Journal Title 0 (0) 1
IOS Press

A Semantic Meta-Model for Data Integration and Exploitation in Precision Agriculture and Livestock Farming

Dimitris Zeginis [a,b,*], Evangelos Kalampokis [a,b], Raul Palma [c], Rob Atkinson [d] and Konstantinos Tarabanis [a,b]

[a] *Information Technologies Institute, Centre for Research & Technology – Hellas, Greece*
[b] *Information Systems Lab, University of Macedonia, Greece*
E-mails: zeginis@uom.edu.gr, ekal@uom.edu.gr, kat@uom.edu.gr
[c] *Poznań Supercomputing and Networking Center - PSNC, Poland*
E-mail: rpalma@man.poznan.pl
[d] *Open Geospatial Consortium*
E-mail: ratkinson@ogc.org

Abstract. At the domains of agriculture and livestock farming a huge amount of data are produced through numerous heterogeneous sources including sensor data, weather/climate data, statistical and government data, drone/satellite imagery, video, and maps. This plethora of data can be used at precision agriculture and precision livestock farming in order to provide predictive insights in farming operations, drive real-time operational decisions, and redesign business processes. The predictive power of the data can be further boosted if data from diverse sources are integrated and processed together, thus providing more unexplored insights. However, the exploitation and integration of agricultural data is not straightforward since they: i) cannot be easily discovered across the numerous heterogeneous sources and ii) use different structural and naming conventions hindering their interoperability. The aim of this paper is to firstly study the characteristics of agricultural data and the user requirements related to data modeling and processing from nine real cases at the agriculture, livestock farming and aquaculture domains and then propose a semantic meta-model that is based on W3C standards (DCAT, PROV-O and QB vocabulary) in order to enable the definition of metadata that facilitate the discovery, exploration, integration and accessing of data in the domain.

Keywords: Semantic model, Metadata, Data Integration, Precision agriculture, Precision livestock farming, DCAT

1. Introduction and motivation

Today, the agriculture and livestock farming sectors produce huge amounts of heterogeneous data [1, 2]. Examples of these data include IoT sensor data measuring soil electrical conductivity [3], drone/satellite imagery data presenting the state of crops at different parts of a field [4] and video data monitoring animal behaviour [5, 6]. Precision agriculture and precision livestock farming make intense use of these data accompanied with other data such as weather data, statistical and government data, and maps to gain insights, make predictions, drive real-time operational decisions and redesign business processes regarding, e.g., disease [7], pests and weeds control, fertilization, harvest, irrigation, and seeding [8], as well as animal behaviour recognition [5, 6, 9] and animal body weight measurement [10].

The full potential of precision agriculture and precision livestock farming can be explored if data from diverse heterogeneous sources are processed together, thus providing more unexplored insights. For example,

*Corresponding author. E-mail: zeginis@uom.edu.gr.

the processing of satellite imagery and weather data about the same time period and geographic area enables the prediction of crop growth and yield as well as the identification the best harvesting period [11]. However, these data are usually fragmented and come from heterogeneous sources [12, 13] using different standards, structures and units (e.g., streaming sensor data and weather data from meteorological institutes). Thus, due to their heterogeneity and fragmentation it is not straightforward to identify agriculture or livestock farming data to be processed together (e.g. identify weather data for a specific time period and geographic area) and if done so, it is difficult to actually combine them [14].

Towards this direction this paper proposes a semantic meta-model for heterogeneous data integration and exploitation in precision agriculture and precision livestock farming. The model re-uses W3C standards such as DCAT model [15], PROV-O [16] and QB vocabulary [17] in order to boost data interoperability and data sharing in the domain. The aim of the proposed model is to serve as a common reference model for: i) the alignment of agricultural and livestock farming data in order to tackle heterogeneity issues and ii) the semantic annotation of data in order to facilitate data identification and exploration. In particular, the model can be used to create metadata (e.g. spatial/temporal coverage of the data, structure of the data) that support:

- The on-demand data discovery and exploration. For example enable the identification of data that address specific criteria e.g. data of area X at the time frame [2018 - 2019] that contain sensor-generated data related to soya yield cultivation.
- Data interoperability by aligning/mapping the structure of the data to the model. For example, align the dimensions (e.g. time, geography) and measures (e.g. temperature, weight) of the data to the dimensions/measures defined by the model.
- Data access. The model contains structural and access metadata that enable the querying/accessing of data based on the metadata. For example enable the formulation and execution of queries (e.g. SQL) based on the metadata.

The rest of the paper is organised as follows, section 2 presents background and related work, section 3 presents the methodology followed in this paper in order to create the model, section 4 presents the model specification including the relevant stakeholders, data and requirements, section 5 presents the conceptualisation and implementation of the model, section 6 applies and demonstrates the use of the model at agricultural and livestock farming data and finally section 7 concludes the paper and discusses interesting points.

2. Background and related work

This section presents existing vocabularies that enable the definition of metadata about datasets (section 2.2) and domain (agriculture and livestock farming) specific models and controlled vocabularies that can be used to populate the metadata (section 2.3). The presented vocabularies and models follow the semantic web and linked data philosophy and principles presented at section 2.1.

2.1. Semantic web and linked data

The term semantic web refers to an extension of the current "web of documents" in order to build a "web of linked data". In order to achieve the vision of linked data, a set of principles have been proposed to “publish data on the web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external datasets, and can in turn be linked to from external datasets” [18]. Semantic Web and linked data are empowered by technologies such as RDF [19], OWL [20], SPARQL [21] and SHACL [22] in order to create vocabularies and publish/access/validate data on the web.

More specifically, RDF is a W3C standard model for data interchange on the Web. It uses URIs to name things and their relationships that are expressed as "triples" <X, Y, Z> (e.g. <John, Cultivates, Soya>). RDF has many serialization formats such as RDF/XML, Turtle and JSON-LD. The examples in this paper use the Turtle format that is more "human readable". The W3C Web Ontology Language (OWL) is a logic-based language that can model ontologies supporting inference. OWL extends RDF in order to express complex knowledge about things. SPARQL is a language to express queries across diverse RDF data sources. Finally, Shapes Constraint Language (SHACL) is a language for validating RDF data against a set of conditions.

The vocabularies and models presented in the following sections as well as the model proposed in this paper adhere to the semantic web and linked data concepts, principles and technologies.

2.2. Metadata vocabularies

Data Catalog Vocabulary (DCAT) [15] is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web. By using DCAT to define metadata of data catalogs, publishers increase discoverability and enable applications to easily consume metadata from multiple catalogs. It further enables decentralized publishing of catalogs and facilitates federated dataset searches across sites. DCAT defines three main classes: i) the dcat:Catalog that represents the catalog, ii) the dcat:Dataset that represents a dataset in a catalog and iii) the dcat:Distribution that represents an accessible form of a dataset (e.g. downloadable file, API).

Diverse extensions of DCAT have been proposed to cover the needs of different domains. The proposed extensions include: i) the DCAT Application Profile (DCAT-AP) [23] for data portals in Europe that enables cross-data portal search for datasets, ii) the StatDCAT Application Profile (StatDCAT-AP) [24] that aims at providing a commonly-agreed dissemination vocabulary for statistical open data and iii) the GeoDCAT-AP [25] Application Profile that aims at making geospatial information better searchable across borders and sectors.

The EU INSPIRE directive [26] proposes specific rules to make spatial data available covering diverse aspects including the definition of metadata and the interoperability of spatial data. The spatial data considered under the directive is extensive and includes among others the agriculture domain. Regarding metadata, INSPIRE defines the elements that should be used for documenting a dataset. The INSPIRE metadata elements have been aligned to ISO 19115/ISO 19119 as well as with DCAT.

The RDF Data Cube (QB) vocabulary [17] is a W3C standard for publishing statistical data on the Web using the linked data principles. The core class of the vocabulary is the qb:DataSet that represents a cube, which comprises a set of dimensions - qb:Dimension Property (e.g. time, geography), measures - qb:Measure Property (e.g. temperature, weight) and attributes - qb: AttributeProperty (e.g. unit of measurement). The declaration of the dimensions, attributes, and measures is done at the qb:DataStructureDefinition, which defines the structure of the cube. Usually the values of the dimensions, attributes, and measures are populated using predefined code lists. A set of best practices for using the QB vocabulary is also proposed [27] and is considered at the definition of the proposed model in this is paper.

The PROV-O [16] W3C recommendation describes provenance in terms of relationships between three main types of concepts: i) prov:Entity, which represents physical, digital, or other types of things, ii) prov:Activity, which occur over time and can generate (prov:wasGeneratedBy) entities and iii) prov:Agent, which are responsible for activities occurring (prov:was AssociatedWith) and entities existing (prov:wasAtt ributedTo).

The Observation and Measurements (O&M) conceptual model [28] is an OGC specification for observations and features involved in sampling when making observations. These provide models for the exchange of information describing observation acts and their results.

The Semantic Sensor Network (SSN) ontology [29] is used for describing actuators, sensors and their observations, the involved procedures, the studied features of interest, the samples used to do so, and the observed properties. SSN is a domain-independent model that supports a wide range of use cases e.g. satellite imagery, large-scale scientific monitoring, observation-driven ontology engineering, and the Web of Things. Such use cases are implemented with domain specific subclasses of abstract concepts such as "procedures" and "observableProperties". SSN is aligned with PROV-O and O&M models.

2.3. Domain specific models and controlled vocabularies

The Agricultural and Aquaculture Facilities (AF) [30] model is used to define geographical information of entities under the Agriculture and Aquaculture scope. AF is based on the Activity Complex model [31] proposed by INSPIRE that avoids specific thematic connotations e.g. AF contain concepts such as the "holding", "site", "location" etc. However, AF also includes an extended model to represent domain information e.g. plots, agri-buildings, installations, irrigation and drainage, farm animals and animal health.

FOODIE ontology [32] provides an application vocabulary that enable the definition of data and metadata related to farm management. The main concept of the ontology is "Plot" that is a continuous area of agricultural land with one type of crop species, cultivated by one user applying one farming mode. One lower level than Plot is the "Management Zone", which enables a more precise description of the land characteristics in fine-grained areas.

AGROVOC [33] is a controlled vocabulary defined by the Food and Agriculture Organization (FAO) of the United Nations that includes concepts related to food, nutrition, agriculture, forestry, fisheries, techniques of plant cultivation etc. More than half of the concepts fall under the top concept "organism", which confirms how AGROVOC is largely oriented towards the agricultural and livestock farming sectors.

