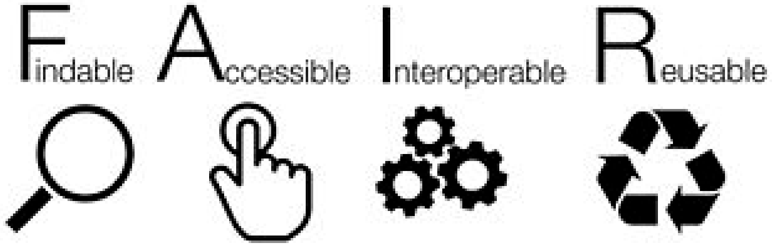
DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

SEMANTIC TECHNOLOGIES FOR OPEN SCIENCE

THE PHILOSOPHY OF OPEN SCIENCE



Scientific knowledge should be shared public good, accessible to all and not restricted by paywalls, proprietary systems, or institutional silos. By making research outputs openly available, reproducible, and reusable, open science aims to accelerate discovery, foster trust, ensure benefits of science reach society as a whole.

Yes, scientific data should be FAIR. But this is an understatement.

Not only data, all research outputs/ artefacts should be FAIR/ openly accessible. This includes publications, code, models, experimental protocols, results of experiments…

We need to represent, annotate and store them, so that they can be found and re-used. All outputs of the scientific process must be formally described and recorded.

This is good for collaboration between scientists, whether they are humans or machines.

AUTOMATION AND FORMALISATION OF SCIENCE

Increasing effort is invested into automating (parts of science)

Robot scientists (aka self-driving labs) are computer/ robotic systems, capable of originating their own experiments, physically executing them, interpreting the results, and then repeating the cycle.

To automate an activity, we need to be able to thoroughly understand it and describe it

To automate science, we need to formalize it

- The goal of science is to increase our knowledge of the natural world through the performance of experiments
- This knowledge should be expressed in formal logical languages
- Formal languages promote semantic clarity, which supports free exchange of scientific knowledge and simplifies scientific reasoning

Ontologies provide controlled vocabularies for describing experiments and their outcomes

SEMANTIC TECHNOLOGIES ARE AN ENABLER

Both for Open Science

And for Automation of Sciences

E.g., controlled vocabularies for describing experiments and their outcomes are essential

- *For documenting and reproducing experiments (by humans)*
- *For automating the design and execution of experiments*

LABORATORY AUTOMATION

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Laboratory Robotics technology is steadily advancing.

Almost everything that a human can do in the lab can now be automated.

Robots can work longer, faster, more accurately than humans.

CLOUD LABORATORY AUTOMATION

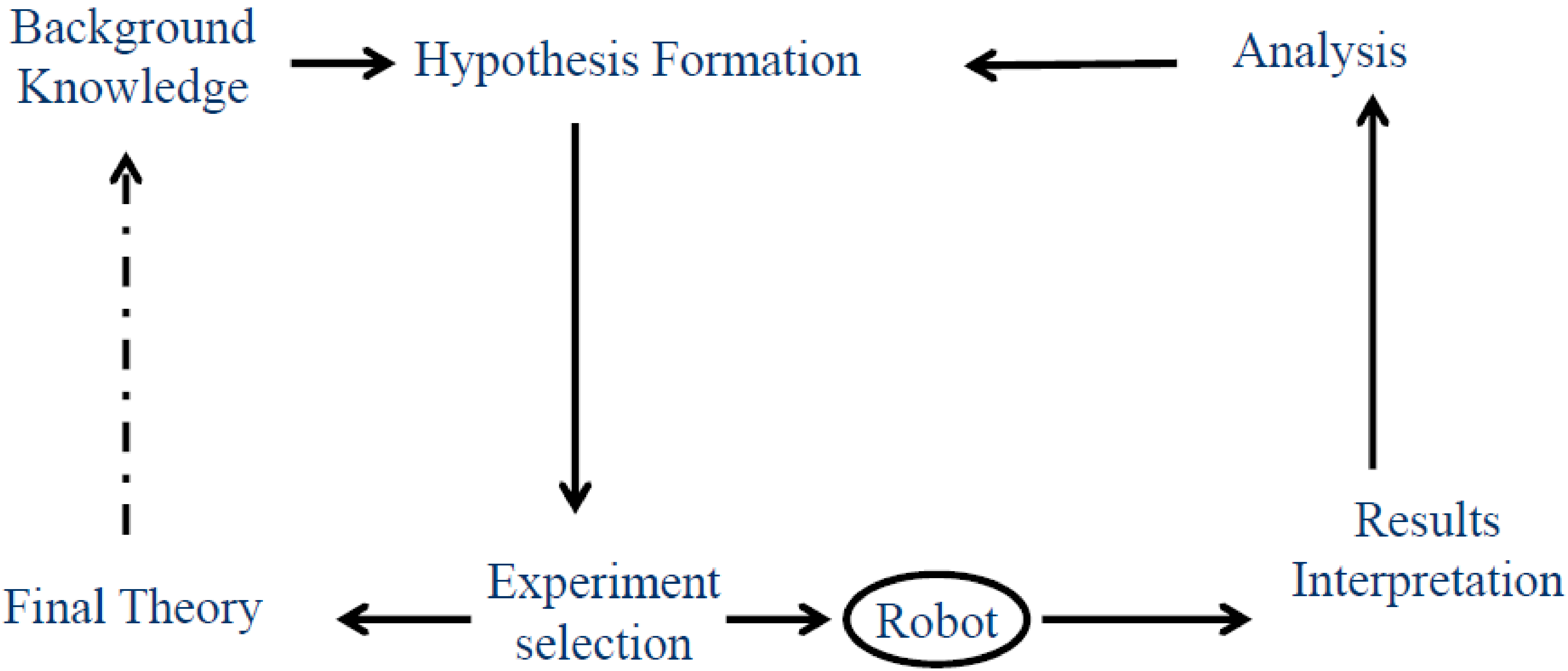
Idea:

- Laboratory automation is expensive and difficult to use.
- Therefore put automation in the cloud.
- Customers use an API to design the experiments.
- Benefits of scale.

Two main companies: Emerald Cloud Services, Transcriptics.

ROBOT SCIENTISTS

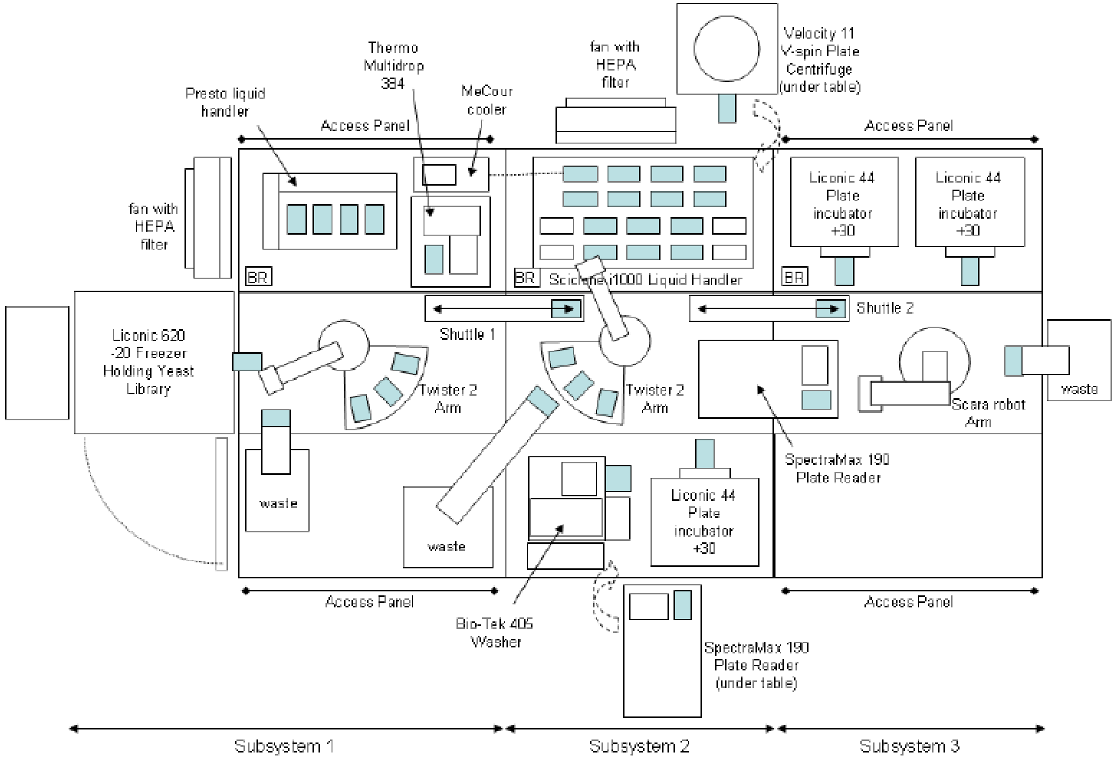
- Computer systems capable of 1) originating their own experiments, 2) physically executing them, 3) interpreting the results, and then repeating the cycle.



