DATA SCIENCE WORKFLOWS FOR THE CANDELA PROJECT

Mihai Datcu[1], Corneliu Octavian Dumitru[1], Gottfried Schwarz[1], Fabien Castel[2], and Jose Lorenzo[3]

[1]Remote Sensing Technology Institute, German Aerospace Center, Wessling 82234, Germany
(email: corneliu.dumitru@dlr.de; gottfried.schwarz@dlr.de; mihai.datcu@dlr.de)

[2]ATOS France SA, Les Espaces St Martin, 6 Impasse Alice Guy, 31024 Toulouse, France
(email: fabien.castel@atos.net)

[3]ATOS SPAIN SA, Calle Albarracín 25, 28037 Madrid, Spain (email: jose.lorenzo@atos.net)

ABSTRACT

This paper describes CANDELA - Copernicus Access Platform Intermediate Layers Small Scale Demonstrator - a general platform for the handling, analysis, and interpretation of Earth observation satellite images, mainly exploiting big data of the European Copernicus Programme. Its workflow allows the selection of satellite images, the generation of local image patch descriptors, the ingestion of image and descriptor data into a common database, the assignment of semantic content labels to image patches, and the search and retrieval of similar content-related image patches.

Index Terms**—**Data science, Earth observation, Copernicus data, data mining, data fusion.

1. INTRODUCTION

With the advent of the Copernicus Programme with its wealth of open data, the Earth observation (EO) application and service development domain is increasingly adopting big data technologies. This adoption is first related to efficient data storage and processing infrastructures, but most importantly to data analytics and application development concepts. Efficient data retrieval, mining augmented with machine learning techniques, and interoperability are key in order to fully benefit from the available assets, create more value and subsequently economic growth and development of the European member states [1-2].

CANDELA - Copernicus Access Platform Intermediate Layers Small Scale Demonstrator - aims at building a platform that delivers building blocks and services which enable users to quickly use, manipulate, explore and process Copernicus data. The main objective of CANDELA is to bridge the gap between big data technology and the EO data user community. While the objective is very ambitious, the pragmatic approach that we follow when building CANDELA makes it reachable. With the right blend between solid, operational existing assets and innovative tool integration, the platform shall help current and future Copernicus users to take a leap and profit from big data technology to maximize value creation.

In addition to an existing set of tools that our consortium already implemented for the platform [2], CANDELA will enable users to integrate already existing building blocks with a homogeneous, powerful and operational platform, opening up collaboration possibilities to research new approaches and offerings. This approach is highlighted by the development of scenarios such as urbanization, vineyard development, or forest disaster monitoring that will not only contribute as validation scenarios, but that constitutes real commercial, operational scenarios with existing customers.

The goal of one of CANDELA's work packages (WPs) is to develop the data science workflows and data analysis tools needed for implementing the functionality needed for the practical use cases of CANDELA. Each of the data science workflows will require configuration of data and optimization of the overall processes which will be performed through this task.

In this paper, we report about the data science workflows of the CANDELA project. The measured performance of the platform/system is beyond the scope of this paper.

2. DATA SET DESCRIPTION

Our main data sets extracted from different instruments are Earth surface images of the European Copernicus Programme, namely Sentinel-1 and Sentinel-2 images. While Sentinel-1 is an active twin satellite synthetic aperture radar (SAR) system, each of the Sentinel-2 twin satellites carries a passive optical multi-color imager. All instruments have been designed, calibrated, and are being operated by the European Space Agency (ESA) [3-4].

There are many reasons why we advocate the use of Sentinel-1 and Sentinel-2 images.

Firstly, we can recognize different target area details in overlapping radar and optical images complementing each other with rapid succession.

Secondly, individually selectable Sentinel-1 and Sentinel-2 images can be rectified and co-aligned by publicly available toolbox routines offered by ESA allowing a straightforward image comparison.

doi:10.2760/848593

Thirdly, all Sentinel instruments are well-documented, and typical data sets are already well understood within the remote sensing community. Many publications describe newly discovered Earth surface characteristics derived from the individual instruments.