Animal Trait and Animal Health ontologies is a collection of ontologies related to livestock farming: i) ATOL (Animal Trait Ontology for Livestock) [34] is an ontology of characteristics defining phenotypes of livestock in their environment, ii) EOL (Environment Ontology for Livestock) [35] is an ontology that describes environmental conditions of livestock farms e.g. feeding modalities, the environment, the structure of livestock farms and iii) AHOL (Animal Health Ontology for Livestock) [36] is an ontology that describes production diseases (associated symptoms, the affected organism, the organism causing the disease).

OWL-Time [37] is an ontology for describing the temporal properties of resources in any data. The ontology provides a vocabulary for expressing information about durations, and about temporal position including date-time information. The main class of the ontology is the "TemporalEntity" that has two subclasses: "Interval" and "Instant". Intervals have some extent, while Instants are point-like in that they have no interior points.

QUDT units ontology [38] provides semantic specifications for units of measure, quantity kind, dimensions and data types. QUDT semantics are based on dimensional analysis expressed in the OWL Web Ontology Language (OWL). The dimensional approach relates each unit to a system of base units.

3. Methodology

The development of the proposed model adopts a "meet-in-the-middle" approach [39] where concepts emerge both in a bottom-up (i.e. analyzing the domain) and top-down (i.e. analyze and integrate existing ontologies, vocabularies and models) fashion. The methodology (figure 1) focuses on a collaborative development that entails the active engagement of domain experts i.e. agriculture and livestock farming stakeholders. More specifically the methodology comprises of the following phases:

- **Specification** (Section 4). At this step the model stakeholders are defined (section 4.1), the domain data are identified and analysed (section 4.2) and the user functional and non-functional data are identified (section 4.3). Towards this direction, a set of interviews and surveys with the stakeholders have been performed.
- **Conceptualization** (Section 5.1). This step identifies the concepts and relations of the semantic model. Concepts emerge by analyzing the domain (e.g. existing agricultural and livestock farming data, functional/non-functional requirements) considering also existing ontologies, models and vocabularies like the ones presented at sections 2.2 and 2.3. The aim is to reuse as much as possible these vocabularies to build the model. The output of this phase is a conceptual model comprising all identified concepts and relationships in a human-readable form (e.g. class diagram). For simplicity and space reasons the figure of the conceptual model is not presented in the paper since it is similar and includes less information than the implementation model (see figure 2) e.g. the mapping of the concepts to existing vocabularies is not part of the conceptual model.
- **Formalization and Implementation** (Sections 5.1 and 5.2). This step transforms the conceptual model into a formal or semi-computable model. The identified concepts are mapped to existing standards and vocabularies (section 5.1). This activity builds a computable model in an ontology language. The implementation language selected is RDF. Additionaly, this step presents the implementation choices in order to formalise the model as a reusable profile of DCAT (section 5.2).
- **Application** (Section 6). This step adopts a twofold approach: i) apply the model at agricultural and livestock farming data to see if it is sufficient and covers all the needs and ii) use the model by performing queries for data retrieval and by developing applications for data exploration.

The methodology has been applied in two rounds within the EU funded project CYBELE [1] that deals with precision agriculture and precision livestock farming. There was enough time between the two rounds enabling the stakeholder to use the model. The feedback of the first round was exploited in order to extract new requirements and improve the model. This

[1] https://www.cybele-project.eu/

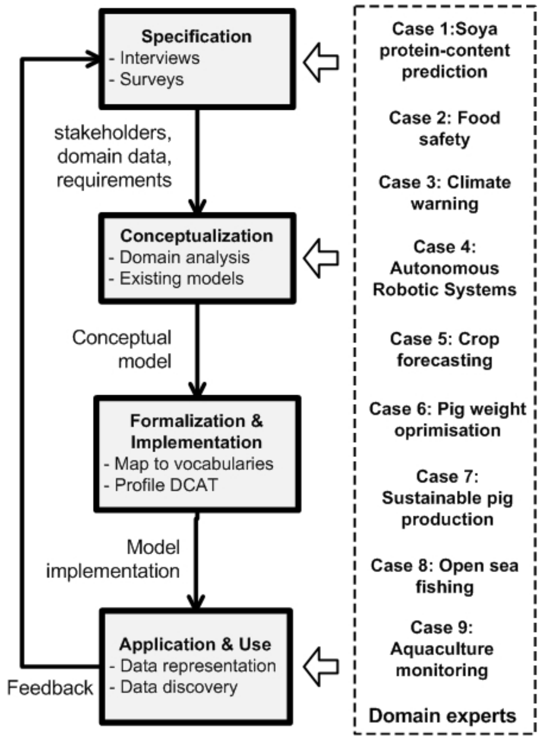

Fig. 1. Methodology to develop the model

paper presents the final model including the complete set of stakeholders, data and requirements used at the second iteration of the methodology. Nine real cases were explored at the domains of agriculture, livestock farming and aquaculture in order to collect relevant data and user requirements. The cases are further described at section 4.1. The model was applied at these cases and also acted as a central artifact that drove the development of data-centric applications as part of an ecosystem.

4. Specification: Identifying Agricultural metadata requirements

This section identifies the cases and requirements that should be covered by the model. Towards this directions it presents the model stakeholders through nine agriculture, livestock farming and aquaculture cases (section 4.1), it identifies and analyses domain data i.e. data exploited at the nine cases (section 4.2) and identifies the functional and non-functional requirements (section 4.3).

4.1. Stakeholders

This section briefly presents the nine real cases that are supported by the model. These cases are further examined in the next sections in order to identify the data and requirements that need to be covered.

Case 1: Organic Soya yield and protein-content prediction. The aim of this case is to predict yield and protein-content maps based on satellite imagery and additional information concerning electromagnetic soil scans, drone images and sensory data. This will enable e.g. farmers to separate A class from B class soya and sell them separately at higher prices.

Case 2: Food Safety. The aim of this case is to assist food safety experts with advanced data analysis and risk prediction for the food supply chain. This includes the prediction of food recalls in the supply chain of various products and the prediction of prices for agricultural products. The data exploited at the case include food recalls, border rejections, lab testing data, fraud cases, production data and trade data.

Case 3: Climate Services for Organic Fruit Production. The aim of this case is to develop an end-to-end frost and hail early warning system for the protection of organic fruit from extreme weather events by mitigating/preventing damages and injuries in sensitive crops. The data exploited at the case include agro-climatic data, weather forecasts, phenology data and satellite images.

Case 4: Autonomous Robotic Systems within Arable Frameworks. The aims of this case are: i) to support decisions and plans for harvest or fertilizer applications and ii) to provide yield predictions and information of yield distributions across the field. The data exploited at the case include soil chemical analysis, hypespectral images of soil, drone images and satellite images.

Case 5: Optimizing computations for crop yield forecasting. The aim of this case is to predict farmer's crop yields at a high spatial resolution in order to improve the quantity of the produced yields, improve resources management, decrease production costs and decrease yield losses. The data exploited at the case include crop data, soil data, weather forecasts, historic yield data, parcel specific data and satellite images.

Case 6: Pig Weighing Optimisation. The aim of the case is to estimate and track the live weight of pigs in a pen based on video images. Convolutional neural networks (CNNs) will be developed that will use video images of pigs in a pen in order to measure the mean weight and the standard deviation of the weight of pigs in the pen. The growth curve estimated by the CNNs in previously developed models for early warning of diarrhoea will also be incorporated.

Case 7: Sustainable Pig Production. The aims of this case are to: i) improve carcass and meat quality by measuring pork water holding capacity and using hyper spectral imaging to measure meat quality and ii) improve health, welfare and production by improving warning systems and detect anomalies. The data exploited at the case include climate sensor data, feed/medicine registration, human observations, weighting data, flow measurements, slaughterhouse data (sensors for carcass grading, infrared sensor, hyperspectral images).

Case 8: Open Sea Fishing. The aim of this case is to monitor fish stocks and manage commercial fishing vessels in order to prevent overfishing. Three different data sources will be exploited: i) vessels' electronic logbooks that comprise daily landing data of commercial fish stocks, ii) a board system that collects location data from a GPS logger, weights of the landings per species group, vessel speed, fuel consumption and tractive power and iii) a visual-based processing of the catch using an RGB camera. The above data will also be combined with meteorological data, habitat maps and environmental data from satellite based imaginary systems.

Case 9: Aquaculture monitoring and feeding optimization. The aim of this case is to develop an efficient feed management system including the estimation of fish growth, cage alignment and dead fish identification. The data exploited at the case include aerial images of fish farms taken from drones, weather information and sensor measurements (mainly related to oxygen and current speed).

4.2. *Precision agriculture & livestock farming data*

Based on the descriptions of the nine cases presented at subsection 4.1, it is obvious that different types of agricultural and livestock farming data from diverse sources are available. Such data can be grouped into the following broader categories:

- **Sensor data** are continuously collected through dedicated hardware and produce spatiotemporal measurements e.g. measure the temperature and humidity at a specific location and time. Sensors produce large volume of data since measurements are repeated regularly (e.g. every 1 minute).
- **Earth observations** e.g. satellite images, drone aerial images, hyper-spectral images, RGB images. This type of data can produce huge volume of spatiotemporal data since they provide high resolution images of the earth.
- **Video** e.g. video data from pig pens to monitor pigs behaviour or RBG video data on conveyor belt to automatically sort the fish catches. This type also produces huge volume of data.
- **Crowd-sourced data** and human observations are collected through manual measurements and inspections (e.g. health inspection at livestock farms). Usually these data are not of big volume, but need to be combined with other data e.g. sensor data, to support decision making at precision agriculture and precision livestock farming.
- **Forecasts** e.g. for weather, prices, production. These data are also of spatiotemporal nature and usually are not of big volume. They can be combined with other data to facilitate decision making at precision agriculture and precision livestock farming.
- **Maps** can be combined with other data to provide easily interpretable results and visualization e.g. show sensor measurements on a map.
- **Statistical and government data** e.g. daily landing data of commercial fish stocks, price data, trade data, food recalls, border rejections, fraud cases. Usually these data are not of big volume, but need to be combined with other data e.g. sensor data, to support decision making at precision agriculture and precision livestock farming.
- **Location data** e.g. location data of the fishing fleet from a Vessel Monitoring System. Like the sensor data these data can be of large volume since the location of the vessels is updated regularly

Based on their structure, the above categories can be separated at two main clusters. The first cluster includes the structured data (sensor data, crowd-sourced data, forecasts, statistical data, location data) that have a well-defined structure and are available as e.g. CSV files, JSON files, relational databases. The second cluster includes the unstructured data (earth observations, videos, maps) that do not have a structure and are available as e.g. video or image files. Each cluster has different characteristics that need to be expressed by the model, but they also share some common. More specifically, both clusters have some generic characteristics that need to be expressed (e.g. title, licence, format, geographical coverage) that will facilitate the exploration of the data (e.g. find data for a specific geographical area). However, regarding the structural data, there is also a need to describe their structure enabling in this way a more fine-grained data exploration (e.g. find data that measure the temperature in a specific geographical area) and also facilitate the access/querying of the data (e.g. formulate SQL queries).

