validation (included associated metadata or nomenclature).

It should be noticed that the system will support to start without this in-situ data until the crop-type processing starts. It will be also possible to provide different versions of the in-situ data along the monitoring period to support update of the in-situ campaign.

Additional inputs can be provided by the user:

- Additional Landsat 8 L1T product or Landsat 8 Surface Reflectance product (TBD).
- Its own Digital Elevation Model (provided it is more accurate than the default one proposed by the system).
- A previous crop mask.
- A previous crop type map.
- A stratification map suitable to identify different agro-ecological systems within the Area of Interest
- Additional system parameters stored into an XML file: duration of the integration of the biophysical indicator product or size of the composite window for example.
- Additional image processing parameters stored into an XML file, for each module.

During the pre-operational phase, the Sen2-Agri system will support also EO data from the TDS for testing purpose.

A Graphical User Interface (GUI) will be available to define and handle all the requested parameters. Its design will be done to simplify the setting of the mandatory parameters and provide all the advanced ones into a separate view. This interface will also control the start of the system after the user has validated all the parameters. These parameters are stored into XML files which are moved into the system when the processing starts. These parameters replace the default parameters. This interface indicates also the main directories or files used by the system: for example in which directory the output products will be available or where the user can find the log file.

3.1.3 Iterative product generation

For some multi-temporal output products, the processing could be done iteratively to integrate the useful information into the output product or into an intermediate product at each time step. For example the composite product should be generated iteratively to avoid to store all the L2A data of the composite window. With the same principle we will compute uniform time sampling data before the end of the considered period for the crop type or crop mask product.

With this approach we will try to minimize data storage and smooth the CPU and memory usage over the monitoring period. Moreover some masks about the time series quality (number of acquisition with snow for each pixel for example) can be computed iteratively.

3.1.4 Temporal resampling

Due to the different orbits of the multiple Sentinel-2 images requested to build a single Sen2-Agri product, all the pixels in this product may not have the same time sampling. Moreover the product pixels can be covered by cloud or cloud shadows and, as consequence, their values discarded which should generate gaps into the time series. Therefore each pixel of the input time series has its own time sampling.

For supervised classification algorithms, learning a model based on time series with gaps is complex and restricts the validity of the model. This is why we propose to fill the gaps and sample with a uniform step each pixel of the area of interest. With this approach a model learned on a set of samples can be applied to the whole area.

This strategy also allows to deal with the mixing between Landsat and Sentinel-2 data into the time series. For each common band, the interpolation at each time step will be made with the two closer values. The availability of Landsat data will increase the density of the time series and so reduce the errors made during the interpolation.

3.2 L3 composite product

The composite product will be generated by an iterative processing which can be split into tiles for parallelization purpose as described in Figure 3-2.

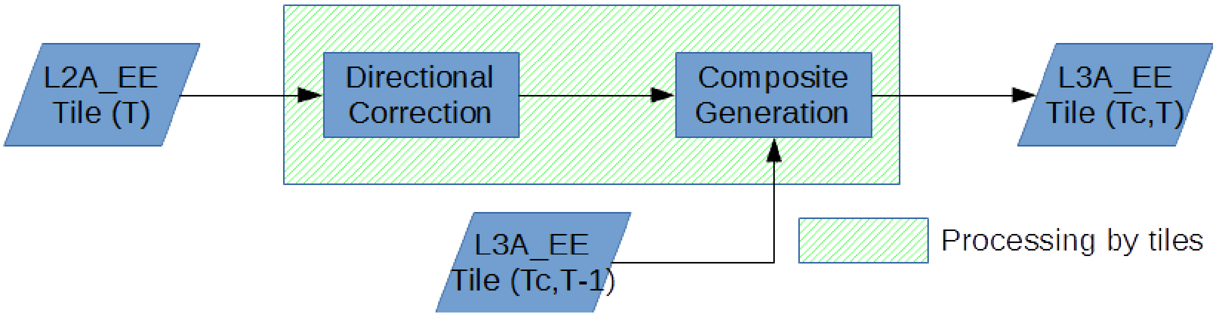

Figure 3-2: Logical data flow of the composite module.

This module will take as input each internal L2A tile product (image and mask at each resolution) available in the L2A directory at time *t*. It will take also the temporary output product generated at the previous date. This product will be updated with the new data until the composite window is complete. Each sub-part of the module will be described more precisely into the composite ATBD.

An additional step will be performed after this module: formatting the tile into a global product. This product will be formatted according to the PSD.

3.3 L4 crop mask product

The crop mask product will be generated by two methods according to the availability of auxiliary data. These auxiliary data could be either a reference layer or reference samples. The reference layer is a previous crop mask generated during the previous season while the reference samples are provided by in-situ measurement during the undergoing one. In this last case, we prefer use the learning model defined during of the former crop season because of the difficulty to get the in-situ data before the end of the current season.