Furthermore, the long-term operations of the Sentinel satellites allow the interpretation of image time series, or even the combination of time series data with external supplementary data via additional data mining/data fusion tools [5-7].

Besides these data sets, we include other 3rd party EO data sets as specified by CANDELA users (*e.g.*, TerraSAR-X, WordView, and Landsat).

3. DATA MINING AND DATA FUSION COMPONENTS OF THE CANDELA PLATFORM

The *CANDELA - Copernicus Access Platform Intermediate Layers Small Scale Demonstrator* - platform allows the prototyping of EO applications by applying interactive data mining and knowledge discovery functions to satellite images in order to select appropriate products in large data archives. It also helps to detect objects or structures, and to classify their land cover categories. The system allows ingesting Synthetic Aperture Radar (SAR) and multispectral images (*e.g.,* Sentinel-1, Sentinel-2, WorldView-2, TerraSAR-X). For other types of data, the main requirement is that the input data are provided in GeoTIFF format.

The main components of the CANDELA system are depicted in Figure 1 and are the following ones: Data Model Generation (DMG), DataBase Management System (DBMS), Image Search and Semantic Annotation, and Multi-Knowledge and Querying.

a) Data Model Generation

The **data mining** image ingestion can be seen as a processing chain, which, in our case, is managed by the Data Model Generation (DMG). This component is responsible for extracting the basic primitive features from the EO images (*e.g.,* SAR or multispectral images), generating tiles and their corresponding high-resolution visual quick-looks, and storing all the generated information into a database. The information is stored into an XML file. Finally, this file is automatically transformed into SQL statements and inserted into the DBMS.

It's important to mention here that for the DMG module, the input EO images should be GeoTIFF files (*e.g.*, Sentinel-1, TerraSAR-X), and their associated metadata XML files, widely used remote sensing data standards. An exception are the Sentinel-2 images, which are composed of several quadrants; here it is necessary to specify the folder of the quadrant-image that shall be processed.

Before starting the ingestion, it is necessary to verify that a freely available and thus popular MonetDB database installation is running. Once this has been verified, the next step is to select the **input images** and the **output location,** together with the size of the image patches and the **number of grid levels** (*e.g.,* 1, 2 or 3). The relation between the size of the image patches and their grid level is that, in case of grid level 1, an image is divided into patches with the specified size. If the grid level is 2, the same patch, from the previous step, is further divided into four patches with half of the previous size. Therefore, the number of patches of grid level 2 will by four times the number of patches of grid level 1. This procedure is repeated for grid level 3, etc.

In addition, for each generated image patch a visual quick-look file is created in JPEG format. In the case of SAR images, their brightness is adjusted to create this JPEG file. In the case of multispectral images, the RGB bands are used to create the JPEG quick-look file.

In case of Sentinel-2, since the data come in JPEG2000 format, an intermediary step is needed in order to convert the JPEG2000 format to GeoTIFF format necessary for the DMG input.

Within the ingestion, the **sensor type** can be chosen by the user from the following sensor types that correspond to the type of images one would like to ingest: **TSX** for TerraSAR-X data, **S1A** for Sentinel-1A/1B data, **S2A** for Sentinel-2A/2B data, or **OPT** for WorldView-2 data or other multispectral data (all in GeoTIFF format).

Depending on the type of input data and the envisaged application, different **feature extraction methods** can be applied. In the current version we implemented: **Gabor filters**, **Weber local descriptors**, and **histograms** [8]. The feature extraction methods are classified and used according to the input data. As for example, for **SAR images** one can use **Gabor filters** with the two options **Gabor linear moments** and **Gabor logarithmic cumulants** or **adaptive Weber local descriptors**, while for **multispectral images** one can use either **Weber local descriptors** or **multispectral histograms**. The size of the feature vector depends on the combination of the selected parameters (*e.g.*, for Gabor linear moments the mean and variance of 4 scales and 5 orientations gives us a feature vector of 2×4×5=40 parameters). Typically, data ingestion and patch tiling take together 1.5 ms per patch of 256×256 pixels, while feature extraction requires 2 ms per patch (for 40 parameters).

