Development and Use of Predictive Toxicology Applications

OpenTox Workshop

19 September 2010

Rhodes, Greece

Barry Hardy (Douglas Connect)

OpenTox

SEVENTH FRAMEWORK PROGRAMME

Collaborating Partners

In Silico Toxicology, Switzerland

Douglas Connect, Switzerland

Albert Ludwigs University Freiburg, Germany

Ideaconsult, Bulgaria

National Technical University of Athens, Greece

Istituto Superiore di Sanità, Italy

Fraunhofer Institute for Toxicology & Experimental Medicine, Germany

Technical University of Munich, Germany

David Gallagher, UK

Institute of Biomedical Chemistry of the Russian Academy of Medical Sciences, Russia

Seascape Learning & JNU, India

OpenTox

OpenTox Advisory Board

- European Centre for the Validation of Alternative Methods
- Pharmatrope
- Bioclipse
- U.S. Environmental Protection Agency
- U.S. Food & Drug Administration
- Nestlé
- Roche
- AstraZeneca
- LHASA
- Leadscope
- University of North Carolina
- EC Environment Directorate General
- Organisation for Economic Cooperation & Development
- CADASTER
- Bayer Healthcare

OpenTox

Journal of Cheminformatics Publication

Collaborative development of predictive toxicology applications
Journal of Cheminformatics 2010, 2:7 doi:10.1186/1758-2946-2-7

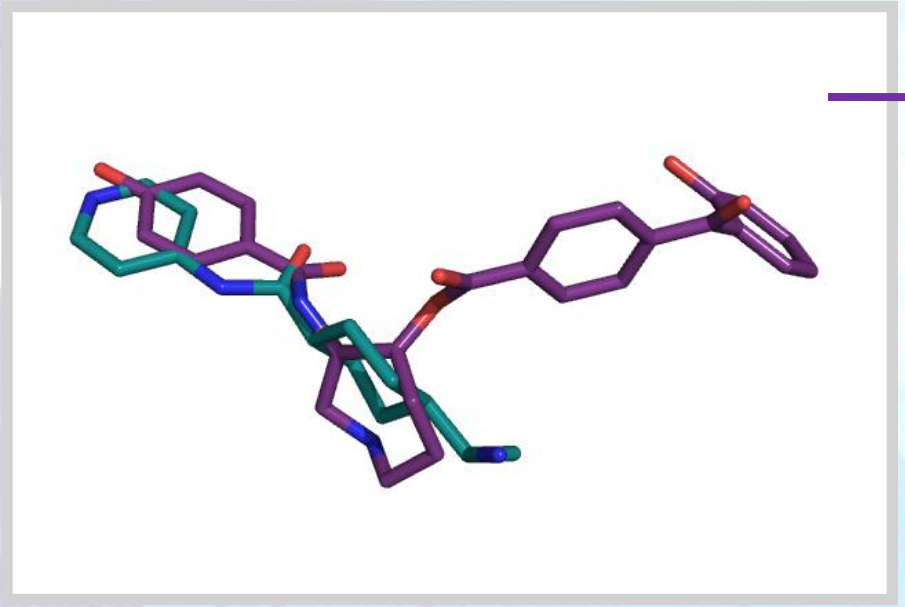
Barry Hardy, Nicki Douglas, Christoph Helma, Micha Rautenberg, Nina Jeliazkova, Vedrin Jeliazkov, Ivelina Nikolova, Romualdo Benigni, OlgaTcheremenskaia, Stefan Kramer, Tobias Girschick, Fabian Buchwald, JoergWicker, Andreas Karwath, Martin Gutlein, Andreas Maunz, Haralambos Sarimveis, Georgia Melagraki, Antreas Afantitis, Pantelis Sopasakis, David Gallagher, Vladimir Poroikov, Dmitry Filimonov, Alexey Zakharov, Alexey Lagunin, Tatyana Gloriozova, Sergey Novikov, Natalia Skvortsova, Dmitry Druzhilovsky, Sunil Chawla, Indira Ghosh, Surajit Ray, Hitesh Patel and Sylvia Escher

Open Access publication available at
www.jcheminf.com/content/2/1/7

OpenTox

Collaborative Predictive Toxicology Challenge

Input Structure



VO →

Out - Toxic or Not?

- LD50
- Liver Toxicity
- Secondary Metabolites
- Bioavailability
- Mutagenicity
- Carcogenicity
- ReproductiveToxicology
- Skin Irritation
- Aqua Toxicity
- Combined predictions for arrays of mutiple end points

REACH
(enRegistrement, Evaluation et Autorisation des substances CHimiques)

Driver →

Increasing demands on industry to satisfy safety evaluation and risk assessment required by REACH legislation. (Over 140k cmpds registered).

OpenTox

ToxPredict

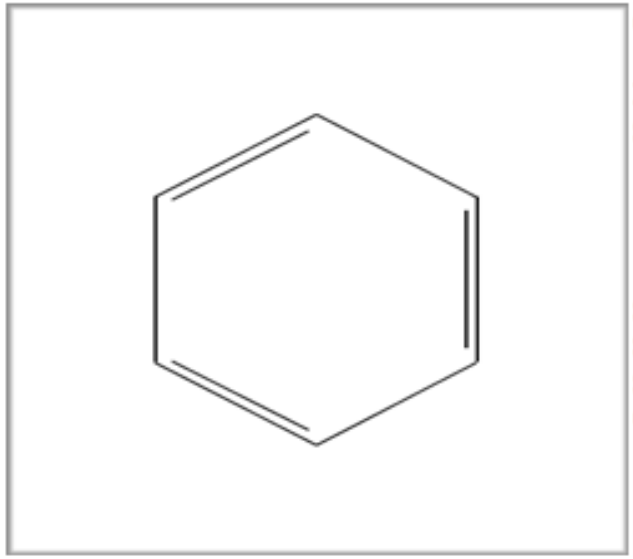
OpenTox demo application

Welcome, **guest**
Help
Admin

Step 1: Search Select structure(s) | **Step 2: Verify structure** Verify structure | **Step 3: Models** Select prediction models | **Step 4: Estimate** Estimate | **Step 5: Results** Display results

This page lists your ToxPredict workflow results for the structure(s) you have selected and the model prediction(s) you have chosen to run. You could also retrieve the ToxPredict report in various other formats, e.g. **SDF**, **CML**, **SMI**, **PDF**, **CSV**, **ARFF**, **RDF/XML** or **RDF/N3**.

Download as



CAS RN	71-43-2
EINECS	200-753-7
IUPAC name	benzene
Synonym	(6)annulene; benzine; Benzol; Benzolene; bicarburet of hydrogen; carbon oil; Coal naphtha; cyclohexatriene; mineral naphtha; motor benzol; nitration benzene; Phene; Phenyl hydride; pyrobenzol.
Synonym	21742.0
Synonym	Benzene
Synonym	benzene
Quality label	OK

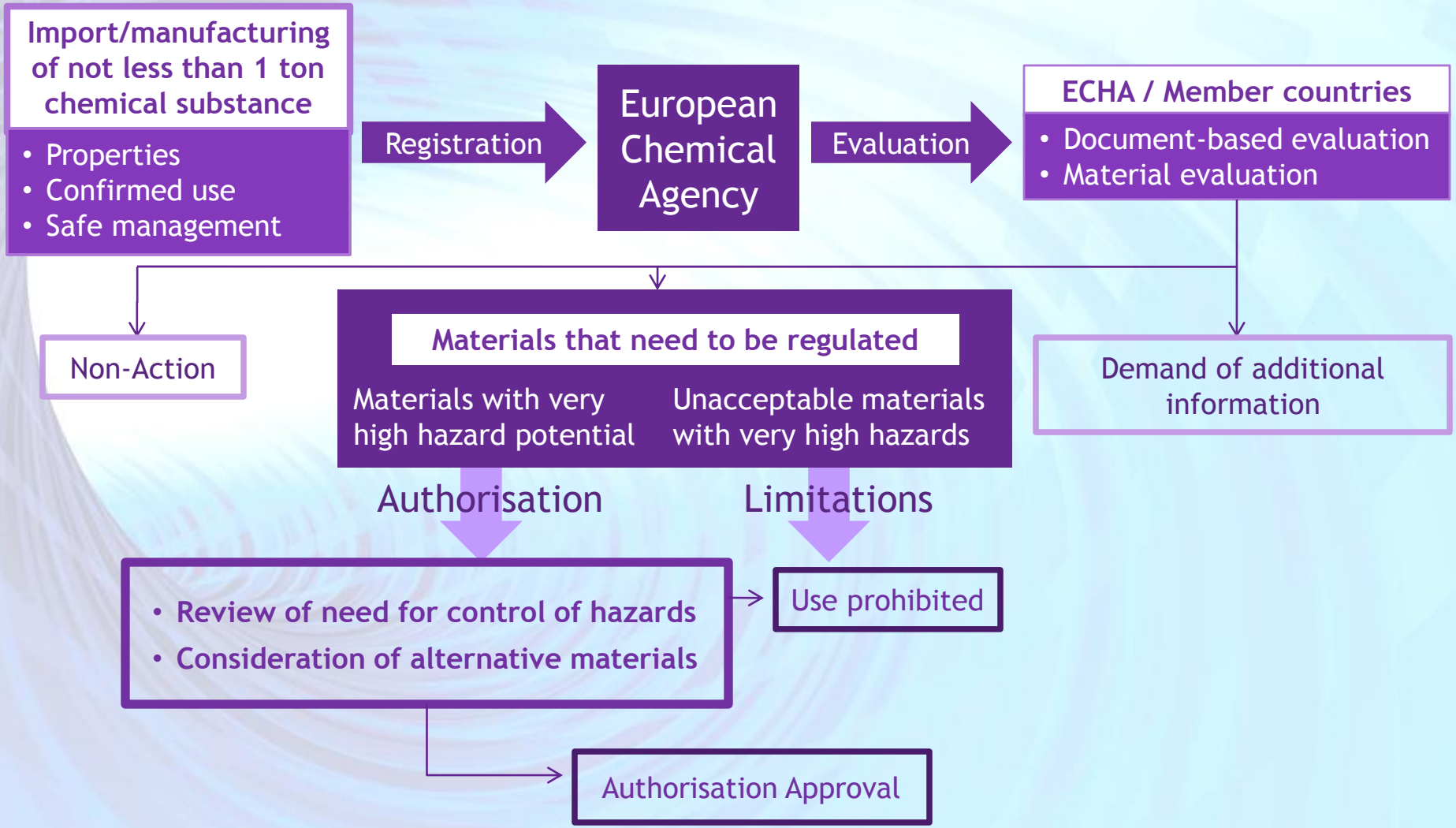
MolecularWeight MolecularWeight

MW	78.1112

REACH

OpenTox

REACH Registration



OpenTox

Challenges to Integrated Resources & Applications

- Database silos
- Missing information
- Varying quality
- Hard to integrate data
- Hard to integrate models
- No common framework
- Lack of standards
- Lack of validation
- Complex subject
- Application difficult
- Lack of transparency
- Interdisciplinary collaboration

OpenTox

Value is in Linking

The key idea of Google's founders in creating their search engine:

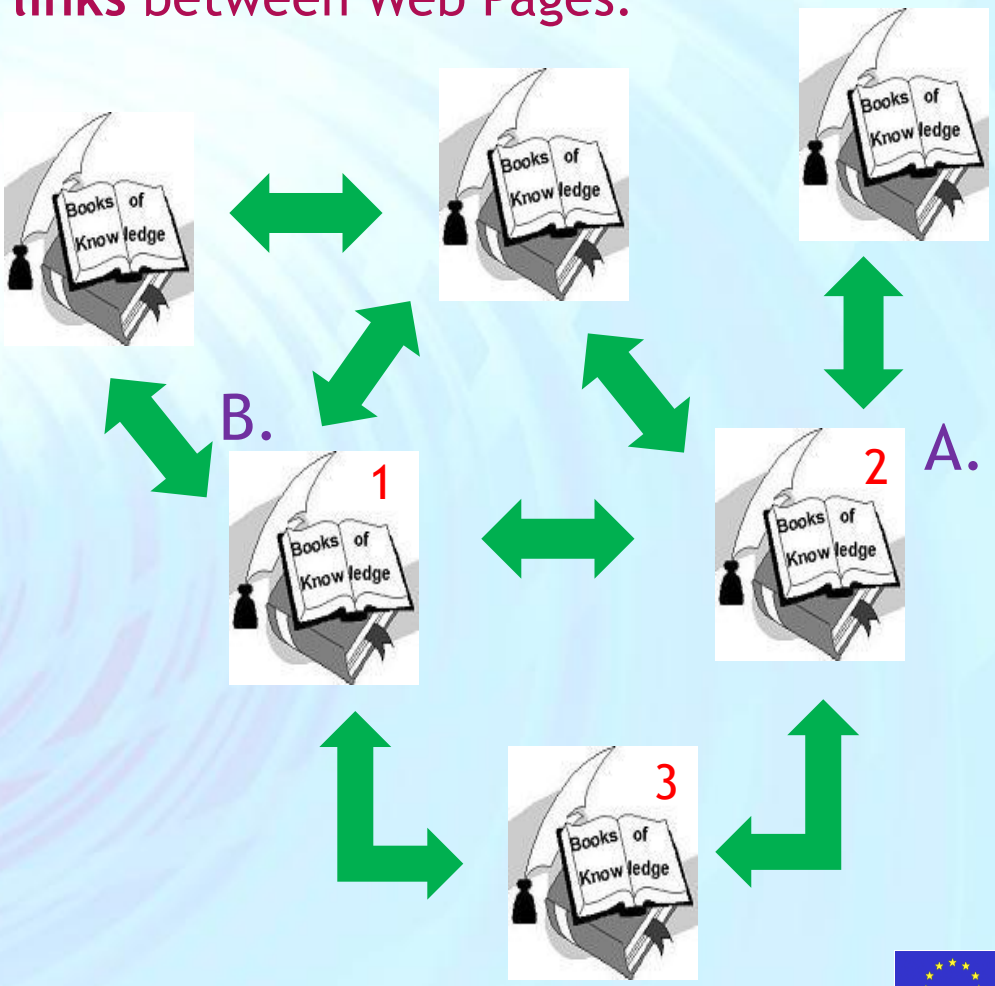
There is useful knowledge in the **links** between Web Pages.

Page Ranking

A page is ranked higher in a search if:

A. it has more connections to it than other pages

B. the pages connecting to it have higher ranking themselves



OpenTox

Linked Data enables Knowledge Creation, Combination and Analysis

Linked Data is a term used to describe the exposing, sharing, and connecting of data on the Semantic Web using:
URIs a generic means to identify entities in the world
HTTP a simple yet universal mechanism for retrieving resources
RDF a generic graph-based data model with which to structure and link data

Linked Data needs:

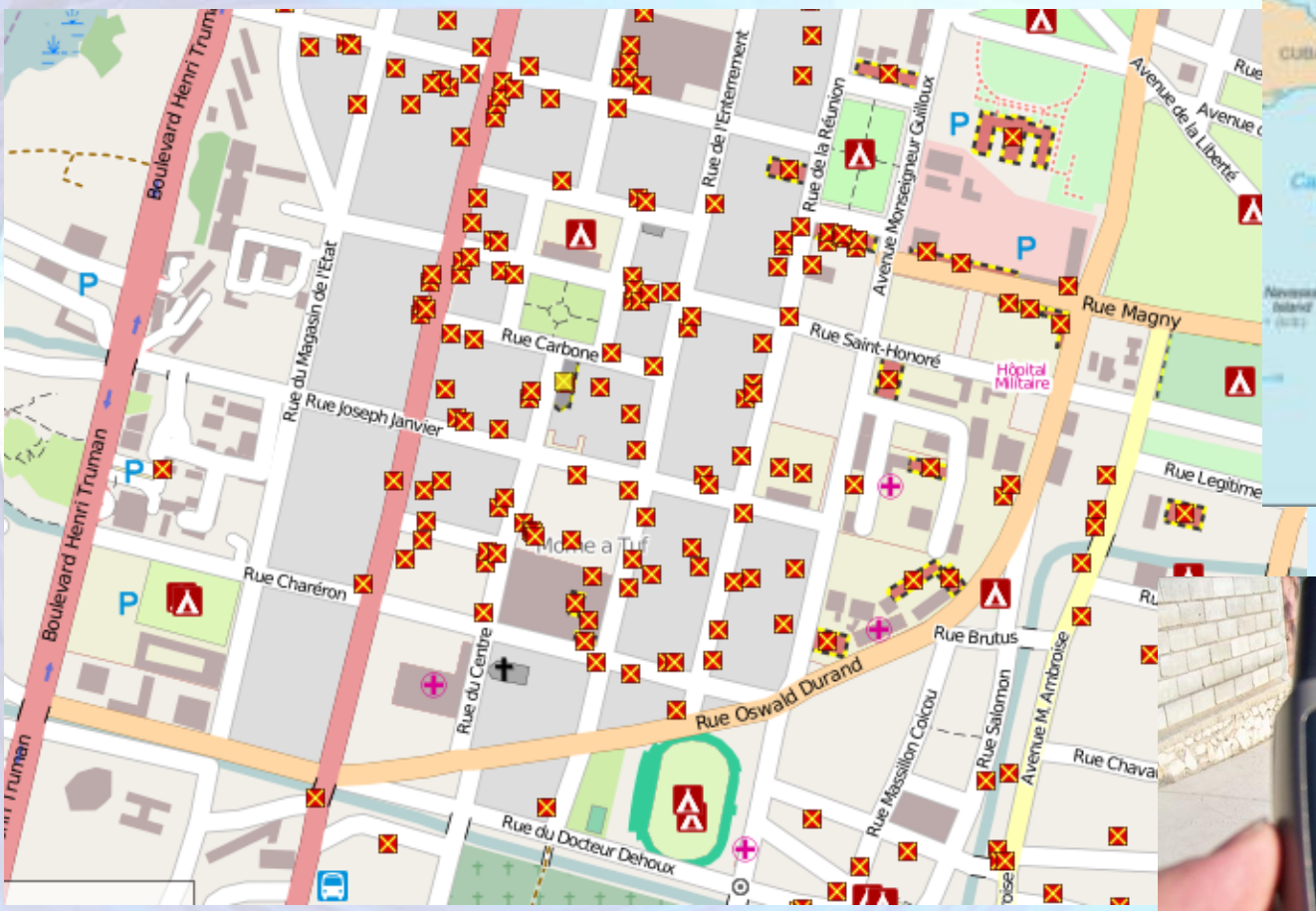
1. Provision of a **URI** that describes a Data Resource
2. Use of **HTTP** to retrieve useful data from the **URI**
3. A Data Format described with standardised semantics (so relationships are enabled) e.g. **RDF**
4. Data should provide links to other Data (through **URIs**)

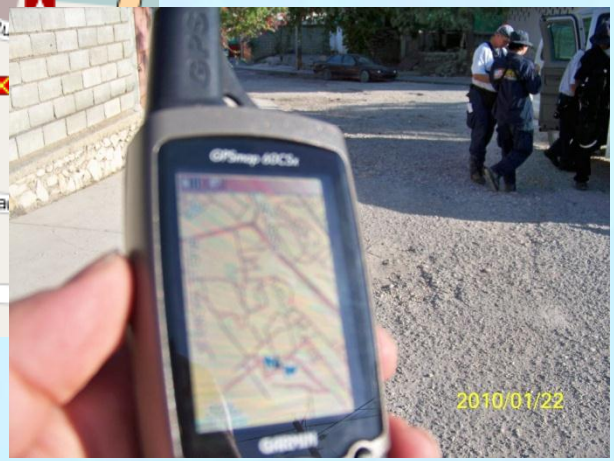
DBpedia = Linked Data approach applied to Wikipedia

Linked Data approach can also be applied to other resource types e.g., for algorithms or models as done in OpenTox…

OpenTox

Solution created by Linked Open Data, Web Applications and Crowdsourcing



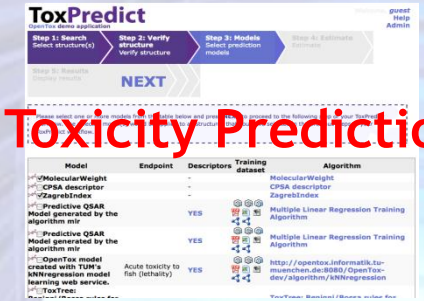


Haiti Earthquake Crisis Response (2010)

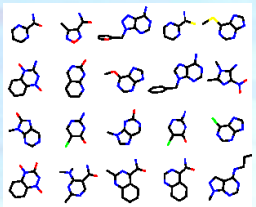
wiki.openstreetmap.org

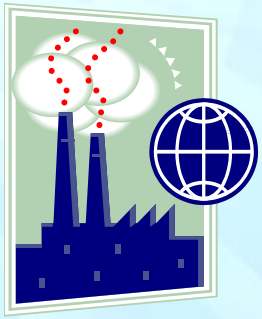
OpenTox

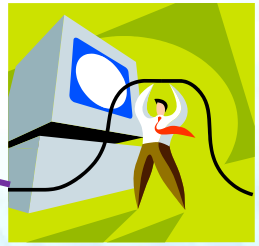
Accelerating Knowledge Flows in Predictive Toxicology

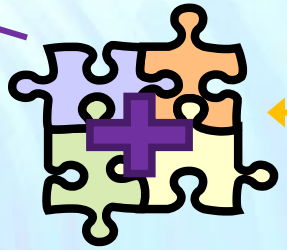


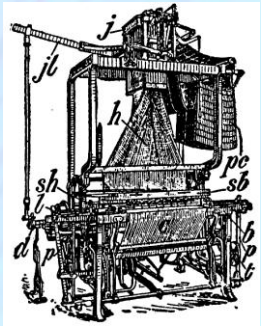
Toxicity Predictions











OpenTox

OpenTox is an Integrating Framework

Framework	• Toxicity Data (Linked) • *in silico* models • Validation & Reporting • Interpretation aids
Diverse Access	• Toxicolog, Biolog, Chem - ists • Computational Scientists • Interfaces for new algorithm development & integration
Interoperability	• Promote Standards • Core Open Source Components • Support Ontologies & Integration of Multiple Resources

OpenTox

	OECD Principle	OpenTox addresses Validation Principles by...
1	Defined Endpoint	providing a unified source of well defined and documented toxicity data with a common vocabulary
2	Unambiguous Algorithm	providing transparent access to well documented models and algorithms as well as to the source code
3	Defined Applicability Domain	integrating tools for the determination of applicability domains during the validation of prediction models
4	Goodness-of-fit, robustness and predictivity	providing scientifically sound validation routines for the determination of errors and confidences
5	Mechanistic interpretation (if possible)	integrating tools for the inference, correlation or prediction of toxicological mechanisms and the recording of opinions and analysis in reports

OpenTox

Overview of Application Programming Interfaces

Report
GET
POST
PUT
DELETE

AppDomain
GET
POST
PUT
DELETE

Dataset
GET
POST
PUT
DELETE

Validation
GET
POST
PUT
DELETE

Model
GET
POST
PUT
DELETE

Feature
GET
POST
PUT
DELETE

Compound
GET
POST
PUT
DELETE

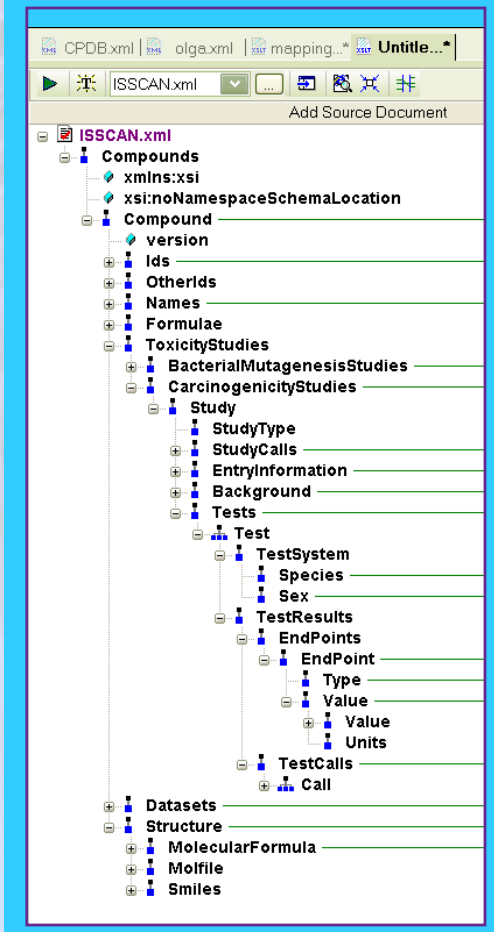
Ontology
GET
POST
PUT
DELETE

Algorithm
GET
POST
PUT
DELETE

OpenTox

Toxicological Endpoint Ontology Development

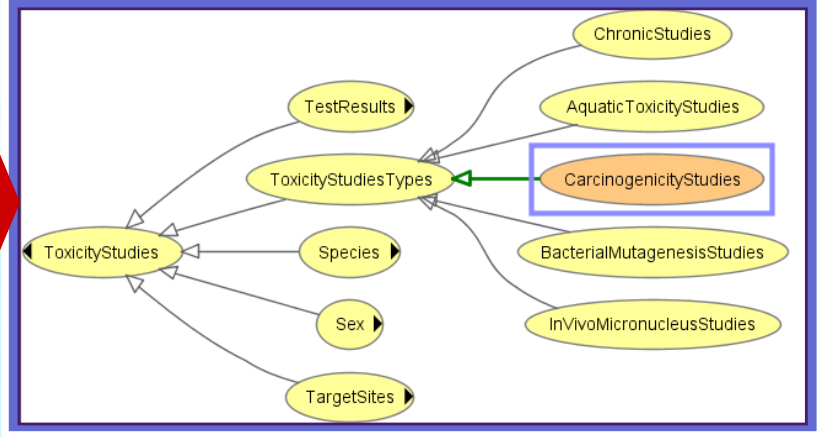
ToxML schema



Other publicly available resources:
DSSTox, GoReni (ITEM), ISSCAN ...

OpenTox Toxicological Endpoint Ontology

Ontology Development



Collaborative Protégé Environment

Re-use of terms defined in neighbouring ontologies (e.g. OBO)

OpenTox

OpenToxipedia

OpenTox

Barry Hardy Log out Quicktools Site Setup Help

Site Map Accessibility Contact Data

Search Site

Home Toxicity Prediction OpenTox Blog People Partners Development OpenToxipedia

User Guidance Latest Entries A B C D E F G H I J K L M N O P Q R S T U V W X Y Z by Categories Entries OpenToxipedia

You are here: Home » OpenToxipedia

Contents View Edit Rules Sharing History

Actions ▾ Display ▾ Add new... ▾ State: Published ▾

OpenToxipedia

by **Barry Hardy** — last modified Sep 03, 2009 01:09 PM

OpenTox Community Resource for Toxicology Vocabulary and Ontology

OpenTox is supporting the creation and curation of OpenToxipedia, a community-based predictive toxicology knowledge resource. All members of the community are welcome to provide entries, suggested definiton edits or additional information to entries in the resource.

OpenTox is supporting the application and development of the **ToxML** standard for representation of toxicology data, the **OECD principles for (Q)SAR model validation**, and the use of the **OECD HT** standard for regulatory reporting purposes.

OpenToxipedia provides here a Vocabulary Resource of toxicology terminology. We hope you find the resource useful and consider contributing to terms and their content.

Guidance for Vocabulary Resource entries

www.opentox.org/opentoxipedia

OpenTox

OpenTox: Databases

Chemical compounds - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://apps.ideaconsult.net:8180/ambit2/query/smarts?type=smiles&search=[*]OC(%3DO)[%236%3BH1]%3D[%236%3BH1]c1ccccc1&t

Disable Cookies CSS Forms Images Information Miscellaneous Outline Resize Tools View Source Options

Chemical compounds

ToxPredict TTC Depiction Datasets Chemical compounds Similarity Substructure Algorithms References Features Templates Models Ontology RDF playground Help

ambit

SMARTS [*]OC(=O)[#6;H1]=[#6;H1]c1ccccc1 Draw substructure

Keywords

Search

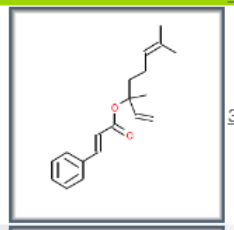
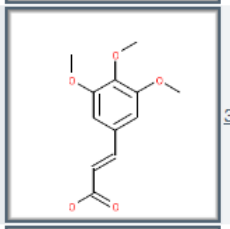
Search for substructure and properties

This site and AMBIT REST services are under development!