For the first method, the algorithm will extract from the image series and a reference map, a set of samples which identifies the *Crop* and *No crop* classes. This method will use the same classifier than the second method. The classification step will be common between the two methods. The Figure 3-3 summarizes the logical flow of data between the two methods.

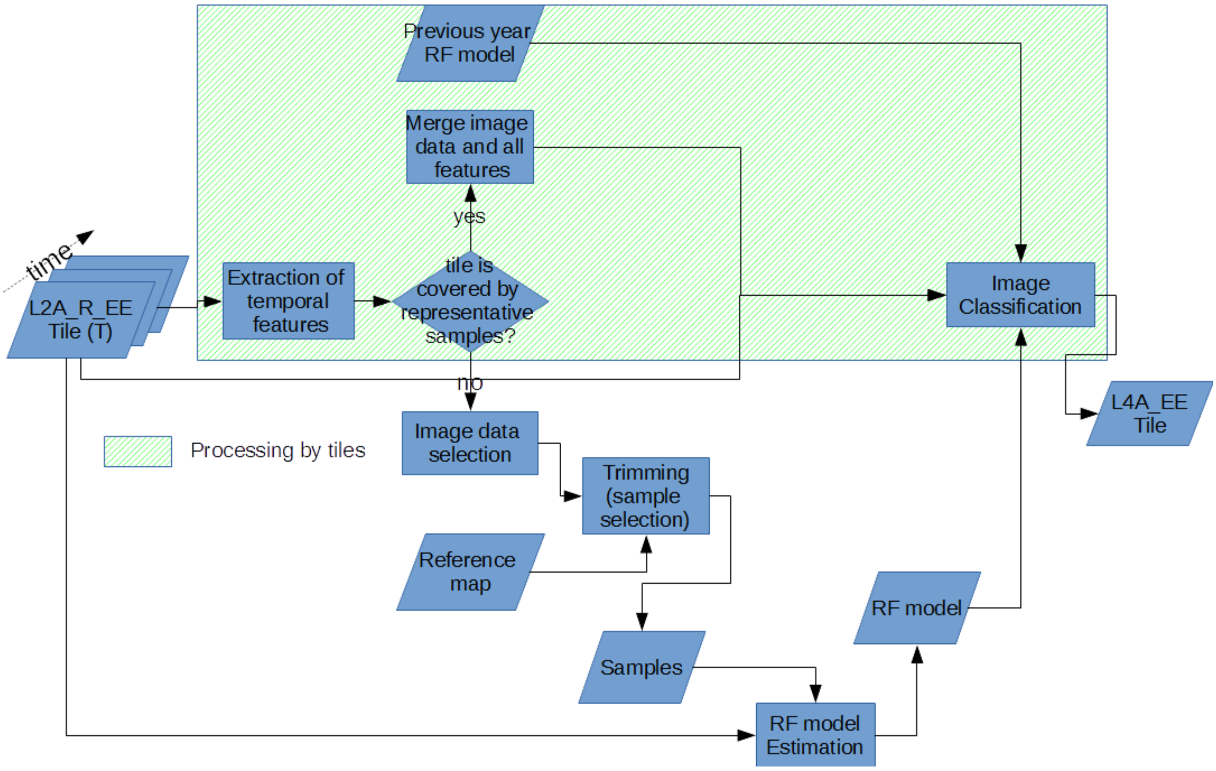

Figure 3-3 : Logical data flow of the composite module.

3.4 L4 crop type product

The crop type product will be generated into three main steps:

- The temporal resampling,
- The feature extraction,
- The classification as described in Figure 3-4,

We will try to turn the temporal resampling into an iterative scheme which is not described into the Figure 3-4 for the moment. This temporal resampling will be done for each band of each tile and so will be easily parallelized. The feature extraction will be made before each crop type estimation and currently not stored into the system because it will generate a large data volume and it is very fast to compute (TBC). The classification will be split into two sub-steps: model estimation limited to tiles covered by samples and image classification which applies the model to all the tiles.

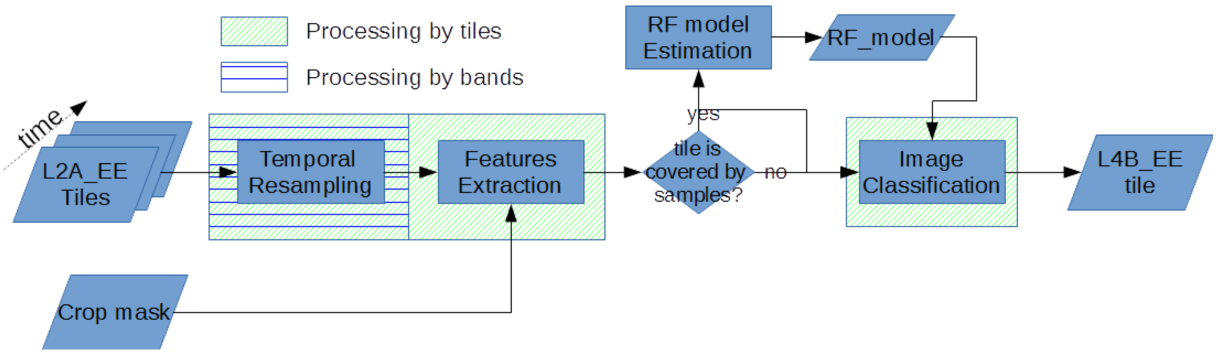

Figure 3-4: Logical data flow of the crop type module.