The data can also be classified based on their domain of coverage to domain specific and cross domain. The domain specific include: i) Agriculture data e.g. crop data, protein content, soil chemical analysis, yield maps, ii) Food safety data e.g. food recalls, border rejections, fraud cases, production data, lab testing data, iii) Livestock farming data e.g. pig weight, livestock health, slaughterhouse data, iv) Fishing data e.g. fish behaviour data, landing data of fish stocks, v) Aquaculture data e.g. water info data (temperature, quality, current speed). While the cross-domain data include the: i) Climate and weather data e.g. temperature, wind speed, humidity and ii) Satellite & aerial image data.

4.3. User Requirements

This section presents the requirements for the proposed semantic model in terms of functional requirements (i.e. activities that can be facilitated by the use of the model) and non-functional requirements (i.e. aspects/concepts that need to be covered by the model). In order to identify the requirements a set of interviews and surveys with the stakeholders have been performed within the CYBELE project that led to the definition of specific use cases and functional/non-functional requirements. Part of the use cases and the identified requirements are related to the semantic model (other areas related to precision agriculture and precision livestock farming e.g. HPC and big data, are also covered but are out of the scope of this paper).

The identified functional requirements that are related to the model can be summarized to the following:

- Find and locate data based on diverse criteria e.g. time/geographical coverage, structure.
- Use and query various types of data from different data sources including e.g. geospatial data, time series.
- Combine data from different and heterogeneous data sources.

The identification of the non-functional requirements of the model is based on the thorough analysis of the data used at the nine cases considering diverse characteristics/properties (e.g. data format, language, theme, temporal/spatial coverage, structure). The characteristics considered for the analysis are based on the DCAT model. Table 1 presents the result of the analysis (e.g. what formats are used by the data? what is the temporal/spatial coverage of the data?) and the stemming non-functional requirements. The following paragraphs present some interesting insights of the data analysis.

The analyzed data cover a broad range of thematic areas including specific cultivations (e.g. soya), cultivation activities (e.g. applying fertilizer, harvesting), livestock farming activities (e.g. feeding), fish farming activities (e.g. catches) and weather/climate data.

The temporal and spatial coverages of the data are expressed in different granularities. The time can be expressed in years or days while the granulatity of the spatial dimension can be the country (e.g. Spain), group of countries (e.g. European Union), land geographic area (e.g. Central America), sea area (e.g North East Atlantic), specified coordinates (e.g. a point), specific area (e.g. a polygon), specific site (e.g. farm site).

The structure of the data use a broad range of dimensions (e.g. time, geography) and measures (e.g. temperature, weight). The measures may vary on the aggregation function (e.g. min/max/average temperature) or on the measurement subject (e.g. water/air/soil temperature). Thus, a broad range of measurement variations may occur e.g. min water temperature, max air temperature.

Finally, the datasets can be created as a result of activities such as observation (e.g. through sensors, autonomous vehicles, human inspection, satellites, aerial drones), forecasting (e.g. weather, price, production) and fusion of pre-existing datasets.

5. Model conceptualization and implementation

This section presents the conceptualization and implementation (section 5.1) of the model by identifying the concepts and their relations and then mapping them to existing standard and vocabularies. The section also presents the decision choices for the formalisation of the model as a reusable profile of DCAT (section 5.2) and details about the publication of the model using the FAIR principles (section 5.3).

Table 1

Non-functional requirements as a result of agricultural and livestock farming data analysis.

Property	Analysis result	Non-functional requirement
Format	Datasets are available in multiple structured (e.g. CSV,JSON, XLSX, GeoTIFF) and unstructured (e.g. TXT, JPG, PNG) data formats.	Express the dataset format. Support structured and unstructured data formats.
Language	Datasets are expressed in diverse languages (e.g. English, Greek, French).	Express the dataset language. Support datasets in different languages.
Thematic area	A broad range of areas are covered: i) specific cultivation e.g. soya, ii) cultivation activities e.g. applying fertilizer, harvesting, iii) livestock farming activities e.g. feeding, iv) fish farming activities e.g. catches and v) weather/climate data.	Express the dataset thematic area. Support diverse thematic areas related to agriculture, livestock farming, fish farming and weather/-climate data.
Update rate	Datasets are updated in diverse rates e.g. every minute, daily, weekly, monthly, quarterly, seasonal, annually, sexennial, never.	Express the dataset update rate. Support diverse dataset update rates (every minute, daily, weekly, monthly, quarterly, seasonal, annually, sexennial, never).
Temporal coverage	The temporal coverage is expressed as a time range having a min/max value. The granularity is the year (e.g. 2010 – 2020) or the day (e.g. 1/1/2020 – 31/10/2020).	Express the dataset temporal coverage. Support temporal coverages in diverse granularities (e.g. year, date).
Minimum temporal step	The datasets use diverse temporal steps (minimum time period resolvable in the dataset) e.g. day, 10 days, plant growth stages, streams.	Express the dataset minimum temporal step. Support a wide range of temporal steps within dataset.
Spatial coverage	The spatial coverage is expressed in diverse granularities including the country (e.g. Spain), group of countries (e.g. European Union), land geographic areas (e.g. Central America), sea areas (e.g North East Atlantic), specified coordinates, specific polygon, specific sites (e.g. farm site).	Express the dataset spatial coverage. Support a wide range of spatial coverages including predefined areas (e.g. countries, group of countries, land/sea geographic areas) and dynamic areas (e.g. coordinates, polygons).
Data standard	Datasets follow diverse data standards models, schema or ontologies e.g. Data Cube vocabulary, INSPIRE, EPSG 28992 coordinate system	Express the standard the dataset is based on. Support diverse standards for datasets.
Structure (dimension, measure)	The structured data contain a broad range of dimensions (e.g. time, geography) and measures (e.g. temperature, weight). The measures may vary on the aggregation function (e.g. min/max/average temperature) or on the measurement subject (e.g. water/air/soil temperature).	Express the structure of datasets including their dimensions and measures. Support diverse dimensions and measures related to agriculture and livestock/fish farming.
Units of measure	Measurements are expressed in a variety on units e.g. meter/centimetre/millimetre for distances, gram/kg/tonne for weight, celsius/fahrenheit for temperature, degree/rad for angles.	Express the unit of measure. Support a broad range of units.
License	Datas are available under different licenses including e.g. Creative Commons, custom.	Express the dataset license. Support diverse types of license.
Activity	The datasets are created as a result of an activities including: i) Observation e.g. through sensors, autonomous vehicles, human inspection, satellites, aerial drones and ii) Forecasting e.g. weather, price, production, iii) The fusion of two pre-existing datasets	Express the activity that created the dataset. Support diverse types of activities including observation, forecasting and fata fusion.
Data service (how the data is accessible)	Datasets are available through APIs, Download URLs, SPARQL endpoints, Databases. At CYBELE all datasets will be stored at a central platform and made available through a database.	Express the service that makes the dataset available. Support different types of services. Facilitate the querying of datasets available through databases.
Publisher	Datasets are published by diverse organizations e.g. the CYBELE demo partners	Express the dataset publisher. Support diverse publishers.
Issued, modified	Datasets are issued/modified at specific points in time.	Express the date/time the datasets are issued/modified.
Spatial resolution	Datasets have different minimum spatial separation of items within the dataset e.g. 30 meters.	Express the dataset minimum spatial separation.
Access rights	Datasets are available as public/open data, private data, and restricted data.	Express the dataset access rights. Support diverse types of access rights (e.g. open, private, restricted).
Byte size	Datasets have different sizes in terms of bytes	Express the dataset byte size.

5.1. The model conceptualization and implementation

Based on the functional and non functional requirements (section 4) four main categories of metadata need to be covered by the model:

- Descriptive metadata: describe the overall features of datasets (e.g. title, language). This category of metadata facilitates the discovery and exploration of data.
- Structural metadata: describe the schema and internal structure of a dataset (e.g. dimensions/measures). This category of metadata facilitates the fine-grained data discovery and exploration based on their structure and enables data interoperability by aligning the structure of the data to the model. This kind of metadata are only applicable to structured data (e.g. CSV, JSON files).
- Provenance metadata: provide information about the origins of the data (e.g. the activity that generated the data). This kind of metadata facilitate the discovery and exoloration of data based on their origin (e.g. search for sensor-generated or crowdsourced data).
- Access metadata: describe the way to access the data (e.g. Database/Table where the dataset is stored). This information accompanied with the structural metadata facilitates the formulation of queries (e.g. SQL).

These categories are aligned with the categories proposed by the W3C Data on the Web Best Practices [40].

The functional and non-functional requirements presented at section 4 were further specialized to competency aspects that should be considered at the design of the model. The competency aspects define what should the model be able to express. Table 2 presents these competency aspects, the model concepts that occur for each of them and their mapping to existing vocabularies. The vocabularies used for the mapping are the DCAT, Dublin Core Metadata Terms (dct) [41], RDFS [42], PROV-O, QB and SDMX[2]. In some cases no relevant concepts where identified at existing vocabularies, thus new concepts are defined using the prefix "cybele".

