OpenTox framework for predictive toxicology

Семинар-работна среща по Биоматематика

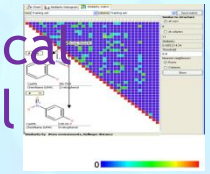
May 26-27, 2010

Др. Нина Желязкова
IdeaConsult

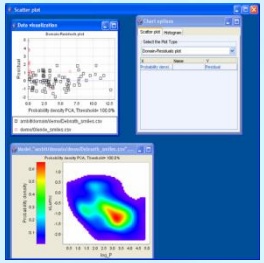
Introduction

- Develops and maintains several well-known open source software products, in particular
  - Toxtree (used in the application of the threshold of toxicological concern)
  - Toxmatch (for encoding and applying chemical similarity indices)
  - Ambit (a QSAR decision support system, including generic database management, structure conversions and searching, as well as applicability domain assessment).
- These tools have been carefully designed and developed in close co-operation both with academia, regulatory bodies and industry.

- Toxtree 1.60 - estimates toxic hazard by applying a decision tree approach
- Toxmatch 1.06 – A chemical similarity evaluation tool



- Ambit Discovery
- Ambit Database Tools

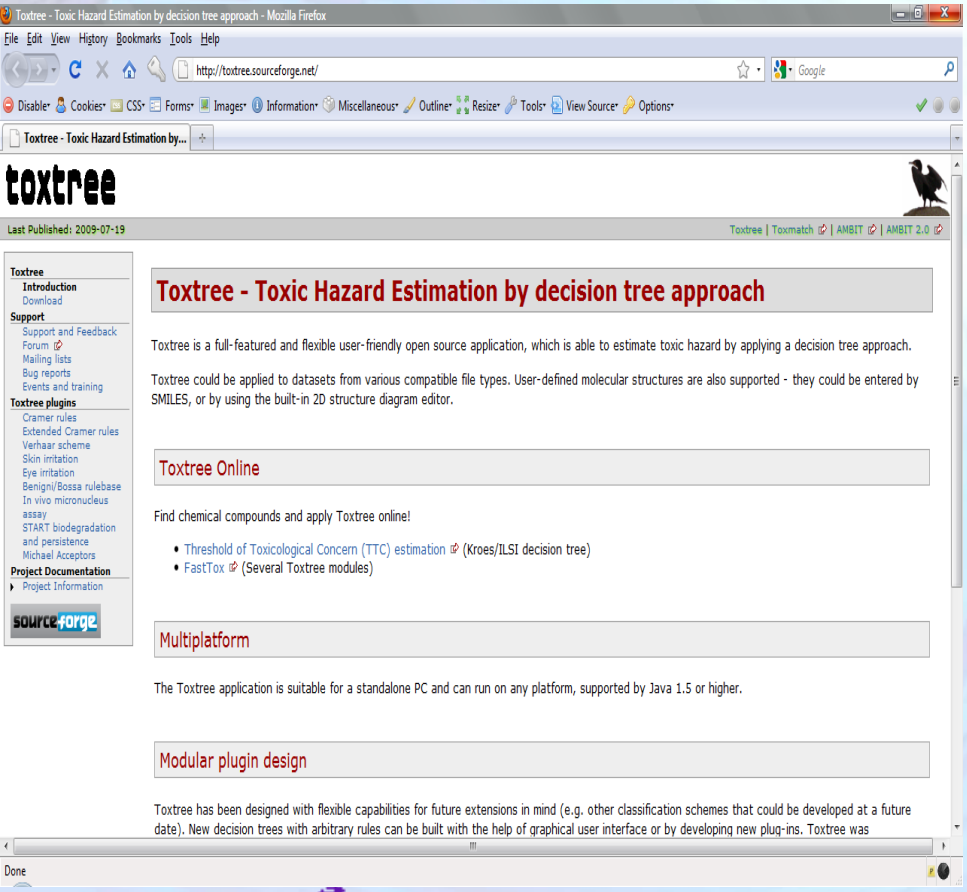


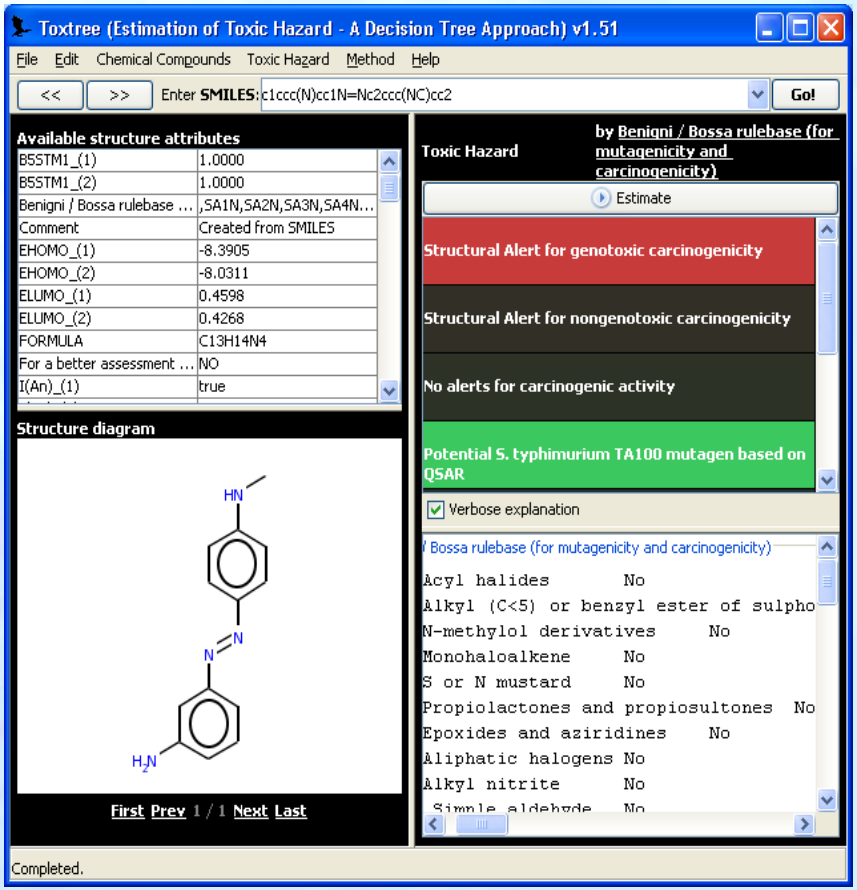
- QMRF repository
- Ambit XT
- Partner in OpenTox FP7 project
- Partner in CADASTER FP7 project

Toxtree 1.60

http://toxtree.sourceforge.net

Applied in industry and academy worldwide



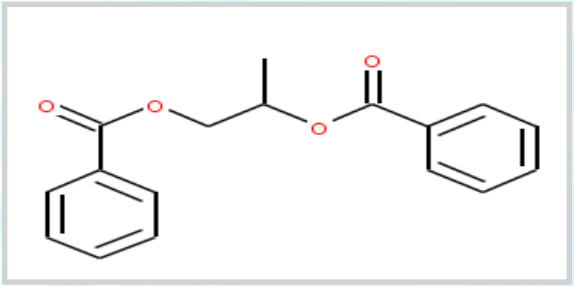


OpenTox

Toxtree internals

Series of questions are applied on query compound and also on reaction products

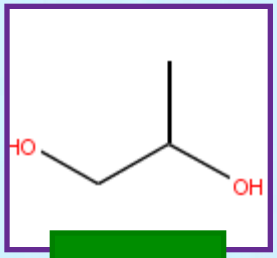
Q1.Normal constituent of the body No
Q2.Contains functional groups associated with enhanced toxicity No
Q3.Contains elements other than C,H,O,N, divalent S No
Q5.Simply branched aliphatic hydrocarbon or a common carbohydrate No
Q6.Benzene derivative with certain substituents No
Q7.Heterocyclic No
Q16.Common terpene No
Q17.Readily hydrolysed to a common terpene No
Q19.Open chain No
Q23.Aromatic Yes
Q27.Rings with substituents Yes
Q28.More than one aromatic ring Yes
Q29.Readily hydrolysed Yes



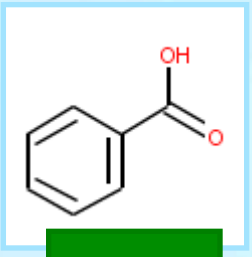
Query compound

Q19.Open chain Yes
Q20.Aliphatic with some functional groups Yes
Q21.3 or more different functional groups No
Q18.One of the list (see explanation) No **Low (Class I)**

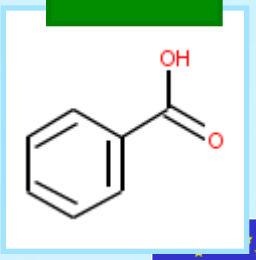
Hydrolysis products



Low (Class I)



Low (Class I)

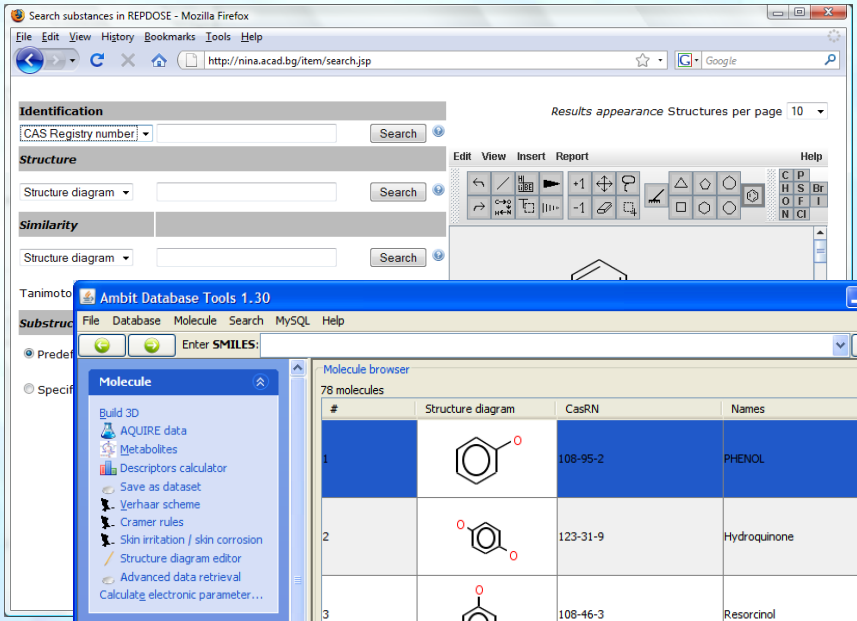


Q30.Aromatic Ring with complex substituents No
Q18.One of the list (see explanation) No **Low (Class I)**

OpenTox

http://ambit.sourceforge.net

AMBIT : *Building blocks for a (Q)SAR decision support system*



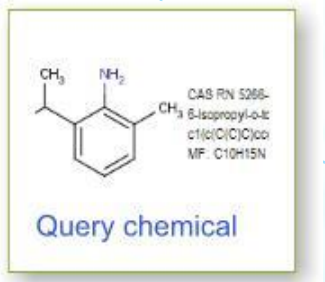
OpenTox

A workflow in AMBIT XT

1. Load chemical
Search by identifier or specify structure

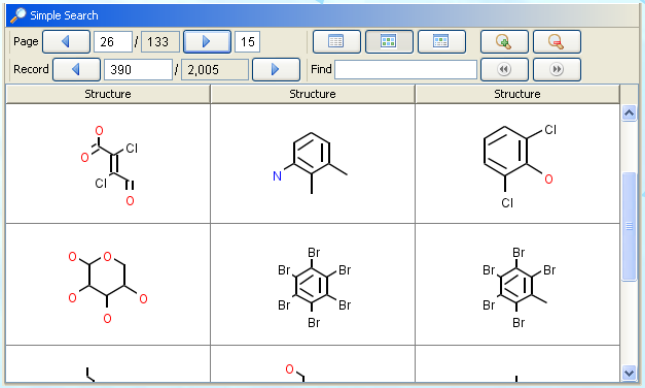
2. Load data
Properties related to the selected endpoint

