# Towards an Ontology of Data Mining Investigations

Panče Panov[1], Larisa N. Soldatova[2], and Sašo Džeroski[1]

[1] Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana, Slovenia
{Pance.Panov,Saso.Dzeroski}@ijs.si
[2] Aberystwyth University, Penglais, Aberystwyth, SY23 3DB, Wales, UK
lss@aber.ac.uk

**Abstract.** Motivated by the need for unification of the domain of data mining and the demand for formalized representation of outcomes of data mining investigations, we address the task of constructing an ontology of data mining. In this paper we present an updated version of the OntoDM ontology, that is based on a recent proposal of a general framework for data mining and it is aligned with the ontology of biomedical investigations (OBI) . The ontology aims at describing and formalizing entities from the domain of data mining and knowledge discovery. It includes definitions of basic data mining entities (e.g., datatype, dataset, data mining task, data mining algorithm etc.) and allows extensions with more complex data mining entities (e.g. constraints, data mining scenarios and data mining experiments). Unlike most existing approaches to constructing ontologies of data mining, OntoDM is compliant to best practices in engineering ontologies that describe scientific investigations (e.g., OBI ) and is a step towards an ontology of data mining investigations. OntoDM is available at:http://kt.ijs.si/panovp/OntoDM/.

## 1 Introduction

Traditionally, ontology has been defined as the philosophical study of what exists: the study of kinds of entities in reality, and the relationships that these entities bear to one another [21]. In recent years use of term ontology has become prominent in the area of computer science research and the application of computer science methods in management of scientific and other kinds of information. In this sense the term ontology has the meaning of a standardized terminological framework in terms of which the information is organized.

The ontological problem is adopting a set of basic categories of objects, determining what kinds of entities fall within each of these categories of objects, and determining what relationships hold within and among different categories in the ontology. The ontological problem for computer science is identical to many of the problems in philosophical ontology, and the success of constructing such an ontology is achievable by applying methods, insights and theories of philosophical ontology. When one sets out to construct an ontology then, what one is doing is designing a representational artifact that is intended to represent the

J. Gama et al. (Eds.): DS 2009, LNAI 5808, pp. 257–271, 2009.

universals and relations amongst universals that exist, either in a given domain of reality (e.g data mining domain) or across such domains.

The engineering of ontologies is still a relatively new research field and some of the steps in ontology design remain manual and can be considered as an art by itself. Recently there was a significant progress in automatic ontology learning [14], application of text mining [17], and ontology mapping [13]. However the construction of a good quality ontology with the use of automatic and even semi-automatic techniques still requires manual definition of the key upper level entities of the domain of interest. Good practices in ontology development are: following an upper level ontology as a template, the use of formally defined relations between the entities and not allowing multiple inheritances [25].

In the domain of data mining and knowledge discovery, researchers have tried to construct ontologies describing data mining entities that were targeted to solve specific problems. Most of the developments are with the aim of automatic planning of data mining workflows [1,30,11,8]. Some of the developments are concerned with description of data mining services on the GRID [6,5].

Current proposals for ontology of data mining are not based on upper level categories nor have used a predefined set of relations based on a upper level ontology. Most of the semantic representations for data mining proposed so far are based on so called light-weight ontologies [15]. Light-weight ontologies are often shallow, without rigid relations between the defined entities, but they are relatively easy to develop by semi/automatic methods and they still greatly facilitate computer applications. The reason why these type of ontologies are more frequently developed then heavy-weight ontologies is that process of development is more difficult and time consuming. In contrast to many other domains, data mining requires elaborate inference over its entities, and hence requires rigid heavy-weight ontologies with the aim of improving the KDD (Knowledge Discovery in Databases) process and providing support for development of new data mining approaches and techniques.

While KDD and data mining have enjoyed great popularity and success in recent years, there is a distinct lack of a generally accepted framework that would cover and unify the data mining domain. The present lack of such a framework is perceived as an obstacle to the further development of the field. In [29], Yang and Wu collected the opinions of a number of outstanding data mining researchers about the most challenging problems in data mining research. Among the ten topics considered most important and worthy of further research, the development of an unifying framework for data mining is listed first. One step towards developing a general framework for data mining is constructing an ontology of data mining.

In this paper we propose an extended and updated version of the ontology of data mining named OntoDM. Our ontology design takes into consideration the best practices in ontology engineering. We use an upper level ontology BFO (Basic Formal Ontology)[1] to define the upper level classes, the OBO Relational

[1] BFO: `http://www.ifomis.org/bfo`

Ontology (RO)[2] to define the semantics of the relationships between the data mining entities, and provide is-a completeness and single is-a inheritance for all DM entities. We also developed our ontology in the most general fashion in order to be able to represent the complex entities in data mining that are becoming more and more popular research areas such as mining structured data and constraint-based mining.