The main classes of the model are: i) the Dataset, that is a collection of data published by a specific publisher (person or organization), ii) the Catalog/Repository, that is a collection of metadata about Datasets, iii) the Activity, that represents the way/method the Dataset was generated involving Agents e.g. human, sensor, iv) the Distribution, that represents an accessible form of a Dataset e.g. downloadable file, Data Service, Database and v) the Structure, that includes structural information of Datasets (Dimensions and Measures). The complete model is depicted at figure 2

Each of these main classes have extra properties. For example the Dataset has properties including the Theme (e.g. pig farming), Language, Issuing/Modification date, Update frequency, Spatial/Temporal coverage, Spatial/Temporal resolution, Access rights, Standard and Web page. These properties of the dataset can be used for data discovery and exploration e.g. identify data about pig farming.

In order to define the structure of a dataset including the Dimensions (e.g. time) and Measures (e.g. temperature), the model uses the classes qb:DimensionProperty and qb:MeasureProperty and the relevant properties qb:dimension and qb:measure. However, the qb:dimen sion and qb:measure properties have the qb:Component Specification as domain the and cannot be used directly at the dcat:Dataset. Thus, the model includes also two auxiliary classes from the QB vocabulary the qb:ComponentSpecification and the qb:DataStructure Definition, the later representing the structure of the dataset. The association of the dataset with its structure is done through the property qb:structure. The use of this property on an individual entails that it is a member of the class qb:Dataset. So, the datasets should also be members (isa) of the class qb:Dataset.

The definition of the dataset's structure accompanied with the definition of access metadata for datasets which are distributed through a database (cybele:Database, cybele:Table) enables the formulation/execution of queries (e.g. SQL). More specifically, the formulation/execution of queries can be done as follows:

- Information about the database where the dataset is stored and the way to connect to the databases is provided through the cybele:Database.
- Information about the actual table where the dataset is stored is provided through the cybele:Table.
- Information about the table fields is provided through the structure of the dataset (qb:Data StructureDefinition). In this case, the label of the qb:DimensionProperty or the qb:MeasureProperty should be identical with the corresponding fields

[2] https://joinup.ec.europa.eu/collection/linked-open-vocabularies/solution/sdmx-rdf-vocabulary

Table 2

Model competency aspects, concepts and mapping to vocabularies

	Competency aspect	Concept	Map to vocabularies
Descriptive metadata	Search for datasets registered at a catalog	Dataset, Catalog	dcat:Dataset, dcat:Catalog
	Dataset contains data about a specific cultivation (e.g. soya) or livestock	Theme (e.g. cultivation, livestock)	dcat:theme
	Dataset contains data about a specific theme e.g. weather data, price data	Theme (e.g. weather, price)	
	Dataset contains data about cultivation/farming activities e.g. irrigation, applying fertilizer, harvesting, feeding, weighing, slaughter	Theme (e.g. the farming activity)	
	Dataset is published by an organization	Publisher	dcat:publisher
	Dataset contains data in a specific language e.g. English	Language	dct:language
	Dataset is issued/modified at e.g. 1/1/2019	Issuing, modification date	dct:issued, dct:modified
	Dataset is updated e.g. monthly	Update frequency	dct:accrualPeriodicity
	Dataset contains data with temporal coverage e.g. [1/1/2017 - 31/12/2017]	Temporal coverage	dct:temporal
	Dataset contains measurements with temporal spacing e.g. one hour (measurements are repeated every one hour)	Temporal resolution	dcat:temporal Resolution
	Dataset contains data with spatial coverage e.g. an area defined by a polygon	Spatial coverage	dct:spatial
	Dataset contains measurements minimum distance between items e.g. 30 meters	Spatial resolution	dcat:spatialResolution InMeters
	Dataset has specific access rights e.g. open access, restricted	Access rights	dct:accessRights
	Dataset conforms to a model/schema/ontology/view/profile e.g. Data Cube vocabulary	Standard	dct:conformsTo
	Dataset is accessed through a web page	Web page	dcat:landingPage
Access metadata	Dataset is distributed under a specific license	Distribution, license	dcat:Distribution, dct:license
	Dataset is distributed in a specific format e.g. CSV, XML, Json	Format	dcat:mediaType
	Dataset's distribution is e.g. 100 MB	Size	dcat:byteSize
	Dataset distribution can be downloaded through a URL	Download URL	dcat:downloadURL
	Dataset is distributed through a service e.g API	Data service	dcat:DataService
	Data service is accessed through an endpoint URL	Endpoint URL	dcat:endpointURL
	Datasets distribution is accessible through a Data Base	Database	cybele:Database
	Database is accessible through a URL using some connection info e.g. username, password	Connection URL	cybele:connection String
	Dataset is accessible through a specific database table	Table, table name	cybele:Table, cybele:tableName
Structural	Dataset measures e.g. NDVI	Measurement	qb:MeasureProperty.
	Dataset has specific dimensions e.g. time, geography	Dimension	qb:DimensionProperty
	Dataset's dimension/measure has a specific range type	Range	rdfs:range
	Dataset uses a unit of measure e.g. prices in euro	Unit of measure	sdmx:unitMeasure
Provenance	Dataset is the result of an activity that involves e.g. sensors, humans, satellites	Activity, agent (human, hardware)	prov:Activity, prov:Agent
	Dataset is the result of an aggregation activity of other data (e.g. raw data)		

of the table where the dataset is stored. For example if the label of a qb:DimensionProperty is “Geography” then the field of the database table that stores the geographical information should also have the same name.

In case that a dataset is not distributed through a database but is available as a download file or through an API, other modeling options should be followed. Specifically, if the data is available as a download file, then the property dcat:downloadURL of the class dcat:Distribution should be used, while if the dataset is distributed through an API then the property dcat:endpointURL of the class dcat:DataService should be used.

Finally, the provenance metadata of the dataset are provided through the classes prov:Activity and prov:Agent. The first defining the activity that generated the dataset, while the secong defining the agent that is involved at the generation of the dataset. The association of these classes with the dataset requires that the dataset is also member (isa) of the class prov:Entity

5.2. Formalising the model as a profile of DCAT

Whenever a “general-purpose” standard model (e.g. DCAT) is applied to an application domain (e.g. agriculture), a set of implementation choices are made. This set of choices can be explicitly named and described as an "application profile" of the standard model. This paper proposes a semantic model implemented as an application profile of DCAT for the agricultural domain. The specific implementation choices of the proposed model are related to the :

- Optional elements of DCAT that are mandatory at the proposed model.
- Additional constraints and rules that are applied to the proposed model.
- Controlled vocabularies that should be used as the range of properties.

For example, the use of the qb:DataStructureDefinition to describe data structure is not part of DCAT, but allowable under the semantic modelling approach. To effectively profile DCAT Datasets to allow the qb:sructure property to link to qb:DataStructureDefinition, a SHACL constraint was used:

```
dcat:Dataset sh:property [
    a sh:PropertyShape ;
    sh:path qb:structure ;
    sh:class qb:DataStructureDefinition ;
    sh:maxCardinality 1 ] .
```

An important aspect of defining a profile is to be able to make statements about use of a generic property, without fundamentally altering the semantics of that property in a way which would preclude two different models from co-existing. By making statements using SHACL, the domain of application of the SHACL rules can be managed independently of the original model (i.e DCAT) enabling the combination of models without introducing conflicting statements.

The previous example could have been expressed in OWL using `dcat:Dataset owl:equivalent Class qb:Dataset` or by `dcat:Dataset dfs:subClassOf qb:Dataset`. This however would mean that another system with different expectations about DCAT datasets would potentially fall foul of expectations about qb:Datasets. It would also require application of OWL and/or RDFS reasoning across the entire combined model, whereas the SHACL constraints can be applied only when circumstances apply (via either additional filter conditions or business logic about what rules to apply, and when).

Additionally, regarding the controlled vocabularies, AGROVOC [33] is proposed to be used for the dcat:theme of datasets. For example, the URI `<http://aims.fao.org/aos/agrovoc/c_14477>` correspond to "soya beans" and can be used as a theme of datasets related to soya cultivation.

The proposed model is an application profile of DCAT, but also re-uses other standard vocabularies including the QB vocabulary, PROV-O, SKOS, FOAF and DC Terms. The relationship of the model to other standards is expressed using the PROFILES vocabulary [43] as shown in the following RDF code.

```
<http://w3id.org/cybele/model>
  prof:isProfileOf
    <https://w3id.org/cybele/proxy/cube>,
    <https://w3id.org/cybele/proxy/dcat>,
    <https://w3id.org/cybele/proxy/dcterms>,
    <https://w3id.org/cybele/proxy/foaf>,
    <https://w3id.org/cybele/proxy/prov>,
    <https://w3id.org/cybele/proxy/skos> ;
  prof:isTransitiveProfileOf
    <http://purl.org/dc/terms>,
    <http://purl.org/linked-data/cube>,
    <http://www.w3.org/2004/02/skos/core>,
    <http://www.w3.org/ns/dcat>,
    <http://www.w3.org/ns/prov>,
    <http://xmlns.com/foaf/0.1> ;
```

Note that each of the vocabularies used by the model have a "proxy" profile generated for it (e.g. https://w3id.org/cybele/proxy/foaf) in order to allow

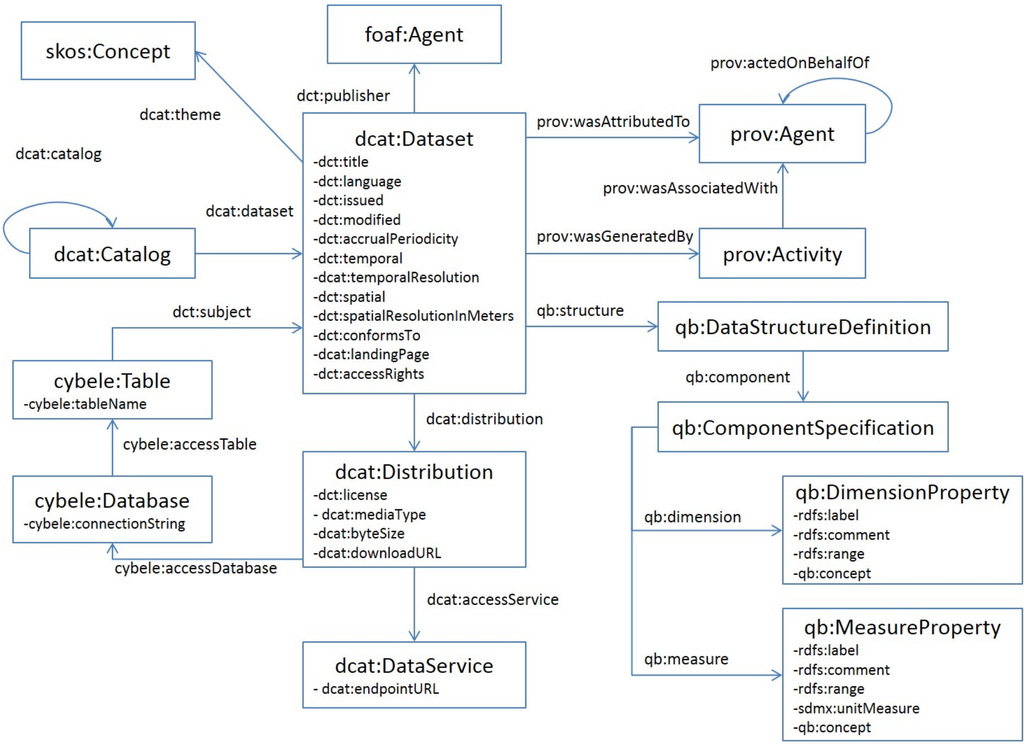

Fig. 2. Semantic Model for Precision Agriculture and Precision Livestock Farming

the referenced models to be accessed with a set of implementation resources describing which parts of these models are used. The next section further discusses the principles of open and stable access to model resources.