3. Predict properties
Calculate descriptors and properties not available in the database



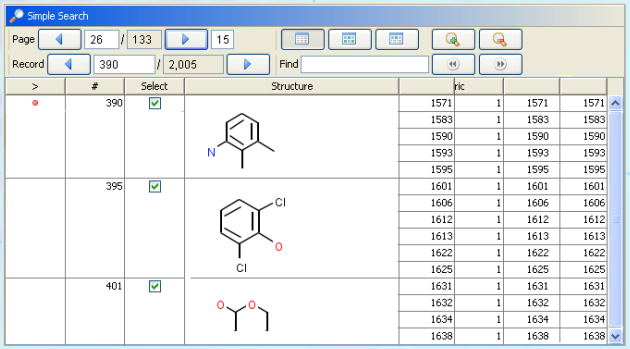
4. Find similar substances
Substances with similar structure and descriptor values

5. Add more substances
Specify similar substances by hand



6. Load data
Retrieve properties related to the selected endpoint

7. Analyse data
Calculate probability that the datapoint is questionable and further testing might be valuable



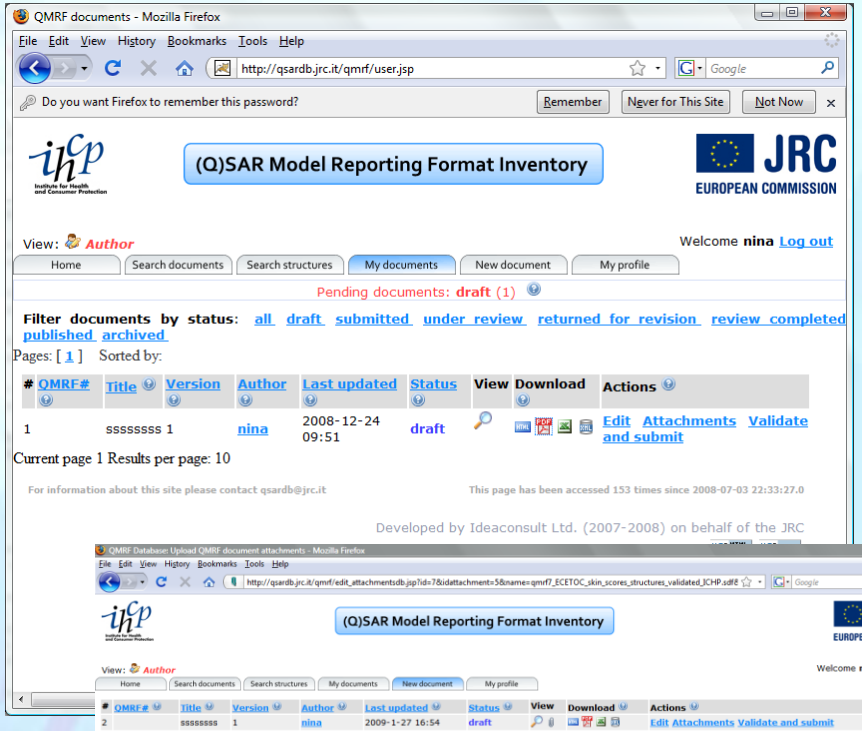
8. Generate report
Data matrix and probabilities

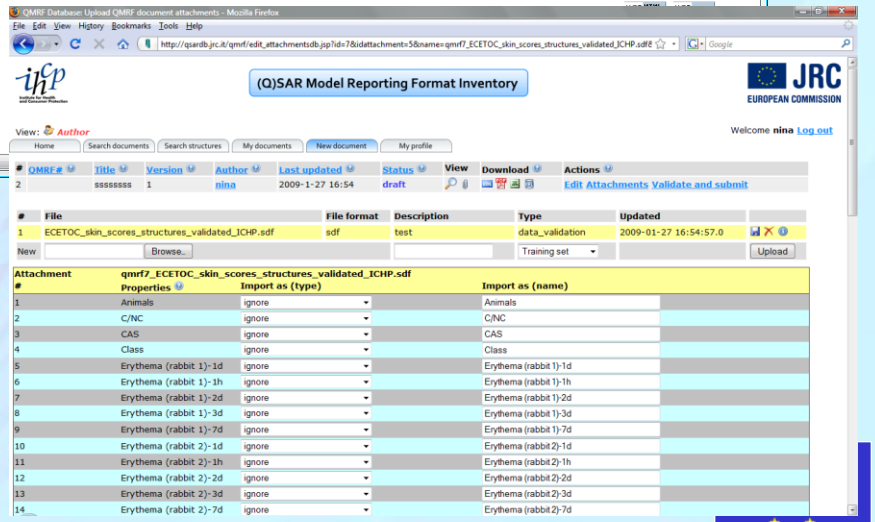
OpenTox

AMBIT Extensions

http://qsardb.jrc.it

- EC JRC Institute for Health and Consumer Protection (former European Chemicals Bureau) - a reference site for retrieving robust summaries of (Q)SAR models in QSAR Model Reporting Format (QMRF)
- AMBIT 2.0
- RESTful Web services
- Custom extensions for third parties



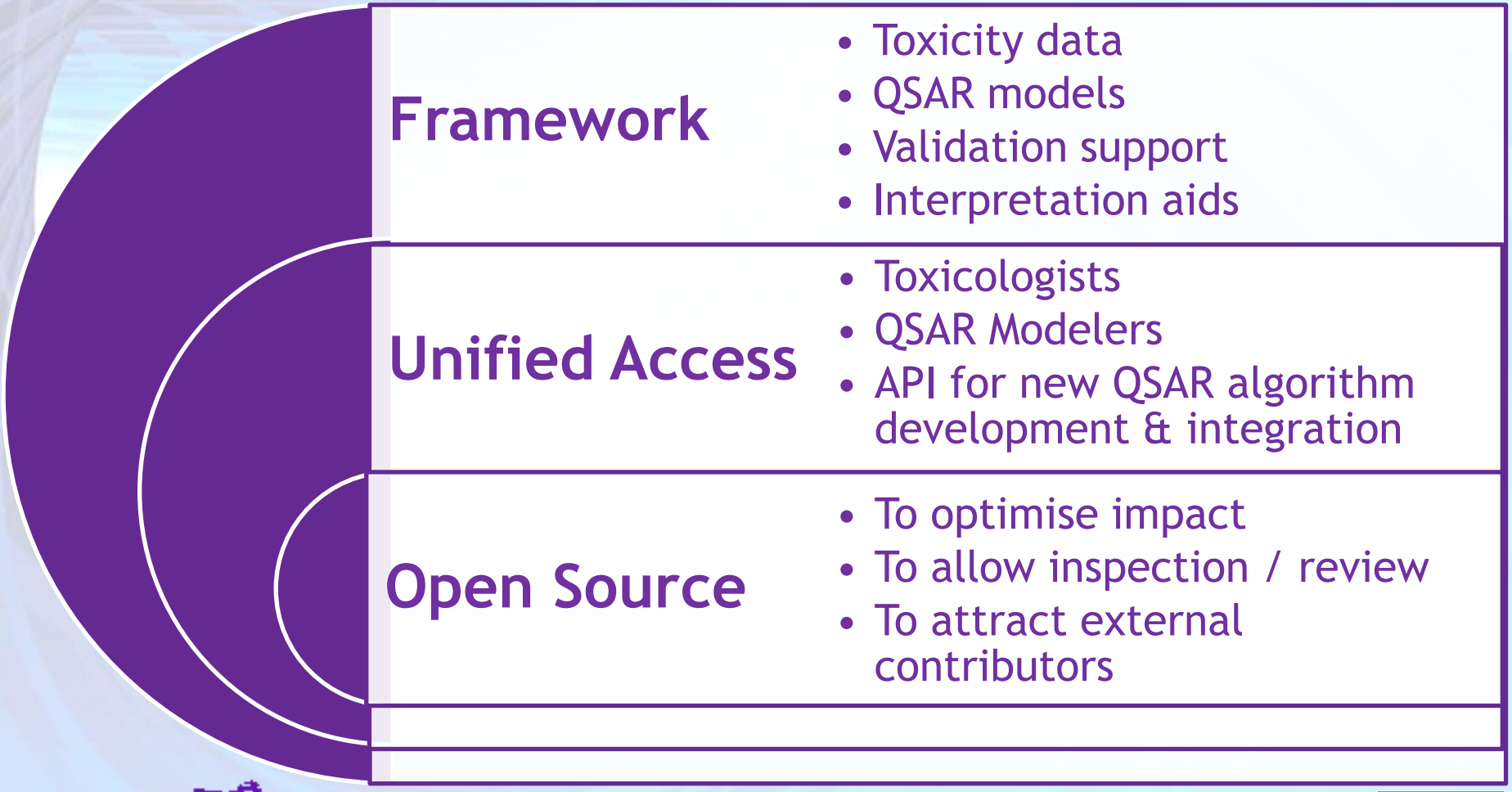


OpenTox summary

- HEALTH-2007-1.3.3 Promotion, development, acceptance and implementation of QSARs (quantitative structure-activity relationship) for toxicology
- 11 partners; http:/opentox.org
- The overall objective of the proposed project is to develop a framework, that provides a unified access to toxicity data, predictive models, procedures supporting validation and additional information that helps with the interpretation of predicted results.
- The OpenTox framework is accessible at three levels:
  - A simple and intuitive interface for toxicological experts, that provides unified access to predictions, toxicological data, models and supporting information
  - An expert interface for the streamlined development and validation of new models
  - An application programming interface (API) for the development, integration and validation of new algorithms

OpenTox

OpenTox Is A Framework



OpenTox

Strategic Context/Goals

- REACH: possible reduction of test animals by using existing experimental data in conjunction with QSAR
- Also practical needs: **reporting** and form filling
- By the OECD principles, a number of requirements to a framework like OpenTox arise

	OECD Principle	OpenTox addresses by...
1	**Defined Endpoint**	well defined data and access to data, unambiguous vocabulary
2	**Unambiguous Algorithm**	providing unified access to documented models and algorithms as well as to the source code of their implementation
3	**Defined Applicability Domain**	integrating tools for the determination of applicability domains and considering these during the validation of (Q)SAR models
4	**Goodness-of-fit, robustness and predictivity**	providing scientifically sound validation routines for the determination of these measures
5	**Mechanistic interpretation (if possible)**	prediction of toxicological mechanisms; **inclusion of biological data; extensibility**

OpenTox

Consequences for Requirements on OpenTox Framework

User Requirements		Software Requirements
Umambiguous data	⇒	*formal way of representing information about **data***
Unambiguous access	⇒	*well-defined interfaces*
Transparency of computational tools	⇒	*formal way of representing information about **methods**, well-defined interfaces*
Variety of user groups	⇒	*simplicity and modularity of design*
Need to integrate various resources (e.g., databases, prediction methods, models, …) to make meaningful predictions	⇒	*distributed architecture, interoperability*
Need to integrate biological information	⇒	*again, modularity of design, extensibility*

OpenTox

Challenges for integration platform in predictive toxicology: The Data

- Flat text files (legacy formats)
- Data type is implicit and part of the data (no metadata)
  - > *<TargetSites_Mouse_Male>*
  - *no positive results*
- Multiline records
  - > *<Endpoint>*
  - *TD50*
  - *Tumor Target Sites*
- Records duplication
  - > <Species>
  - rat
  - 489 occurrences in CPDBAS_v5c_1547_29Apr2008.sdf
- Some records have internal structure (subrecords)
  - > <Species>
  - rat, mouse
- Records order matters

Good
- Simplicity
- Manual inspection
- Sequential processing

Bad
- Data standardization
- Data summarization
- Data duplication
- Metadata

Ugly
- Random access processing
- Huge volumes processing

einecs_structures_V13Apr07.sdf 135MB
- Records modification
- Errors !

