Semantically Enriched Crop Type Classification and Linked Earth Observation Data to Support the Common Agricultural Policy Monitoring

Maria Rousi, Vasileios Sitokonstantinou, Georgios Meditskos , Ioannis Papoutsis , Ilias Gialampoukidis , Alkiviadis Koukos, Vassilia Karathanassi, Thanassis Drivas, Stefanos Vrochidis , Charalampos Kontoes, and Ioannis Kompatsiaris , *Senior Member, IEEE*

Abstract—During the last decades, massive amounts of satellite images are becoming available that can be enriched with semantic annotations for the creation of value-added earth observation products. One challenge is to extract knowledge from the raw satellite data in an automated way and to effectively manage the extracted information in a semantic way, to allow fast and accurate decisions of spatiotemporal nature in a real operational scenario. In this work, we present a framework that combines supervised learning for crop type classification on satellite imagery time-series with semantic web and linked data technologies to assist in the implementation of rule sets by the European common agricultural policy (CAP). The framework collects georeferenced data that are available online and satellite images from the Sentinel-2 mission. We analyze image time-series that cover the entire cultivation period and link each parcel with a specific crop. On top of that, we introduce a semantic layer to facilitate a knowledge-driven management of the available information, capitalizing on ontologies for knowledge representation and semantic rules, to identify possible farmers noncompliance according to the Greening 1 (crop diversification) and SMR 1 rule (protection of waters against pollution caused by nitrates) rules of the CAP. Experiments show the effectiveness of the proposed integrated approach in three different scenarios for crop type monitoring and consistency checking for noncompliance to the CAP rules: the smart sampling of on-the-spot checks; the automatic detection of CAP's Greening 1 rule; and the automatic detection of susceptible parcels according to the CAP's SMR 1 rule.

Index Terms—Crop type classification, data fusion for decision-making, European Union (EU) common agricultural policy (CAP)

Maria Rousi, Georgios Meditskos, Ilias Gialampoukidis, Stefanos Vrochidis, and Ioannis Kompatsiaris are with the Centre for Research and Technology Hellas, Information Technologies Institute, 57001 Thermi-Thessaloniki, Greece (e-mail: mariarousi@iti.gr; gmeditsk@iti.gr; heliasgj@iti.gr; stefanos@iti.gr; ikom@iti.gr).

Vasileios Sitokonstantinou is with the National Observatory of Athens, 118 10 Athens, Greece, and also with the Laboratory of Remote Sensing, National Technical University of Athens, 15790 Zografou, Greece (e-mail: vsito@noa.gr).

Ioannis Papoutsis, Alkiviadis Koukos, Thanassis Drivas, and Charalampos Kontoes are with the National Observatory of Athens, 118 10 Athens, Greece (e-mail: ipapoutsis@noa.gr; akoukos@noa.gr; tdrivas@noa.gr; kontoes@noa.gr).

Vassilia Karathanassi is with the Laboratory of Remote Sensing, National Technical University of Athens, 15790 Zografou, Greece (e-mail: karathan@survey.ntua.gr).

noncompliance checking, linking earth observation (EO) data and web content, semantic enrichment.

I. INTRODUCTION

IN RECENT years, a massive quantity of georeferenced data is generated from many different sources like human activity and earth observation (EO), *in situ* sensors, satellite missions (e.g., Copernicus), and mobile phones. The semantic enrichment and linking of these free and open data of this scale, frequency, and quality constitute a fundamental challenge for interoperability and automation in decision-making. EO data become useful only when analyzed together with other sources of data (e.g., geospatial data, *in situ* data) and turned into actionable information and knowledge for decision-making. In this context, linked data[1] is a data paradigm that studies how one can make resource description framework (RDF) [1], [2] data available on the web and interconnect it with other data with the aim to increase its value. In the last few years, linked geospatial data has received attention as researchers have started tapping the wealth of geospatial information available on the web using semantic web technologies [3], [4]. Nevertheless, there are only a handful of applications that showcase the semantic integration of linked EO and non-EO products. The scalability to accommodate big linked EO data also remains an open issue [5].

One of the domains that is already heavily dependent on the effective and efficient knowledge extraction from EO data is the control of the common agricultural policy (CAP) [6]. The European Union (EU), through the CAP, aims at increasing the European agricultural productivity under sustainable practices while at the same time making sure that the farmers maintain a decent standard of living.[2] It is the EU's aim to reinforce the competitiveness of European agriculture, whilst maintaining and strengthening its sustainability. This manifests as a major priority, with CAP's annual budget amounting to approximately 59 billion Euros. The Integrated Administration and Control System (IACS) of the CAP consumes the majority of its annual budget. The IACS functions as the management system for the CAP payments, and is implemented by the national

---

[1][Online]. Available: https://www.w3.org/standards/semanticweb/data
[2][Online]. Available: http://esa-sen4cap.org/

paying agency of each EU member state (MS).[3] The CAP legal framework is transitioning to its new form, the post-2020 CAP reform, which aims to modernize and simplify the current operating model.[4] Based on the post-2020 CAP ambitions and toward the direction of the so-called monitoring approach for the implementation of IACS, EO has been identified as a key enabler.

Multiple EC funded projects have employed Copernicus data, using advanced ICT and artificial intelligence technologies, to address the monitoring of the CAP. The RECAP project[5] has been one of the first to develop Copernicus based machine learning pipelines to assist the paying agencies in reducing the costs and increasing the efficiency of the control of CAP's Cross-Compliance. Additionally, the Sen4CAP project,[6] building on the legacy of RECAP, has focused on reducing the costs of IACS toward the post-2020 CAP objectives; exploring the applicability of an evidence-based monitoring approach. To the best of our knowledge, none of the existing approaches are able to support the following CAP scenarios to deliver true business value. The selected scenarios require significant human resources, with on-the-spot checks (OTSCs) on only a small sample of the farmers' applications.

A. Smart Sampling of OTSCs

Farmers declare the cultivated crop type for their arable land each year, around the months of May and June. Paying agencies are responsible for validating the declarations to then grant the requested subsidy to the farmer. The MSs randomly sample and inspect 1%–5% of the total number of declarations. It is necessary to automatically monitor the farmers' declarations, using and linking additional and online available data, in order to create a targeted, instead of random, sample set that requires OTSC.

B. Automatic Detection of CAP's Greening 1 Rule

This rule seeks to improve biodiversity and reduce soil erosion by imposing limits in the size and number of the different cultivations in a farm. Specifically, the farmers that own 10–30 ha of arable farm should grow at least two different crop types, whereas farmers that own more than 30 ha should grow at least three different crop types. In the first case, the main crop should not cover more than 75% of the land, whereas for the latter case the two main crops should additionally not exceed 95% of the total land.[7] This rule requires accurate and semantically enriched geospatial data, so as to both detect correctly the crop types and perform semantic reasoning to infer the consistency between the farmers' declaration and the Greening 1 rule.

C. Automatic Detection of Susceptible Parcels According to CAP's SMR 1 Rule

This particular CAP requirement expects from the farmers, among other things, to perform a risk assessment on the susceptibility of their parcel to contribute nitrate-rich soil to nearby surface waters. The farmer should account for the slope of land, the ground cover, the proximity to surface water, weather conditions, soil type and conditions and the presence of land drains.[8] The SMR 1 requirement defines buffer zones, which shall be respected in terms of fertilizer spreading. Specifically, 1) manufactured fertilizer spreading should be at least 2 m from surface water and 2) organic manure spreading should be at least 10 m from surface water.[9] Therefore, measuring the proximity of the parcel boundaries to the nearest surface water is of great significance both for the inspections of the paying agency, but also for the farmer who wishes to better comply with the requirement. To that end, the semantic data fusion of Sentinel images and other linked open geospatial data would lead to a more efficient monitoring of SMR 1.

In this work, we adopt and extend the linked open EO data life cycle paradigm [3], and provide for the first time an end-to-end implementation to address the operational needs of the CAP. Motivated by the above-mentioned use cases and their inherent requirements to effectively and intelligently combine and interlink various geospatial data sources (e.g., land parcel identification system (LPIS), hydrographic network, Natura2000 zones, etc.), we instantiate the linked open EO data life cycle in the domain of CAP. To this end, we propose a hybrid data- and knowledge-driven framework, developing concrete CAP-related scenarios and demonstrate automatic pipelines for satellite data processing, content extraction, semantic annotation and transformation to RDF, interlinking layer, validation and querying of linked open spatiotemporal data.

Our contributions are summarized as follows.
1) In the context of CAP monitoring, to check compliance with Greening-1 requirements, taking into account satellite-derived products and ancillary geospatial data.
2) We demonstrate the use of spatial relationships in LOD (GeoSPARQL) toward assessing vulnerable parcels according to CAP SMR-1 specifications.
3) We propose the smart sampling scheme, i.e., the use of spatiotemporal queries to define a new, educated sampling of parcels that need to be checked for compliance with CAP rules with in-field visits.
4) We evaluate our framework under the light of national scale application, in line with post-2020 CAP monitoring needs. Therefore, we discuss scalability implications for both the knowledge extraction from satellite imagery module and the semantic reasoning framework.

The rest of the article is structured as follows. Section II presents related technologies considering both the EO-based

---

[3][Online]. Available: https://ec.europa.eu/info/food-farming-fisheries/key-policies/common-agricultural-policy/financing-cap/controls-and-transparency/managing-payments_en

[4][Online]. Available: https://www.consilium.europa.eu/el/policies/cap-future-2020/

[5][Online]. Available: https://www.recap-h2020.eu/

[6][Online]. Available: http://esa-sen4cap.org/

[7][Online]. Available: https://ec.europa.eu/info/food-farming-fisheries/key-policies/common-agricultural-policy/income-support/greening_en

[8]GCCE 2017 v1.0, "The guide to cross compliance in England 2017," produced by the Department for Environment, Food and Rural Affairs.

[9][Online]. Available: https://www.gov.uk/guidance/using-nitrogen-fertilisers-in-nitrate-vulnerable-zones
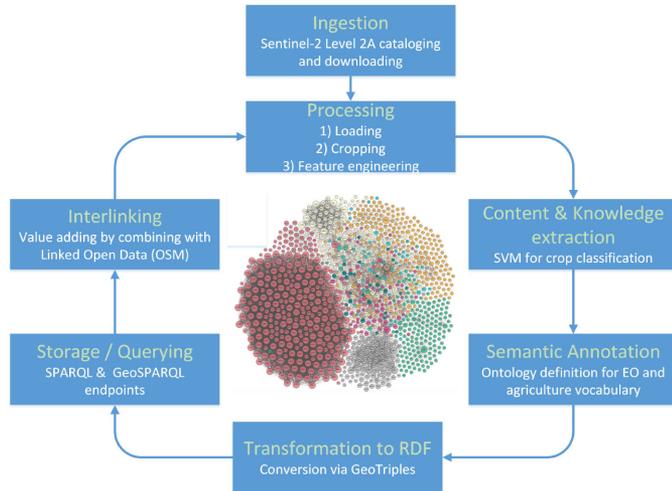
Fig. 1. Main stages of the complete life cycle of linked open EO data for the monitoring of the CAP.

CAP monitoring and semantic web technologies. The technologies are grouped under a common framework, aiming to cover the whole life cycle of the linked EO data for the control of the CAP. Section III describes the proposed semantically enriched crop type classification model for checking the compliance of farmers' declarations to the CAP regulations. Section IV presents experiments and results regarding our proposed methodology, along with the area of interest (AOI), the considered CAP scenarios, their implementations and results regarding the effectiveness and efficiency of our pipeline. Finally, Section V concludes this article.

II. LIFE CYCLE OF LINKED EO DATA FOR THE CONTROL OF THE CAP

Existing works for the monitoring of CAP mainly focus on knowledge extraction technologies, whereas the semantic technologies cover more generic agricultural needs. However, our main objective is to cover the complete life cycle of linked open EO data paradigm, as originally discussed in [3], following a multidisciplinary approach. The stages of the life cycle of linked EO data for the control of the CAP are presented in Fig. 1. The first step covers the content extraction machine learning methodologies so as to get new information layers out of the large streams of raw satellite data. The second step involves the standardized data representation and ontological modeling for semantic annotation. The semantic annotation is based on semantic web technologies, which are being adapted and developed under the EO and agricultural domains. The next step of the life cycle regards the transformation of the extracted content into RDF, allowing the population of the knowledge base (triplestore) to perform semantic queries that offer a better knowledge of the data (storage/querying). Applying useful interconnections in the semantic data using external datasets can additionally enrich the content and extract hidden knowledge (interlinking).

A. Content and Knowledge Extraction: Crop Classification

Over the last decades there have been multiple studies that have utilized EO data to extract high-level thematic knowledge for the agricultural land. Recently and since the introduction of the Sentinel missions, there have been a plethora of scientific publications that have exploited either Sentinel-1 or Sentinel-2 imagery, or in certain cases both, to classify crop types. The high spatial and temporal resolution of the Sentinel missions make them ideal for constructing dense image time-series of high quality that capture all the phenological stages of the different crop types and thus allowing for their accurate discrimination.

The state of the art in EO-based and specifically the Sentinel-based crop classification has advanced significantly over the last years, with the majority of publications reaching optimal accuracy levels (>85%) for multiclass problems. The relative differences in the published approaches are based on the nature and level of specificity of the investigated crop classes, the computational complexity restrictions, the scale of application and the ground truth information that is available for training and validation.

In order to reduce the computational complexity of crop classification and develop scalable solutions, multiple studies have followed object-based image analysis approaches. For instance, Lebourgeois *et al.* [7] have segmented their image stack into objects using spectral segmentation techniques on Very High Resolution imagery, whereas Sitokonstantinou *et al.* [8] made use of the LPIS to partition their feature space into parcel objects.

Synthetic aperture radar and optical imagery, retrieved from Sentinel-1 and Sentinel-2 missions, respectively, have been used either individually or combined. In [9], a combination of both Sentinel-1 and Sentinel-2 is used in order to create very dense time-series, thus alleviating the cloud coverage limitations. In [10], Arias *et al.* employ solely Sentinel-1 data, suggesting a weather-independent crop classification scheme for the monitoring of the CAP, hence accounting for northern European countries that suffer from year-round cloud coverage. Other studies focus on generating multiple diverse features from Sentinel imagery, beyond the most common spectral bands and vegetation indices (VI). Feng *et al.* [11] and Akbari *et al.* [12] create deep feature spaces, additionally including variations of VI, texture and phenology parameters. Such methods are shown to be particularly useful in classifying spectrally heterogeneous crop classes, i.e., vegetables.