TWO ROBOT SCIENTISTS

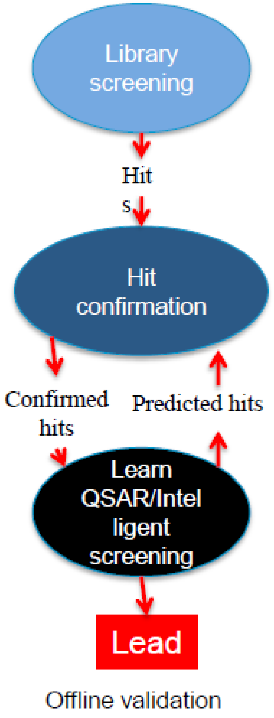
DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

- Adam: Yeast biology



- Eve: Drug design/repurposing

EVE: ROBOT SCIENTIST FOR QSAR

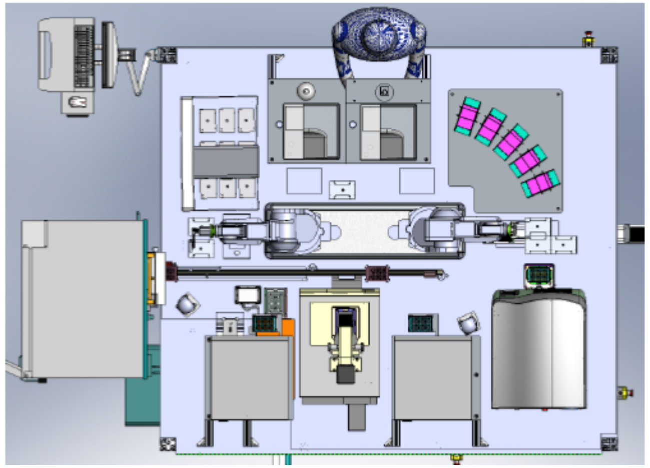


- Standard library screening is brute force:
- Eve uses intelligent screening

- In the standard "pipeline" the 3 processes are not integrated.
- In Eve automated and integrated.

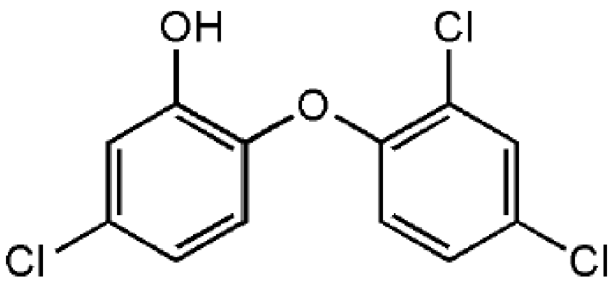
Highlights of Eve's hardware:

- Acoustic liquid handling
- High throughput 384 well plates
- Two industrial robot arms
- Automated 60x microscope
- Liquid handlers, fluorescence readers, barcode scanners, dry store, incubator, tube decapper ...



EVE’S ACHIEVEMENT

Triclosan Repositioned for Malaria



Simple compound

Known to be safe – used in toothpaste.

Targets both DHFR and FAS-II – well established targets.

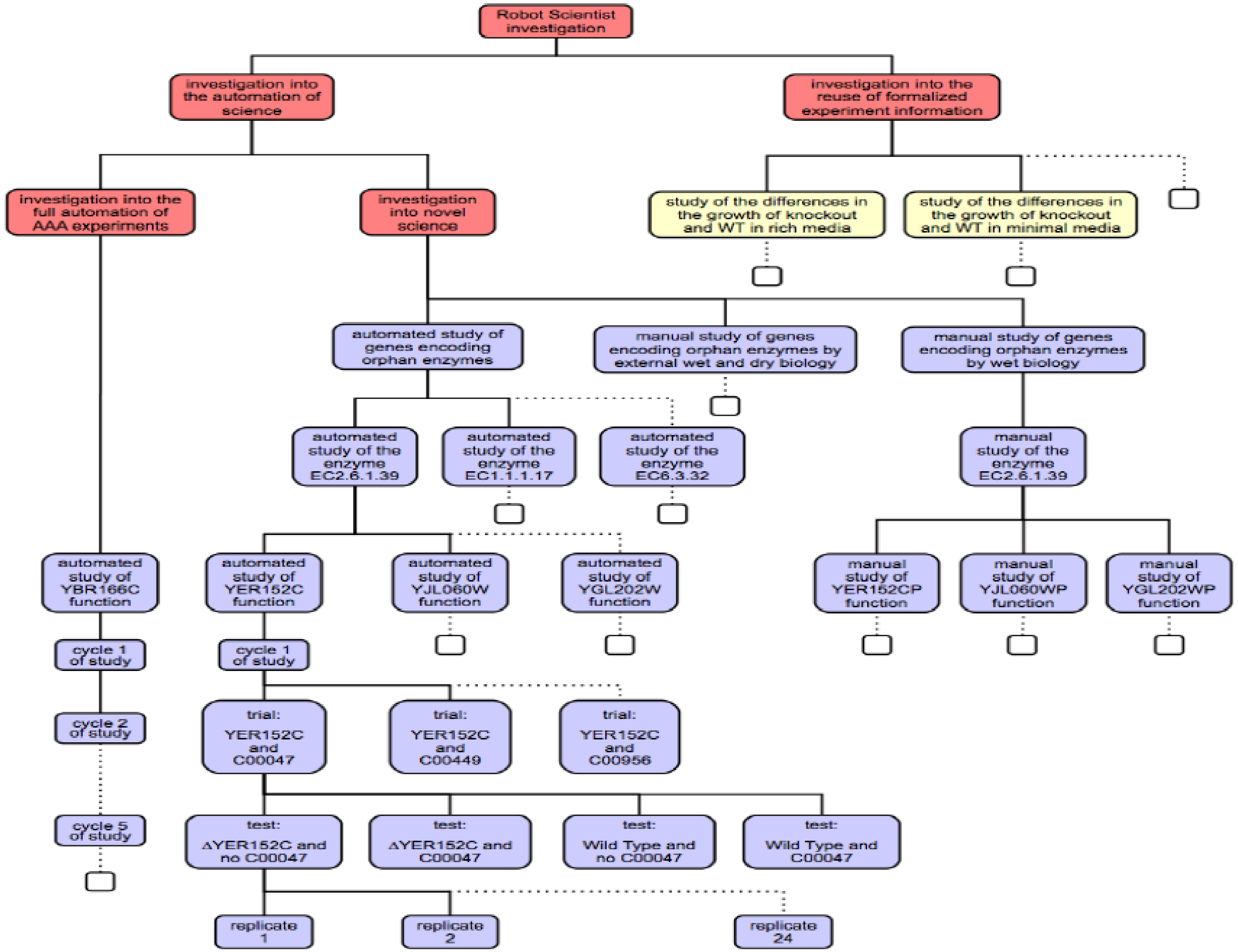
Demonstrated activity using multiple wet experimental techniques.

Works against wild-type and drug-resistant *Plasmodium falciparum*, and *Plasmodium vivax*.