OpenTox

Challenges for integration platform in predictive toxicology: Chemical structures and Models

- Journal publications
- Commercial software
- Open source software

- No common means of sharing models

- Recent attempt to formalize description of models (QMRF by JRC)

Good

- Many models available
- >20 000 QSARs
- >5000 descriptors

Bad

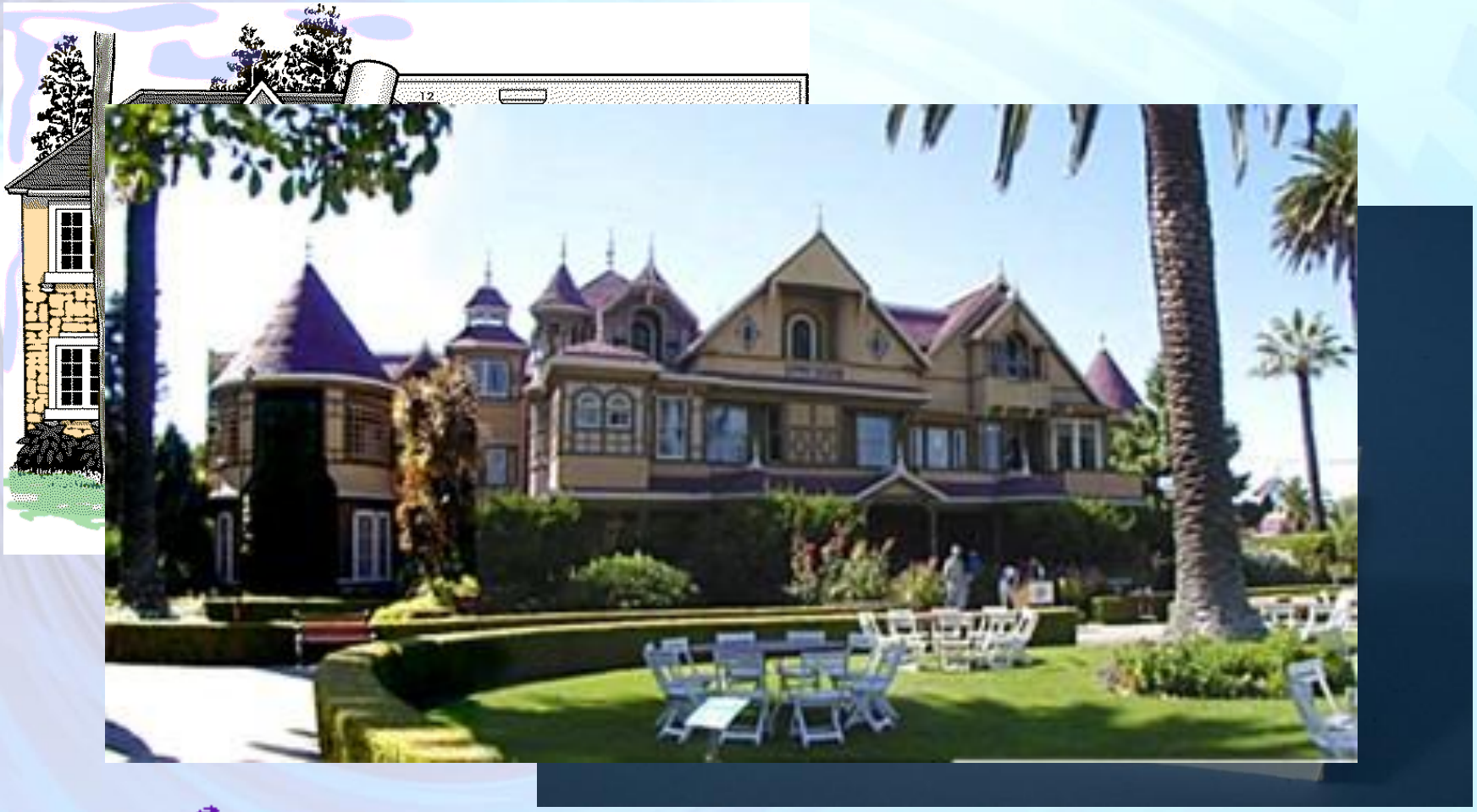
- Hard or impossible to reproduce published models
- Ambiguity in chemical structures
- Errors in chemical structures

Ugly

- Information loss and efforts duplication
- Same chemical structures are submitted to (almost) the same processing

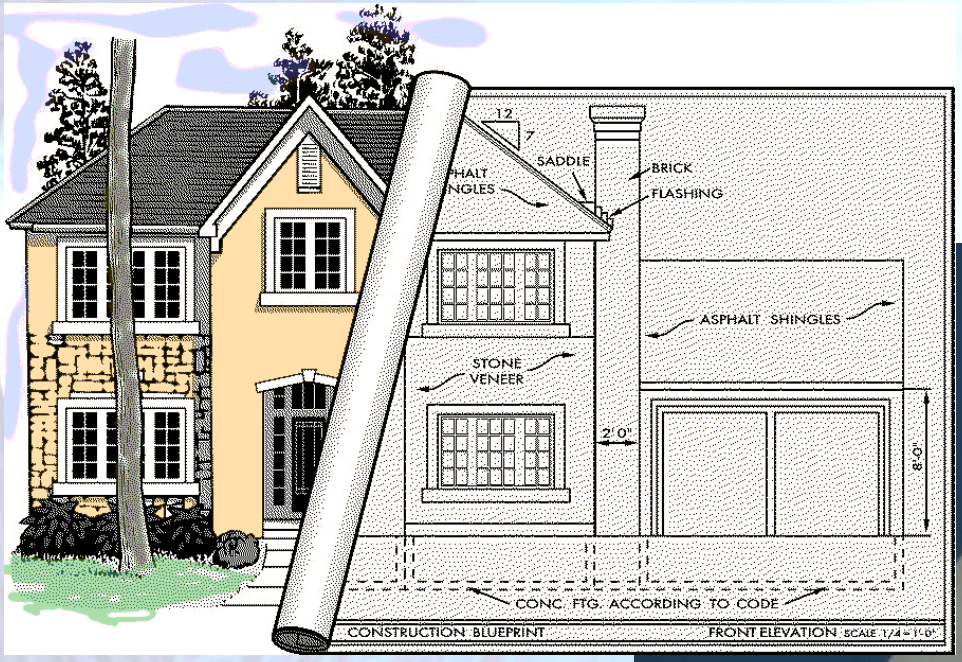
OpenTox

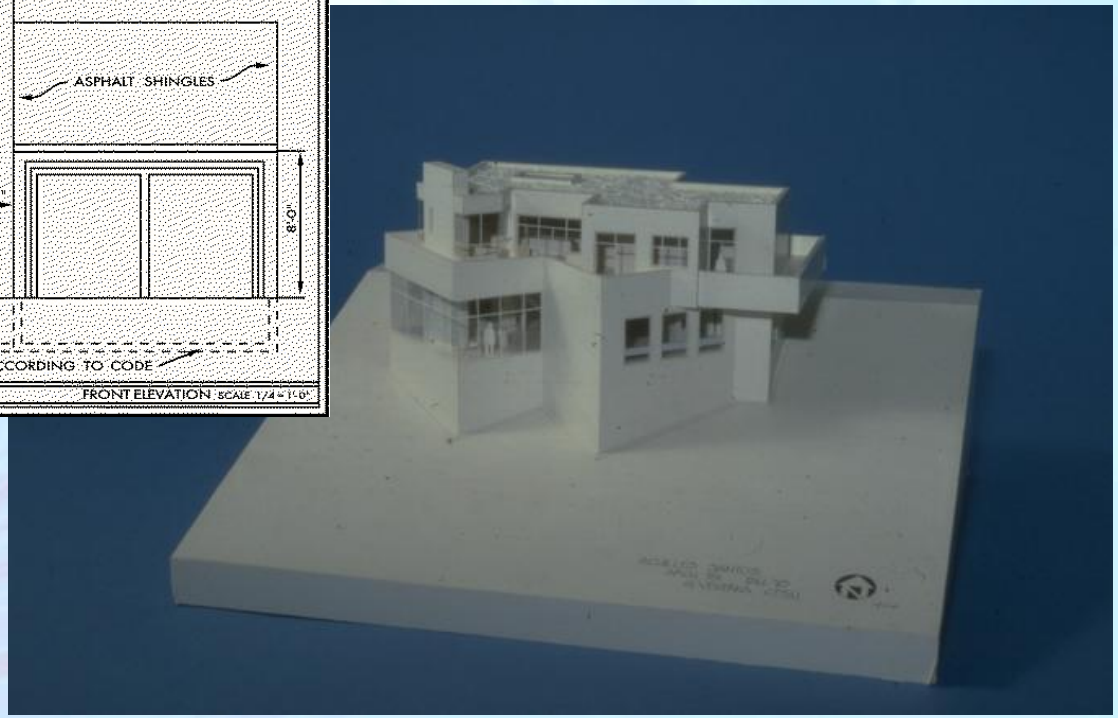
Software Architecture



OpenTox

Software Architecture





To solve a wide range of modeling tasks: needs buildings blocks (modularity) and well-defined ways of combining them

Technological choices…

OpenTox

- Distributed Web services for predictive toxicology
- REST technology
  - Every object has an unique URI
  - URIs are dereferensable
  - Multiple representation of an object is encouraged (e.g. RDF, but also others)
  - Fixed operations – GET, PUT, POST, DELETE
- Every object has RDF representation
  - Compounds
  - Datasets
  - Compound properties
  - Prediction algorithms
  - Models
  - Validation statistics
  - Reports
- Ontologies: Opentox.owl, Blue Obelisk algorithm ontology, OpenTox algotihm types ontology, OpenTox endpoints ontology, based on ECHA endpoints classification; specific endpoints ontologies, developed by ISS & ITEM

Representational State Transfer (REST): What and Why?

What?