With respect to the classification methods employed, both supervised and semisupervised learning approaches can be found in the literature. In [13], for example, Tatiana Solano-Correa *et al.* have combined a hierarchical correlation clustering with an artificial neural network. The vast majority of studies, however, make use of supervised learners, such as support vector machines (SVM) and random forest (RF) ([8], [7], [11], [12], [9], [14], [15]). Their effectiveness stems from their ability to accurately describe the nonlinear relationships between crops' physical condition and their spectral characteristics while being particularly insensitive to noise and overfitting. Finally, there are important studies that have used convolutional neural networks or recurrent neural networks or a combination of both [16],

which allow the learning of time and space correlation over the Sentinel time-series, thus reducing manual feature engineering.

In [8], Sitokonstantinou *et al.* have developed a scalable crop identification scheme, employing a second-order polynomial SVM on a time-series of Sentinel-2 data. Sitokonstantinou *et al.* [8] have additionally performed an extensive comparison between SVM and RF, them being the most widely used classifiers for crop mapping problems. The results showcased the superiority of SVM over RF for the classification of multiple and spectrally similar classes. This conclusion is additionally supported by Zhang *et al.* [17].

This study builds on the methods and results that were published in [8], following the state-of-the-art in crop classification as described earlier. The crop classification method of this study, however, has been applied to three different areas of interest, of diverse characteristics, thus proving its transferability. Finally, we perform multiple crop classifications, starting from very early in the year; therefore with truncated feature spaces. Nonetheless, the results, even early in the year, are satisfactory for the purposes of smart sampling the CAP OTSCs.

B. Semantic Web Technologies

Web Ontology Language (OWL) [18] is an ontology language that provides classes, properties, and individuals under the semantic web aspect. Ontologies offer the taxonomy of semantic objects and the relationship between them. RDF [19], [20] is the W3C recommendation standard that offers data representation under subjectpredicateobject standard, which is known as triples. Each subject is a resource and each object can be either a resource, a value, or an empty node. Predicates or properties express the relationship between a specific subject and object. Data expressed in this format are saved into RDF triplestores, named Knowledge Graphs.

SPARQL Protocol and RDF Query Language (SPARQL) [21] is the most popular querying language for the retrieval and manipulation of data in RDF format. SPARQL offers a wide range of query forms and operators to access and retrieve the data. stSPARQL [22],[23] is a SPARQL version that applies semantic queries into data in stRDF format. Such formats offer representation and querying of thematic and spatial data, which contain a temporal dimension. GeoSPARQL [23] focuses more on geospatial data querying, by providing a wide list of functions to support semantic queries execution over geometry and feature objects. Topological relationships are also taken into account.

Reasoning [24] is the procedure of inferring logical consequences based on asserted facts or axioms. In RDF graphs, reasoning takes advantage of data triples using in many cases different data sources to specify the rules that can lead to useful knowledge extraction. Reasoning with rules is usually based on description logics (DLs). DLs [25] constitute a family of logic-based representation formalisms and are usually used to represent well-structured knowledge over the application domain. Its name comes from a combination of descriptions, which are the expressions namely predicates, and the fact that they support logic-based semantics. DLs are strongly associated with

structuring ontology languages such as OWL, but are also widely used in application domain.

The vision of linked data is associated with the transformation of data into RDF formats. Data in this format can be published on the web and linked with other existing data that come from different sources [20]. Linked data are easily accessible using semantic queries. The main advantage of semantics is that they have the means to create intelligent interconnections over objects that come from heterogeneous sources as they support better information management, complexity limitation, and useful inferences extraction [26]. In this work, we provide a list of functions and semantic queries, using semantic web technologies, to support three CAP-related scenarios in real operational problems, that require content extraction and semantic linking of data for compliance checking.

C. Semantic Annotation Under the EO and Agriculture Domains

Building appropriate ontologies to describe the different aspects of EO and agriculture are presented in this section. EO ontologies focus more on the environmental monitoring domain.

The ontology presented in [27] deals with hydrological monitoring issues and captures the main components of hydrological monitoring, which are the events, the sensors, and the observations. Sensors and observations are divided into many subcategories such as physical and meteorological, whereas events are associated with any hydrological cycle change. Modular environmental monitoring ontology [28] extends the abovementioned ontology as, except from sensor and observation data, it provides a structure to model a plethora of different aspects that are identified on an emergency situation under the environmental monitoring domain. The ontology provides the structures to represent environmental features (procedure and material), physical conditions (disaster), and spatiotemporal information (geolocation and time).

An agricultural ontology representation method is described in [29]. The suggested model contains information such as cultivation and processing practices, storage, pests control, genetic attributes, etc. OntoCrop ontology [30] offers knowledge representation for common cultivation practices, pests control and in general the crops physiology. Each plant is characterized by properties such as name, growth stage, type of infection, infected part, information about disorders. Agriculture ontology for the purpose of agriculture Internet of Things [31] presents a more product-oriented view of agricultural products containing information related to the product, the seeding procedures, the physical conditions, the phase, the location, and temporal dimensions. The main purpose of this ontology is to support healthy food management. The ontology presented in [32] offers a uniform representation of text classification and concept extraction results. The ontology matches specific concepts into ontology classes, which include many different types of products such as agricultural, planting, livestock, fishery and agricultural material.

Most works that have been previously mentioned describe either EO or agricultural data. What is actually missing is a

combination of both describing information from EO with information from the agricultural domain. In this work, we reuse and extend the ontology found in [8] that describes the relationship among crop types, families and season, and create a combination with GeoSPARQL vocabulary that represents geospatial-related data such as points in polygon geometry.

D. Transformation Into RDF

Another widely investigated issue is combining semantics with EO data to discover hidden knowledge. This section describes some frameworks that deal with data transformation into semantic format, integration and searching.

Intelligent interactive image knowledge retrieval [33] is a framework that utilizes EO data archives and applies image segmentation (PCA kernel approach) and classification techniques (SVM learning method). From the semantic aspect, the system achieves high-level query processing into context information from distributed data archives. Domain-specific ontologies provide the appropriate structures to integrate heterogeneous data sources in order to support complex semantic queries. A hybrid ontology approach has been used to integrate data coming from different ontologies. Semantic restrictions have been applied using DL reasoning to determine the conditions under which an instance will belong to a class.

GeoTriples [23] is a tool that deals with the geospatial data transformation into semantic RDF format. The system gets as an input a file in various formats and creates a mapping that is based on GeoSPARQL vocabulary, using RML and R2RML rules (mapping generator). Users have the opportunity to define the rules if needed. Initial data are transformed into RDF graph format using the RML rules defined in previous phase (mapping processor). Various RDF syntax formats are supported. Querying is also available in a relational database using R2RML mapping (stSPARQL/GeoSPARQL evaluator).

In this study, we use the GeoTriples tool as a basis in order to transform shapefile data into RDF format under the GeoSPARQL standard for semantic representation.

E. Storage and Querying

In this work, we choose to handle the three different CAP scenarios using semantic technologies. The problem could have been solved using relational databases, though this selection would be accompanied with an inflexible data schema and higher execution times [34]. Additionally, the relationships between the entities handled in this work are quite complex to be represented using SQL keys.[10] Information is coming from three layers and, with the usage of semantics, are combined in the most effective way, whereas OWL 2 RL rules are used to enrich the data [35].

RDF triplestores are semantic databases that offer data saving in semantic graph format. Strabon [36] is a geospatialoriented RDF triplestore that offers a broad amount of querying functions over georeferenced information, supporting both stSPARQL and GeoSPARQL. GraphDB[11] [37] is also a popular triplestore that supports saving and querying over georeferenced and nongeoreferenced semantic data, supporting native OWL 2 reasoning. It is considered as one of the best triplestores available in terms of storage, supported functionalities, performance, and execution time [38]. Other RDF triple stores that provide geospatial support include RDF4J,[12] Virtuoso[13] [39], OntopSpatial[14] [40], Oracle spatial and Graph,[15] AllegroGraph,[16] Stardog,[17] uSeekM,[18] and Parliament.[19]

The storage and query capabilities of our framework capitalize on an existing RDF triple store, on top of which SPARQL and GeoSPARQL standards are used to form the queries that support the rules of agriculture policies. The current implementation uses the GraphDB semantic graph database, taking full advantage of the provided dashboard to explore and manage the RDF repositories. It also supports different reasoning profiles, such as OWL 2 Rule (RL) reasoning, allowing us to use off-the-shelf reasoning on top of our domain ontology. It is worth mentioning, however, that since our framework capitalizes on existing, well-known standards (RDF, OWL, SPARQL, etc.), it is interoperable and it does not depend on specific implementations. For example, it requires minor updates to migrate to different triple stores, according to the application requirements, such as Strabon and AllegroGraph, or to use different SPARQL query engines.

F. Interlinking

To exploit the wealth of data, there comes the need of generating intelligent interconnections between different datasets. In the literature, many systems have been implemented dealing with this issue but due to the vast heterogeneity of data, using existing systems into new datasets does not work in most cases. Interlinking is achieved based on geospatial data characteristics in some cases [4], whereas in others specific mechanisms have been developed to meet the needs of the data [41].

The system that is presented in [4] receives data from heterogeneous sources such as meteorological, health and EO. It creates an appropriate RDF representation and associations between specific characteristics. Data interlinking is achieved by calculating the similarity between different datasets. In [41], a system that integrates EO data to support data management is introduced. The system receives data from different data sources and has two different functionalities. For data that are related with China multiple components have been developed to adapt in different interfaces, whereas for international data the GEO DAB agent is used. PREDICAT [28] is a system that focuses on natural catastrophes prediction. PREDICAT uses different ontologies to semantically represent the data that are pertinent to the system (semantic layer). The system overcomes data heterogeneity and provides a common structure of interconnected objects

---

[10][Online]. Available: https://www.sqlshack.com/understanding-benefits-of-graph-databases-over-relational-databases-through-self-joins-in-sql-server/
[11][Online]. Available: http://graphdb.ontotext.com/

[12][Online]. Available: https://rdf4j.org/
[13][Online]. Available: https://virtuoso.openlinksw.com/
[14][Online]. Available: http://ontop-spatial.di.uoa.gr/
[15][Online]. Available: https://www.oracle.com/database/technologies/spatialandgraph.html
[16][Online]. Available: https://allegrograph.com/
[17][Online]. Available: https://www.stardog.com/
[18][Online]. Available: https://www.openhub.net/p/useekm
[19][Online]. Available: https://github.com/SemWebCentral/parliament

containing spatiotemporal information (data integration layer), whereas a reasoner and a decision-maker are implemented to provide the appropriate responses to the user (data processing layer). In CANDELA project[20] semantic search is supported on EO images and other associated metadata, using existing technologies, namely GeoSPARQL, OWL-Time, SOSA, DCAT, and PROV-O. These technologies support a monitoring use case in agriculture, where the impact of a storm on vineyards is measured by first extracting knowledge from Sentinel images and then semantically fusing them with weather reports in the same AOI.[21] However, the semantic search targets mainly to offer a mature solution for insurance companies, whereas it is not straightforward to monitor CAP-related regulations that require to be checked in terms of the farmers' compliance to CAP rules.

G. Complete Life Cycle for CAP Monitoring Using Semantic Technologies and Linked Open EO Data

Despite the fact that a lot of progress has been achieved in different aspects of the individual components mentioned above (e.g., semantic annotation has been implemented under the environmental monitoring domain [27], [28], [42], data integration has been widely investigated (interlinking) [28], [33], [43]–[46]), not much effort has been achieved in implementing a system that supports the *complete life cycle*. The major challenge, which we address in this article, is mostly related with interlinking phase and dealing with the heterogeneity of data (e.g., sensors [46], aerial and satellite imagery [43], OpenStreetMap data [43], [45], Google Earth imagery [45]) and defining the ways to exploit these data and enhance knowledge discovery [33].

Contrary to the presented approaches in this section, in this work we focus on the linked open EO data life cycle paradigm proposed in [3], aiming to support impactful use cases in CAP. To the best of our knowledge, this is the first attempt to reuse and adapt the proposed architecture to the domain of CAP monitoring. The proposed framework implements a hybrid scheme of data analysis and annotation: the results of a data-driven crop classification framework are semantically annotated and interlinked in order to foster advanced interpretation, such as improving classification accuracy through domain knowledge, and querying solutions. We demonstrate the added value and feasibility of our approach in a number of challenging use cases in CAP monitoring.

III. METHODOLOGY

The overall framework of our proposed methodology is presented in Fig. 2. The layers are the image analysis layer, the mapping layer, the data ingestion and reasoning layer and the query processing layer. The knowledge extraction phase of the life cycle, which was presented in Section II, consists of the image analysis layer using machine learning techniques for the content extraction. The semantic web technologies in the context of EO and Agriculture domains involve also the semantic

[20][Online]. Available: http://candela-h2020.eu/
[21][Online]. Available: http://candela-h2020.eu/content/semantic-search-v2
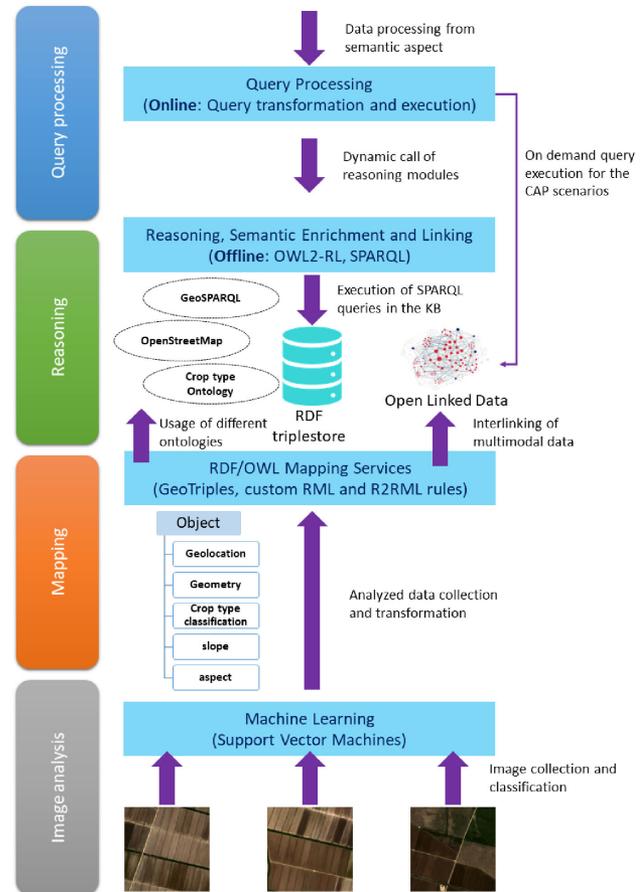
Fig. 2. System architecture overview.

annotation and transformation into RDF of the Mapping layer. The data ingestion and reasoning layer populates the knowledge base with the extracted knowledge, for storage and querying in a standard data representation model. Finally, the interlinking is done as part of the query processing layer that allows for checking the compliance of the farmers' declarations to the CAP regulations under the reasoning mechanism of the previous layers, using the open linked data paradigm. These layers are presented in detail in the following sections.