In previous work [16] we presented an initial version of OntoDM sufficient for the representation of data mining tasks and complex data types. The ontology is based on the proposal for a general framework for data mining presented in [9]. The initial version of OntoDM was using the philosophy of Ontology of Scientific Experiments (EXPO) [26] and ontology of biomedical investigations (OBI)[3] for identification and organization of entities in a *is-a* class hierarchy.

The version described in the current paper has been sufficiently updated in several ways. First, the structure of the ontology was aligned with the top level structure of the OBI ontology. This procedure requested revising the representation of some data mining entities and also introduced new entities in the ontology (e.g., the entity data mining algorithm was split into three entities each capturing different dimension of a description; algorithm specification, algorithm implementation and algorithm description). Second, we extended the set of relations used in the initial version with relations defined in the OBI ontology in order to express the relations between informational entities, entities that are realized in a process and processes. Finally, we extended the OBI classes with data mining specific classes for describing complex entities (e.g., data mining scenarios, queries).

The rest of the paper is structured as follows: Section 2 provides the background for this work. Section 3 presents the ontology design principles and we provide a detailed description of the alignment with OBI ontology and description of upper level classes and relations. Section 4 presents an example of representation of a data mining algorithm in OntoDM based on the alignment with OBI ontology and Section 5 discusses the representation of complex data mining entities. In Section 6 we give a roadway for future research and development of the ontology.

## 2 Background

### 2.1 Motivation

The motivation for developing an ontology of data mining is multi-fold. Firstly, as it was mentioned in the introduction, the area of data mining is developing rapidly and one of the most challenging problems deals with developing a general framework for data mining. By developing an ontology of data mining we are taking one step towards solving this problem. The ontology would define and formalize what are the basic entities (e.g., dataset, data mining algorithm) in

[2] RO: `http://www.obofoundry.org/ro/`
[3] OBI: `http://obi-ontology.org/`

data mining and define the relations between the entities. After the basic entities are identified and defined, we can build upon them and define more complex entities (e.g. data mining query, data mining scenario and experiment). All the defined data mining entities organized in the form of an ontology would be a backbone of the systems for automated data mining.

Secondly, there exist several proposals for ontologies of data mining but all of them are light-weight, aimed at covering a particular use-case in data mining, are of a limited scope and highly use-case dependent. Data mining is a domain that needs a heavy-weight ontology with a broader scope, where much attention is paid to the rigorous meaning of each entity, semantically rigorous relations between entities and compliance to an upper level ontology and the domains of application (e.g., biology, environmental sciences).

Finally, an ontology of data mining should define what is the minimum information required for the description of a data mining investigation. Biology is leading the way in developing standards for recording and representation of scientific data and biological investigations (e.g., already more than 50 journals require compliance of papers reporting microarray experiments to the Minimum Information About a Microarray Experiment - MIAME standard). The researchers in the domain of data mining should follow this good practice and the ontology of data mining would support development of standards for performing and recording of data mining investigations.

### 2.2 State-of-the-Art

**Formalizing scientific investigations.** In recent years, there is an increased need for formalized representations of the domain of data mining and formal representation of outcomes of research in general. There exist several formalisms for describing scientific investigations and outcomes of research. In this part we will focus on two proposals that are relevant for describing data mining investigations: Ontology for Biomedical Investigations (OBI) and Ontology of Scientific Experiments (EXPO).

*Ontology of biomedical investigations - OBI.* The OBI(`http://obi-ontology.org/`) ontology aims to provide a standard for the representation of biological and biomedical investigations. OBI is developed through collaboration of 19 biomedical communities (transcriptomics, proteomics, metabolomics, etc.). They are developing a set of universal terms that are applicable across various biological and technological domains and domain specific terms relevant only to a given domain. The ontology supports consistent annotation of biomedical investigations regardless of particular field of the study. It aims to represent design of an investigation, the protocols and used instrumentation, used materials, generated data and type of analysis performed on it.

The OBI ontology employs rigid logic and semantics as it uses an upper level ontology BFO and the RO relations to define the top classes and a set of relations. OBI defines occurrences (processes) and continuants (materials, instruments, qualities, roles, functions) relevant to biomedical domains. OBI is fully

compliant with the existing formalisms in biomedical domains. OBI is a part of OBO Foundry [22] which requires all member ontologies follow the same design principles, the same set of relations, the same upper ontology, and to define a single class only once within OBO to facilitate integration and automatic reasoning.

The Data Transformation Branch is an OBI branch with the scope of identifying and ontologising entities and relations to describe processes which produce output data given some input data, and the work done by this branch is related to the proposal presented in this paper.

