Cross-Sectorial Semantic Model for Support of Data Analytics in Process Industries

Authors:

Martin Sarnovsky, Peter Bednar, Miroslav Smatana

Date Submitted: 2019-07-30

Keywords: process industry, semantic annotation, data analytics, semantic model

Abstract:

The process industries rely on various software systems and use a wide range of technologies. Predictive modeling techniques are often applied to data obtained from these systems to build the predictive functions used to optimize the production processes. Therefore, there is a need to provide a proper representation of knowledge and data and to improve the communication between the data scientists who develop the predictive functions and domain experts who possess the expert knowledge of the domain. This can be achieved by developing a semantic model that focuses on cross-sectorial aspects rather than concepts for specific industries, and that specifies the meta-classes for the formal description of these specific concepts. This model should cover the most important areas including modeling the production processes, data analysis methods, and evaluation using the performance indicators. In this paper, our primary objective was to introduce the specifications of the Cross-sectorial domain model and to present a set of tools that support data analysts and domain experts in the creation of process models and predictive functions. The model and the tools were used to design a knowledge base that could support the development of predictive functions in the green anode production in the aluminum production domain.

DOI of Published Version: https://doi.org/10.3390/pr7050281 processes

MDPI

Article

Cross-Sectorial Semantic Model for Support of Data Analytics in Process Industries

Martin Sarnovsky *, Peter Bednar and Miroslav Smatana

Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University Kosice, Letna 9, 040 01 Kosice, Slovakia; peter.bednar@tuke.sk (P.B.); miroslav.smatana@tuke.sk (M.S.)

* Correspondence: martin.sarnovsky@tuke.sk

Received: 21 March 2019; Accepted: 6 May 2019; Published: 13 May 2019

Abstract: The process industries rely on various software systems and use a wide range of technologies. Predictive modeling techniques are often applied to data obtained from these systems to build the predictive functions used to optimize the production processes. Therefore, there is a need to provide a proper representation of knowledge and data and to improve the communication between the data scientists who develop the predictive functions and domain experts who possess the expert knowledge of the domain. This can be achieved by developing a semantic model that focuses on cross-sectorial aspects rather than concepts for specific industries, and that specifies the meta-classes for the formal description of these specific concepts. This model should cover the most important areas including modeling the production processes, data analysis methods, and evaluation using the performance indicators. In this paper, our primary objective was to introduce the specifications of the Cross-sectorial domain model and to present a set of tools that support data analysts and domain experts in the creation of process models and predictive functions. The model and the tools were used to design a knowledge base that could support the development of predictive functions in the green anode production in the aluminum production domain.

Keywords: semantic model; data analytics; semantic annotation; process industry

1. Introduction

Process industries have characterized a significant share of European industry, involving the intense use of raw resources and energy. Therefore, these industries provide a context where even small optimizations can lead to large savings in terms of both economic and environmental costs [1,2]. This is especially true for specific industrial processes, such as aluminum smelting or injection moulding, characterized by production in high volumes and divided among many distributed production units. Predictive modeling can be effective for optimizing processes in this context. The application of predictive models is not straightforward for several reasons, including interoperability issues among existing software systems. As a consequence, the deployment of predictive functions in these production environments at a sustainable cost or with sufficient reliability is not always feasible. As process industries usually integrate various software systems and incorporate a wide range of technologies on different execution layers [3,4], proper representation of the information and underlying data obtained from these systems is necessary. Such semantic models should provide machine-readable conceptualizations that describe the available and needed assets, equipment, and production processes involved in the specific production, as well as the functionality and information flow between these assets. In the process industries, production processes must be accurately modeled to achieve the optimization goals set by the companies.

In this paper, we describe a semantically-enriched model for process industries. Semantic techniques are used to define, inter-link, and share distributed models describing all the key aspects of the multi-scale processes of interest. First, the models include all system-specific classes existing at different layers and then all the different classes describingfú the same or related information are logically inter-linked by means of semantic web standards. This results in an open, extendable semantic description of all data entities handled by all systems available in the production site. The proposed semantic model was designed in the context of the MOdel based coNtrol framework for Site-wide OptimizatiON of data-intensive processes (MONSOON), a SPIRE (Sustainable Process Industry through Resource and Energy Efficiency) research project, which aims to develop an infrastructure with the main objective of establishing a data-driven methodology supporting the identification and exploitation of optimization potentials by applying model-based predictive controls in the production processes [5]. The platform was evaluated in two domains: Aluminum production and plastic injection. The proposed architecture consists of two main components, as depicted in Figure 1.

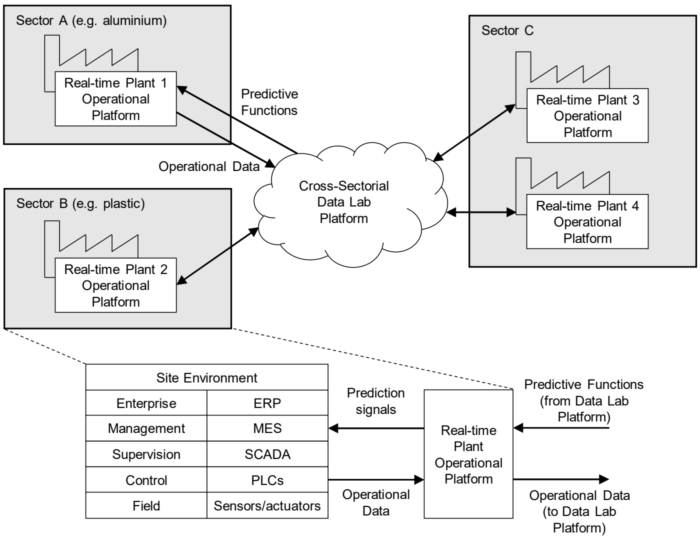

Sector A (e.g. aluminium)
Real-time Plant 1 Operational Platform
Predictive Functions
Operational Data
Sector C
Real-time Plant 3 Operational Platform
Real-time Plant 4 Operational Platform
Cross-Sectorial Data Lab Platform
Sector B (e.g. plastic)
Real-time Plant 2 Operational Platform
Site Environment
Enterprise
ERP
Management
MES
Supervision
SCADA
Control
PLCs
Field
Sensors/actuators
Prediction signals
Operational Data
Real-time Plant Operational Platform
Predictive Functions (from Data Lab Platform)
Operational Data (to Data Lab Platform)

Figure 1. High-level architecture of the MONSOON platform.

The Cross-Sectorial Data Lab (hereafter called the Data Lab) is the core component of the MONSOON platform. It is a collaborative environment built on top of the big data frameworks and technologies and used to collect, store, and process large volumes of data obtained from multiple production sites from different process industries. In the Data Lab, the data scientists process and analyze the production data and develop predictive functions to optimize the production processes. Data Lab also provides a set of tools to evaluate those models, simulate the functions performance in testing environment, and deploy them into the production environment.

The Real-Time Plant Operation Platform is a set of components used during run-time in a particular plant. The components are used to integrate various sensors and systems deployed on-site (including PLC, MES, SCADA, and ERP) and provides a common interface for communication with these systems. The components are then used to transfer relevant data from the production site to the Data Lab. Predictive models developed in Data Lab are deployed in these components and are integrated within the on-site environment.

This paper is organized as follows. Section 2 describes the motivation for this work from the perspective of data analysis process and its relation to the proposed semantic model. Then, an overview of existing standards and solutions relevant to our work in Section 3 and is followed by the description of the methodology adopted in development of the semantic model in Section 4. Section 5 introduces the model, describes its structure and main components, and demonstrates its usage. The following chapters then present the model evaluation and the software tools using the presented semantic model. The Appendix A contains several examples from the demonstration scenarios from one of the MONSOON's application domains.

The terms *model* and *modeling* in this paper have different meanings:

(a) the *semantic model* is the formal specification of the concepts from the given domain;
(b) the *data analysis model* is the result of applying the data modeling techniques (e.g., predictive function).
(c) the *process model*, i.e., model, specifies the decomposition of the process to steps and describes the steps' data and execution dependencies (usually based on graphical notation in the form of flowcharts).

The aim of this study was to propose a formalized semantic model that represents the knowledge exchanged by domain experts and data scientists. Parts of this semantic model also cover process modeling for production processes and describe metadata about the data analysis models (i.e., predictive functions).

2. Motivation

This section describes the main objectives of the proposed Cross-sectorial domain model in the context of the general data analysis process. Several standard methodologies, such as CRISP-DM or SEMMA [6–8], break the process into several major phases.

The initial phase called *Problem understanding*, focuses on understanding the objectives and requirements of data analysis from a business perspective, i.e., how data analytics methods can optimize the production process. During this phase, the domain experts in cooperation with the data scientists analyze the overall production process and its steps (production segments) and specify which key-performance indicators (KPIs) will be optimized. This knowledge is then converted into the data analysis problem definition (e.g., classification, regression, anomaly detection, etc.) by data scientists. The subsequent *Data understanding* phase starts with data collection and proceeds with activities to ensure familiarity with the data, to identify data quality problems, to provide the first insights into the data, and to detect interesting subsets to form hypotheses for hidden information. During this phase, data scientists, with the cooperation of domain experts, select and describe the subset of relevant data required to achieve optimization objectives specified in the problem understanding phase. The data are described using the already known dependencies between the data and KPIs.

The *Data understanding* phase is followed by the *Data preparation* phase, which covers all activities necessary to construct the final dataset (data that will be fed into the data analysis tools), and the modelling phase, where the various data analysis techniques (such as decision trees, neural networks, logistic or linear regression models, etc.) are selected and applied, and their parameters are calibrated to optimal values. While these two phases are essential for the overall results of the data analysis process, they are more technical and mostly involve activities of data scientists, with some exceptions, e.g., to collect initial raw data from the production site environment, data scientists have to cooperate with the site IT-specialists responsible for the setup of the data integration component.

The result of the *Modelling phase* is the set of data analytics models (i.e., predictive functions) that appear to have high quality from a data analysis perspective. This quality is usually evaluated by applying the model to the independent testing dataset not used during the building of the model, and computing the evaluation statistics such as classification accuracy (number of tested records with the correctly predicted value by the model/total number of records in the testing dataset). Before proceeding to the final deployment of the model, it is essential to evaluate the model more thoroughly and review the steps executed to construct the model, to ensure it adequately achieves the business objectives [9]. Within the *Evaluation phase*, the quality of the models evaluated on the testing dataset is projected onto the business-oriented KPIs specified in the problem understanding phase. As such, domain experts can directly interpret the evaluation results of the overall data analysis and assess the quality of the constructed models from a business perspective.

From the perspective of semantic modeling, the most critical steps are Problem understanding, Data understanding, and Evaluation phases. These steps define the interface between the domain experts and data scientists, and they are the most intensive regarding knowledge creation and sharing.

The overall motivation to develop the Cross-sectorial semantic model was to improve the communication between data scientists and domain experts during these phases and to formally capture and externalize communicated or created knowledge. Such a semantic model should enable cross-sectorial sharing of knowledge about the optimization of the production processes by data analytics methods and predictive functions (Figure 2). The model itself should represent the common language between domain experts and data scientists, which would provide the necessary concepts, which are important to capture in context of data analytics in particular domain. The model should assist the domain experts to externalize their tacit domain-specific knowledge (about production processes, resources, etc.) into a more comprehensive form, and also to capture all necessary aspects of the domain-specific knowledge, which could be necessary when applying the data analytics in particular domain. which would be represented in the terminology used by the data scientists. Such knowledge sharing should support and improve the Problem and Data understanding phases of the data analytics process by providing a common ground when describing the domain and related problem and data aspects. On the other hand, the data scientists could use the semantic model to define the predictive functions to solve the data analytical tasks in particular domain and to present to the domain experts, how such functions influence the domain KPIs.