A. Satellite Image Analysis for the Monitoring of the CAP

Paying agencies of EU MSs, usually receive the annual subsidy applications in May or June. The paying agency inspectors require the information of the cultivated crop type, even as early as May. This way, inspectors can select and organize their OTSCs, which follow in the coming months. Additionally, crop classification results, received prior to the annual farmer declarations, can assist as an alerting mechanism during the application process. The image analysis layer is the first layer of our proposed pipeline (see Fig. 2), where a Sentinel-based crop classification system for the monitoring of the CAP is developed.

The AOI is located in northeastern Spain and specifically the district of Navarra (see Fig. 3). The AOI covers the agricultural land surrounding the city of Pamplona, capital of Navarra.
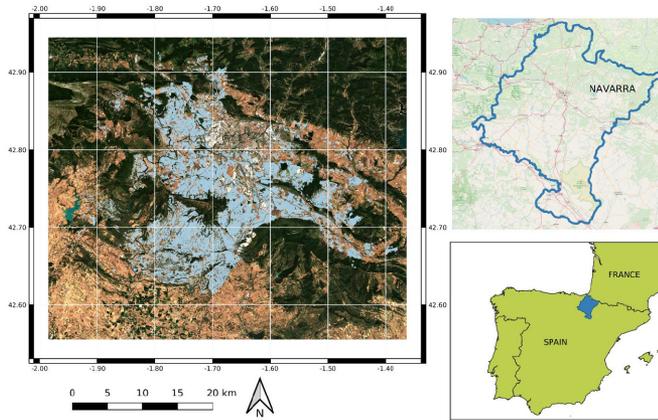
Fig. 3. Study area located in northeastern Spain and specifically in the Navarra district. The parcels of interest are shown in light blue color.

Fig. 4. Timeline of the growth cycle of major crops in Navarra, reworked from [8], together with the acquisition dates of the Sentinel-2 images.

In detail, the dataset includes 9052 parcels and amounts to approximately 215 km² of total land area. The Northern part of Navarra is surrounded by the Pyrenees Mountains, as they stretch southward from France. The landscape of the district is a mixture of forested mountains and watered valleys, whereas the agricultural land is characterized by substantial fragmentation [47]. This study builds upon the crop classification results of [8], as described in Section II. SVM-based crop maps are produced, including the crop types of soft wheat (50%), barley (26%), oats (8.4%), maize (1.4%), sunflower (3.2%), vineyards (1.3%), broad beans (4.5%), rapeseed (5.4%), and cherry trees (0.2%). The aforementioned crop types are the lowest level of ontology, as shown in Fig. 5.

The dataset for training the SVM classifier is based on the LPIS, which includes the parcel polygons in vector format and the associated farmer declaration for the 2018 CAP subsidy applications. The parcel polygons are used for segmenting the stack of Sentinel imagery to objects. The LPIS was provided by INTIA,[22] a public company, part of the Department of Rural Development, Environment and Local Administration of Spain. INTIA serves the role of paying agency for the district of Navarra, performing all CAP compliance inspections for the area. INTIA has additionally provided the timeline of growth for the major crops of the area. Fig. 4 illustrates the acquisitions of Sentinel-2 images, spanning over the entirety of the various crop cycles. The feature space used for the crop classification includes the Sentinel-2 images for the acquisitions depicted in Fig. 4. The acquisitions have been selected to have minimal cloud coverage over the AOI. All spectral bands, except B09 and B10, were used, along with the VI normalized difference vegetation index, normalized difference water index (NDWI), and plant senescence reflectance index. In this study, NDWI is used as defined by Gao [48]. Sentinel-2 images are atmospherically corrected to bottom of atmosphere reflectances using the Sen2Cor tool, and all bands are resampled to 10 m spatial resolution. The feature space comprises parcel entities described by the mean value, for all features, of the pixels that fall within their LPIS boundaries.
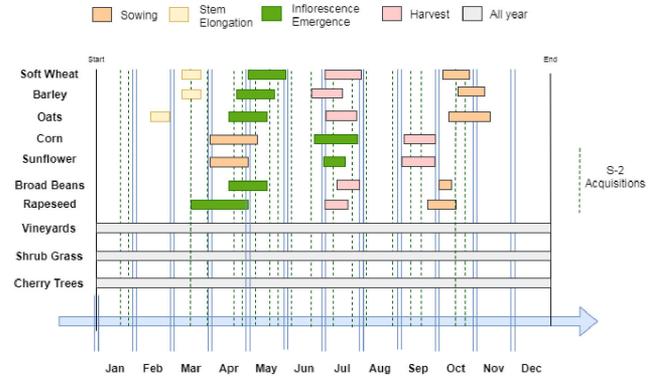
The proposed methodology is based on a traffic light system approach. Specifically, each parcel is categorized into four groups, each offering different levels of confidence. These categories comprise the green, yellow, red, and unreliable classes, indicating high to low levels of confidence, in that order. The categorization of each parcel is based on the difference between the two highest SVM scores. This study focuses predominantly on the green category, namely the decision of highest confidence. It is on those most confident samples that we then record the mismatches of model predictions and the farmer declarations. An alarm mechanism is then introduced, identifying the green parcels that have been systematically misclassified (mismatch of declaration and prediction) during the cultivating season.

In the proposed algorithm (Algorithm 1), the alarms of potential false declarations are detected for any time instance throughout the year, with variable accuracy considering the satellite imagery available to date. The X_train and X_test are the training and test feature spaces, respectively. The feature spaces are dynamically populated with all new acquisitions. Therefore, when the algorithm is executed ($currentDate$), it uses the up to date feature spaces as input, containing imagery until the latest available acquisition

$$acqDate = acqDate_k, k = 1, \ldots, \lambda, \ldots, A \qquad (1)$$

where A is the index to the latest acquisition prior to $currentDate$ and λ is the index to the acquisition that defines the starting feature space, early in the year.

The algorithm iterates A-λ times, each time recording the misclassifications (mis). Misclassifications, in this context and as previously stated, refer to the mismatch between the SVM model's prediction and the farmers' declaration. For each iteration, a second-order polynomial SVM model is trained based on $X_\text{train}(t)$ (2) and the SVM scores are computed after applying the model to $X_\text{test}(t)$ (3). $X_\text{train}(t)$ and $X_\text{test}(t)$ are the training and test data for each iteration, and f is the number of individual features for each acquisition. The farmer declarations, as part of the annual subsidy application for the CAP, are used for labeling the parcels and thus training the model. A stratified random split was performed to split the samples into 30% and 70% subsets for $X_\text{train}(t)$ and $X_\text{test}(t)$, respectively. This amounts to 2716 parcels, which have been used for training. All classification metrics that

[22][Online]. Available: https://www.intiasa.es/en/

Algorithm 1: Smart Sampling.

Input: $X_{train} = \{(x_i, d_i), x_i \in R^l, d_i \in \{1, 2, \ldots, m\}, i = 1, \ldots N\}$, $X_{test} = \{(x_i, d_i), x_i \in R^l, d_i \in \{1, 2, \ldots, m\}, i = 1, \ldots, M\}$, $acqDate = \{acqDate_k, k = 1, \ldots, \lambda, \ldots, A\}$, number of iterations $t = 0$, the acquisition number that bounds that starting feature space λ, the date the algorithm is executed $currentDate$, persistence threshold for each iteration $Pt = 0$, number of misclassifications $mis = \{mis_i, i = 1, \ldots M\}$

Output: Alarms for potential breaches of compliance

  while $acqDate_t \leq currentDate$ **do**

    $X_{train}(t) = \{(x_i, d_i), x_i \in R^{(acqDate_{(t+\lambda)} \cdot f)}, d_i \in \{1, 2, ..m\}, i = 1, ...N\}$

    $X_{test}(t) = \{(x_i, d_i), x_i \in R^{(acqDate_{(t+\lambda)} \cdot f)}, d_i \in \{1, 2, ..m\}, i = 1, ...M\}$

    $h(\overrightarrow{x}) = \sum \alpha_i y_i (\overrightarrow{x_i} \cdot \overrightarrow{x} + b)^2 + b$ *(Train SVM)*

    $\pi_d = P(d|X_{test(\tau)}) = \frac{1}{(1+e^{(A*f(X_{test(t)})+B)})}$
    (Calculate SVM Scores)

    $score = \max \pi_d - \max(\pi_d - \max \pi_d)$

    if $t \bmod 2 = 1$ **then**

      $P_t = P_t + 1$

    end if

    $alarms = \{\}$

    for $i = 1$ to M **do**

      if not $score \geq threshold$ **then**

        continue   (Bypass unreliable decisions)

      end if

      if parcel is misclassified**then**

        $mis_i = mis_i + 1$

      end if

      if $mis_i \geq P_t$ **and** $mis_i > 0$ **then**

        $alarms = alarms \cup i$

      end if

    end for

    $t = t + 1$

  end while

---

are presented in later sections have been averaged for 20 random splits of different seeds. The percentage of training samples was ultimately set to 30% after experimenting with larger datasets, which have provided only a marginal increase in performance

$$X_{\text{train}}(t) = \{(x_i, d_i), x_i \in R^{(acqDate_{(t+\lambda)} \cdot f)}, d_i \in \{1, 2, ..m\},$$
$$i = 1, ...N\} \qquad (2)$$

$$X_{\text{test}}(t) = \{(x_i, d_i), x_i \in R^{(acqDate_{(t+\lambda)} \cdot f)}, d_i \in \{1, 2, ..m\},$$
$$i = 1, ...M\} \qquad (3)$$

where x_i is the feature representation of the *i*th out of *N* parcels, belonging to $R^{(acqDate_{(t+\lambda)} \cdot f)}$. The superscript represents the dimensionality of the feature space. In each iteration, starting with $t = 0$, the feature space comprises the starting feature space, i.e., the one including all features, f, of all acquisitions until $acqDate_\lambda$, plus all features, f, of acquisitions $acqDate_{\lambda+t}$;

Algorithm 2: Smart Sampling Fine Tuning.

Input: *alarms* from Algorithm 1, $Y_{test} = \{(y_i \in 1, 2, ..m\}$ the actual estimations of the classifier, $D_{test} = \{(d_i \in 1, 2, ..m\}$ the declared labels

Output: Updated alarms for potential breaches of compliance

  $updatedAlarms = \{\}$

  $i = 1$

  $n = \text{size}(alarms)$

  while $i \leq n$

    ifseason of $y_i \neq$ season of d_i **then**

      $updatedAlarms = updatedAlarms \cup i$

    end if

    $i = i + 1$

  end while

---

d_i is the label for each parcel, ranging from 1 to m (=10), representing the different crop types.

The difference between the two highest per class scores $P(d|X_{\text{test}(t)})$ for each sample is recorded as the overall score value (4) for the selection of the most confident decisions against a *threshold*. These parcels constitute the *green* labels in the aforementioned defined traffic light system. We denote by π_d the difference scores, i.e., $\pi_d = P(d|X_{\text{test}(t)})$, so the overall score is given by

$$score = \max \pi_d - \max(\pi_d - \max \pi_d). \qquad (4)$$

The algorithm returns the misclassifications of the last iteration, namely the confident decisions of mismatch between the prediction and the declaration, which are classified as $alarms$. $alarms$ are the samples that have been found misclassified at least P_t times, in all previous iterations. P_t is varying based on the time within the year the algorithm is executed

$$P_t = \sum_{t=1}^{A-\lambda} t \mod 2. \qquad (5)$$

Early classifications are characterized by limited reliability, as the imagery included the training datasets does not cover the entirety of crops' growth cycle. For this reason, Algorithm 2 can be optionally used to further refine the selected alarms.

Algorithm 2 uses the alarms of Algorithm 1 as input and returns an updated set of alarms. In order to increase the reliability of the smart sampling algorithm, we select alarms for which the crops are classified to a type of a completely different crop season class that then one of the type declared.

B. Mapping Layer for Semantic Representation

An important aspect of the framework is the representation of the available information, e.g., crop classification results, as well as capturing of domain knowledge needed to further correlated results. For the former, we use GeoTriples to transform data into the RDF format, whereas for the latter we developed a domain-specific ontology. The metadata and analysis results
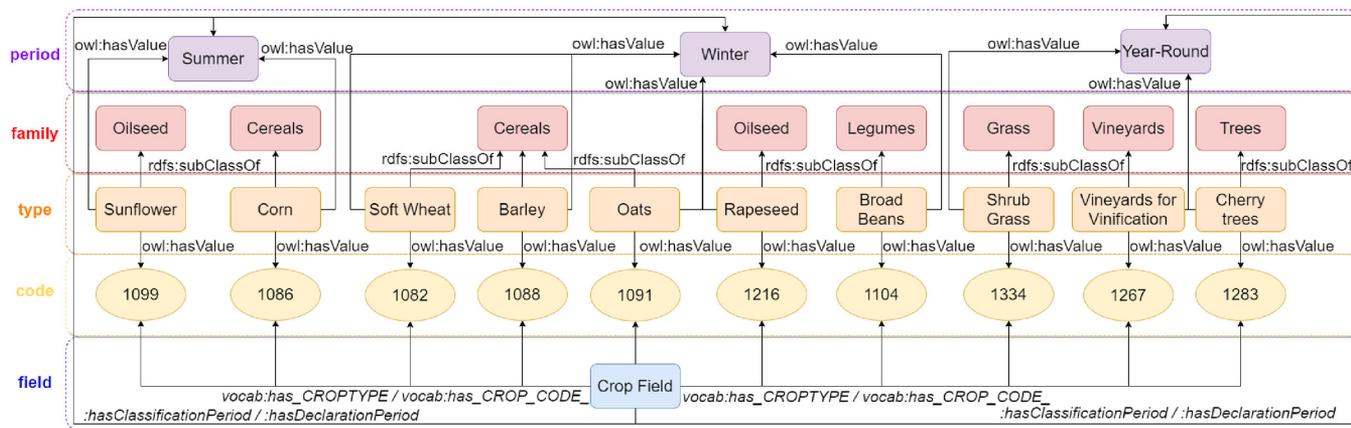
Fig. 5. Crop type ontology (knowledge extracted from [8]).

are integrated using geotagged information and analyzed using semantic queries.

As far as the domain ontology is concerned, it consists of the following three layers.

I) Data, which describe crop land taxonomy. Fig. 5 shows the relationship between period, crop family, crop type, and crop code classes. More specifically, each crop type (crop type classes) has a unique crop code (crop code property) and belongs to a specific crop family (family classes). Crop families are connected with the season that each crop thrives (period classes). Each crop thrives in a different season, whereas some crops seem to thrive in year-round basis. Same crop families may belong to different seasons when they represent different crop types.