*Ontology of experiments EXPO and LABORS.* The formal definition of experiments for analysis, annotation and sharing of results is a fundamental part of science practice. A generic ontology of experiments EXPO [26] tries to define the principal entities for representation of scientific investigations. The EXPO ontology is of a general value in describing experiments from various areas of research. This was demonstrated with the use of the ontology for the description of high-energy physics and phylogenetics investigations [26].The ontology uses a subset of SUMO[4] suitable for scientific representations as an upper level ontology and a minimized set of relations in order to provide compliance with the existing formalisms. An ontology LABORS is an extension of EXPO for the description of automated investigations (the Robot Scientist Project [5]).

LABORS defines such research units as investigation, study, test, trial, replicate which are required for the description of complex multilayered investigations carried out by a robot. For example an investigation resulted in a fully automatic discovery of new gene functions consists of >10,000 such research units [12]. LABORSs logical definions of the research units properties, hypotheses, results, conclusions and data base of the experimental observations and results are translated into datalog for the reasoning over all data and metadata.

*Ontology of experiment actions - EXACT.* An ontology of experiment actions (EXACT) [24] aims to provide a structured vocabulary of concepts for the description of protocols in biomedical domains. The main contribution of this ontology is the formalizing biological laboratory protocols in order to enable repeatability and reuse of already published experiment protocols. This work is related with the descriptions of data mining scenarios and workflows.

**Describing data mining entities.** Main developments in description of data mining entities in a form of an ontology are in the area of semi automatic data mining workflow construction and description of data mining services and resources on the GRID. Other research includes description of machine learning experiments in context of experiment databases and identification of entities using collection of data mining literature. We will briefly describe all the mentioned approaches.

[4] SUMO: `http://www.ontologyportal.org/`

[5] `http://www.aber.ac.uk/compsci/Research/bio/robotsci/`

*Describing data mining workflows.* In [1] the authors propose a prototype of an Intelligent Discovery Assistant (IDA) which provides users with systematic enumerations of valid data mining processes (sequences of data mining operators) and effective rankings of the processes by different criteria, in order to facilitate the choice of data mining processes to execute to solve a concrete data mining task. This automated system takes the advantage of an explicit ontology of data mining operators (algorithms). The ontology that is designed is a light-weight ontology that contains only a hierarchy of data mining operators divided into three main classes: preprocessing operators, induction algorithms and post processing operators. The leaves of the hierarchy are the actual operators. The ontology does not contain information about the internal structure of the operators and the taxonomy is produced only according to the role that the operator has in the knowledge discovery process.

In [11] the authors build upon the work presented in [1] and propose an intelligent data mining assistant that combines planning and meta-learning for automatic design of data mining workflows. A knowledge driven planner relies on a knowledge discovery ontology [1], to determine the valid set of operators for each step in the workflow. The probabilistic meta-learner is proposed for selecting the most appropriate operators by using relational similarity measures and kernel functions based on past data mining experiments.

The work in [30] also addresses the problem of semiautomatic design of workflows for complex knowledge discovery tasks. The idea is to automatically propose workflows for the given type of inputs and required outputs of the discovery process. This is done by formalizing the notions of a knowledge type and data mining algorithm in the form of an ontology. The planning algorithm accepts task descriptions expressed using the vocabulary of the ontology.

*Describing data mining services and resources.* In [5] the authors introduce an ontology-based framework for automated construction of complex interactive data mining workflows as a means of improving productivity of GRID-enabled data systems. For this purpose they develop a data mining ontology which is based on concepts from industry standards like: predictive model mark-up language (PMML)[6], WEKA [28] and Java data mining API.

In the context of GRID programming in [6] the authors propose a design and implementation of an ontology of data mining. The motivation for building the ontology comes from the context of the author's work in Knowledge GRID [7]. The main goals of the ontology are to allow the semantic search of data mining software and other data mining resources and to assist the user by suggesting the software to use on the basis of the user's requirements and needs. The proposed DAMON (DAta Mining ONtology) ontology is built through a characterization of available data mining software.

In [8] the authors introduce a semantic based, service oriented framework for tools sharing and reuse, in order to give support for the semantic enrichment through semantic annotation of KDD tools and deployment of tools as web

[6] `http://www.dmg.org/`

services. For describing the domain the authors propose an ontology named KD-DONTO which is developed having in mind the central role of a KDD algorithm and their composition similar to work in [1,30].

*Experiment databases.* As data mining and machine learning are experimental sciences, lot of insight of the performance of a particular algorithm is obtained by implementing it and studying how it behaves on different datasets. In [2,3] the authors propose an experimental methodology based on experiment database in order to allow repeatability of experiments and generalizability of experimental results in machine learning. In [27] the authors propose an XML based language for describing classification and regression experiments. In this process the authors identified the main entities for describing a machine learning experiment, which is the first step towards including the experimental entities in an ontology.