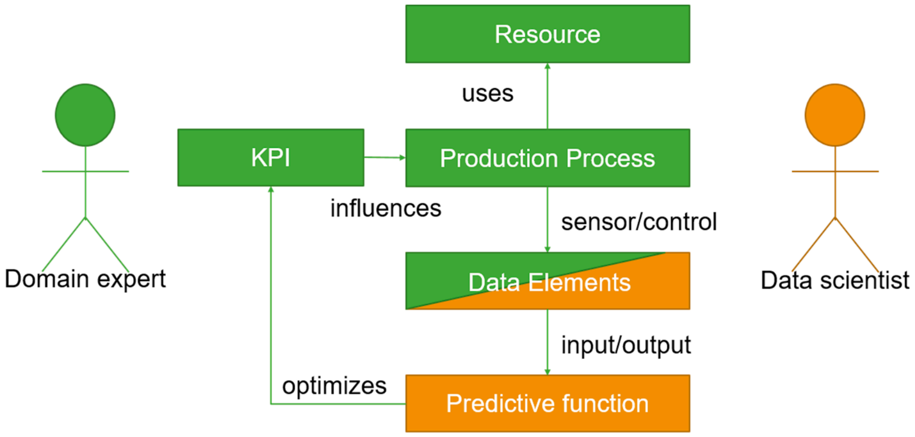


Figure 2. The interaction between the domain expert and data scientist in the data mining process.

The main knowledge management objectives of the Cross-sectorial domain model are:

(1) Improve communication between data scientists and domain experts during the phases of problem understanding, data understanding, and evaluation by creation of a shared knowledge base.
(2) Allow within- and cross-sectorial sharing of knowledge about the optimization of the production processes using data analytics methods and predictive functions.
(3) Automatize evaluation of predictive functions by automatic computation of the impact on the key performance indicators (KPIs) from evaluation statistics.
(4) Allow "what if" simulation scenarios and automatic optimization of the production process during the modelling and evaluation phases.
(5) Automatize validation of data dependencies during the deployment of the predictive function in the operation environment.

Besides the knowledge management objectives, another objective of the semantic model design should be automatized evaluation of predictive functions by computation of the impact on the KPIs from evaluation statistics. Modeling in combination with related tools should enable development of what if simulation scenarios and automatic optimization of production process during the modelling and evaluation phases and automatize validation of data dependencies during the deployment of the predictive function in the operational environment. To achieve these objectives, the semantic model should support:

(1) Presentation in a human-readable form using modeling tools that allow creation, visualization, querying, filtering, and navigation in the semantic model elements by data scientists and domain experts.

(2) A machine-readable format with formally defined semantics, which supports automatic validation of deployment, quality evaluation of the predictive functions, and simulation and automatic optimization of the production processes.

As the main motivation was to develop an approach applicable to more industrial domains, the proposed semantic model does not aim to specify the concepts for a specific industry domain, but instead focuses on cross-sectorial aspects. From that point of view, the Cross-sectorial domain model can rather be considered as a meta-model that determines meta-classes for the formal description of these particular concepts. Such meta-classes should cover all important aspects, common for the process industries, in the context of optimization of production processes using data analytics. Therefore, the Cross-sectorial domain model should include the elements to describe the production itself, including the processes and their performance, equipment and personnel, concepts enabling to describe any kind of data obtained from the production and concepts used to describe the predictive models used in optimization. Therefore, we decided to formally divide the semantic model into the three respective modules: production process modeling, data and KPI modeling, and modeling of predictive functions. The semantic model and respective modules will be in more detail, as described in Section 5.

To ensure the aspect of the presentation in human-readable form and to support the adoption of the semantic model in the real-world scenarios, a set of modelling tools were designed as a part of the MONSOON platform. Such tools utilize the semantic model as a basis and provide the graphic user interface for creation of the industrial domain models. Semantic framework, a tool specifically designed to capture the domain knowledge, is described in Section 7.

3. Related Work

3.1. Standards for Modeling Manufacturing Processes

In this section, we describe two types of conceptual information models in the industrial domain: Models based on industrial standards and semantic models. In the first group, ANSI/ISA-95 is a well-established international standard for developing an automated interface between enterprise and manufacturing control systems [10]. Data models of this standard are implemented in Business To Manufacturing Markup Language (B2MML), an XML-based set of schemas for modeling manufacturing processes [11]. B2MML is frequently used to integrate business systems (e.g., ERP systems) with manufacturing systems (e.g., control systems). In general, B2MML allows describing the manufacturing processes as well as material, personnel, equipment, and physical assets related to the production. B2MML enables us to describe how to produce the product, to specify the resources needed in the production, and to define the production schedule and performance. ISA-88 is another standard addressing batch production processes [12]. Both ISA-95 and ISA-88 are overlapping and inconsistent in several aspects. ISA-88 is better in terms of conceptualization and flexibility, whereas ISA-95 is applicable to a broader spectrum of processes. Therefore, activities involving the integration of both standards have been explored [13], including the development of upper-level ontology approach to map the concepts of both standards [14].

Several applications of semantic and knowledge technologies to the modeling of manufacturing processes have been presented and developed. Lemaignan et al. [15] introduced MASON, an upper ontology that represents a common semantic model of the manufacturing processes and environment. The primary objective of the ontology was to help estimate the manufacturing costs. The ontology describes the most important concepts of the industrial processes but only uses the cost factor as a method of their evaluation. Kharlamov et al. [16] used an OWL 2 to develop an ontology for the manufacturing and energy production sectors. The authors developed a set of tools for engineers with little knowledge of semantic technology in the creation of ontology-based models and in populating them with data. In Pakonen et al. [17], fuzzy ontologies were used in the industrial domain to improve information retrieval from the plant knowledge base consisting of reports obtained from different plant systems. With the emergence of the Industry 4.0 concept, several other ontological approaches have been created. Ontologies have been used to represent manufacturing knowledge [18], to describe industrial production processes [19], and to explain the capabilities of manufacturing resources [20]. An important objective of the use of the ontological approach is to support the interoperability between the industrial standards [21,22].

The main objective of the cross-sectorial domain model is to capture the common generic concepts related to the process industries. This approach leads to the design of a meta-model, which would enable the development of specific domain models for particular industries. From that perspective, the use of a standard, such as B2MML, as the basis for the semantic model proved to be useful.

3.2. Standards for Modeling Predictive Functions

The need has increased for formalized representations of the data analytics processes and for the formal representation of outcomes of those processes in general. Several formalisms describe scientific investigations and findings of research available, but most of them are specific to a domain, such as biomedicine. Examples of such formalisms include the ontology of biomedical investigations (OBI) [23] or ontology of experiments (EXPO) [24]. These ontologies specify useful concepts that describe general processes producing output data given some input data and formalize outputs and results of the data analytics investigations. They are aligned with the general upper-level ontologies such as SUMO [25] or DOLCE [26].

Semantic technologies are also applied directly to formalize knowledge about data analytics processes and to discover knowledge in databases. The initial goal of this effort was to build an intelligent data mining assistant that combines planning and meta-learning for the automatic design of data mining workflows. The assistant relies on the formalized ontologies of data mining operators that specify constraints, required inputs, and provided outputs for various operations in the data pre-processing, modeling, and validation phases. Examples of the ontologies for data mining and data analytics include Data Mining OPtimization Ontology (DMOP) [27] and Ontology of Data Mining (OntoDM) [28]. DMOP ontology provides concepts for the description of the input analyzed data, data pre-processing operations, data mining algorithms, and models and analysis results. OntoDM provides a unified framework for data mining, and contains definitions of the basic data mining concepts together with the concepts describing scientific investigations. This was further extended by expose ontology [29]. Expose provides a formal domain model for a database consisting of data mining experiments based on OntoDM and extends it with the data mining algorithms' description from DMOP. Another extension is a description of experiments (e.g., algorithm parameters, execution environment, methods of evaluation, etc.). OntoDM and Expose are primarily used to provide a controlled vocabulary for data analytics investigations. Other studies [30–32] presented an interesting approach for developing data science ontology, used to understand the data analytics code using dataflow graphs.

Formalized ontologies, such as DMOP or OntoDM, have already been built by the conceptualization of existing industry standards. These standards include XML-based or JSON-based formats for exchanging of data mining processes and models such as PMML [33], PFA [34], or ONNX, or interfaces of common software libraries for data mining such as Rapid-miner, Weka, scikit-learn, or Keras. Alignment of semantic models with the data analytics technologies is particularly important for interoperability with existing tools, which will allow seamless integration of semantic technologies with the data analytics components of the platform.

4. Methodology

For the design of the semantic model, we adopted a methodology that aims to facilitate the development of production process and data analytics ontologies so that the ontology developer can justify the rationale behind the involved decisions. The base of the proposed methodology is SMOL [35], which was extended using the concepts introduced by Grüninger and Fox [36]. The final proposed methodology emphasizes conceptualization based on the usage scenarios for the knowledge model. The methodology consists of the following steps:

(1) Methodology strategy selection. In this phase, the best methodological strategy is selected according to the available or recoverable data from the existing knowledge sources related to the specific domain, or from the information from the opinion of domain experts. For our case, we adopted the middle-out strategy: We started with the conceptualization of the main domain concepts, which were subsequently mapped to the upper-level ontology and further specified with new domain sub-concepts.

(2) Specification of use cases and query requirements. In this phase, we specified base use cases outlining how the knowledge system implementing the semantic models are used within the scope of the whole data analytics platform. From each use case, we defined the requirements of the semantic models specified in the form of competency questions expressed in natural or pseudo-natural language (with limited question structure). The competency questions in this phase are informal since they are not expressed in formalized ontology language. Examples of the questions are shown in Section 6.

(3) Knowledge structure construction. Once the informal competency questions have been proposed for the new ontology, the ontology terminology (classes, relations, and instances) are specified using some first-order logical language. This language must provide the necessary vocabulary to formally restate the informal competency questions [37], i.e., after conceptualization, the developed ontology is accompanied by a set of formal competency questions. After the base concepts of the proposed semantic model are specified, the concepts were mapped and refined according to the selected upper-ontologies.

(4) Knowledge structure validation and reorganization. The formal competency questions associated with the developed ontology represent how to evaluate the ontology to verify that it is adequate. Each competency question is formally evaluated to recognize if the proposed structure of the knowledge model is consistent and complete. In the case where some knowledge is missing or inconsistent, changes in the ontology structure are proposed by introducing the new classes, relations, or instances, or by restructuring the existing hierarchies of concepts.

Finally, the formal competency questions were also mapped to the web service interfaces of the implemented tools and the underlying JavaScript Object Notation (JSON) data model implementing the semantic models. A part of the competency question was then implemented in the form of software integration tests to validate the functionality of the final implemented knowledge system.

5. Cross-Sectorial Domain Model

The main tool introduced by the cross-sectorial domain model to improve communication between users is a shared vocabulary collaboratively created by domain experts and data scientists. The domain concepts are described as the vocabulary entries that unambiguously define their meanings. Each entry has an assigned primary label presented as the unique title representing the concepts for the user. Optionally, each entry can list additional labels such as synonyms, acronyms, or abbreviations representing the same concept in the particular natural language(s). Besides the definition of the meaning, vocabulary can be used as the Knowledge Organization System (KOS), i.e., the entries can be used as the indexing keywords for the organization of any documents or other resources (e.g., datasets, scripts, etc.), which contain relevant information about the domain or solved business problem. The concepts in the vocabulary are further organized in a hierarchical structure (taxonomy) from broader to narrower concepts. The taxonomy structure allows efficient navigation in the domain knowledge and retrieval of relevant information.