II) The crop type classification data that contain a linking of metadata, which describe field information like parcel identifier, geometry, slope, aspect and classification scores for all different crop types.

III) Data collected from OpenStreetMap, containing water and waterways information in geospatial format, i.e., the geometries of hydrographic network objects.

Fig. 5 depicts the relationship between crop fields, periods, families, types, and codes. Crop field is a class that corresponds to the crop fields that have been identified in a classification run. Each crop type is a subclass of a specific family and related to a specific crop code. All periods, families and types are different classes in the ontology. Crop codes are the main interconnection point between the crop fields and the characteristics of each crop type. OWL property restrictions have been identified to automatically detect the period of the crop code declaration or classification using the value of the crop code. These relationships are useful for extending the dataset described in layer II of the ontology (crop type classification data). For instance, in the example presented in Listing 2 the crop code that has been identified by classification is 1082 and the crop field is associated with the classification period winter. The crop code that the farmer declared is 1334, whose classification period is

year-round. OWL2 RL is used to make useful inferences and apply reasoning rules into crop type classification data.

Listing 1: Example OWL2 RL.

```
:SoftWeat
  a owl:Class ;
  rdfs:subClassOf :Cereals ;
  rdfs:subClassOf [
      a owl:Restriction ;
      owl:hasValue :winter ;
      owl:onProperty :hasClassifica-
tionPeriod ;
    ] ;
  owl:equivalentClass [
      a owl:Restriction ;
      owl:hasValue "1082" ;
      owl:onProperty vo-
cab:has_CROPTYPE ;
    ] ;.
```

In the example presented in Listing 1, the classification period and hyperclass have been automatically assigned to the crop field, taking advantage of the dynamics of OWL2 RL. T-Box reasoning (in the form of OWL 2 RL entailment rules supported by the GraphDB implementation) is applied to infer that soft wheat is a cereal, crop type "1082" corresponds to soft wheat and classification period "winter," etc. More specifically, the rdfs:subClassOf has been used to assign the crop family values, whereas owl:Restrictions have been used to assign the value "winter" on the property hasClassificationPeriod when the value of the property vocab:has_CROPTYPE is "1082."

The transformation of crop type classification data and data collected from OpenStreetMap into the RDF model has been supported by GeoTriples. The tool accepts data in shapefile format and automatically produces an RDF mapping language (RML) file containing the rules that the RDF file should satisfy. Then, it produces the RDF serialization that satisfies the

Listing 2: Example of Crop Field Information as Described in RDF Combining the Crop Type Ontology (I) and the Results of Crop Type Classification (II).

```
@prefix :        <http://mklab.iti.gr/
 ontologies/croptypes/>.
@prefix owl:   <http:
 //www.w3.org/2002/07/owl#>.
@prefix xsd:    <http:
 //www.w3.org/2001/XMLSchema#>.
@prefix vocab: <http://example.com/
 ontology#>.
@prefix rdfs:  <http:
 //www.w3.org/2000/01/rdf-schema#>.
@prefix fa:     <http://example.com/
 farmer_ontology#>.
@prefix map:    <http://example.com/#>.
@prefix geo: <http:
 //www.opengis.net/ont/geosparql#>.

<http://example.com/parcels_
 classification_wscores_v3/Geometry/14356>
        a        geo:Geometry ;
        geo:asWKT
                "<http://www.opengis.
net/def/crs/EPSG/0/4326>
                MULTIPOLYGON (((-
1.6927107344688275 42.649935712372795,
                ..., -
1.692387508984902  42.65003808978244,
                -
1.692421288894873 42.65001269177615,
                -
1.6925231315390605 42.649985701208124,
                -
1.6927107344688275 42.649935712372795))
                )"^^geo:wktLiteral.

<http://example.com/parcels_
 classification_wscores_v3/id/14356>
        a                       vo-
cab:parcels_classification_wscores_v3 ;
        fa:hasOwner          fa:farmid5 ;
        vo-
cab:has_ASPECT          5.33798E1 ;
        vo-
cab:has_CROPTYPE        "1082" ;
        vo-
cab:has_CROP_CODE_      1334 ;
        vo-
cab:has_ID              773280 ;
        vo-
cab:has_SLOPE           5.0291E0 ;
        vo-
cab:has_scores_t_1      7.511E-3 ;
        vo-
cab:has_scores_t_2      4.33932E-1 ;
        vo-
cab:has_scores_t_3      5.4892E-2 ;
        vo-
cab:has_scores_t_4      1.41E-2 ;
        vo-
cab:has_scores_t_5      4.9191E-2 ;
        vo-
cab:has_scores_t_6      3.6199E-2 ;
        vo-
cab:has_scores_t_7      2.17041E-1 ;
        vo-
cab:has_scores_t_8      2.802E-2 ;
        vo-
cab:has_scores_t_9      3.8657E-2 ;
        vo-
cab:has_scores_typ      1.20457E-1 ;
        :hasClassificationPe-
riod   :winter ;
        :hasDeclarationPe-
riod     :year-round ;
        geo:hasGeometry
    <http://example.com/parcels_
 classification_wscores_v3/Geometry/
 14356>.
```

RML rules. The well-known text representation of coordinate reference systems standard has been reused to capture location-related information.

For the classification results, GeoTriples tool is used multiple times to convert the results of each run into RDF format. Each file has a unique name and because of that, new instances are created in the knowledge base. In the end of this procedure, the Knowledge Base contains many different instances of the same parcel having in common the parcel identifier. In such way, we keep crop type classification data of past runs in the knowledge base, which can be aligned with new classification data.

C. Data Ingestion and Querying Layer for Storage and Semantic Enrichment

Semantic enrichment aims at interconnecting and further enriching the contents of the generated knowledge graphs, applying semantic rules. The focus is given on improving the smart sampling methodology, executing a set of queries (rules) to improve the selection process of OTSCs. In the following, we describe the specifics of the approach, presenting the defined rules.

With every new image acquisition, a new crop classification is performed and the classification results are dynamically populating the knowledge base. Procedural code is used to run semantic queries in sequence and pass the needed values from past queries (parcel, value1, value2) into next ones (Listings 3–8). The reason for this decision is that SPARQL lacks in terms of arguments saving or passing into next queries and these calculations are better expressed using many SPARQL queries, improving the execution time. The queries that are presented in this section are running for each parcel instance. The parcel instances are

retrieved by the query depicted in Listing 3. A parcel instance (described as $< parcel >$) may be, for example, http://example.com/parcels_classification_wscores_v8/id/1109. The first step (Listing 4) is to retrieve the two highest classification scores among the ten different crop types, as these are defined in the ontology (see Fig. 5). This is done in order to compute the *score* value in Algorithm 1. This process takes place for each of the different classification instances in the knowledge base. In this query, typical SPARQL functions are used, such as BIND, to group different type results under the same variable, and ORDER BY to arrange the results in descending order.

Listing 3: Semantic Query to Retrieve All Parcel Instances for the Latest Classification Run.

```
PREFIX vo-
cab: <http://example.com/ontology#>
SELECT * WHERE {
    ?parcel a vo-
cab:parcels_classification_wscores_v8.
}
```

The difference between the two largest values (Listing 5) is calculated using the results of the previous query (value1, value2). If the difference is bigger than a specific threshold, the query in Listing 6 is used to mark the parcel as "green," i.e., parcels with a high confidence that the prediction is correct.

In addition, each time a green parcel is identified, the query in Listing 7 marks the parcels that have been incorrectly classified. In the end, the query in Listing 8 marks the parcels that the season of the declaration does not agree with the season of the classification, based on the domain ontology (see Fig. 5). SPARQL INSERT function is used to enrich the parcels with "misclassification," "green parcel," and "same season" information. All in all, semantic enrichment tries to enrich smart sampling, taking into account the provided classification results and domain knowledge about crop types, so as to detect parcels that have a high probability to have a false crop type declaration by the farmers.

D. Query Processing Layer for Interlinking Spatial Data

By capturing data in the RDF model space, we enable spatial relationships-based querying and easy integration with other data sources, such as linked data. We demonstrate the query answering capabilities of the framework, as well as the ability to integrate external datasets, by defining queries to detect possible noncompliance of the farmers according to the specified rules.

1) Greening 1 Requirement: The query in Listing 9 calculates the number of different crop types that a farmer cultivates and the total area of their farm. More specifically, the query detects a breach of compliance when the farmers own a total farm area between 10 and 30 ha and cultivate at least two different crop types, but the dominant one is more than 75% of the total farm area. The crop type as it is calculated in the SVM classification is expressed in the mapping via the vocab:has_CROPTYPE property. In the RDF space, we use count(distinct ?ctype) as ?count to detect the number of the

Listing 4: Semantic Query to Retrieve the Two Highest Classification Score Values Per Parcel.

```
PREFIX vo-
cab: <http://example.com/ontology#>
select ?max_types where {
    <parcel> vo-
cab:has_scores_t_1 ?type_1.
    <parcel> vo-
cab:has_scores_t_2 ?type_2.
    <parcel> vo-
cab:has_scores_t_3 ?type_3.
    <parcel> vo-
cab:has_scores_t_4 ?type_4.
    <parcel> vo-
cab:has_scores_t_5 ?type_5.
    <parcel> vo-
cab:has_scores_t_6 ?type_6.
    <parcel> vo-
cab:has_scores_t_7 ?type_7.
    <parcel> vo-
cab:has_scores_t_8 ?type_8.
    <parcel> vo-
cab:has_scores_t_9 ?type_9.
    <parcel> vo-
cab:has_scores_typ ?type_10.
    {BIND(?type_1 as ?max_types)}
    union
    {BIND(?type_2 as ?max_types)}
    union
    {BIND(?type_3 as ?max_types)}
    union
    {BIND(?type_4 as ?max_types)}
    union
    {BIND(?type_5 as ?max_types)}
    union
    {BIND(?type_6 as ?max_types)}
    union
    {BIND(?type_7 as ?max_types)}
    union
    {BIND(?type_8 as ?max_types)}
    union
    {BIND(?type_9 as ?max_types)}
    union
    {BIND(?type_10 as ?max_types)}
}
ORDER BY DESC (?max_types)
LIMIT 2
```

different crop type values that are detected for each farmer (GROUP BY ?owner). Additional queries are applied to also check, for instance, the farmers that have less than three different crop types cultivated in a total farm area of more than 30 ha if: $\{sum > 300000\} \wedge \{count < 3\} \wedge \{max_{croptype}[1] + max_{croptype}[2] > 0.95 * sum\}$ as it has been described in Section I. The query (Listing 10) further supports the Greening

Listing 5: Semantic Query to Detect the Parcels That the Difference From the Two Highest Score Values Is Above Threshold.

```
PREFIX xsd: <http:
 //www.w3.org/2001/XMLSchema#>
SELECT * where {
    BIND (xsd:double(value1)-
xsd:double(value2)
       AS ?result)
    FIL-
TER (?result>"0.5"^^xsd:double)
}
```

Listing 6: Semantic Query to Mark Green Parcels That Satisfy the Conditions of the Two Previous Queries.

```
PREFIX vo-
 cab: <http://example.com/ontology#>
PREFIX xsd: <http:
 //www.w3.org/2001/XMLSchema#>
PREFIX map: <http://example.com/#>
INSERT {
    <parcel> map:risk "green parcel".
} WHERE {
    <parcel> ?p ?o.
}
```

Listing 7: Semantic Query to Mark Parcels as Misclassified When Parcels Have Been Marked as Green and Declaration Does Not Agree With Classification.

```
PREFIX vo-
 cab: <http://example.com/ontology#>
PREFIX xsd: <http:
 //www.w3.org/2001/XMLSchema#>
PREFIX map: <http://example.com/#>
INSERT {
    <parcel> map:misclassification "mis-
classification".
} WHERE {
    <parcel> map:risk "green parcel".
    <parcel> vo-
cab:has_CROP_CODE_ ?decl.
    <parcel> vo-
cab:has_CROPTYPE ?class.
    FILTER(?decl!=xsd:integer(?class))
}
```

Listing 8: Marking Parcels Where Declaration and Classification Belong to Different Seasons.

```
PREFIX vo-
 cab: <http://example.com/ontology#>
PREFIX xsd: <http:
 //www.w3.org/2001/XMLSchema#>
PREFIX map: <http://example.com/#>
PREFIX owl: <http:
 //www.w3.org/2002/07/owl#>
PREFIX rdfs: <http:
 //www.w3.org/2000/01/rdf-schema#>
PREFIX : <http://mklab.iti.gr/
 ontologies/croptypes/>
INSERT {
    <parcel> map:same_season "false".
} WHERE {
    <parcel> vocab:has_ID ?id.
    <parcel> vo-
cab:has_CROP_CODE_ ?decl.
    <parcel> vo-
cab:has_CROPTYPE ?class.

    ?type a owl:Class.
    ?type rdfs:subClassOf ?object1.
    ?object1 owl:hasValue ?code.
    ?type rdfs:subClassOf ?object2.
    ?object2 owl:hasValue ?period1.
    ?period1 a :Period.
    FIL-
TER regex (str(?code),str(?decl))

    ?type2 a owl:Class.
    ?type2 rdfs:subClassOf ?object3.
    ?object3 owl:hasValue ?code2.
    ?type2 rdfs:subClassOf ?object4.
    ?object4 owl:hasValue ?period2.
    ?period2 a :Period.
    FILTER regex (str(?code2),?class)
    FILTER (?period1!=?period2)
}
```

field, whereas the sum of the fields area using the results of GeoSPARQL area calculations. All operations are implemented using SPARQL functions such as count, distinct and sum.

2) SMR 1 Requirement: Another important information for the end users is the distance of the parcels from the hydrographic network objects. This is in accordance with the requirements of SMR1, as described in Section I. In order to effectively identify the parcels susceptible to contribute nitrate-rich soil to nearby surface water, a filtering mechanism takes place, accounting for the slope and aspect of the parcel. Results are important for both farmers and paying agencies that want to check the compliance according to SMR 1 requirement.

Listing 11 presents the query that lists the parcel instances where the distance from surface waters is lower than 10 m,

1 requirement by detecting the farmers that grow less than three different crop types in a total area of more than 30 ha and farmers that grow less than two different crop types in a total area of 10–30 ha.