Retrieve data

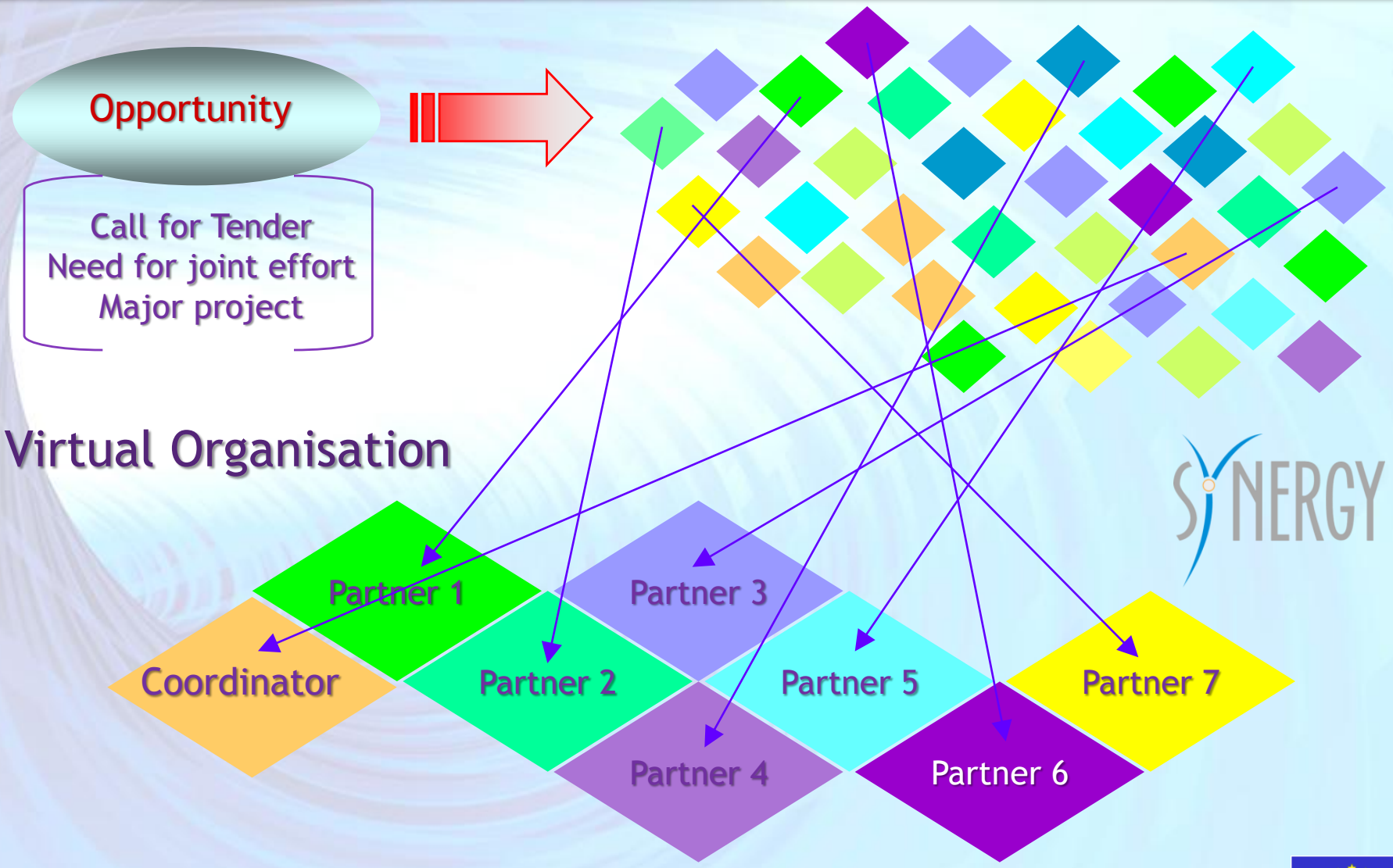
Search results SMARTS [*]OC(=O)[#6; Download as Max number of hits: 100

- Default
- Identifiers
- Datasets
- Models
- Endpoints
- All descriptors
- pKa
- Molecule size
- Electronic descriptors (PM3 optimized structure)
- Electronic descriptors (original structure)
- Toxtree: Cramer rules
- http://apps.idea

#	Compound	ECHA REGISTRATION DATE	ECHA CasRN	ECHA EC	ECHA Names	ECHA SYNON Names	ECHA SYNON Names	ECHA SYNON Names	ECHA SYNON Names	ECHA SYNON Names	ECHA SYNON Names
1		30.11.2010	78-37-5	201-110-3	linalyl cinnamate						
2		30.11.2010	90-50-6	201-999-8	3,4,5-trimethoxycinnamic acid						

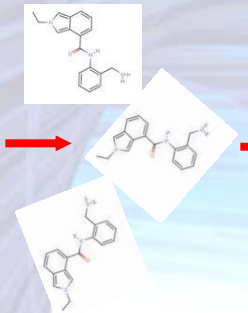
OpenTox

Creation of VO from Collaboration Pool



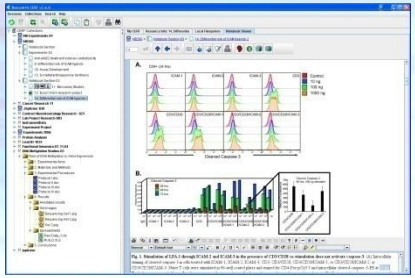
OpenTox

Synergy Drug Design Collaboration Pilot

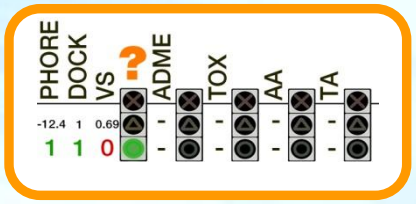


Screened Library

Data

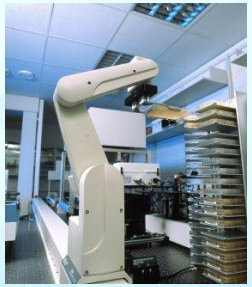


CERF

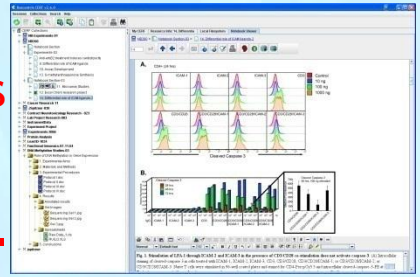


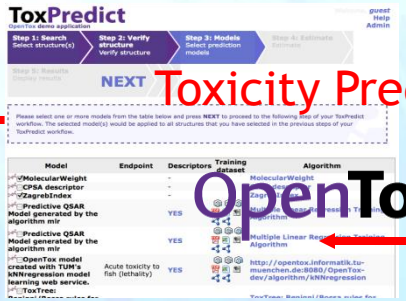
Refine Predictions

SYNERGY



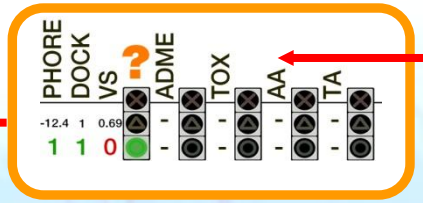
Data



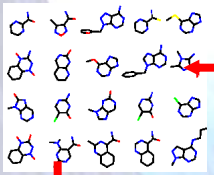


Toxicity Predictions

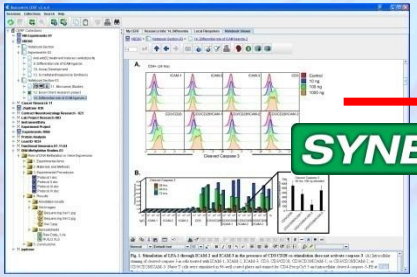
OpenTox



SYNERGY



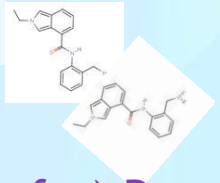
Toxicity Assays



SYNERGY

Decision Dashboard

	PREDICTIVE												
	AFFINITY				ADME								
Compound	PHORE	VS	DOCK	Affinity prediction Traffic Light	ADME	ADME Traffic Light	TOX	Overall Prediction Traffic Light	Affinity Assay	Affinity Assay Traffic Light	Tox Assay	Final Traffic Light
001	1	0	1	"G"	1	"G"	1	"G"	0	"R"	n.d.	"R"
002	0	1	0	"R"	n.d.	"R"	n.d.	"R"	n.d.	"R"	n.d.	"R"
003	1	1	0	"Y"	1	"G"	1	"G"	1	"G"	1	"G"
004	1	1	1	"G"	1	"G"	1	"G"	1	"G"	0	"R"
004	0	1	1	"G"	1	"G"	0	"R"	n.d.	"R"	n.d.	"R"



(Safer) Drug Leads

OpenTox

Recording of Collaborative R&D

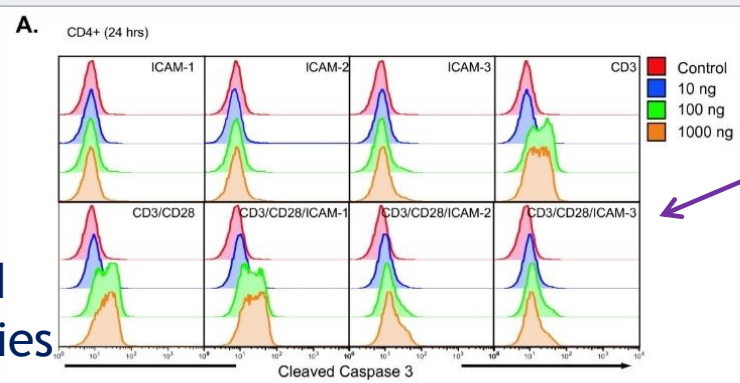
Rescentris CERF v2.6.0

Sessions Collections Search Help

CERF Collections
- NB-Experiments-01
- NB260
  - Notebook Section
  - Experiments-02
    - Anti-erbB2 treatment induces cardiotoxicity
    - 9. Differential role of ICAM ligands
    - 10. Assay Development
    - 13. ß-methylanthraquinone Synthesis
  - Notebook Section-03
    - 11. Microarray Studies
    - 12. Excel Chem research project
    - 14. Differential role of ICAM ligands-2
- Cancer Research 11
- JSpitzner-038
- Contract Neurotoxicology Research - 023
- Lab Project Research-003
- InstrumentData
- Experiment Project
- Experiments 2006
- Protein Analysis
- Lead ID-1034
- Functional Genomics-07-11-04
- DNA Methylation Studies-03
  - Role of DNA Methylation in Gene Expression
    - 1. Experimental Aims
    - 2. Materials and Methods
    - 3. Experimental Procedures
      - Protocol I.doc
      - Protocol II.doc
      - Protocol III.doc
      - Protocol IV.doc
    - 4. Results
      - Annotated results
      - Gel Images
        - Sequencing Gel 1.jpg
        - Sequencing Gel 2.jpg
        - Gel 3.jpg
      - Spreadsheets
        - Raw Data_1.xls
        - RUILI2.XLS
    - 5. Conclusions
- jspitzner

My CERF | Resource Info: 14. Differentia | Local Filesystem | Notebook Viewer

NB260 > Notebook Section-03 > 14. Differential role of ICAM ligands-2



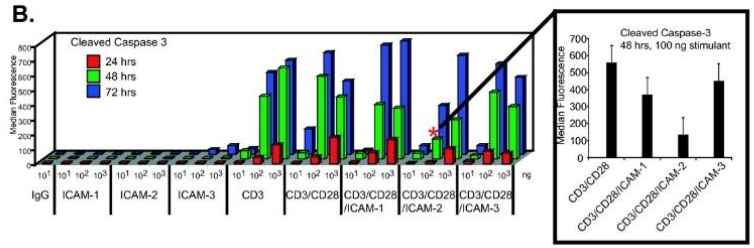


Fig. 1. Stimulation of LFA-1 through ICAM-2 and ICAM-3 in the presence of CD3/CD28 co-stimulation does not activate caspase-3. (A) Intracellular staining of cleaved-caspase 3 in cells treated with ICAM-1, ICAM-2, ICAM-3, CD3, CD3/CD28, CD3/CD28/ICAM-1, or CD3/CD28/ICAM-2, or CD3/CD28/ICAM-3. Naïve T cells were stimulated in 96-well coated plates and stained for CD4-PercpCy5.5 and intracellular cleaved-caspase-3-PE at 24 hrs.

Controlled Vocabularies

Visualisation

Collaborative Electronic Laboratory Notebook (ELN)

OpenTox

OpenTox - Synergy Predictive Toxicology VO Pilot Strategy Development & Case Study

1
- Data Mining of Human Adverse Drug Events
- Data Mining of Literature Knowledge

2
- Creation of Mechanism-based Hypothesis
- Selection of Biological Pathways & Targets

3
- Selection of Compounds
- Prediction of Metabolites of Compounds

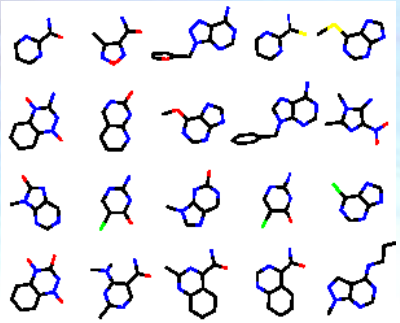
4
- Selection of *in vitro* assays relevant to Mechanism
- Selection and integration of Toxicity Data

5
- Creation of Predictive Toxicology Model including Model Validation and Applicability Domain

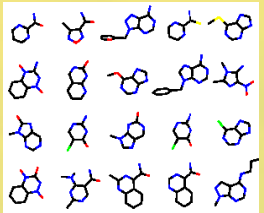
6
- Selection of Low and High Content Assays for Testing in Cell Lines
- Analysis of Results

OpenTox

1. A library of compounds is entered to the ELN



ELN



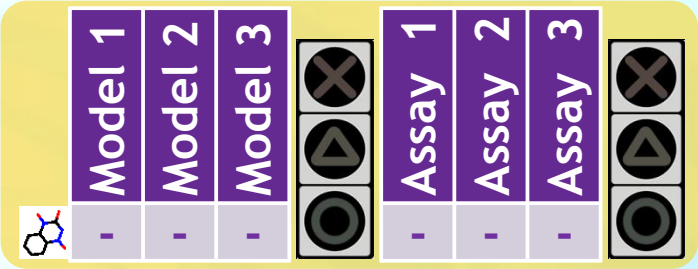
Synergy

OpenTox

OpenTox

2. Each compound is assigned a data structure in ELN

ELN

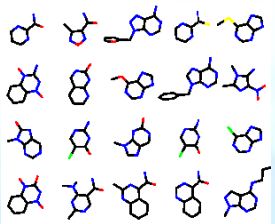


Synergy

OpenTox

OpenTox

3. ELN passes compounds to OpenTox and SYNERGY



ELN

	Model 1	Model 2	Model 3	Assay 1	Assay 2	Assay 3
	-	-	-	-	-	-

Synergy

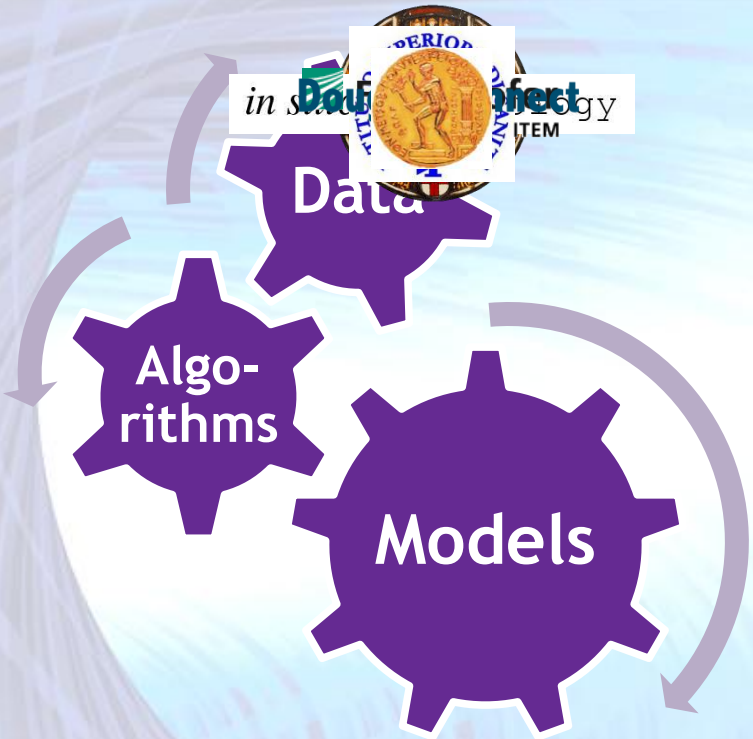
	Model 1	Model 2	Model 3	Assay 1	Assay 2	Assay 3
	-	-	-	-	-	-

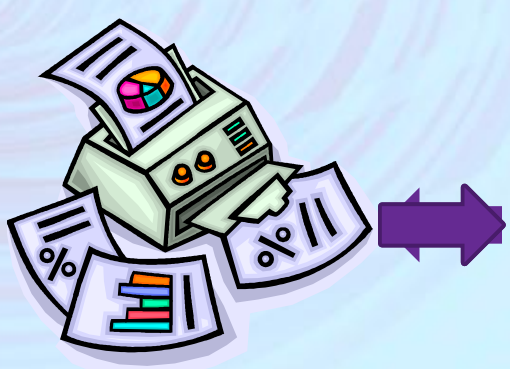
OpenTox

	Model 1	Model 2	Model 3	Assay 1	Assay 2	Assay 3
	-	-	-	-	-	-

OpenTox

4. OpenTox computes toxicity predictions





ELN

	Model 1	Model 2	Model 3		Assay 1	Assay 2	Assay 3	
	-	-	-		-	-	-	

Synergy

	Model 1	Model 2	Model 3		Assay 1	Assay 2	Assay 3	
	-	-	-		-	-	-	

OpenTox

	Model 1	Model 2	Model 3		Assay 1	Assay 2	Assay 3	
	1	0	1		-	-	-	

OpenTox

5. OpenTox sends back a report to ELN

ELN

	Model 1	Model 2	Model 3		Assay 1	Assay 2	Assay 3	
	1	0	1		-	-	-	

Synergy

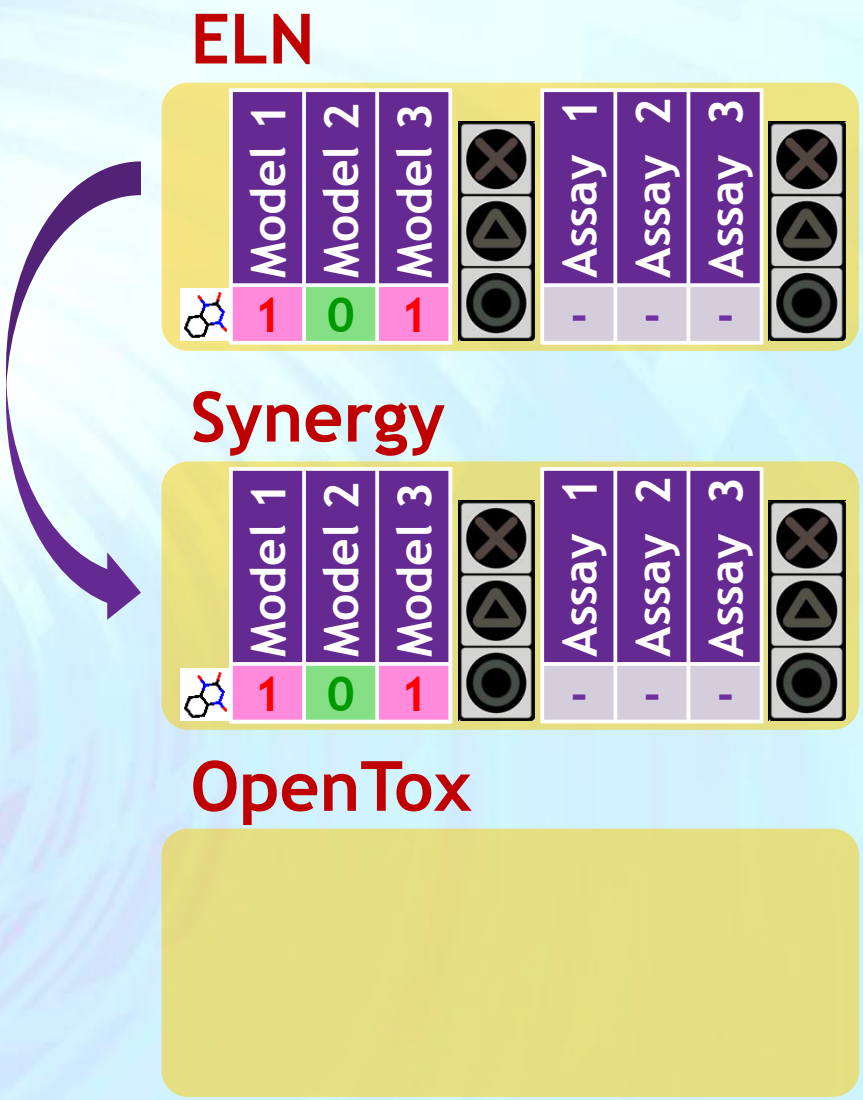
	Model 1	Model 2	Model 3		Assay 1	Assay 2	Assay 3	
	-	-	-		-	-	-	

OpenTox

	Model 1	Model 2	Model 3		Assay 1	Assay 2	Assay 3	
	1	0	1		-	-	-	

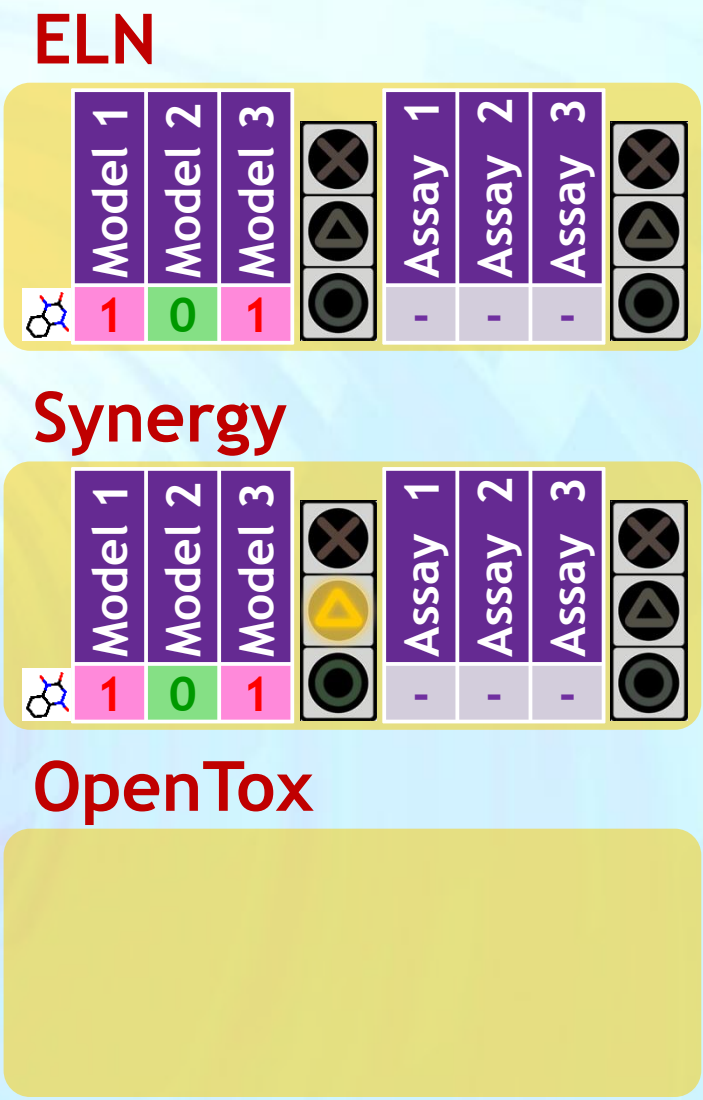
OpenTox

6. ELN sends the results to SYNERGY



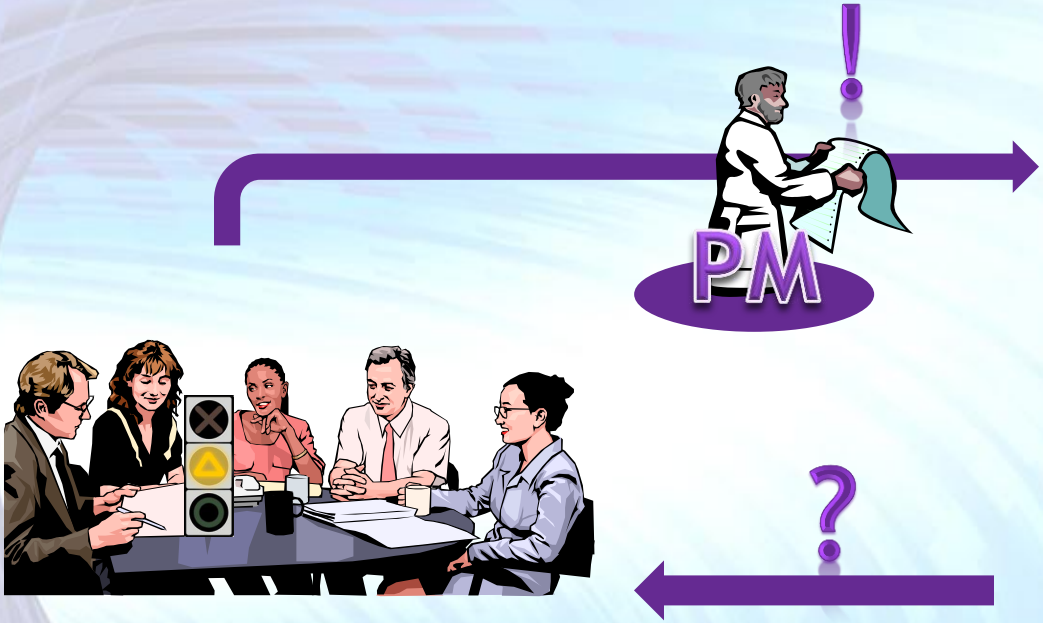
OpenTox

7. SYNERGY applies the Recommendation Rules



OpenTox

8. Inconclusive data → SYNERGY calls a meeting



ELN

Model 1	Model 2	Model 3	Assay 1	Assay 2	Assay 3
1	0	1	-	-	-

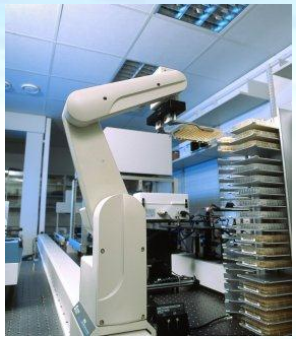
Synergy

Model 1	Model 2	Model 3	Assay 1	Assay 2	Assay 3
1	0	1	-	-	-

OpenTox

OpenTox

9. Experimental assays confirm toxicity



ELN

	Model 1	Model 2	Model 3		Assay 1	Assay 2	Assay 3	
	1	0	1		-	1	1	

Synergy

	Model 1	Model 2	Model 3		Assay 1	Assay 2	Assay 3	
	1	0	1		-	-	-	

OpenTox

Development and Use of Predictive Toxicology Applications

OpenTox Workshop

19 Sept. 2010, Rhodes, Greece

OpenTox Framework Design

Christoph Helma
(in silico toxicology)

OpenTox

Initial Motivation

- Predictive Toxicology applications need common components, e.g.
  - Access to datasets
  - Algorithms for descriptor calculation and model building
  - Validation routines
- These components have to be reimplemented for every new application
- If we had these components readily available we could
  - Quickly build new applications for specific purposes
  - Experiment with new combinations of algorithms
  - Speed up method develoment and testing
  - …

OpenTox

OpenTox Components

Compounds: Structures, names, ...

Features: Chemical and biological (toxicological) properties, substructures, ...

Datasets: Relationships between compounds and features

Algorithms: Instructions for solving problems

Models: Algorithms applied to data yield models which can be used for predictions

Validation: Methods for estimating the accuracy of model predictions

Reports: Report predictions and models e.g. to regulatory authorities

Tasks: Handle long running calculations

Authentification and Authorisation: Protect confidential data

Requirements

- Platform independence
- Interoperability for communication with external programs and data sources
- Transparency for scientific and regulatory credibility
- Open for future extensions

OpenTox

Technological Choices

- Webservices
- Comunication through well defined interfaces
- Ontologies for the exchange of knowledge and data
- Use and promote open standards
- Open source components

OpenTox

Representational State Transfer (REST)

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 and local systems**
- Language independent

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

OpenTox

Ontologies

What?

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

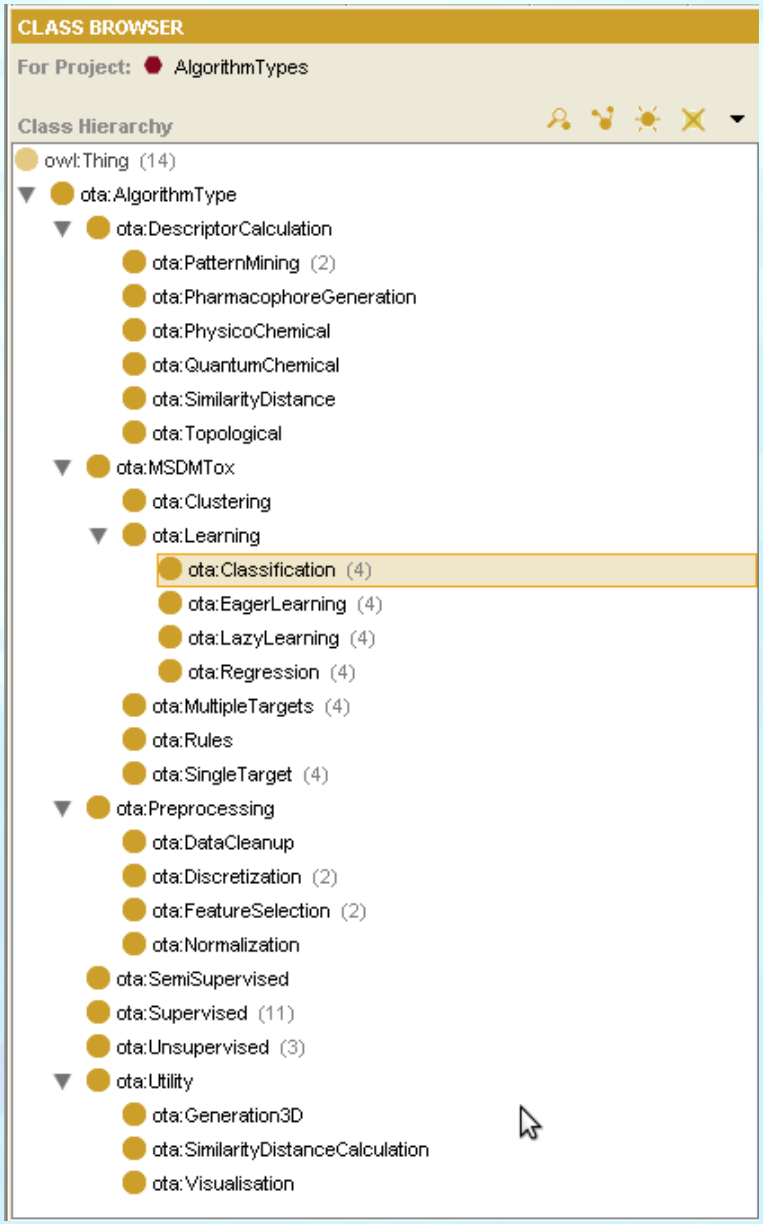
Why?

