Exposé: An ontology for machine learning experimentation

Joaquin Vanschoren, K.U.Leuven (Belgium), U. Leiden (The Netherlands)
Larisa Soldatova, University of Aberystwyth (UK)

DM Ontology Jamboree 2010
Exposé: An ontology for machine learning experimentation

Joaquin Vanschoren, K.U.Leuven (Belgium), U. Leiden (The Netherlands)
Larisa Soldatova, University of Aberystwyth (UK)

DM Ontology Jamboree 2010
Overview

Ontology lessons
Exposé ontology
Use cases
Ontology lessons
What did we learn from other ontologies
Ontology design

• Start from accepted classes & properties (top-level ontologies, e.g. OBI, RO)

• If possible, reuse prior ontologies to build on their knowledge/consensus

• Use ontology design patterns: reusable patterns for recurrent problems
  - http://ontologydesignpatterns.org

• Check clarity, consistency, extensibility, minimal commitment
Ontology recap:
OntoDM (Panov et al., ’09,’10)

• Aim: unified framework for DM research, builds on BFO
Ontology recap:
OntoDM (Panov et al., ’09,’10)

- Aim: unified framework for DM research, builds on BFO
Ontology recap: DMOP (Hilario et al., ’09)

- Model internal structure of learning algorithms
Ontology recap: DMOP (Hilario et al., ’09)

• Model internal structure of learning algorithms
Ontology recap:
DMWF (Kietz et al., ’09)

- Reason about KD operators: in/outputs, conditions/effects (SWRL rules)
Ontology recap:
DMWF (Kietz et al., ’09)

- Reason about KD operators: in/outputs, conditions/effects (SWRL rules)
Ontology recap:
DMWF (Kietz et al., ’09)

• Reason about KD operators: in/outputs, conditions/effects (SWRL rules)
Ontology recap:
DMWF (Kietz et al., ’09)

- Reason about KD operators: in/outputs, conditions/effects (SWRL rules)
Ontology recap:
DMWF (Kietz et al., ’09)

• Reason about KD operators: in/outputs, conditions/effects (SWRL rules)

"RapidMiner.ID3":

Supercalss:
ClassificationLearning and (uses exactly 1 AttributeValueDataTable) and
(produces exactly 1 Model) and
(simpleParameter1(name="minimal size for split") exactly 1 integer) and
(simpleParameter2(name="minimal leaf size") exactly 1 integer) ...

Condition:
(AttributeValueDataTable and MissingValueFreeData and
(inputAttribute only (hasAttributeType only Categorial)) and
(targetAttribute exactly 1 (hasAttributeType only Categorial)) )(?D),
noOfRecords(?D,?Size), ?P1 is ?Size / 100
→ uses(this,?D), simpleParameter2(this,?P1)

Effect:
uses(this,?D), hasFormat(?D,?F), inputAttribute(?D,?IA),targetAttribute(?D,?TA),
→ new(?M,?D), DecisionTree(?M), produces(this,?M), hasFormat(?M,?F),
inputAttribute(?M,?IA),predictedAttribute(?M,?TA),
Ontology recap: EXPO (Soldatova and King, ’06)

- Make goal and structure of scientific experiments more explicit.
Ontology recap: EXPO (Soldatova and King, ’06)

- Make goal and structure of scientific experiments more explicit

Diagram: Experiment diagram with labels for experiment, hypothesis, result, model, design, variable, and admin info.
Exposé
an ontology for data mining experimentation
Context

- Giant, public database(s) of data mining experiments
- We need:
  - Common language to share experiments (through DM tools)
  - Intuitive ways to store and query experimental results
- We want:
  - Interoperable ontology: OntoDM for top-level, DMOP for detailed properties of learning algorithms
  - Driven by actual experiments submitted to database
    - New algorithms -> ideally, described by author
    - Instances automatically extracted from database
Problem 1: Experiments

What is a machine learning experiment?
What do we need to know about it?
Exposé: Experiments

hp: has participant
hd: has description

Experiment workflow

KD workflow

Workflow:
has inputs, outputs, operators (participants)
Exposé: Experiments

Workflow: has inputs, outputs, operators (participants)
Exposé: Experiments

Workflow: has inputs, outputs, operators (participants)
Problem 2: Algorithms

When talking about an algorithm, what is meant?