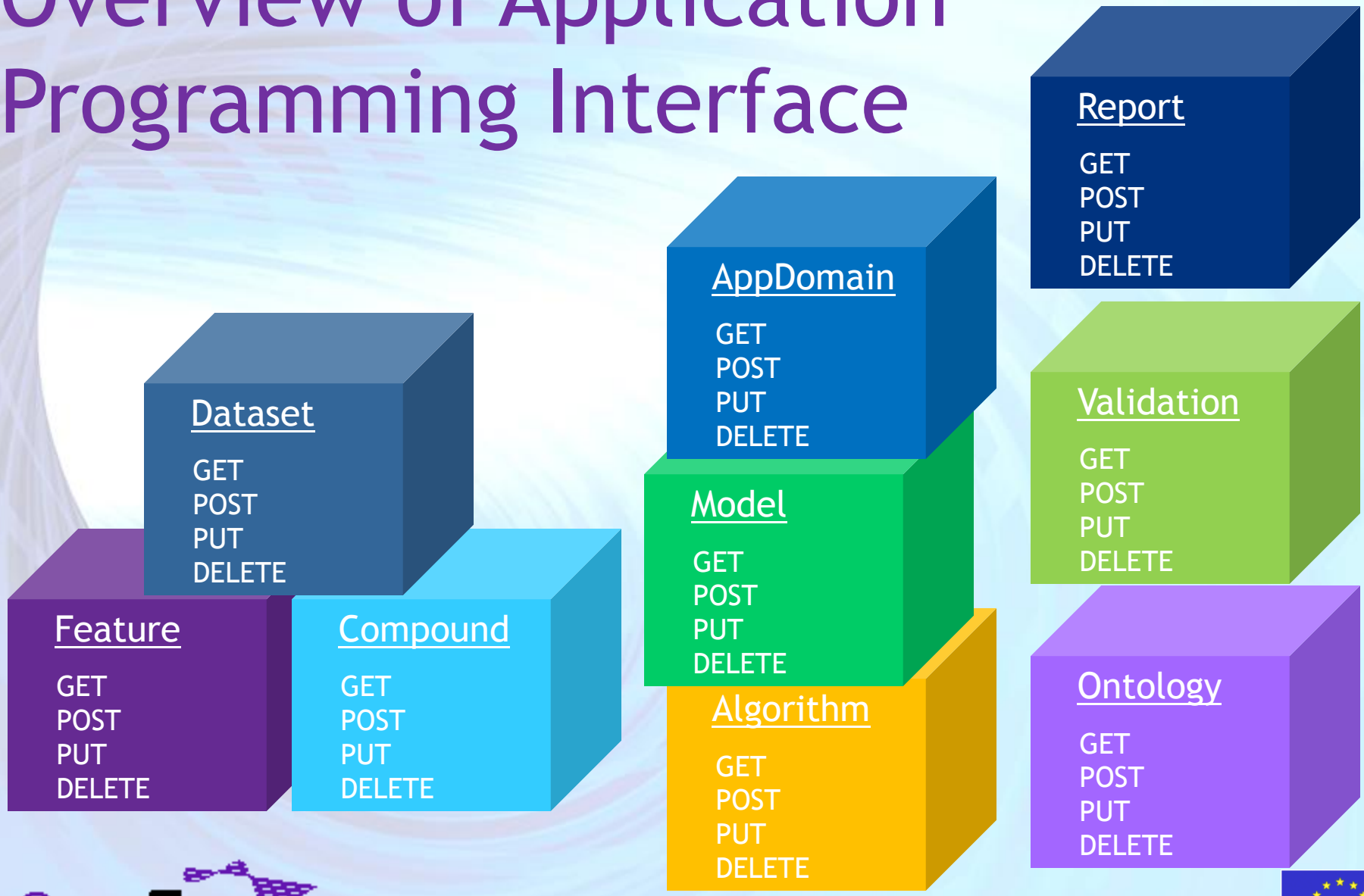
- Architectural style for distributed information systems on the Web
- Simple interfaces, data transfer via **hypertext transfer protocol (HTTP)**, stateless client/server protocol
  - GET, POST, PUT, DELETE
- Each **resource** is **addressed** by its own **web address**

Why?

- **Lightweight** approach to **web services**
- **Simplifies/enables** development of **distributed systems**
- (More or less) language independent/installation-free

OpenTox

Overview of Application Programming Interface

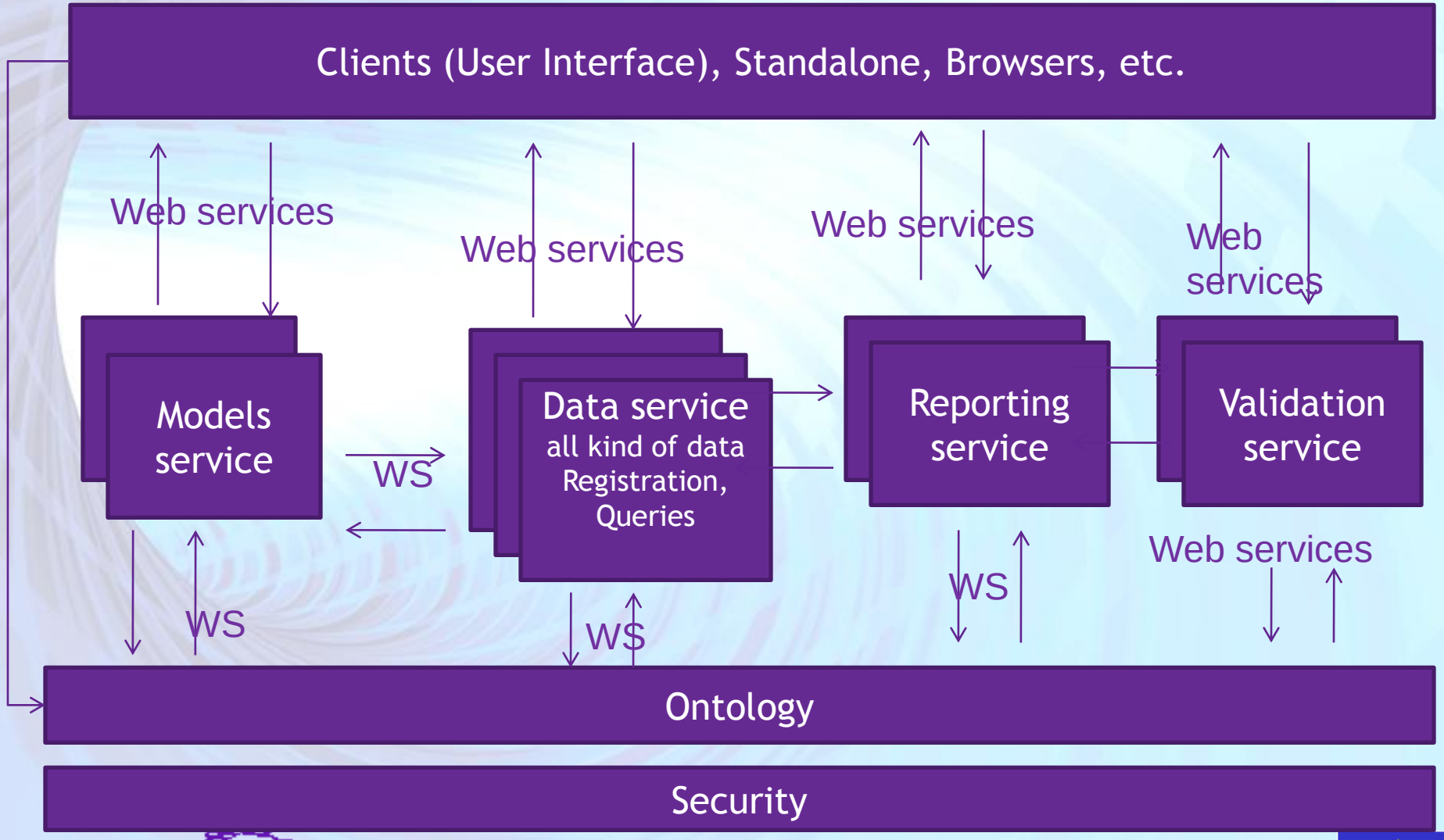


OpenTox

Interface Definitions

Description	Method	URI	Parameters	Result	Status codes
Retrieve SPARQL query results	GET	/ontology	**?query**=SPARQL_QUERY (mandatory)	RDF representation of the query results.	200,404,500
Predefined query to retrieve all models	GET	/ontology/models		RDF representation of all models.	
Predefined query to retrieve all endpoints	GET	/ontology/endpoints		RDF representation of all endpoints.	
Predefined query to retrieve all algorithms	GET	/ontology/algorithms		RDF representation of all algorithms.	
Submit SPARQL query and/or OpenTox service URL	POST	/ontology	**uri**[]=URL of a OpenTox RDF resource **query**=SPARQL_QUERY	RDF representation of the query results, if **query** is specified. if **uri[]** is specified, the server retrieves a RDF representation and adds it to the RDF storage, thus making it available for the subsequent queries.	200,404,500,502

Information exchange based on standardized ontologies and REST web services. All the information from different services might be used separately or via an ontology repository with reasoning capabilities



OpenTox

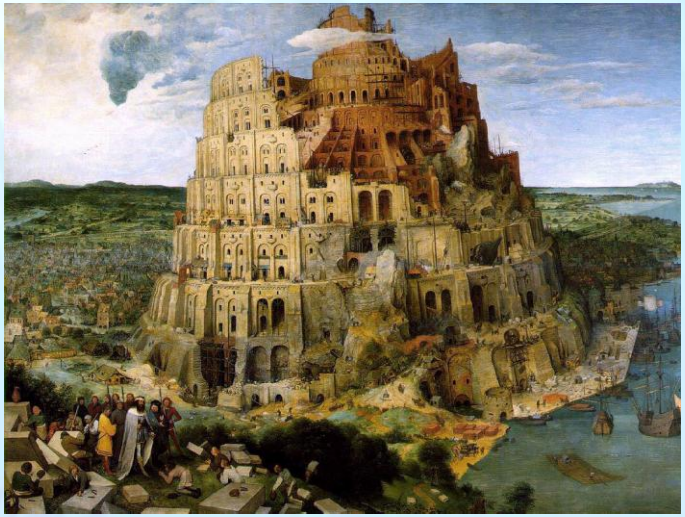
Ontologies: What and Why?

What?

- **Formal, shared conceptualization** of a **domain**

Why?

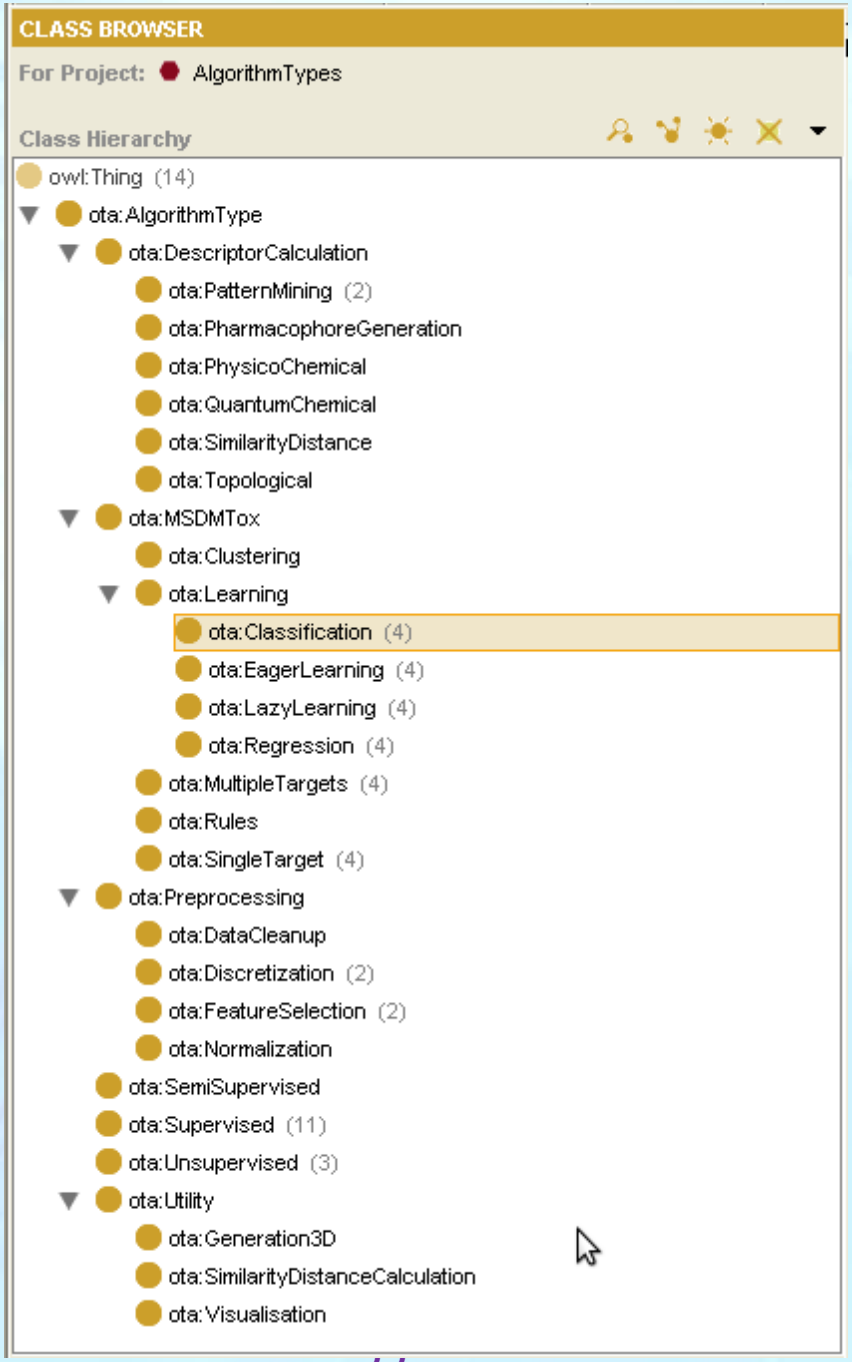
- Distributed services **need** to be able to „talk to each other“, i.e. have a **common understanding** of endpoints, any type of property, methods, etc.