The structure of the cross-sectorial domain model vocabulary is based on the Simple Knowledge Organization System (SKOS) standard with the extensions defined in the JSKOS standard, which define the JSON (JavaScript Object Notation) format for the knowledge organization systems such as taxonomies. All domain concepts are represented as the objects of the JSKOS concept type, which formally defines fields of the vocabulary entries and their relationships, e.g., narrower or broader, previous/next, related, etc. The structure of the controlled vocabulary is domain independent, and the proposed JSKOS format can be used to describe concepts in any domain. For the application in the process industry domain, the cross-sectorial domain model defines the core of the vocabulary with the main basic concepts for the representation of the main entities in the process industry. The top concepts can be divided into:

(1) Production process modeling, which specifies the decomposition of production processes into segments (production phases) and describes resources required for production;
(2) Data modeling, which specifies concepts for the description of the data elements and key performance indicators; and
(3) Predictive function modeling, which specifies concepts for the description of the data analytics models.

Besides the shared vocabulary, the execution workflow of the production processes and their decomposition to activities and sub-processes are usually modeled using graphical notation in the form of various schemas or workflow diagrams. The structure of the vocabulary specified by the cross-sectorial domain model allows linking each graphical object from the diagram to the definition entry in the shared vocabulary, i.e., it is possible to navigate from the process model to the definition of the related concepts or from the concept (e.g., data element, KPI, or predictive function) to the relevant part of the graphical process model. As the reference graphical notation for the modeling of the production processes, the cross-sectorial domain model adopts the subset of the BPMN 2.0 standard.

To summarize, the overall structure of the proposed cross-sectorial domain model is presented in Figure 3. The bottom layer provides interoperability with the linked data and semantic web technologies based on the RDF and JSON-LD standards. Over this standard knowledge model, we defined the meta-model for the knowledge organization system based on the SKOS and JSKOS standards. Using this meta-model, we identified common vocabulary for the process industry domain, which describes production processes and production resources. The domain concepts are complemented by the general data analytics concepts adopted from the DMOP ontology, which are used to describe the data elements, KPIs, and predictive functions. Finally, the definition of the production processes is extended with graphical notation based on the simplified BPMN diagrams.

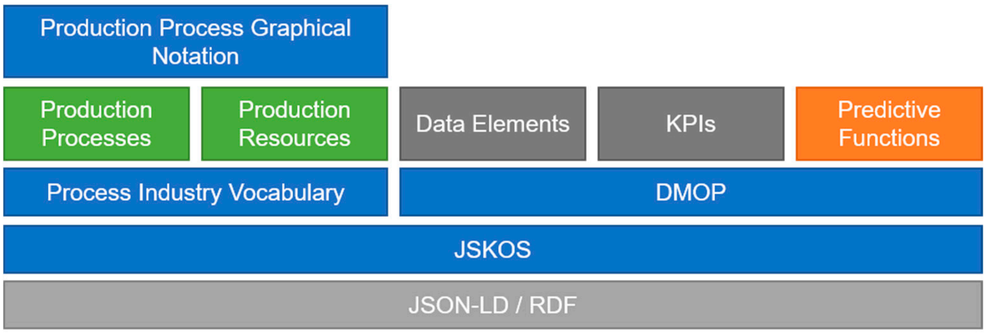


Figure 3. Overall structure of the cross-sectorial domain model.

5.1. Production Process Modeling

5.1.1. Core Concepts

Besides the description of the properties of concepts from the production industry domain, the cross-sectorial domain model provides core vocabulary that consists of the main categories dividing all domain concepts, production processes, and production resources. The vocabulary is represented as a JSKOS concept scheme, and each category is represented as a JSKOS concept. Note that it is necessary to distinguish types for the domain concepts (production process and production resource,) and category concepts (production process concept and production resource concept). All specific concepts of the given type are linked as the narrower concepts of the corresponding category. Each specific process concept (e.g., Green Anode Production) has a production process type and is linked as the narrower concept to the production process concept category. The structure of the core concept scheme is presented in Figure 4.

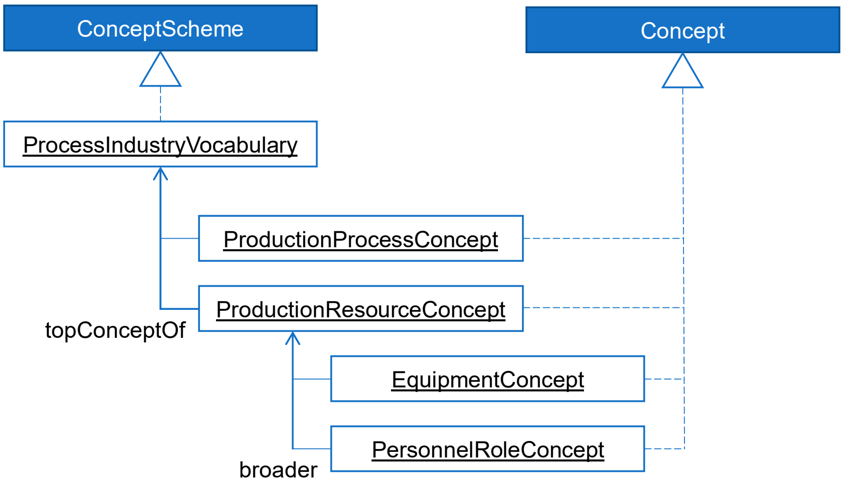


Figure 4. Core concepts scheme.

5.1.2. Production Process and Production Segment

Production process and process segments (Figure 5) concepts represent the decomposition of the overall production process into particular production steps. The production process is a container composed of production segments (specific steps in the process) ordered in the execution workflow. The production process can be further decomposed to sub-processes. This is represented by the compound production segment, which delegates execution to the underlying sub-process. The sub-processes refer to the parent process on the upper level of the process hierarchy by the inverse *partOf* relation. The taxonomy of the production process concepts corresponds to the decomposition of the production processes to sub-processes. The *broader* relation is equivalent to the *partOf* field, whereas *narrower* specifies an inverse relationship.

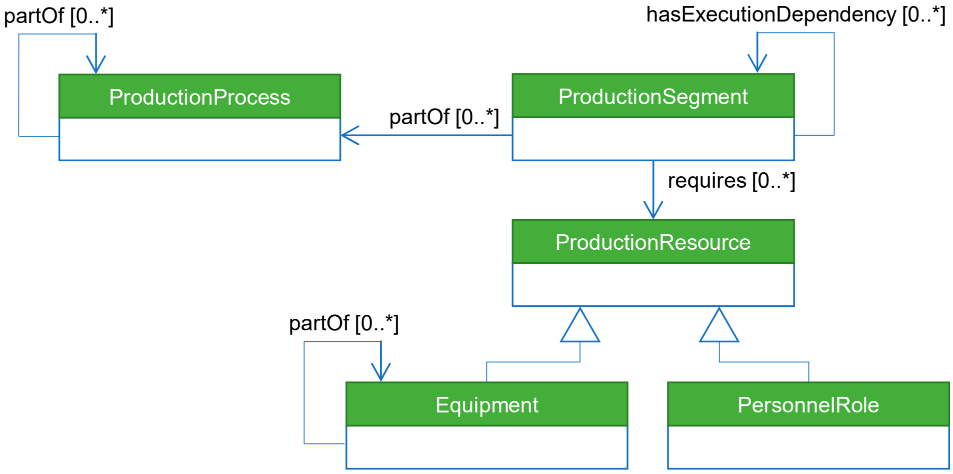


Figure 5. Process modeling concepts.

The process segment is the logical grouping of personnel and equipment resources required to complete a production step. It defines what types of personnel and equipment are needed, and can describe specific resources, e.g., some particular kind of equipment for specific segment. The *partOf* field, in this case, refers to the production process that consists of this production segment. Process segments are ordered in the execution workflow by an execution dependency relationship. The execution workflow order between the production segments (steps) is represented by the *hasExecutionDependency* relation, which specifies a condition when a segment can be executed only after the previous segment is finished. The ordering of the production segment concepts can also be expressed using the *previous* and *next* fields. *previous* is equivalent to *hasExecutionDependency*, and *next* represents an inverse relationship. The list of production resources (equipment or personnel role concepts) required to perform this production segment is represented using the *requires* field.

The cross-sectorial domain model is not limited to a particular graphical notation and can be used to describe processes using various schemas, workflow diagrams, or dynamic visualizations presented in the process visualization software (e.g., InTouch screens). One of the suitable forms for modeling the knowledge about the production processes involves using diagrams where the process steps, their execution dependencies, and the involved resources are represented as the interconnected graphical objects. As the reference graphical notation for the implementation of the MONSOON semantic framework, we adopted a subset of the BPMN 2.0 standard [38]. Examples are provided in Section 8.

5.1.3. Production Resources

Production resources are conceptually divided into equipment resources and person roles (Figure 4). The common super-type for these concepts is the production resource concept, which represents resources required to complete a particular production step. Grouping of resources with similar characteristics and purposes can be described by resource classes. Any piece of the production resource may be a member of zero or more classes. The semantic model does not prescribe any classification scheme for the resources but defines a simple knowledge organization system (SKOS) [39] as the preferred concept schema for specification of classification thesauri, subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabulary. Additionally, each production resource can be annotated with the relevant data elements describing its parameters and capabilities (Figure 6).

Equipment concept covers any equipment-related entities, e.g., production sites and areas, production units, production lines, work, and process cells. Equipment may include other equipment. For example, a production line may be comprised of work cells. Each of them may be defined in the model as a separate equipment element with distinct properties and capabilities. The person concept covers any human roles involved in the production step and describes their capabilities. Properties and capabilities of equipment or human roles are specified as the linked data elements.

Both equipment and personnel concepts can be organized in concept hierarchies. The taxonomy of equipment concepts corresponds to the decomposition of the equipment to sub-pieces or the equipment can be categorized into more generic classes.

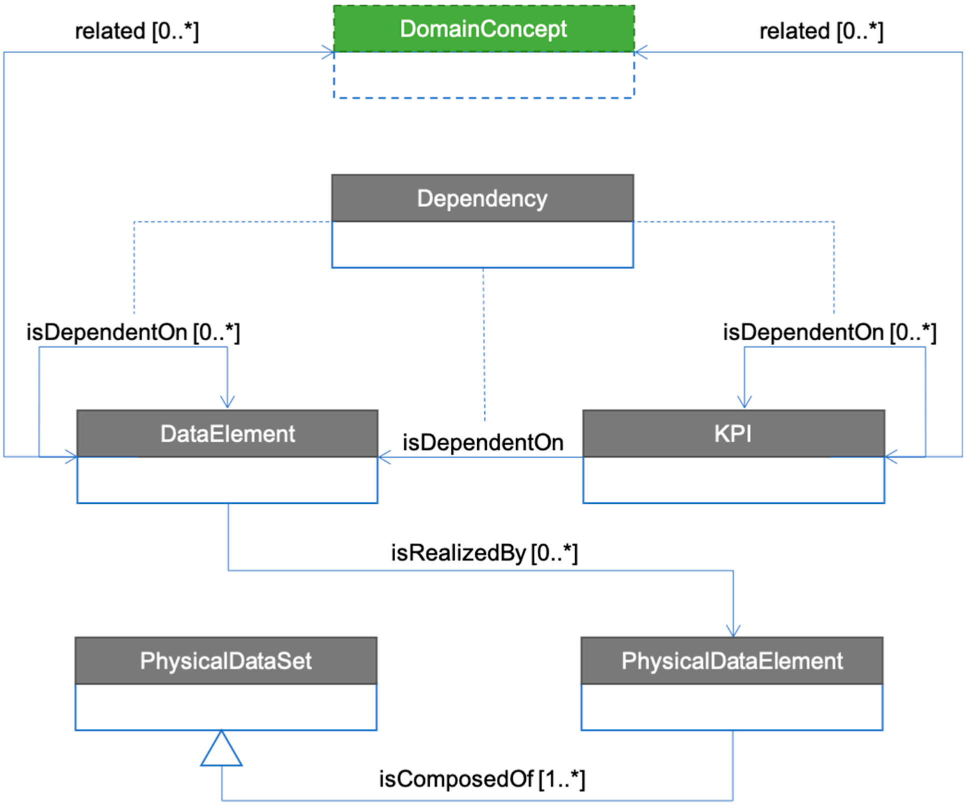


Figure 6. Data modeling concepts.