5.3. *Publication of the model using FAIR principles*

The FAIR principles ("Findable, Accessible, Interoperable, Reusable") provide a basis for model publication requirements to support interoperability in a domain over time. Reusability is dependent on stability, and a stable home for such models is predicated on the semantic relationships between the model and the available governance and resourcing models in the domain. In this case, the semantic model proposed in this paper describes interoperability constraints for spatio-temporal data and conveniently falls under the interest domain of the Open Geospatial Consortium (OGC). This has represented a valid challenge to the OGC's own development of Linked Data publication approaches, and hence has been used to drive additional Linked Data infrastructure to support the nexus between OGC activities and the wider community of implementers of OGC specifications. The result was the development of the "OGC Definitions Server" that is a semantic publishing framework applying the FAIR principles.

The community node of the "OGC Definitions Server" provides a catalog of profiles and is maintained as an interoperable adjunct to its normative Linked Data published content. Interaction with the server is seamless via URI redirection, while browse and search options, and potentially various UI theme styles, for each node is distinct.

The semantic model proposed in this paper and the supporting implementation resources (e.g. "proxy" profiles of existing vocabularies) are published through the "OGC Definitions Server" using a combination of Linked Data technologies, APIs and relevant open standards. The published model is made available through a dereferenceable URI:

```
http://w3id.org/cybele/model
```

Various alternative representation of the model are available and can be accessed via the URI of the model using "Content negotiation by profile" (e.g. http://w3 id.org/cybele/model?_profile=owl). The alternative model representations are:

- The "SKOS Concept Scheme" contains a representation of all the models' concepts in a hierarchical way.
- The "Class Diagram" provides an overview diagram of the model.
- The "JSON-LD" representation of the model.
- The "JSON Schema" view of the classes and properties of the model.
- The "FeatureType" representation that uses the ISO19109 meta-model.
- The "OWL" representation that contains all profile-specific statements and all the statements from the imported vocabularies needed for a self-contained model.
- The "SHACL" representation contains a set of constraints describing the profiled classes in the model. This provides additional detail about constraints often only available implicitly.

The publication of the model according to the FAIR principles addresses the most common barrier to reuse of common models in that they are not necessarily directly accessible, and discovery of the resources needed to implement them is often difficult. It also provides an opportunity to easily compare the model with other application domains and determine the extent of shared usage and hence interoperability.

Finally, in order to understand how similar agriculture and livestock farming datasets may be compared and be interoperable is a complicated task. It requires comparison of the details of each dataset, which may be serialised in many different forms, and distributed across a range of imported or implicitly referenced resources. To solve this problem and facilitate the comparison of agriculture and livestock farming datasets, the datasets could declare conformance to the proposed model. Thus, a statement such as <DATASET_URI> dct:conformsTo <http://w3id.org/cybele/model> can be made to more accurately describe the interoperability of datasets.

6. Application of the Model

This section applies and uses the proposed model in the context of the nine cases for precision agriculture and livestock farming presented at section 4.1. Section 6.1 introduces some agriculture and livestock farming datasets and demonstrates how these can be described using the semantic model. Section 6.2 presents examples for data retrieval using SPARQL queries over data generated and stored based on the model, while section 6.3 introduces a semantic REST API that is built on top of the model and facilitates the data exploration.

The semantic REST API is part of an ecosystem of applications built on top of the model within the CYBELE project. The model acts as a central artifact that drives the development of applications related to: i) data alignment: a component has been developed that maps the structure of the datasets to the model in order to facilitate their interoperability, ii) annotation: a component has been developed that enables the semi-automatic generation of metadata like the those presented at section 6.1 and iii) discovery/exploration: a component has been developed that uses the REST API to explore the datasets and generate SQL queries based on the metadata. The complete description of the above components and their architecture is out of the scope of this paper.

6.1. Applying the model at domain datasets

Dataset 1 (case 7): a time series dataset about the feeding process in a farm in Belgium for the day 02/02/2019. The dataset is published by the Flanders Research Institute for Agriculture, Fisheries and Food (ILVO) on 24/03/2021 as a CSV file and is distributed through a relational database (the CSV is inserted at the database using a dedicated software component). The data about individual pigs are automatically collected through sensors (namely sensorA100 and sensorC205) located at the feeding station of a barn. The feeding station contains a platform weighing scale to measure the pig body weight, a feeding tray with weighing scale that measures the content of the feeding tray and an RFID antenna that registers the identity of the pig in the feeding station by scanning its RFID ear tag. One line of data is provided per feeding visit.

The structure of the dataset comprises the following fields: i) Location - the feeding location inside the barn e.g. pen number, feeding station number, ii) Responder - the pig RFID identifier), iii) Animal number - a short pig identifier iv) Life number - an optional additional pig identifier, v) Time - the timestamp of the beginning of the feeding visit, vi) Duration - the duration of the feeding visit in seconds, vii) Feed intake - the amount of feed provided during feeding visit (precision 1 gram) and viii) Weight - median of pig weights measured during feeding visit (precision 500 grams).

Listing 1 presents some general descriptive metadata for the dataset e.g. the publisher, date issued, title, language. Some of the metadata use primitive values (e.g. date, string) while other use URIs to link to concepts defined at controlled vocabularies (dct:spatial, dct:language, dct:accessRights, dct:accrualPeriodicity) or to link to complex structures within the dataset description (e.g. dct:temporal, qb:structure). The controlled vocabularies used are defined by the European Union[3]. The examples use the prefixes `eu-lang`, `eu-country`, `eu-accessRight`, `eu-license`, `eu-fileType` for the code lists defined by the European Union that correspond to the language, country, access rights, license and file type accordingly.

Listing 1: Dataset 1 general description

```
<https://w3id.org/cybele/dataset/id/d1>
 a dcat:Dataset, qb:DataSet;
 dct:issued "2021-03-24"^^xsd:date ;
 dct:publisher
     <https://w3id.org/cybele/org/ILVO> ;
 dct:title "Feed intake records" ;
 dct:description "Data about the feeding
     process in a Belgian farm" ;
 dct:language eu-lang:ENG ;
 dct:temporal
     <https://w3id.org/cybele/temp/id/d1>;
 dct:spatial eu-country:BEL ;
 dct:accrualPeriodicity
     <http://purl.org/cld/freq/irregular> ;
 dct:accessRights eu-accessRigh:PUBLIC ;
 dcat:distribution
     <https://w3id.org/cybele/dist/id/d1> ;
 qb:structure
     <https://w3id.org/cybele/dsd/id/d1>;
 prov:wasGeneratedBy
     <https://w3id.org/cybele/activity/id/d1>;
 dct:subject
     <http://w3id.org/cybele/tbl/d1> .
```

Listing 2 defines the temporal coverage of Dataset 1, that is the whole day 02/02/2019. The temporal coverage is defined as a time space (dct:PeriodOfTime) with a specific beginning (dcat:startDate) and end (dcat:endDate).

Listing 2: Dataset 1 temporal coverage

```
<https://w3id.org/cybele/temp/id/d1>
   a dct:PeriodOfTime ;
   dcat:endDate
       "2019-02-02T23:50:00"^^xsd:dateTime ;
   dcat:startDate
       "2019-02-02T00:00:00"^^xsd:dateTime .
```

[3]https://op.europa.eu/en/web/eu-vocabularies/authority-tables

Listing 3 defines the Dataset 1 provenance information, namely the sensor reading activity that generated it, along with the sensors (sosa:Sensor) used to carry out this activity. Note that sosa:Sensor is defined at the SSN ontology as a subclass of prov:Agent.

Listing 3: Dataset 1 provenance information

```
<https://w3id.org/cybele/activity/id/d1>
   a prov:Activity ;
   rdfs:label "sensor reading activity" ;
   prov:wasAssociatedWith [
     a prov:Agent, sosa:Sensor;
     rdfs:label "sensorA100"] ;
   prov:wasAssociatedWith [
     a prov:Agent, sosa:Sensor;
     rdfs:label "sensorC205"] .
```

The dataset is stored and made accessible through a database. Listing 4 defines information about the distribution (e.g. dct:license, dcat:mediaType), the specific database (cybele:Database) and table (cybele:Table) where the dataset is stored. For the license and format values the description uses predefined controlled vocabularies defined by the European Union. The definition of all the available tables at the database is done through the property cybele:accessTable. The example uses only one table, however more tables may exist. The association of the dataset to the specific table where a dataset is stored is done through the property dct:subject of the dcat:Dataset.