OpenTox

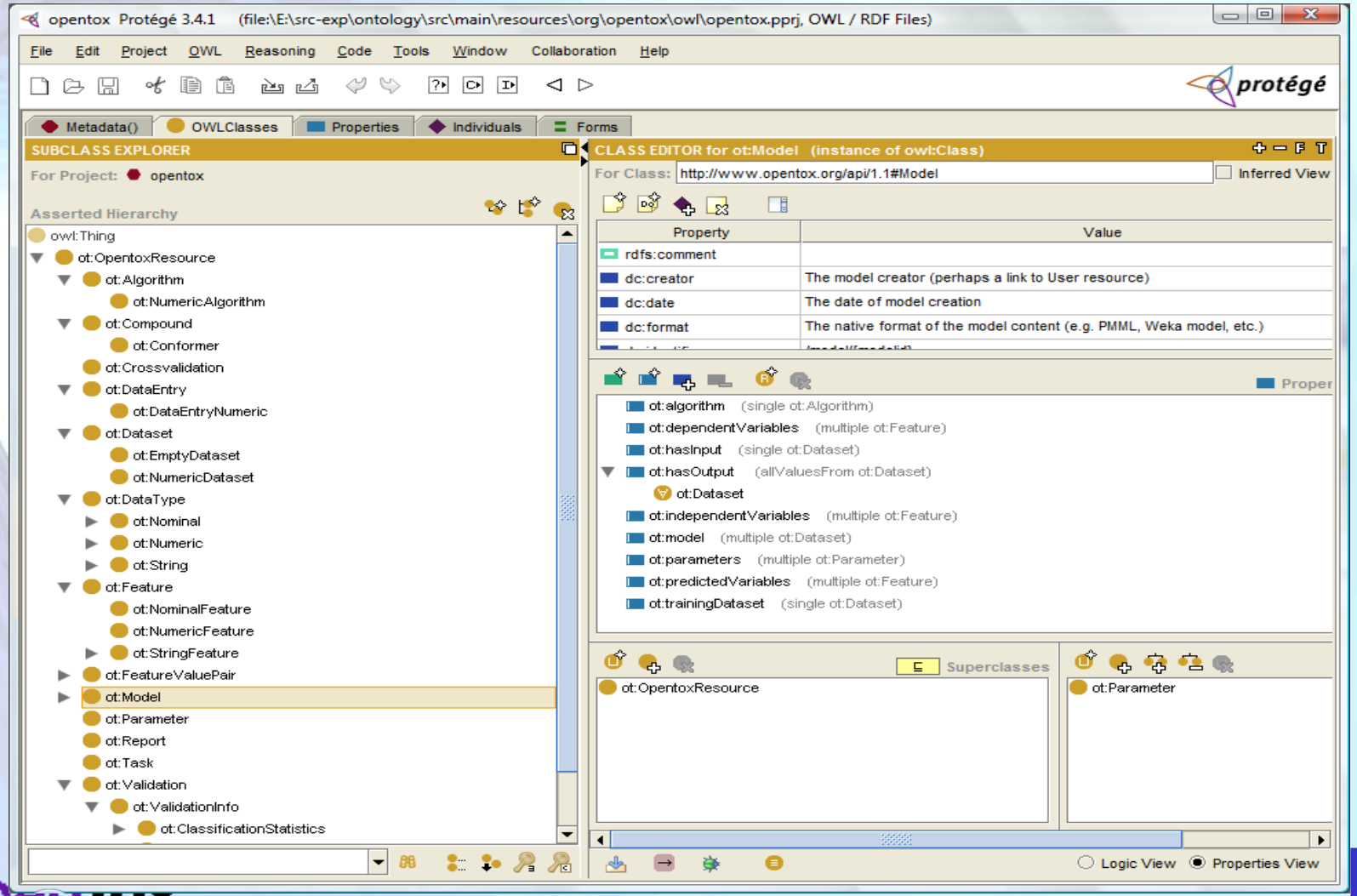
Ontologies

- Standards: **OWL** as representation language and **SPARQL** as query language
- There are many proposals for tox ontologies out there
- Our strategy: use existing work and standards wherever possible
- However, new ontology, e.g., for algorithms

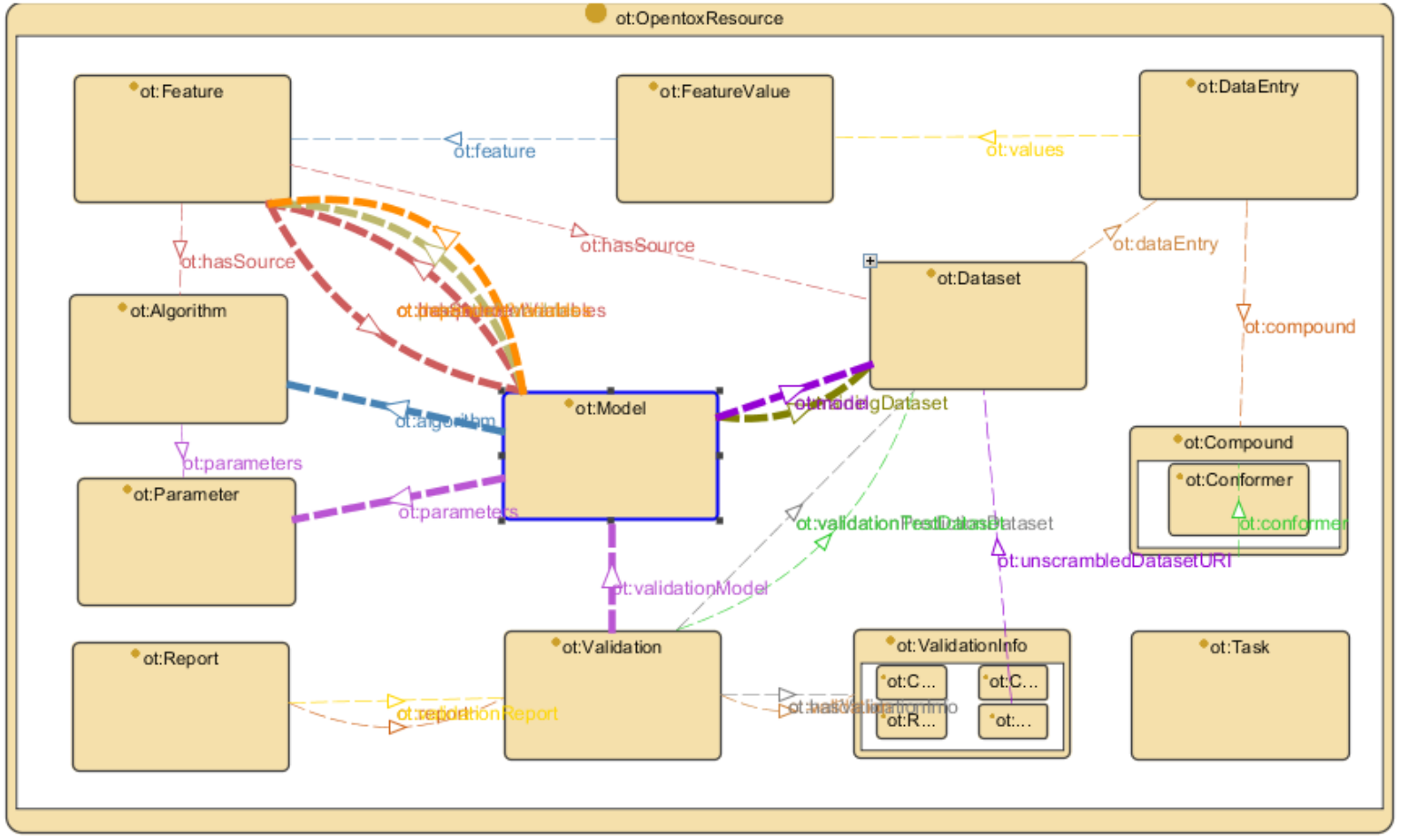


OpenTox

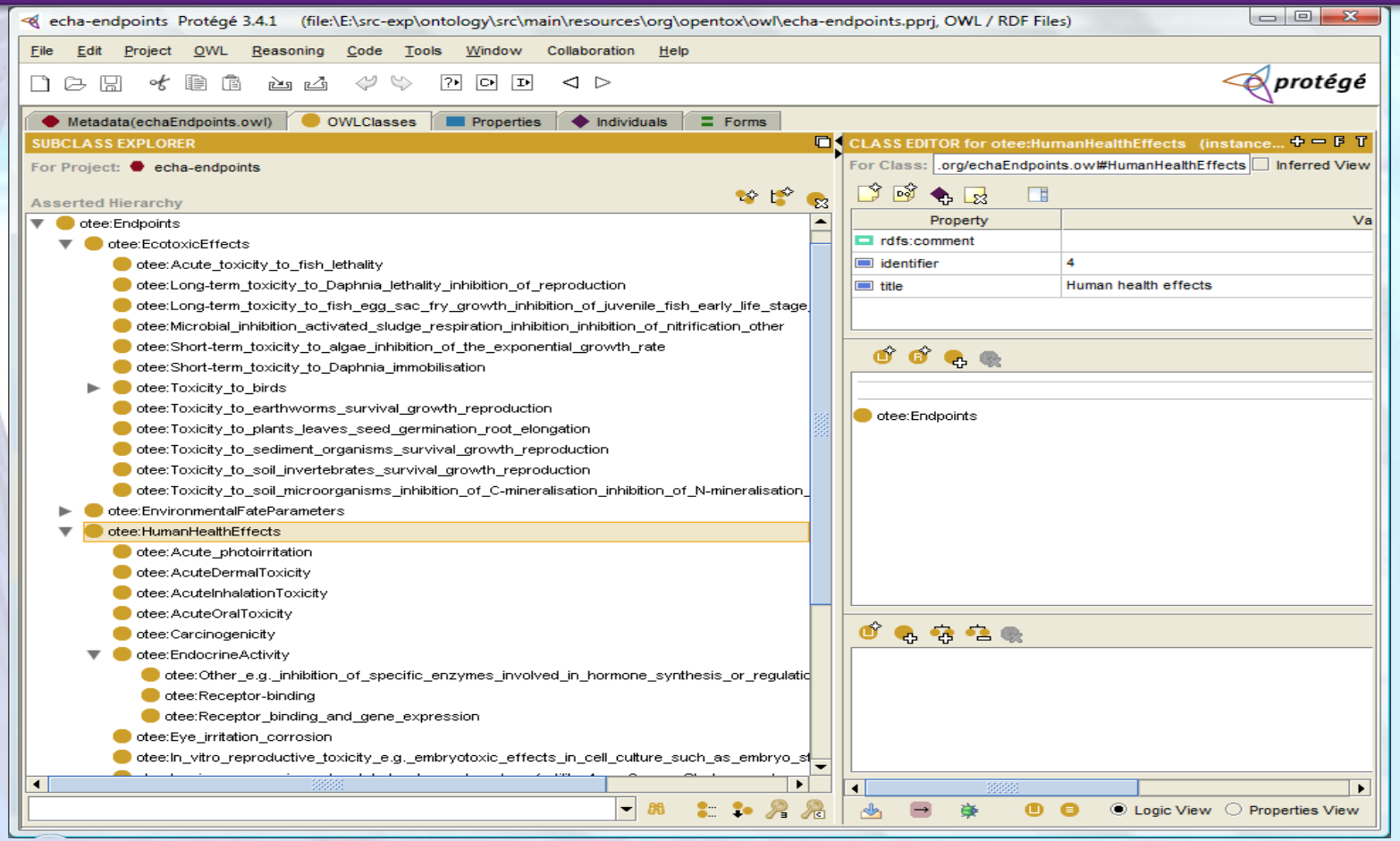
OpenTox object ontology
http://www.opentox.org/api/1.1/opentox.owl