*Identification of data mining entities using collections of DM literature.* In [18] the authors survey a large collection of data mining and knowledge discovery literature in order to identify and classify the data mining entities into high-level categories using grounded theory approach and validating the classification using document clustering. As a result of the research study the authors have identified eight main areas of data mining and knowledge discovery: data mining tasks, learning methods and tasks, mining complex data, foundations of data mining, data mining software and systems, high-performance and distributed data mining, data mining applications and data mining process and project.

## 3 OntoDM Design and Description

Our ontology of data mining (OntoDM) aims to provide a structured vocabulary of entities sufficient for the description of data mining scenarios and workflows. OntoDM aims to follow the OBO Foundry principles[7] in ontology engineering that are widely accepted in the biomedical domains. The main OBO Foundry principles state that "the ontology is open and available to be used by all", "is in a common formal language", "includes textual definition of all terms", "uses relations which are unambiguously defined", "is orthogonal to OBO ontologies" and "follows a naming convention" [20]. In this way, OntoDM will be built on a sound theoretical foundation, will be compliant with other (e.g., biological) domains and can be widely re-usable. Our ontology intends to be compatible with other formalisms, to share and reuse already formalized knowledge. OntoDM is available at: `http://kt.ijs.si/panovp/OntoDM/`.

OntoDM is expressed in OWL-DL and is being developed using the Protege ontology editor[8]. It consists of three main components: classes, a hierarchical structure (*is-a* relations) of classes and relations (other than *is-a* relations) between instances. All three major components are described in the following subsections.

[7] OBO Foundry: `http://ontoworld.org/wiki/OBO_foundry`
[8] Protege: `http://protege.stanford.edu`

### 3.1 Identifying Basic Data Mining Entities

OntoDM is based on the proposal of a general framework for data mining by Džeroski [9]. From the framework proposal we identified a set of basic entities of data mining. The basic entities identified are the following (please consult [9] for a detailed description of the entities):

- `dataset`, which consists of `data items`;
- `datatype`, which can be `primitive` (`nominal`, `boolean`, `numeric`), or `structured` (`set`, `sequence`, `tree`, `graph`);
- `data mining task`, which includes `predictive modeling`, `pattern discovery`, `clustering` and `probability distribution estimation`;
- `generalization`, the output of a data mining algorithm, which can be: `predictive model`, `pattern`, `clustering`, `probability distribution`;
- `data mining algorithm`, which solves a data mining task and produces generalizations from a dataset and includes components of algorithms such as: `distance function`, `kernel function`, `refinement operator`;
- `function`, which can be: an `aggregation function`, `prototype function`, `evaluation function`, `cost function` etc;
- `constraint`, which include `evaluation` and `language constraint` (`hard constraint`, `soft constraint`, `optimization constraint`) and
- `data mining scenarios`, related to `queries` and `inductive queries`.

The entities listed above are used to describe different dimensions of data mining. These are all orthogonal dimensions and different combinations among these should be facilitated. Through combination of these basic entities, one should be able to describe most of the diversity present in data mining approaches today.

### 3.2 Upper Level Concepts

In the initial version of the ontology [16] the structure was grounded by the following upper level classes: <*informational entity*>, <*agregate*>, <*procedure*>, <*process*>, <*quality*>, <*representation*> and <*role*>.

In this version of the ontology we mapped the entities more closely to the structure of the OBI ontology. We use BFO upper level classes to represent entities which exist in the real world (i.e., processes, informational entities created in human brain), and in addition we use extensions of EXPO <*abstract entity*> to represent mathematical entities. Recently, due to the limitations of BFO in dealing with information, an Information Artifact Ontology (IAO) has been proposed as a spin-off of the OBI project[9]. Currently IAO is available only in a draft version, but we have included the most stable and relevant classes into OntoDM.

Figure 1 shows the part of the OntoDM class hierarchy. The OntoDM ontology contains 292 classes (including imported upper level classes), and all of the OntoDM classes are extensions of the upper level classes from BFO, OBI, IAO, and EXPO.