This module is restricted to tiles including by crops formerly identified within the crop mask. The crop mask will also be used to limit the area concerned by the computation of features and the classification. Each sub-part of the module will be described more precisely into the composite ATBD.

An additional step will be performed after this module: formatting the tile into a global product. This product will be formatted according to the PSD.

3.5 L3 biophysical product

The L3 biophysical product generation will be split into two processing chains:

- One for the multi-temporal LAI generation
- Another one for the phonology metrics.

The multi-temporal LAI will be parallelized by tiles due to the fact that LAI estimation depends on the acquisition conditions (solar and view angles). Moreover we propose to use temporal information derived from the models computed previously. These models compose a database of models by tiles. Each sub-part of the module will be described more precisely into the biophysical product ATBD.

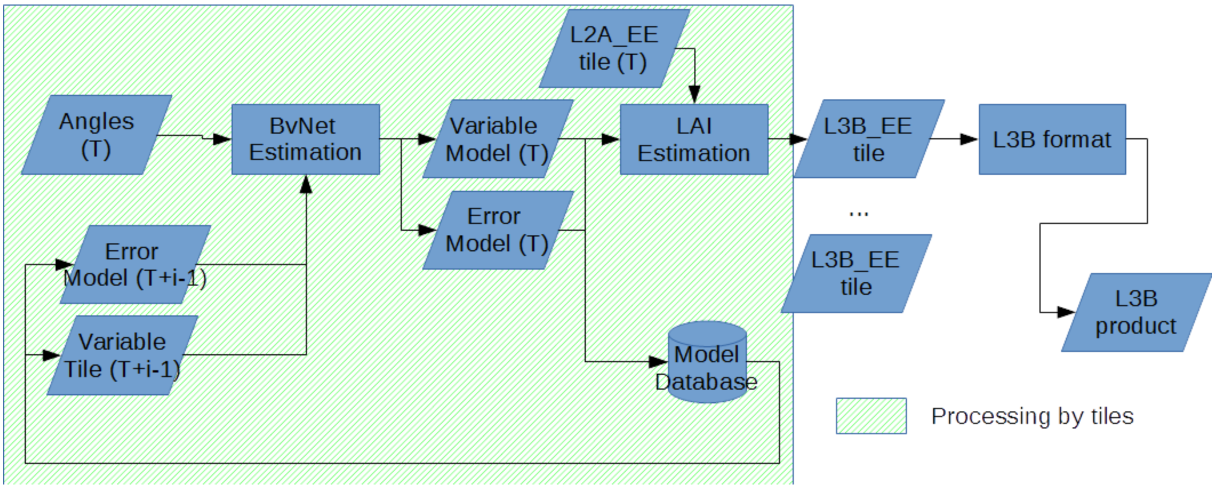

Figure 3-5 : Logical data flow of the biophysical indicator module.

Another product is expected at the end of the season to meet the user requirement about the need of metrics on the phenological status. The logical data flow of the module is described in Figure 3-6. All the sub-part of the module can be split by tiles and by dates to decrease the processing time. We can also compute the NDVI after each acquisition and store the result into the cold storage of the system. At the end of the season, the NDVI time series will be moved to the warm storage to be processed. As input of this module we should also consider the final crop mask because the metrics estimation can be done only on pixel with vegetation time profile.

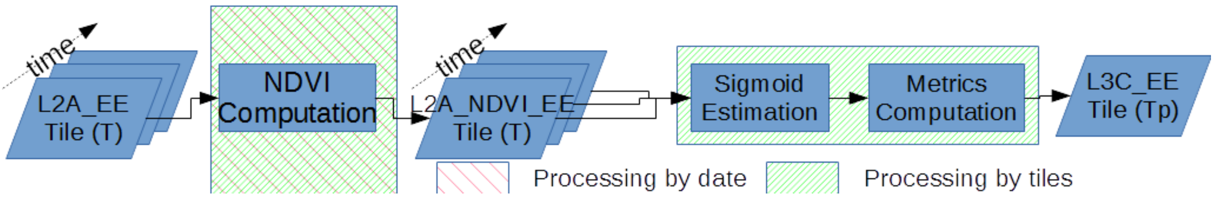

Figure 3-6 : Logical data flow of the NDVI metrics module

Each sub-part of the module will be described more precisely into the biophysical product ATBD.

4 Requirements coverage

This section will be completed after the PDR meeting after the discussion about the benchmarking.

4.1 Partially covered requirements

TBD

4.2 Not covered requirements

TBD

4.3 Not applicable requirements

TBD