The area of the fields is calculated using GeoSPARQL ext:area function, taking advantage of the polygon points coordinates. The number of different crop types is calculated using the result of the SVM classification prediction of crop type on the

Listing 9: Semantic Query to Extract Possible Noncompliance in the Greening 1 Requirement for the Farmers Owning a Total Farm Area Between 10 and 30 Ha and Cultivating at Least Two Different Crop Types, Where the Dominant one Is More Than 75% of the Total Farm Area.

```
PREFIX geo: <http:
 //www.opengis.net/ont/geosparql#>
PREFIX ext: <http:
 //rdf.useekm.com/ext#>
PREFIX fa: <http://example.com/farmer_
 ontology#>
PREFIX vo-
cab:   <http://example.com/ontology#>
PREFIX xsd: <http:
 //www.w3.org/2001/XMLSchema#>
select * where {
{
   select ?owner (sum(?area) as ?max)
?sum where {
        { se-
lect ?owner (count(distinct ?ctype) as
?count)
        (sum(?area) as ?sum) where {
                ?field fa:hasOwner
?owner.
              ?field vo-
cab:has_CROPTYPE ?ctype.
              ?field
geo:hasGeometry ?geo.
              ?geo geo:asWKT ?polygon.
              BIND(ext:area(?polygon)
as ?area).
              ?owner a fa:Farmer.
        }
        GROUP BY ?owner
        HAV-
ING (?sum > 100000 && ?sum <= 300000
&& ?count>=2)
     }
     ?field fa:hasOwner ?owner.
     ?field vocab:has_CROPTYPE ?ctype.
     ?field geo:hasGeometry ?geo.
     ?geo geo:asWKT ?polygon.
     BIND(ext:area(?polygon) as ?area).
     ?owner a fa:Farmer.
     }
     GROUP BY ?owner ?ctype ?sum
     ORDER BY DESC (?max)
  }
  FILTER (?max>0.75*?sum)
}
```

Listing 10: Semantic Query to Extract Possible Noncompliance in the Greening 1 Requirement for the Farmers That Grow Less Than Three Different Crop Types in a Total Area of More Than 30 Ha and Farmers That Grow Less Than Two Different Crop Types in a Total Area of 10-30 Ha.

```
PREFIX geo: <http:
 //www.opengis.net/ont/geosparql#>
PREFIX ext: <http:
 //rdf.useekm.com/ext#>
PREFIX fa: <http://example.com/farmer_
 ontology#>
PREFIX vo-
cab:   <http://example.com/ontology#>
PREFIX xsd: <http:
 //www.w3.org/2001/XMLSchema#>
select ?owner (count(distinct ?ctype)
 as ?count)
        (sum(?area) as ?sum) where {
     ?field fa:hasOwner ?owner.
     ?field vocab:has_CROPTYPE ?ctype.
     ?field geo:hasGeometry ?geo.
     ?geo geo:asWKT ?polygon.
     BIND(ext:area(?polygon) as ?area).
     ?owner a fa:Farmer.
}
GROUP BY ?owner
HAVING ((?sum > 300000 && ?count<3) ||
    (?sum > 100000 && ?sum <= 300000
 && ?count<2))
```

which is the buffer for organic manure application. Since both parcel (LPIS) and hydrographic network data[23] (data from OpenStreetMap) contain geospatial information (e.g., multipolygon, polygon, etc.), the distance is calculated using geof:distance function of GeoSPARQL. The function accepts two geometries and calculates the shortest distance between any two points of the specified geometries. A filtering mechanism selects the fields according to their distance from the hydrographic network objects, slope and aspect as described in Section IV-C3. The GeoSPARQL function ext:closestPoint is used to compute the closest points of each geometry compared to the other geometry. A string replacement pattern is used in order to retrieve the coordinates of the two points, which are utilized to compute the angle of the two points in degrees. GraphDB math functions are used to achieve such computations.

3) Smart Sampling: The query in Listing 13 supports the retrieval of the results of Section III-C. When querying for the parcels to be inspected through OTSC, at any given time in the year, all past classification instances until that point are used. More specifically, the query takes advantage of all past classification decisions for each parcel to ensure that the prediction is indeed a misclassification. The threshold, above which a parcel is considered to be persistently misclassified, is

---

[23][Online]. Available: https://download.geofabrik.de/europe/spain.html

Listing 11: Semantic Query to Detect Susceptible Parcels According to the SMR 1 Requirement Taking Into Account the Slope, the Aspect, the Angle, and the Distance of the Parcel From Hydrographic Network Objects.

```
PREFIX geof: <http://www.opengis.net/
def/function/geosparql/>
PREFIX geo: <http:
//www.opengis.net/ont/geosparql#>
PREFIX uom: <http:
//www.opengis.net/def/uom/OGC/1.0/>
PREFIX map: <http://example.com/#>
PREFIX ogc: <http:
//www.opengis.net/ont/geosparql#>
PREFIX ofn: <http:
//www.ontotext.com/sparql/functions/>
PREFIX xsd: <http:
//www.w3.org/2001/XMLSchema#>
PREFIX ext: <http:
//rdf.useekm.com/ext#>
PREFIX vo-
cab: <http://example.com/ontology#>
select ?parcel ?distance where {
  BIND(<parcel> AS ?parcel).
  ?parcel a vo-
cab:parcels_classification_wscores_v8.
  ?a vocab:has_ID ?id.
  ?parcel vocab:has_SLOPE ?slope.
  FILTER(?slope>12)
  ?parcel ogc:hasGeometry ?s.

  ?s geo:asWKT ?o.
  ?fGeom geo:asWKT ?fWKT.
  ?fGeom map:containsWater ?b.
  FILTER (?fGeom != ?s).
  FILTER NOT EXISTS {
    ?s map:containsWater ?wa.
  }
  BIND(geof:distance(?o, ?fWKT) as
?distance).

  FILTER(?distance<=0.1)
  BIND (ext:closestPoint(?o, ?fWKT) as
?clpoint1)
  BIND (ext:closestPoint(?fWKT, ?o) as
?clpoint2)

  BIND( re-
place( str(?clpoint1), "^[^0-9\\.-
]*([-]?[0-9\\.]+)
    .*$", "$1" ) as ?long )
  BIND( replace( str(?clpoint1), "^.*
    ([-]?[0-9\\.]+)[^0-
9\\.]*$", "$1" ) AS ?lat )

  BIND( re-
place( str(?clpoint2), "^[^0-9\\.-
```

```
]*([-]?[0-9\\.]+)
    .*$", "$1" ) AS ?long2 )
  BIND( replace( str(?clpoint2), "^.*
    ([-]?[0-9\\.]+)[^0-
9\\.]*$", "$1" ) AS ?lat2 )
  BIND (xsd:double(?lat)-
xsd:double(?lat2) AS ?x)
  BIND (xsd:double(?long)-
xsd:double(?long2) AS ?y)

  BIND (ofn:atan2(?x, ?y) AS ?atan2)
  ?parcel vocab:has_ASPECT ?aspect.
  BIND(IF(ofn:toDegrees(?atan2)>
"0"^^xsd:double,
    ofn:toDegrees
(?atan2),ofn:toDegrees(?atan2)+
    "360"^^xsd:double) AS ?toDegrees )
  BIND(IF(?aspect-45>0, ?aspect-
45, ?aspect+360-45) AS ?min)
  BIND(IF(?aspect+45>360, ?aspect-
360+45, ?aspect+45) AS ?max)

  FILTER(?min<?toDegrees &&
?toDegrees<?max)
}
LIMIT 1
```

defined in Listing 12 and is dynamically updated given the different parcel classes, that correspond to different times of the year, which exist in the triple store. The different classification runs are saved as instances of vocab:parcels_classification_wscores_v1 for the first run, vocab:parcels_classification_wscores_v2 for the second, etc. In such a way, using the FILTER function, we select the number of different runs that currently exist in the triple store by detecting the number of different classes that contain the word "parcels. The query in Listing 13 retrieves all the parcels that have been misclassified in past classifications more times than the defined threshold. Then the associated season for the crop type prediction of the latest available classification is compared with the associated season type of the declarations using the results of Listing 8. If there is a disagreement, the parcel is selected as an alert and candidate for OTSC. This query further narrows the smart sampling filter and is recommended for query executions early in the year (months before July), when the classification results are less trustworthy. More information about this functionality is described in Algorithm 2.

IV. EXPERIMENTS AND RESULTS

A. Hyperparameter Optimization and Performance of the SVM Classifier

The crop classification model was built using the SVC function from the scikit-learn library[24] of Python. The hyperparameters were optimized using grid search over a range of values for each parameter, using fivefold cross-validation.

Listing 12: Semantic Query to Calculate the Smart Sampling Threshold According to the Number of Different Parcel Classes, That Correspondto Different Times of the Year, Which Exist in the Triplestore.

```
PREFIX xsd: <http:
 //www.w3.org/2001/XMLSchema#>
PREFIX ofn: <http:
 //www.ontotext.com/sparql/functions/>

SELECT ?final_threshold WHERE {
    {
        SELECT (count(distinct ?par-
cel_class) as ?count)
        WHERE {
            ?parcel_instance a ?par-
cel_class.
            FILTER con-
tains(str(?parcel_class), "parcels")
        }
    }
    BIND(IF(ofn:floorMod(?count,2)=
"1"^^xsd:int,
        ((?count+"1"^^xsd:int)/2)-
"1"^^xsd:int,
        ?count/2) AS ?threshold )
    BIND(IF(?count="1"^^xsd:double,
"0"^^xsd:int,
        ?threshold) AS ?threshold )
    BIND(IF(?threshold>"5"^^xsd:double,
"5"^^xsd:int,
        ?threshold) AS ?fi-
nal_threshold )
}
```

Listing 13: Semantic Query to Check Persistent Misclassifications Using the Smart Sampling Threshold and the Number of Past Classification Instances That Have Been Misclassified Provided That in the Latest Classification Run the Period of the Farmer Declaration Does Not Agree With the Period of Classification.

```
PREFIX map: <http://example.com/#>
PREFIX vo-
 cab: <http://example.com/ontology#>
PREFIX xsd: <http:
 //www.w3.org/2001/XMLSchema#>
SELECT ?id (count(distinct ?par-
 cel) as ?count) WHERE {

    ?parcel vocab:has_ID ?id.
    ?parcel map:misclassification ?o.

    ?latest_parcel a vo-
cab:parcels_classification_wscores_v8.
    ?latest_parcel vocab:has_ID ?id.
    ?lat-
est_parcel map:same_season "false".
}
GROUP BY ?id
HAVING (?count>final_threshold)
ORDER BY desc (?count)
```

is extended cloud coverage throughout most of the season, resulting in sparse time-series of cloud-free Sentinel-2 imagery. In total, 10, 11, and 14 crop classes have been classified in the Spanish, Greek and Lithuanian cases, respectively.

Validated results were consolidated based on OTSCs that were performed by the respective paying agencies of the three AOIs, during the 2018 subsidy applications. In the case of the Spanish AOI, out of the 107 randomly selected parcels for inspection, 105 were classified correctly. In Greece, inspectors visited only parcels classified with high confidence, namely of high posterior probability for the classification decision, to crop types other than the one declared. These instances are considered as potential breaches of compliance. It was shown that 76 out of 85 inspected parcels were indeed wrongly declared and correctly classified by our model. Finally, in Lithuania, the validated dataset acquired through the inspections resulted in an overall accuracy of 76.2% in late June and 80% in late August out of 3319 parcels inspected. The results revealed the dependencies of the crop classification model performance, on the percentage of truthful declarations, the cloud coverage and the parcel shape and size.

In the Spanish AOI for which we focus in this work all these dependencies were optimal, i.e., more than 97% of truthful declarations, limited cloud coverage, and an average parcel size of 2 ha. Hence more than 90% classification accuracy was achieved.

Our model evaluation analysis also revealed that classification decisions for larger parcels and parcels with straighter borders

The hyperparameter combination that was selected was the one that produced the highest overall accuracy. For the penalty parameter of the error term, C, we tested values within the range $[2^{-2}, \ldots, 2^9]$. For the kernel coefficient, namely the gamma parameter, the optimal value was examined within the range $[10^{-4}, \ldots, 10^1]$, whereas for the independent term of the kernel function, coef0, we searched the range $[10^{-3}, \ldots, 10^2]$. The best combination consisted of $C = 4$, $\gamma = 0.001$ and $coef_0 = 10$.

Finally, the SVC function includes a parameter called *class_weight*, which was set to "balanced." Its impact is that it sets the parameter C of each class i to $class\_weight[i] * C$, where $class\_weight[i]$ is inversely proportional to class frequencies of the input data. This way the negative effects of an unbalanced dataset, such as the one used for this study, are ameliorated.

We evaluated our core classification model [8] for three different agricultural areas, in Greece, Lithuania and Spain, for which we had independent *in situ* validation data. These areas present various challenges. In the case of Greece, the agricultural landscape is significantly fragmented, resulting in small and narrow parcels, occupied by mixed pixels. In Lithuania, there

TABLE I
ACCURACY OF CROP TYPE CLASSIFICATION FOR DIFFERENT PARCEL
AREA RANGES

Spain			
		Accuracy	% parcels
Large	>1ha	96.70	44.85
Medium	0.35ha<x<1ha	94.44	23.85
Small	<0.35ha	87.02	31.30

TABLE II
RELATIONSHIP BETWEEN SVM SCORES AND OVERALL ACCURACY

Spain		
SVM score	Accuracy	% parcels
>0.85	97.88	77.43
0.7<x<0.85	89.46	10.27
<0.7	66.32	11.30

TABLE III
PA AND UA FOR THE CLASSIFICATION OF TEN CROPS TYPES USING THE FULL
SEASON TIME-SERIES OF SENTINEL IMAGERY

Class	UA (%)	PA (%)
Soft wheat	94.00	95.42
Maize	94.98	94.36
Barley	93.09	93.69
Oats	92.52	88.40
Sunflower	94.87	92.85
Rapeseed	95.34	92.08
Broad beans	94.84	89.40
Shrub grass	84.15	80.70
Vineyards	85.50	87.76
Cherry trees	83.22	81.98

tend to have higher accuracy than smaller parcels or parcels with more irregularly shaped boundaries. The parcel area is important since accuracy depends on the number of image pixels that fall within the parcel boundaries. Sentinel-2's 10 m pixel size equates to 50 image pixels in 0.5 ha of land. Our analysis shows that having 50 pixels of information provides accurate results, whereas for smaller parcels the decision is both less confident, namely of lower SVM score, and less accurate (see Tables I and II).

In Table II is observed that there is a strong correlation between the SVM score and accuracy. Indeed, the subset of parcels with an SVM score larger than 0.85, achieves an overall accuracy of more than 97%. On the other hand, the subset of parcels with SVM scores less than 0.7, achieves an overall accuracy of 66%.

B. Scenarios for the Control of the CAP

1) Scenario 1. Smart Sampling of OTSCs: The paying agency inspectors search for parcels of farmers that have potentially falsely declared the cultivated crop type. The parcels prone to noncompliance are dynamically provided, starting from late June until the end of the cultivation period, to allow the inspectors to better target their inspections.