[9] IAO:`http://code.google.com/p/information-artifact-ontology/`

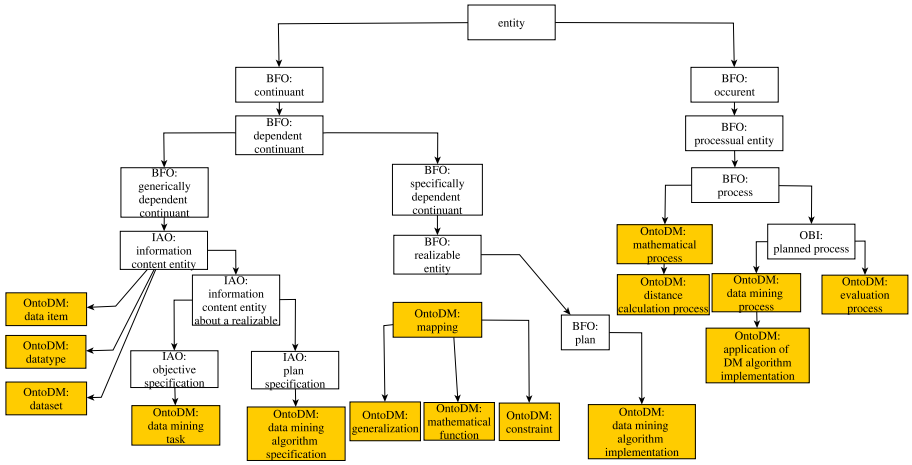


**Fig. 1.** Part of the OntoDM class hirearchy (*is-a* hirearchy): OntoDM classes are extensions of BFO, OBI, IAO and EXPO top level classes

### 3.3 Alignment of OntoDM with OBI

*Information content entity.* The class <*information content entity*> was recently introduced into OBI and denotes all entities that are generically dependent on some artifact and stand in relation of aboutness to some entity. In the domain of data mining we have identified and extended the <*information content entity*> class with the following sub-classes: <*datatype*>, <*data item*> and others. The class <*dataset*> is an information content entity that is an aggregate of data items.

*Realizable entity and information entity about a realizable.* Realizable entities include all entities that can be executed (manifested, actualized, realized) in concrete occurrences (e.g processes). Realizable entities are entities of a type whose instances are typically such that in the course of their existence they contain periods of actualization, when they are manifested through processes in which their bearers participate.

We have identified and extended the class <*realizable entity*> and its subclasses <*plan*>, <*role*>, <*function*> with data mining specific entities. Basic realizable data mining classes are: <*generalization*>, <*data mining algorithm implementation*>, <*constraint*>, <*mathematical function*>, <*query*>, <*data mining scenario*>. Here we just briefly describe <*generalization*> and <*data mining algorithm implementation*>.

The class <*generalization*> represent entities that are products of a data mining process (e.g., the application of a data mining algorithm implementation on a concrete dataset with concrete parameter settings) and includes entities:<*predictive model*>, <*pattern*>, <*clustering*> and <*probability distribution*>. These entities are realized in the <*generalization interpretation process*> where an input to a process is a <*data item*> and the output is a result of

applying of the generalization to the data item (e.g.,the prediction of a predictive models).

The class <*data mining algorithm implementation*> is a subclass of the class <*plan*>. It describes a concrete implementation of a <*data mining algorithm specification*>, subclass of <*plan specification*> and is realized though a data mining process <*application of data mining algorithm*>.

Information entities that concern a realizable entity include: objective specification, plan specification, action specification, etc. A plan specification includes parts such as: objective specification, action specifications and conditional specifications. When concretized, it is executed in a process in which the bearer tries to achieve the objectives, in part by taking the actions specified. An objective specification describes an intended process endpoint.

We have identified and extended the <*information entity about a realizable*> and its subclasses, <*objective specification*> and <*plan specification*>, with data mining specific entities. Basic information entities about a realizable are: <*data mining task*>, subclass of <*objective specification*>, and <*data mining algorithm specification*>, which is a subclass of <*plan specification*>.

*Process.* Process entities represent occurrences that have a specified beginning and end. A planned process is the realization of a plan borne by an agent that initiates this process in order to bring about the objective(s) specified as part of the plan specification. Process entities have as participants continuants and can be also performed by an agent. In the case of data mining, processes have inputs and outputs that can be informational entities and realizable entities. We have identified and extended the <*process*> and <*planned process*> classes with data mining specific classes. Basic data mining process entities described in our ontology include: <*application of a data mining algorithm implementation*>, <*evaluation process*>, <*distance function calculation*> etc.

### 3.4 Ontological Relations

The consistent use of rigorous definitions to characterize formal relations is a major step towards enabling the achievement of interoperability among ontologies in support of automated reasoning across data derived from multiple domains. For, if a fruitful exchange of information to be possible between such ontologies and the data annotated with their terms, each of the system involved must treat the relations in the same way. A relational expression must always stand for one and the same relation, even if it is used in multiple ontologies.

The OntoDM ontology includes and different types of formaly defined ontological relations in order to achieve the desired level of expressiveness. The initial version of the ontology [16] included: fundamental relations (*is-a*, *part-of*), relations from RO [23] *has-participant*, *has-agent*, relations from EXPO/LABORS [26] (*has-representation*), relations from EXACT[24](*has-information*) and relations from OBI (*has-role*, *has-quality*, *has-specified-input*,*has-specified-output*).

The fundamental relations *is-a* and *has-part* are used to express subsumption and part-whole relationships between entities. The relations *has-participant*

and *has-agent* express the relationship between a process and participants in a process, that can be passive or active. Other relations, *has-specified-input* and *has-specified-output*, are specific for relating data mining processes with special types of participants that are inputs and outputs of the data mining process. These two relations have been recently introduced in the OBI ontology.