5.2. *Data Modeling*

Data modeling elements represent the main concepts used to describe the data and relationships between the physical data stored in the Data Lab, particular data elements of those datasets (e.g., attributes), and how they are related to overall KPIs specified for a particular process segment. High-level descriptions of those elements were introduced by Sarnovsky et al. [5]. The next sections provide a more in-depth description.

5.2.1. Data Elements

The main concepts for modeling the data elements are divided into two levels of abstractions: logical and physical. The logical data element (data element concept in the model, Figure 6) concept specifies role, type, and quantity (e.g., units of measurements) of each data element. A role describes how the data elements are treated in the production step. In general, we recognize two different roles: Input and output. Data elements with the input role can be selected as the input of the predictive functions and describe, for example, measurements or input control signals. Data elements with output roles describe diagnostic or control signals, which can be optimized as the outputs of the predictive functions. Each data element object can have both input and output roles specified, e.g., a control signal can serve as an output for the predictive control or as an input for the predictive maintenance. The type of the data element specifies the values of the data element and denotes real intervals (continuous values), ordered sets (ordinal values), or unordered sets (nominal values). Data elements can be related to the production process or equipment, or any other domain concept.

Physical data elements are used to link the data elements to the physical representation of the data in the cross-sectorial Data Lab, e.g., structured records or files. A (logical) data element can be mapped to multiple physical data elements depending on the data pre-processing methods applied to the data during the process of data acquisition and integration. Each physical data element has a specified data type that represents how the data are physically encoded in the records. The data type can be atomic (e.g., byte, integer, double, string) or composed (e.g., array, enumeration, map or union of types). Multiple physical data elements (fields) can be grouped into one physical data set that corresponds to the one relational table, collections of objects in NoSQL database, or flat structured file with the tabular or transaction data. The physical location of the dataset is represented as the IRI. For physical data sets, it is possible to infer datatype schema (represented, for example, as the composed avro datatype), which is the composition of datatypes of the grouped physical data elements.

A simple yet effective specification for physical data elements could be provided by the media types for sensor measurement lists (SenML) [40], which is a data model serializable with limited processing capabilities that is able to fit a large amount of data in a contextually representative format that is readable and writable by machines and humans. This specification defines media types for representing simple sensor measurements and device parameters in the SenML based on JSON, Concise Binary Object Representation (CBOR), eXtensible Markup Language (XML), and Efficient XML Interchange (EXI). Su et al. [41] showed how to transform data expressed with SenML into the RDF to support semantic interoperability. An alternative to the SenML specification described above is the OGC SensorThings API standard specification [42]. Both SenML specification and the OGC SensorThings API standard specification provide a JSON-based method of modeling physical and logical data elements and were thus further investigated within the cross-sectorial domain model.

5.2.2. KPIs

KPI concepts are used to model the overall performance parameters for a given production process. The KPIs are linked to the concepts of process segments or are specified for the whole production process. They represent metrics that can be used for visualization, assessment, and management of the performance or impact of the specific operations within the process industries. Since the KPIs must be measurable, they inherit the main properties (i.e., data type and quantity) from the (logical) data elements (as described in the previous section). KPIs can be further organized into a hierarchy of KPI categories according to the specified classification schema, such as the classification of environmental impact defined in ISO 14031 [43].

Besides the organization of the KPIs in the classification hierarchies, causal relationships between the KPIs can be expressed by the non-hierarchical KPI dependency relationships. Such relations can represent the transformation of one KPI to another (e.g., transformation of energy or material savings to environmental impact KPI, such as level of emissions). Using the KPI dependency relations, KPIs can also be inferred from the evaluation results, which estimate the performance of the predictive functions (see the description of the predictive functions concepts in the following subsection). Using the composition of the KPI dependency relations, it is possible to evaluate the impact of the application of a predictive function from various perspectives, such as product quality, environmental impact, material/energy consumption, or effective usage of the production resources.

5.3. Predictive Functions Modeling

As mentioned in Section 4, the majority of the ontological models used to describe data mining models support the overall data mining workflows. Several other models are available that provide more complete frameworks for data mining, e.g., OntoDM, DMOP or Exposé ontologies, but those provide rather general concepts related to the data mining tasks, datasets, etc., as well as more in-depth and detailed entities related to algorithms or experiments. Regarding the description of the data mining models, DMOP covers a detailed description of the data mining algorithms and its internal structure.

In the cross-sectorial domain model, we decided to use the concepts from DMOP ontology. Figure 7 depicts the main concepts re-used from DMOP ontology and their linking to the rest of the cross-sectorial domain model. The model training task specifies a particular step in the data mining process, and it is realized by an algorithm. The algorithm concept from DMOP (or its sub-classes) describes the particular learning algorithm. The algorithm is applied to the data and produces a model, e.g., predictive model training algorithm transforms input data to an output predictive model. Algorithm characteristics and parameters then specify various aspects of the model induction algorithm, e.g., type of the solved problem, tolerance to noise in the input data, tolerance to missing values, etc., and are represented by *hasQuality* and *hasParameter* properties, respectively.

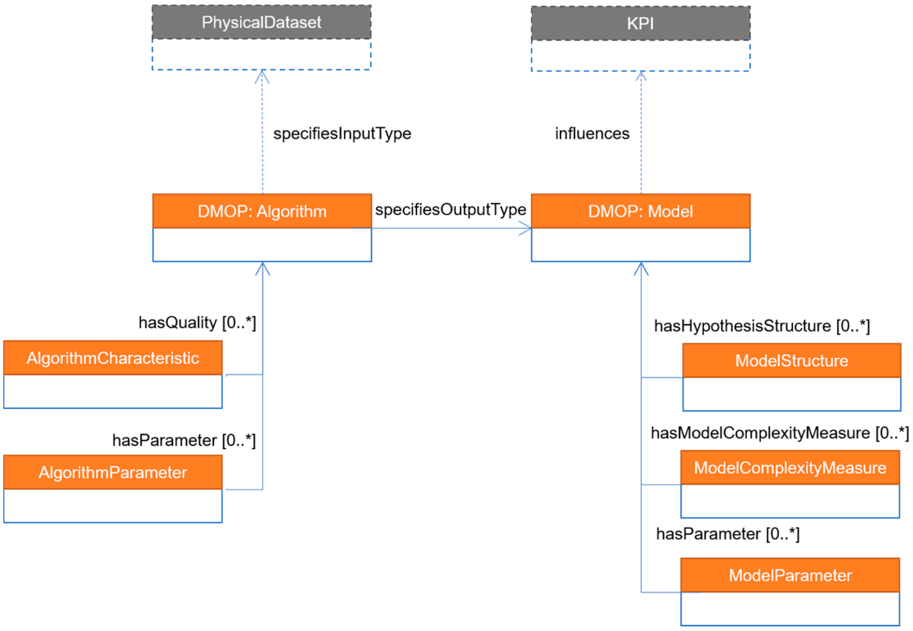


Figure 7. Predictive functions modeling concepts.

The model then represents the output of a learning algorithm. A model concept represents a logical definition of the predictive function. Model class (or its subclasses) then specify the model structure and model parameter. The model structure describes the nature of the model, e.g., logical structure (tree, rules, etc.) or mathematical expression (e.g., kernel functions, etc.). Model parameter specifies the various parameters for a given model type. Each specific model type has a specific set of model parameters. Such parameters can include, for example, weights in neural network models, threshold splits in tree models, etc. The model complexity measure class then contains concepts to quantify the model complexity. Similarly, specific model types with specific parameters have assigned particular complexity measures. In the case of the tree classification model, this could be the depth of a tree or a number of leaves; in the case of support vector machine (SVM), this could be the model number of support vectors or the sum of their weights. DMOP provides a concept hierarchy for both models and algorithms. The trained model can be used, when deployed on-site, in real-world application scenarios. Such scenarios are usually designed to optimize one or more KPIs for the given process. Using the cross-sectorial model, it is possible to specify the *influences* relationship between the trained model and a KPI related to the production process.

6. Evaluation of the Semantic Models

Various methodologies for ontology evaluation are available that specify existing approaches for different types of the ontologies [44–46]. Bandeira. et al. [47] divided the evaluation techniques according to the ontology type (application, task, domain, and top-level ontologies) and presented the main evaluation methodology applicable for a specific ontology type. Different groups of evaluation methods are used based on what should be evaluated and when (in what stage of ontology design) [48]. In our approach, we decided to use the approach described by Goméz-Peréz et al. [49]. The primary reason for evaluating the model was driven by specified use cases within the application domains of the MONSOON project.

The cross-sectorial domain model can be considered a domain ontology, which can be evaluated against a frame of reference. This can be achieved by formulating a set of competency questions—requirements from the real-world. As the evaluation is an iterative process, it can be performed starting from the ontology development to achieve a complete model. In this case, competency questions can represent a part of the functional requirements that can be used during the design and must be satisfied by the implemented model. To use them during the evaluation phase, there is a need to determine how to properly formulate the competency questions and how to measure if the question can be successfully answered using the developed ontology [50].

If we consider a competency question as a natural language formulation that should be answered by the ontology, then its ability to answer those questions can be measured by a possibility of specifying of a single query (or multiple queries) as an expression of competency questions using a chosen query language and ontology concepts and relationships. To collect a proper set of competency questions, a different set of real-world scenarios covering most of the ontology use cases must be considered, but invalid or questions should be removed. This includes redundant and/or incomplete questions, and questions that cannot be answered due to limitations of used formalism. Using the approach previously described [46], we designed a set of competency questions that were common for both aluminum and plastic application domains and categorized them according to different patterns and competency questions archetypes. A list of those competency questions in the design phase for all three developed modules was established. Table 1 summarizes the competency questions and defines which of the patterns the competency questions cover.

Table 1. Informal queries and their patterns. CE: Class Expression, OPE: object property, DP: datatype property, KPIs: key performance indicators.

Informal Question	Pattern	Predicate Arity	Question Type
Which data elements are related to the given process segment?	Which [CE1][OP][CE2]	2	Selection
Which KPIs are relevant to the given production?	Which [CE1][OP][CE2] or [OP][CE3]	3	Selection
Which predictive function can be applied to optimize the given process segment?	Which [CE1][OP][CE2]	2	Selection
Which predictive function can be applied to optimize the given KPIs?	Which [CE1][OP][CE2][OP][CE3]	3	Selection
Which data elements are influencing the given data element?	Which [CE1][OP][CE2]	2	Selection
What is the impact of the predictive function quality on the given KPIs?	What is [DP][CE1][OP][CE2	3	Counting question
Which datasets can be used to build the given predictive function?	Which [CE1][OP][CE2]	3	Selection
Can we validate that all data elements required to apply predictive function are available?	How [CE1][OP][CE2]	2	Binary

The answerability of the ontology can be evaluated by testing how the competency questions can be formulated by the chosen query language using ontology concepts and relations. Table 2 summarizes them and presents their representation in a query language that represents the ability of the ontology module to solve and answer the questions it was designed to answer. For that purpose, we formulated the questions using the SPARQL query language [51] and using the concepts and relations from the presented semantic model.