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

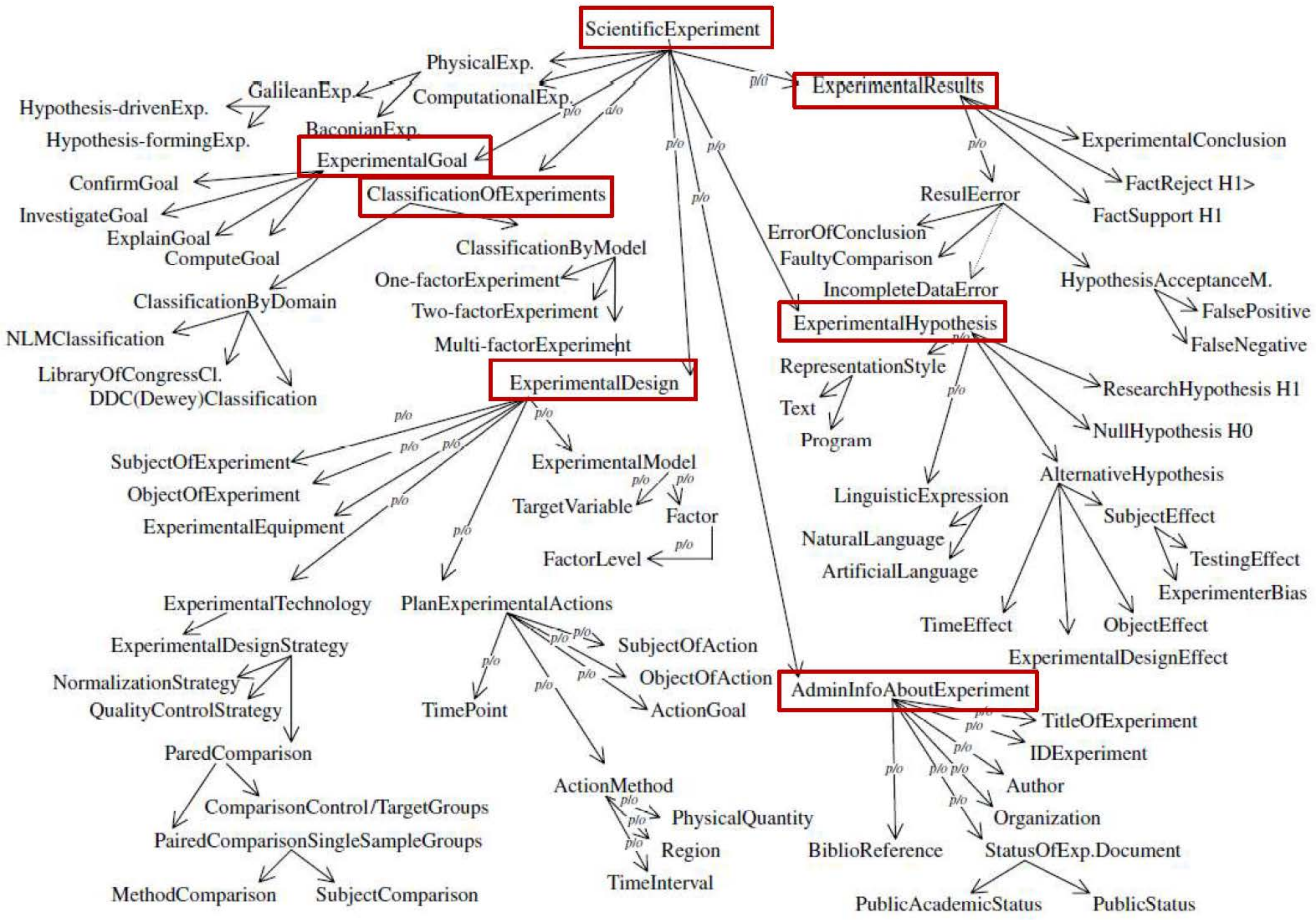
ONTOLOGIES FOR SCIENCE

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute



ONTOLOGIES FOR SCIENCE: ONTOLOGY OF SCIENTIFIC EXPERIMENTS

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute



SCIENTIFIC EXPERIMENT

- Scientific experiment
  - Research method which permits the investigation of cause-effect relations between known and unknown (target) variables of the field of study (domain). An experimental result cannot be known with certainty in advance.
- Types of scientific experiments
  - Physical experiment - scientific experiment which investigates cause-effect relations between known and unknown variables by manipulating the real-world (physical) domain.
  - Computational experiment - scientific experiment which investigates cause-effect relations between known and unknown (target) variables by manipulating the computational (non-physical) domain adequate to the real-world domain.

SCIENTIFIC INVESTIGATION

- Scientific investigation
  - Systematic examination, careful and detailed research of a problem of interest
- Scientific investigation is a complex process that includes several important parts:
  - information gathering
  - problem analysis
  - hypothesis formation
  - experimental design
  - executing a scientific experiment
  - results evaluation
  - results interpretation

FORMALISATION OF SCIENCE

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

- The goal of science is to increase our knowledge of the natural world through the performance of experiments
- This knowledge should be expressed in formal logical languages
- Formal languages promote semantic clarity, which in turn supports the free exchange of scientific knowledge and simplifies scientific reasoning

WHAT ARE ONTOLOGIES

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Ontology is a formal explicit specification of a common conceptualization of a domain of interest.

formality	expressed in a knowledge representation language that provides a formal semantics
explicitness	states knowledge explicitly to make it accessible for machines
being shared	reflects an agreement on a domain conceptualisation among people in a community
conceptuality	specifies knowledge in a conceptual way in terms of symbols that represent concepts and their relations
domain specificity	specifications in an ontology are limited to knowledge about a particular domain

COMPONENTS OF AN ONTOLOGY

- Concepts / classes
  - They represent ontological categories that are relevant to the domain of interest
- Relations
  - They semantically connect concepts and individuals by defining their relationships
- Individuals / instances
  - They represent named and recognizable objects from the domain of interest

TYPES OF ONTOLOGIES

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

- Top-level ontologies
  - Upper ontologies or foundational ontologies
  - They describe very abstract and general concepts
  - They can be used for many domains and applications
  - Well thought out and extensively axiomatized
  - Examples of such ontologies: BFO, SUMO, DOLCE, GFO
- Domain ontologies
  - They cover knowledge of a specific domain
  - They have a very narrow and definite focus
  - They align with top-level ontologies

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

ONTOLOGIES FOR OPEN COMPUTER SCIENCE

OntoDM:
AN ONTOLOGY OF DATA MINING

- Entities in OntoDM = Domain terms:
- First step in ontology development
- OntoDM follows principles from a proposal for a general framework for Data Mining
  - Mining different types of structured data in a uniform fashion
  - Different types of data mining tasks
  - Different types of generalizations (patterns/models)
- Onto DM covers
  - Data
  - Data Mining
  - The process of Knowledge Discovery in Databases (of which DM is a part)

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

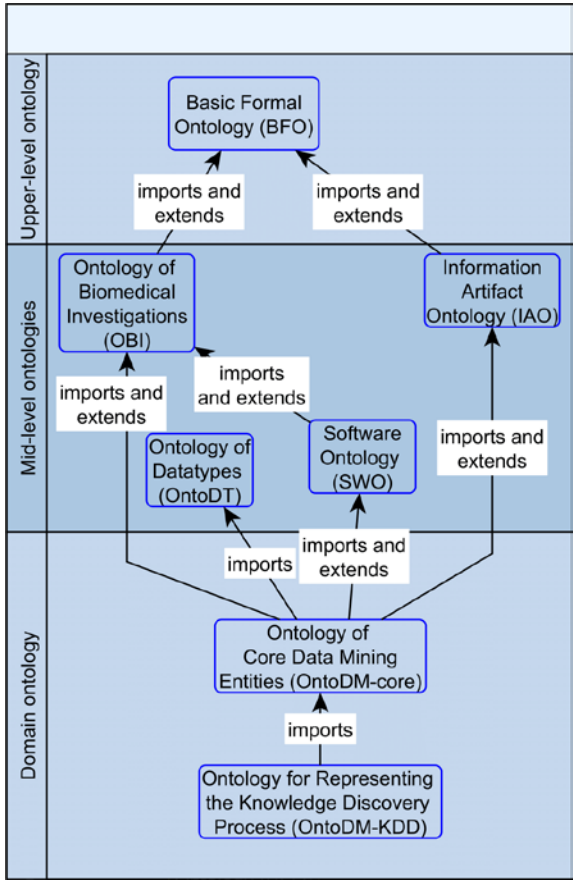
THE BASIC CONCEPTS OF DATA MINING

- **Data:** Structured data types (sets, sequences, trees)
- **Generalizations:** Predictive models, clustering, … (linear equations, decision trees, …)
- **Data mining tasks**: Given data, find generalizations that are valid in the data (satisfy a set of constraints)
- **Constraints** (Language-size of tree, Evaluation-acc.)
- **DM Algorithms**: Solve data mining tasks

ONTO-DM ONTOLOGY MODULES

- OntoDM is a modular ontology
- Modules can be used together or independently depending on the use case