OpenTox object ontology
http://www.opentox.org/api/1.1/opentox.owl

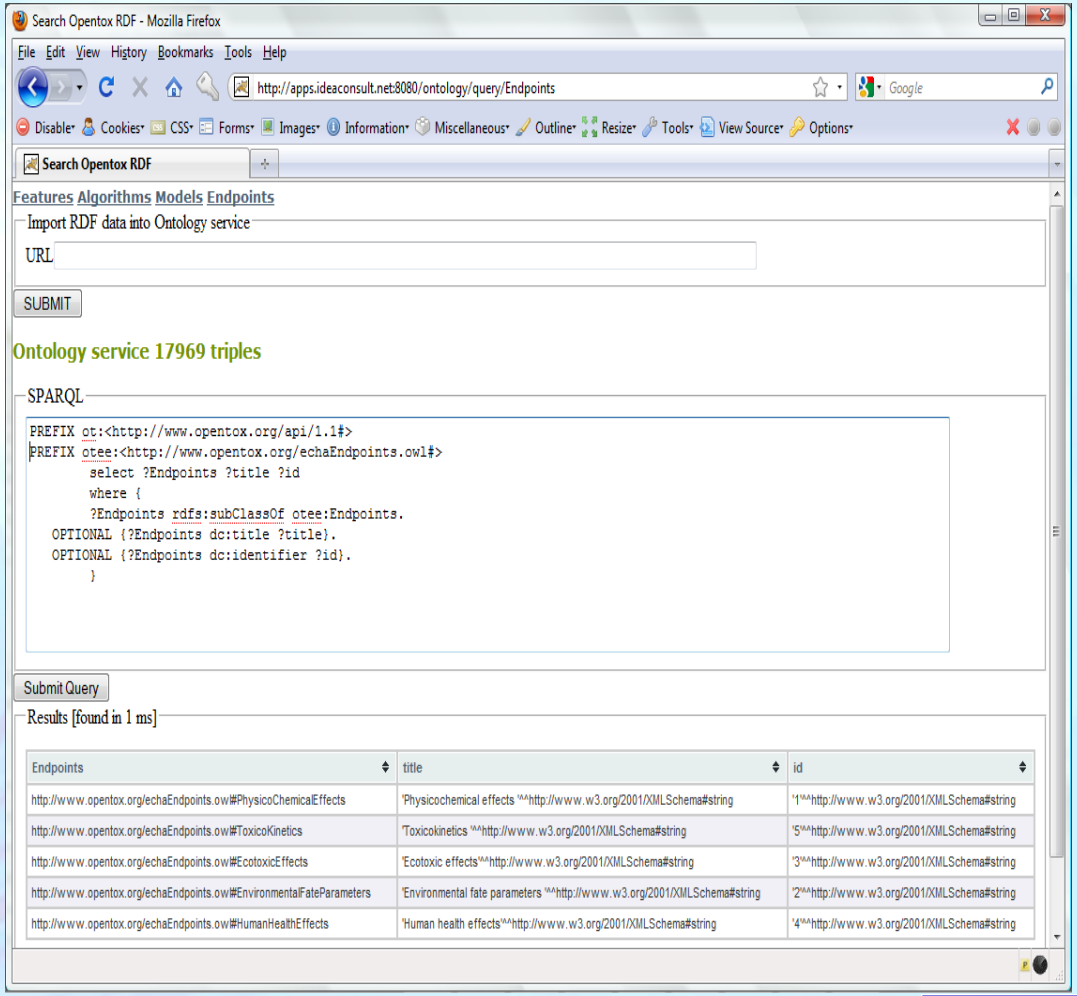


Toxicity endpoints ontology, based on ECHA classification http://www.opentox.org/api/1.1/echa-endpoints.owl



Ontology service

- RDF triple storage
- REST interface for registration of OpenTox objects
- SPARQL query



Current State of the System

Web services online:

- AlgorithmWS: NTUA, TUM, IDEA, IST
- ModelWS: NTUA, TUM, IDEA, IST
- FeatureWS: NTUA, TUM, IDEA, IST, ALU-FR
- CompoundWS: NTUA, TUM, IDEA, IST
- ValidationWS: ALU-FR
- DatasetWS: NTUA, TUM, IDEA, IST, ALU-FR
- OntologyWS: IDEA

OpenTox

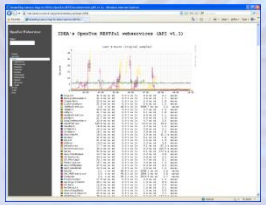
OpenTox database supported via Dataset service

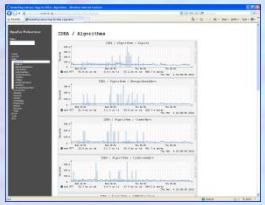
Dataset	OK	Probably OK	Probably ERROR	Unknown	Probably ERROR%
ECHA list of pre-registered substances	N/A	N/A	N/A	N/A	N/A
Chemical Identifier Resolver	67779	5314	3638	3471	4.75%
ChemIDplus	64802	7986	921	1745	1.24%
ChemDraw	17918	1147	502	478	2.57%
JRC PRS list	61332	4833	4022	2880	5.83%
ISSCAN	931	50	98	62	9.40%
CPDBAS	778	37	0	693	0%
DBPCAN	60	2	0	147	0%
EPAFHM	281	5	0	331	0%
KIERBL	102	1	0	175	0%
IRISTR	346	16	0	177	0%
FDAMDD	213	19	1	983	0.08%
ECETOC skin irritation	158	12	0	5	0%
Skin sensitisation (LLNA)	160	7	4	38	1.95%
Bioconcentration factor (BCF) Gold Standard Database	N/A	N/A	N/A	N/A	N/A

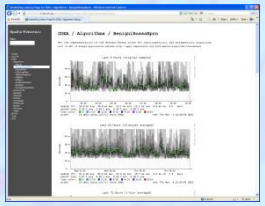
OpenTox

Implementation: key facts (1)

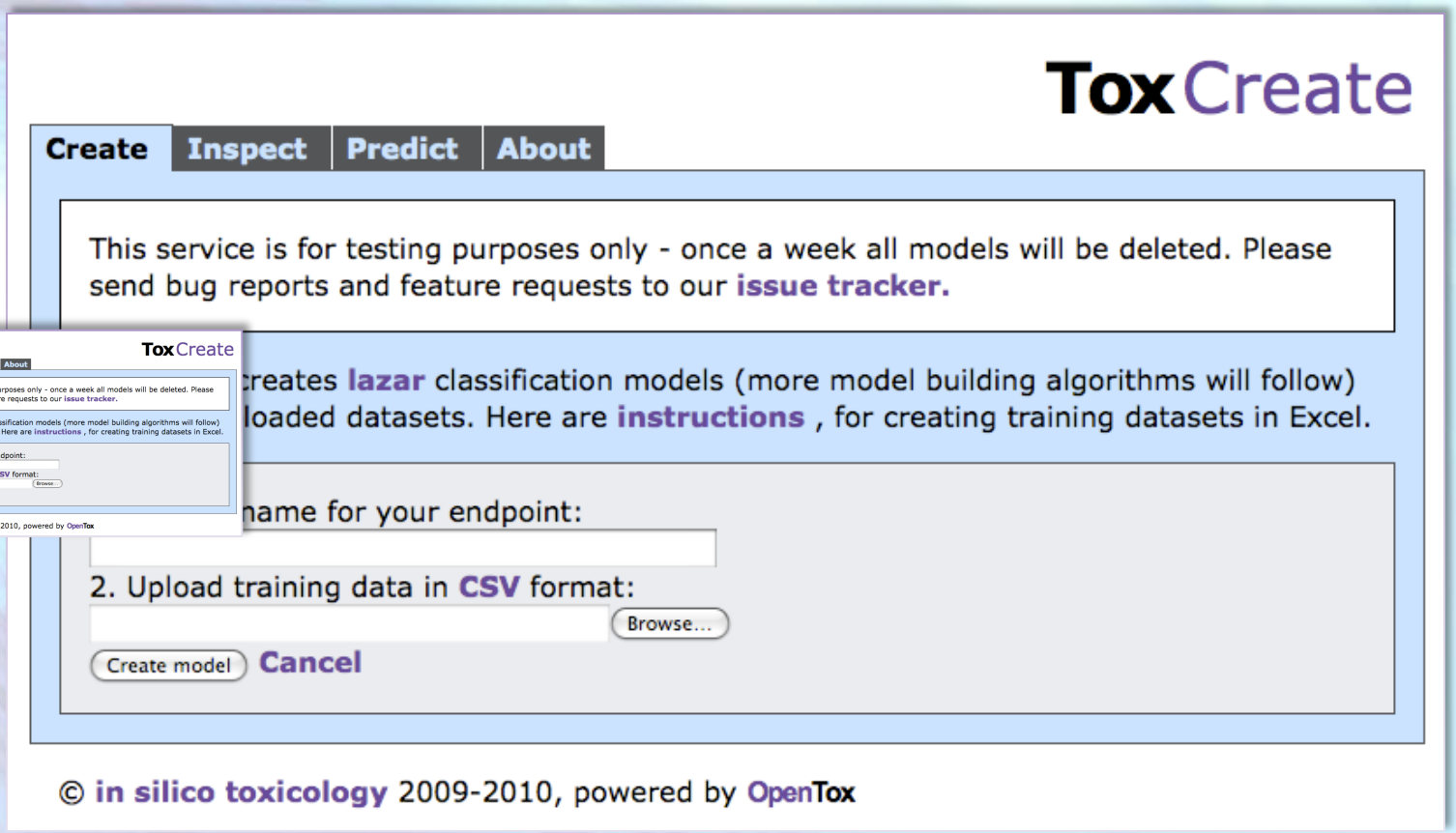
- Seven independent open source implementations of different OpenTox API-v1.1 subsets have been developed in three different programming languages (Java – 4, Ruby - 2, Delphi – 1);
- Continuous availability and performance monitoring of a selected subset of OpenTox webservices (including integration aspects), running and gathering detailed statistics since December 2009;
- **Next major milestone – design and implementation of authentication and authorisation (AA) support**;