Listing 4: Dataset 1 distribution through a Database

```
<https://w3id.org/cybele/dist/id/d1>
  a dcat:Distribution ;
  dct:license eu-license:APACHE_2_0;
  dcat:mediaType eu-fileType:CSV ;
  dcat:byteSize "20".
  cybele:accessDatabase
    <https://w3id.org/cybele/db/leanxcale>.
<https://w3id.org/cybele/db/leanxcale> a
    cybele:Database ;
  rdfs:label "LeanXcale DB" ;
  cybele:connectionString
    "jdbc:leanxcale://127.0.0.1:1529/CYBELE;
     user=cybele" ;
  cybele:accessTable
      <http://w3id.org/cybele/tbl/d1>.
<http://w3id.org/cybele/tbl/d1>
  a cybele:Table ;
  cybele:tableName "VELOS_ILVO".
```

Listing 5 defines partially the structure of the dataset. For space reasons, only two dimensions (location and time) and one measure (feed intake) are defined. The definition of the rest dimensions/measures can be done in a similar way. In order to enable the generation and execution of SQL queries based on the metadata, the label (rdfs:label) of the dimension/measure should be the same as the corresponding field at the database table where the dataset is stored. For example, the field that stores the feed intake at the database table should use the name "feed_intake". The range (rdfs:range) of dimensions and measure (e.g. xsd:dateTime, xsd:int) is also defined in order to facilitate the processing of the data. Additionally, the unit of measure (sdmx-attribute:unitMeasure) is defined in order to facilitate the transformation of data to other metric systems or granularities (e.g. transorm kilograms to grams or pounds).

Listing 5: Dataset 1 structure (partial)

```
<https://w3id.org/cybele/dsd/id/d1>
  a qb:DataStructureDefinition;
  qb:component
   [qb:dimension
    <https://w3id.org/cybele/dim/location>];
   [qb:dimension
    <https://w3id.org/cybele/dim/time>];
   [qb:measure
    <https://w3id.org/cybele/mes/feedIntake>].
<https://w3id.org/cybele/dim/location>
  a qb:DimensionProperty, rdf:Property ;
  rdfs:subPropertyOf sdmx-dimension:refArea;
  rdfs:label "location";
  rdfs:range xsd:int;
  rdfs:comment "The location";
  qb:concept sdmx-concept:refArea.
<https://w3id.org/cybele/dim/time>
  a qb:DimensionProperty, rdf:Property ;
  rdfs:subPropertyOf
       sdmx-dimension:timePeriod;
  rdfs:label "time";
  rdfs:range xsd:dateTime;
  rdfs:comment "The time" ;
  qb:concept sdmx-concept:timePeriod .
<https://w3id.org/cybele/mes/feedIntake>
  a qb:MeasureProperty, rdf:Property ;
  sdmx-attribute:unitMeasure unit:GM;
  rdfs:label "feed_intake" ;
  rdfs:range xsd:int;
  rdfs:comment "Amount of feed provided".
```

Dataset 2 (case 3): contains weather data at the granularity of day captured by an Agri-climatic station in the province of Valencia, during the period 01/02/2016-29/02/2016. The dataset is published by GMV Aerospace and Defence company in Spain on 23/12/2019 and is updated monthly. The dataset is available in CSV format and contains dimension and measures including: i) IdProvincia (province identifier), ii) Time of observation, iii) TempMax (max air temperature) and iv) HoraTempMax (time of max air temperature). The dataset contain more measures (e.g. the moisture, wind speed) that are omitted for space reasons.

Listing 6 presents some general descriptive metadata for the dataset. In this dataset the temporal coverage (February 2016) is defined using the reference.data.gov.uk Time Interval vocabulary (the example uses the prefix `uk-month`). The temporal resolution (i.e. the minimum time resolvable in the dataset) is the "day" that is expressed through the string "P1D" which follows a specific format defined by the xsd:duration. The update rate (dct:accrualPeriodicity) of the dataset-as-a-whole is the "month" that is expressed using the Dublin Core Collection Description Frequency Vocabulary[4]. The spatial coverage is defined using the NUTS classification[5] (Nomenclature of territorial units for statistics) that is a hierarchical system for dividing up the economic territory of the EU (Valencia corresponds to the code ES52).

Listing 6: Dataset 2 description

```
<https://w3id.org/cybele/dataset/id/d2>
  a dcat:Dataset, qb:DataSet;
  dct:issued "2019-12-23"^^xsd:date ;
  dct:publisher
       <https://w3id.org/cybele/org/id/GMV> ;
  dct:title "Weather data from the province
       of Valencia - February 2016" ;
  dct:temporal uk-month:2016-02;
  dct:spatial nuts:ES52 ;
  dcat:temporalResolution "P1D" ;
  dct:accrualPeriodicity
       <http://purl.org/cld/freq/monthly> ;
  qb:structure
       <https://w3id.org/cybele/dsd/id/d2>.
```

Listing 7 defines partially the structure of the dataset. Two dimensions, one measure and one attribute are used. Note that the time dimension URI is the same as the one used at listing 5 in order to facilitate the interoperability between datasets. However, the URIs used for the geographical dimensions are not

[4]https://www.dublincore.org/specifications/dublin-core/collection-description/frequency/

[5]https://ec.europa.eu/eurostat/web/nuts

the same since they have different ranges (rdfs:range). The range of the dimension "IdProvincia" is the "Unitary Authority" as defined by the "Administrative geography and civil voting area ontology"[6], while the range of the dimension "location" at listing 5 is int (e.g. feeding station number). In this case, interoperability is achieve indirectly by defining both dimensions (IdProvincia, location) as sub-properties of the sdmx-dimension:refArea and by associating them with the same concept (sdmx-concept:refArea). The attribute "HoraTempMax" is used to define the time of max air temperature. In a similar way attributes can be used to associate the aggregated values of measures such as min or max within a specific time period e.g. day, month, year.

Listing 7: Dataset 2 structure (partial)

```
<https://w3id.org/cybele/dsd/id/d2>
 a qb:DataStructureDefinition;
 qb:component
  [qb:dimension
   <https://w3id.org/cybele/dim/IdProvincia>];
  [qb:dimension
   <https://w3id.org/cybele/dim/time> ];
  [qb:measure
   <https://w3id.org/cybele/mes/TempMax>];
  [qb:attribute
   <https://w3id.org/cybele/att/HoraTempMax>].
<https://w3id.org/cybele/dim/IdProvincia>
   a qb:DimensionProperty, rdf:Property ;
   rdfs:subPropertyOf sdmx-dimension:refArea;
   rdfs:label "IdProvincia" ;
   rdfs:range admingeo:UnitaryAuthority>;
   rdfs:comment "The reference province" ;
   qb:concept sdmx-concept:refArea .
<https://w3id.org/cybele/mes/TempMax>
   a qb:MeasureProperty, rdf:Property ;
   rdfs:subPropertyOf sdmx-measure:obsValue;
   sdmx-attribute:unitMeasure unit:DEG_C;
   rdfs:label "TempMax" ;
   rdfs:range xsd:double;
   rdfs:comment "Max Air Temperature".
<https://w3id.org/cybele/att/HoraTempMax>
   a qb:AttributeProperty ;
   rdfs:label "HoraTempMax"@en ;
   rdfs:range xsd:date;
   rdfs:comment "Time of max air temp.".
```

Dataset 3 (case 5): the organization Wageningen University & Research (WUR) has published a dataset with the output of the crop simulation WOFOST (WOrld FOod STudies) model that performs quantitative analysis of the growth and production of annual field crops. The model has run for a specific field (with ID 7977923) with maize crop in Netherlands for the year 2018. In this example the spatial coverage corresponds to the geometry of the field 7977923, which is provided in a WKT format [44]. WKT supports geospatial positions expressed in coordinate reference systems. The description of the dataset structure is omitted, as it is similar to the examples above.

[6]http://data.ordnancesurvey.co.uk/ontology/admingeo/

Listing 8 provides the general description of the dataset. The temporal coverage of the dataset (2018) is defined using the reference.data.gov.uk controlled vocabulary, while the spatial coverage is defined at the listing 9. The dataset is the output of the WOFOST model, thus it conforms (dct:conformsTo) to the structure formally defined by the model.

Listing 8: Dataset 3 description

```
<https://w3id.org/cybele/dataset/d3>
 a dcat:Dataset, qb:DataSet;
 dct:publisher
     <https://w3id.org/cybele/org/id/WUR> ;
 dct:title "WOFOST output dataset - 2018" ;
 dct:temporal uk-year:2018;
 dct:conformsTo
   <https://wofost.readthedocs.io/en/latest/>;
 dct:spatial
     <https://w3id.org/cybele/spatial/d3> ;
 dcat:distribution
     <https://w3id.org/cybele/dist/de> ;
 qb:structure
     <https://w3id.org/cybele/dsd/id/d1>.
```

Listing 9 provides the spatial coverage of the dataset. The spatial coverage is defined as a dct:Location that in turn is associated (locn:geometry) with the WKT polygon.

Listing 9: Dataset 3 spatial information description

```
<https://w3id.org/cybele/spatial/d3>
 a dct:Location ;
 locn:geometry """MULTIPOLYGON (((
   6.083006 51.15162 , 6.082502 51.151949 ,
   6.082028 51.15225 , 6.081948 51.152313 ,
   6.081888 51.15237 , 6.081842 51.152431 ,
   6.081755 51.152895 , 6.081723 51.153143 ,
   6.081669 51.1536 , 6.081641 51.153755 ,
   6.081558 51.154111 , 6.081474 51.154572 ,
   ...)))"""^^geosparql:asWKT .
```

Dataset 4 (case 5): the organization Wageningen University & Research (WUR) has published a dataset with the Normalized Difference Vegetation Index (NDVI) for some areas in the Netherlands for the year

2018. NDVI [45] is an index produced by satellite images and can be used to estimate the density of green on an area of land. Low values of NDVI correspond to barren areas (rock, sand, snow), moderate values represent shrub and grassland, while high values indicate tropical rainforests. The NDVI values of the dataset are derived from Sentinel 2 satellite images (band 2,3 and 4) at 10m and 25m resolution. The dataset is provided as a raster image (GeoTIFF format).

Listing 10 presents the geneneral description of the dataset. An interesting remark is the use the dcat:spatialResolutionInMeters in order to define the minimum resolution of the NDVI measurements (i.e. 10m). Note that this dataset is not associated to a structure (qb:structure) since it is available as an unstructured image file.