- Distributed services **need** to be able to "talk to each other", e.g. have a **common understanding** of endpoints, properties, methods, etc.
- Allows us to integrate existing knowledge from many related domains

Ontologies

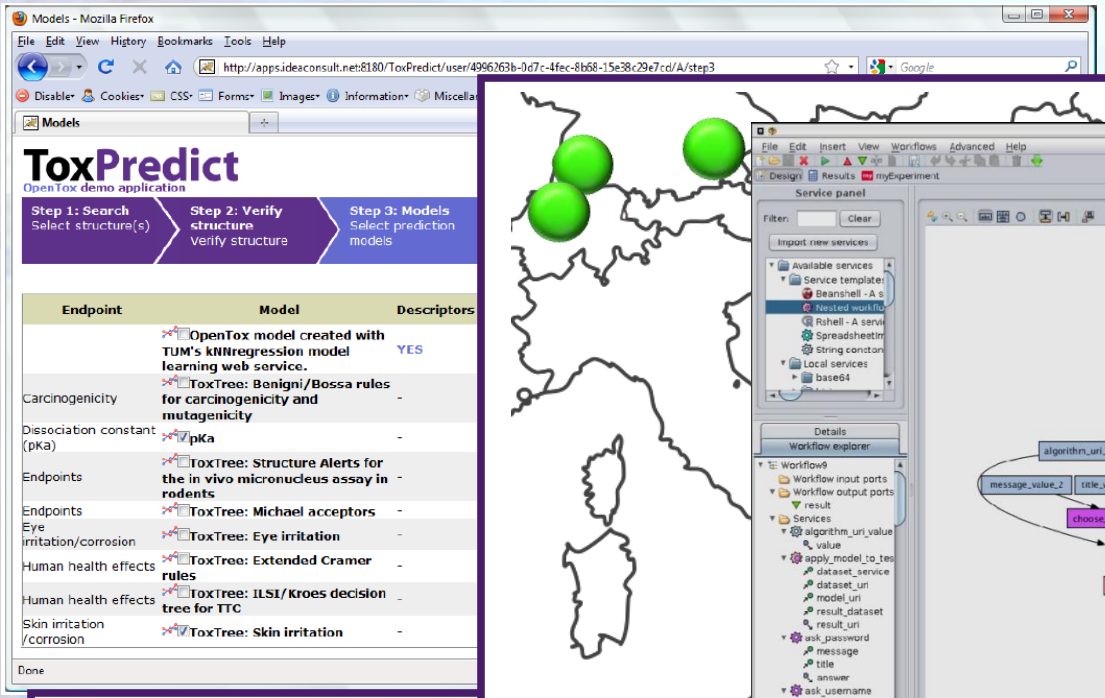
- Standards: **OWL-DL** as representation language and **SPARQL** as query language
- There are many ongoing biological ontology projects
- Our strategy: use existing work and standards wherever possible
- However, there are new ontology needs for OpenTox applications, e.g. for algorithms, toxicological endpoints

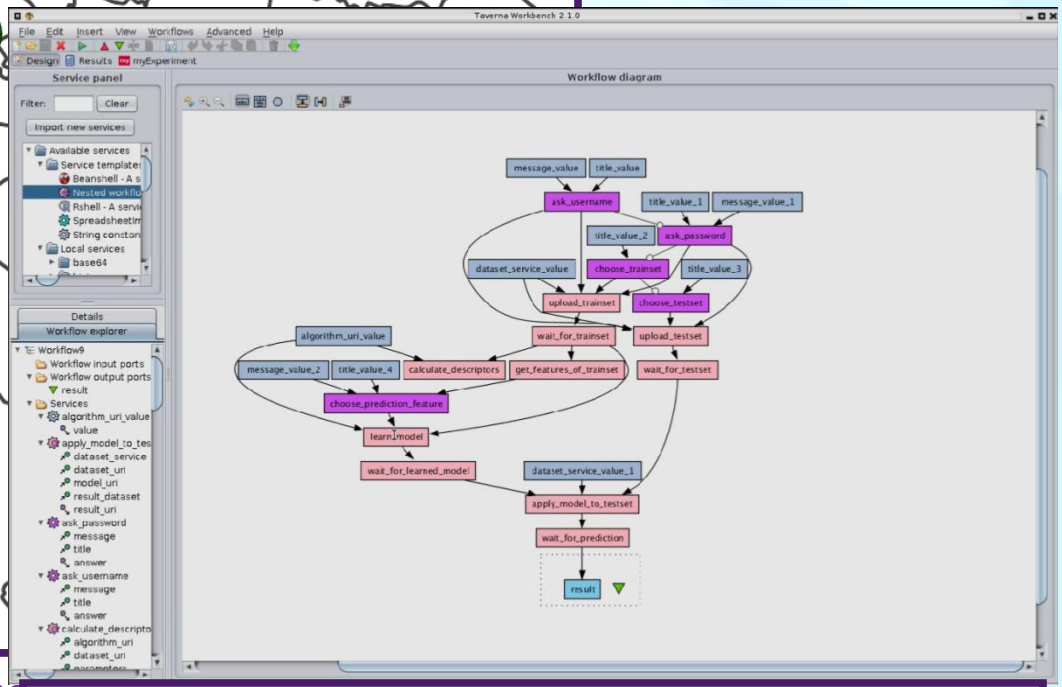
OpenTox
Ontology Working Group



OpenTox

What can you do with OpenTox





Simple building of pred
applications based
methods and datab

Distributed applica
wide range of data
methods

Integration into workflow systems for computational biology

OpenTox

Summary

- OpenTox is a framework for predictive toxicology
- Designed for language independence, interoperability, transparancy and extensibility
- Implemented as open source REST webservices
- Exchange of data and knowledge with ontologies (OWL-DL)
- OpenTox components: Compound, Feature, Dataset, Algorithm, Model, Validation, Report, Task, Authentification and Authorisation
- Documentation: www.opentox.org/dev

OpenTox

Development and Use of Predictive Toxicology Applications

An OpenTox Workshop

19 Sep 2010, Rhodes, Greece

Application Programming Interfaces

presented by Nina Jeliazkova
(Ideaconsult Ltd., Bulgaria)

OpenTox

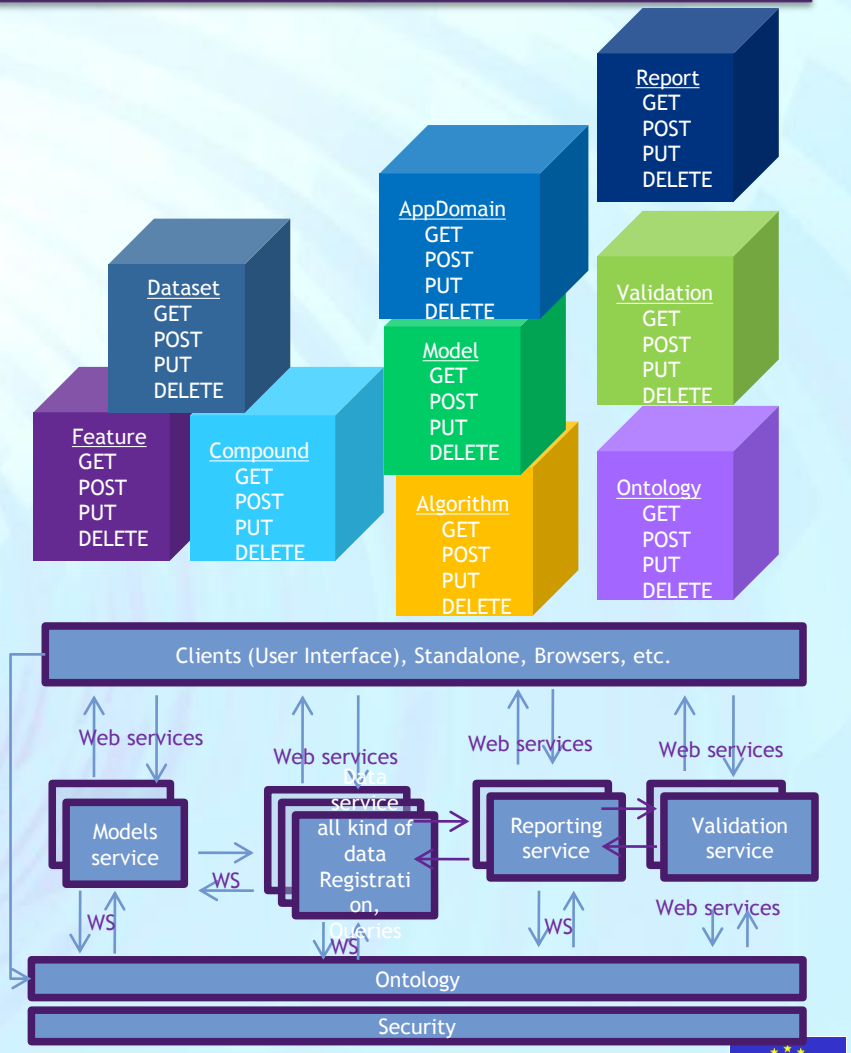
SEVENTH FRAMEWORK PROGRAMME

Framework design rationales

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*

The framework

- OpenTox API
  - The way applications talk to each other
  - The way developers talk to applications
  - http://opentox.org/dev/apis/api-1.1
- The basic building blocks:
  - data, chemical structures, algorithms and models.
- Functionality offered
  - build models,
  - apply models,
  - validate models,
  - access and query data in various ways.
- Technologies
  - REST style web services
  - RDF for description of resources
  - Links to existing and newly developed ontologies (mainly to describe metadata) about resources



Representational State Transfer (REST)

A software architecture style, defined by Roy Fielding in his PhD thesis (2000). Many services worldwide offer REST API. There are (currently) **no standards** for RESTful applications, but merely **design guides.**

Design principles:

- Resource oriented
  - Every object (resource) is named and addressable (e.g. HTTP URL) Example: **http://example.opentox.com/model/myBestModel** , **http://example.opentox.com/compound/50-00-0**
  - RESTfull API design starts by identifying most important objects and groups of objects, supported by the software system and proceeds by defining URL patterns.
- Transport protocol
  - HTTP is the most popular choice of transport protocol, but other protocols can be used as well
- Operations
  - All resources (nouns) support the same fixed and universal number of operations (verbs). HTTP (GET, POST, PUT, DELETE) operations are the common choice, when the transport protocol is HTTP.
- Hypermedia as the Engine of Application State
  - All resources should be reachable via a single (or minimum) number of entry points into RESTful applications. Thus, a representation of a resource should return hypermedia links to related resources
- Error codes (for each resource/operation pair)
  - HTTP status codes (e.g. 200 OK, 400 Bad Request, 404 Not found, etc.) are usually used

OpenTox resources (1)

OpenTox considers the following set of entities as essential building blocks:

- Structures of **chemical compounds**
- **Properties and identifiers** of chemical compounds
- **Datasets** of chemical compounds and various properties (measured or calculated)
- **Algorithms**
  - Data processing algorithms
    - Algorithms generating certain values, based on chemical structure (e.g. descriptor calculation)
    - Data preprocessing (e.g. Principal component analysis, feature selection)
    - Structure processing (e.g. structure optimization)
    - Algorithms, relating set of structures to another set of structures (e.g. similarity search or metabolite generation)
  - Machine learning algorithms
    - Supervised (e.g. Regression, Classification)
    - Unsupervised (e.g. Clustering)
  - Prediction algorithms, defined by experts (e.g. series of structural alerts, defined by human experts , not derived by learning algorithms)

OpenTox resources (2)

- **Models** are generated by respective algorithms, given specific parameters
  - Statistical models are generated by applying statistical/machine learning algorithms to specific dataset and parameters
  - Models can be other than statistical, e.g. expert defined rules, quantum mechanical calculations, metabolite generation, etc. The intention of the framework is to be generic enough to accommodate varieties of predictive models.
- **Validation** provides procedures independent of model building facilities (e.g. crossvalidation) and generates relevant statistics.
- **Reports**
  - Various types of reports might be generated, using building blocks above (e.g. validation report can be generated using validation object, a model and a dataset).
- In addition, the following components are introduced:
  - **Task** (asynchronous processing of computationally intensive tasks)
  - **Authentication and authorization** (Ensuring secure access to sensitive resources)
  - **Ontology service** (provides an RDF storage and SPARQL endpoint for resources registration)

Resources identification

All resources are identified via unique web address, assigned according to the URL templates

Component	Description	URL Template (example)
Compound	Representations of chemical compounds	http://host:port/compound/{compoundid}
Feature	Properties and identifiers	http://host:port/feature/{featureid}
Dataset	Encapsulates set of chemical compounds and their property values	http://host:port/dataset/{datasetid}
Model	OpenTox model services	http://host:port/model/{modeld}
Algorithm	OpenTox algorithm services	http://host:port/algorithm/{algorithmid}
Validation, Report	A validation corresponds to the validation of a model on a test dataset.	http://host:port/validation/{validationid} http://host:port/report/{reportid}
Task	Asynchronous jobs are handled via an intermediate Task resource. A resource, submitting an asynchronous job should return the URI of the task.	http://host:port/task/{taskid}
Ontology service	Provides storage and SPARQL search functionality for objects, defined in OpenTox services and relevant ontologies	http://host:port/ontology
Authentication and authorisation	Granting access to protected resources for authorised users	http://host:port/opensso http://host:port/opensso-pol

OpenTox REST operations

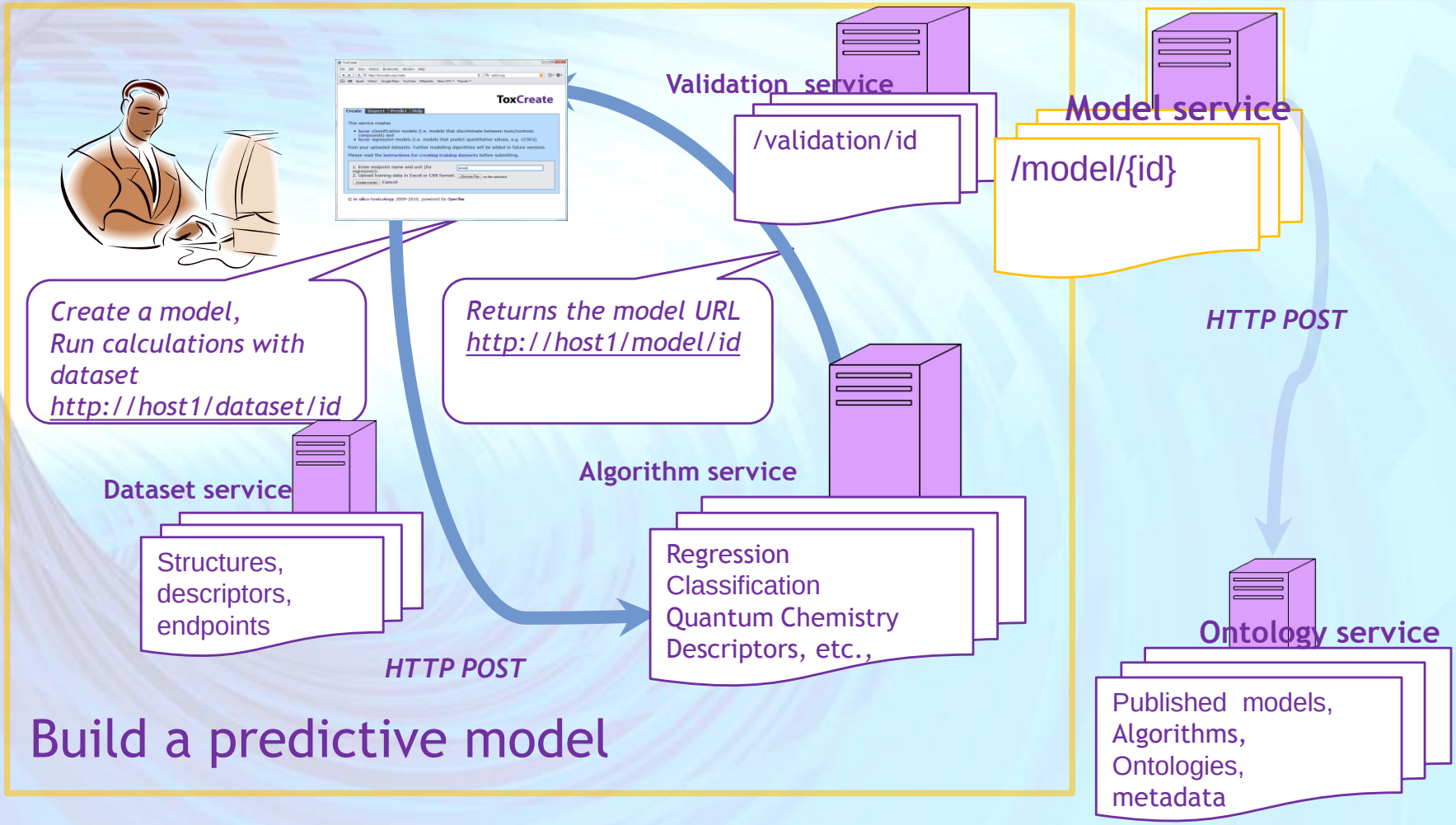
Individual resources (e.g. a dataset or a model)

- URI template http://host:port/{resource}/{resourceid} , e.g. http://host:port/model/{model_id} or http://host:port/dataset/{dataset_id}
- GET – retrieve representation of the resource
- PUT – update representation of the resource
- POST :
  - replace representation of the resource with a new one (e.g. replace the dataset with new content)
  - initiate calculations, based on this resource (e.g. submit dataset URI to an algorithm resource and obtain a model URI as a result)
- DELETE – delete the resource

Collections of resources (e.g. list of all available models, or datasets)

- URI template http://host:port/{resource} , (e.g. http://host:port/model or http://host:port/dataset)
- GET – retrieve representation of multiple resources (e.g. retrieve all available algorithms)
- PUT - N/A
- POST – create new resource and return its URI (e.g. create a new dataset by submitting new dataset content to the dataset service)
- DELETE – N/A

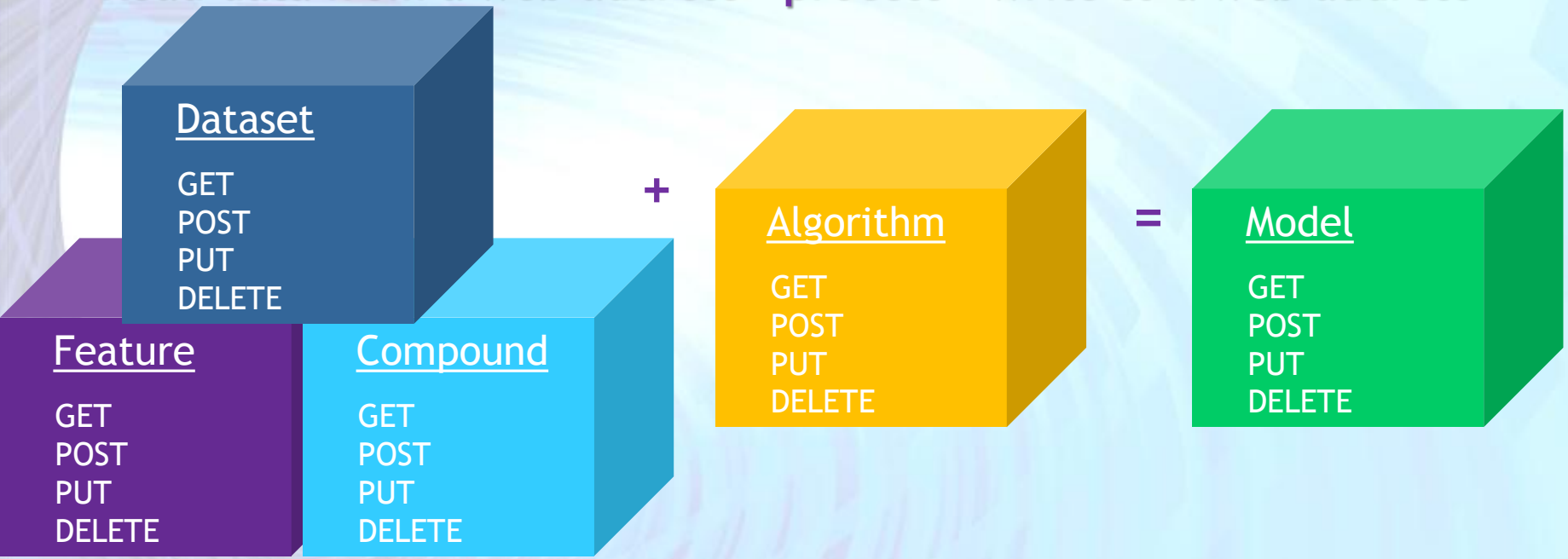
Build a predictive model



OpenTox

Uniform approach to models creation

Read data from a web address - process - write to a web address



http://myhost.com/algorithm/neuralnetwork

http://myhost.com/dataset/trainingset1

http://myhost.com/model/predictivemodel1

Use an algorithm to build a model

- An algorithm is applied by submitting HTTP POST to the algorithm URI and providing required parameters.
- A common required parameter is **dataset_uri=http://host:port/dataset/{datasetid}**, which specifies the data set to be operated on.
- HTTP POST in REST style services returns URI of the result, and not the content of the result.
- The algorithm services are designed to store the results into a dataset service and return the URL of the resulted dataset.
- In case of slow calculations a Task URI, instead of the dataset URI is returned

```
$ curl -H "Accept:text/uri-list" -X POST -d
'dataset_uri=http://apps.ideaconsult.net:8080/ambit2/dataset/1037' -d
'prediction_feature=http://apps.ideaconsult.net:8080/ambit2/feature/26
701' -d
'dataset_service=http://apps.ideaconsult.net:8080/ambit2/dataset'
http://opentox.informatik.tu-muenchen.de:8080/OpenTox-
dev/algorithm/J48  -iv
* Connected to opentox.informatik.tu-muenchen.de (131.159.28.16) port
8080 (#0)
➢POST /OpenTox-dev/algorithm/J48 HTTP/1.1
➢> Host: opentox.informatik.tu-muenchen.de:8080
➢Accept: */*
> Content-Type: application/x-www-form-urlencoded
< HTTP/1.1 202 Accepted
< Date: Sat, 31 Jul 2010 14:46:38 GMT
< Location: http://opentox.informatik.tu-muenchen.de:8080/OpenTox-
dev/task/acdf6eac-d5a2-402c-a4e2-06cd7e3ca1b5
< Accept-Ranges: bytes
< Server: Noelios-Restlet-Engine/1.1.snapshot
< Content-Type: text/uri-list;charset=ISO-8859-1
< Content-Length: 99
<
* Connection #0 to host opentox.informatik.tu-muenchen.de left intact
* Closing connection #0
http://opentox.informatik.tu-muenchen.de:8080/OpenTox-
dev/task/acdf6eac-d5a2-402c-a4e2-06cd7e3ca1b5
```

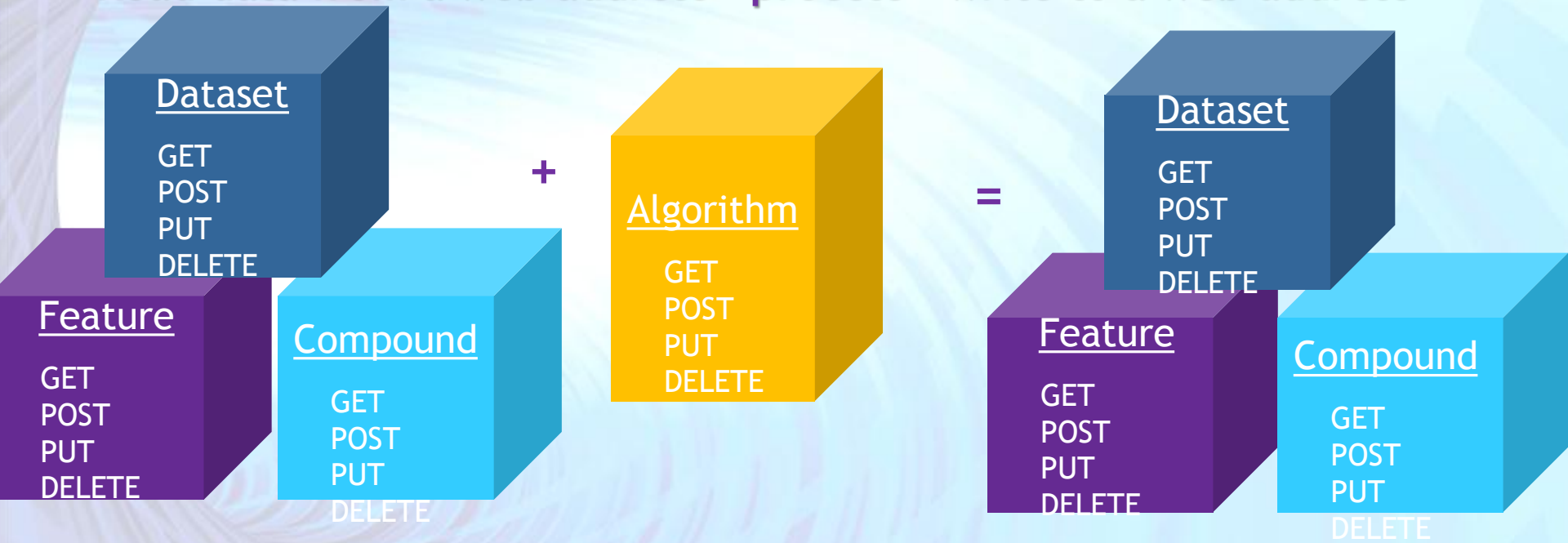
Resources: The model

- When task URI is returned , the returned status code is HTTP 202 Accepted, instead of HTTP 200 OK.
- This tells the client the processing is not completed and the client need to poll the task URI until OK code is returned
- The final result, returned by Example 25 is the URI of the new model http://opentox.informatik.tu-muenchen.de:8080/OpenTox-dev/model/TUMOpenToxModel_j48_48 .
- To obtain prediction results POST a dataset to the model URI

```
$ curl -iv -H "Accept:text/uri-list" http://opentox.informatik.tu-muenchen.de:8080/OpenTox-dev/task/acdf6eac-d5a2-402c
* About to connect() to opentox.informatik.tu-muenchen.de port 8080 (#0)
*   Trying 131.159.28.16... connected
* Connected to opentox.informatik.tu-muenchen.de (131.159.28.16) port 8080 (#0)
> GET /OpenTox-dev/task/acdf6eac-d5a2-402c-a4e2-06cd7e3ca1b5 HTTP/1.1
> User-Agent: curl/7.18.2 (x86_64-pc-linux-gnu) libcurl/7.18.2 OpenSSL/0.9.8g
zlib/1.2.3.3 libidn/1.8 libssh2/0.18
> Host: opentox.informatik.tu-muenchen.de:8080
> Accept:text/uri-list
>
< HTTP/1.1 200 OK
< Date: Sat, 31 Jul 2010 14:47:22 GMT
Date: Sat, 31 Jul 2010 14:47:22 GMT
< Location: http://opentox.informatik.tu-muenchen.de:8080/OpenTox-dev/model/TUMOpenToxModel_j48_48
< Vary: Accept-Charset, Accept-Encoding, Accept-Language, Accept
< Accept-Ranges: bytes
< Server: Noelios-Restlet-Engine/1.1.snapshot
< Content-Type: text/uri-list;charset=ISO-8859-1
< Content-Length: 86
<
* Connection #0 to host opentox.informatik.tu-muenchen.de left intact
* Closing connection #0
http://opentox.informatik.tu-muenchen.de:8080/OpenTox-dev/model/TUMOpenToxModel_j48_48
```

Uniform approach to data processing (e.g. Descriptors calculation)

Read data from a web address - process - write to a web address



http://myhost.com/algorithm/{descriptorX}

http://myhost.com/dataset/trainingset1

http://myhost.com/dataset/results

Uniform approach to models validation and report generation

Read data from a web address - process - write to a web address

http://myhost.com/dataset/trainingset1

http://myhost.com/dataset/predictedresults1

Validation report

http://myhost.com/validation

Validation

GET
POST
PUT
DELETE

+

Dataset

GET
POST
PUT
DELETE

=

Report

GET
POST
PUT
DELETE

Model generating predictions

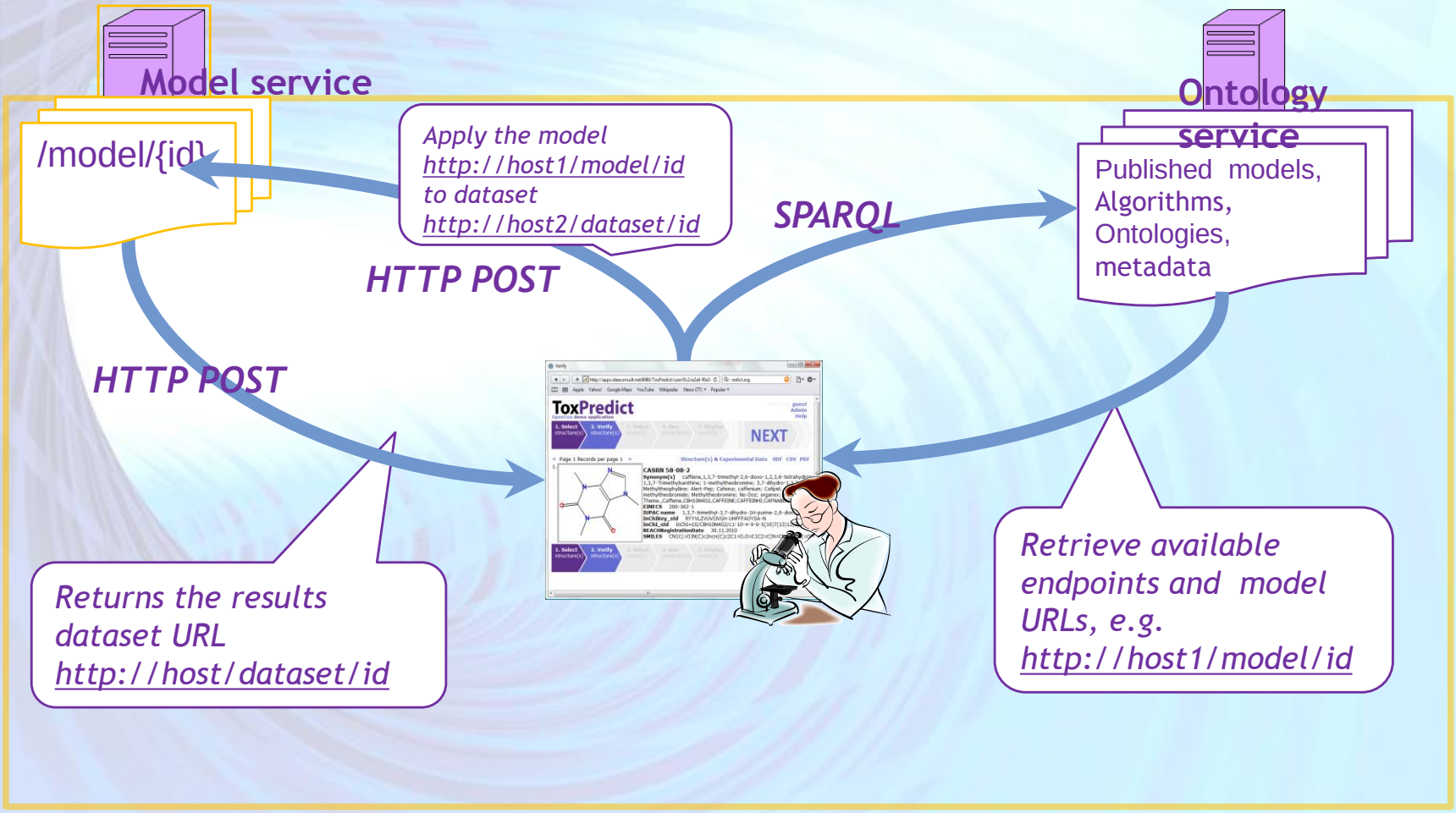
Model

GET
POST
PUT
DELETE

http://myhost.com/report/1

http://myhost.com/model/predictivemodel1

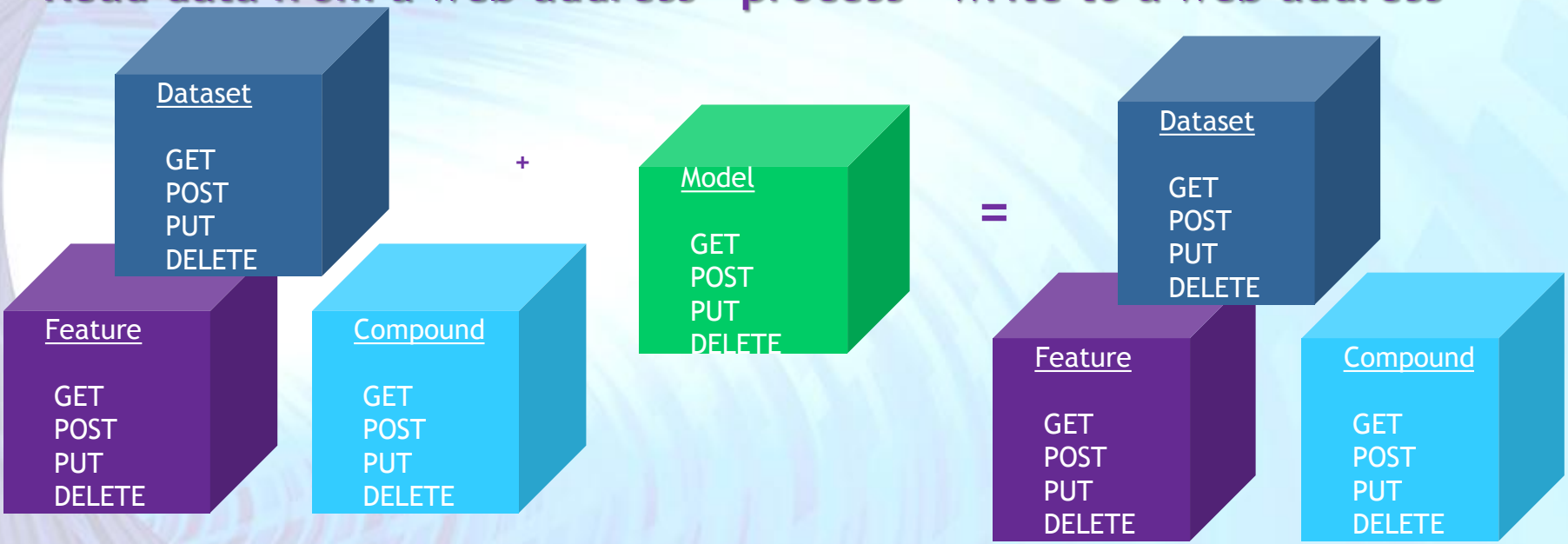
Apply predictive models



OpenTox

Uniform approach to model prediction

Read data from a web address - process - write to a web address



http://myhost.com/model/predictivemodel1

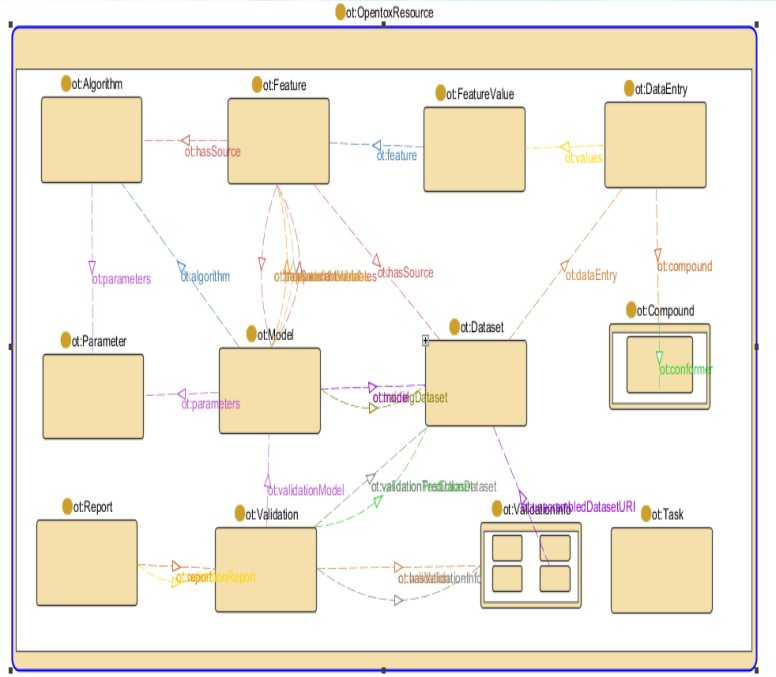
http://myhost.com/dataset/id1

http://myhost.com/dataset/results1

RDF - Resources representation

- The opentox.owl ontology
  - A common OWL data model of all OpenTox resources
  - Describes OpenTox resources
  - Describes relationships between them
  - Generates object's RDF representations.
- RDF/XML representation is mandatory for OpenTox resources.
- Uniform approach to data representation
  - Calculated and measured properties of chemical compounds are represented in an uniform way
  - Linked to the resource used for data generation
  - Annotated via ontology entries
  - Model representations link to algorithms and data used