The relation between an entity and a dependent continuant is expressed via the relation *bearer-of* (defined in the OBI ontology) and this relation is more general and replaces the relations *has-role* and *has-quality* used in the inital version of the ontology.

For expression of informational properties of entities we are using the relation *has-information* and for expression of a representational properties of entities we use the relation *has-representation*, both defined in the EXAT and EXPO/LABORS ontologies.

In this version of the ontology we include relations for expressing relationships between: a process and realizable entity (*realizes*), a planned process and objective specification (*achieves-planned-objective*) and informational entity about a realizable and a realizable entity (*is-concretized-as*). These relations are defined in the OBI ontology.

## 4 The Example Representation of a Data Mining Algorithm

In this section we give an example of the representation of a concrete algorithm using the OntoDM ontology terms (see Figure 2). We describe how to represent the well known C4.5 algorithm [19] for learning decision tree predictive models and its concrete implementation in the WEKA data mining system [28].

When describing a data mining algorithm, one has to have in mind three different aspects. First aspect is the data mining algorithm specification, e.g. *<c45 algorithm specification>*, which is a subclass of the *<information entity>* class about a realizable entity that describes declarative aspects of an algorithm, e.g. has as a part *<predictive modeling>* information about a data mining task in hand. The second aspect is the concrete implementation of an algorithm, e.g. *<wekaJ48 algorithm implementation>*, which is a realizable entity. The third aspect is the process aspect where we describe an application of a concrete data mining algorithm (e.g *<application of wekaJ48>*) on a dataset under concrete algorithm parameter settings. It is necessary to have all three aspects represented separately in the ontology as they have distinctly different nature and this will facilitate different usage of the ontology. The process aspect can be used for constructing data mining workflows and definition of participants of workflows and its parts; the specification aspect can be used to reason about components of data mining algorithms; the implementation aspect can be used for search over implementations of data mining algorithms and to compare various implementations.