Table 2. Competency questions common for both domains and their SPARQL representation.

Competency Question	SPARQL Query

Which data elements are related to the given process segment?	SELECT ?element WHERE {
	?element csdm:type DataElement.
	?element csdm:isRelatedTo <segment ID>.}

Which KPIs are relevant to the given production?	SELECT ?kpi WHERE { ?kpi csdm:type KPI. ?kpi csdm:relatedTo ?id. ?id csdm:type ?type.} FILTER(?type IN (ProductionSegment, ProdutionProcess))

Which predictive function can be applied to optimize the given process segment?	SELECT ?function WHERE {
	?function csdm:influences ?kpi
	?kpi csdm:relatedTo <segment ID>.}

Which predictive function can be applied to optimize the given KPIs?	SELECT ?function WHERE {
	?function dmop:specifiesOutputType ?model
	?model csdm:influences <KPI ID>.}

Which data elements are influencing the given data element?	SELECT ?element WHERE {
	?dep csdm:source <element ID>.
	?dep csdm:target ?element.}

What is the impact of the predictive function quality to the given KPIs?	SELECT ?dep WHERE {
	?dep csdm:source ?stats.
	?dep csdm:target <KPI ID>.
	?stats csdm:influences <function ID>.}

Which data sets can be used to build the given predictive function?	SELECT ?dataset WHERE {
	?dataset csdm:isComposedOf ?physelm.
	?element csdm:isRealizedBy ?physelm.
	{<function ID> dmop: specifiesInputType ?element}

Can we validate that all data elements required to apply predictive function are available?	SELECT ?physelm WHERE { ?elm csdm:isRealizedBy ?physelm <function ID> dmop:specifiesInputType ?element.} It has to be a binding to physical element (i.e., not empty result) for all logical inputs of the predictive functions

Besides the evaluation of the ontology expressiveness using competency questions, we also designed and performed tests aimed to evaluate how the designed ontology can support communication between the domain experts and data scientists and how the ontology can be used for automatization of the data mining tasks. In the first case (evaluation of the efficiency of communication), we performed two experiments [52].

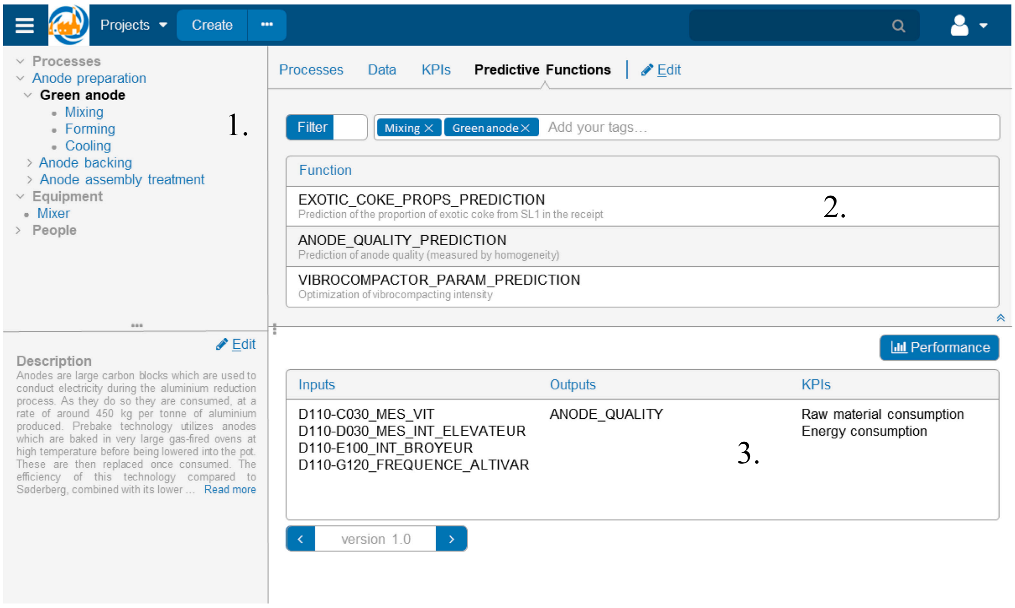
In the first experiment, we divided the domain experts into two groups: (1) Documenting the results of the communication using the documentation based on the textual documents and the graphical schemas for the process models, and (2) using the semantic framework to capture and directly formalize the domain knowledge in the cross-sectorial domain model. Both groups of domain experts were supported by the same team of data scientists. The goal of both groups was to complete the phases of problem understanding and data understanding (modelling phase and model evaluation phase were not covered in this experiment). After the data scientists and domain experts produced documentation and semantic models, the textual documents were manually annotated by the data scientists to extract the vocabulary of domain concepts described in the documents and relationships between these concepts explicitly defined in the text sentences. The list of concepts extracted from the textual documentation documented by the first group was then compared with the concepts modelled directly using the cross-sectorial domain model by the second group of domain experts. The comparison was then evaluated using various metrics, such as precision and recall, to evaluate the coverage of the knowledge in both forms of the documentation. The conclusion was that the semantic model fully covers the textual documentation (i.e., good recall) with some issues related to the process modelling (i.e., lower precision for process models). The issues were mainly related to using the wrong terminology for some process activities (e.g., using of a term describing the tool instead of a verb term describing the tool usage) and wrong decomposition of the processes to sub-activities and sub-processes.

In the second experiment, we divided the data scientists into two groups and both groups were working with the same group of domain experts from the two different domains (i.e., aluminum production and plastic production). The task for the groups was again complete the problem and to understand data phases in cooperation with the domain experts to use semantic models to document the domain knowledge. Each group of data scientists were working on one domain: one group on the aluminum use case and one on the plastic case. After the groups finished the phases of problem and data understanding, the groups exchanged their domains and continued with the phases of data pre-processing, modelling, and validation. One group that described the problem and data for the plastic domain the completed the pre-processing, modelling, and data evaluation for aluminum domain and vice versa. Note that the data scientists did not have any previous experience with both domains, so their understanding of the problem and data was based solely on the documentation expressed by the other group. In this case, we measured how much information was missing in the semantic model by formulating the list of competency questions asked by each group during the phases of pre-processing, modelling, and validation. The result was that both groups were able to correctly identify the problem and how it can be transformed to the data mining task, which predictive functions will be built from data, and which KPIs will be optimized by these functions. The missing knowledge formulated in the competency questions was mainly related to the dependencies between the data attributes and to a dependency between the KPIs in the validation phase. However, all missing competency questions were covered in the semantic model; it was possible to add missing dependencies without extending the cross-sectorial domain model conceptualization.

7. Semantic Modeller

To improve the effectiveness of communication between the experts and to implement the shared knowledge base using the semantic model described in the previous sections, we designed and implemented the prototype of the semantic modeling tools customized for the proposed model. The tool is primarily aimed to be used collaboratively by the domain experts and data scientists. It consists of a graphic interface enabling to describe the particular industrial domain characteristics, e.g., draw the diagrams of the processes, describe the equipment and personnel related to the processes and process steps, annotate the data produced by those elements, etc. On the other hand, besides the knowledge base in human-readable form stored in the tool, also machine-readable representation of the domain model is created on the back-end.

The web-based user interface of the semantic tools is presented in Figure 8. The layout of the interface is divided into the navigation panel (1), detailed contextual view (2), and main content view (3). The navigation panel consists of the navigation tree presenting the domain concepts organized according to the main types of views (processes, equipment, and people). The detailed contextual view provides an additional description of the domain concepts selected in the navigation panel. The navigation panel also provides the context for the main view and filter information related to the selected domain concepts (i.e., the structure of the domain part of the semantic model is always visible for the experts for navigation). The main views correspond to the main concepts and present the semantic properties and relations for processes, data (elements), KPIs, and predictive functions.



Description

Anodes are large carbon blocks which are used to conduct electricity during the aluminium reduction process. As they do so they are consumed, at a rate of around 450 kg per tonne of aluminium produced. Prebake technology utilizes anodes which are baked in very large gas-fired ovens at high temperature before being lowered into the pot. These are then replaced once consumed. The efficiency of this technology compared to Søderberg, combined with its lower ... Read more

Function
EXOTIC_COKE_PROPS_PREDICTION Prediction of the proportion of exotic coke from SL1 in the receipt
ANODE_QUALITY_PREDICTION Prediction of anode quality (measured by homogeneity)
VIBROCOMPACTOR_PARAM_PREDICTION Optimization of vibrocompacting intensity

Inputs	Outputs	KPIs
D110-C030_MES_VIT D110-D030_MES_INT_ELEVATEUR D110-E100_INT_BROYEUR D110-G120_FREQUENCE_ALTIVAR	ANODE_QUALITY	Raw material consumption Energy consumption

Figure 8. The user interface of the modelling tools designed for the proposed semantic model. The figure shows predictive functions view.

7.1. Process View

In the process view, the user can use graphical process editor to model production processes. The graphical notation is based on the simplified BPMN diagrams that consist of the main building blocks: start/end node, activity (representing the step of the production process), generic gateway (branching of the process to parallel phases), and sub-processes. Process view is synchronized with the navigation bar, i.e., the user can select an element in the navigation bar, and process view will switch to the process/sub-process and select the element. A description of the selected element is provided in the detailed contextual view.

7.2. Data and KPIs View

The data and KPIs view provides a list of data elements and KPIs defined for the specified part of the production process or resource. For data elements, it provides information about the role of the data attribute, which can be used as the input/output to the predictive function, measurement units, and data type. The important part of the view is the filter, which allows selecting only the subset of the elements relevant to the specified domain concepts. As such, users can create a contextual view of the specific data analytics task targeted at the optimization of the production step or equipment. Additionally, the user can annotate each data element with arbitrary tags and use these annotations to further filter the presented information.

7.3. Predictive Functions View

Similar to the data view, predictive functions view provides the list of semantic elements describing the predictive functions specified by the data scientists. A list of functions can be filtered according to the context specified by tags and domain concepts. For the selected function, the view provides an overview of the main information about the function such as table summarizing inputs, outputs, and influenced KPIs.

8. Case Study

This section presents the application of the cross-sectorial domain model to one of the project's application domains: aluminum production. The MONSOON aluminum use case was evaluated at AP Dunkerque smelter, France, which has the highest primary aluminum production in the EU.

Currently, the aluminum production process can be considered highly-mature, but several areas still could benefit from the application of modern data analytics.

The aluminum production process depicted in Figure 9 consists of several main steps. Aluminum is produced using an electrolysis process in the potline. The potline consists of a set of several hundred pots where liquid aluminum is being produced using electrolysis. The main inputs to the process are electricity, alumina (Al_2O_3, obtained from bauxite), anodes (produced on-site, in the carbon plant), and an electrolytic bath [53]. Each pot is equipped with anodes and cathodes. Anodes are consumed during the process, and therefore they have to be continually replaced. Anodes and their quality are some of the most critical and controllable inputs for the electrolysis. When using the cross-sectorial domain model and simple BPMN graphical notation as a production process consisting of process segments, top-level aluminum production process could be modeled, as depicted in Figure 10.