All OpenTox components are defined by OWL ontology
http://opentox.org/api/1.1/opentox.owl
All resources are subclasses of ot:OpenToxResource



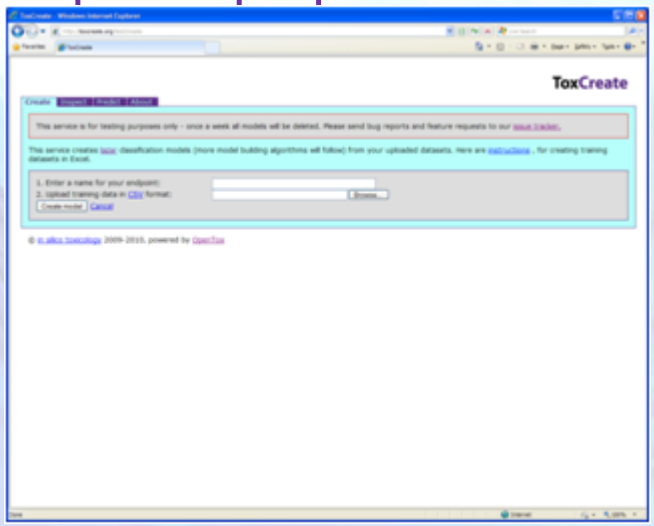
Services implementation by partner and service type

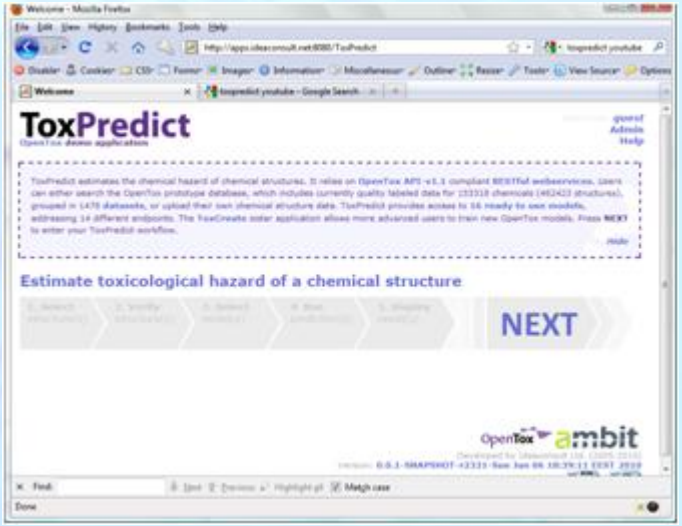
All components are implemented as REST web services.
There could be multiple implementations of same type of components.
(Subset of) services could be hosted by the same provider, or by multiple providers on separate locations.

Partner No. /Service type	Compound	Dataset	Feature	Algorithm (processing)	Algorithm (model)	Model	Validation	Report	Task	Autherntication and Authorisation service	Ontology service
2	Y	Y		Y	Y	Y			Y		
3	Y	Y	Y	Y	Y	Y			Y		Y
5				Y	Y	Y					
6							Y	Y	Y	Y	
7					Y	Y			Y		
10						Y					

Demo applications

- Two end user oriented demo applications, making use of OpenTox webservices, have been developed, deployed and are available for testing – **http://toxcreate.org** and **http://toxpredict.org** ;
- ToxCreate creates models from user supplied datasets;
- ToxPredict uses existing OpenTox models to estimate chemical compound properties





ToxPredict: a detailed case study

- 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 datasets, or upload their own chemical structure data. ToxPredict provides access to **14 ready to use models**, addressing **13 different endpoints**
- ToxPredict uses the following OpenTox webservices: Compound, Feature, Dataset, Algorithm, Model , Task and Ontology;
- more details on its interactions with OpenTox webservices which are taking place behind the scenes and without requiring any end-user intervention;

ToxPredict: Step 1 (Select structure(s))

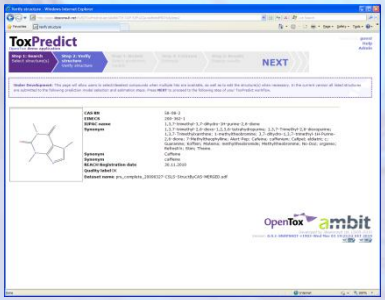
Find structure by name, registry number, SMILES, InChI, structure, substructure, similarity…

ToxPredict Web Application

OT Dataset API ***HTTP GET***

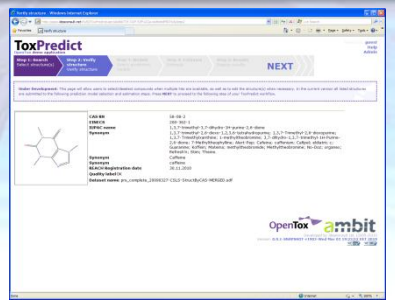
OT Dataset Service

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 2 (Verify structure(s))



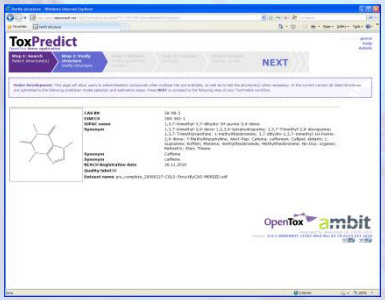
Select and/or edit structure(s)

ToxPredict Web Application

OT Dataset API ***HTTP GET***
(***HTTP POST*** for structure editing)

OT Dataset Service

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



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

ToxPredict: Step 3 (Select model(s))

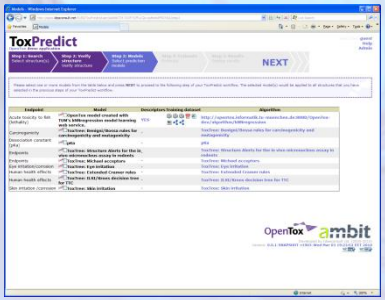
What prediction models are available? Is there a model for endpoint X?

ToxPredict Web Application

HTTP GET SPARQL query

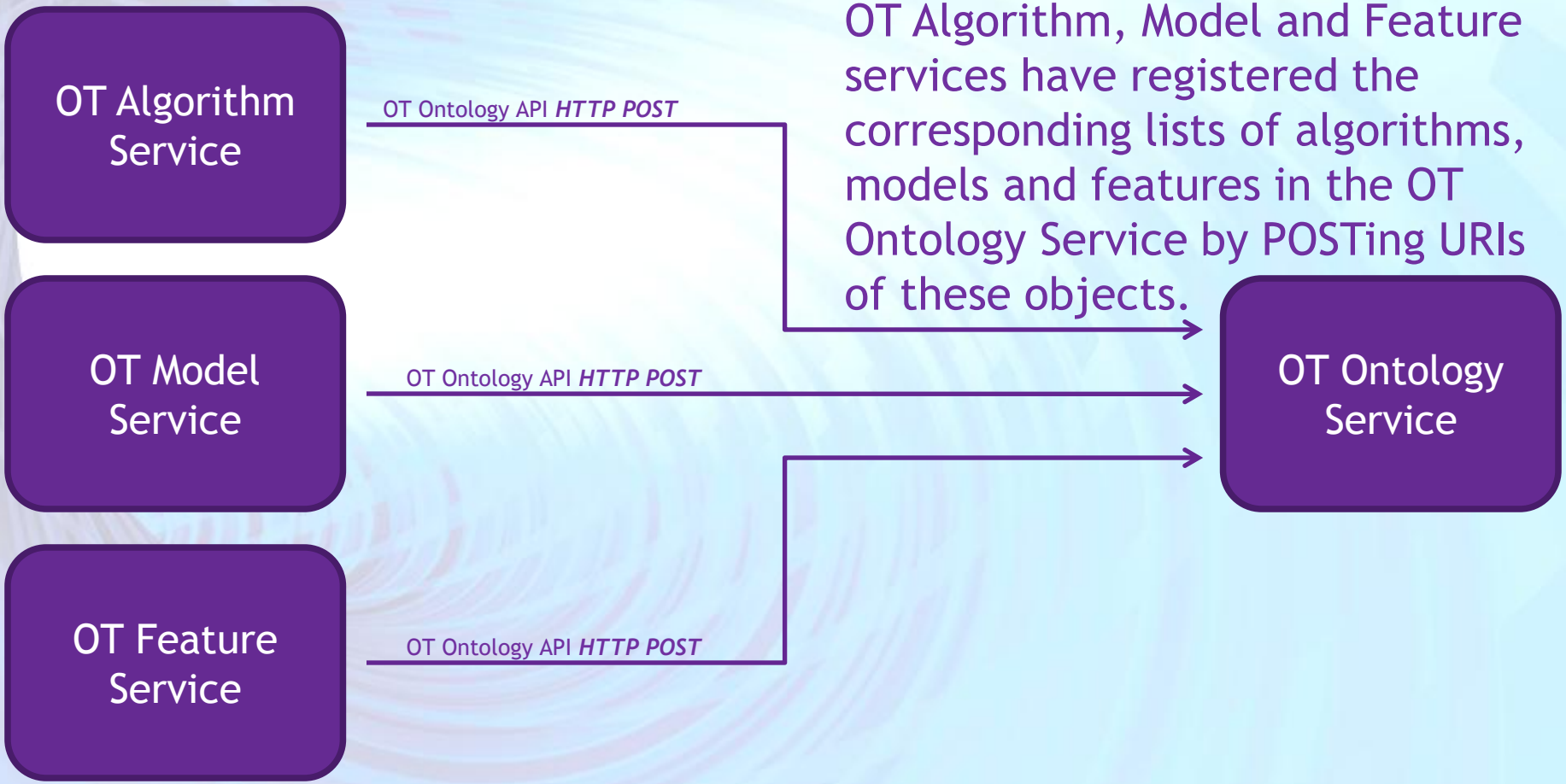
OT Ontology Service

application/sparql-results+xml



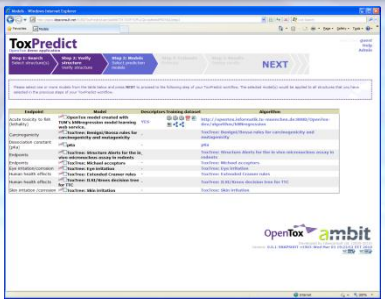
Here is the list of model URIs and related endpoints and algorithms in SPARQL format.

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)



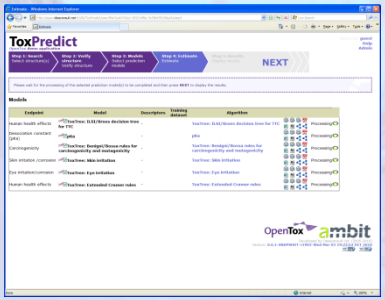
Run the selected models.

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

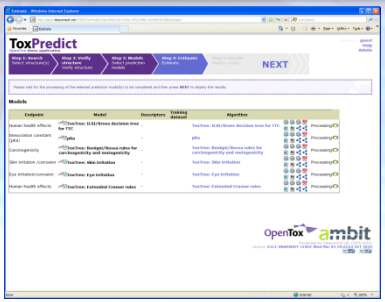


The calculation will take a while; here is a task URI, which can be queried for processing status.

Create a new task.

OT Task Service

ToxPredict: Step 4 (Estimate)



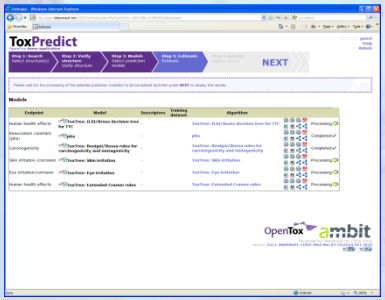
Is the task completed?

ToxPredict Web Application

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

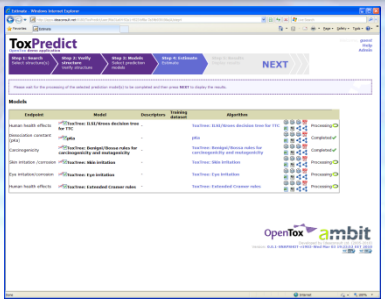
OT Task Service

HTTP code 202 "Accepted for processing"
Returns Task URI



Not yet, but processing has finished and the results have been posted to the OT Dataset service; here is a task URI for the dataset import.

ToxPredict: Step 4 (Estimate)



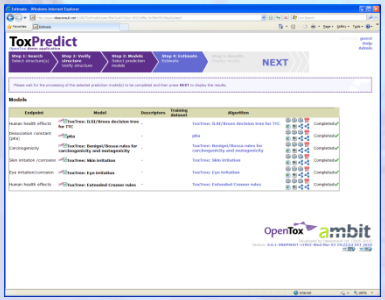
Is the task completed?

ToxPredict Web Application

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

OT Task Service

HTTP code 200 "OK"
Returns Dataset URI



Yes, here is the dataset URI of the results.

ToxPredict: Step 4 (behind the scenes)

Does the model make use of descriptors?

ToxPredict Web Application

OT Ontology API ***HTTP GET SPARQL*** query

OT Ontology Service

application/sparql-results+xml

Yes, here is a list of features and algorithms, used to calculate descriptors for the training set.

ToxPredict: Step 4 (behind the scenes)

Are the descriptor values for query compounds already calculated and available in the OT database?

ToxPredict Web Application

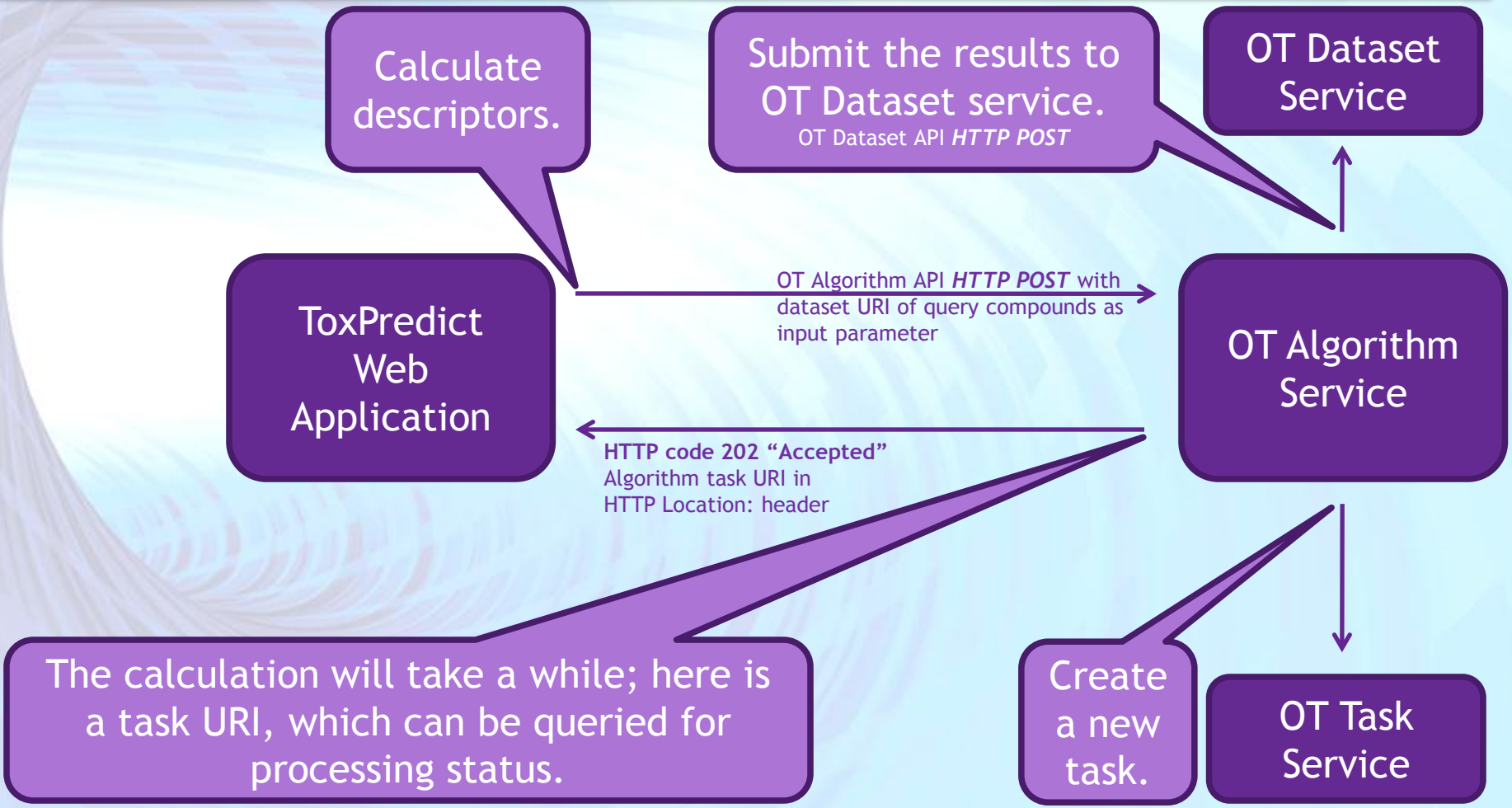
OT Dataset API ***HTTP GET***

OT Dataset Service

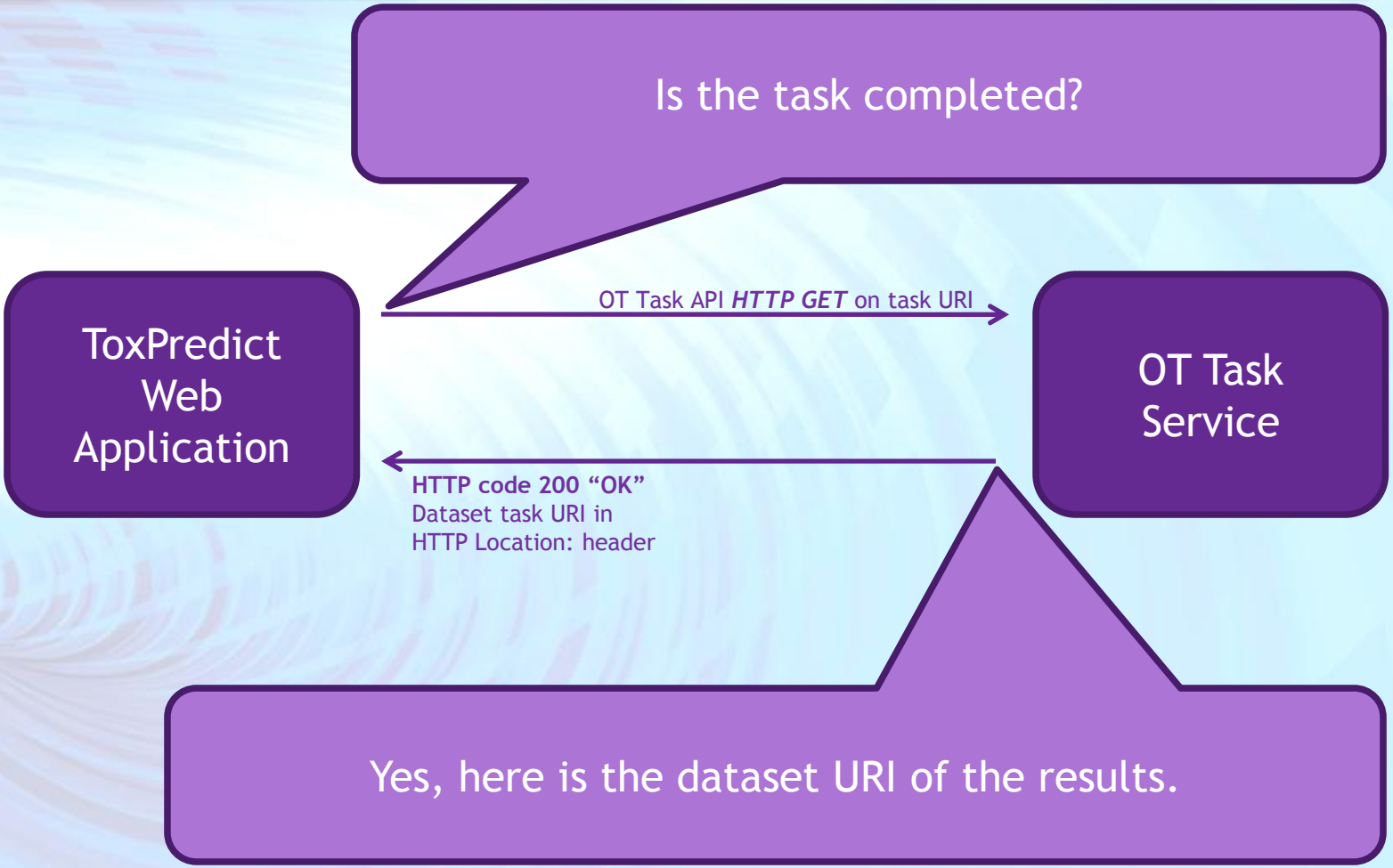
HTTP code 404 "Not Found"

No, the descriptor values for query compounds are not available in the OT database.

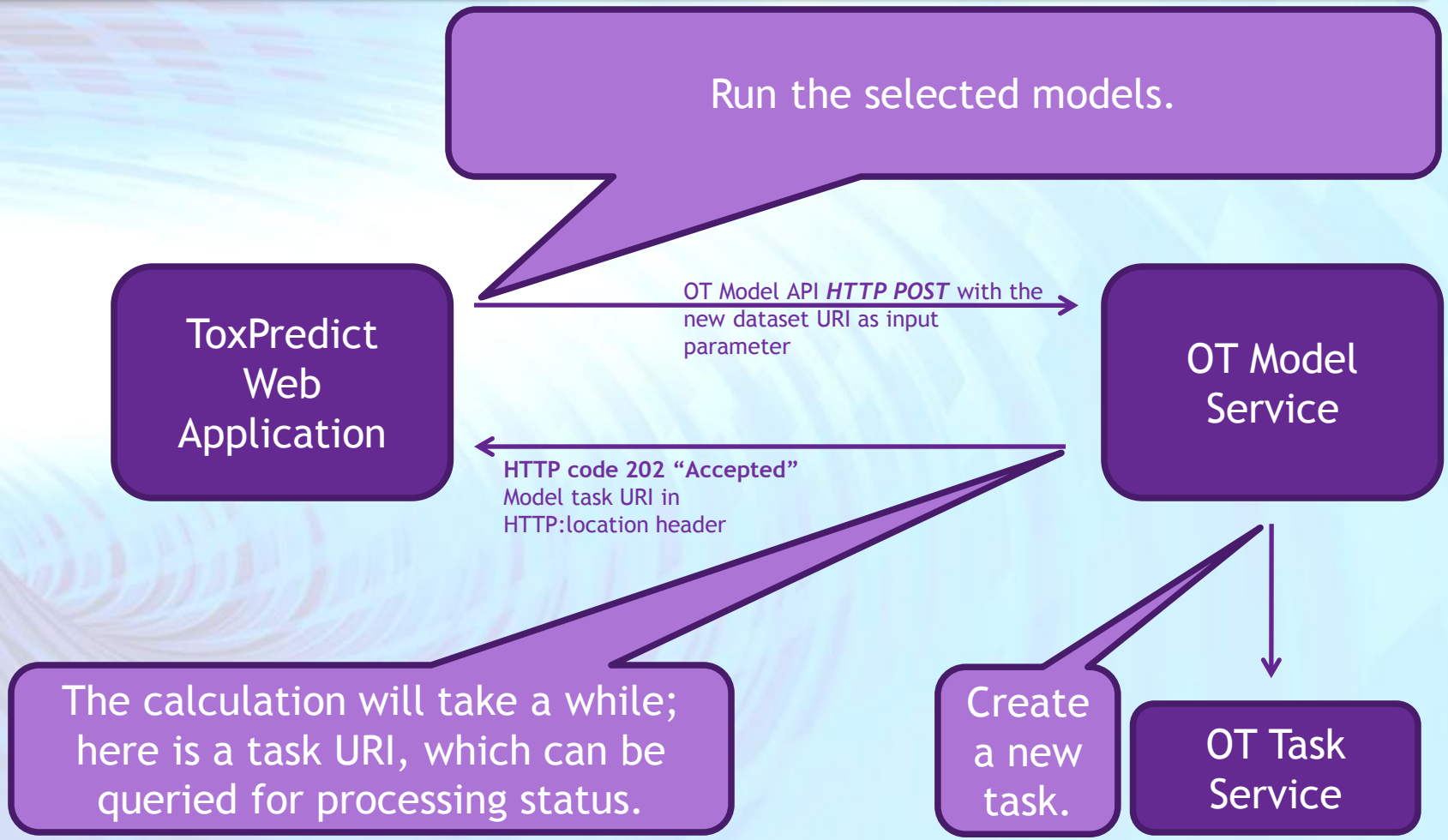
ToxPredict: Step 4 (behind the scenes)



ToxPredict: Step 4 (behind the scenes)

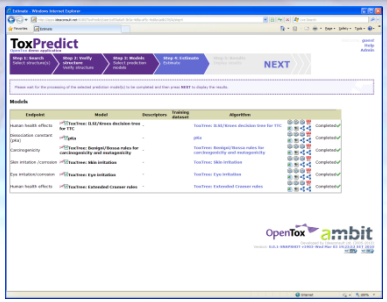


ToxPredict: Step 4 (behind the scenes)



… the remaining part of the Estimation step has been described

ToxPredict: Step 5 (Display results)



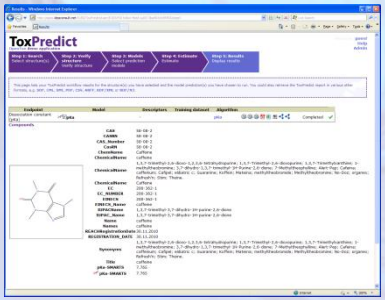
Retrieve calculation results from the final dataset URI, obtained in Step 4 (Estimation).

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.