Farmer profile: Dan is a farmer. He has into his possession a field containing barley, whereas he has declared that the field's crop type is maize.

Analysis results: His field has been selected for OTSC because the crop classification has classified the cultivated crop type as barley, with high confidence. Even though the classification is not particularly trustworthy, being performed in early June, the parcel is marked as high risk by the smart sampling algorithm, as the prediction belongs to a different season class, namely winter.

Decision: The paying agency inspector needs to check the field, as there is a strong possibility that the farmer has wrongly declared the cultivated crop type.

2) Scenario 2. Greening 1: Crop Diversification: As previously stated, paying agencies need an automated system to detect noncomplying farmers with respect to the Greening 1 requirement.

Farmer profile: Bob is a farmer. He has three fields in his possession covering an area of 27.4066 ha. The total area of

the fields that contain soft wheat is 27.1058 ha, whereas for the crops that contain rapeseed the total area is 0.3008 ha.

Analysis results: According to our analysis there seem to be two fields with soft wheat and one with rapeseed. The farmer does not comply with the rule because even though he cultivates two different crops, the total area of the soft wheat parcels exceeds the 75% of the total area of the fields that the farmer has into his possession.

Decision: The paying agency inspectors need to check these fields because there is a strong possibility that the farmer is not complying with the Greening 1 requirement.

3) Scenario 3. Detecting Parcels Prone to Noncompliance to the SMR-1 Requirement: The paying agency inspectors need to establish the compliance to the SMR 1 defined buffer zones. In order to do that the distance from the parcels to nearby surface waters is calculated. The slope and aspect of the parcels are also taken into account in order to establish if there is an actual risk for runoff.

Farmer profile: Lucy is a farmer. She has into her possession a field whose distance from surface waters is almost 3 m. The slope of the field is $15°$, therefore of high runoff risk, and the aspect $251°$, namely of western orientation. The aspect of the proximity line is $267°$.

Analysis results: Since the parcel appears to be very close to surface waters, within the SMR-1 buffer, and the difference of the parcel and proximity line aspects is within the range of potential runoff, the analysis marks it as high risk.

Decision: The paying agency inspectors need to monitor this parcel because there is a strong risk for nitrate-rich runoff to surface waters.

C. Implementation of the CAP Scenarios

1) Implementation of Scenario 1. Smart Sampling: In this scenario, the potential of better targeted, smart OTSCs is explored, exploiting accurate crop classification results, early in the year. This is achieved through the proposed semantic enrichment of classification results and the pertinent smart sampling query. The idea is to provide alarms, starting from early summer, when declarations are usually received, and dynamically update those using progressively larger feature spaces that include new Sentinel-2 acquisitions.

Table III summarizes the producer's accuracy (PA) and user's accuracy (UA) for the predictions of all crop classes. PA is

the percentage of correctly classified parcels against the total number of parcels of a given class. On the other hand, UA is the percentage of correctly classified parcels for a given class against the total number of parcels classified to that class. The metrics in Table III refer to the classification results produced using the entirety of acquisitions, 24/01/2018–21/10/2018, and have been averaged over 20 iterations of random training dataset splits. Additionally, the PA and UA results address the lowest level of crop taxonomy, as shown in Fig. 5.

Table III summarizes excellent classification performance for most of the crop types. Specifically, it is observed that the classifier achieves PA and UA values of more than 80% for all crop classes while reaching values as high as 95%. It is also worth noting the excellent performance for the classes oats, soft wheat, and barley. These crop types belong to both the cereals and winter superclasses, having similar spectral and phenological characteristics. Nonetheless, the classifier appears to discriminate among them easily. Finally, the weakest performances are observed for the classes of shrub grass, vineyards, and cherry trees. These are all year-long crop types, with no significant phenological characteristics to assist the classifier. Additionally, the shrub grass class, which is an ambiguous crop description, is characterized by diverse spectral characteristics from parcel to parcel.

The metrics shown in Table III theoretically allow for the effective sampling of OTSCs. However, the results presented have been produced using images until the end of October, when the notion of smart sampling becomes obsolete, as the inspections would have preceded. For this reason, the crop classification is performed at multiple instances throughout the year, starting as early as May 4. As it would be expected, the classification results of reduced feature spaces, when executed early in the year, would achieve suboptimal results of low reliability. In this regard, per class scores are computed for each sample, using class membership probability estimates [49]. Therefore each sample is associated with ten different membership probabilities, as many as the crop classes involved. In order to select the most reliable decisions, the difference between the two highest scores for each sample is recorded ($score$ in Algorithm 1). This is achieved via the semantic enrichment mechanism, as described in Figs. 4 and 5. Fig. 6 illustrates those differences in the form of a histogram, using 100 bins, for the entire dataset. The algorithm classifies most of the data with high reliability. The ones that are classified with low score are most likely crops that belong to the same family and some of them with similar spectral signatures are expected to be misclassified at some level.

The SVM algorithm is trained and tested for progressively larger feature spaces, whose classification results populate the knowledge base. The evolution of accuracy is shown in Fig. 7, using the F1-score metric [50]. F1-score, defined as $2 \times \left(\frac{\text{PA} \times \text{UA}}{\text{PA} + \text{UA}} \right)$, providing an overall accuracy metric that accounts for both PA and UA. The first point on the *x*-axis of Fig. 7 refers to the F1-score for the classification that uses images until the 4th of May. For each next run onward, features keep on populating the feature space, with every new acquisition. It is observed that the larger feature spaces result in better classification results. Specifically, feature spaces that include the July 23rd acquisition, appear to approach optimal F1-scores for most crop classes,
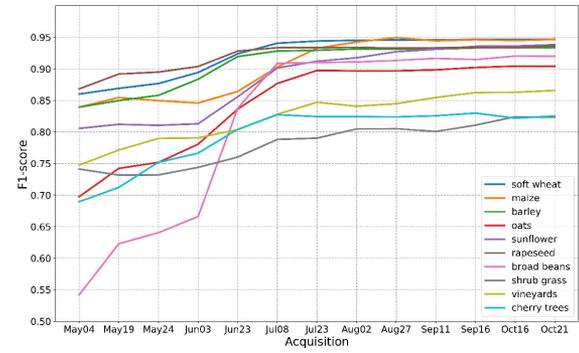
Fig. 6. Histogram of differences between the two highest scores of each sample decision. The closer is the difference to 1.0, the lower the classification uncertainty.
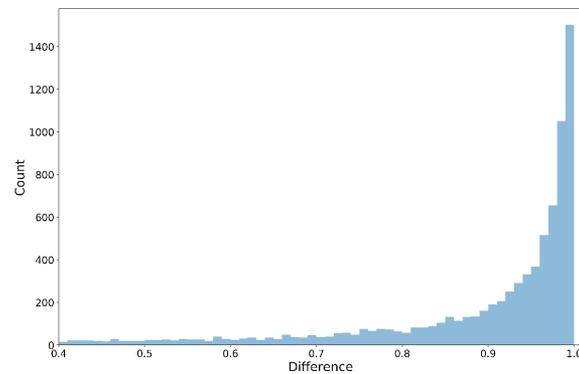
Fig. 7. Evolution of F1-score for crop type classification with progressively larger feature spaces.
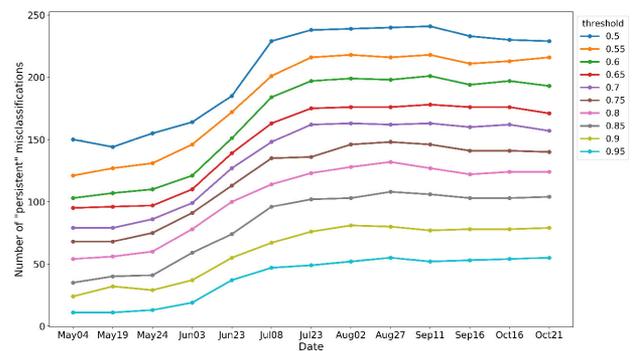
Fig. 8. Evolution of the number of constant persistent misclassifications for a varying threshold of the green category.

creating a plateau on the evolution curve. The majority of these crops are harvested from late June to early July, and thus our model can separate the classes more comfortably, when features that cover this period are added to the feature space.

Accurate crop classification enables the monitoring of the CAP rules and allows for efficient decision-making on the farmers' compliance. Toward this direction, the inspected parcels were assorted based on the previously described traffic light system. Two highest prediction probabilities or SVM scores (see Fig. 8) are taken into consideration to pinpoint the parcels of the
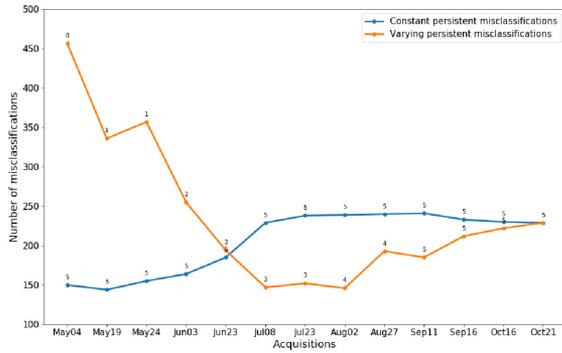
Fig. 9. Evolution of the number of misclassifications under constant (blue line) and varying thresholds (orange line). The varying threshold increases by one every two image acquisitions. The thresholds define which parcels are the misclassifications for each classification instance looking at the number of times they were found misclassified in previous runs.

TABLE IV
LIST OF FARMERS THAT SEEM TO BE NONCOMPLIANT TO THE CROP
DIVERSIFICATION REQUIREMENT AND INFORMATION ABOUT
THEIR CROP FIELDS

Farmer id	crop types count	sum area of parcels	main crop(s) area	crop types
37	1	245008		Soft Wheat
1005	2	309458		Soft Wheat, Barley
1007	1	191124		Soft Wheat
1133	2	274066	271058	Soft Wheat, Rape
1229	2	407609		Soft Wheat, Barley
1319	1	103471		Barley
1306	1	302232		Soft Wheat
1860	2	287242	278455	Soft Wheat, Rape
2677	3	577902	576071	Rape, Soft Wheat, Vinification vineyard
3131	3	614529	586915	Soft Wheat, Rape, Barley

highest probability of noncompliance, for which the algorithm is assumed to predict with high reliability.

At first, for each run with a different feature space in the knowledge base, the set of green misclassifications is recorded, namely the reliable instances for which the prediction does not match the declaration. Then, the set of persistent green misclassifications is computed using the semantic query in Fig. 13. Persistence refers to the number of times a sample has been found misclassified in the different classification iterations. In this study, there are two different thresholds of persistence. First, if a sample has been misclassified more than five times, in all different runs, from May to late October, then it is considered to be a validated alert. This is called constant persistent misclassification, as shown in Fig. 9, and is assumed to function as the validation dataset, against which performance metrics are computed.

Fig. 8 presents the number of constant persistent misclassifications relative to the increasing number of features, for different threshold values of what is considered a green parcel. The latter translates to varying values of the difference between the two highest scores for each sample. As expected, lower thresholds lead to higher number of misclassifications. It can also be observed that for each threshold that was tested, the pattern of the plot line is the same. The number of misclassifications is constant during May, then it presents an increase at the start of June, which begins to stabilize toward the start of August. It can be concluded that it is difficult for the algorithm to identify truly mislabeled data, really early in the year, but it seems to perform better with an increasing number of features. This improvement, which starts around June, is justified because most of the crops that were examined, are harvested at June, after which our model can classify the data with higher confidence.

However, in order to calculate the persistence of a parcel as mentioned above, a full Sentinel-2 series, for any given year of inspection, is required. Therefore, the constant persistent misclassifications are merely used as validation dataset of truly wrongly declared parcels, and are not part of the smart sampling algorithm. Thus, in order to be able to identify wrongly declared samples, in real scenarios, the total number of times that a green

parcel has been misclassified, until the current run, is calculated. Then, we set a different threshold for each iteration (P_t in Algorithm 1), which is calculated using the semantic query in Fig. 12. This is referred to as threshold of varying persistent misclassifications.

Fig. 9 displays the number of misclassified parcels using both constant and varying persistence thresholds and a constant green threshold equal to 0.5. The threshold value was defined based on the *a priori* knowledge of the percentage of false declarations, annually. INTIA has stated that usually no more than 3% of the agricultural parcels in Navarra are falsely declared. Additionally, INTIA, as acting paying agency for the Navarra region, is obligated to conduct randomly selected inspections for at least 1% of applications. Therefore, using 0.5 as the threshold for the green category, provides 267 constant persistent misclassifications, which amount to 3.1% of the dataset; satisfying both the expected false declarations percentage and the minimum number of mandated OTSCs. Inspecting Fig. 9, each point, of both curves, is associated with the persistence number that is used to count the misclassifications.

2) Implementation of Scenario 2. Crop Diversification: Table IV contains a list of farmers that seem to not comply with the rule described in Figs. 9 and 10. The first column contains the farmer identifier that corresponds to a specific farmer, the second contains the count of different crop types that the specific farmer cultivates in the fields that he has into his possession and the third the sum area of the parcels that the farmer has into his possession. The fourth column contains the area of the main crop for the cases that the farmer cultivates more than two crop types in an area of 10–30 ha but the main crop exceeds the 75% of the total area. It also contains the sum of the two main crops for the cases that the farmer cultivates more than three crop types in an area of more than 30 ha but the sum of two main crops exceeds the 95% of total area. The last column contains the crop types that the farmer is cultivating in their farm.

Fig. 10 presents the percentage of farmers that seem to be complying and noncomplying to the Greening 1 requirement. The number of farmers that there is no need to be checked is 2610 (82%), whereas the number of potentially noncomplying farmers is 703 (18%). More specifically, 506 (13%) farmers own 10–30 ha of arable farm from which 54 (1%) grow less than two
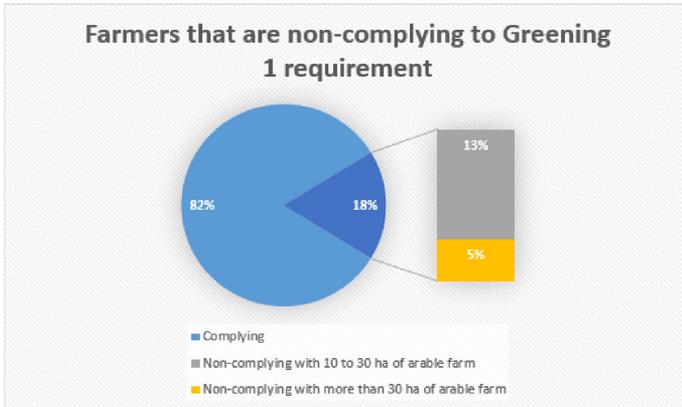
Fig. 10. Farmers distribution under Greening 1 requirement according to their compliance and total arable farm area that have under their possession.