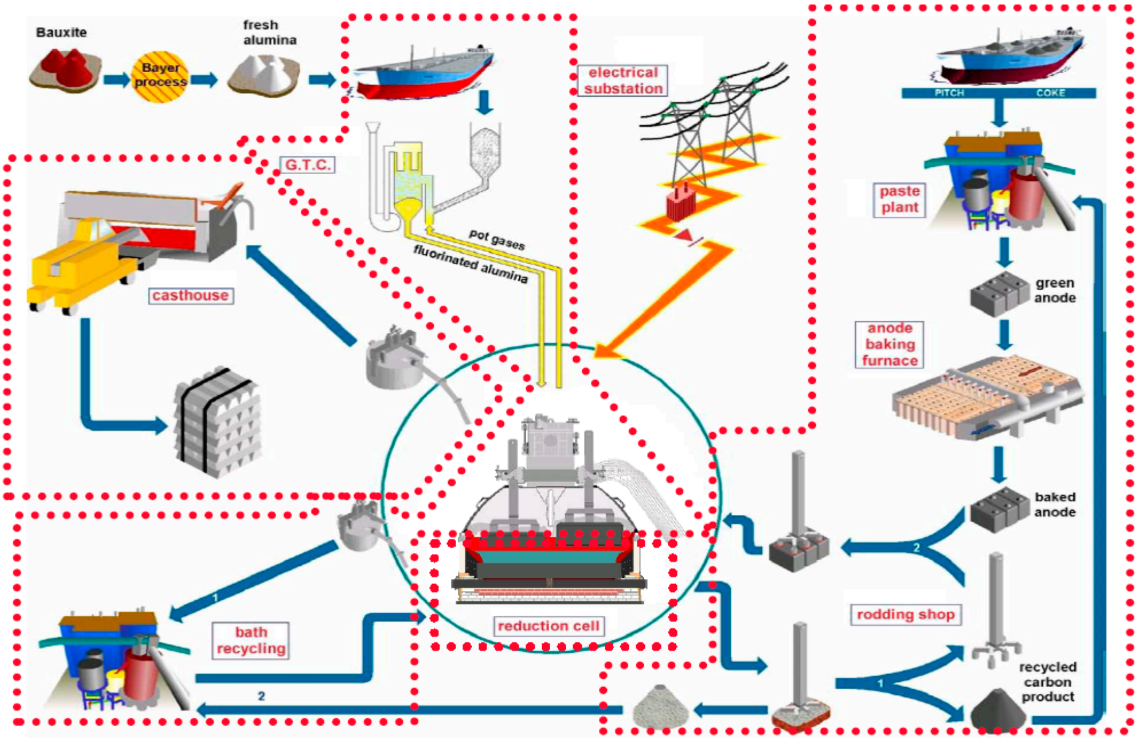


Figure 9. Aluminum production process.

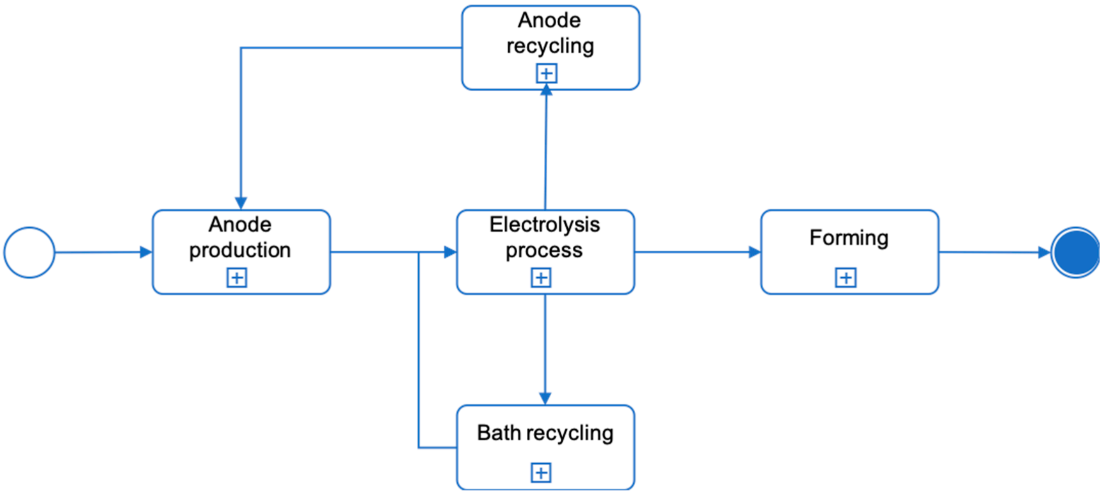


Figure 10. Top-level aluminum production process model.

Anodes are produced in the carbon plant. Anode production consists of three stages:

(1) Green anodes production involves preparation of paste and forming of the anodes before their baking in the furnace.
(2) Anode baking: baked anodes are produced from green anodes in chambers of the baking furnace.
(3) Anode rodding: rodded anodes are produced from baked anodes by splicing them together with stems, producing the final product of the carbon plant stage, and are distributed to electrolysis phase.

Anodes production can be modeled as a process segment as a separate sub-process, as depicted in Figure 11.

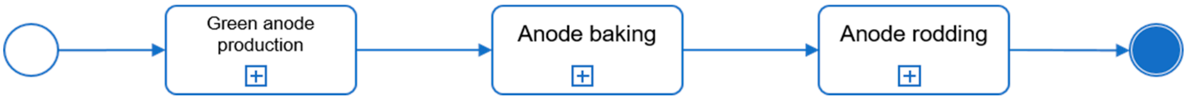


Figure 11. Anode production process model.

At the production segment level, each of the anode production process segments can be hierarchically modeled in a similar fashion as the separate processes. The following example in Figure 12 presents the sub-process diagram for green anode production segment. Green anodes are produced from calcined petroleum-coke, coal-tar pitch and recycled scraps, and anode butts. Such ingredients are prepared and dosed. The material is then processed (including milling, screening, proportioning, and mixing of the grain), which results in a paste used to form the anodes.

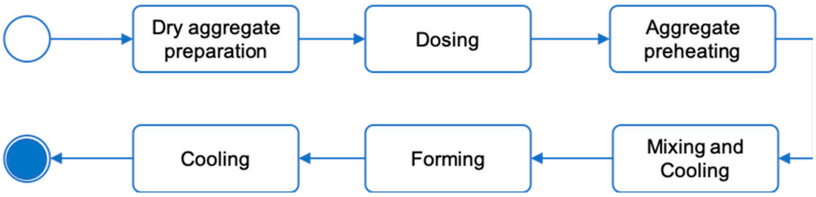


Figure 12. Green anode production process model.

Each of the process steps on each level of the structure can be described by specifying the relation with the equipment used in this process. We can specify the particular device and its relation to the process segment (e.g., the buss mixer is related to the mixing phase) and define the concept hierarchy for the devices (e.g., define the mixer as a hydraulic device type). Equipment can be described using several parameters (input or output). As an example, the buss mixer device can be characterized using a mixer power parameter. A parameter, in general, is specified by corresponding data element, which defines the units, description, type, and tags for full-text search in the modeler tool (e.g., the mixer power attribute is an input, which means that it can be set up by personnel on the plant site). Then, the parameter can be tied (using the *represent* relation) to the concrete data element in a specific data set. The data set is described similarly and uses the concepts from DMOP ontology.

A KPI is described using a specification of a dependency relation between the data element and a KPI element. In this scenario, we specified an energy consumption KPI that represents the overall energy consumption of the green anode production process. The buss mixer, as a device involved in one of the process' segments, influences the given KPI. This contribution can be represented using the definition of a dependency between the energy consumption element and the buss mixer power consumption data element.

The random forest (RF) classification model can be used in a predictive maintenance scenario to anticipate the breakdown of a buss mixer during its operation. This model can be described by the DMOP concepts using the semantic model. The model itself is an output of a RF algorithm (univariate decision tree algorithm) applied on a concrete dataset (Dataset1) and specified using a set of different settings, e.g., maximum depth of trees in forest and minimal number of records in a split or leaves, and has several quality attributes. The algorithm was applied using concrete values of the algorithm parameters, which are specified as individuals of the *Parameter* class, e.g., *Min_samples_leaf* parameter instance was set to an integer value of 1, the *Min_samples_split* parameter was set to 5, and *Max_depth* was set to none (which specifies that the parameter was not considered in model training). The model was properly trained to anticipate the breakdowns of the buss mixer, which directly influences the energy consumption of the device (and therefore the process as a whole), as depicted in Figure 13.

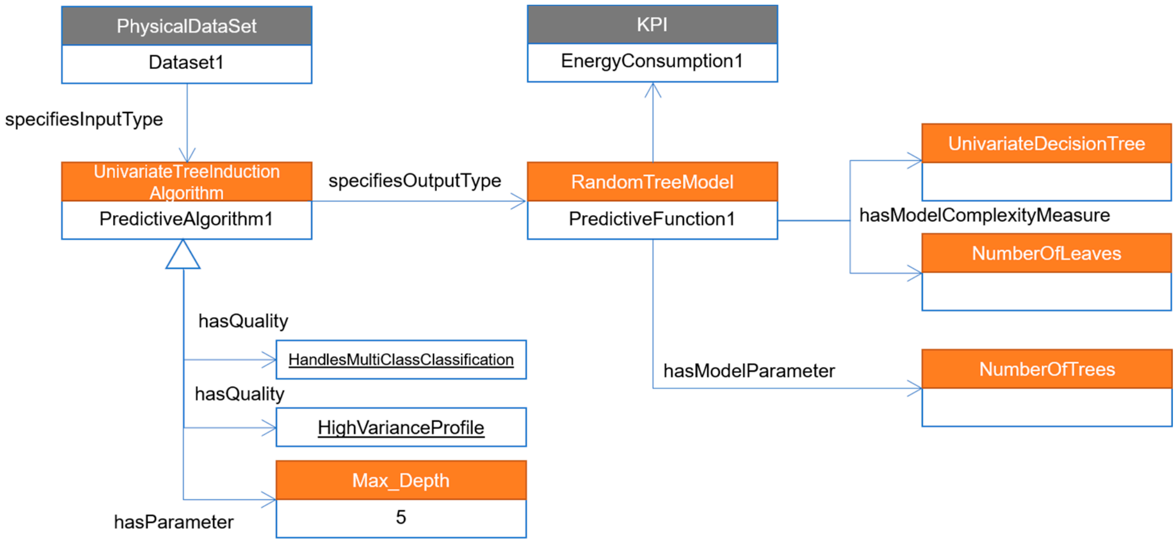


Figure 13. Random forest predictive function model.

The described scenario and the presented diagrams can be represented using the semantic model in a machine-readable format in JSON-LD notation. Appendix A contains the example JSON-LD documents for each of the presented elements.

9. Conclusions

In this paper, we proposed a unified ontology for the formalization of data analytics processes applied for optimization in the process industry. This ontology promotes the sharing of best practices and results within and between domain experts and data scientists, reducing both the duplication and loss of knowledge. It is an essential step in formalizing the data analytics methods and how they can be applied to the optimization of production processes. The proposed ontology has modular architecture divided into a domain-specific part and data analytics part. The domain-specific part specifies concepts for the description of production processes, equipment, human resources, and overall KPIs. The data analytics part specifies concepts for data representation, pre-processing, and modeling of the predictive functions. Both parts are interlinked in one unified semantic framework. The proposed semantic framework was developed using sound methodology with emphasis on the usability of the semantic models in various user scenarios covering most of the data analytics phases from problem understanding to model validation. Besides the formal specification of the semantic models, we also designed and implemented modeling tools that allow collaborative semantic modeling and sharing of knowledge between domain experts and data scientists. An important feature of the modeling tools is that they are designed for users without skills in semantic technology. In future work, we would like to further extend our model with functional modeling of the dependencies between data elements and KPIs. Such extensions will allow automatic reasoning and computation of the impact of the data analytics methods on the overall performance of the production process assessed by the specified KPIs.