Development and Use of Predictive Toxicology Applications

An OpenTox Workshop

19 Sep 2010, Rhodes, Greece

Ontology and Data Schema

Olga Tcheremenskaia

Romualdo Benigni

(Istituto Superiore di Sanità, Rome, Italy)

OpenTox

SEVENTH FRAMEWORK PROGRAMME

Ontology: presentation plan

Introduction, Tools, Principles

- What is an ontology?
- OWL and OBO languages, Protégé Editor
- Biomedical ontologies
- OBO Foundry Principles

Incorporation of Ontologies into OpenTox

- Why we need an ontology for the OpenTox?
- Principles of the OpenTox toxicological endpoint ontology development

ToxML data standard

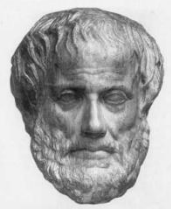
- ToxML data standard
- Toxicological Databases mapping onto ToxML scheme
- ToxML in ontology development

Ontology Development

- Collaborative Protégé Server
- Toxicological Endpoints Ontology structure
- Organs ontology
- OpenToxipedia: community knowledge resource

OpenTox

What Is An Ontology?
Ontologies in Information Science



- **Ontology** (Socrates & Aristotle 400-360 BC): The study of being
- Explicit description of the conceptualization of a domain

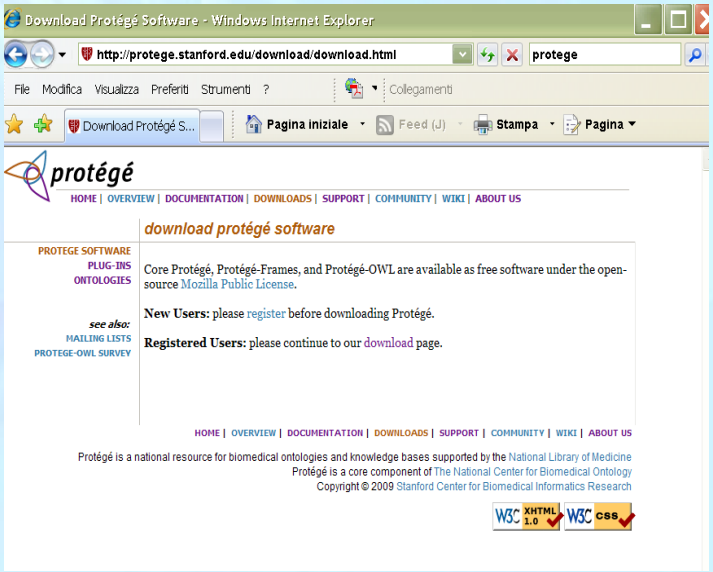
- **Ontology in Information Science** is a base for **Semantic Web** (a group of methods and technologies to allow machines to understand the meaning - or "semantics" - of information)
- **Ontology Components**
  - **concepts** – set of entities within a domain
  - **relations** – interaction between concepts or concept's properties
  - **instances** – concrete examples of concepts of the domain
  - **axioms** – explicit rules to constrain the use of concepts
- **Role of Ontologies**
  - To share **common understanding** of the structure of descriptive information
    - among people
    - among software agents
    - between people and software
  - To enable **reuse** of domain knowledge
  - To introduce **standards** to allow interoperability

OpenTox

OWL : The Web Ontology Language
Protégé Ontology Editor

- **OWL: the Web Ontology Language** is the World Wide Web Consortium (W3C) standard
- **OWL components**:
  - a formal description of concepts, terms, and relationships within a given knowledge domain
  - formal computational definitions
  - tools for reasoning
- **OWL tutorial** published by The University of Manchester "A Practical Introduction to Ontologies & OWL" http://www.co-ode.org/resources/tutorials/intro/

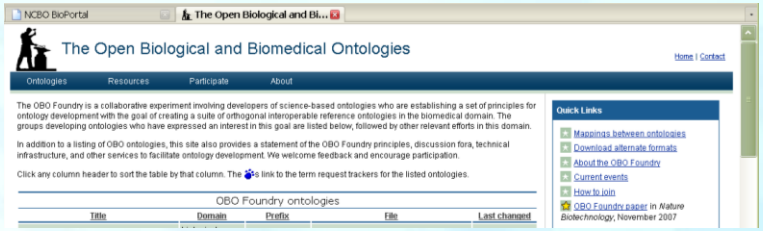
- **Protégé** is a free, open source **ontology editor** and knowledgebase framework (http://protege.stanford.edu)
  - provides tools for visualizing ontologies as well as for constructing them
  - facilitates **OWL ontology development** and maintenance
  - allows an **automated reasoning**: RACER is a reasoner frequently used with Protégé (www.sts. tu-harburg.de/~r.f.moeller/racer/)



OpenTox

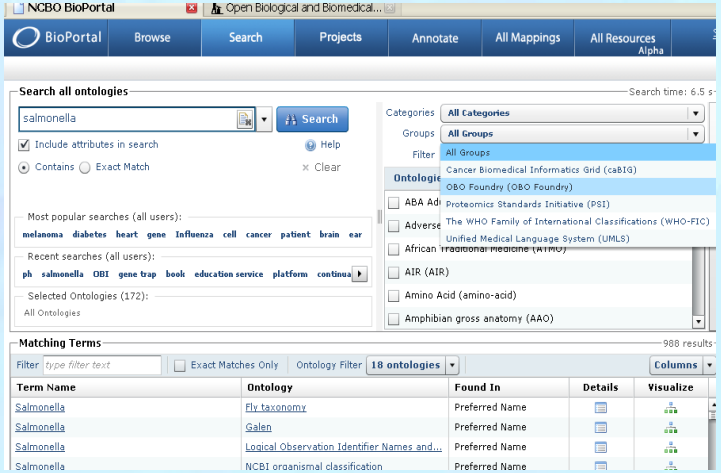
Public Biomedical Ontologies Resources

- **The Open Biomedical Ontologies (OBO):** ontology development is regulated within the OBO Foundry, which defines a set of shared principles governing ontology development http://obofoundry.org



- **OBO Foundry principles and criteria:**
  - The ontology is **open** and available to be used by all.
  - The ontology is in a **common formal language**.
  - The developers of the ontology agree in advance **to collaborate** with developers of other OBO Foundry ontology **where domains overlap**.
  - **Orthogonality**: for any particular domain, there is community convergence on a single controlled vocabulary.

- **Web-based ontology portal: BioPortal** (www.bioontology.org/tools/portal/bioportal.html) allow users to browse, search, and visualize ontologies.



- **Ontology Lookup Service:** accessing ontologies www.ebi.ac.uk/ontology-lookup/

OpenTox

Biomedical Application of Ontologies and Examples

- **Possible biomedical applications of ontologies:**
  - Search and query of heterogeneous biomedical data
  - Data exchange among applications
  - Information integration
  - Representation of encyclopedic knowledge
  - Computer reasoning with data

- **A growing community interest in using and producing biomedical ontologies.**

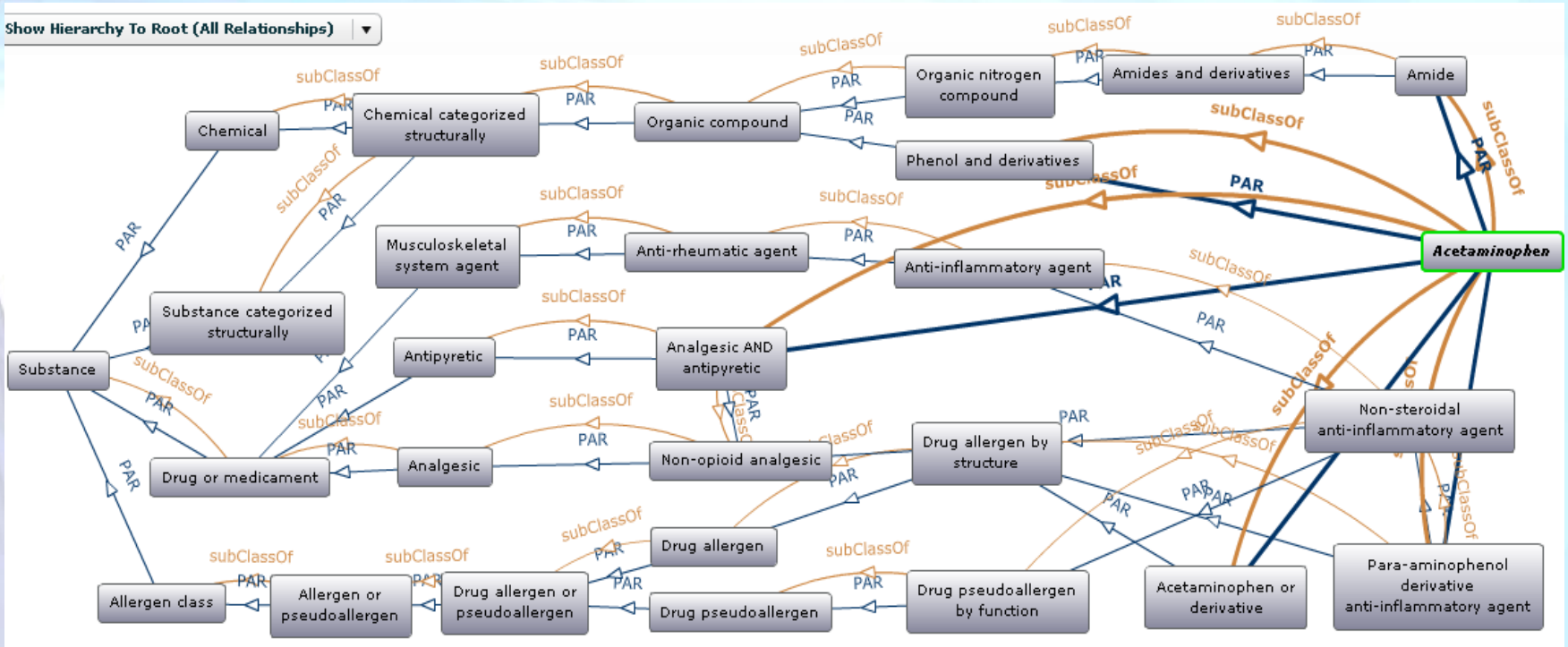
- **Examples of most important ontology projects:**
  - Gene Ontology (GO)
  - Chemical entities of biological interest (ChEBI)
  - Ontology for biomedical investigations (OBI)
  - NCI Thesaurus
  - Foundational Model of Anatomy (FMA)
  - Mouse Adult Gross Anatomy (MA)
  - Mouse gross anatomy and development (EMAP)

- **Toxicology domain – no ontology available**

OpenTox

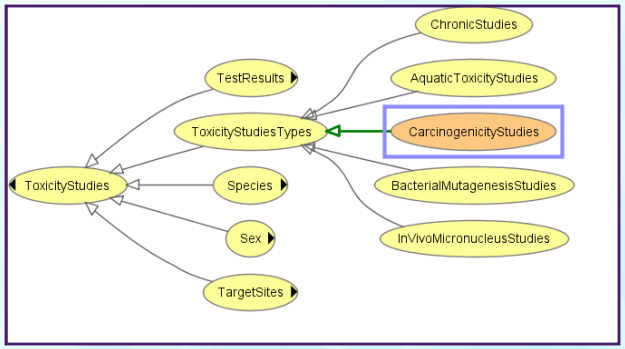
Example of ontological representation of terms

SNOMED CT (**S**ystematized **N**omenclature of **Med**icine - **C**linical **T**erms):
Paracetamol (Acetaminophen) example



OpenTox Toxicological Endpoint Ontology

- **Why we need an ontology?**
  - 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
- **Methodology**
  - Starting from 5 toxicological endpoints
  - following OBO Foundry principles



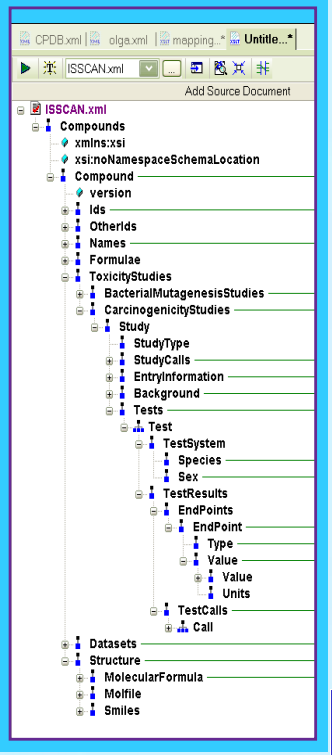
Other freely available resources: DSSTOX, GoReni (ITEM), etc

Protégé, free open source OWL (Web Ontology Language) editor

OpenTox Toxicological Data Ontology

Re-use of pieces or terms defined in neighboring ontologies (OWL and OBO)

ToxML scheme



OpenTox

Toxicological data: needs for standards ToxML scheme

- Need for data standards for automatic data integration
  - Example: **Carcinogenic Activity**

CPDBAS: Carcinogenic Potency Database
http://www.epa.gov/ncct/dsstox/sdf_cpdbas.html#SDFFields

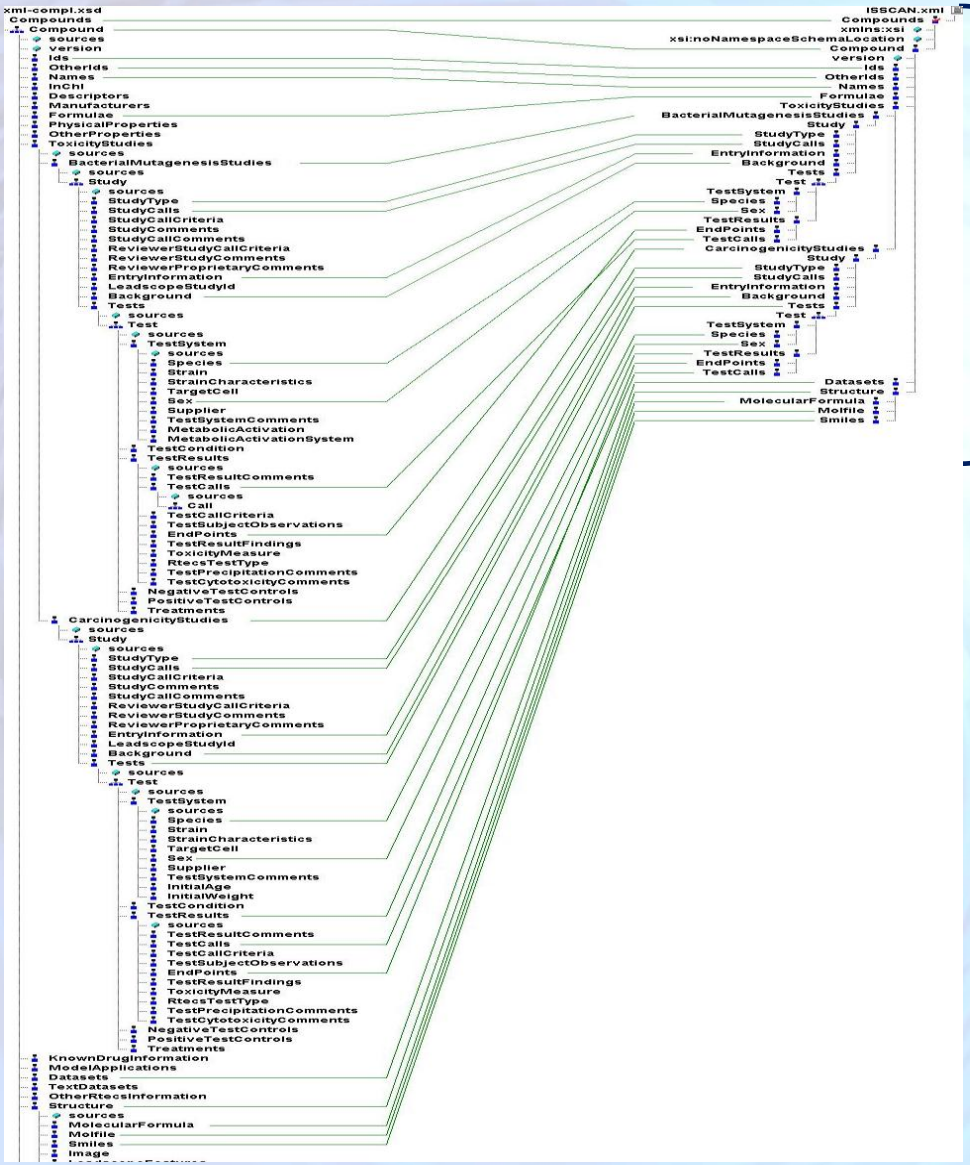
ISSCAN: Chemical Carcinogens Database
http://www.iss.it/ampp/dati/cont.php?id=233&lang=1&tipo=7

- **ToxML** (http://www.leadscope.com/toxml.php)
  - is a public initiative led by scientists at LeadScope, Inc
  - controlled vocabularies and XML scheme for storing chemical toxicity data
  - the latest version of ToxML public schema (April 7th , 2009)

- It is supported by OpenTox for interoperable data communications between services

OpenTox

Mapping of the ISSCAN entry - ToxML xsd scheme



CAS	67-66-3
Substance ID	4
ChemName	Chloroform
Synonyms	Formyl trichloride; methane trichloride; methenyl trichloride; Methyl trichloride; R 20; r 20 (refrigerant); Refrigerant R20; trichloroform; Trichloromethane
SAL	negative
Canc	positive
Reference	CPDB
TD50_Rat	262
TD50_Mouse	90.3
Rat_Male_Canc	positive
Rat_Male_NTP	ND
Rat_Female_Canc	positive
Rat_Female_NTP	ND
Mouse_Male_Canc	positive
Mouse_Male_NTP	ND
Mouse_Female_Canc	positive
Mouse_Female_NTP	ND
MolWeight	119.38
Formula	CHCl3
SDF file	Connection table
SMILES	ClC(Cl)Cl

OpenTox

ToxML: mapping of toxicological data
Needs for extensions

- The latest version of public ToxML scheme (Last updated April 7th, 2009) has been studied in terms of their suitability to map the content of databases candidate for the OpenTox database using Stylus Studio 2008 XML Enterprise software
  - 1. ISSCAN database
  - 2. In vivo micronucleus database (in development at the ISS);
  - 3. Bacterial mutagenesis database (to be developed at the ISS)
  - 4. RepDose ITEM database (ITEM)

- OpenTox Toxicological Ontology: ToxML integration
  - OWL Ontology: Classes and Hierarchies on the base of ToxML scheme

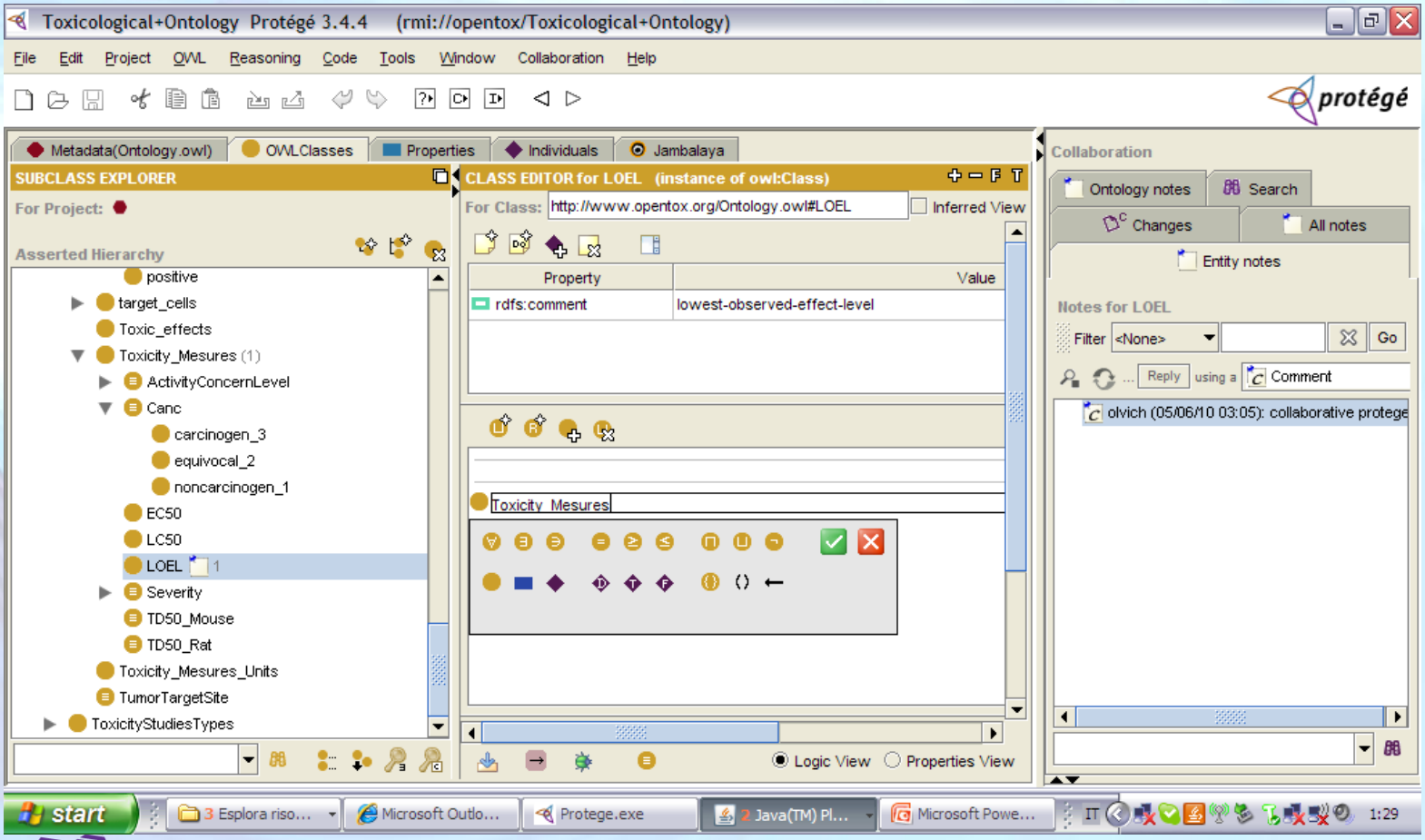
  - Extension needs:
    - Free text field: TargetSite – Organs Ontology and Effects Ontology to avoid a huge variability of terms, as each laboratory/researcher is able to enter his individual description of an observed effect

    - Introduction of relationship between different classes, restrictions: e.g., introducing the property "has_Test_Species" limites the test species suitable for certain toxicity study.

OpenTox

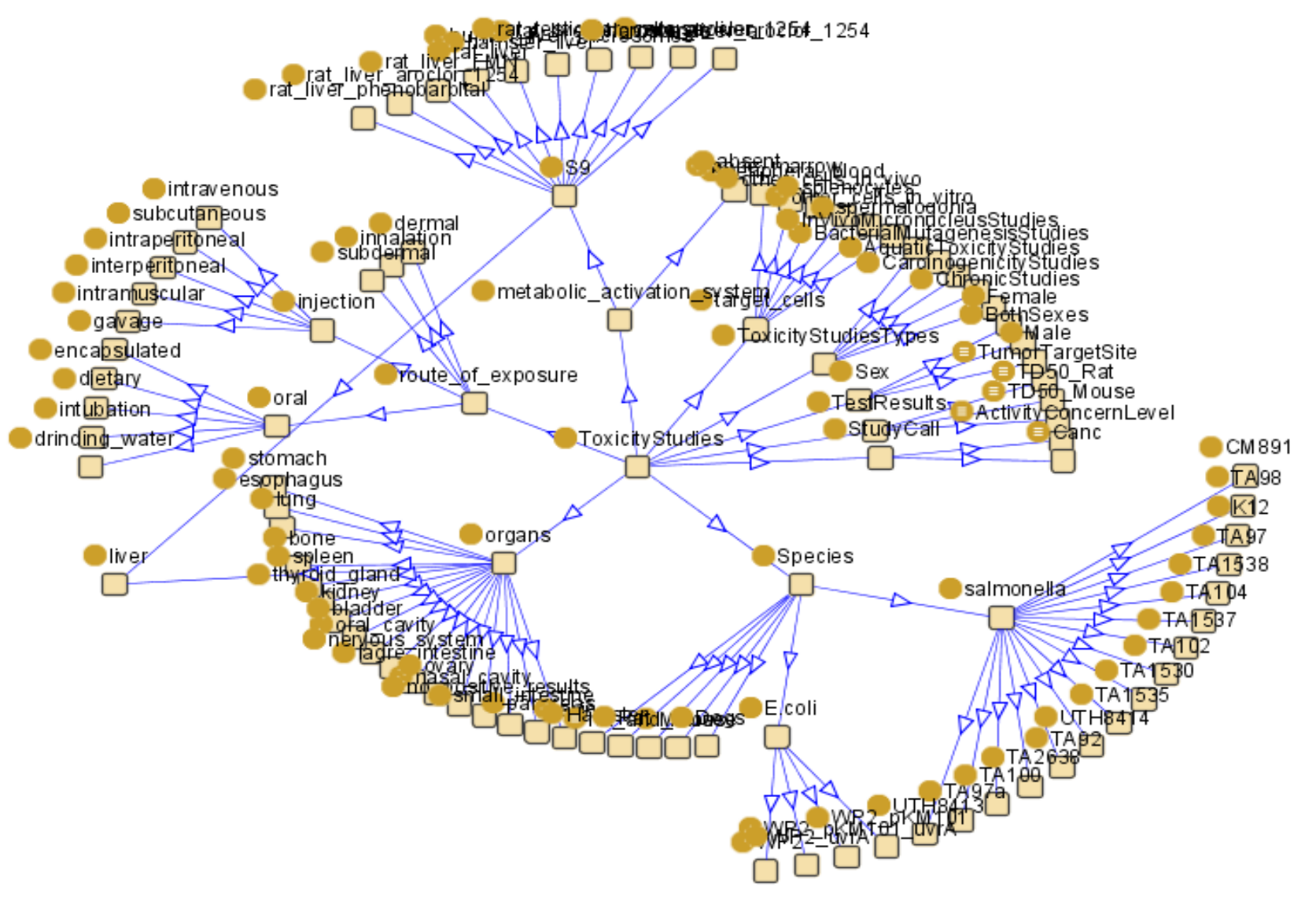
OpenTox: Collaborative Ontology Development

- **Collaborative Protégé: why?**
  - annotating ontology components
  - tracking changes, history of concept
  - discussion, live messages
  - proposals and voting
  - searching and filtering
  - defining users, groups, policies
  - available in multi-user mode



OpenTox

Toxicological Ontology: graphical representation



OpenTox

Ontology For Target Organs

- Contribution of **Fraunhofer Institute for Toxicology and Experimental Medicine (ITEM)**

- Ontology Development on the base of **INHAND** (International Harmonization of Nomenclature ad Diagnostic Criteria for Lesions in Rats and Mice)

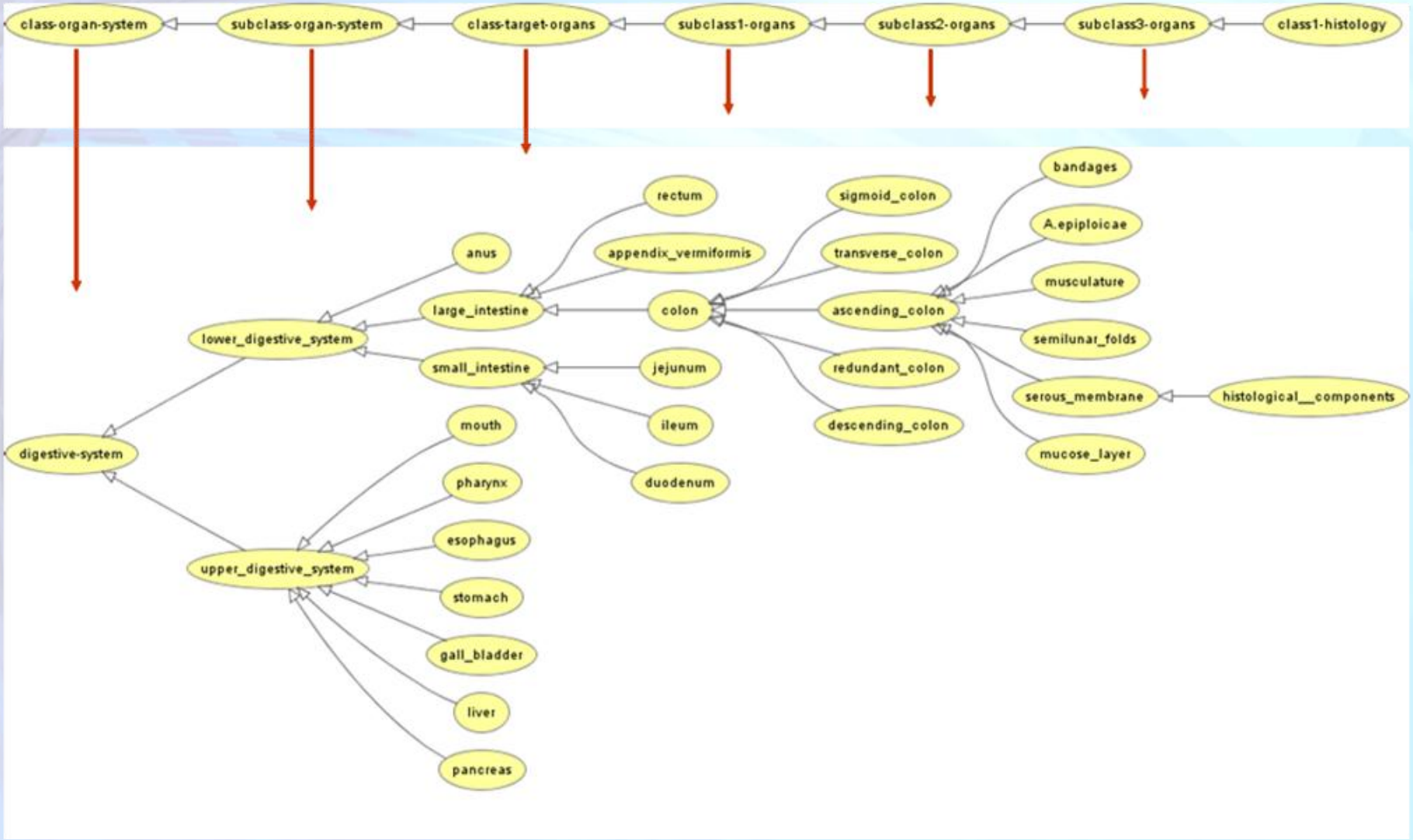
  Organs system (class) – Subclass organs system
  Target organs (class) – Targets organs (subclass 1 to N)
  Histopathology (class) – Histopathology (subclasses if needed)

- Key Features of the ITEM Organs Ontology:
  - more synonyms regarding organ systems, target organs and their subclasses
  - organs are more detailed, up to histological components
  - organs linked to different organ systems (e.g. ovaries)
  - reviewed by pathologists, who have been involved in the INHAND process
  - 12 organ systems fully described

- Perspectives:
  - linking of the organ systems and their components with pathologic effects

OpenTox

Ontology For Target Organs: Digestive system example

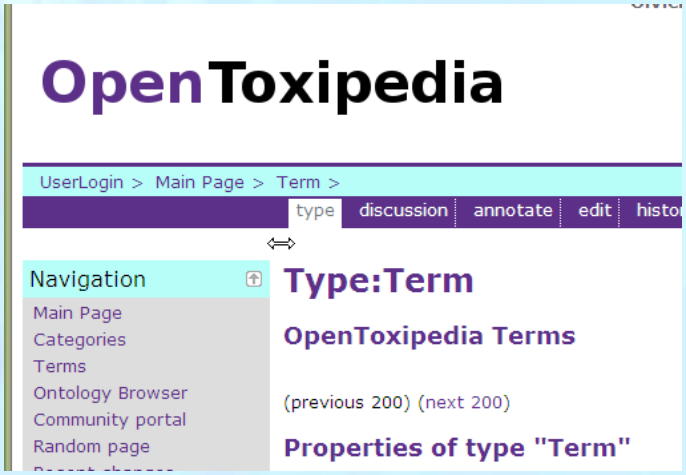


OpenTox

OpenToxipedia: community-based, predictive toxicology knowledge resource

- The OpenToxipedia content has moved to the new Semantic Media Wiki (SMW) platform. You will find a link to it at the common page http://www.opentoxipedia.org
  - Community based collaborative database
  - Automatically-generated lists.
  - Improved data structure, semantic terms annotation
  - External reuse: RDF export and import in Protégé Ontology Editor

- **Creating, adding, editing and keeping terms used in toxicology terminology**
- **862 toxicological terms with description and literature references classified into 26 categories**
- **Transparency and scientific basis of information**
- **Curation by the OpenTox toxicologists**



OpenTox

Plan for Future Work

- Continue collaborative ontology development using the Collaborative Protégé server, covering more toxicological endpoints
- Review carefully all existing OBO and Bioportal ontologies, find the overlapping ontology, valuate the possibility of import
- External collaboration, e.g., with Ontology for Biomedical Investigations (OBI) group
- Setup Protege sub project for *in vitro* assays ontology (ToxCast data)
- Upload OpenTox Toxicological Endpoints Ontology to the BioPortal Web Site
- OpenToxipedia: community based development

OpenTox

OpenTox Ontology Working Group

- **External collaboration**
  - Barry Hardy
- **Algorithms and features ontology**
  - Nina Jeliazkova, Vedrin Jeliazkov, Ivelina Nikolova
  - Christoph Helma
  - Tobias Girschick
  - Andreas Karwath
  - Georgia Melagraki
  - Sunil Chawla
  - David Gallagher
- **Collaborative Protégé administrator**
  - Micha Rautenberg
- **Toxicological Endpoint Ontology**
  - Aldo Benigni, Sylvia Escher, Helvi Grimm, Alexey Lagunin, Olga Tcheremenskaia
- **OpenToxipedia**
  - Alexey Lagunin, Sergey Novikov, Natalya Skvortsova

OpenTox

Development and Use of Predictive Toxicology Applications

An OpenTox Workshop

19 Sep 2010, Rhodes, Greece

Data management and integration

presented by Nina Jeliazkova
(Ideaconsult Ltd., Bulgaria)