General algorithm?
Specific implementation? Which version?
When run, which parameters, components?
Exposé: Algorithms

Specification, implementation, application

hp: has participant
hd: has description
ico: is concretization of

Similar to OntoDM
Same for functions and parameters

ico = is concretization of
hp = has participant
Problem 3: Algorithm composition

plug-in functions, kernels, other algorithms

such components play different roles

-> Agent-role pattern
Exposé: Algorithms

hp: has participant
hd: has description
Exposé: Algorithms

hp: has participant
hd: has description
Exposé: Algorithms

- **hp**: has participant
- **hd**: has description

Diagram:
- Learning algorithm
- Algorithm specif.
- Algorithm impl
- Algorithm appl
- Operator
- Function appl
- Algorithm component role
- Algorithm role
- Algorithm function role
- Baselearner
- Data preprocessor
- Kernel
- Distance functions
- Has part
- Has quality
- Has participant
- Has description
- Parameter impl
- Parameter setting
Problem 4: Workflows

Inputs, outputs, operators
Hierarchical: workflows within workflows
Reuse, parameterize common workflows, e.g. k-fold CV
Exposé: workflows

Similar to RapidMiner?
Problem 5: Reuse

How can we make maximal use of existing ontologies?

OBI: top-level
OntoDM: top-level DM concepts
DMO: operators, learning mechanisms
BFO: accepted top-level classes

- thing
- material entity
- planned process
- digital entity
- information content entity
- data quality
- role
- realizable entity
- machine
- workflow
- operator
- implementation
- dataset
- model
BFO: accepted top-level classes
OntoDM: top-level DM concepts

- Quality
- Realizable
- Market
- Planned
- Digital
- Information

BFO

- Objective
- Dataset
- Dataset spec
- Model
- Model spec

OntoDM

- Data quality
- Algorithm quality
- Role
- Machine
- Operator
- Implementation
- Algorithm

ico = is concretization of
hp = has participant
DMO: operators, learning mechanisms

OntoDM

DMOP
Exposé: top level classes

BFO

- thing
- planned process
- digital entity
- information content entity

Objective

- dataset
- dataset spec

OntoDM

- function appl.
- function impl.
- algo appl.
- algo impl.
- algo specif.

DMOP

- param setting
- param impl
- parameter

ico = is concretization of
hp = has participant

p=?

01 10

01 10

01 10
Other aspects
Datasets

data processing
appl

data processing
workflow

KD
workflow

dataset

has input
has output

hp

Datasets
Datasets

data role
realizes
data repository
part of
data specification

has input
has output
data quality
is concretization of
data property

data item
has quality

data instance

has part
data feature

has quality
data feature

instance property

labeling

unlabeled

labeled

nominal
datatime set

value set

numeric

datatime set

qualitative dataset

property

numeric

datatime set

nominal

datatime set

value set

numeric

dataset property

property

information-theoretic

datatime set

property

landmarker

statistics

data processing

appl

hp

KD workflow
Learner evaluation

evaluation function appl.

evaluation function impl.

Evaluation
Experiment context

- singular experiment
- composite experiment
- hp
Exposé: final notes

• In total 860 classes, 32 properties (from RO + DMOP)

• Individuals: all algorithms, preprocessors, evaluation from WEKA
  • actually stored in experiment database
  • should be programmatically added (and updated)

• Written in OWL-DL, using Protégé 4.0

• Can be browsed at:
  • http://expdb.cs.kuleuven.be/expdb/expose.owl
  • http://www.e-lico.eu/OWLBROWSER2/Manage/
Use Cases
Goal: Collaborative experimentation
Now: small-scale, not repeatable, not reusable
Goal: Collaborative experimentation
Now: small-scale, not repeatable, not reusable

new algorithm
Goal: Collaborative experimentation
Now: small-scale, not repeatable, not reusable

datasets
Goal: Collaborative experimentation
Now: small-scale, not repeatable, not reusable
Goal: Collaborative experimentation
Now: small-scale, not repeatable, not reusable
Goal: Collaborative experimentation
Now: small-scale, not repeatable, not reusable
Goal: Collaborative experimentation
Now: small-scale, not repeatable, not reusable
Goal: Collaborative experimentation
Now: small-scale, not repeatable, not reusable
Goal: Collaborative experimentation
Now: small-scale, not repeatable, not reusable
Goal: Collaborative experimentation
Now: small-scale, not repeatable, not reusable