- OntoDT – Ontology of datatypes
- OntoDM-core – Ontology of core data mining entities
- OntoDM-KDD – Ontology of the knowledge discovery process



DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

ONTOLOGY OF DATATYPES – OntoDT

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

- Representation of scientific knowledge about datatypes.
- The notion of a datatype is very important in data mining
  - Characterize the types of data contained in a dataset
  - Applicability of a data mining task on data from a given datatype
  - Applicability of a data mining algorithm on a dataset
- The generic nature of OntoDT enables it to support a wide range of other applications

P. Panov, L. N. Soldatova, S. Džeroski (2015) “Generic ontology of datatypes", Information sciences (in press), doi: 10.1016/j.ins.2015.08.006

ONTOLOGY OF CORE DATA MINING ENTITIES - OntoDM-CORE

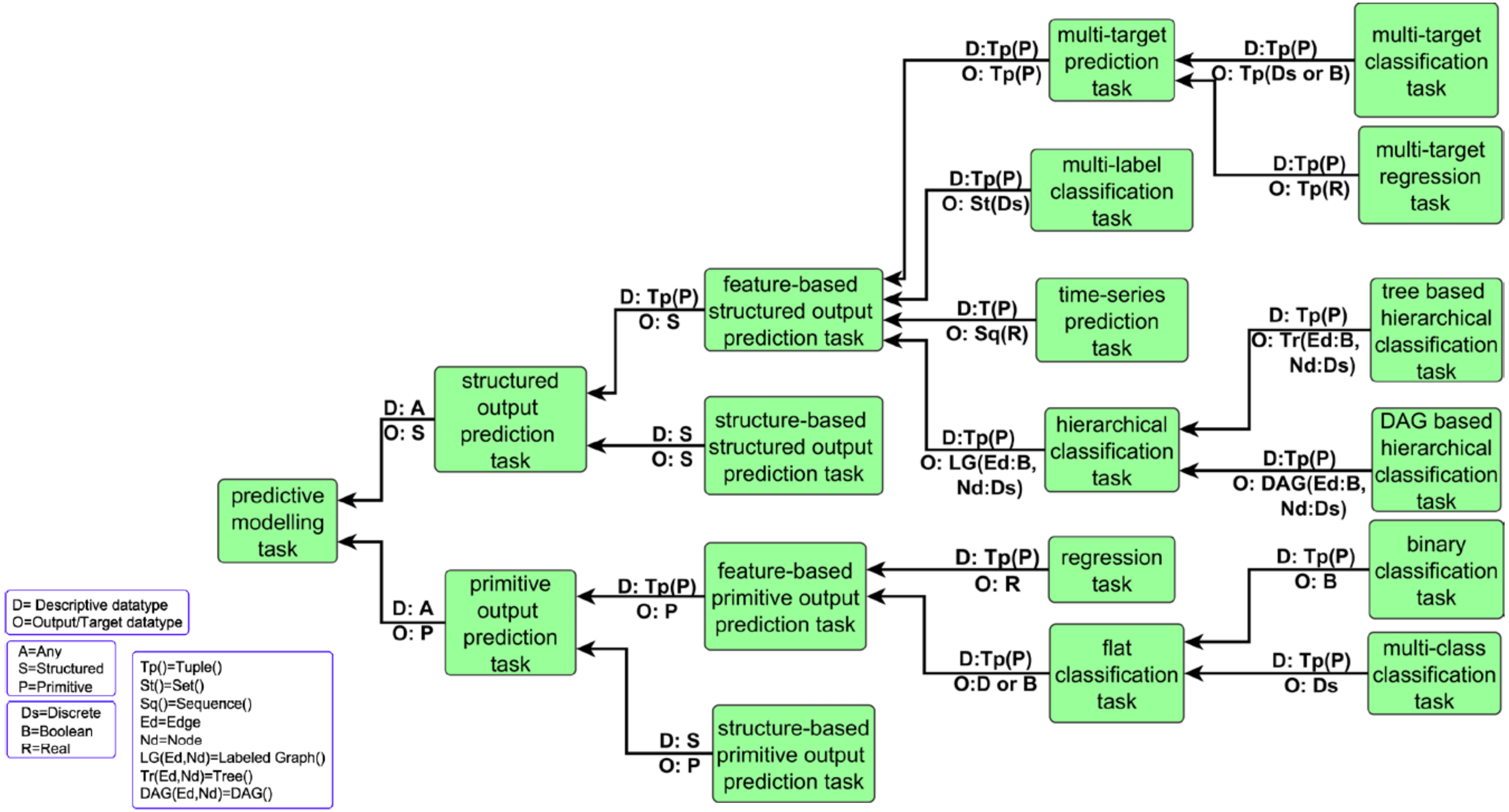
DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

- Based on a proposal for a general framework for data mining OntoDM-core describes the most essential data mining entities
  - Data specification
  - Dataset
  - Data mining task
  - Data mining algorithm
  - Generalizations (patterns, models)
- Taxonomies of datasets, data mining tasks, generalizations, data mining algorithms based on the type of data.
- Representational framework for describing the mining of structured data

P. Panov, L. N. Soldatova, S. Džeroski (2014) "Ontology of core data mining entities", Data mining and knowledge discovery 28(5-6):1222-1265