Listing 10: Dataset 4 description

```
<https://w3id.org/cybele/dataset/d4>
 a dcat:Dataset, qb:DataSet;
 dct:issued "2020-01-20"^^xsd:date ;
 dct:publisher
   <https://w3id.org/cybele/org/id/WUR> ;
 dct:title "NDVI for netherlands - 2018" ;
 dct:language eu-lang:ENG ;
 dct:temporal uk-year:2018;
 dct:spatial eu-country:NLD ;
 dcat:temporalResolution "P1D" ;
 dcat:spatialResolutionInMeters
     "10.0"^^xsd:decimal ;
 prov:wasGeneratedBy
  <https://w3id.org/cybele/activity/id/d4>;
 dcat:distribution
  <https://w3id.org/cybele/distribution/d4>.
```

Listing 11 defines the Dataset 4 provenance information, i.e., the remote sensing process carried out by sentinel-2 satellites that produced the images from which the NDVI values were derived. Note that Sentinel-2 satellite is modeled as a sosa:Platform that can host other entities e.g. sensors.

Listing 11: Dataset 4 provenance information

```
<https://w3id.org/cybele/activity/id/d4>
  a prov:Activity ;
  rdfs:label "remote sensing activity" ;
  prov:wasAssociatedWith [
    a prov:Agent, sosa:Platform;
    rdfs:label "Sentinel-2 satellites"] .
```

6.2. *Data discovery using SPARQL queries*

This section presents some data discovery and exploration SPARQL queries that can be created based on the model and the dataset descriptions provided at the previous sub-section. The queries can be tested directly at the SPARQL endpoint that hosts the generated data: https://www.foodie-cloud.org/sparql.

Query 1: retrieves all datasets distributed in CSV format. This query can be used for data exploration e.g. for applications that can process specific file formats. The query searches at the name of the format the string "CSV". An alternative would be to search using a predefined URI from a code list e.g. `eu-fileType:CSV`

Listing 12: Query 1: Retrieves all datasets distributed in CSV format

```
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX dcat: <http://www.w3.org/ns/dcat#>
PREFIX skos:
     <http://www.w3.org/2004/02/skos/core#>
PREFIX bif: <bif:>

SELECT DISTINCT ?s
WHERE {
   ?s a dcat:Dataset.
   ?s dcat:distribution ?dist.
   ?dist dcat:mediaType ?media .
   ?media skos:prefLabel ?mediaName
   FILTER (bif:contains(?mediaName, "CSV"))
}
```

Query 2: retrieves all datasets with spatial coverage intersecting a particular polygon. This query is applicable to datasets with spatial coverage expressed as polygons (e.g. listing 9) and can be used for data exploration e.g. identify datasets for a specific area of interest.

Listing 13: Query 2: Retrieves datasets with spatial coverage intersecting a particular polygon

```
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX dcat: <http://www.w3.org/ns/dcat#>
PREFIX locn: <http://www.w3.org/ns/locn#>
PREFIX geo:
    <http://www.opengis.net/ont/geosparql#>
PREFIX geof:
    <http://www.opengis.net/def/function/
    geosparql/>
PREFIX bif: <bif:>

SELECT ?s
WHERE {
```

```
    ?s a dcat:Dataset.
    ?s dct:spatial ?spatial.
    ?spatial a dct:Location .
    ?spatial locn:geometry ?geometry .
    FILTER(geof:sfIntersects(
     bif:st_geomFromText(?geometry),
         "POLYGON((2.823460137618956
         53.35228692539731,8.140842950118955
         53.35228692539731,8.140842950118955
         50.870213151481515,2.823460137618956
         50.870213151481515,2.823460137618956
         53.35228692539731))"^^geo:wktLiteral))
         .
}
```

Query 3: retrieve all datasets that measure temperature (e.g. max/min air temperature, sea temperature) and show the name of the measure and the spatial location associated (which can be a named place or some geometry). The query searches at the comment of the measure the string "temperature" in order to identify all temperature-related measures. An alternative would be to search using a URI e.g. `https://w3id.org/cybele/mes/TempMax` but in this case only datasets that measure the max air temperature would be returned. This query can be used for fine-grained data discovery and exploration based on the dataset structure.

```
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX dcat: <http://www.w3.org/ns/dcat#>
PREFIX qb:
    <http://purl.org/linked-data/cube#>
PREFIX locn: <http://www.w3.org/ns/locn#>
PREFIX rdfs:
    <http://www.w3.org/2000/01/rdf-schema#>
PREFIX bif: <bif:>

SELECT distinct ?s ?measureName
    (COALESCE(?geometry, ?spatialName,
    ?spatial) AS ?spatialLocation)
WHERE {
   ?s a dcat:Dataset.
   ?s dct:spatial ?spatial.
   ?s qb:structure ?structure .
   ?structure qb:component ?component .
   ?component qb:measure ?measure .
   ?measure rdfs:label ?measureName .
   ?measure rdfs:comment ?measureComment .
   OPTIONAL { ?spatial skos:prefLabel
       ?spatialName .
   FILTER ( lang(?spatialName) = "en" )} .
   OPTIONAL { ?spatial locn:geometry
       ?geometry } .
   FILTER (bif:contains(?measureComment,
       "temperature"))
}
```

Query 4: retrieve all datasets that were generated by SENTINEL-2 satellites. The query searches the string "Sentinel-2" at the name of the sosa:Platform responsible for the generation of the dataset.

```
PREFIX dcat: <http://www.w3.org/ns/dcat#>
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX rdfs:
    <http://www.w3.org/2000/01/rdf-schema#>
PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX bif: <bif:>

SELECT distinct ?s
WHERE {
   ?s a dcat:Dataset.
   ?s prov:wasGeneratedBy ?activity.
   ?activity prov:wasAssociatedWith ?agent .
   ?agent a sosa:Platform .
   ?agent rdfs:label ?agentName .
   FILTER (bif:contains(?agentName,
       "Sentinel-2"))
}
```

6.3. Semantic REST API

The native language to access the RDF data generated based on the model is SPARQL. However, in order to facilitate the access and consumption of data to other components/services a REST API is also provided that returns JSON. The REST API is implemented using GRLC[7] that translates SPARQL queries stored in a Git repository[8] to a REST API on the fly.

The functionality of the API includes: i) `GET /allDatasets`: the retrieval of information about all the datasets, with full structure and including information of referenced resources, ii) `GET /getByKeyword`: the retrieval of information about datasets that contain a specific keyword (e.g. "greece") in any of their labels, iii) `GET /getDatasetStructure`: the retrieval of structural information about a particular dataset, iv) `GET /getDatasetSpatialInfo`: the retrieval of spatial information about a particular dataset and v)) `GET /getDatasetDBDistribution`: the retrieval of the database distribution information of a particular dataset.

Datasets may be registered at different catalogs (dcat:Catalog). The API supports the use of multiple catalogs enabling the retrieval of information across all of them (using federated queries) or from a specific one. For example, the API contains a method that en-

[7] http://grlc.io
[8] https://github.com/cybele-project/metadata

ables the retrieval of all information of datasets that belog to a specific catalog.

All methods of the API can be accessed and tested via the Swagger interface: http://grlc.io/api-git/cybele -project/metadata/ (At the Swagger interface the term "Testbet"is used as an alias to the catalog).

7. Conclusion and discussion

As the global population is growing, there is a push to agriculture and livestock farming domains to be more effective and efficient. Towards this direction, they make intense use of data coming from numerous heterogeneous sources (e.g. sensor data, weather/climate data, statistical & government data, drone/satellite imagery, video, and maps) in order to provide insights and drive operational decisions through precision agriculture and precision livestock farming. However, a further boost can be given to precision agriculture and precision livestock farming if data from heterogeneous sources could be exploited together.

Towards this direction, this paper studies nine real cases at the domains of agriculture, livestock farming and aquaculture in order to identify the data characteristics and user requirements at the domain. Onwards the paper proposes a semantic meta-model for the domain that is based on the W3C standards DCAT, PROV-O and QB in order to facilitate the discovery, exploration, integration and accessing of data.

Although the existing vocabularies can individually express most of the model's concepts there is no single model that can address all the requirements. For example, there is no property to associate a dcat:Dataset with the measures it contains. StatDCAT defines a property to associate only the dataset dimensions, while the QB vocabulary defines the property qb:measure that however is not applicable to dcat:Datasets. Towards this direction the paper proposes a “merging” of the DCAT and the QB vocabularies by associating a qb:DataStructureDefinition at the dataset. This merging is done through the definition of domain specific SHACL constraints that can be managed independently of the original model (i.e DCAT), thus it does not fundamentally alter the semantics of the original model.

The proposed model currently enables the definition of dimensions (qb:DimensionProperty) and measures (qb:MeasurePoroperty). The use of attributes (qb:AttributeProperty) is also possible through the QB vocabulary. The paper presents an example use of attributes (Listing 7) however this potential is not fully explored, thus the qb:AttributeProperty is not part of the proposed model but could be possibly added in the future.

In a similar way, a "merging" of DCAT with PROV-O is proposed in order to allow the definition of provenance metadata for the datasers.

Additionally the model defines concepts (cybele: Database, cybele:Table) that facilitate the access and querying of data stored at relational databases. The authors plan to explore the possibility of supporting more structured or semi-structured types of databases e.g. noSQL or Graph databases. This extension will enable the generation and execution of queries at noSQL databses (e.g. MongoDB) based on the provided metadata.

The proposed model has been tested and demonstrated within the CYBELE H2020 project at the domains of agriculture and livestock farming in order to achieve interoperability and homogenized access to data sources. However it could also be exploited at other domains that have similar data and requirements. The model extends DCAT by enabling the definition of structural metadata (association with the QB vocabulary) and provenance metadata (association with PROV-O ontology). The definition of such metadata could also be beneficial for other domains too.

Acknowledgments

Part of this work was funded by the European Commission within the H2020 Programme in the context of the project CYBELE under grant agreement no. 825355.

References

[1] R. García, J. Aguilar, M. Toro, A. Pinto and P. Rodríguez, A systematic literature review on the use of machine learning in precision livestock farming, *Computers and Electronics in Agriculture* **179** (2020), 105826. doi:https://doi.org/10.1016/j.compag.2020.105826. https://www.sciencedirect.com/science/article/pii/S0168169920317099.

[2] D.I. Patrício and R. Rieder, Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review, *Computers and Electronics in Agriculture* **153** (2018), 69–81. doi:https://doi.org/10.1016/j.compag.2018.08.001. https://www.sciencedirect.com/science/article/pii/S0168169918305829.

[3] L. Hamami and B. Nassereddine, Application of wireless sensor networks in the field of irrigation: A review, *Computers and Electronics in Agriculture* **179** (2020), 105782. doi:https://doi.org/10.1016/j.compag.2020.105782. https://www.sciencedirect.com/science/article/pii/S0168169920310528.