OpenTox

SEVENTH FRAMEWORK PROGRAMME

Outline

- Ontology Server
- Using the Dataset API
- Dataset Integration
- Data quality and accuracy

Component	Description	URL Template (example)
Compound	Representations of chemical compounds	http://host:port/compound/{compoundid}
Feature	Properties and identifiers	http://host:port/feature/{featureid}
Dataset	Encapsulates set of chemical compounds and their property values	http://host:port/dataset/{datasetid}
Model	OpenTox model services	http://host:port/model/{modeld}
Algorithm	OpenTox algorithm services	http://host:port/algorithm/{algorithmid}
Validation, Report	A validation corresponds to the validation of a model on a test dataset.	http://host:port/validation/{validationid} http://host:port/report/{reportid}
Task	Asynchronous jobs are handled via an intermediate Task resource. A resource, submitting an asynchronous job should return the URI of the task.	http://host:port/task/{taskid}
Ontology service	Provides storage and SPARQL search functionality for objects, defined in OpenTox services and relevant ontologies	http://host:port/ontology
Authentication and authorisation	Granting access to protected resources for authorised users	http://host:port/opensso http://host:port/opensso-pol

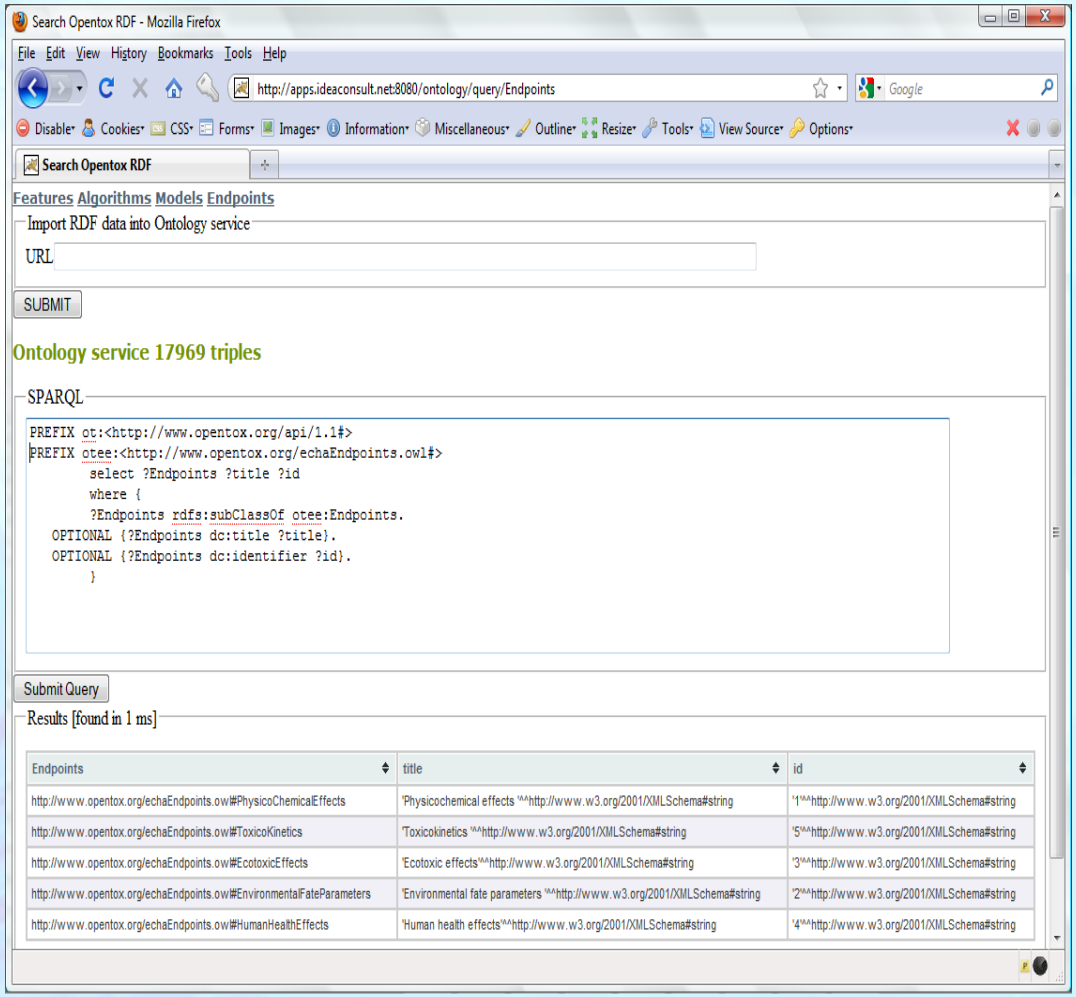
OpenTox

- 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

OpenTox

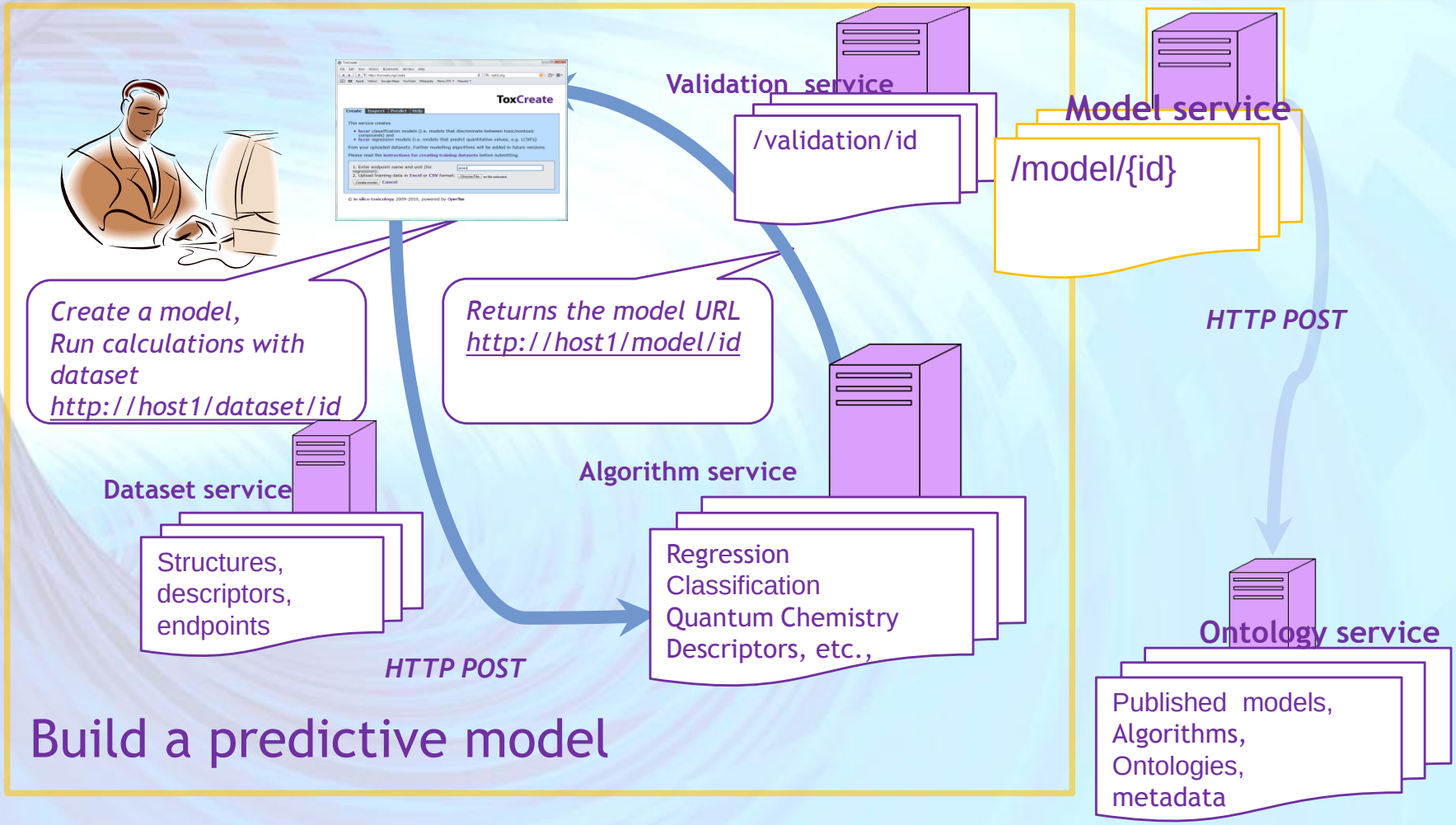
Ontology service

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



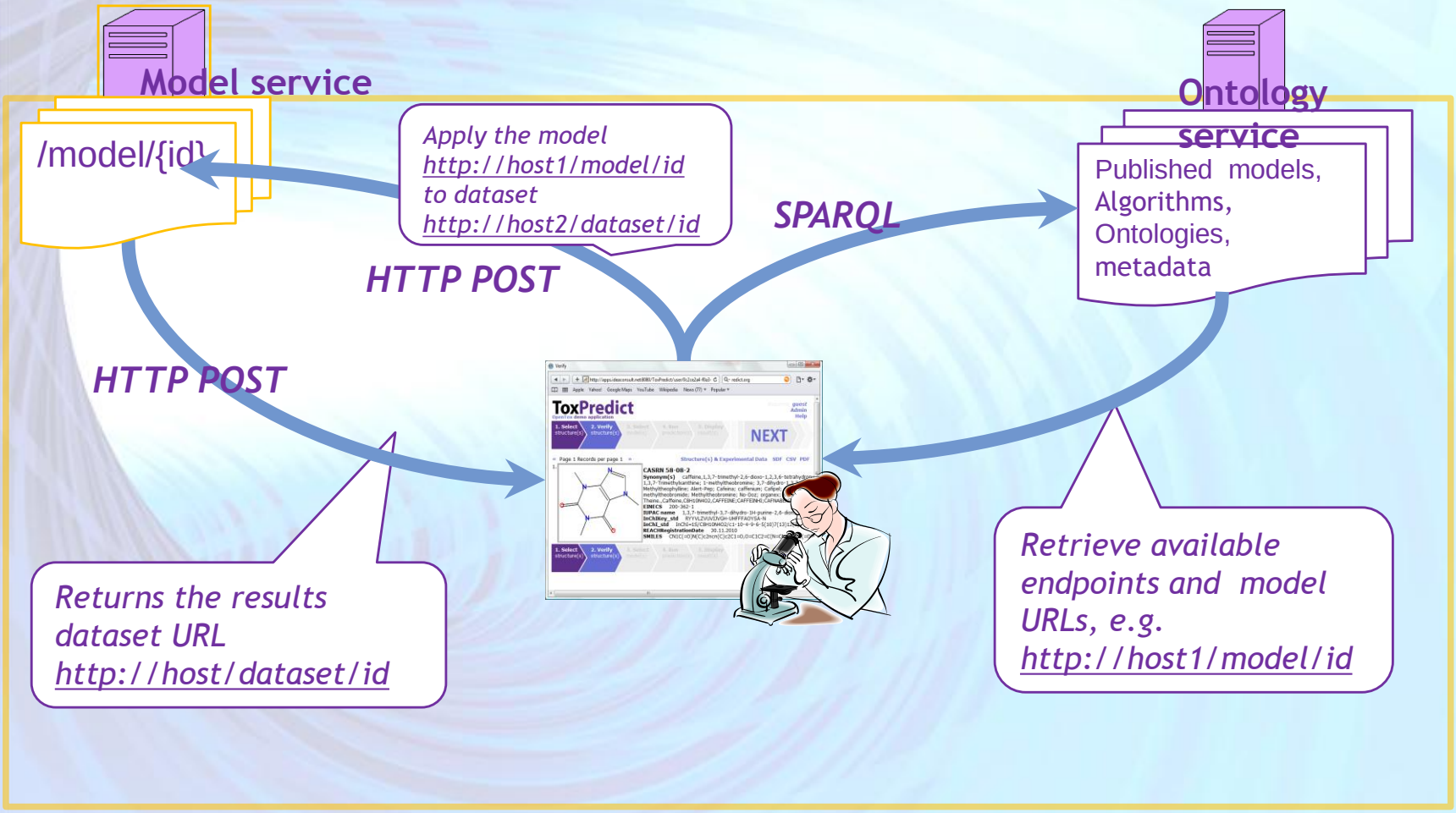
OpenTox

Build a predictive model



OpenTox

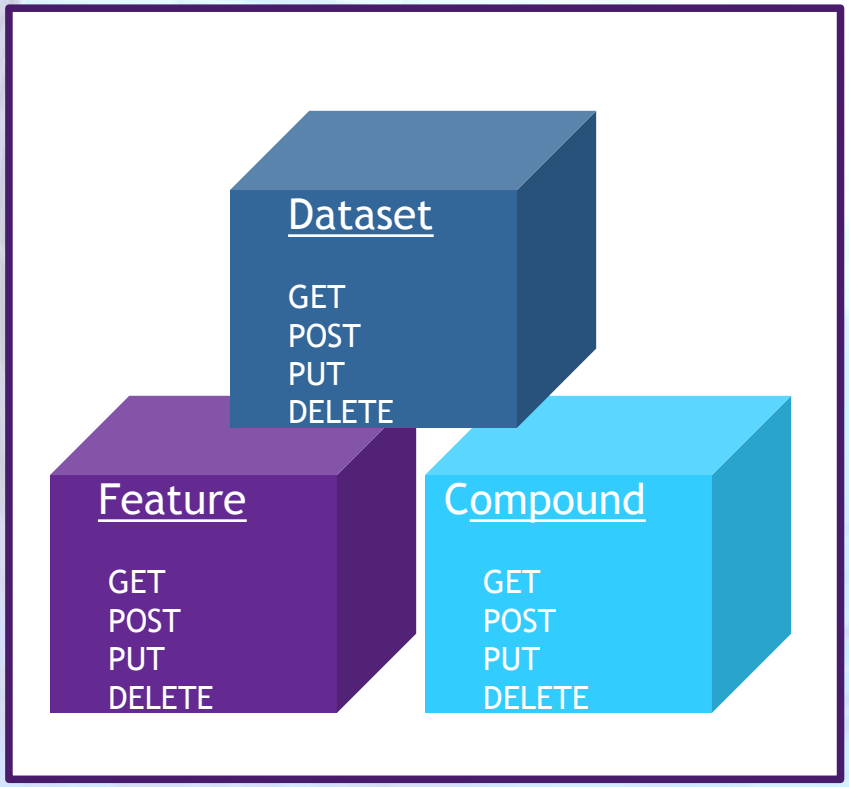
Apply predictive models

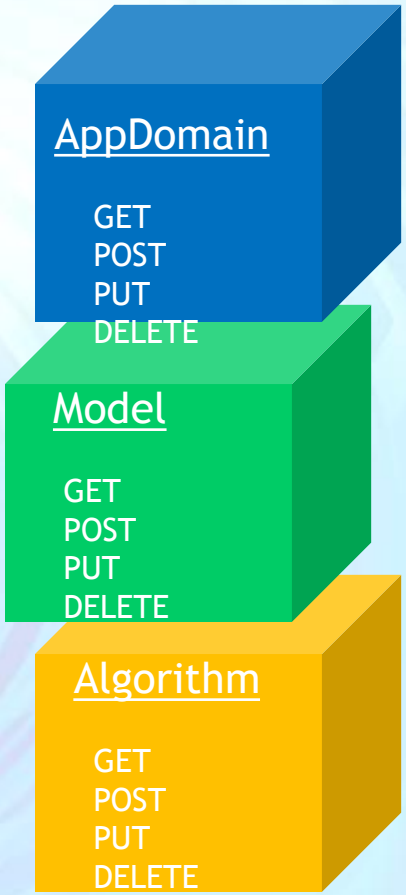


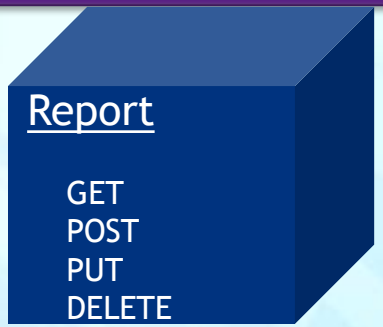
OpenTox

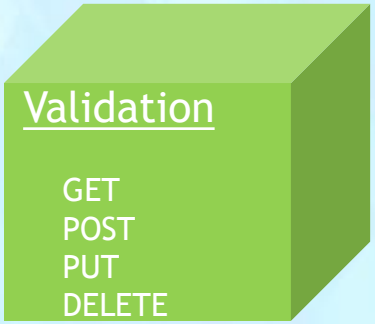
OpenTox API (Application Programming Interface)

- The way applications talk to each other
- The way developers talk to applications









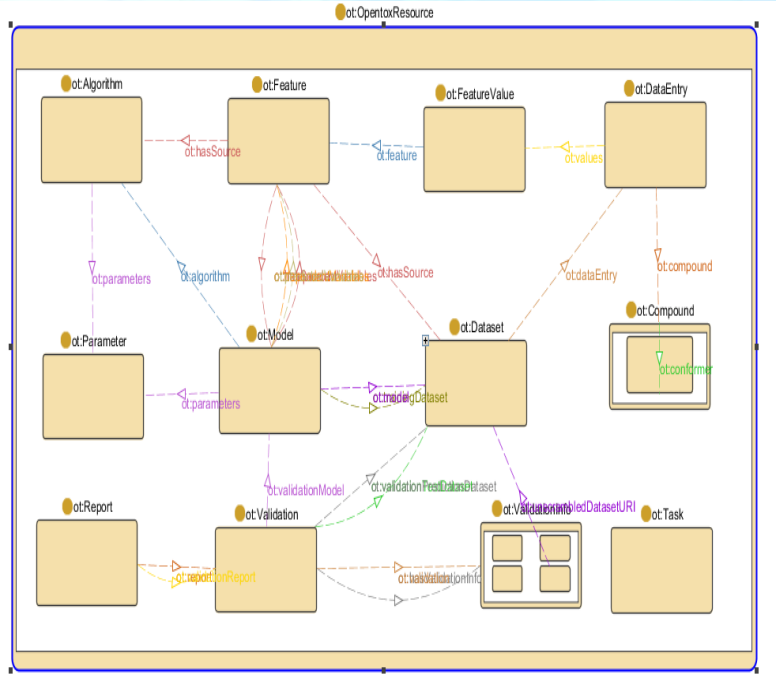
OpenTox

SEVENTH FRAMEWORK PROGRAMME

RDF - Resources representation

- The opentox.owl ontology
  - A common OWL data model of all OpenTox resources
  - Describes OpenTox resources
  - Describes relationships between them
  - Generates object's RDF representations.
- RDF/XML representation is mandatory for OpenTox resources.
- Uniform approach to data representation
  - Calculated and measured properties of chemical compounds are represented in an uniform way
  - Linked to the resource used for data generation
  - Annotated via ontology entries
  - Model representations link to algorithms and data used

All OpenTox components are defined by OWL ontology
http://opentox.org/api/1.1/opentox.owl
All resources are subclasses of ot:OpenToxResource



OpenTox

Resources: Chemical compound

Compound

Provides different representations for chemical compounds with a unique and defined chemical structure.

/compound/{id}

Conformer

/compound/{id}/conformer/{id}

Documentation

http://opentox.org/dev/apis/api-1.1/structure

Representation

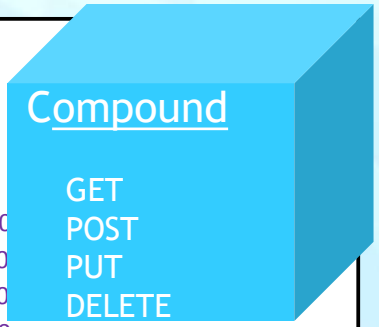
A subclass of ot:OpenToxResource.

Supports different Chemical MIME formats

RDF representation only for specifying owl:sameAs links to external resources

Example 1. Retrieve compound as MOL

```
$ curl -H "Accept:chemical/x-mdl-molfile"
http://apps.ideaconsult.net:8080/ambit2/compound/1
CH2O
APtclcactv09040902283D 0   0.00000     0.00000
  4  3  0  0  0  0  0  0  0  0999 V2000
   -0.6004    0.0000    0.0001 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.6072    0.0000   -0.0004 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.1472    0.9353    0.0016 H   0  0  0  0  0  0  0  0  0  0  0  0
    1.1472   -0.9353    0.0016 H   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  2  0  0  0  0
  2  3  1  0  0  0  0
  2  4  1  0  0  0  0
```



Example 2. Retrieve compound as SMILES

```
$ curl -H "Accept:chemical/x-daylight-smiles"
http://apps.ideaconsult.net:8080/ambit2/compound/1
O=C
```

Example 3. Query compounds

```
$ curl –H Accept:chemical-mime “
http://apps.ideaconsult.net:8080/ambit2/query/compound/{any-identifier-or-keyword}
```

```
$ curl –H Accept:chemical-mime “
http://apps.ideaconsult.net:8080/ambit2/query/smarts?search={smarts}
```

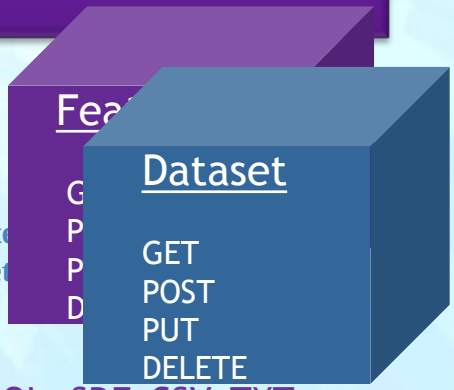
OpenTox

Resources: Dataset

Dataset

Provides access to chemical compounds and their features (e.g. structural, physical-chemical, biological, toxicological properties)

```
@prefix ad:     <http://apps.ideaconsult.net:8080/ambit2/dataset/> .
@prefix af:     <http://apps.ideaconsult.net:8080/ambit2/feature/> .
@prefix ot:     <http://www.opentox.org/api/1.1#> .
...
ad:9  a      ot:Dataset ;
    ot:dataEntry
          [ a      ot:DataEntry ;
            ot:compound
<http://apps.ideaconsult.net:8080/ambit2/compound/413/conformer/409421> ;
            ot:values
                  [ a      ot:FeatureValue ;
                    ot:feature af:21576 ;
                    ot:value "3.309999942779541"^^xsd:double
                  ] ;
            ot:values
                  [ a      ot:FeatureValue ;
                    ot:feature af:21573 ;
                    ot:value "3.0"^^xsd:double
                  ]
          ] ;
```



Operations

- **POST – Upload a dataset**
- **PUT – Update the dataset conte**
- **DELETE – Remove the dataset**

Representation

RDF/XML (mandatory), MOL, SDF, CSV, TXT, ARFF, .. (optional)

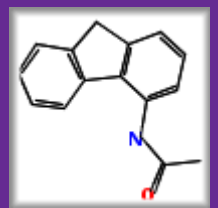
- **The dataset consists of data entries.**
- Each entry is associated with exactly **one chemical compound**, identified by its URI and available via OpenTox Compound service API;
- One and the same compound can be associated with multiple dataset entries;
- Every "column" is associated with a **Feature,** its representation should be available via OpenTox Feature API

OpenTox

OpenTox datasets: Uniform access to data

Everything described by W3C RDF (Resource Description framework)

Compound/ Data	http://myhost.com/feature/21580	http://myhost.com/feature/21589	http://myhost.com/feature/21573	http://myhost.com/feature/21576	http://myhost.com/feature/21588	http://myhost.com/feature/21858	http://myhost.com/feature/22114
http://myhost.com/compound/413	N,N-dimethyl-4-aminoazobenzene	CN(C1=CC=C(C=C1)N=N/C2=CC=CC=C2)C	3	3.31	225.3	YES	3.123
http://myhost.com/compound/44497	4-acetamidofluorene	O=C(Nc3					
...	...						



http://myhost.com/feature/21573

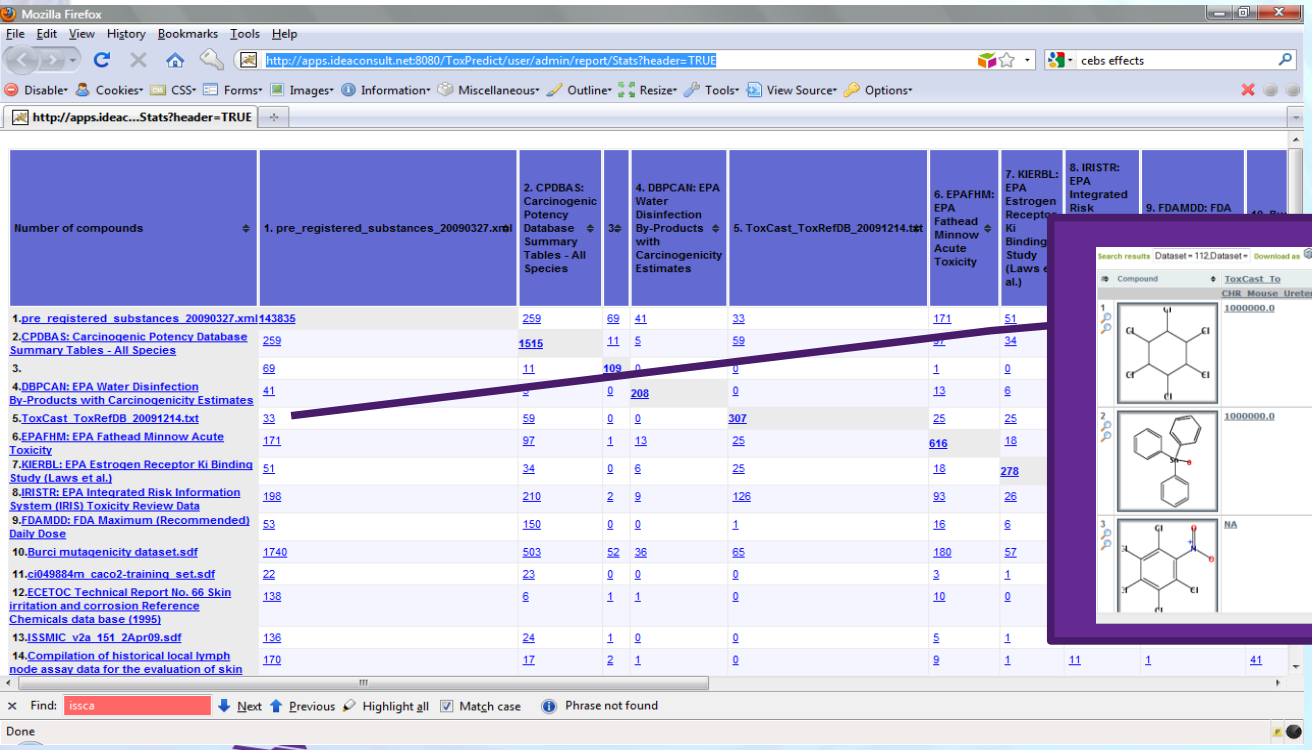
http://myhost.com/feature/21858

```
http://myhost.com/feature/22114
a        ot:Feature , ot:NumericFeature ;
dc:creator
"http://www.blueobelisk.org/ontologies/chemoinformatics-algorithms/#xlogP" ;
dc:title "XLogP" ;
ot:hasSource
<http://myhost.com/algorithm/org.openscience.cdk.qsar.descriptors.molecular.XLogPDescriptor> ;
=        otee:Octanol-water_partition_coefficient_Kow .
```

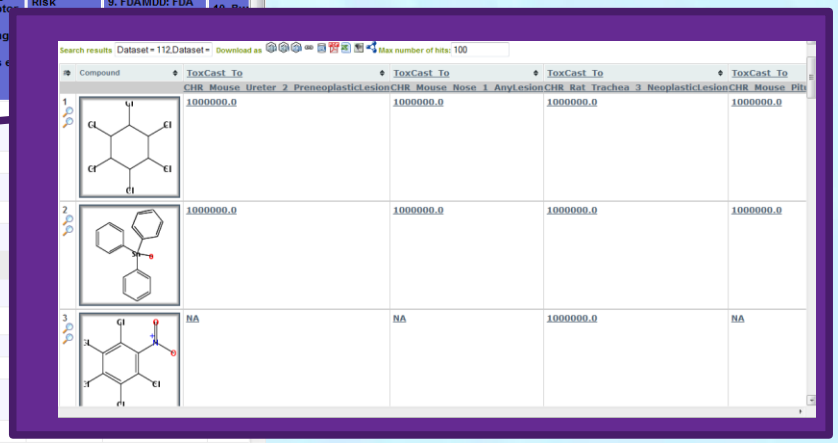
OpenTox

Uniform access to the data

- Datasets can be easily merged, compared , and calculations reproduced, regardless of their physical place.
- The dataset service offers property, compound, substructure and similarity searches via uniform OpenTox Application Programming

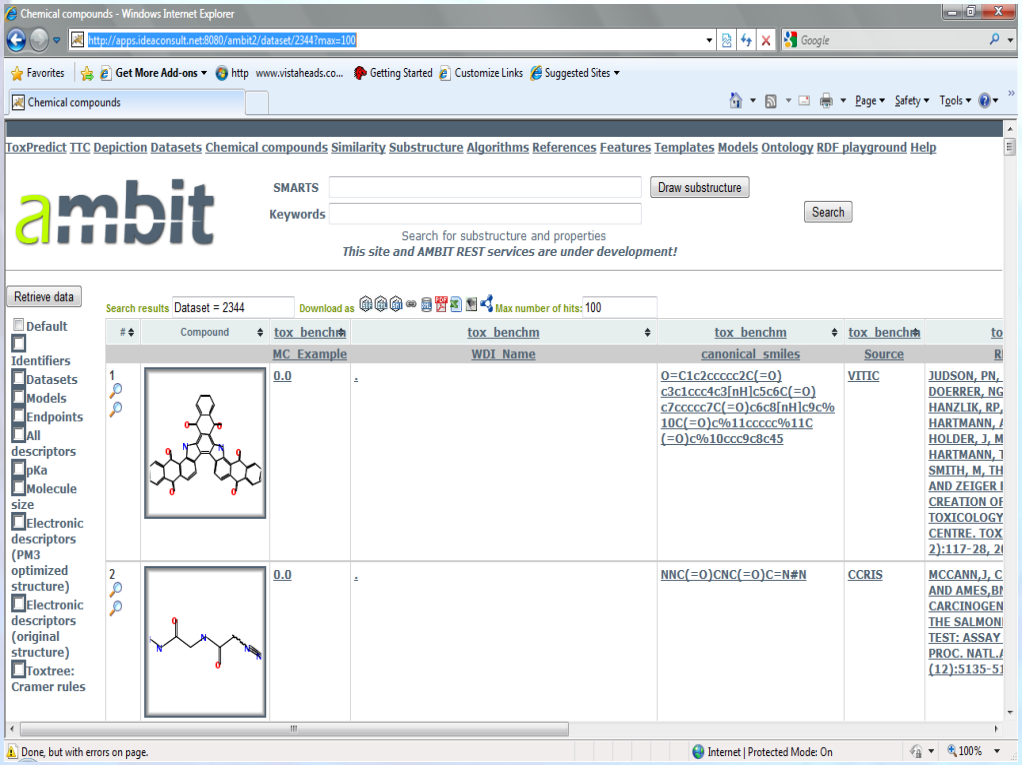


Number of compounds	1. pre_registered_substances_20090327.xml	2. CPDBAS: Carcinogenic Potency Database Summary Tables - All Species	3	4. DBPCAN: EPA Water Disinfection By-Products with Carcinogenicity Estimates	5. ToxCast_ToxRefDB_20091214.txt	6. EPAFHM: EPA Fathead Minnow Acute Toxicity	7. KIERBL: EPA Estrogen Recepto Ki Binding Study (Laws et al.)	8. IRISTR: EPA Integrated Risk	9. FDAMDD: FDA
1.pre_registered_substances_20090327.xml	143835	259	69	41	33	171	51		
2.CPDBAS: Carcinogenic Potency Database Summary Tables - All Species	259	1515	11	5	59	97	34		
3.	69	11	109	0	9	1	0		
4.DBPCAN: EPA Water Disinfection By-Products with Carcinogenicity Estimates	41	2	0	208	0	13	6		
5.ToxCast_ToxRefDB_20091214.txt	33	59	0	0	307	25	25		
6.EPAFHM: EPA Fathead Minnow Acute Toxicity	171	97	1	13	25	616	18		
7.KIERBL: EPA Estrogen Receptor Ki Binding Study (Laws et al.)	51	34	0	6	25	18	278		
8.IRISTR: EPA Integrated Risk Information System (IRIS) Toxicity Review Data	198	210	2	9	126	93	26		
9.FDAMDD: FDA Maximum (Recommended) Daily Dose	53	150	0	0	1	16	6		
10.Burci mutagenicity dataset.sdf	1740	503	52	36	65	180	57		
11.ci049884m_caco2-training_set.sdf	22	23	0	0	0	3	1		
12.ECETOC Technical Report No. 66 Skin irritation and corrosion Reference Chemicals data base (1995)	138	6	1	1	0	10	0		
13.ISSMIC_v2a_151_2Apr09.sdf	136	24	1	0	0	5	1		
14.Compilation of historical local lymph node assay data for the evaluation of skin	170	17	2	1	0	9	1	11	1



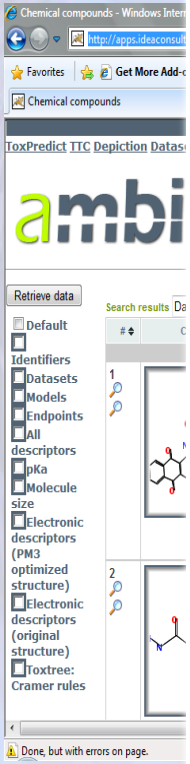
Example: mutagenicity dataset

- http://apps.ideaconsult.net:8080/ambit2/dataset/2344 (the dataset)
- http://apps.ideaconsult.net:8080/ambit2/dataset/2344/metadata (metadata, obviously)



Example: mutagenicity dataset

- Activity




```
@prefix ot:     <http://www.opentox.org/api/1.1#> .
@prefix dc:     <http://purl.org/dc/elements/1.1/> .
@prefix :       <http://apps.ideaconsult.net:8080/ambit2/> .
@prefix ota:    <http://www.opentox.org/algorithmTypes.owl#> .
@prefix otee:   <http://www.opentox.org/echaEndpoints.owl#> .
…
@prefix af:     <http://apps.ideaconsult.net:8080/ambit2/feature/> .

af:28958
    a       ot:Feature , ot:NumericFeature ;
    dc:creator "194.141.0.136" ;
    dc:title "Activity" ;
    ot:hasSource "tox_benchmark_N6512.sdf" ;
    ot:units "" ;
    =       otee:Mutagenicity .

ot:hasSource
    a       owl:ObjectProperty .

ot:units
    a       owl:DatatypeProperty .

ot:Feature
    a       owl:Class .

ot:NumericFeature
    a       owl:Class ;
    rdfs:subClassOf ot:Feature .
```

Query: Is there other mutagenicity data available?