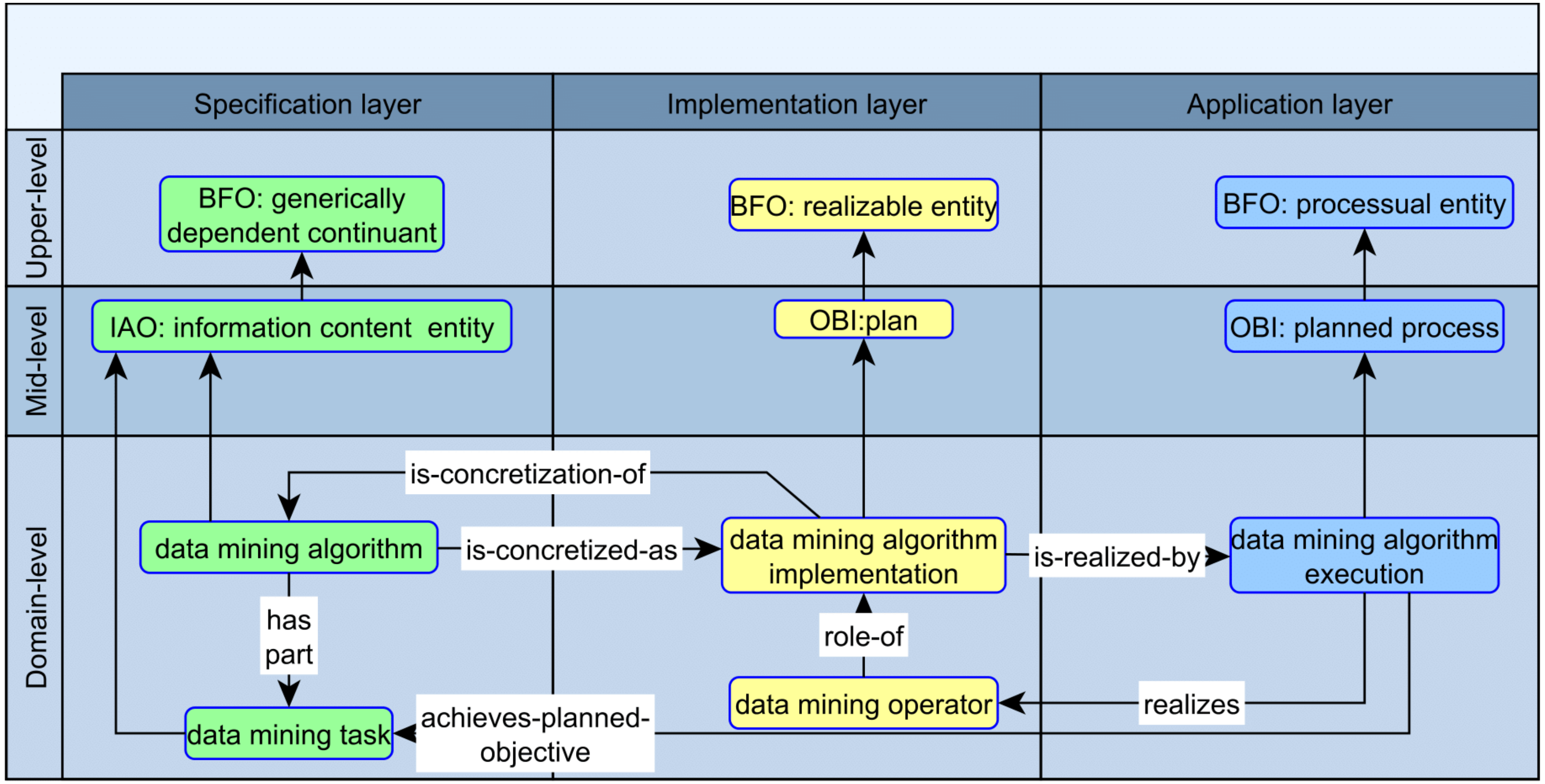
PREDICTIVE MODELING TASKS: TAXONOMY

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

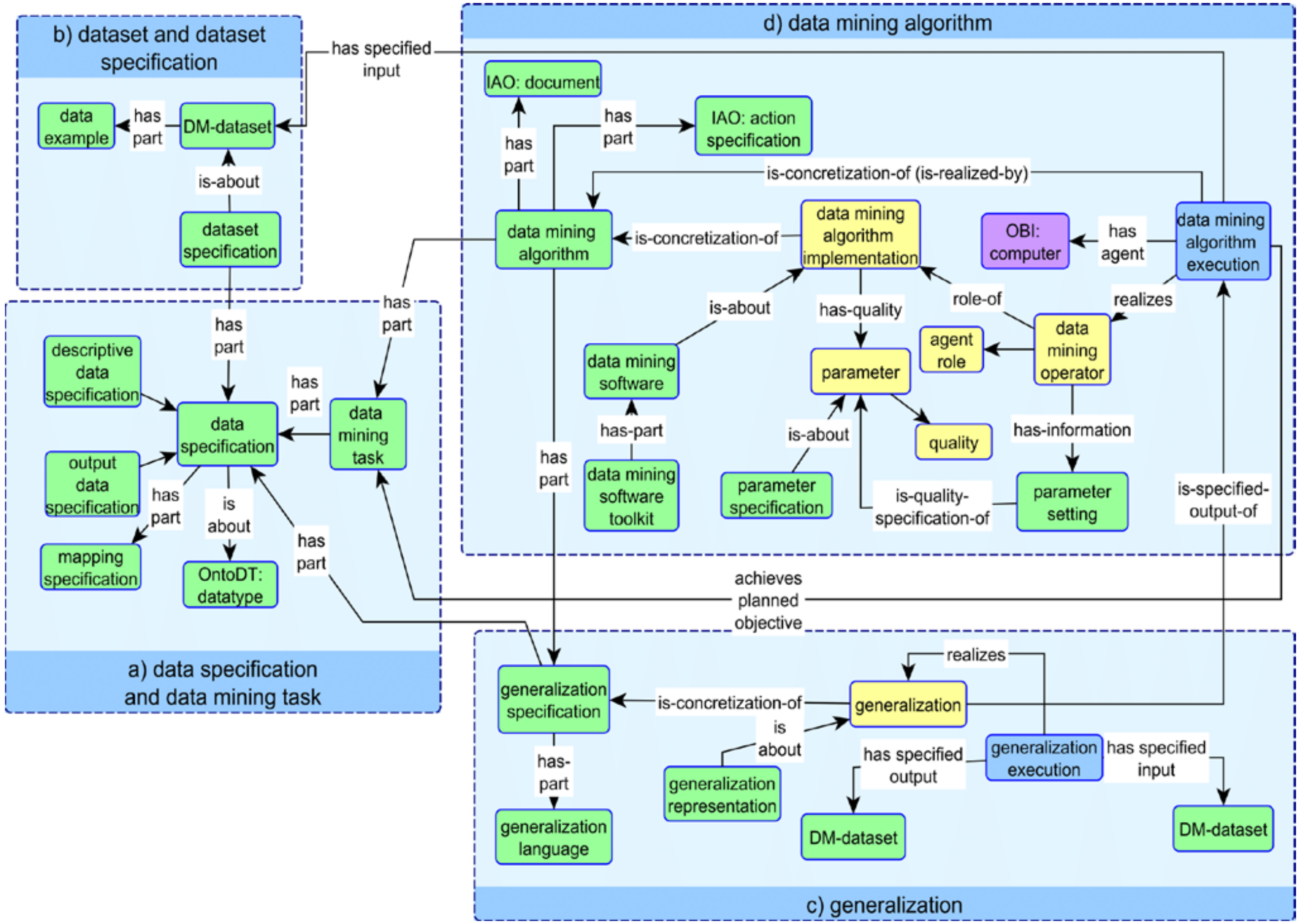


DESIGN STRUCTURE

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute



CORE DATA MINING ENTITIES



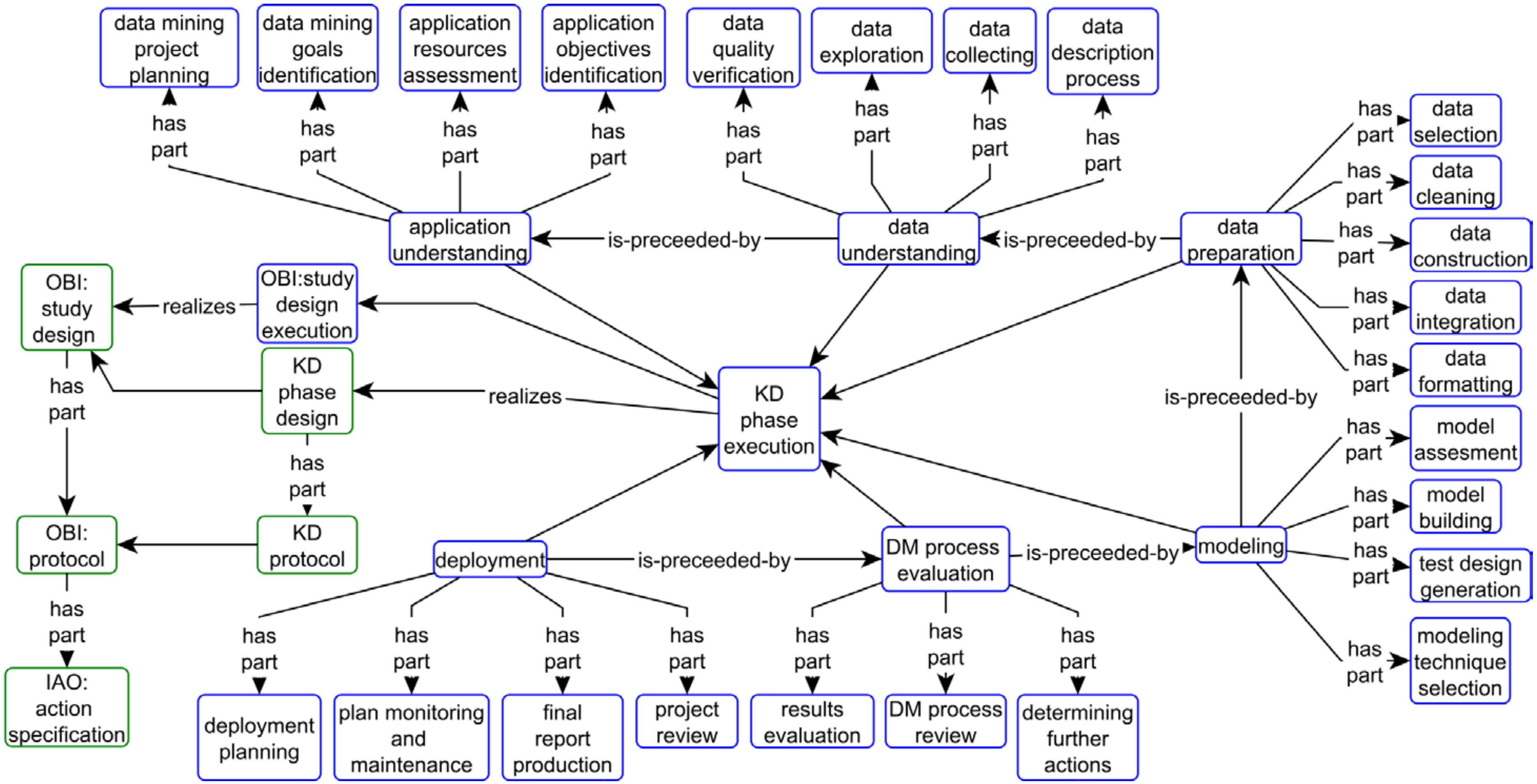
OntoDM-KDD

- Ontology for representing the knowledge discovery (KD) process
- Based on the Cross Industry Standard Process for Data Mining (CRISP-DM)
- Most essential entities for describing data mining investigations
- A taxonomy of KD specific actions and processes
- Specifications of inputs and outputs of processes