Inside the DMG module, there are a number of components that have been developed for **data fusion**. These fusion components are **data fusion ingestion**, **data fusion feature generation,** and **data fusion high-resolution quick-look**.

For each EO product within the data fusion component, there are a number of **features/descriptors** which can be selected by the user. For multispectral products (*e.g.,* Sentinel-2, WorldView) there are three feature algorithms being implemented, namely **multispectral histograms**, **Weber local descriptors**, and **Gabor linear moments** (computed for each band and concatenating the results). For SAR products, the same number of feature algorithms are implemented, namely **Gabor linear moments**, **Gabor log-cumulants**, and **adaptive Weber local descriptors**.

Based on the available features, the user can select one type of feature per sensor (one for multispectral, and one for SAR data) and after that, the features are fused and normalized before being inserted into the DataBase Management System (DBMS).

b) DataBase Management System

Here in the DBMS, the inputs from data ingestion and from the semantic annotation are stored into a relational database structure and act as the core of the system interacting with all components and supporting their functionality. The database is used also for different types of queries.

c) Image Search and Semantic Annotation

This module is used to search for image content and to create semantic annotations of the ingested images. Our implementation is based on the Cascaded Active Learning for Object Retrieval (CALOR) algorithm [9], which contains a Support Vector Machine (SVM) as our active learning and relevance-feedback method in order to allow the inclusion of human expertise in the annotation.

The definition of image semantics is achieved using an interactive loop where human expertise is required to define the most appropriate semantic category and to terminate the loop [10]. The employed CALOR algorithm is based on active learning methods. The idea behind active learning is that a machine learning algorithm can achieve higher accuracy with fewer training labels if it's allowed to choose the data from which it learns.

Another important component is **data fusion**, the generation of high-resolution quick-look data which includes the DMG from each sensor as well as the corresponding high-resolution quick-look image, and generates a single fused quick-look image needed for the image search and semantic annotation module. This module has been developed for semantic annotation of the image content by using machine learning algorithms and human interaction; another benefit of this module is that it can be used for fusion of different semantic labels. These fused semantic labels are saved into the database management system from where the user can run a query using the Multi-Knowledge and Querying module.

d) Multi-Knowledge and Querying

This module reads the required information from all tables in the database.

The query module is an interactive component, which allows the user to better exploit the EO products (*e.g.,* image and metadata). Based on these two types of data (image content and metadata), there are two different queries being available: query by metadata, and query by semantics. These queries can be also combined.

The first type of query is exploiting the entire metadata of each EO product by extracting and storing the information into the DBMS. Depending on the user needs, different metadata parameters can be combined during querying.

The second querying option is a query by semantics. In this case, the user can select one or several labels from a given list in order to perform the query. Please note that the semantic labels are generated *a priori* via the Image Search and Semantic Annotation module. Using this module the quality of the semantic annotation can be verified after the interactive part has been finished. To query new data (Sentinel-1 /-2), these data need to be annotated before.

4. DATA SCIENCE WORKFLOWS

In this section, we explain the functionality of the **data mining** and **data fusion** modules. First, we start by presenting in Figure 2 and 3 the data mining capabilities for two functions, namely data mining exploration and data mining semantic annotation. Similar workflows can be created for data fusion. Our example refers to a single sensor and can be extended to multi-sensor configurations.

After the first stage has been completed (see Figure 2) and the user has a first idea of the content of the data set, now he/she can go ahead and can assign semantic labels to each retrieved category (see Figure 3).

Based on the output of these schemes (Figures 2 and 3), two possible scenarios can be imagined in close connection with known CANDELA use cases. The first one is a data mining query together with data analytics, and the second one is semantic sensor fusion.