Author Contributions: Conceptualization, (M.S.) Martin Sarnovsky and P.B.; Methodology, P.B.; Resources, (M.S.) Martin Sarnovsky; Software, (M.S.) Miroslav Smatana; Supervision, P.B.; Validation, (M.S.) Martin Sarnovsky; Writing—original draft, (M.S.) Martin Sarnovsky and P.B.

Funding: This research was funded by H2020-SPIRE-2016 Project MONSOON “MOdel based control framework for Site-wide OptimizatioN of data-intensive processes” under grant agreement No. 723650.

Acknowledgments: This work was supported by Slovak APVV research grant APVV-16-0213 and SK-AT-2017-0021.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

This section shows the JSON-LD representation for the examples described in Section 8. The following JSON-LD document represents the green anode production sub-process model specifying its particular steps (*narrower* relation).

```
{
“uri”: “ https://monsoon.ekf.tuke.sk/modeller/process-GreenAnodeProduction”,
“@type”: “Process”,
“@context”: “ https://gbv.github.io/jskos/context.json”,
“hasExecutionDependency”: ““,
“partOf”: ““,
“narrower”: [
“ https://monsoon.ekf.tuke.sk/modeller/DryAggregatePreparation”,
“ https://monsoon.ekf.tuke.sk/modeller/Dosing”,
“ https://monsoon.ekf.tuke.sk/modeller/AggregatePreheating”,
“ https://monsoon.ekf.tuke.sk/modeller/MixingAndMilling”,
“ https://monsoon.ekf.tuke.sk/modeller/Forming”,
“ https://monsoon.ekf.tuke.sk/modeller/Cooling”
],
“broader”: ““,
“next”: ““,
“previous”: ““,
“definition”: {
“en”: “Process of manufacturing of green anodes consisting of screening, classification and mixing of the grain into the paste and vibro-compaction of the paste.”
},
“identifier”: “GreenAnodeProduction”,
“prefLabel”: {
“en”: “Green anode production”
}
}
```

The following JSON-LD code describes the mixing and milling step in the green anode production process. Subsequently, we present a model for buss mixer equipment.

```
{
“uri”: “ https://monsoon.ekf.tuke.sk/modeller/MixingAndMilling”,
“@type”: “Process”,
“@context”: “ https://gbv.github.io/jskos/context.json”,
“hasExecutionDependency”: [
“ https://monsoon.ekf.tuke.sk/modeller/DryAggregatePreparation”,
“ https://monsoon.ekf.tuke.sk/modeller/Dosing”,
“ https://monsoon.ekf.tuke.sk/modeller/AggregatePreheating”
```

],
“partOf”: “ https://monsoon.ekf.tuke.sk/modeller/process-GreenAnodeProduction”,
“broader”: “ https://monsoon.ekf.tuke.sk/modeller/process-GreenAnodeProduction”,
“next”: “ https://monsoon.ekf.tuke.sk/modeller/Forming”,
“previous”: “ https://monsoon.ekf.tuke.sk/modeller/AggregatePreheating”,
“definition”: {
“en”: “Mixing and of the coke, recycled anodes with pitch and dedusting products into a uniform anode paste, which will be used to form the anodes. The produced paste is cooled and the subsequently used in Forming phase. Anodes will be formed using vibrocompaction.”
},
“identifier”: “Mixing”,
“related”: “ https://monsoon.ekf.tuke.sk/modeller/equipment-BUSS_Mixer”,

}
{
“uri”: “ https://monsoon.ekf.tuke.sk/modeller/equipment-BUSS_Mixer”,
“@type”: “Equipment”,
“@context”: “ https://gbv.github.io/jskos/context.json”,
“partOf”: “ https://monsoon.ekf.tuke.sk/modeller/MixingAndMilling”,
“narrower”: “”,
“broader”: “ https://monsoon.ekf.tuke.sk/modeller/HydraulicMachine”,
“definition”: {
“en”: “BUSS Mixer is a machine used in mixing phase to process the raw materials with liquid pitch to produce the compact the paste used to form the anodes in Vibrocompactor.”
},
“identifier”: “BUSS_Mixer”,
“prefLabel”: {
“en”: “BUSS Mixer mixing machine.”
},
“related”: “ https://monsoon.ekf.tuke.sk/modeller/MixingAndCooling”
}

The following elements demonstrate the description of the buss mixer power consumption parameter and its relation to the physical data set described using the concepts from DMOP ontology.

{
“uri”: “ https://monsoon.ekf.tuke.sk/modeller/data-BUSS_MIXER_POWER”,
“@context”: “ https://gbv.github.io/jskos/context.json”,
“@type”: “DataElement”,
“unit”: “kW”,
“role”: “Input”,
“prefLabel”: {
“en”: “BUSS mixer power consumption.”
},
“tags”: [
“mixing”,
“green anode production”
],
“definition”: {
“en”: “Data element containing the mean values of mixer power.”
},

```
“identifier”: “BUSS-mixer-power”,
“related”: [
“ https://monsoon.ekf.tuke.sk/modeller/MixingAndCooling”,
“ https://monsoon.ekf.tuke.sk/modeller/equipment-BUSS_Mixer”
]
}
{
“uri”: “ https://monsoon.ekf.tuke.sk/modeller/data-D110-J160_PUISSANCE_MOY_MALAXEUR”,
“@context”: “ https://gbv.github.io/jskos/context.json”,
“@type”: “PhysicalDataElement”,
“represents”: “ https://monsoon.ekf.tuke.sk/modeller/data-BUSS_MIXER_POWER”,
“identifier”: “D110-J160_PUISSANCE_MOY_MAL AXEUR”,
“definition”: {
“en”: “Values of the mean mixer power, time series, acquisition frequency 5s.”
},
“related”: [
“ https://monsoon.ekf.tuke.sk/modeller/equipment-BUSS_Mixer”,
              “ https://monsoon.ekf.tuke.sk/modeller/dataset1”
]
}
{
“uri”: “ https://monsoon.ekf.tuke.sk/modeller/DataSet1”,
“@context”: “ https://gbv.github.io/jskos/context.json”,
“@type”: [
“PhysicalDataset “,
“DMOP: DataSet”
],
“definition”: {
“en”: “Sensor data containing from anode density info from September 2018, dimensions, coke quality, production time, overall 46 attributes.”
},
“isComposedOf”: “ https://monsoon.ekf.tuke.sk/modeller/data-D110-J160_PUISSANCE_MOY_MALAXEUR”,
“location”: [
“hdfs://monsoon.ekf.tuke.sk/data/examples/dataset1.csv”
]
}
```

The following elements present the KPI modeling by specifying the energy consumption KPI and the dependency specification between the given KPI and the buss mixer power consumption data element.

```
{
“@id”: “ https://monsoon.ekf.tuke.sk/modeller/EnergyConsumption1”,
“@type”: “KPI”,
“title”: “Energy consumption of the device.”,
“definition “: “Overall electrical energy consumption of the production process.”,
“quantity”: “kW/h”
}
{
“@type”: “Dependency”,
“definition”: “Higher mean mixer power leads to higher power consumption.”,
“from”: “ https://monsoon.ekf.tuke.sk/modeller/data-BUSS_mixer_power”,
“to”: “ https://monsoon.ekf.tuke.sk/modeller/EnergyConsumption1”,
“directed”: true
}
```

The predictive functions are modeled using concepts from DMOP ontology. This example describes the RF model, which is a result of *PredictiveAlgorithm1* (of *UnivariateDecisionTree* type). The algorithm is applied to specified data and produces a model. The model can influence a given KPI (*EnergyConsumption1* KPI in this case).

```
{
“@id”: “ https://monsoon.ekf.tuke.sk/modeller/PredictiveFunction1”,
“@context”: “ http://www.e-lico.eu/ontologies/dmo/DMOP/DMKB.owl#”,
“@type”: “RandomTreeModel”,
“hasHypothesisStructure”: “UnivariateDecisionTree”,
“hasModelComplexityMeasure”: “NumberOfLeaves”,
“hasParameter”: “NumberOfTrees”,
“title”: “BUSS Mixer power predictive model”,
“definition”: {
“en”: “Model for predicting of BUSS Mixer power consumption during the 8h shifts in Mixing and milling phase. Target attribute represents the 2 values representing normal operation parameter and.”
},
“isRealizedBy”: “ https://monsoon.ekf.tuke.sk/modeller/PredictiveAlgorithm1”,
“influences”: “ https://monsoon.ekf.tuke.sk/modeller/EnergyConsumption1”
}
{
“@id”: “ https://monsoon.ekf.tuke.sk/modeller/PredictiveAlgorithm1”,
“@context”: {
“DOLCE-Lite”: “ http://www.loa-cnr.it/ontologies/DOLCE-Lite#”
},
“@type”: “DMOP: UnivariateTreeInductionAlgorithm”,
“DOLCE-Lite:has-quality”: [
“DMOP: HandlesMulticlassClassification”,
“DMOP: HighVarianceProfile”
],
“DMOP:hasFeatureTestEvaluator”: “DMOP: GiniIndex”,
“DMOP:hasParameter”: [
{
“DMOP:Max_Depth”: {
“@id”: “Max_Depth1”
}
},
“DMOP:InformationGainRatio”
],
“DMOP:specifiesInputType”: “hdfs://monsoon.ekf.tuke.sk/data/examples/dataset1.csv”,
“DMOP:specifiesOutputType”: “ http://www.e-lico.eu/ontologies/dmo/DMOP/DMKB.owl# PredictiveFunction1”,
“title”: “Predictive Algorithm1”,
“definition”: {
```

```
"en": "Random Forest algorithm used to train the Predictive Function for prediction of the BUSS Mixer energy consumption for green anode production process optimization."
}
}
```

The model is described using model parameters, and the algorithm by algorithm parameters. Both are represented by ontology instances; the following example presents the instantiation of maximum depth of tree parameter set to a concrete value.

```
{
"@id": " https://monsoon.ekf.tuke.sk/modeller/Max_Depth",
"@context": {
"DMOP": " http://www.e-lico.eu/ontologies/dmo/DMOP/DMKB.owl#"
},
"@type": "AlgorithmParameter",
"hasDataValue": 5
}
```