P. Panov, L. N. Soldatova, S. Džeroski (2013) "OntoDM-KDD : ontology for representing the knowledge discovery process", 16th International Conference on Discovery science (DS 2013), Lecture notes in computer science vol. 8140, pg 126-140

OntoDM-KDD BASIC STRUCTURE

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute



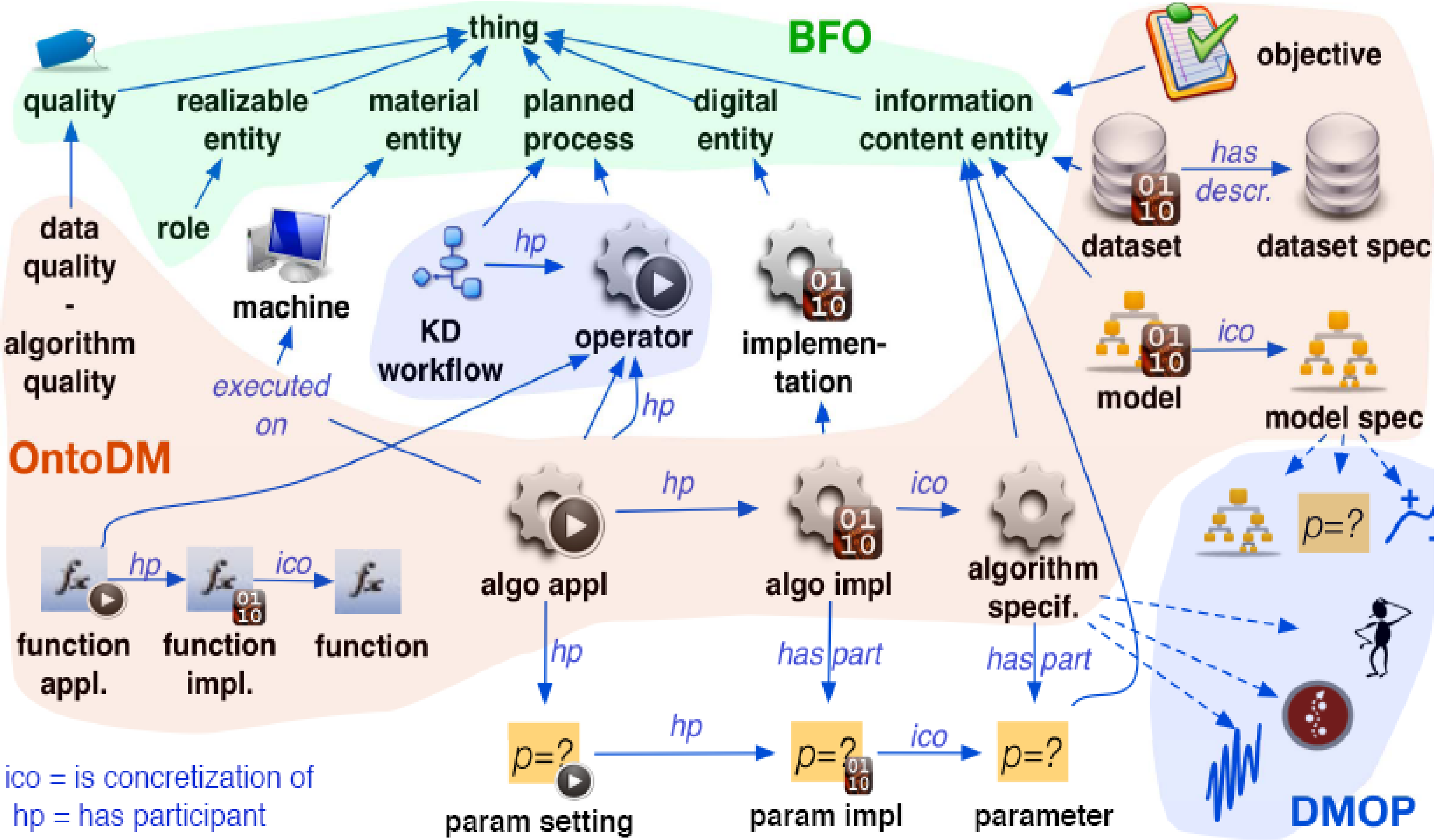
ONTODM USE CASE: EXPERIMENT DATABASES FOR ML

- Store the results of machine learning experiments
  - Many different machine learning algorithms
  - Run on many different datasets
  - With many different parameter settings
  - Described in enough detail to be reproducible

- The machine learning experiments are described
  - In a mark-up language (EXP-ML), which is based on
  - An ontology (EXPOSE), whish in turn is largely based on OntoDM and imports many of its classes

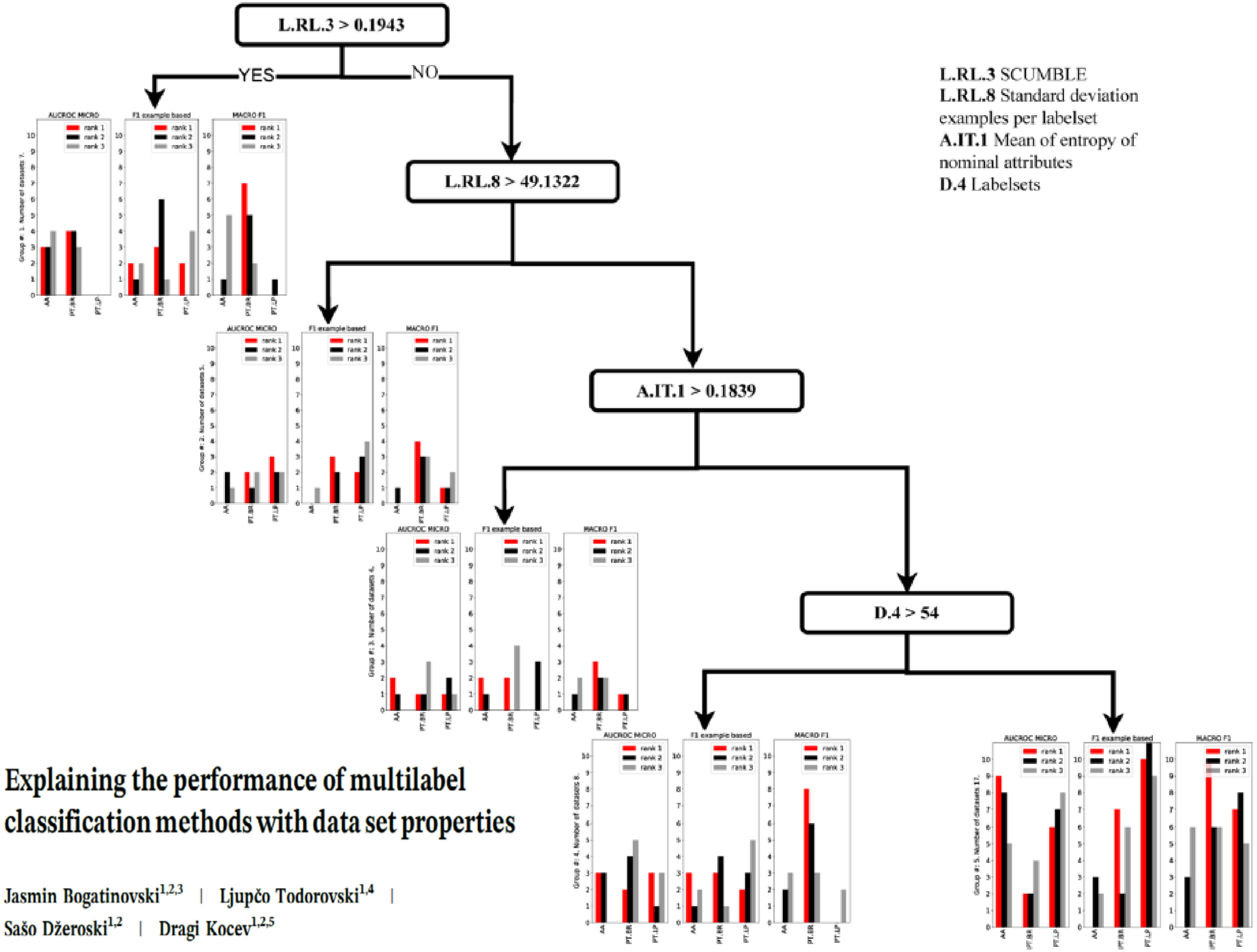
- Initial version: http://expdb.cs.kuleuven.be
- Current version: http://www.openml.org/
- OpenML data used for meta-learning, driving AutoML