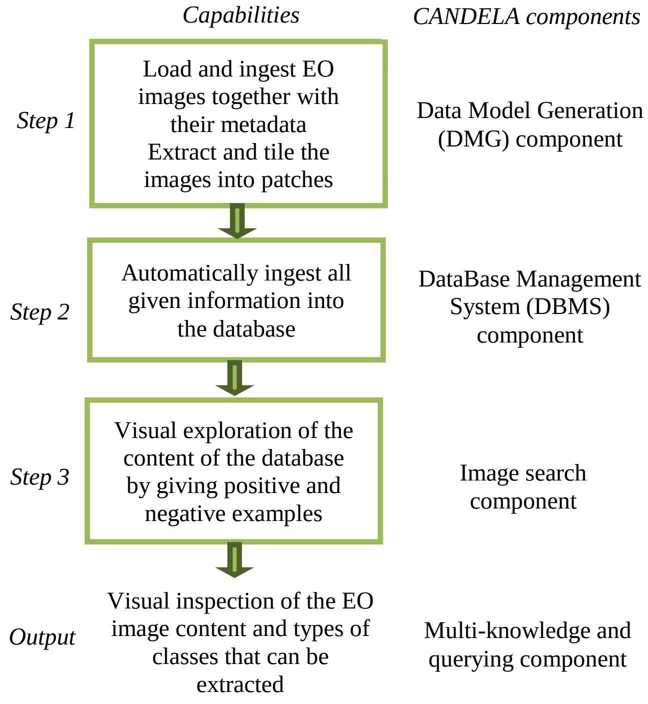
Capabilities
CANDELA components
Step 1
Load and ingest EO images together with their metadata Extract and tile the images into patches
Data Model Generation (DMG) component
Step 2
Automatically ingest all given information into the database
DataBase Management System (DBMS) component
Step 3
Visual exploration of the content of the database by giving positive and negative examples
Image search component
Output
Visual inspection of the EO image content and types of classes that can be extracted
Multi-knowledge and querying component

Figure 2. Data mining exploration.

5. DISCUSSIONS

After the final testing of our system has been completed, we will select a dataset for which the output is known and we will try to find similar systems, if such systems exist, to compare the results.

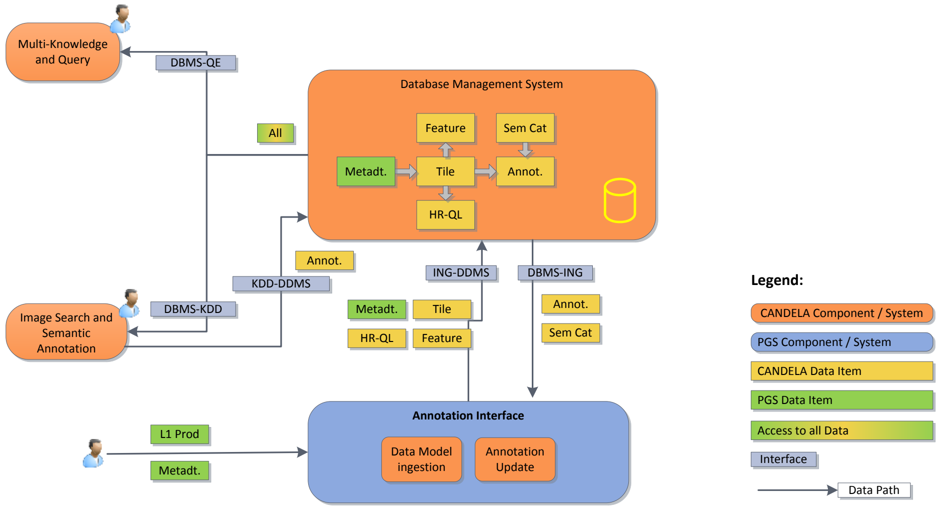

Figure 1. Components of the CANDELA - Copernicus Access Platform Intermediate Layers Small Scale Demonstrator - platform.

In addition, we are very much interested in comparable approaches implemented by other institutions.