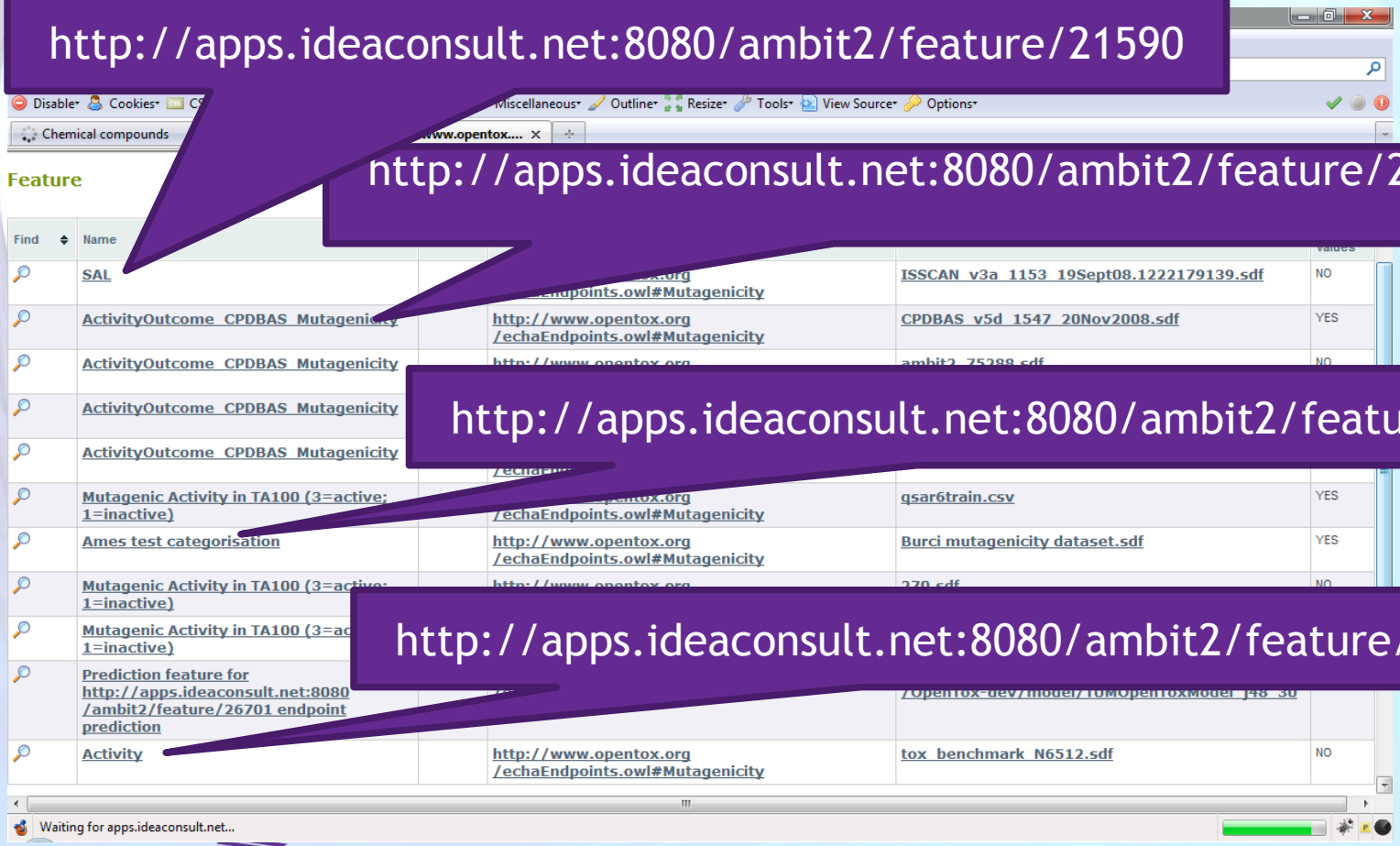
- http://apps.ideaconsult.net:8080/ambit2/feature?sameas=http%3A%2F%2Fwww.opentox.org%2FechaEndpoints.owl%23Mutagenicity

http://apps.ideaconsult.net:8080/ambit2/feature/21590

http://apps.ideaconsult.net:8080/ambit2/feature/21611

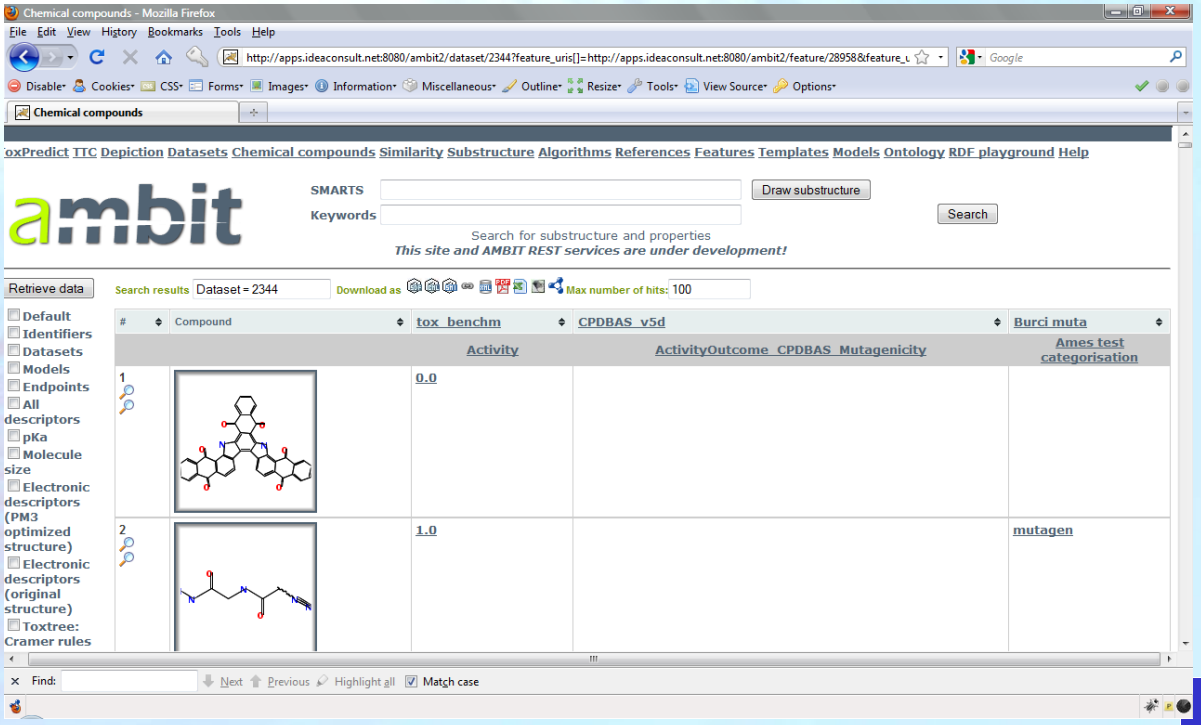
http://apps.ideaconsult.net:8080/ambit2/feature/26221

http://apps.ideaconsult.net:8080/ambit2/feature/28958



Merge mutagenicity data

- http://apps.ideaconsult.net:8080/ambit2/dataset/2344?feature_uris[]=http://apps.ideaconsult.net:8080/ambit2/feature/28958&feature_uris[]=http://apps.ideaconsult.net:8080/ambit2/feature/21611&feature_uris[]=http://apps.ideaconsult.net:8080/ambit2/feature/26221&feature_uris[]= http://apps.ideaconsult.net:8080/ambit2/feature/21590



Dataset : metadata and features

Description	URI Template
Retrieve entire dataset content. If uri-list, retrieve only compound URIs	http://host:port/dataset/{id}
Retrieve representation of features (columns) of the dataset	http://host:port/dataset/{id}/feature
Retrieves dataset metadata (name, etc.)	http://host:port/dataset/{id}/metadata

```
$ curl -H "Accept:application/rdf+xml" http://apps.ideaconsult.net:8080/ambit2/dataset/9/metadata
<rdf:RDF
xmlns:ot="http://www.opentox.org/api/1.1#"
……
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
 xml:base="http://apps.ideaconsult.net:8080/ambit2/">
<ot:Dataset rdf:about="dataset/9">
  <dc:source>ISSCAN_v3a_1153_19Sept08.1222179139.sdf</dc:source>
  <dc:publisher>somebody</dc:publisher>
  <rdfs:seeAlso>
    <bx:Entry rdf:about="reference/20117">
      <rdfs:seeAlso>http://www.epa.gov/NCCT/dsstox/sdf_isscan_external.html</rdfs:seeAlso>
      <dc:title>ISSCAN_v3a_1153_19Sept08.1222179139.sdf</dc:title>
    </bx:Entry>
  </rdfs:seeAlso>
  <dc:title>ISSCAN: Istituto Superiore di Sanita, CHEMICAL CARCINOGENS: STRUCTURES AND EXPERIMENTAL DATA</dc:title>
 </ot:Dataset>
</rdf:RDF>
```

Data publishing

1)POST a file with chemical structures and properties to OpenTox dataset service.

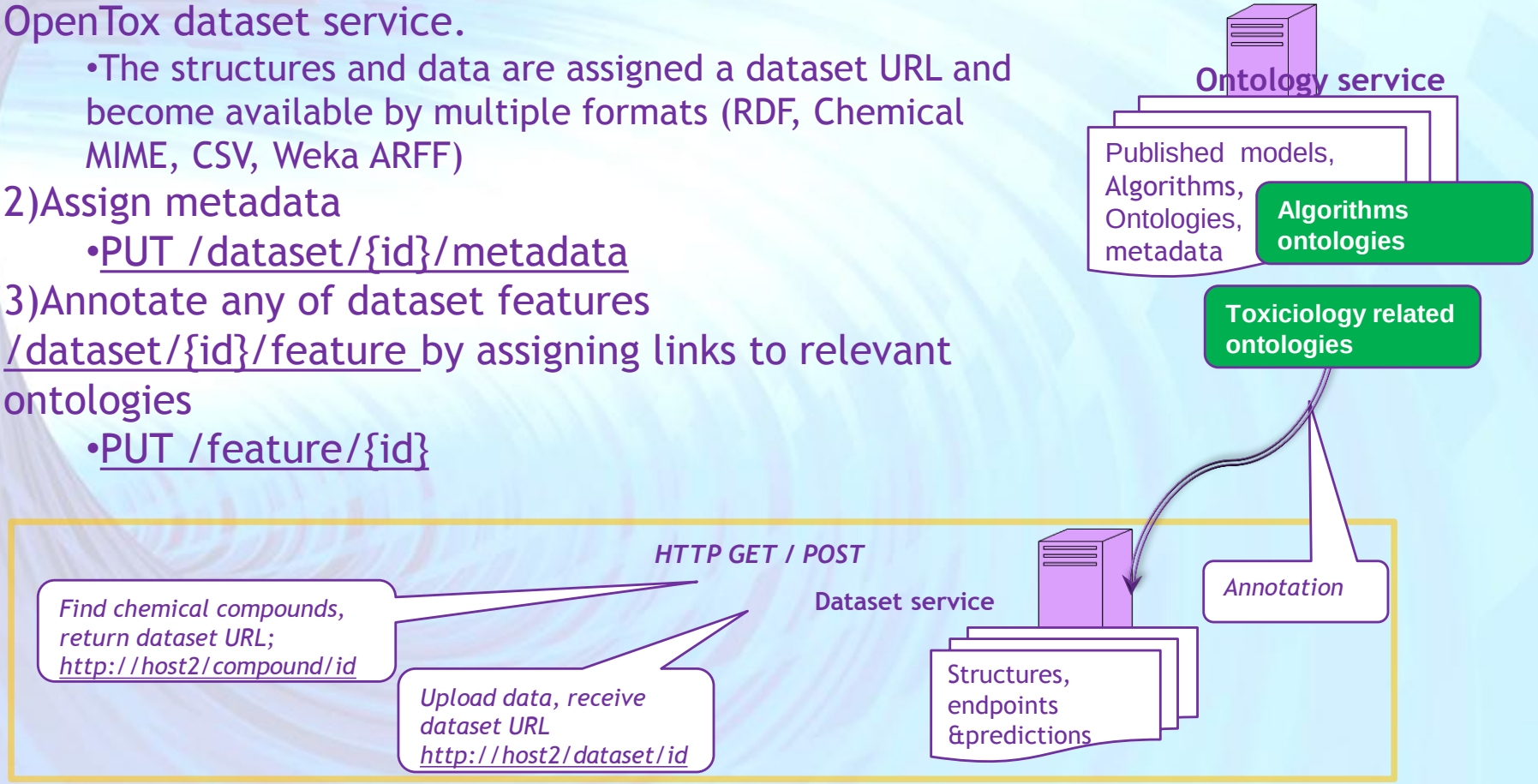
- The structures and data are assigned a dataset URL and become available by multiple formats (RDF, Chemical MIME, CSV, Weka ARFF)

2)Assign metadata

- PUT /dataset/{id}/metadata

3)Annotate any of dataset features /dataset/{id}/feature by assigning links to relevant ontologies

- PUT /feature/{id}

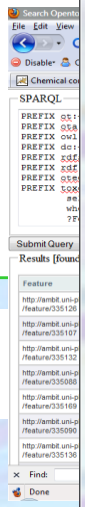


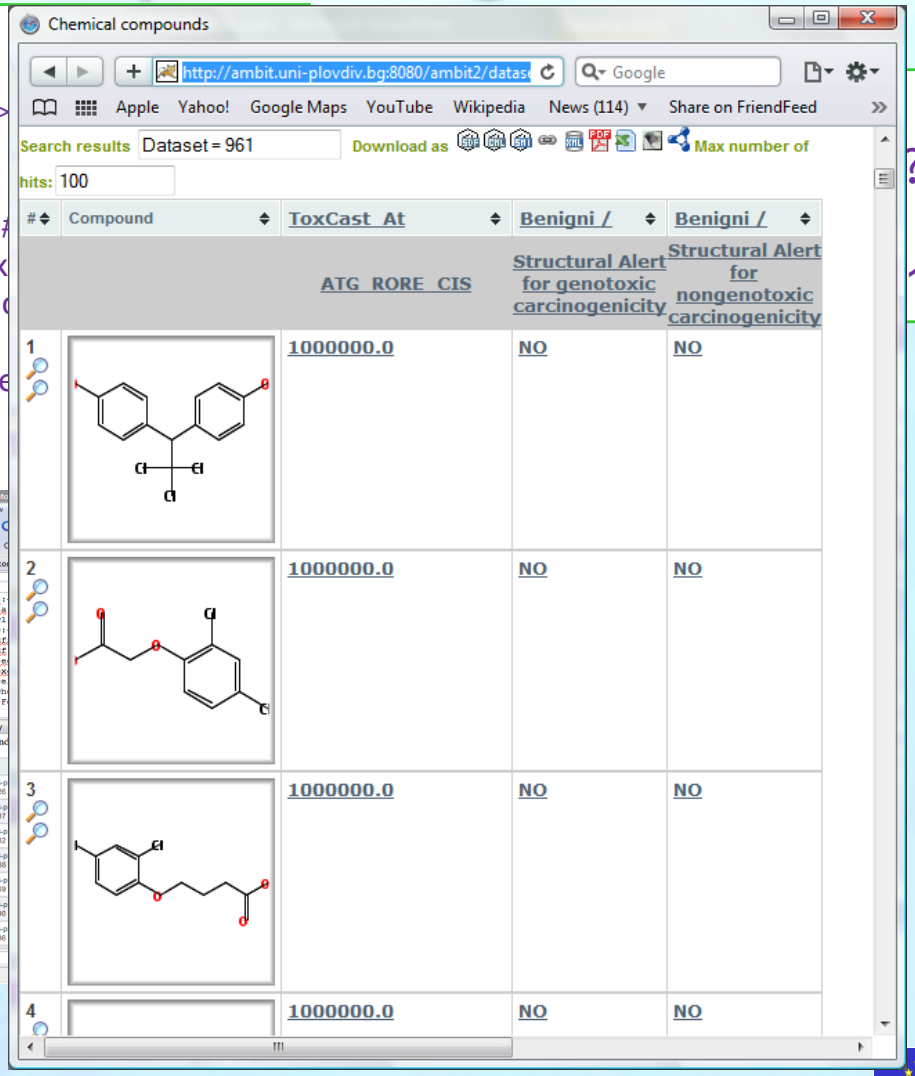
OpenTox

Dataset and Ontology - find an assay, linked to specific gene

```
PREFIX ot:<http://www.opentox.org/api/1.1#>
PREFIX ota:<http://www.opentox.org/algorithms.owl#>
PREFIX owl:<http://www.w3.org/2002/07/owl#>
PREFIX dc:<http://purl.org/dc/elements/1.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax
PREFIX otee:<http://www.opentox.org/echaEndpoints.
PREFIX toxcast:<http://www.opentox.org/toxcast#>
        select ?Feature ?title ?id ?assay ?geneid ?ge
        where {
        ?Feature rdf:type ot:Feature.
  {?Feature dc:title ?title}.
  {?Feature owl:sameAs ?assay}.
  {?assay toxcast:gene ?geneid}.
  {?assay toxcast:hasProperty ?genename}.
  {?genename rdf:type toxcast:GENE_NAME}.

        }
```





?feat

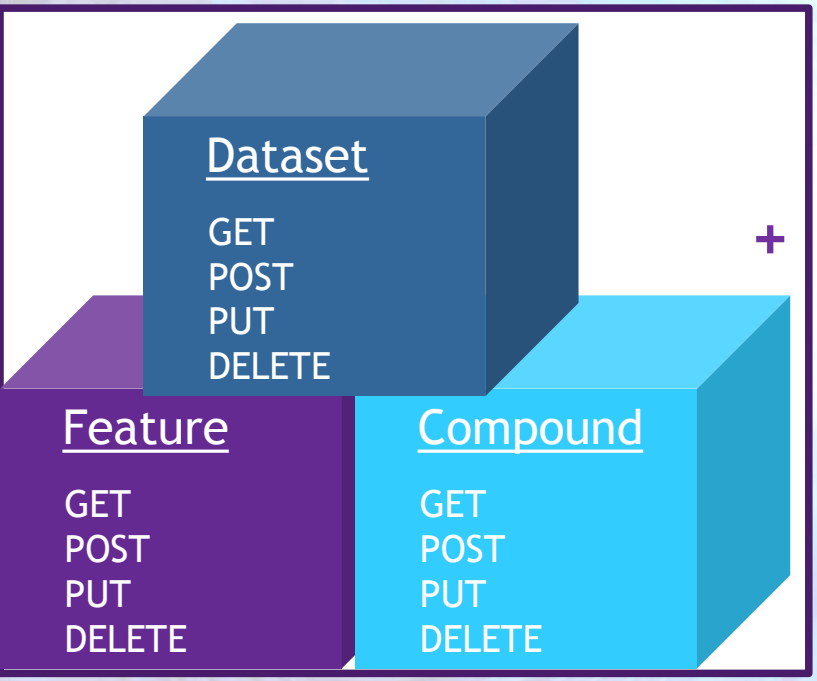
126

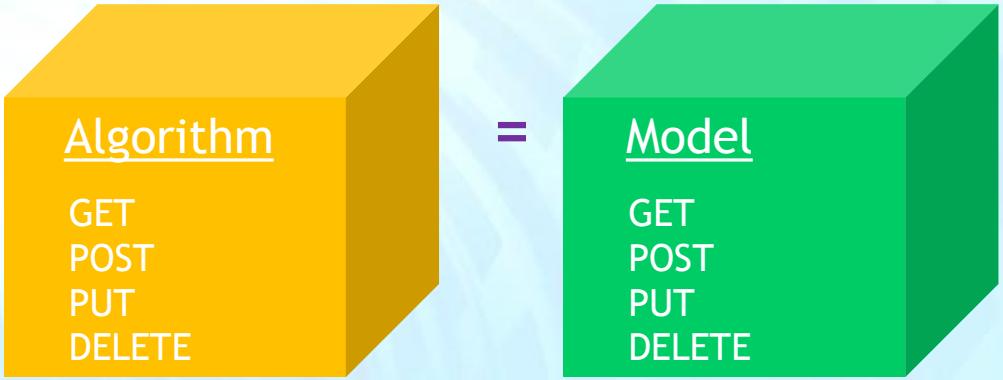
Query an OpenTox ontology service at http://ambit.uni-plovdiv.bg:8082/ontology

OpenTox

OpenTox dataset : create a model

Read data from a web address - process - write to a web address





http://myhost.com/algorithm/neuralnetwork

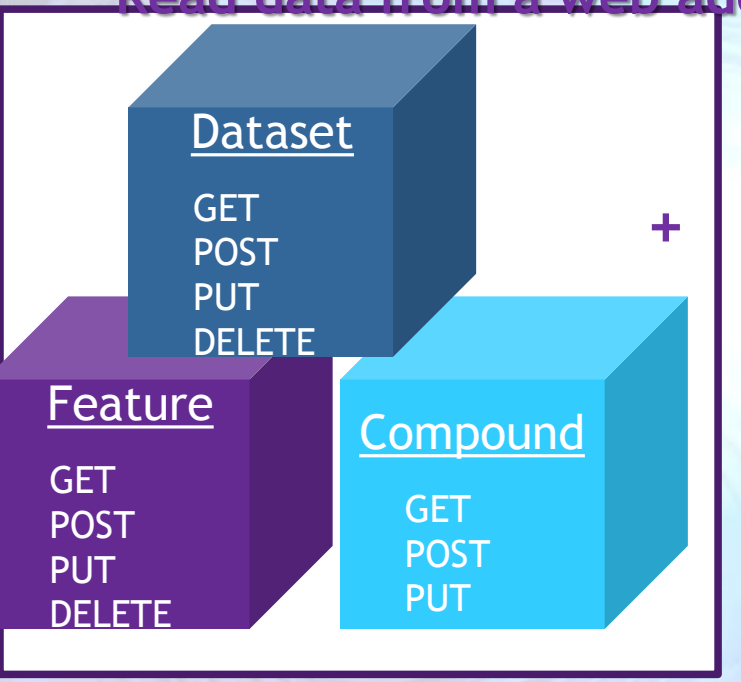
http://myhost.com/dataset/trainingset1

http://myhost.com/model/predictivemodel1

OpenTox

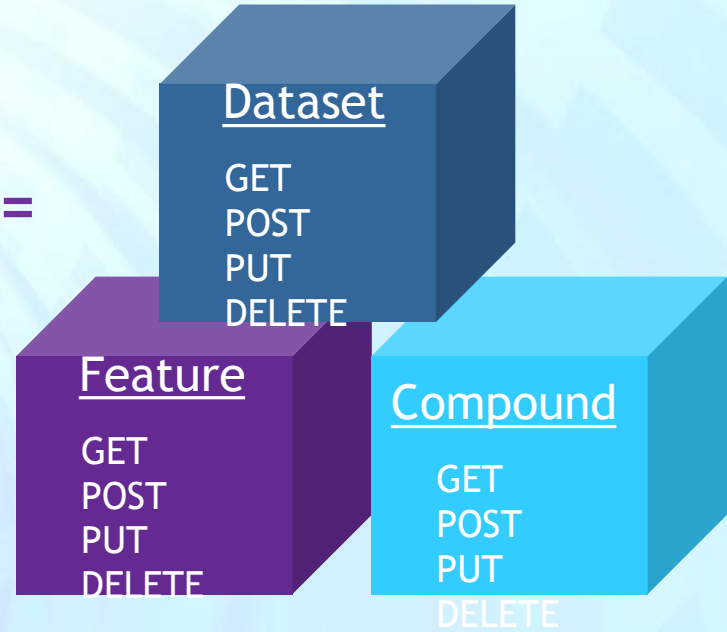
OpenTox dataset : descriptor calculation

Read data from a web address - process - write to a web address



+

=



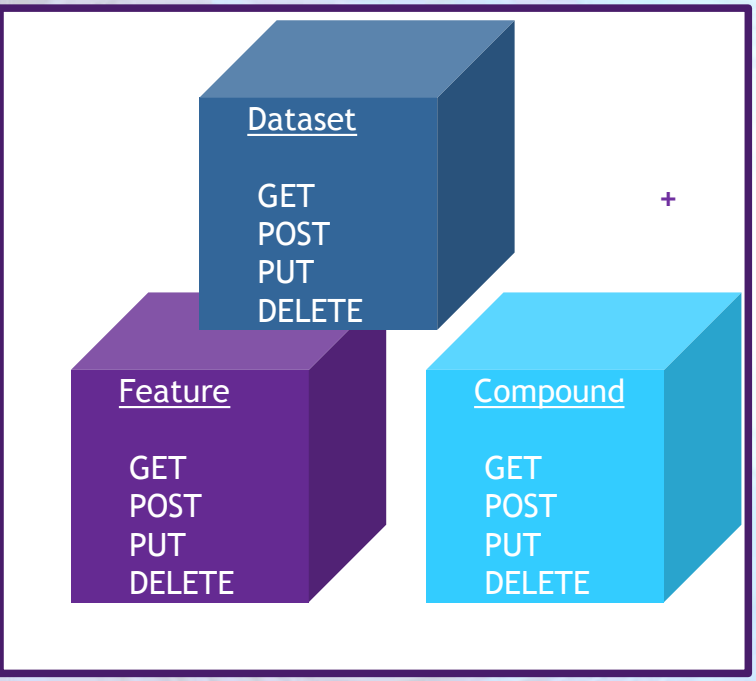
http://myhost.com/algorithm/{descriptorX}

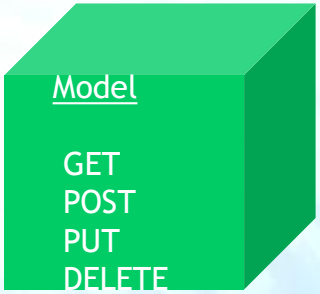
http://myhost.com/dataset/trainingset1

http://myhost.com/dataset/results

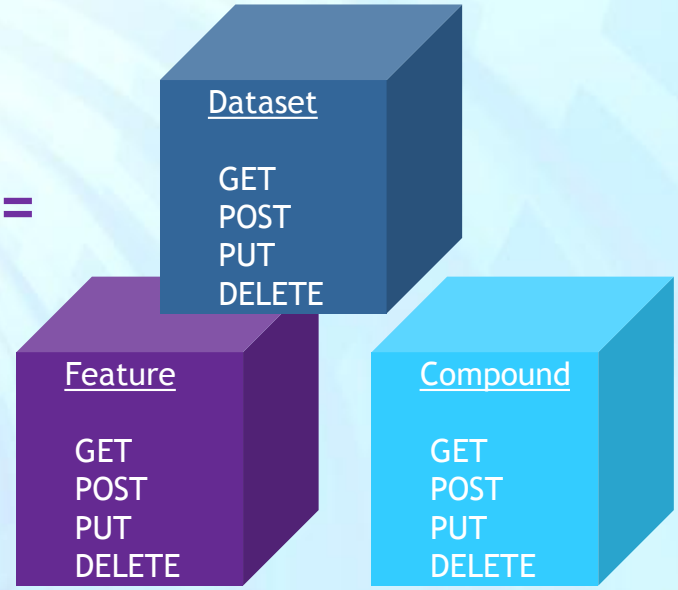
OpenTox dataset: apply a model

Read data from a web address - process - write to a web address





=



http://myhost.com/model/predictivemodel1

http://myhost.com/dataset/id1

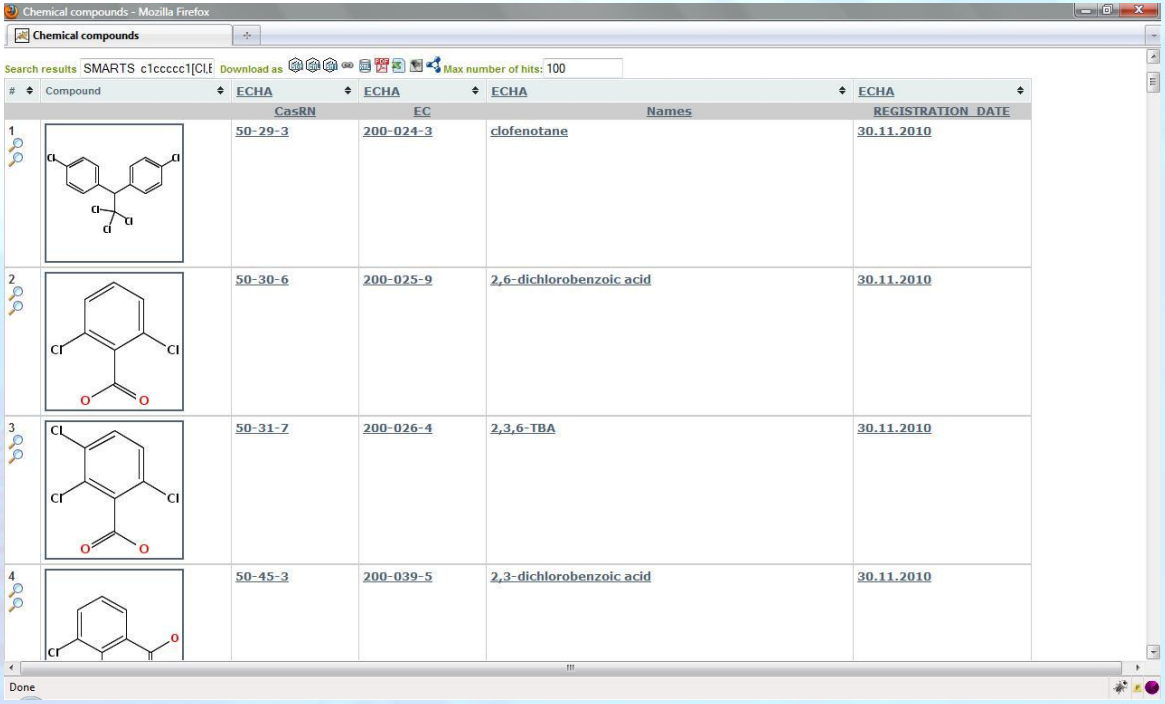
http://myhost.com/dataset/results1

OpenTox

OpenTox datasets : Substructure and similarity search

REST web service interface

http://apps.ideaconsult.net:8080/ambit2/query/smarts?search=c1ccccc1[Cl,Br,F,I]



OpenTox

Datasets : Structure and quality labels

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

Linked resources:
Compound, Algorithm, Model, Dataset, Features

Dataset Resource

http://apps.ideaconsult.net:8080/ambit2/dataset/R545

http://apps.ideaconsult.net:8080/ambit2/dataset/R545	data Entry	compound	http://apps.ideaconsult.net:8080/ambit2/compound/38/conformer/419609
		values	feature: TopoPSA
			Feature value: 6.480000019073486
			type: Feature Value
			feature: nHBDon
			Feature value: 0.0
			type: Feature Value
			feature: caco2
			Feature value: 4.849999904632568
			type: Feature Value
			feature: WNSA-3
			Feature value: -4.374800205230713
			type: Feature Value
			feature: FPSA-2
			Feature value: 0.8797000050544739
			type: Feature Value
		compound	http://apps.ideaconsult.net:8080/ambit2/compound/144824/conformer/419615
		values	feature: TopoPSA
			Feature value: 210.53999328861328
			type: Feature Value
			feature: nHBDon
			Feature value: 5.0
			type: Feature Value
			feature: caco2
			Feature value: -5.920000076293945
			type: Feature Value
			feature: WNSA-3
			Feature value: -64.12879943847656
			type: Feature Value
			feature: FPSA-2
			Feature value: 2.147799968719482
			type: Feature Value

Descriptor resource

Assay resource

Chemical compound

http://apps.ideaconsult.net:8080/ambit2/feature/22213

Name of the algorithm	type	Class
	type	Class
Numeric Feature	type	Class
	subClassOf	
Source	type	ObjectProperty
Units	type	DatatypeProperty
nHBDon	sameAs	http://www.blueobelisk.org/ontologies/chemoinformatics-algorithms/#hBondDonors
	Title	nHBDon
	Source	http://apps.ideaconsult.net:8080/ambit2/algorithm/org.openscience.cdk.qsar.descriptors.molecular.HBondDonorCountDescr
	Units	
	type	Numeric Feature

Blue Obelisk algorithms ontology

Regression
Classification
Quantum Chemistry Descriptors,etc.