EXPOSE / OntoDM



ico = is concretization of
hp = has participant

OntoDM USE: META-LEARNING IN MLC BENCHMARKING



Explaining the performance of multilabel classification methods with data set properties

Jasmin Bogatinovski[1,2,3] | Ljupčo Todorovski[1,4] | Sašo Džeroski[1,2] | Dragi Kocev[1,2,5]

Predict relative performance of classes of MLC methods from dataset properties

- Algorithm adaptation (AA)
- Problem transformation (PT)
  - Binary relevance (BR)
  - Label powerset (LP)

Trying to understand

- Which algorithms
- Work best for which classes of problems
- According to which metrics

EXAMPLE QUERY WITH OntoDM

Find all algorithms that solve the task of multi-target prediction (MTP), produce a generalization expressed in the language of PCTs as output, and are applicable to the EDM dataset

dataMiningAlgorithm
algorithm_s:clus-Bagging-MTC-PCTs
algorithm_s:clus-RandomSubspace-MTC-PCTs
algorithm_s:clus-SubBag-MTC-PCTs
algorithm_s:clus-RandomForest-MTC-PCTs
algorithm_s:clus-MTC-PCTs
algorithm_s:clus-MTR-PCTs
algorithm_s:clus-Bagging-MTR-PCTs
algorithm_s:clus-RandomSubspace-MTR-PCTs
algorithm_s:clus-RandomForest-MTR-PCTs
algorithm_s:clus-SubBag-MTR-PCTs

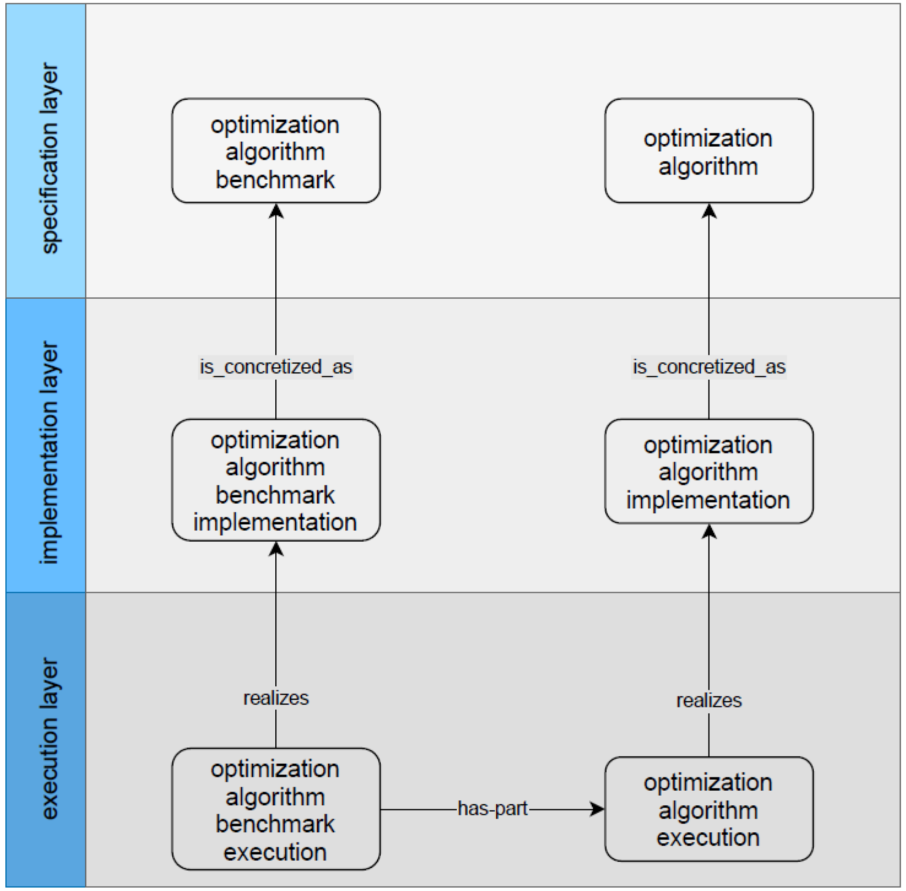
```
SELECT DISTINCT ?dataMiningAlgorithm
WHERE { ?dataMiningAlgorithm  RO:has_part  ?dataMiningTask .
   ?datasetSpecification RO:has_part  ?outputDataSpecification .
   ?generalizationSpecification RO:has_part  OntoDM-clus:OntoDM_
   ?datasetSpecification RO:has_part  ?descriptiveDataSpecification .
   ?dataMiningTask     RO:has_part  ?outputDataSpecification .
   ?generalizationSpecification RO:has_part  ?descriptiveDataSpecification .
   ?dataMiningAlgorithm RO:has_part  ?generalizationSpecification .
   ?generalizationSpecification RO:has_part  ?outputDataSpecification .
   ?dataMiningTask      RO:has_part  ?descriptiveDataSpecification .
   ?datasetSpecification OBO:IAO_0000136  ?datasetInstance .
   ?datasetInstance  rdf:type  OntoDM-clus:OntoDM_clus_00266 .
   ?outputDataSpecification rdf:type  OntoDM-core:OntoDM_000027 .
   ?datasetSpecification rdf:type  OntoDM-core:OntoDM_000031 .
   ?descriptiveDataSpecification rdf:type  OntoDM-core:OntoDM_000247 .
   ?dataMiningTask     rdf:type  OntoDM-core:OntoDM_114697 .
   ?dataMiningAlgorithm rdf:type  OntoDM-core:OntoDM_000038 .
 }
```

OPTION: OPTIMIZATION ALGORITHM BENCHMARKING ONTOLOGY

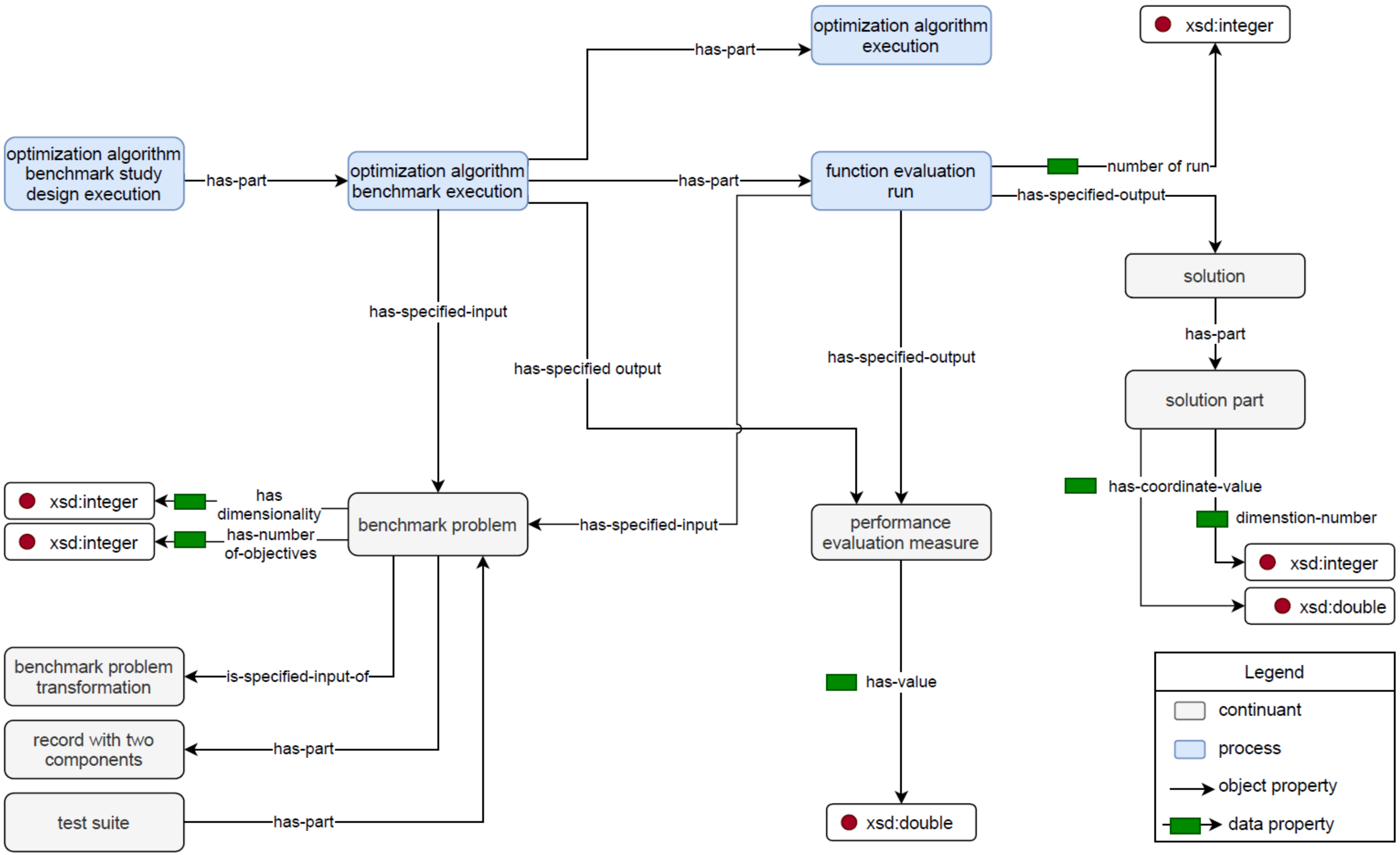
- Follows the same principles as for OntoDM
- Standardizes and formalizes knowledge from the domain of benchmarking optimization algorithms
- Provides the vocabulary needed for semantic annotation of the core entities involved in the benchmarking process
  - Algorithms
  - Problems
  - Evaluation measures