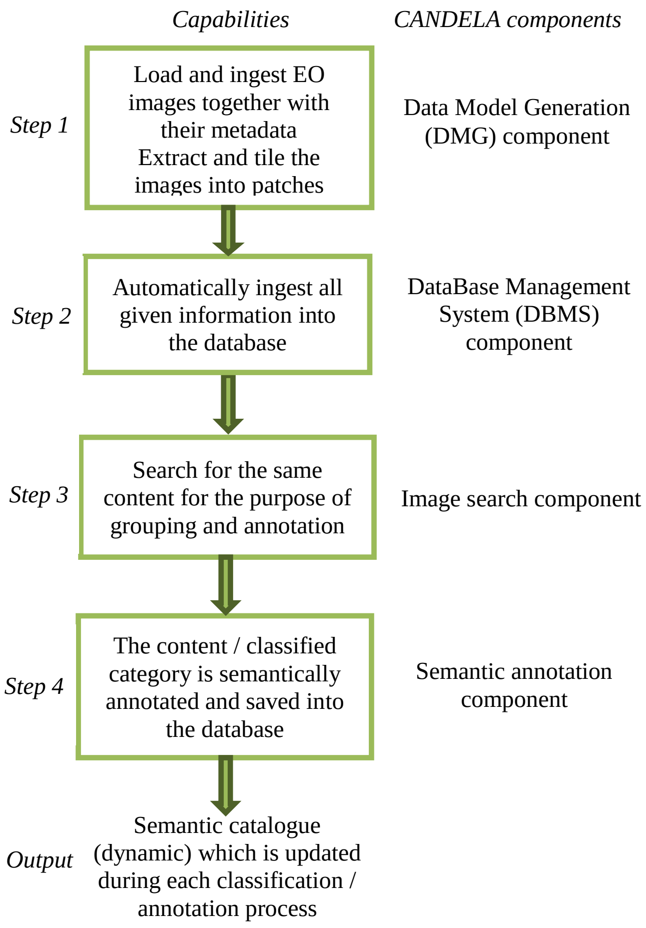
Capabilities
CANDELA components
Step 1
Load and ingest EO images together with their metadata
Extract and tile the images into patches
Data Model Generation (DMG) component
Step 2
Automatically ingest all given information into the database
DataBase Management System (DBMS) component
Step 3
Search for the same content for the purpose of grouping and annotation
Image search component
Step 4
The content / classified category is semantically annotated and saved into the database
Semantic annotation component
Output
Semantic catalogue (dynamic) which is updated during each classification / annotation process

Figure 3. Data mining semantic annotation.

6. ACKNOWLEDGEMENTS

This work was supported by the CANDELA - Copernicus Access Platform Intermediate Layers Small Scale Demonstrator H2020 research and innovation project under grant agreement No. 776193.

7. REFERENCES

[1] CANDELA: Copernicus Access Platform Intermediate Layers Small Scale Demonstrator H2020 proposal, 2017.

[2] CANDELA project, 2018. [Online]. Available: http://www.candela-h2020.eu/.

[3] ESA Sentinel-1, 2018. [Online]. Available: https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-1.

[4] ESA Sentinel-2, 2018. [Online]. Available: https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-2.

[5] Living Planet Symposium, 2016. [Online]. Available: http://lps16.esa.int/.

[6] Big Data from Space Conference, 2017. [Online]. Available: https://earth.esa.int/web/guest/events/all-events/-/article/conference-on-big-data-from-space-bids-17.

[7] IGARSS, 2018. [Online]. Available: https://www.igarss2018.org/Tutorials.asp#FD-6.

[8] D. Espinoza-Molina, V. Manilici, S. Cui, Ch. Reck, M. Hofmann, C.O. Dumitru, G. Schwarz, H. Rotzoll, and M. Datcu, *"Data Mining and Knowledge Discovery for the TerraSAR-X Payload Ground Segment"*, PV 2015, Darmstadt, Germany, 2015.

[9] P. Blanchart, M. Ferecatu, S. Cui and M. Datcu, *"Pattern retrieval in large image databases using multiscale coarse-to-fine cascaded active learning,"*Selected Topics in Applied Earth Observations and Remote Sensing, vol. 7, no. 4, pp. 1127-1141, 2014.

[10] C. Dumitru, G. Schwarz and M. Datcu, *"SAR Land Cover Datasets for Benchmarking,"* IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 11, no. 5, pp. 1571-1592, 2018.