Fig. 11. Example of parcel that is susceptible to runoff according to the SMR 1 requirement. The parcel slope aspect (arrow) and the location of the nearby Regata de Larrea river are also shown.

different crop types and 452 (14%) grow two or more different crop types but the main crop covers more than 75% of the land. On the other hand, 189 (5%) farmers own more than 30 ha of arable farm from which 68 (2%) grow less than three different crop types and 121 (3%) grow three or more different crop types but the two main crops cover more than 95% of the land. To perform these calculations we used the data of the classification run from early July (jul08).

3) Implementation of Scenario 3. SMR 1: In the SMR 1 requirement, we select the parcels that their distances from surface are under 10 m based on the organic manure spreading buffer. From these only parcels with slope higher than 12° are recorded, according to the SMR1 specifications. Additionally, in order to select only parcels with an actual risk for runoff, the aspect of the proximity line ($\alpha_{degrees}$), namely the orientation of the line connecting the parcel to the water object, needs to fulfill the relationship

$$aspect - 45 < \alpha_{degrees} < aspect + 45 \quad (6)$$

where the aspect is given per parcel and the proximity line aspect is calculated using the following formula:

$$\alpha_{radians} = \text{atan}_2(x_1 - x_2, y_1 - y_2) \quad (7)$$

$$\alpha_{degrees} = \alpha_{radians} \times 180°/\pi. \quad (8)$$

Fig. 11 shows the visual representation of the parcel that was mentioned in Section IV-B3. According to the aforementioned rules, this parcel is of high risk.

Fig. 12 presents the percentage of parcels that seem to be of low and high risk based on the SMR 1 requirement. The number of low-risk parcels that there is no need to be checked is 14407 (96%), whereas the number of high-risk parcels is 630 (4%).

D. Smart Sampling Accuracy

Table V presents the PA and UA of the misclassified green parcels, for varying and constant (5) persistence, with the latter functioning as the ground truth. To calculate the PA, the number of correctly misclassified green parcels of varying persistence (P_t) is divided by the total number of misclassified parcels in
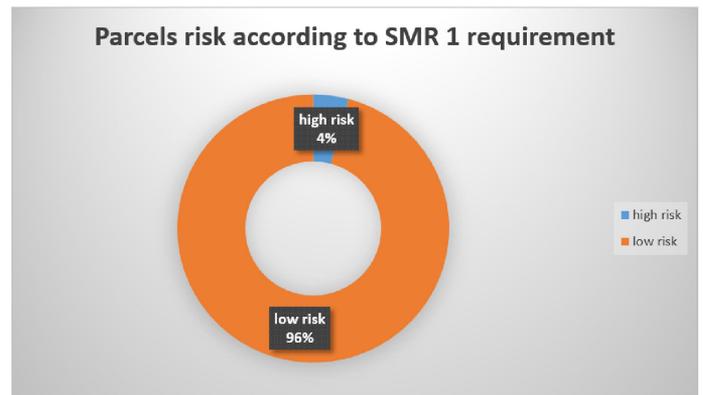
Fig. 12. Distribution of low-risk and high-risk parcels according to the SMR1 specifications.

TABLE V
PA AND UA EVOLUTION OF THE SMART SAMPLING ALGORITHM

Date	PA (%)	UA (%)
May04	56.18	32.82
May19	50.19	39.88
May24	52.81	39.50
Jun03	48.69	50.98
Jun23	50.94	70.10
Jul08	53.93	97.96
Jul23	56.93	100.00
Aug02	54.68	100.00
Aug27	71.54	98.96
Sep11	69.29	100.00
Sep16	79.40	100.00
Oct16	83.15	100.00
Oct21	85.77	100.00

TABLE VI
PERCENTAGES OF GREEN MISCLASSIFICATIONS FOR EACH SEASON CLASS AND
UA OF PERSISTENT ALARMS FILTERED THROUGH ALGORITHM

Date	Total	Winter	Summer	All Year	Alarms	Persistent Alarms UA (%)
May04	457	418	3	1	35	57.14
May19	336	302	3	1	30	63.33
May24	357	321	4	2	30	60.00
Jun03	255	226	4	1	24	70.83
Jun23	194	167	4	1	22	68.18
Jul08	147	125	3	1	18	94.44
Jul23	152	133	2	1	16	100.00
Aug02	146	128	1	1	16	100.00
Aug27	193	168	0	2	23	91.30
Sep11	185	164	0	2	19	100.00
Sep16	212	182	0	2	28	100.00
Oct16	222	186	1	2	33	100.00
Oct21	229	190	1	2	36	100.00

Fig. 13. Smart sampling alerts at different instances throughout the year of inspection. The red parcels indicate to the paying agency inspector where to target their inspections. Early in the year the alerts are fewer, stemming from less reliable classifications. With more images as we move along in time, classifications become more reliable and thus more alerts are identified.

the validation dataset (267), whereas for the UA it is divided by the total number of misclassified green parcels of varying persistence.

Table V summarizes that suboptimal PAs and UAs are achieved for classifications early in the year. This indicates that the smart sampling algorithm will select erroneously a significant percentage of the suggested OTSCs. Nevertheless, the smart sampling precision, indicated by the UA, reaches near perfect values, from early July08 onward.

In an attempt to further filter the OTSCs selection of the suboptimal smart sampling results, found early in the year, we exploit the season type of the crop taxonomy, as described in Fig. 5. Each of the crops that were examined belongs to one of the three season types and that is summer, winter, and year-round. Based on that, a false declaration would be more likely, if the prediction of the algorithm belongs to a different season class as compared to the corresponding season class of the declaration. Table VI presents the number of the green misclassifications that come from each of the different season types as well as the total sum of them, along with the number of green misclassifications for which the predicted crop season differs from the declared one, identified as an alarm. Finally, the last column indicates the percentage of those alarms that belong to the validation dataset of more than five persistent misclassifications.

Comparing the "Persistent Alarms UA" column of Table VI and the "UA" column of Table III reveals increase in the precision of smart sampling for runs that take place as early as May. Special attention is given to UA instead of PA, as it demonstrates the reliability of the system. Since the inspections are not exhaustive but rather sampled based, it is more important to ensure that most alerts are indeed wrong declarations. In Fig. 13 is displayed the evolution of the smart sampling alerts for different classifications throughout the year. It can be observed that for runs until June 23, the alerts are only few as they are passed through the Algorithm 2 filter. It can be seen that even though some alerts only appear early in the year, the critical mass of them can be identified from as early as July.

TABLE VII
MEAN EXECUTION TIME OF EACH SCENARIO ACCORDING TO THE SEMANTIC
QUERIES DESCRIBED IN SECTION III-D

Query	Execution time (sec) actual size	Execution time (sec) reduced size
Smart Sampling	0.4	0.4
Greening 1 requirement	7.35	0.023
SMR 1 requirement	1.16	0.016
Number of triples	7.369.055	1.260.600

E. Scalability

Table VII presents the execution time for each one of the rules that are described in this work. The execution time for each rule is calculated as the mean time of each scenario of the queries presented in Section III-D. Results show that the query that takes the most time to run is the SMR 1 requirement, taking into account that this query runs per parcel, compared to the other queries that run for the whole dataset. The reason is that this query requires multiple calculations between polygons such as distance and aspect, which are very time-consuming. The second column presents the execution time in seconds for the actual size of the dataset, whereas the last one presents the execution time in a reduced dataset size in order to further understand the scalability.

Fig. 14 presents the mean execution time of SMR 1 requirement for each parcel according to the risk type (low, high). The execution time is extremely low when the parcel does not satisfy the filters (aspect, distance, slope), whereas when the parcel satisfies the filters the execution time is significantly higher. Despite this fact, the mean execution time of all parcels is significantly low, taking into account the large number of low-risk parcels.

F. CAP Monitoring in Practice

The results in Section IV-C described the implementation of the practical applications of the proposed system, which have been showcased in the form of the three scenarios of Section IV-B. Scenario 2 on Greening 1 and Scenario 3 on SMR-1

Fig. 14.   SMR 1 requirement mean execution time for each parcel according to the risk type (low, high) compared with total mean execution time.

requirements, respectively, are performed once for every year of inspection. Therefore, these two scenarios infer a single cycle of execution, with reference to Fig. 1. On the other hand, the Smart Sampling application (Scenario 1) is an iterative process. The first crop classification is performed in early May, coinciding with the CAP subsidies applications commencement, with subsequent iterations producing progressively more accurate crop type maps with every new Sentinel-2 acquisition. Hence, there is a full-cycle execution (see Fig. 1) with every newly acquired image, from ingestion to interlinking.

According to the results in Table VI, the issued alarms are adequately trustworthy to suggest potential OTSCs from as early as July, with a UA of more than 94%. From then on, the paying agency inspectors can have targeted OTSCs that increase in number with every new iteration, following their summer-long inspection process. Alternatively, the results from May until early July can be used in assistance of the farmer application process. The applicants and the paying agencies can have an indication of potential noncompliance even at the application stage, thus allowing the farmer to make timely changes to their application.

V. Conclusion and Future Work

In this article, we presented a semantic-oriented framework for knowledge discovery using a supervised classification in the CAP domain. The main focus is given on the detection of possible violations according to the declaration of the farmers, the Greening 1 requirement and SMR 1 requirement. The framework can strongly assist in decision-making issues by providing helpful information to paying agency inspectors and environmental consultant to detect possible breaches of compliance. The proposed solution relies on data coming from sentinel images and open data (e.g., OpenStreetMap). Common denominator between the two datasets is the provision of georeferenced information. Data refer to a region in northeastern Spain. The SVM classification method has been applied to classify the cultivated crop types for multiple instances throughout the cultivation season. The data, which are in Shapefile format, are converted into Turtle RDF format using the GeoTriples tool. Results are saved in GraphDB triplestore. Semantic queries are executed to enrich

the data with information about possible farmers noncompliance according to agricultural policies.

In this work, we have shown how the paying agencies of the CAP can benefit from the exploitation of big Copernicus data. We showcased how with only freely available satellite data and ancillary LOD one can provide actionable information. Combining the state of the art in EO-based crop classification, semantic enrichment and linking free and open data has facilitated the development of an end-to-end system, from data acquisition to CAP related decision-making. The main innovations of the presented methodology include its reusability and transferability, using predominantly open data and requiring minimal finetuning when applied to other regions, and scalability, accounting for all big data considerations and choosing the computationally efficient alternative every step of the way, toward the monitoring approach of the new CAP.

Future work includes the investigation of similar datasets that correspond to other regions to apply the agricultural policies rules. The proposed framework can also be extended to similar problems, beyond the control of the EU CAP, using semantic enrichment and reasoning to support farmers in monitoring their crops, insurance companies to assess the risk in a specific AOI, and public agencies that monitor the sustainability of rural areas as a consequence of climate change.

References

[1] B. McBride, "The resource description framework (RDF) and its vocabulary description language RDFs," in *Handbook on Ontologies*. Berlin, Germany: Springer, 2004, pp. 51–65.

[2] N. Kolbe, S. Kubler, J. Robert, Y. Le Traon, and A. Zaslavsky, "Linked vocabulary recommendation tools for Internet of Things: A survey," *ACM Comput. Surv.*, vol. 51, no. 6, pp. 1–31, 2019.

[3] M. Koubarakis *et al.*, "Managing big, linked, and open earth-observation data: Using the TELEIOS/LEO software stack," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 3, pp. 23–37, Sep. 2016.

[4] Y. Zhu *et al.*, "Multidimensional and quantitative interlinking approach for linked geospatial data," *Int. J. Digit. Earth*, vol. 10, no. 9, pp. 923–943, 2017.

[5] M. Koubarakis, K. Bereta, G. Papadakis, D. Savva, and G. Stamoulis, "Big, linked geospatial data and its applications in earth observation," *IEEE Internet Comput.*, vol. 21, no. 4, pp. 87–91, Jul. 2017.

[6] G. Pe'Er *et al.*, "A greener path for the EU common agricultural policy," *Science*, vol. 365, no. 6452, pp. 449–451, 2019.

[7] V. Lebourgeois, S. Dupuy, É. Vintrou, M. Ameline, S. Butler, and A. Bégué, "A combined random forest and OBIA classification scheme for mapping smallholder agriculture at different nomenclature levels using multisource data (simulated Sentinel-2 time series, VHRS and DEM)," *Remote Sens.*, vol. 9, no. 3, 2017, Art. no. 259.

[8] V. Sitokonstantinou *et al.*, "Scalable parcel-based crop identification scheme using Sentinel-2 data time-series for the monitoring of the common agricultural policy," *Remote Sens.*, vol. 10, no. 6, 2018, Art. no. 911.

[9] J. Brinkhoff, J. Vardanega, and A. J. Robson, "Land cover classification of nine perennial crops using Sentinel-1 and-2 data," *Remote Sens.*, vol. 12, no. 1, 2020, Art. no. 96.

[10] M. Arias, M. A. Campo-Bescós, and J. Álvarez-Mozos, "Crop type mapping based on Sentinel-1 backscatter time series," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 6623–6626.

[11] S. Feng, J. Zhao, T. Liu, H. Zhang, Z. Zhang, and X. Guo, "Crop type identification and mapping using machine learning algorithms and Sentinel-2 time series data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3295–3306, Sep. 2019.

[12] E. Akbari, A. D. Boloorani, N. N. Samany, S. Hamzeh, S. Soufizadeh, and S. Pignatti, "Crop mapping using random forest and particle swarm optimization based on multi-temporal Sentinel-2," *Remote Sens.*, vol. 12, no. 9, 2020, Art. no. 1449.

[13] Y. Tatiana Solano-Correa, F. Bovolo, and L. Bruzzone, "A semi-supervised crop-type classification based on Sentinel-2 NDVI satellite image time series and phenological parameters," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 457–460.

[14] S. S. Ray and Neetu, "Exploring machine learning classification algorithms for crop classification using Sentinel 2 data," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. XLII-3/W6, pp. 573–578, 2019.

[15] J. Singh, A. Mahapatra, S. Basu, and B. Banerjee, "Assessment of Sentinel-1 and Sentinel-2 satellite imagery for crop classification in indian region during Kharif and Rabi crop cycles," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 3720–3723.

[16] V. Mazzia, A. Khaliq, and M. Chiaberge, "Improvement in land cover and crop classification based on temporal features learning from Sentinel-2 data using recurrent-convolutional neural network (R-CNN)," *Appl. Sci.*, vol. 10, no. 1, 2020, Art. no. 238.

[17] J. Zhang, Y. He, L. Yuan, P. Liu, X. Zhou, and Y. Huang, "Machine learning-based spectral library for crop classification and status monitoring," *Agronomy*, vol. 9, no. 9, 2019, Art. no. 496.