- A lot of work, limits depth
- Results cannot be reused by others (have to be repeated)
- Hard to repeat experiments from descriptions in papers!
Goal: Collaborative experimentation
Now: small-scale, not repeatable, not reusable

- A lot of work, limits depth
- Results cannot be reused by others (have to be repeated)
- Hard to repeat experiments from descriptions in papers!

The journal system is perhaps the most open system for the transmission of knowledge that could be built ... with 17th century media.

Nielsen (APS Physics 2008)
Data mining as an e-science
Ontologies: experiments shared, run automatically
Data mining as an e-science
Ontologies: experiments shared, run automatically

- Share experiments
  - Internet = large, collaborative workspace
Data mining as an e-science
Ontologies: experiments shared, run automatically

• Share experiments
  • Internet = large, collaborative workspace

• Store them in **experiment databases**
  • Ensure reproducibility
  • Reuse millions of prior experiments
  • Use all info on algorithms, datasets
  • Results universally accessible and useful
e-Sciences
Astrophysics: Virtual Observatories

Target coordinates (J2000):
RA: 05 34 32.0
Dec: +22 00 51

Red: POSSII/F/DSS2
Green: (average)
Blue: POSSII/J/DSS2

Available images:
2MASS/H (fits)
2MASS/J (fits)
2MASS/K (fits)
IRAS-IRIS/100MU (fits)
IRAS-IRIS/12MU (fits)
IRAS-IRIS/25MU (fits)
IRAS-IRIS/60MU (fits)
POSSII/E/DSS1
POSSII/O/DSS2
# e-Sciences

## Bio-informatics: Micro-array Databases

![Micro-array Databases](image.png)

<table>
<thead>
<tr>
<th>ID</th>
<th>Title</th>
<th>Assays</th>
<th>Species</th>
<th>Date</th>
<th>Processed</th>
<th>Raw</th>
<th>Atlas</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-GEOID-12748</td>
<td>Weighted Gene Coexpression Network Analysis Identifies Biomarker...</td>
<td>7</td>
<td>Mus musculus</td>
<td>2009-09-25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-GEOID-14335</td>
<td>Whole genome profiling of fibroblasts from Diamond-Blackfan Anemia</td>
<td>10</td>
<td>Homo sapiens</td>
<td>2009-09-25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-GEOID-14561</td>
<td>Expression data of murine GPI-deficient bone marrow cells in a...</td>
<td>17</td>
<td>Mus musculus</td>
<td>2009-09-25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-GEOID-17170</td>
<td>A systems genetics approach implicates USF1, FAO53 and other...</td>
<td>70</td>
<td>Homo sapiens</td>
<td>2009-09-25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-GEOID-17300</td>
<td>Overexpression of USF1 in HEK293T cells</td>
<td>6</td>
<td>Homo sapiens</td>
<td>2009-09-25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-GEOID-17994</td>
<td>Expression Profiling of brain samples from wt and SCA3 tg anima...</td>
<td>19</td>
<td>Mus musculus</td>
<td>2009-09-25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-GEOID-17995</td>
<td>Role of ICOS:ICOSL interaction in acute GVHD</td>
<td>7</td>
<td>Mus musculus</td>
<td>2009-06-25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-GEOID-18000</td>
<td>Gene expression comparision of drug-resistant Penc1/mock cells...</td>
<td>2</td>
<td>Homo sapiens</td>
<td>2009-09-25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-GEOID-18026</td>
<td>Analysis of chronic lymphocytic leukemia CLL cells and normal B...</td>
<td>7</td>
<td>Homo sapiens</td>
<td>2009-09-25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-GEOID-18043</td>
<td>Priming integrin alpha5 promotes human mesenchymal stromal c...</td>
<td>12</td>
<td>Homo sapiens</td>
<td>2009-09-25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-GEOID-13524</td>
<td>Transcription profiling of rat nucleus accumbens of alcohol-prefer...</td>
<td>29</td>
<td>Rattus norvegicus</td>
<td>2009-09-24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-GEOID-4773</td>
<td>Transcription profiling of human SK-N-MC cell line model of Park...</td>
<td>21</td>
<td>Homo sapiens</td>
<td>2009-09-22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-GEOID-6285</td>
<td>Transcription profiling of brains of mice fed four different diets...</td>
<td>24</td>
<td>Mus musculus</td>
<td>2009-09-22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-GEOID-10748</td>
<td>Transcription profiling of rat brain treated with D-serine</td>
<td>24</td>
<td>Rattus norvegicus</td>
<td>2009-09-22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-TABN-555</td>
<td>Transcription profiling of rat to investigate technical and biologica...</td>
<td>96</td>
<td>Rattus norvegicus</td>
<td>2009-09-21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-BLGS-79</td>
<td>Transcription profiling of Staphylococcus aureus to subinhibitory...</td>
<td>30</td>
<td>Staphylococcus aureus</td>
<td>2009-09-18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-GEOID-11974</td>
<td>Re-sequencing of rice seed</td>
<td>1</td>
<td>Oryza sativa</td>
<td>2009-09-16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-GEOID-12297</td>
<td>Re-sequencing of human post mortem cerebellum reveals alternative...</td>
<td>20</td>
<td>Homo sapiens</td>
<td>2009-09-16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-GEOID-13750</td>
<td>Re-sequencing of yeast ribosome bound sequences</td>
<td>8</td>
<td>Saccharomyces cerevisiae</td>
<td>2009-09-16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-GEOID-14605</td>
<td>Re-sequencing of mouse single cell - wild-type eses...</td>
<td>6</td>
<td>Mus musculus</td>
<td>2009-09-16</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4310 experiments, 100545 assays. Displaying experiments 1 to 25. Pages: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35.
e-Sciences
Bio-informatics: Micro-array Databases
Collaborative Experimentation
Why?
Collaborative Experimentation