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

OPTION: DESIGN PATTERN



OPTION ONTOLOGY: CORE ENTITIES



QUERIES THAT CAN BE ASKED

- For a given algorithm (e.g., CMA-ES) and fixed budget scenario (e.g., between 1000 - 2000 function evaluations), return the noise-free fitness - Fopt calculated for a set of benchmark problems (e.g., f1 and f7) and a set of problem instances (e.g., the first 5 instances).
- For a given algorithm (e.g., CMA-ES) and fixed target scenario, return the number of function evaluations needed to reach the target.
- For multiple algorithms and fixed budget return the noise-free fitness - Fopt calculated for a set of benchmark problems (e.g., f1-f5).
- For a set of instances from a given benchmark problem, return all noise-free fitness - Fopt values for all (or a set of) algorithms with a fixed budget of 1000 function evaluations.
- For a given study, return the available provenance information.

EXAMPLE SPARQL QUERY

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

For all algorithms included in the study with the DOI number 10.1145/2739482.2768467 and for a fixed budget scenario with 1000-2000 function evaluations, return the *noise-free fitness* - *Fopt* performance evaluation measure calculated on the first 5 instances of the f1 and f7 benchmark problems from the BBOB benchmark suite.

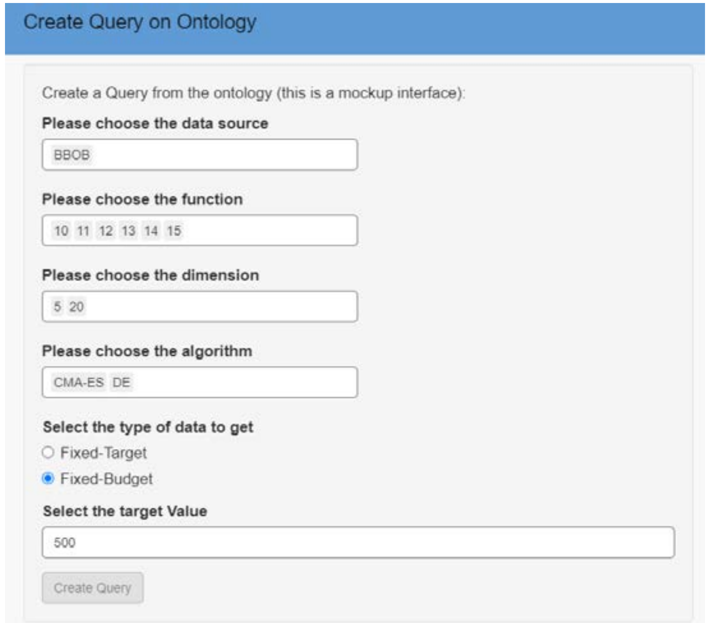
```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-n
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT ?function_evaluation_run_label ?num_function_evaluation_run ?problem ?problem_instance_id  ?dimensionality ?noise_free_fitness
WHERE {
  ?optimization_algorithm_benchmark_execution rdfs:label "10.1145/2739482.2768467_optimization_algorithm_benchmark_execution".
  ?optimization_algorithm_benchmark_execution <http://purl.obolibrary.org/obo/BFO_0000051> ?function_evaluation_run .
  ?function_evaluation_run rdf:type <http://w3id.org/ontoopt/function_evaluation_run> .
  ?function_evaluation_run rdfs:label ?function_evaluation_run_label .
  ?function_evaluation_run <http://purl.obolibrary.org/obo/OBI_0000299> ?performance_meassure.
  ?function_evaluation_run <http://w3id.org/ontoopt/number_of_run> ?num_function_evaluation_run .
  filter (?num_function_evaluation_run > 1000 && ?num_function_evaluation_run < 2000).
  ?function_evaluation_run <http://purl.obolibrary.org/obo/OBI_0000293> ?problem.
  ?problem rdf:type ?problem_class .
  filter (?problem_class in (<http://w3id.org/ontoopt/COCO_benchmark_problem_f1>, <http://w3id.org/ontoopt/COCO_benchmark_problem_f7>)).
  ?problem <http://w3id.org/ontoopt/has_dimensionality> ?dimensionality .
  filter (xsd:integer(?dimensionality) in ("5"^^xsd:integer, "10"^^xsd:integer)) .
  ?problem <http://purl.org/dc/elements/1.1/identifier> ?problem_instance_id .
  filter (?problem_instance_id < 6).
  ?performance_meassure rdf:type <http://w3id.org/ontoopt/noise-free_fitness_-_Fopt> .
  ?performance_meassure <http://w3id.org/ontoopt/has_value> ?noise_free_fitness .
}
```

ANSWER TO THE SPARQL QUERY

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

	function_evaluation_run_label	num_function_evaluation	problem	problem_instance_	dimensionality	noise_free_fitness
1	"10.1145/2739482.2768467_CMA-MSR_f1_1_dim10_run1122_fer"	"1122"^^xsd:long	<http://w3id.org/ontoopt/COCO_benchmark_problem_f1_instance_1_dim_10>	"1"^^xsd:int	"10"^^xsd:int	"6.200845621E-4"^^xsd:double
2	"10.1145/2739482.2768467_CMA-MSR_f1_1_dim10_run1258_fer"	"1258"^^xsd:long	<http://w3id.org/ontoopt/COCO_benchmark_problem_f1_instance_1_dim_10>	"1"^^xsd:int	"10"^^xsd:int	"2.324759161E-5"^^xsd:double
3	"10.1145/2739482.2768467_CMA-MSR_f1_1_dim10_run1412_fer"	"1412"^^xsd:long	<http://w3id.org/ontoopt/COCO_benchmark_problem_f1_instance_1_dim_10>	"1"^^xsd:int	"10"^^xsd:int	"1.723421161E-5"^^xsd:double
4	"10.1145/2739482.2768467_CMA-MSR_f1_1_dim10_run1584_fer"	"1584"^^xsd:long	<http://w3id.org/ontoopt/COCO_benchmark_problem_f1_instance_1_dim_10>	"1"^^xsd:int	"10"^^xsd:int	"1.58360497E-6"^^xsd:double
5	"10.1145/2739482.2768467_CMA-MSR_f1_1_dim10_run1778_fer"	"1778"^^xsd:long	<http://w3id.org/ontoopt/COCO_benchmark_problem_f1_instance_1_dim_10>	"1"^^xsd:int	"10"^^xsd:int	"1.801939646E-7"^^xsd:double

Query composition tools



TOWARDS OPEN COMPUTER SCIENCE

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Of course, data should be FAIR and shared

But it should be used not only to achieve better performance on a given problem

But rather to understand:

- Which algorithms
- Work best for which classes of problems
- According to which metrics