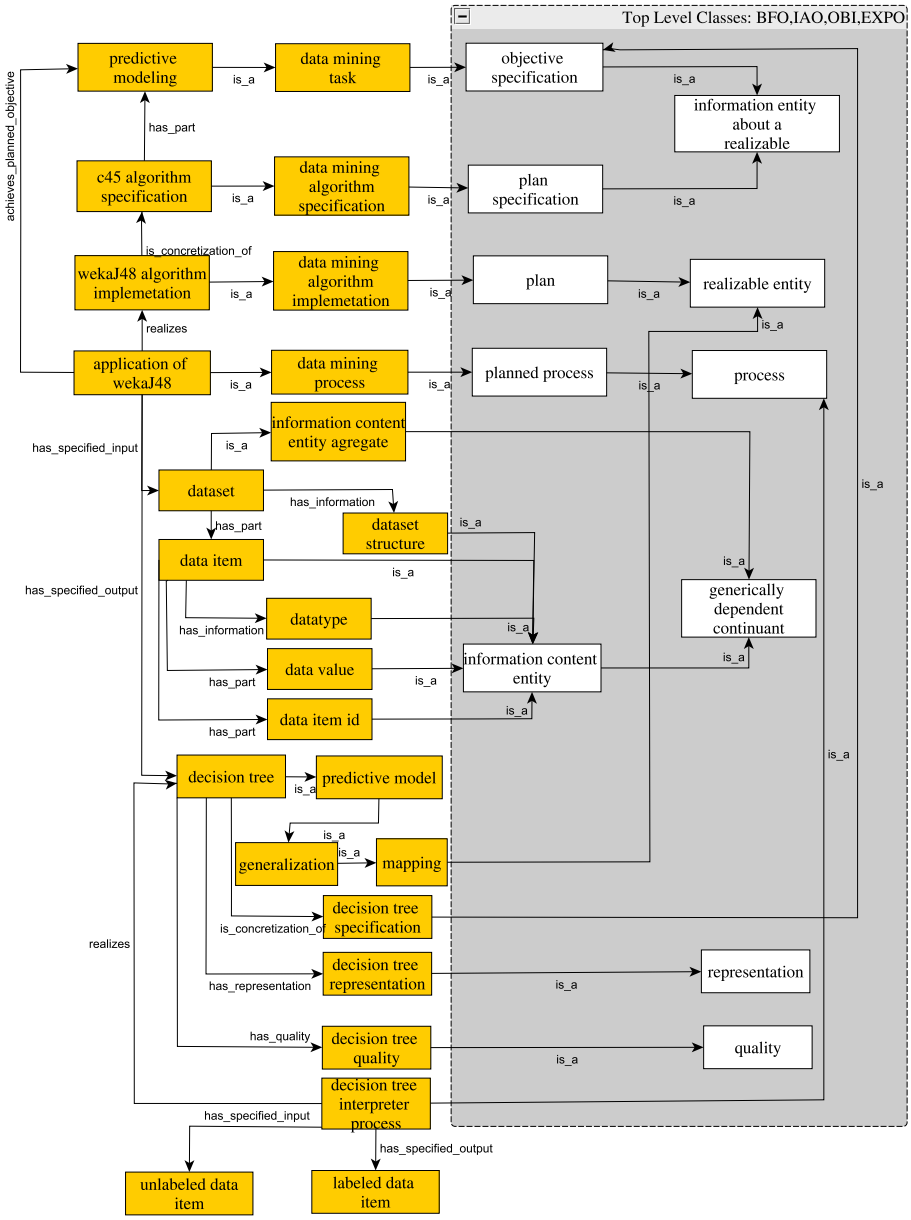


**Fig. 2.** Example - representation of a data mining algorithm (weka.J48)

The relations between the classes representing different aspects of data mining algorithm are as follows:
*<wekaJ48 alg. implementation> is-concretization-of <c45 alg. specification>*
*<wekaJ48 application > realizes <wekaJ48 alg. implementation>*
*<wekaJ48 application > achieves planned objective <predictive modeling>*

Figure 2 presents the process aspect of a data mining algorithm in more detail. Each process has defined input and output entities which are linked to the process via *has-specified-input* and *has-specified-output* relations correspondingly. An input to an application of data mining algorithm is a dataset and parameter values and as output we get a generalization (e.g., *<decision tree>*). A dataset has as parts data items that are characterized with a datatype (e.g., *<tuple of primitives>*). In the case of propositional learning, the datatype of data items is a tuple of primitive data types (nominal values, numeric values, boolean values). A generalization entity has also two aspects. One is connected with looking at it as a data structure and in that case we have a generalization specification (e.g. *<decision tree specification>*) and generalization representation (e.g. *<decision tree representation>*). Another aspect is the functional aspect, when we apply a concrete generalization to a new data item (e.g., prediction using a decision tree). In this case a generalization is realized through a generalization interpreter process (e.g. *<decision tree interpreter process>*) where the input to the process is an unlabeled data item and the output is a labeled data item.

## 5 Complex Data Mining Entities in OntoDM

Our proposal for an ontology of data mining includes descriptions of basic data mining entities. These basic entities are to define more complex entities e.g., entities from the area of inductive databases. The concept of an inductive database [10] employs a database perspective on knowledge discovery, where the knowledge discovery process is composed of query sessions. In this case ordinary queries can be used to access and manipulate the data, while inductive queries (data mining queries) can be used to generate (mine), manipulate and apply generalizations.

Real life applications of data mining typically require interactive sessions and involve formulation of a complex sequence of inter-related inductive queries, which we call a KDD scenario [4]. KDD scenarios can be described at different level of detail and precision and can serve multiple purposes. At the most detailed level of description, KDD scenarios can serve to document the exact sequence of data mining operations undertaken by a human analyst on a specific task. At higher level of abstraction, the scenarios enable the re-use of already performed analyses,e.g., on a new dataset of the same type. The explicit storage and manipulation of scenarios would greatly facilitate the KDD process in whole. Our proposed ontology can be used for formalizing and describing KDD scenarios at various levels of abstraction.

## 6 Conclusion and Further Work

In this paper we present updated and modified version of the OntoDM ontology, which is based on a recent proposal of a general framework for data mining, and includes definitions of basic data mining entities and it also allows for the definition of more complex entities, e.g., constraints in constraint-based data mining, sets of such constraints (inductive queries) and data mining scenarios (sequences of inductive queries).

OntoDM is general-purpose and has been designed with as broad as possible use in mind and can be used to support a number of relevant activities, such as describing data mining services and resources, data mining experiments/investigations, as well as data mining scenarios/workflows.

The ontology OntoDM as presented here is in its early stages of development and hence much work remains to be done. We first need to populate the proposed classes of data mining entities with individuals, identify shortcomings of our ontology in the process and refine the structure of OntoDM as needed in order to describe different aspects of data mining.

Formalizing the knowledge about the domain of data mining and building of a heavy weight ontology of data mining is a time and resource consuming task and should be a community effort. Our goal is to have a mature ontology of data mining that is sufficient and expressive enough to describe the current trends in data mining. This would be also be a helpful step in developing standards for data mining and would lead towards an ontology of data mining investigations.