References

1. Thomas, A.; Lamouri, S. Industrial Management in the Process Industry. *IFAC Proc. Vol.* **1998**, *31*, 841–846. [CrossRef]
2. Reh, L. Challenges for process industries in recycling. *China Particuol.* **2006**, *4*, 47–59. [CrossRef]
3. Van Donk, D.P.; Fransoo, J.C. Operations management research in process industries. *J. Oper. Manag.* **2006**, *24*, 211–214. [CrossRef]
4. Scott-Young, C.; Samson, D. Project success and project team management: Evidence from capital projects in the process industries. *J. Oper. Manag.* **2008**, *26*, 749–766. [CrossRef]
5. Sarnovsky, M.; Bednar, P.; Smatana, M. Big Data Processing and Analytics Platform Architecture for Process Industry Factories. *Big Data Cogn. Comput.* **2018**, *2*, 3. [CrossRef]
6. Shearer, C.; Watson, H.J.; Grecich, D.G.; Moss, L.; Adelman, S.; Hammer, K.; Herdlein, S. The CRISP-DM model: The New Blueprint for Data Mining. *J. Data Warehous.* **2000**, *5*, 13–22.
7. Shafique, U.; Qaiser, H. A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). *Innov. Space Sci. Res.* **2014**, *12*, 217–222.
8. Azevedo, A.; Santos, M.F. KDD, SEMMA and CRISP-DM: A parallel overview. In Proceedings of the IADIS European Conference Data Mining, Amsterdam, The Netherlands, 24–26 July 2008; pp. 182–185.
9. Wirth, R. CRISP-DM: Towards a Standard Process Model for Data Mining. In Proceedings of the Fourth International Conference on the Practical Applications of Knowledge Discovery and Data Mining, Manchester, UK, 11–13 April 2000; pp. 29–39.
10. ISA. *Enterprise—Control System Integration Part 1: Models and Terminology*; ISA: Research Triangle Park, NC, USA, 2000; 1999; ISBN 1-55617-727-5.
11. Gould, L.S. B2MML Explained. *Automot. Des. Prod.* **2007**, *119*, 54.
12. American National Standard. *ANSI/ISA-88.01 Batch Control Part 1: Models and Terminology*; ISA: Research Triangle Park, NC, USA, 1995; Volume 1, ISBN 1-55617-562-0.
13. Vegetti, M.; Henning, G. ISA-88 formalization. In A step towards its integration with the ISA-95 standard. In Proceedings of the CEUR Workshop Proceedings, Riva del Garda, Italy, 19 October 2014; 2014; Volume 1333.
14. Vieille, J. A Meta-Model for Leveraging the ISA-88/95/106 Standards. Available online: https://www.researchgate.net/publication/296332226_A_meta-model_for_leveraging_the_ISA-8895106_standards (accessed on 20 March 2019).
15. Lemaignan, S.; Siadat, A.; Dantan, J.Y.; Semenenko, A. MASON: A proposal for an ontology of manufacturing domain. In Proceedings of the DIS 2006: IEEE Workshop on Distributed Intelligent Systems—Collective Intelligence and Its Applications, Prague, Czech Republic, 15–16 June 2006; Volume 2006, pp. 195–200.

16. Kharlamov, E.; Grau, B.C.; Jiménez-Ruiz, E.; Lamparter, S.; Mehdi, G.; Ringsquandl, M.; Nenov, Y.; Grimm, S.; Roshchin, M.; Horrocks, I. Capturing industrial information models with ontologies and constraints. In Proceedings of the Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Kobe, Japan, 17–21 October 2016; 2016; Volume 9982, pp. 325–343.
17. Pakonen, A.; Tommila, T.; Hirvonen, J. A fuzzy ontology based approach for mobilising industrial plant knowledge. In Proceedings of the Proceedings of the 15th IEEE International Conference on Emerging Technologies and Factory Automation, ETFA 2010, Bilbao, Spain, 13–16 September 2010.
18. Cheng, H.; Zeng, P.; Xue, L.; Shi, Z.; Wang, P.; Yu, H. Manufacturing Ontology Development Based on Industry 4. In 0 Demonstration Production Line. In Proceedings of the 2016 Third International Conference on Trustworthy Systems and their Applications (TSA), Wuhan, China, 18–22 September 2016; pp. 42–47.
19. Gurjanov, A.V.; Zakoldaev, D.A.; Shukalov, A.V.; Zharinov, I.O. The ontology in description of production processes in the Industry 4.0 item designing company. *J. Phys. Conf. Ser.* **2018**, *1059*, 012010. [CrossRef]
20. Järvenpää, E.; Siltala, N.; Hylli, O.; Lanz, M. The development of an ontology for describing the capabilities of manufacturing resources. *J. Intell. Manuf.* **2019**, *30*, 959–978. [CrossRef]
21. Fraga, A.L.; Vegetti, M.; Leone, H.P. Semantic Interoperability among Industrial Product Data Standards using an Ontology Network. In Proceedings of the 20th International Conference on Enterprise Information Systems, Madeira, Portugal, 21–24 March 2018; SCITEPRESS—Science and Technology Publications: Funchal, Portugal, 2018; pp. 328–335.
22. Giustozzi, F.; Saunier, J.; Zanni-Merk, C. Context Modeling for Industry 4.0: An Ontology-Based Proposal. *Procedia Comput. Sci.* **2018**, *126*, 675–684. [CrossRef]
23. Bandrowski, A.; Brinkman, R.; Brochhausen, M.; Brush, M.H.; Bug, B.; Chibucos, M.C.; Clancy, K.; Courtot, M.; Derom, D.; Dumontier, M.; et al. The Ontology for Biomedical Investigations. *PLoS ONE* **2016**, *11*, e0154556. [CrossRef]
24. Soldatova, L.N.; King, R.D. An ontology of scientific experiments. *J. R. Soc. Interface* **2006**, *3*, 795–803. [CrossRef]
25. Pease, A.; Niles, I.; Li, J. The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications. *Imagine* **2002**, *28*, 7–10.
26. Masolo, C.; Borgo, S.; Gangemi, A.; Guarino, N.; Oltramari, A.; Schneider, L. DOLCE: A descriptive ontology for linguistic and cognitive engineering. *WonderWeb Proj. Deliv. D17 V2, Tech. Rep.* **2002**, *1*, 2–3.
27. Keet, C.M.; Ławrynowicz, A.; D'Amato, C.; Kalousis, A.; Nguyen, P.; Palma, R.; Stevens, R.; Hilario, M. The Data Mining OPtimization Ontology. *J. Web Semant.* **2015**, *32*, 43–53. [CrossRef]
28. Panov, P.; Soldatova, L.; Džeroski, S. OntoDM-KDD: Ontology for representing the knowledge discovery process. In Proceedings of the Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Paphos, Cyprus, 28–30 November 2012; 2013; Volume 8140, pp. 126–140.
29. Vanschoren, J.; Soldatova, L.N. Exposé: An ontology for data mining experiments. In Proceedings of the SoKD 2010—Third Generation Data Mining Workshop at ECML PKDD, Barcelona, Spain, 20–24 September 2010; 2010; pp. 31–46.
30. Patterson, E.; Baldini, I.; Mojsilović, A.; Varshney, K.R. Semantic Representation of Data Science Programs. In Proceedings of the Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–18 July 2018; International Joint Conferences on Artificial Intelligence Organization: Stockholm, Sweden, 2018; pp. 5847–5849.
31. Patterson, E.; McBurney, R.; Schmidt, H.; Baldini, I.; Mojsilovic, A.; Varshney, K.R. Dataflow representation of data analyses: Toward a platform for collaborative data science. *IBM J. Res. Dev.* **2017**, *61*, 9:1–9:13. [CrossRef]
32. Patterson, E.; Baldini, I.; Mojsilovic, A.; Varshney, K.R. Teaching machines to understand data science code by semantic enrichment of dataflow graphs. *arXiv* **2018**, arXiv:1807.05691.
33. Pechter, R. What's PMML and what's new in PMML 4.0? *ACM SIGKDD Explor. Newsl.* **2009**, *11*, 19. [CrossRef]

34. Pivarski, J.; Bennett, C.; Grossman, R.L. Deploying Analytics with the Portable Format for Analytics (PFA). In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD '16, San Francisco, CA, USA, 13–17 August 2016; ACM Press: New York, NY, USA, 2016; pp. 579–588.
35. Gil, R.; Martin-Bautista, M.J. SMOL: A systemic methodology for ontology learning from heterogeneous sources. *J. Intell. Inf. Syst.* **2014**, *42*, 415–455.
36. Fox, M.S.; Gruninger, M. Enterprise modeling. *AI Mag.* **1998**, *19*, 109.
37. Panov, P.; Soldatova, L.N.; Džeroski, S. Generic ontology of datatypes. *Inf. Sci.* **2016**, *329*, 900–920. [CrossRef]
38. Aagesen, G.; Krogstie, J. BPMN 2.0 for Modeling Business Processes. In *Handbook on Business Process Management 1*; vom Brocke, J., Rosemann, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2015; pp. 219–250. ISBN 978-3-642-45099-0.
39. Miles, A.; Pérez-Agüera, J.R. SKOS: Simple Knowledge Organisation for the Web. *Cat. Classif. Q.* **2007**, *43*, 69–83. [CrossRef]
40. Soediono, B. Media Types for Sensor Markup Language (SENML). *Netw. Work. Group Ietf* **2013**, *53*, 160.
41. Su, X.; Zhang, H.; Riekki, J.; Keränen, A.; Nurminen, J.K.; Du, L. Connecting IoT sensors to knowledge-based systems by transforming SenML to RDF. *Procedia Comput. Sci.* **2014**, *32*, 215–222. [CrossRef]
42. Huang, C.Y.; Wu, C.H. Design and implement an interoperable Internet of Things application based on an extended OGC sensorthings API Standard. In Proceedings of the International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences—ISPRS Archives, Prague, Czech Republic, 12–19 July 2016; Volume 41, pp. 263–266.
43. Bennett, M.; James, P. ISO 14031 and the Future of Environmental Performance Evaluation. In *Sustainable Measures: Evaluation and Reporting of Environmental and Social Performance*; Greenleaf Publishing Limited: Yorkshire, UK, 1999; pp. 75–97. ISBN 978-1-907643-19-4.
44. Bilgin, G.; Dikmen, I.; Birgonul, M.T. Ontology Evaluation: An Example of Delay Analysis. *Procedia Eng. Amst. Neth.* **2014**, *85*, 61–68. [CrossRef]
45. Brank, J.; Grobelnik, M.; Mladenić, D. A survey of ontology evaluation techniques. In Proceedings of the Conf. Data Min. Data Wareh. 2005; pp. 166–170.
46. Gomez-Perez, A. Some ideas and examples to evaluate ontologies. In Proceedings of the CAIA '95—11th Conference on Artificial Intelligence for Applications, Los Angeles, CA, USA, 20–23 February 1995; IEEE Comput. Soc. Press; pp. 299–305.
47. Bandeira, J.; Bittencourt, I.I.; Espinheira, P.; Isotani, S. FOCA: A Methodology for Ontology Evaluation. *arXiv* **2016**, arXiv:161203353.
48. Bouiadjra, A.B.; Benslimane, S.M. A framework for evaluating and ranking ontologies. *Int. J. Metadata Semant. Ontol.* **2013**, *8*, 155. [CrossRef]
49. Gómez-Pérez, A. Evaluation of ontologies. *Int. J. Intell. Syst.* **2001**, *16*, 391–409. [CrossRef]
50. Ren, Y.; Parvizi, A.; Mellish, C.; Pan, J.Z.; Van Deemter, K.; Stevens, R. Towards competency question-driven ontology authoring. In *Proceedings of the Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 2014 May 25*; Springer: Cham, Switzerland, 2014; Volume 8465, pp. 752–767.
51. Harris, S.; Seaborne, A.; SPARQL 1.1 Overview 2013. W3C recommendation 21 March 2013. Available online: https://www.w3.org/TR/sparql11-overview/ (accessed on 20 March 2019).
52. Sarnovsky, M.; Bednar, P.; Miksa, T. Semantic model for description of process industries domain, Knowledge modelling for data analytical processes. In Proceedings of the Knowledge Modelling for Data Analytical Processes workshop; TU Kosice, Slovakia, 27–29 June 2018; 2018; pp. 13–17.
53. Chareyre, M.; Jolas, J.-M.; Praizelin, N.; Guillaud, V.; Richiardone, M.; Muhammad, A.; Schlutter, R.; Gelhen, M.; Dias, M.; Silva, A.; et al. Initial Process Industry Domain Analysis and Use Cases—Project Deliverable 2016. Available online: https://www.spire2030.eu/sites/default/files/users/user475/D2.2%20Process%20Industry%20Domain%20Analysis%20and%20Use%20Cases%20v1.3.pdf (accessed on 20 March 2019).