Why?

Reproducibility

Good science
Collaborative Experimentation

Why?

**Reproducibility**
Good science

**Quick, easy analysis**
Querying: Answer questions
Test hypotheses
Collaborative Experimentation

Why?

- **Reproducibility**
  - Good science

- **Quick, easy analysis**
  - Querying: Answer questions
  - Test hypotheses

- **Reuse**
  - Save time & energy
  - (e.g. benchmarking)
Collaborative Experimentation

Why?

- **Reproducibility**
  - Good science

- **Quick, easy analysis**
  - Querying: Answer questions
  - Test hypotheses

- **Reuse**
  - Save time & energy
  - (e.g. benchmarking)

- **Generalizability:**
  - Plug into prior results: larger studies
Collaborative Experimentation

**Why?**

- **Reproducibility**
  - Good science

- **Quick, easy analysis**
  - Querying: Answer questions
  - Test hypotheses

- **Integration**
  - Data mining tools
  - Import/export

- **Reuse**
  - Save time & energy
    - (e.g. benchmarking)

- **Generalizability:**
  - Plug into prior results: larger studies
Collaborative Experimentation

Why?

Reproducibility
Good science

Quick, easy analysis
Querying: Answer questions
Test hypotheses

Integration
Data mining tools
import/export

Reuse
Save time & energy
(e.g. benchmarking)

Reference
‘Map’ of known approaches
Compare to state-of-the-art
Includes negative results

Generalizability:
Plug into prior results: larger studies
Use Case 1

Describe experiments in a common language
-> sharing or running experiments on grid
Use Exposé to define common language: ExpML
Use Exposé to define common language: ExpML
Use Exposé to define common language: ExpML
Use Exposé to define common language: ExpML
Use Exposé to define common language: ExpML
Use Exposé to define common language: ExpML
ExpML: a markup language for DM experiments