[4] Y. Inoue, Satellite- and drone-based remote sensing of crops and soils for smart farming – a review, *Soil Science and Plant Nutrition* **66**(6) (2020), 798–810. doi:10.1080/00380768.2020.1738899.

[5] Q. Yang and D. Xiao, A review of video-based pig behavior recognition, *Applied Animal Behaviour Science* **233** (2020), 105146. doi:https://doi.org/10.1016/j.applanim.2020.105146. https://www.sciencedirect.com/science/article/pii/S0168159120302343.

[6] C. Chen, W. Zhu and T. Norton, Behaviour recognition of pigs and cattle: Journey from computer vision to deep learning, *Computers and Electronics in Agriculture* **187** (2021), 106255. doi:https://doi.org/10.1016/j.compag.2021.106255. https://www.sciencedirect.com/science/article/pii/S0168169921002726.

[7] N. Hatton, A. Sharda, W. Schapaugh, D. Van der Merwe et al., Remote thermal infrared imaging for rapid screening of sudden death syndrome in soybean, in: *2018 ASABE Annual International Meeting*, American Society of Agricultural and Biological Engineers, 2018, p. 1.

[8] I. Cisternas, I. Velásquez, A. Caro and A. Rodríguez, Systematic literature review of implementations of precision agriculture, *Computers and Electronics in Agriculture* **176** (2020), 105626. doi:https://doi.org/10.1016/j.compag.2020.105626. https://www.sciencedirect.com/science/article/pii/S0168169920312357.

[9] D. Berckmans, Precision livestock farming technologies for welfare management in intensive livestock systems, *Scientific and Technical Review of the Office International des Epizooties* **33**(1) (2014), 189–196.

[10] R. Dohmen, C. Catal and Q. Liu, Computer vision-based weight estimation of livestock: a systematic literature review, *New Zealand Journal of Agricultural Research* **0**(0) (2021), 1–21. doi:10.1080/00288233.2021.1876107.

[11] R.A. Schwalbert, T. Amado, G. Corassa, L.P. Pott, P.V.V. Prasad and I.A. Ciampitti, Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil, *Agricultural and Forest Meteorology* **284** (2020), 107886. doi:https://doi.org/10.1016/j.agrformet.2019.107886. https://www.sciencedirect.com/science/article/pii/S0168192319305027.

[12] R. Lokers, R. Knapen, S. Janssen, Y. van Randen and J. Jansen, Analysis of Big Data technologies for use in agro-environmental science, *Environmental Modelling & Software* **84** (2016), 494–504. doi:https://doi.org/10.1016/j.envsoft.2016.07.017.

[13] A. Kamilaris, A. Kartakoullis and F.X. Prenafeta-Boldú, A review on the practice of big data analysis in agriculture, *Computers and Electronics in Agriculture* **143** (2017), 23–37. doi:https://doi.org/10.1016/j.compag.2017.09.037.

[14] V.M. Ngo and M.-T. Kechadi, Crop Knowledge Discovery Based on Agricultural Big Data Integration, in: *Proceedings of the 4th International Conference on Machine Learning and Soft Computing*, ICMLSC 2020, Association for Computing Machinery, New York, NY, USA, 2020, pp. 46–50–. ISBN 9781450376310. doi:10.1145/3380688.3380705.

[15] R. Albertoni, D. Browning, S. Cox, A.G. Beltran, A. Perego and P. Winstanley, Data Catalog Vocabulary (DCAT) - Version 2, 2020, https://www.w3.org/TR/vocab-dcat-2/.

[16] T. Lebo, S. Sahoo and D. McGuinness, PROV-O: The PROV Ontology, 2013, https://www.w3.org/TR/prov-o/.

[17] R. Cyganiak and D. Reynolds, The RDF Data Cube Vocabulary, 2014, https://www.w3.org/TR/vocab-data-cube/.

[18] C. Bizer, T. Heath and T. Berners-Lee, Linked data-the story so far, *Semantic Services, Interoperability and Web Applications: Emerging Concepts* (2009), 205–227.

[19] R. Cyganiak, D. Wood and M. Lanthaler, RDF 1.1 Concepts and Abstract Syntax, W3C, 2014, https://www.w3.org/TR/rdf11-concepts/.

[20] W3C, OWL 2 Web Ontology Language, 2012, https://www.w3.org/TR/owl2-overview/.

[21] W3C, SPARQL 1.1 Overview, 2012, https://www.w3.org/TR/sparql11-overview/.

[22] H. Knublauch and D. Kontokostas, Shapes Constraint Language (SHACL), W3C, 2017, https://www.w3.org/TR/shacl/.

[23] B.V. Nuffelen, DCAT Application Profile for data portals in Europe Version 2.0.1, 2020.

[24] N. Sofou and A. Dragan, StatDCAT-AP – DCAT Application Profile for description of statistical datasets Version 1.0.1, 2019.

[25] GeoDCAT-AP: A geospatial extension for the DCAT application profile for data portals in Europe, 2016.

[26] Establishing an Infrastructure for Spatial Information in the European Community (INSPIRE), 2007.

[27] E. Kalampokis, D. Zeginis and K. Tarabanis, On modeling linked open statistical data, *Journal of Web Semantics* **55** (2019), 56–68. doi:https://doi.org/10.1016/j.websem.2018.11.002. http://www.sciencedirect.com/science/article/pii/S1570826818300544.

[28] OGC Abstract Specification Geographic information — Observations and measurements, 2013.

[29] A. Haller, K. Janowicz, S. Cox, D.L. Phuoc, K. Taylor and M. Lefrancois, Semantic Sensor Network Ontology, 2017, https://www.w3.org/TR/vocab-ssn/.

[30] Data Specification on Agricultural and Aquaculture Facilities – Technical Guidelines, 2013.

[31] INSPIRE Data Specifications – Base Models – Activity Complex, 2012.

[32] R. Palma, T. Reznik, M. Esbri, K. Charvat and C. Mazurek, An INSPIRE-based vocabulary for the publication of Agricultural Linked Data, in: *Proceedings of the OWLED Workshop: OWL Experiences and Directions, collocated with the 14th International Semantic Web Conference (ISWC-2015), Bethlehem PA, USA*, 2015.

[33] AGROVOC, 2021, https://www.fao.org/agrovoc.

[34] C. Hurtaud, J. Bugeon, O. Dameron, A. Fatet, I. Hue, M.-C. Meunier-Salaün, M. Reichstadt, A. Valancogne, J. Vernet, J. Reecy, C. Park and P.-Y. Le Bail, ATOL: A new ontology for livestock, in: *ICAR 2011: "New technologies and new challenges for breeding and herd management"*, Bourg-en-Bresse, France, 2011, Session 2 : phenotyping of complex traits. https://hal.archives-ouvertes.fr/hal-01189518.

[35] L. Joret, J. Bugeon, J. Aubin, J.P. Blancheton, M. Hassouna, C. Hurtaud, S.S. Kaushik, F. Médale, M.-C. Meunier-Salaün, J. Vernet, A. Wilfart, J.-Y. Dourmad and P.-Y. Le Bail, EOL: a new ontology for livestock system and rearing conditions, in: *64. Annual Meeting of the European Federation of Animal Sci-*

ence (EAAP), Book of Abstracts of the 64th Annual Meeting of the European Federation of Animal Science, Vol. 19, Wageningen Academic Publishers, Nantes, France, 2013, p. 660 p. https://hal.inrae.fr/hal-02745982.

[36] M.-C. Salaun, J. Yon, P.-Y. Le Bail and M. Reichstadt, Animal Health Ontology for Livestock, 2019, https://doi.org/10.15454/KKZ3TS.

[37] S. Cox and C. Little, Time Ontology in OWL, 2020, https://www.w3.org/TR/owl-time/.

[38] Quantities, Units, Dimensions and Data Types Ontologies (QUDT), 2021, http://qudt.org/.

[39] D. Zeginis, A. Hasnain, N. Loutas, H.F. Deus, R. Fox and K. Tarabanis, A collaborative methodology for developing a semantic model for interlinking Cancer Chemoprevention linked-data sources, *Semantic Web* **2**(5) (2014), 127–142.

[40] B.F. Loscio, C. Burle and N. Calegari, Data on the Web Best Practices, W3C, 2017, https://www.w3.org/TR/dwbp/.

[41] DCMI Metadata Terms, 2020, https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.

[42] D. Brickley and R.V. Guha, RDF Schema 1.1, 2014, https://www.w3.org/TR/rdf-schema/.

[43] R. Atkinson and N. Car, The Profiles Vocabulary, 2019, https://www.w3.org/TR/dx-prof/.

[44] R. Lott, K. Ryden, M. Desruisseaux, M. Hedley and C. Heazel, Geographic information — Well-known text representation of coordinate reference systems, 2019, http://www.opengis.net/doc/is/wkt-crs/2.0.6.

[45] J. Weier and D. Herring, Measuring Vegetation (NDVI & EVI), 2000.

[46] H. Hanke and D. Knees, A phase-field damage model based on evolving microstructure, *Asymptotic Analysis* **101** (2017), 149–180.

[47] E. Lefever, A hybrid approach to domain-independent taxonomy learning, *Applied Ontology* **11**(3) (2016), 255–278.

[48] P.S. Meltzer, A. Kallioniemi and J.M. Trent, Chromosome alterations in human solid tumors, in: *The Genetic Basis of Human Cancer*, B. Vogelstein and K.W. Kinzler, eds, McGraw-Hill, New York, 2002, pp. 93–113.

[49] P.R. Murray, K.S. Rosenthal, G.S. Kobayashi and M.A. Pfaller, *Medical Microbiology*, 4th edn, Mosby, St. Louis, 2002.

[50] E. Wilson, Active vibration analysis of thin-walled beams, PhD thesis, University of Virginia, 1991.

[51] T. Harmon, C. Kvien, D. Mulla, G. Hoggenboom, J. Judy, J. Hook et al., Precision agriculture scenario, in: *NSF workshop on sensors for environmental observatories. Baltimore, MD, USA: World Tech. Evaluation Center*, 2005.

[52] E. Kalampokis, A. Karamanou and K.A. Tarabanis, Interoperability Conflicts in Linked Open Statistical Data, *Inf.* **10**(8) (2019), 249. doi:10.3390/info10080249.