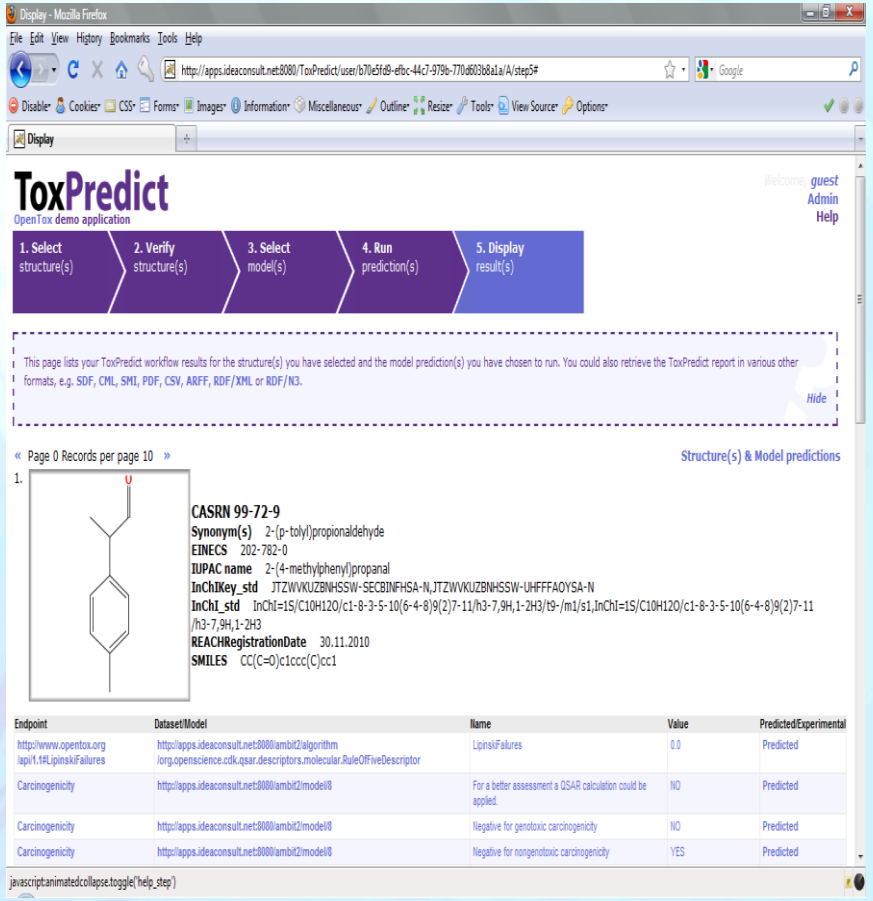
Use case 1): ToxCreate



OpenTox

Use case 2: ToxPredict http://toxpredict.org

- ToxPredict estimates the chemical hazard of chemical structures. It relies on [OpenTox API-v1.1] compliant RESTful webservices.
- Users can either search the OpenTox prototype database, which includes currently quality labelled data for **~150,000 chemicals**, grouped in more than a dozen datasets, or upload their own chemical structure data. ToxPredict provides access to **14 ready to use models**, addressing **14 different endpoints** (and growing!);



ToxPredict: Step 1 (Select structure(s))

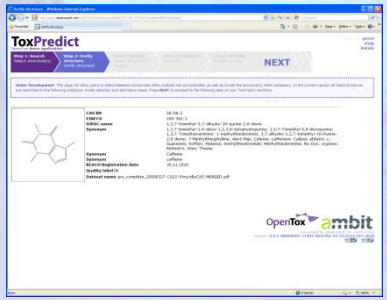
Find structure by name, registry number, SMILES, InChI, structure, substructure, similarity…

ToxPredict Web Application

OT Dataset API ***HTTP GET***

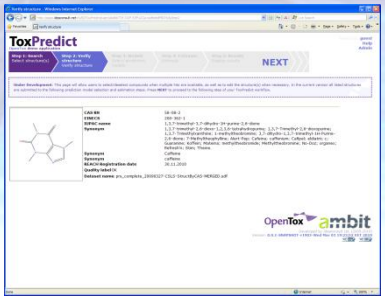
OT Dataset Service

text/uri-list,
application/rdf+xml,
chemical/x-daylight-smiles
chemical/x-mdl-sdfile,…



Here is the list of structures as URI links, RDF, MOL or SMILES.

ToxPredict: Step 3 (Select model(s))



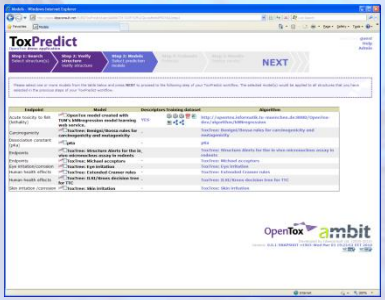
What prediction models are available? Is there a model for endpoint X?

ToxPredict Web Application

HTTP GET SPARQL query

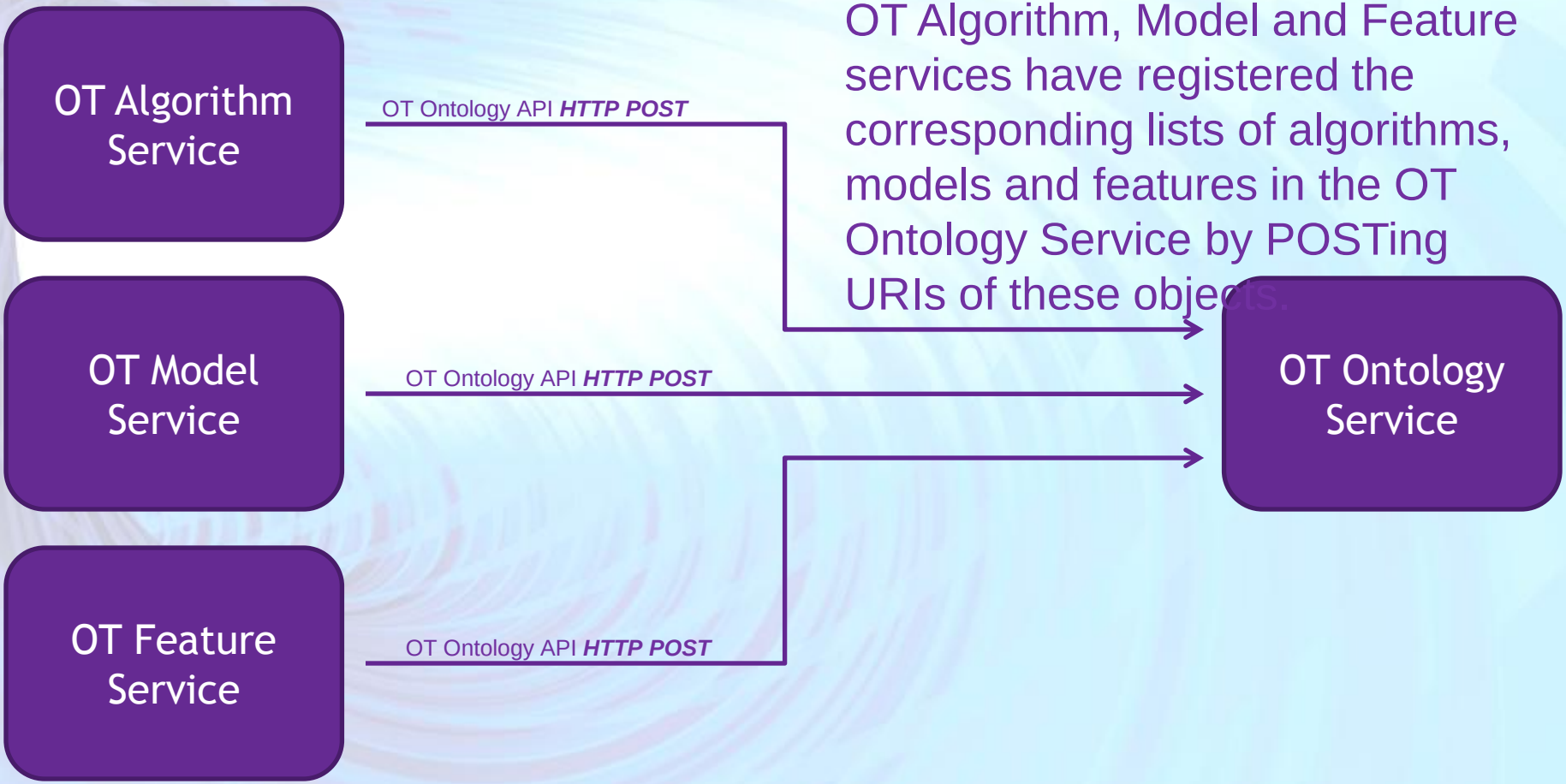
OT Ontology Service

application/sparql-results+xml



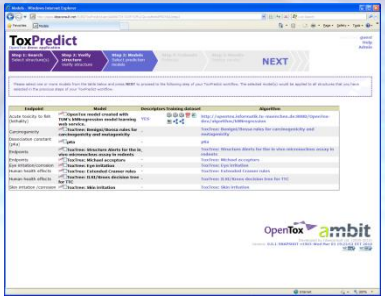
Here is the list of model URIs and related endpoints and algorithms in SPARQL format.

ToxPredict: Step 3 (behind the scenes)



OT Algorithm, Model and Feature services have registered the corresponding lists of algorithms, models and features in the OT Ontology Service by POSTing URIs of these objects.

ToxPredict: Step 4 (Estimate)



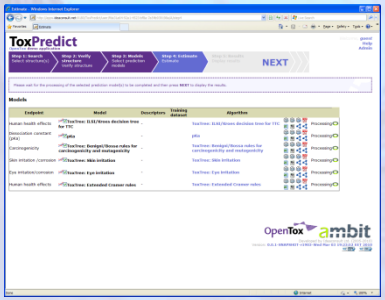
Run the selected models.

ToxPredict Web Application

OT Model API ***HTTP POST*** with parameter dataset URI from steps 1-2

OT Model Service

HTTP code 202 "Accepted"
Model task URI in HTTP:location header

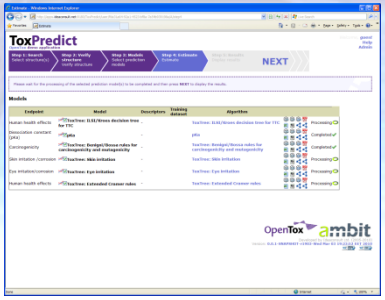


The calculation will take a while; here is a task URI, which can be queried for processing status.

Create a new task.

OT Task Service

ToxPredict: Step 4 (Estimate)



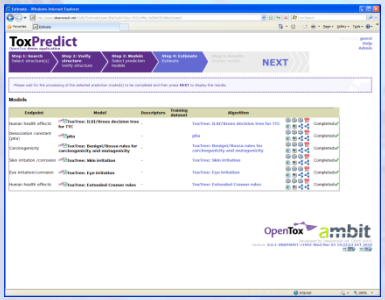
Is the task completed?