## References

1. Bernstein, A., Provost, F., Hill, S.: Toward intelligent assistance for a data mining process: An ontology-based approach for cost-sensitive classification. IEEE Trans. on Knowl. and Data Eng. 17(4), 503–518 (2005)
2. Blockeel, H.: Experiment databases: A novel methodology for experimental research. In: Bonchi, F., Boulicaut, J.-F. (eds.) KDID 2005. LNCS, vol. 3933, pp. 72–85. Springer, Heidelberg (2006)
3. Blockeel, H., Vanschoren, J.: Experiment databases: Towards an improved experimental methodology in machine learning. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) PKDD 2007. LNCS (LNAI), vol. 4702, pp. 6–17. Springer, Heidelberg (2007)
4. Boulicaut, J.-F., Klemettinen, M., Mannila, H.: Modeling KDD processes within the inductive database framework. In: Data Warehousing and Knowledge Discovery, pp. 293–302 (1999)
5. Brezany, P., Janciak, I., Tjoa, A.: Ontology-Based Construction of Grid Data Mining Workflows. In: Data Mining with Ontologies: Implementations, Findings and Frameworks. IGI Global (2007)
6. Cannataro, M., Comito, C.: A data mining ontology for grid programming. In: Proceedings of (SemPGrid2003), pp. 113–134 (2003)
7. Cannataro, M., Talia, D.: The knowledge GRID. Commun. ACM 46(1), 89–93 (2003)

8. Diamantini, C., Potena, D.: Semantic annotation and services for KDD tools sharing and reuse. In: ICDMW 2008, Washington, DC, USA, 2008, pp. 761–770. IEEE Computer Society Press, Los Alamitos (2008)
9. Džeroski, S.: Towards a general framework for data mining. In: Džeroski, S., Struyf, J. (eds.) KDID 2006. LNCS, vol. 4747, pp. 259–300. Springer, Heidelberg (2006)
10. Imielinski, T., Mannila, H.: A database perspective on knowledge discovery. Comm. Of The ACM 39, 58–64 (1996)
11. Kalousis, A., Bernstein, A., Hilario, M.: Meta-learning with kernels and similarity functions for planning of data mining workflows. In: Proceedings of the Second PlanLearn Workshop 2008, pp. 23–28 (2008)
12. King, R.D., et al.: The Automation of Science. Science 324(5923), 85–89 (2009)
13. Lister, A., Lord, Ph., Pocock, M., Wipat, A.: Annotation of SMBL models through rule-based semantic integration. In: Proc. of Bio-ontologies SIG/ ISMB 2009 (2009)
14. Malaia, E.: Engineering ontology: domain acquisition methodology and practice. VDM Saarbrucken (2009)
15. Mizoguchi, R.: Tutorial on ontological engineering - part 3: Advanced course of ontological engineering. New Generation Comput 22(2) (2004)
16. Panov, P., Džeroski, S., Soldatova, L.: OntoDM: An ontology of data mining. In: ICDMW 2008, pp. 752–760 (2008)
17. Cimiano, P., Buitelaar, P. (eds.): Ontology learning and population: bridging the gap between text and knowledge. IOS Press, Netherlands (2008)
18. Peng, Y., Kou, G., Shi, Y., Chen, Z.: A descriptive framework for the field of data mining and knowledge discovery. International Journal of Information Technology & Decision Making (IJITDM) 7(04), 639–682 (2008)
19. Quinlan, R.: C4.5: programs for machine learning. Morgan Kaufmann, San Francisco (1993)
20. Schober, D., Kusnierczyk, W., Lewis, S.E., Lomax, J.: Towards naming conventions for use in controlled vocabulary and ontology engineering. In: Proceedings of BioOntologies SIG, ISMB 2007, pp. 29–32 (2007)
21. Smith, B.: Ontology. In: Blackwell Guide to the Philosophy of Computing and Information, pp. 155–166. Oxford Blackwell, Malden (2003)
22. Smith, B., et al.: The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology 25(11), 1251–1255 (2007)
23. Smith, B., et al.: Relations in biomedical ontologies. Genome Biology 6(5) , (2005)
24. Soldatova, L., Aubrey, W., King, R.D., Clare, A.: The exact description of biomedical protocols. Bioinformatics, 24(13) (2008)
25. Soldatova, L., King, R.D.: Are the current ontologies in biology good ontologies? Nature Biotechnology 23(9), 1095–1098
26. Soldatova, L., King, R.D.: An ontology of scientific experiments. Journal of the Royal Society Interface 3(11), 795–803 (2006)
27. Vanschoren, J., Blockeel, H., Pfahringer, B., Holmes, G.: Experiment databases: Creating a new platform for meta-learning research. In: Proceedings of the Second PlanLearn Workshop 2008, pp. 10–15 (2008)
28. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. (June 2005)
29. Yang, Q., Wu, X.: 10 challenging problems in data mining research. International Journal of Information Technology and Decision Making 5(4), 597–604 (2006)
30. Zakova, M., Kremen, P., Zelezny, F., Lavrač, N.: Planning to learn with a knowledge discovery ontology. In: Proceedings of the Second Planning to Learn Workshop, pp. 29–34 (2008)