- Share DM experiments, XML-based
ExpML: a markup language for DM experiments

- Share DM experiments, XML-based

```xml
<expml>
  <dataset id='d1'>
    <learner_evaluation id='e1' input_data='d1'>
      <learner_impl name=... version=...>
        <parameter_setting name='P' value='100'/>
      </learner_impl>
      <learner_appl role='base-learner'>
        ...
      </learner_appl>
      <performance_estimation_appl>
        ...
      </performance_estimation_appl>
      <model_evaluation_function_appl>
        ...
      </model_evaluation_function_appl>
      <model_evaluation_result output_of='e1'>
        <evaluation name='accuracy' value='0.99'>
          ...
        </evaluation>
      </model_evaluation_result>
    </learner_evaluation>
  </dataset>
</expml>
```
ExpML: a markup language for DM experiments

- Share DM experiments, XML-based

```xml
<expml>
  <dataset id='d1'>
    <learner_evaluation id='e1' input_data='d1'>
      <learner_impl name=... version=...>
        <parameter_setting name='P' value='100'/>
        <learner_appl role='base-learner'>
          ...
        </learner_appl>
      </learner_evaluation>
      <performance_estimation_appl>...
      <model_evaluation_function_appl>...
      <model_evaluation_result output_of='e1'>
        <evaluation name='accuracy' value='0.99'>...
      </model_evaluation_result>
    </learner_evaluation>
  </dataset>
</expml>
```

<table>
<thead>
<tr>
<th>ontology</th>
<th>XML</th>
</tr>
</thead>
<tbody>
<tr>
<td>has-part, has-participant</td>
<td>XML subelement</td>
</tr>
<tr>
<td>has-description</td>
<td>(required) attribute</td>
</tr>
<tr>
<td>has-quality</td>
<td><code>property</code> subelement</td>
</tr>
<tr>
<td>is-concretization-of</td>
<td>implementation_of attribute</td>
</tr>
<tr>
<td>part-of</td>
<td>attributes</td>
</tr>
<tr>
<td>has-specific-input</td>
<td>input_data attribute</td>
</tr>
<tr>
<td>has-specified-output</td>
<td>output_of attribute</td>
</tr>
</tbody>
</table>
Use Case 2

Collect experiments in a database to query all empirical results
ExpDB: a database to share experiments
Experiment Database

>650,000 experiments, 54 algorithms, >87 datasets, 45 evaluation measures, 2 data processors, bias-variance analysis
>650,000 experiments, 54 algorithms, >87 datasets, 45 evaluation measures, 2 data processors, bias-variance analysis

<table>
<thead>
<tr>
<th>Table Name</th>
<th>Columns</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>experiment</td>
<td>eid, laid, did</td>
<td>Details of experiments</td>
</tr>
<tr>
<td>learner_application</td>
<td>laid, lpid, sid</td>
<td>Properties of learners</td>
</tr>
<tr>
<td>learner_component</td>
<td>laid, caid, role</td>
<td>Role and identity of learners</td>
</tr>
<tr>
<td>learner_parameter</td>
<td>laid, lpid, liid</td>
<td>Parameters associated with learners</td>
</tr>
<tr>
<td>learner_property</td>
<td>lpid, description</td>
<td>Properties of learner properties</td>
</tr>
<tr>
<td>kernel_application</td>
<td>caid, kiid, is_default</td>
<td>Application and kernel properties</td>
</tr>
<tr>
<td>kernel_component</td>
<td>kiid, kid, name</td>
<td>Kernel components and properties</td>
</tr>
<tr>
<td>kernel_implementation</td>
<td>fiid, fid, name</td>
<td>Implementation details of kernels</td>
</tr>
<tr>
<td>learner_implementation</td>
<td>liid, lid, name</td>
<td>Implementation details of learners</td>
</tr>
<tr>
<td>learner_property_value</td>
<td>liid, lpid, value</td>
<td>Values of learner properties</td>
</tr>
</tbody>
</table>

Experiment Database

- [Diagram of database schema]
Use Case 3

Intuitive querying
Query Interface (YouTube “experiment database”)
http://expdb.cs.kuleuven.be
The way ahead

- 3rd generation of tools could make data mining into e-science
  - Experiments shared, reused, run worldwide
  - Repeatable, generalizable, reusable
- Cooperation on a standardized ontology for data mining?
- Automatic ontology extraction: DM paper -> ontology extension
- RDF experiment databases?
- Open problems:
  - Queriable models, auto-population (active meta-learning), quality control
http://expdb.cs.kuleuven.be