ToxPredict Web Application

OT Task API ***HTTP GET*** on task URI

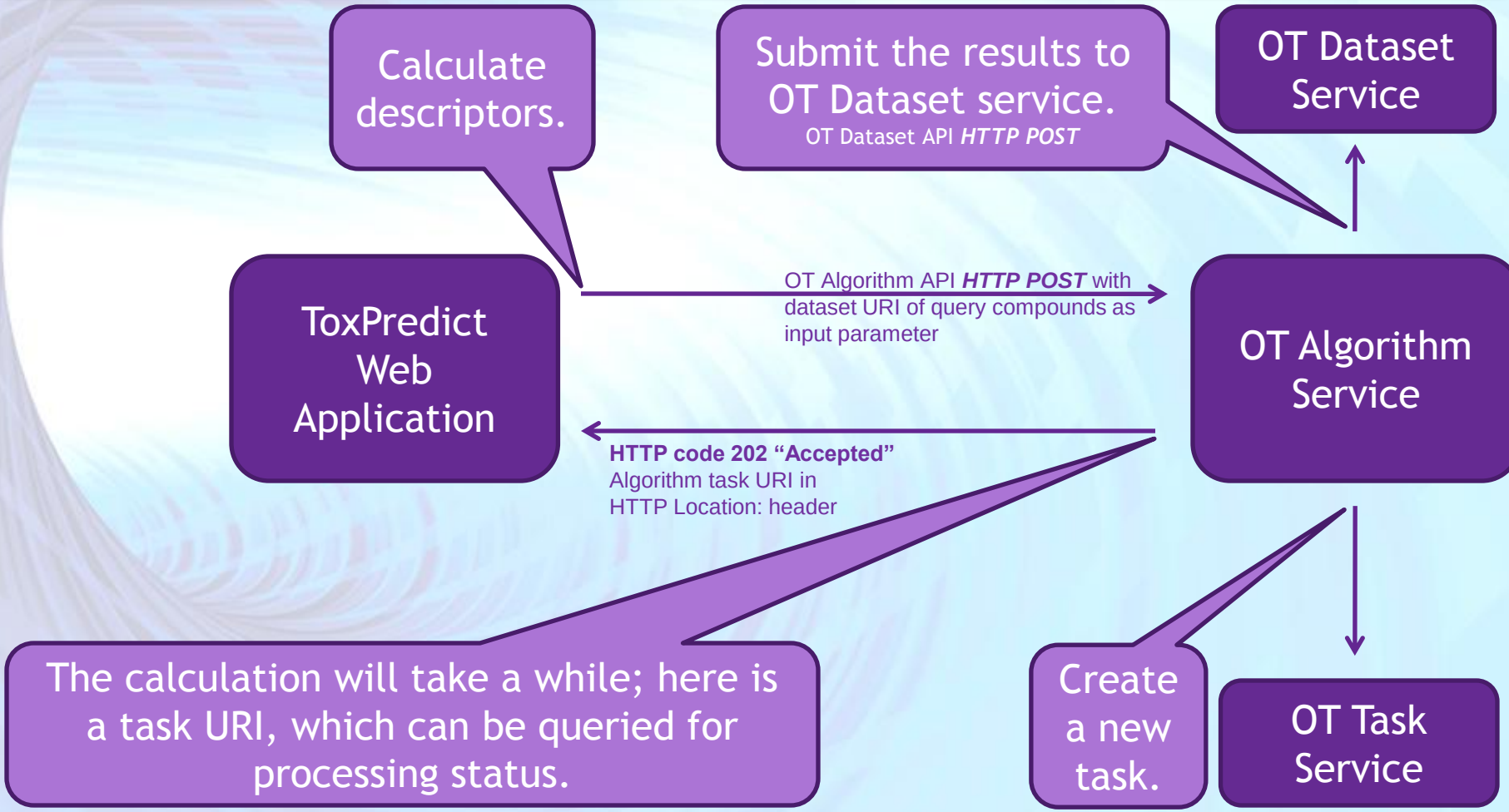
OT Task Service

HTTP code 200 "OK"
Dataset URI in HTTP:location header

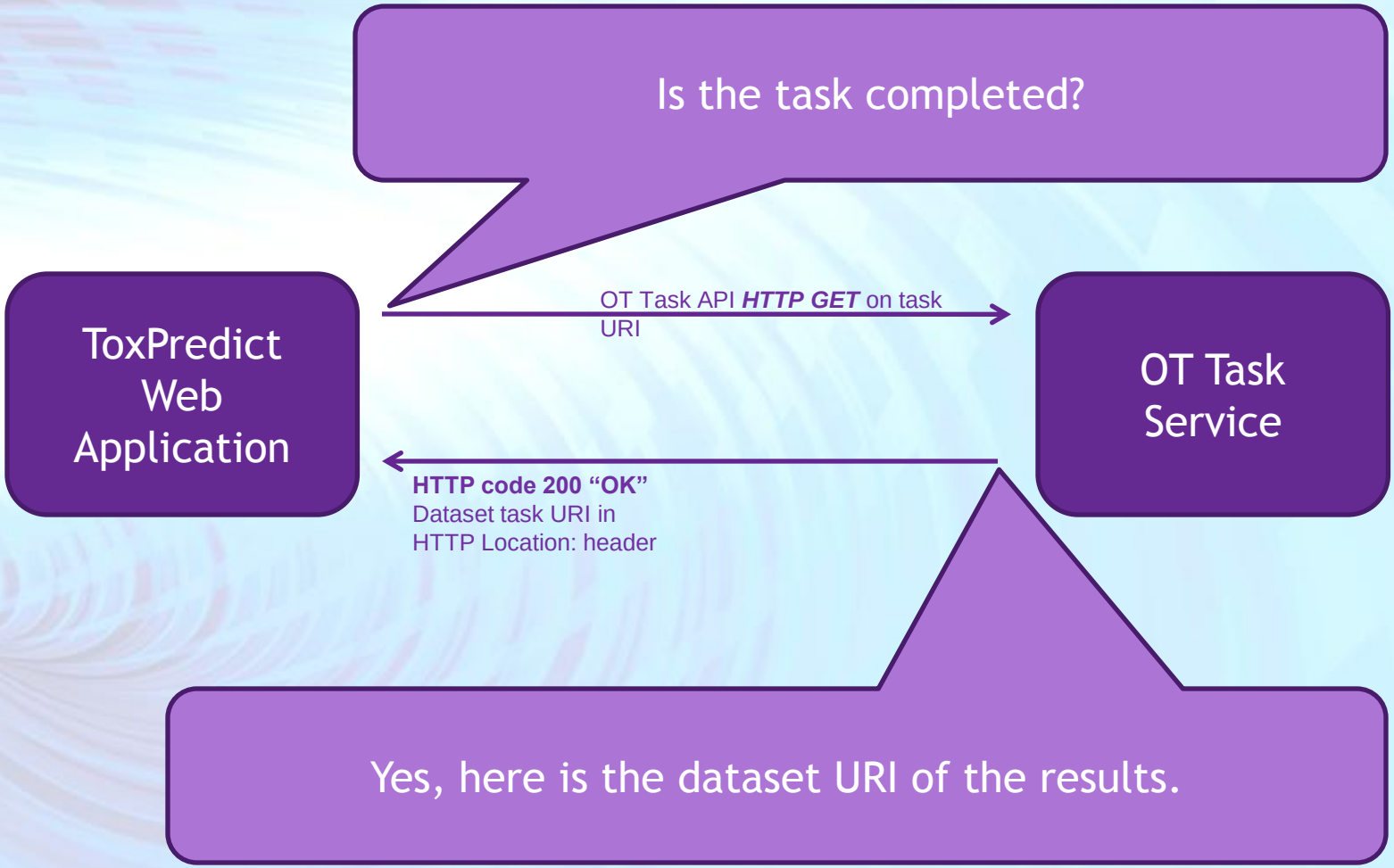


Yes, here is the dataset URI of the results.

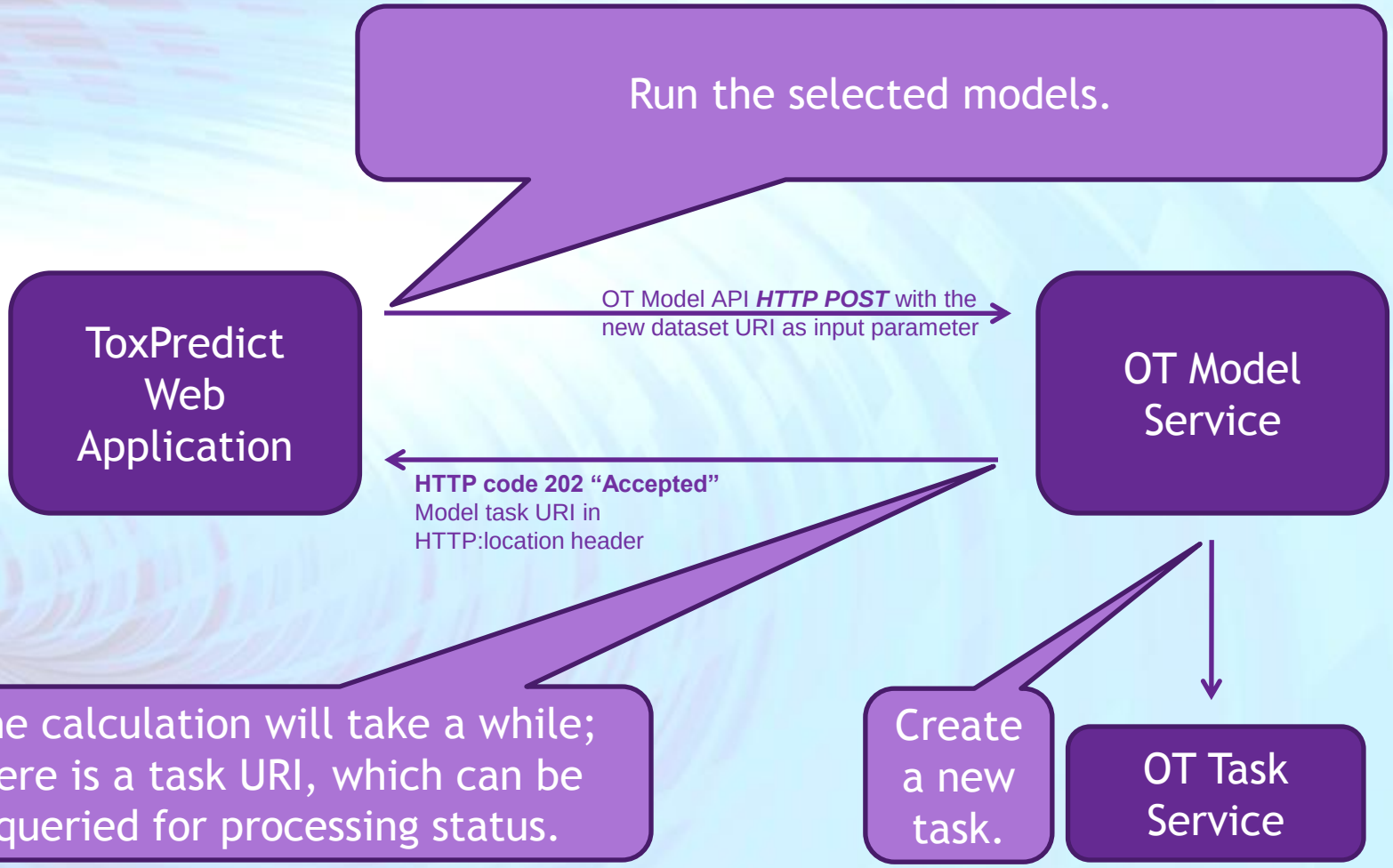
ToxPredict: Step 4 (behind the scenes)



ToxPredict: Step 4 (behind the scenes)

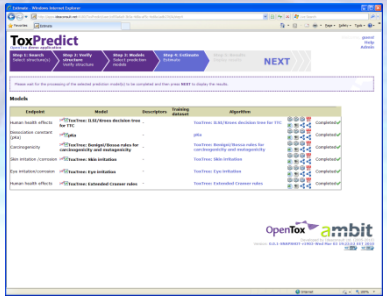


ToxPredict: Step 4 (behind the scenes)



… the remaining part of the Estimation step has been described already in slides 11 and 12

ToxPredict: Step 5 (Display results)



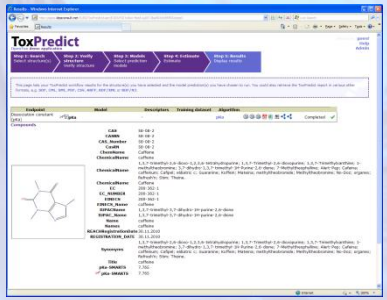
Retrieve calculation results from the final dataset URI, obtained in Step 4 (Estimation).

ToxPredict Web Application

OT Dataset API ***HTTP GET***

OT Dataset Service

application/rdf+xml



Here is the dataset content in RDF format, according to OpenTox.owl and containing estimation results, as well as compound identifiers and experimental data.

Summary and Future Work

- Comprehensive **framework** for **predictive toxicology** that allows you to address a wide range of tasks
  - interface definitions, services (also for: validation, **reporting**, …), use cases
  - **interoperability** and **extensibility by design**
- Further work: e.g.,
  - Infrastructure and API for providing restricted access to data and models
  - Improvements in ToxCreate and ToxPredict use cases
  - Read-across (technical infrastructure available),
  - ToxCast analysis
  - Pathway prediction service: developing a service for predicting transformation products

OpenTox

Do you develop models and algorithms for prediction of biological effects?

It can be made available as OpenTox web service!

Thank You For Your Attention!

OpenTox events:

- 1 day OpenTox workshop at AXLR8 meeting (Potsdam, Germany, May 30 2010)
- eChemInfo hands-on training Cambridge, UK, Aug 2010
- Oral & Poster Presentation @ ACS National Meeting , section *Semantic Web in Chemistry*, Boston, MA, USA, Aug 2010
- 1day OpenTox workshop at EuroQSAR 2010 (Rhodos, Greece, Sep 2010)

OpenTox