[18] D. L. McGuinness and F. Van Harmelen, "Owl web ontology language overview," *W3C Recommendation*, vol. 10, no. 10, p. 2004, 2004.

[19] P. Patel-Schneider and J. Siméon, "The Yin/Yang web: XML syntax and RDF semantics," in *Proc. 11th Int. Conf. World Wide Web*, 2002, pp. 443–453.

[20] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data: The story so far," in *Semantic Services, Interoperability and Web Applications: Emerging Concepts*. Hershey, PA, USA: IGI Global, 2011, pp. 205–227.

[21] J. Pérez, M. Arenas, and C. Gutierrez, "Semantics and complexity of SPARQL," *ACM Trans. Database Syst.*, vol. 34, no. 3, pp. 1–45, 2009.

[22] M. Koubarakis and K. Kyzirakos, "Modeling and querying metadata in the semantic sensor web: The model stRDF and the query language stSPARQL," in *Proc. Extended Semantic Web Conf.*, 2010, pp. 425–439.

[23] K. Kyzirakos *et al.*, "GeoTriples: Transforming geospatial data into RDF graphs using R2RML and RML mappings," *J. Web Semantics*, vol. 52, pp. 16–32, 2018.

[24] A. I. Maarala, X. Su, and J. Riekki, "Semantic reasoning for context-aware Internet of Things applications," *IEEE Internet Things J.*, vol. 4, no. 2, pp. 461–473, Apr. 2017.

[25] F. Baader, I. Horrocks, and U. Sattler, "Description logics," *Found. Artif. Intell.*, vol. 3, pp. 135–179, 2008.

[26] H. V. Dam, J. Engberg, and H. Gerzymisch-Arbogast, *Knowledge Systems and Translation*, vol. 7. Berlin, Germany: Walter de Gruyter, 2011.

[27] C. Wang, W. Wang, and N. Chen, "Building an ontology for hydrologic monitoring," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 6232–6234.

[28] M. Masmoudi *et al.*, "Predicat: A semantic service-oriented platform for data interoperability and linking in earth observation and disaster prediction," in *Proc. IEEE 11th Conf. Service-Oriented Comput. Appl.*, 2018, pp. 194–201.

[29] G. Song, M. Wang, X. Ying, R. Yang, and B. Zhang, "Study on precision agriculture knowledge presentation with ontology," *AASRI Proc.*, vol. 3, pp. 732–738, 2012.

[30] M. T. Maliappis, "Applying an agricultural ontology to web-based applications," *Int. J. Metadata, Semantics Ontologies*, vol. 4, no. 1/2, pp. 133–140, 2009.

[31] S. Hu, H. Wang, C. She, and J. Wang, "AgOnt: Ontology for agriculture Internet of Things," in *Proc. Int. Conf. Comput. Comput. Technol. Agriculture*, 2010, pp. 131–137.

[32] Y. Su, R. Wang, P. Chen, Y. Wei, C. Li, and Y. Hu, "Agricultural ontology based feature optimization for agricultural text clustering," *J. Integrative Agriculture*, vol. 11, no. 5, pp. 752–759, 2012.

[33] S. S. Durbha and R. L. King, "Semantics-enabled framework for knowledge discovery from earth observation data archives," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 11, pp. 2563–2572, Nov. 2005.

[34] G. Jaiswal and A. P. Agrawal, "Comparative analysis of relational and graph databases," *IOSR J. Eng.*, vol. 3, no. 8, pp. 2250–3021, 2013.

[35] G. Meditskos and N. Bassiliades, "Rule-based owl ontology reasoning systems: Implementations, strengths, and weaknesses," in *Handbook of Research on Emerging Rule-Based Languages and Technologies: Open Solutions and Approaches*. Hershey, PA, USA: IGI Global, 2009, pp. 124–148.

[36] K. Kyzirakos, M. Karpathiotakis, and M. Koubarakis, "Strabon: A semantic geospatial DBMS," in *Proc. Int. Semantic Web Conf.*, 2012, pp. 295–311.

[37] R. H. Güting, "GraphDB: Modeling and querying graphs in databases," in *Proc. Very Large Data Bases*, 1994, vol. 94, pp. 12–15.

[38] P. Bellini and P. Nesi, "Performance assessment of RDF graph databases for smart city services," *J. Vis. Lang. Comput.*, vol. 45, pp. 24–38, 2018.

[39] O. Erling and I. Mikhailov, "RDF support in the Virtuoso DBMS," in *Networked Knowledge-Networked Media*. New York, NY, USA: Springer, 2009, pp. 7–24.

[40] K. Bereta, G. Xiao, and M. Koubarakis, "Ontop-spatial: Ontop of geospatial databases," *J. Web Semantics*, vol. 58, 2019, Art. no. 100514.

[41] J. Yang and G. Li, "Earth observation data integration and opening system," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 5461–5484.

[42] E. Kontopoulos *et al.*, "Ontology-based representation of crisis management procedures for climate events," in *Proc. Int. Conf. Inf. Syst. Crisis Response Manage.*, 2018, pp. 1064–1073.

[43] N. Audebert, B. Le Saux, and S. Lefèvre, "Joint learning from earth observation and OpenStreetMap data to get faster better semantic maps," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. Workshops*, 2017, pp. 67–75.

[44] N. Audebert, B. Le Saux, and S. Lefèvrey, "Fusion of heterogeneous data in convolutional networks for urban semantic labeling," in *Proc. IEEE Joint Urban Remote Sens. Event*, 2017, pp. 1–4.

[45] W. Yao, D. Marmanis, and M. Datcu, "Semantic segmentation using deep neural networks for SAR and optical image pairs," in *Proc. Big Data Space*, 2017, pp. 1–4.

[46] D. Hilbring *et al.*, "Harmonizing data collection in an ontology for a risk management platform," in *Proc. Int. Conf. Inform. Environ. Protection*, 2018, pp. 1–6.

[47] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 67, pp. 93–104, 2012.

[48] B.-C. Gao, "NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space," *Remote Sens. Environ.*, vol. 58, no. 3, pp. 257–266, 1996.

[49] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multiclass classification by pairwise coupling," *J. Mach. Learn. Res.*, vol. 5, pp. 975–1005, 2004.

[50] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation," in *Proc. Australas. Joint Conf. Artif. Intell.*, 2006, pp. 1015–1021.

Maria Rousi received the B.Sc. degree in applied informatics from the University of Macedonia, Thessaloniki, Greece, in 2016, and the M.Sc. degree in internet and world wide web from the Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2018.

Since August 2018, she has been a Research Assistant with the Centre for Research and Technology Hellas, Information Technologies Institute, Thermi-Thessaloniki, Greece. Her research interests include information retrieval and analysis on the web, with a particular focus on the knowledge representation and discovery on the semantic web.

Vasileios Sitokonstantinou received the B.Eng. degree in electrical and electronic engineering from Imperial College London, London, U.K., in 2013, the M.Sc. degree in wireless and optical communications from the University College London, London, U.K., in 2014, and the second M.Sc. degree in space science, technology and applications from the University of Peloponnese, Tripoli, Greece, in collaboration with the National Observatory of Athens, Athens, Greece, in 2017.

He is currently a Research Associate with the Institute of Astronomy, Astrophysics, Space Applications and Remote Sensing, National Observatory of Athens, leading a team on agriculture monitoring using big earth observation data and artificial intelligence.

Georgios Meditskos received the B.Sc. degree in informatics and the M.Sc. degree in information systems, and the Ph.D. degree in informatics for his dissertation on "Semantic Web Service Discovery and Ontology Reasoning using Entailment Rules" from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2004, 2007, and 2009, respectively.

Since January 2012, he has been a Postdoctoral Research Fellow with the Information Technologies Institute, Center for Research and Technology Hellas, Thermi-Thessaloniki, Greece. During the last eight years, he has participated in numerous European and national research projects and he is the author of more than 80 publications in refereed journals and international conferences. His research interests revolve around AI, and more specifically on knowledge representation and reasoning in the semantic web (RDF/OWL, rule-based ontology reasoning, combination of rules and ontologies), semantic web services (discovery, composition), and context-based multi-sensor reasoning and fusion in pervasive environments. He has been also involved in the development of semantic interpretation frameworks for the high-level integration, analysis and preservation of heterogeneous contextual information in the healthcare domain.

Ioannis Papoutsis received the Diploma in electrical and computer engineering from the National Technical University of Athens (NTUA), Zografou, Greece, in 2002, the M.S. degree in technologies for broadband communication from the Department of Electronic and Electrical Engineering, University College London, London, U.K., in 2003, and the Ph.D. degree in remote sensing from the School of Rural and Surveying Engineering, NTUA, in 2014.

He is currently an Associate Researcher with the Institute for Astronomy, Astrophysics, Space Applications, and Remote Sensing, National Observatory of Athens, Athens, Greece. His research interests include exploitation, management and processing of big satellite data, and the use of artificial intelligence for earth observation applications.

Ilias Gialampoukidis received the bachelor's degree in mathematics, M.Sc. degree in statistics and modeling, and the Ph.D. degree in mathematics with a special interest in applied mathematics, time series analysis, stochastic modelling, and network analytics, from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2009, 2001, and 2014, respectively.

He is an Interdisciplinary Postdoctoral Researcher with Centre for Research and Technology Hellas, Information Technologies Institute, Thermi-Thessaloniki, Greece, with research interests that involve knowledge discovery and information retrieval, big data analytics, early and late multimodal fusion, supervised (deep) and unsupervised learning, and social media network analytics.

Alkiviadis Koukos received the M.Eng. degree in electrical and computer engineering and the M.Sc. degree in data science and machine learning from the National Technical University of Athens (NTUA), Zografou, Greece, in 2018 and 2020, respectively.

He is a Data Scientist with the Institute of Astronomy, Astrophysics, Space Applications and Remote Sensing, National Observatory of Athens, Athens, Greece, focusing on agriculture monitoring using machine learning techniques and big earth observation data.

Vassilia Karathanassi received the M.Eng. degree in rural and surveying engineering from the National Technical University of Athens, Athens, Greece, in 1984, the D.E.A. "Amanegement - Urbanisme - Geographie" degree from University Paris IV, Paris, France, in 1986, and the Ph.D. degree in remote sensing received from National Technical University of Athens, Remote Sensing Laboratory, Athens, Greece, in 1990.

She has been a Professor with the School of Rural and Surveying Engineering, National Technical University of Athens, Zografou, Greece, since 2000. Her research interests include multispectral and hyperspectral remote sensing, SAR image processing and differential interferometry. Her published research work includes more than 40 papers in peer reviewed scientific journals, more than 60 papers in international conference proceedings, 33 research project reports, and one book chapter. According to Google Scholar, there are more than 1400 citations in her published work.

Dr. Karathanassi is a Reviewer in multiple high impact journals. She has participated in 25 European and national excellence/competitive research projects as a PI, coordinator, or team member.

Thanassis Drivas received the B.Sc. degree in computer science from Athens University of Economics and Business, Athens, Greece, in 2007, and the M.Sc. degree in space science, technology and applications from the University of Peloponnese, Tripoli, Greece, in collaboration with the National Observatory of Athens, Athens, Greece, in 2017.

He has long professional experience in teaching computer science and programming, STEM and robotics in private schools and colleges. At the same time, he has worked as a Web and Back-End Developer in several companies like SaraLee Hellas. Since 2017, he has been a Research Associate with the Institute of Astronomy, Astrophysics, Space Applications and Remote Sensing, National Observatory of Athens, focusing on the design-development of geospatial applications using machine learning techniques and earth observation big data.

Stefanos Vrochidis received the Diploma degree in electrical engineering from Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2000, the M.Sc. degree in radio frequency communication systems from the University of Southampton, Southampton, U.K., in 2001, and the Ph.D. degree in electronic engineering from the Queen Mary University of London, London, U.K., in 2013.

He is currently a Senior Researcher (Grade C) with the Multimedia Knowledge and Social Media Analytics Lab, Centre for Research and Technology Hellas, Information Technologies Institute, Thermi-Thessaloniki, Greece, and the Head of the Multimodal Data Fusion and Analytics Group. His research interests include multimedia analysis and retrieval, multimodal fusion, computer vision, multimodal analytics based on artificial intelligence, semantic web, as well as media and arts, environmental and security applications.

Dr. Vrochidis has participated in more than 35 European and national projects dealing with semantic analysis, multimodal fusion, and retrieval of multimedia information. He has been a member of the organization team of several conferences and workshops relevant to the aforementioned research areas. He has edited two books and authored more than 200 related scientific journal, conference, and book chapter publications. He was a Reviewer in several international journals and as technical program committee in well reputed conferences and workshops.

Charalampos Kontoes (Haris) received the Ph.D. degree in remote sensing of the environment, in 1992, and completed his doctoral studies holding a grant from the European Commission in the Space Applications Institute of the Joint Research Centre (JRC), Ispra, Italy (Environmental Mapping Group, JRC).

He is the Research Director of the Institute for Astronomy and Astrophysics, Space Applications and Remote Sensing, National Observatory of Athens (NOA), Athens, Greece, and Scientific Responsible of the Center for Earth Observation Research and Satellite Remote Sensing BEYOND of NOA (www.beyond-eocenter.eu). Since 1992, he has been managing earth observation operational and research projects at national, European Union, and global level, focusing on sectors as risk assessment and mitigation, risk monitoring and management, environmental resource management, agriculture, land use/land cover, health/epidemiology, renewable energy, in various territorial contexts and scales.

Ioannis Kompatsiaris (Senior Member, IEEE) received the Diploma degree and the Ph.D. degree in coding of stereoscopic image sequences using 3D model from the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, in Thessaloniki, Greece, in 1996 and 2001, respectively.

He is a Research Director of the Centre for Research and Technology Hellas, Information Technologies Institute, Thermi-Thessaloniki, Greece, the Head of the Multimedia Knowledge and Social Media Analytics Laboratory and Deputy Director of the Institute. His research interests include image and video analysis, big data and social media analytics, semantics, human computer interfaces (AR and BCI), eHealth, security and culture applications. He is the coauthor of 129 papers in refereed journals, 46 book chapters, 8 patents, and more than 420 papers in international conferences.

Dr. Kompatsiaris has participated in 59 national and European research programs, in 15 of which he has been the Project Coordinator and in 41 the Principal Investigator, since 2001. He has also been the PI in four research collaborations with industry (Motorola U.S. and U.K.). He has been the Co-Chair of various international conferences and workshops including the 13th IEEE Image, Video, and Multidimensional Signal Processing (2018) Workshop and has served as a Regular Reviewer, Associate and Guest Editor for a number of journals and conferences currently being an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING. He is a member of the Scientific Advisory Board of the CHIST-ERA funding program, an elected member of the IEEE Image, Video and Multidimensional Signal Processing - Technical Committee, and a member of ACM.