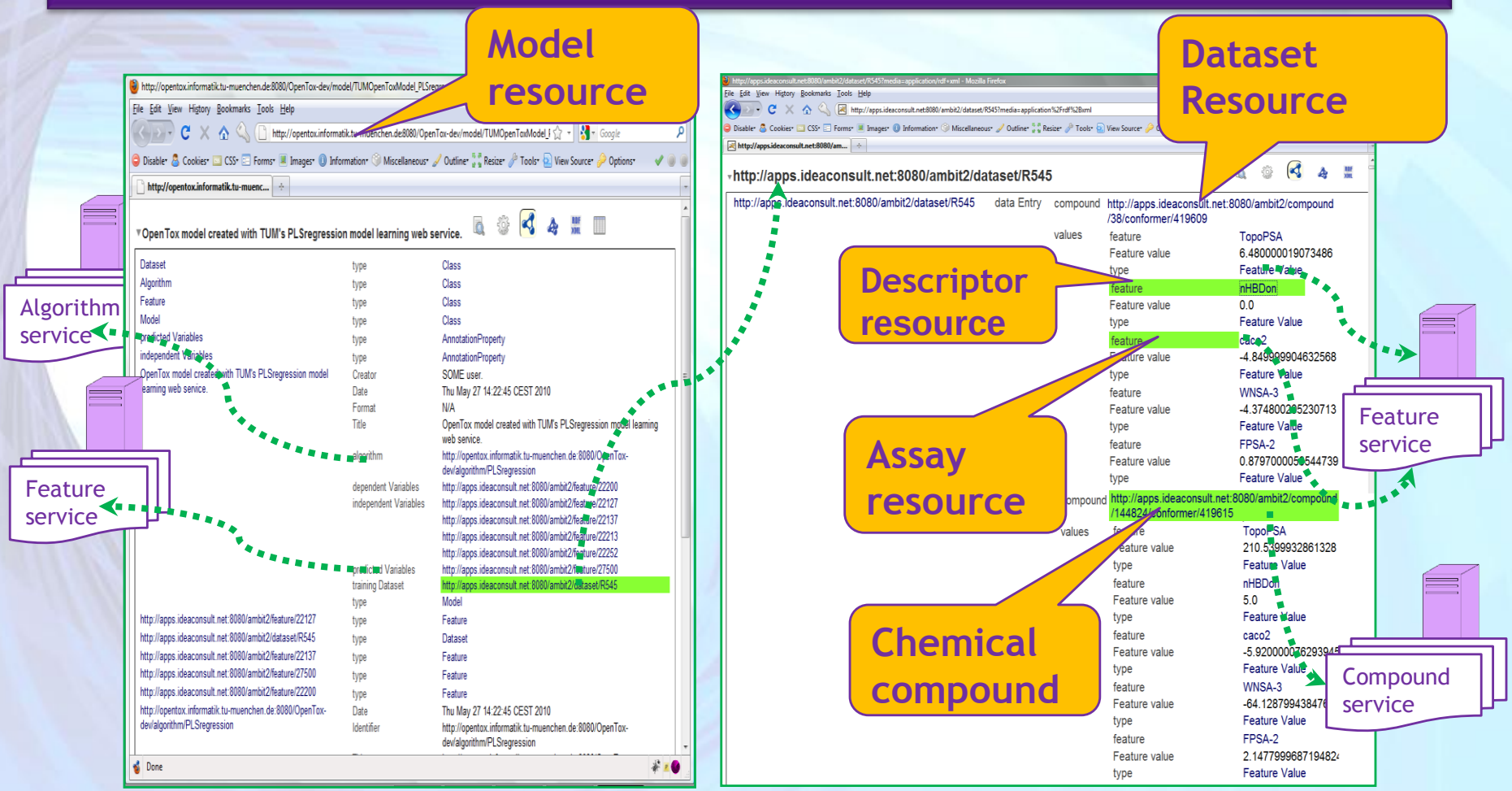
OpenTox algorithm types ontology

http://apps.ideaconsult.net:8080/ambit2/feature/22200

	type	Class
Numeric Feature	type	Class
	subClassOf	
Source	type	ObjectProperty
Units	type	DatatypeProperty
caco2	Title	caco2
	Source	ci049884m_caco2-training_set.sdf
	Units	
	type	Numeric Feature
	sameAs	Gastrointestinal absorption

Toxiciology related ontologies

Linked resources:
Compound, Algorithm, Model, Dataset, Features



Development of Predictive Toxicology Applications

An OpenTox Workshop

19 Sep 2010, Rhodes, Greece

Algorithms

Stefan Kramer

(Technische Universität München, Munich, Germany)

OpenTox

SEVENTH FRAMEWORK PROGRAMME

Overview of Application Programming Interfaces

Report
GET
POST
PUT
DELETE

AppDomain
GET
POST
PUT
DELETE

Dataset
GET
POST
PUT
DELETE

Validation
GET
POST
PUT
DELETE

Model
GET
POST
PUT
DELETE

Feature
GET
POST
PUT
DELETE

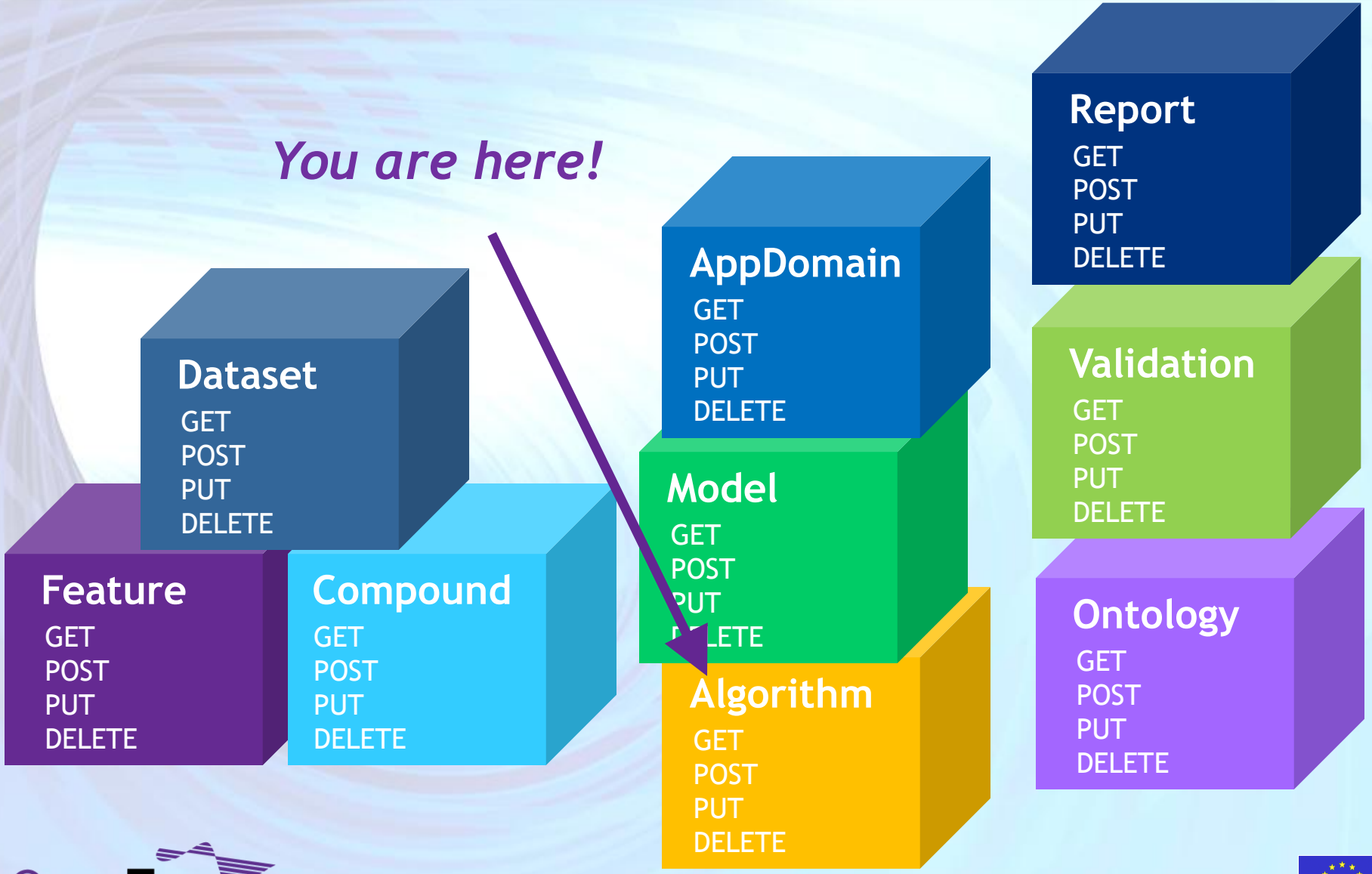
Compound
GET
POST
PUT
DELETE

Ontology
GET
POST
PUT
DELETE

Algorithm
GET
POST
PUT
DELETE

OpenTox

Overview of Application Programming Interfaces



OpenTox

Overview of Algorithms in the OpenTox Framework

- Algorithms for descriptor calculation: generation and selection of features for the representation of chemicals (structure based features, chemical and biological properties),
- Classification and regression algorithms for the creation of (Q)SAR models,
- Algorithms for the aggregation of predictions from multiple (Q)SAR models and endpoints, and aggregation of predictions,
- General purpose algorithms (e.g., for visualization, similarity and substructure queries, applicability domains, read across, ….)

OpenTox

OpenTox Algorithms: Descriptor Calculation and Feature Selection

- Descriptor calculation: services based on
  - OpenBabel
  - JoeLib2
  - CDK
  - multi-level neighborhood of atoms (MNA)
  - substructure/fragment generation ("product line": gSpan, FreeTreeMiner, BBRCs, LastPM; details later)
- Feature selection:
  - service for feature selection based on information gain
  - service for feature selection based on Chi^2 statistics
  - PCA
  - filter pipeline for preprocessing: combining approaches for handling missing values, feature selection, …

OpenTox

OpenTox Algorithms: Classification / SAR

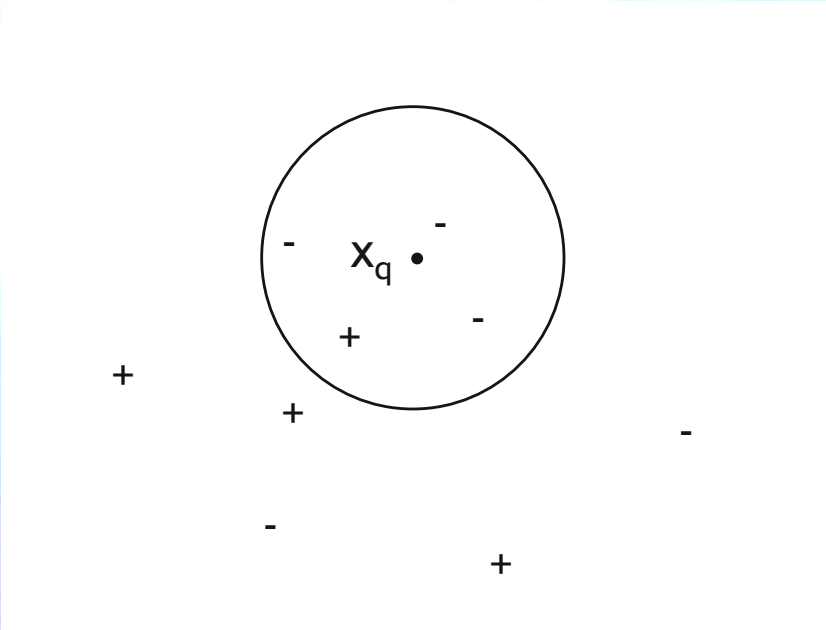
- Simple baseline: k-Nearest Neighbor

Leaving out algorithms/approaches for applicability domain and validation for now

OpenTox

OpenTox Algorithms: Classification / SAR

- Simple baseline: k-Nearest Neighbor



Leaving out algorithms/approaches for applicability domain and validation for now

OpenTox

OpenTox Algorithms: Classification / SAR

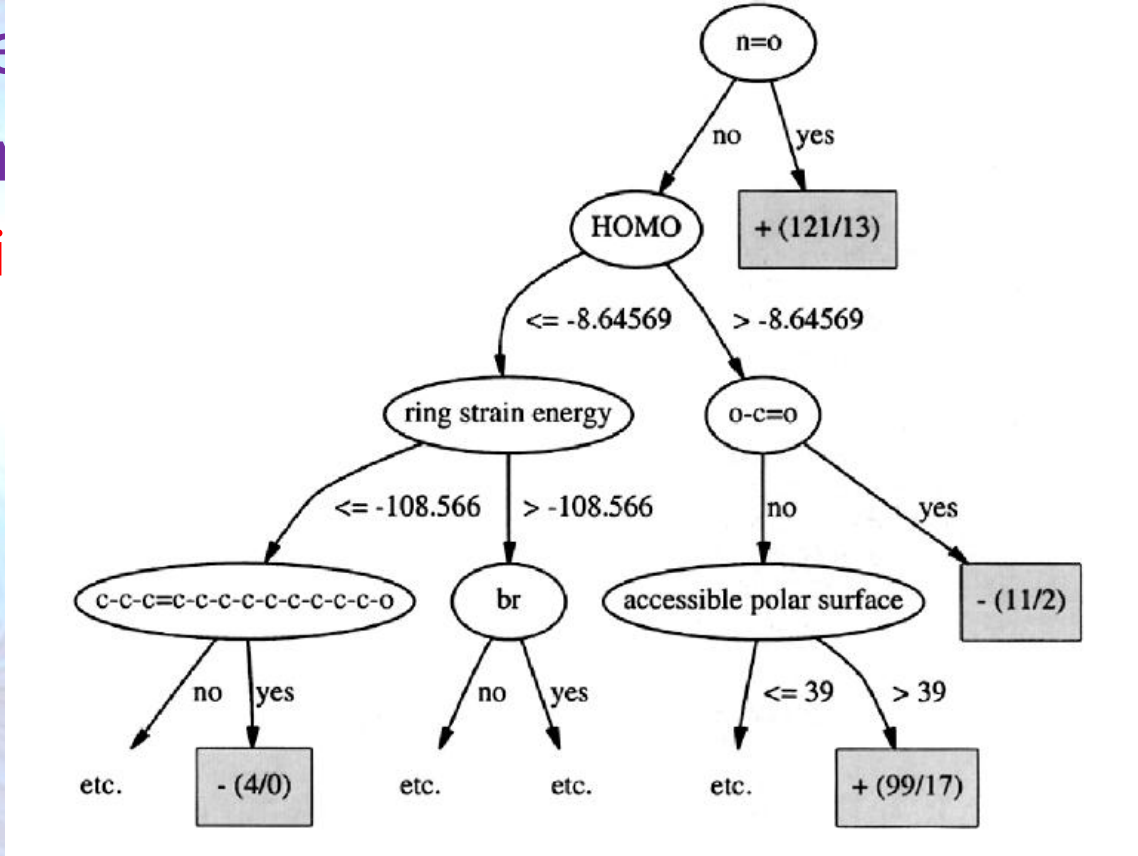
- Simple baseline: k-Nearest Neighbor
- Machine learning algorithms:
  - decision trees (J48)

Leaving out algorithms/approaches for applicability domain and validation for now

OpenTox

OpenTox Algorithms: Classification / SAR

- Simple
- Machi
  - deci



Leaving out algorithms/approaches for applicability domain and validation for now

OpenTox

OpenTox Algorithms: Classification / SAR

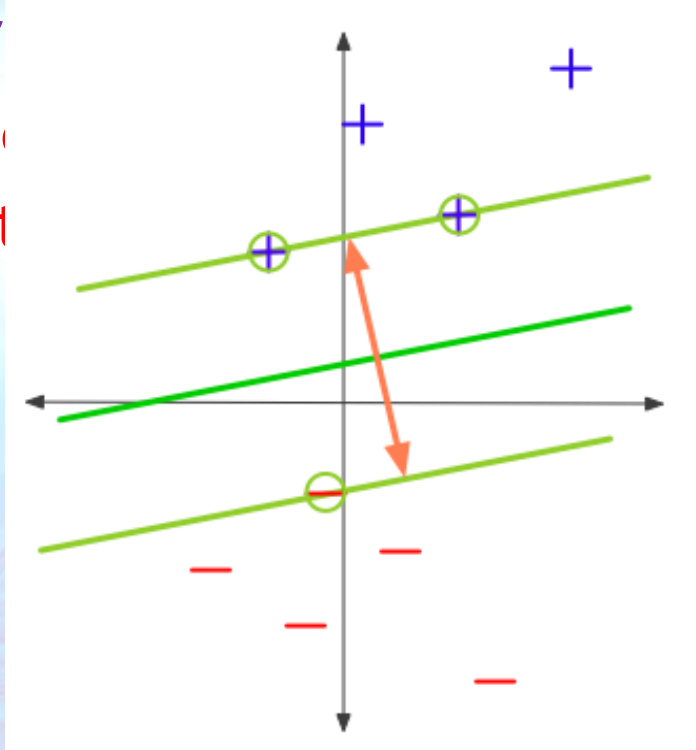
- Simple baseline: k-Nearest Neighbor
- Machine learning algorithms:
  - decision trees (J48)
  - support vector machines (SVMs)

Leaving out algorithms/approaches for applicability domain and validation for now

OpenTox

OpenTox Algorithms: Classification / SAR

- Simple baseline: k-Nearest Neighbor
- Machine lear
  - decision tre
  - support vect



Leaving out algorithms/approaches for applicability domain and validation for now

OpenTox

OpenTox Algorithms: Classification / SAR

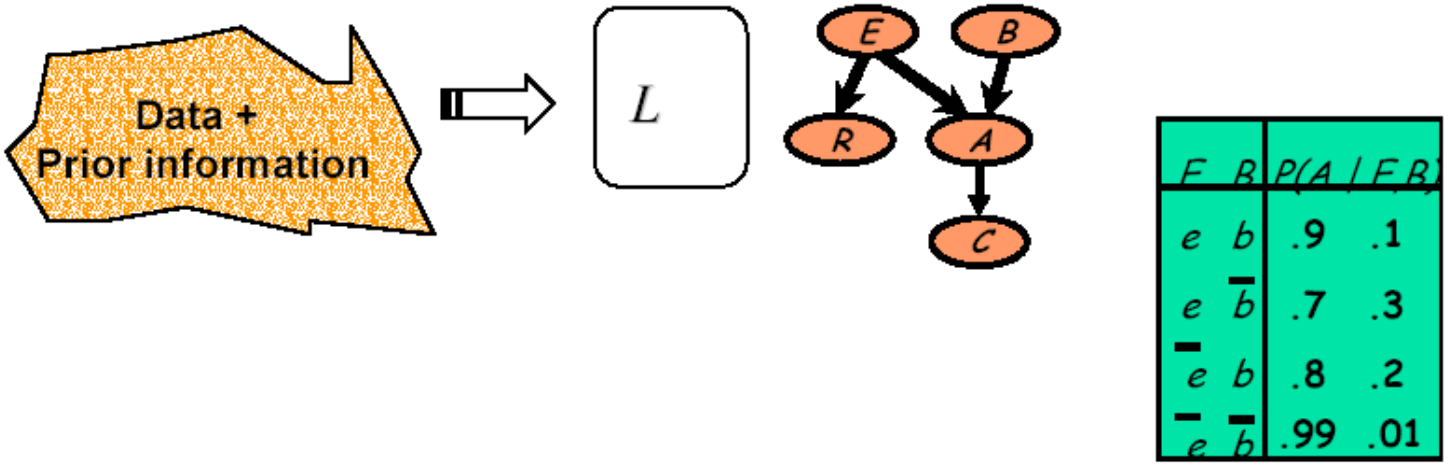
- Simple baseline: k-Nearest Neighbor
- Machine learning algorithms:
  - decision trees (J48)
  - support vector machines (SVMs)
- Probabilistic/graphical models
  - Bayesian network

Leaving out algorithms/approaches for applicability domain and validation for now

OpenTox

OpenTox Algorithms: Classification / SAR

- Simple baseline: k-Nearest Neighbor
- Machine learning algorithms:
  - decision trees (J48)
-



Leaving out algorithms/approaches for applicability domain and validation for now

OpenTox

OpenTox Algorithms: Classification / SAR

- Simple baseline: k-Nearest Neighbor
- Machine learning algorithms:
  - decision trees (J48)
  - support vector machines (SVMs)
- Probabilistic/graphical models
  - Bayesian network

Leaving out algorithms/approaches for applicability domain and validation for now

OpenTox

OpenTox Algorithms: Regression / QSAR

- Simple baseline: k-Nearest Neighbor
- Classical statistical algorithms:
  - multiple linear regression (MLR)
  - partial least squares (PLS)
- Machine learning algorithms:
  - model trees (M5')

OpenTox

OpenTox Algorithms: Regression / QSAR

- Simple baseline: k-Nearest Neighbor
- Classical statistical algorithms:
  - multiple linear regression (MLR)
  -
- Ma
  -

```
log_fluence <= -6.01 :
|   log_hr321 <= -0.112 : LM1
|   log_hr321 >  -0.112 : LM2
log_fluence >  -6.01 :
|   log_hr321 <= 0.0846 : LM3
|   log_hr321 >  0.0846 : LM4

LM1:  log_t90 = -0.879 + 0.0353log_hr321 - 0.373log_fluence
                    + 0.0394mfbmfr_class=inter,long + 0.0327mfbmfr_class=long
LM2:  log_t90 = 0.00965 - 0.138log_hr321 - 0.203log_fluence
                    + 0.0394mfbmfr_class=inter,long + 0.0327mfbmfr_class=long
```

OpenTox

OpenTox Algorithms: Regression / QSAR

- Simple baseline: k-Nearest Neighbor
- Classical statistical algorithms:
  - multiple linear regression (MLR)
  - partial least squares (PLS)
- Machine learning algorithms:
  - model trees (M5')
  - support vector regression
- Probabilistic/graphical models:
  - Gaussian process regression

OpenTox

OpenTox API for Algorithms

Description	Method	URI	Parameters	Result	Status codes
Get URIs of all available algorithms	GET	/algorithm	(optional) ?sameas=URI-of-the-owl:sameAs-entry	List of all algorithm URIs or RDF representation, or algorithms of specific types, if query parameter exists. Returns all algorithms, for which owl:sameAs is given by the query.	200,404,503
Get the ontology representation of an algorithm	GET	/algorithm/{id}	-	Algorithm representation in one of the supported MIME types.	200,404,503
Apply the algorithm	POST	/algorithm/{id}	**dataset_uri** parameter **prediction_feature**, more to be specified and documented by algorithm provider **dataset_service**=dataset serviceuri	*model URI* *dataset URI* *featureURI* Redirect to task URI for time consuming computations.	200,303,404,503

OpenTox

GET http://opentox.informatik.tu-muenchen.de:8080/OpenTox-dev/algorithm/

TUM OpenTox REST web services - Mozilla Firefox

Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe

http://opentox.informatik.tu-muenchen.de:8080/OpenTox-dev/algorithm/

Meistbesuchte Seiten Erste Schritte Aktuelle Nachrichten news.ORF.at sport.ORF.at SPIEGEL ONLINE - Nac... derStandard.at sueddeutsche.de Topt... SPIEGEL ONLINE - Uni...

TUM OpenTox REST web services

Technische Universität München

TUM - OpenTox - REST services 1.1

This site and TUM OpenTox REST web services are under development!
The full API can be found on the opentox.org website

Available algorithms:

- http://opentox.informatik.tu-muenchen.de:8080/OpenTox-dev/algorithm/kNNclassification
- http://opentox.informatik.tu-muenchen.de:8080/OpenTox-dev/algorithm/J48
- http://opentox.informatik.tu-muenchen.de:8080/OpenTox-dev/algorithm/kNNregression
- http://opentox.informatik.tu-muenchen.de:8080/OpenTox-dev/algorithm/PLSregression
- http://opentox.informatik.tu-muenchen.de:8080/OpenTox-dev/algorithm/M5P
- http://opentox.informatik.tu-muenchen.de:8080/OpenTox-dev/algorithm/GaussP
- http://opentox.informatik.tu-muenchen.de:8080/OpenTox-dev/algorithm/LR
- http://opentox.informatik.tu-muenchen.de:8080/OpenTox-dev/algorithm/BayesNet
- http://opentox.informatik.tu-muenchen.de:8080/OpenTox-dev/algorithm/FTM
- http://opentox.informatik.tu-muenchen.de:8080/OpenTox-dev/algorithm/gSpan
- http://opentox.informatik.tu-muenchen.de:8080/OpenTox-dev/algorithm/FTM/{smiles}
- http://opentox.informatik.tu-muenchen.de:8080/OpenTox-dev/algorithm/gSpan/{smiles}
- http://opentox.informatik.tu-muenchen.de:8080/OpenTox-dev/algorithm/CDKPhysChem
- http://opentox.informatik.tu-muenchen.de:8080/OpenTox-dev/algorithm/JOELIB2
- http://opentox.informatik.tu-muenchen.de:8080/OpenTox-dev/algorithm/InfoGainAttributeEval
- http://opentox.informatik.tu-muenchen.de:8080/OpenTox-dev/algorithm/PrincipalComponents
- http://opentox.informatik.tu-muenchen.de:8080/OpenTox-dev/algorithm/ChiSquared

Initial documentation

k-NN Classification

Mozilla Firefox

Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe

http://opentox.informatik.tu-muenchen.de:8080/OpenTox-dev/algorithm/kNNclassification

Meistbesuchte Seiten Erste Schritte Aktuelle Nachrichten news.ORF.at sport.ORF.at SPIEGEL ONLINE - Nac... derStandard.at sueddeutsche.de Topt... SPIEGEL ONLINE - Uni...

http://opentox.inf.../kNNclassification


```
- <rdf:RDF>
    <owl:Class rdf:about="http://www.opentox.org/api/1.1#Algorithm"/>
    <owl:Class rdf:about="http://www.opentox.org/api/1.1#Parameter"/>
  - <ot:Algorithm rdf:about="http://opentox.informatik.tu-muenchen.de:8080/OpenTox-dev/algorithm/kNNclassification">
      <dc:contributor>joerg.wicker@in.tum.de</dc:contributor>
    - <ot:parameters>
      - <ot:Parameter>
          <dc:title rdf:datatype="http://www.w3.org/2001/XMLSchema#string">nearestNeighbourSearchAlgorithm</dc:title>
        - <dc:description rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
            The nearest neighbour search algorithm to use (Default: weka.core.neighboursearch.LinearNNSearch).
          </dc:description>
          <ot:paramScope rdf:datatype="http://www.w3.org/2001/XMLSchema#string">optional</ot:paramScope>
          <ot:paramValue rdf:datatype="http://www.w3.org/2001/XMLSchema#string">LinearNNSearch</ot:paramValue>
        </ot:Parameter>
      </ot:parameters>
    - <ot:isA>
        http://www.opentox.org/algorithms.owl#ClassificationLazySingleTarget
      </ot:isA>
    - <ot:parameters>
      - <ot:Parameter>
          <dc:title rdf:datatype="http://www.w3.org/2001/XMLSchema#string">crossValidate</dc:title>
        - <dc:description rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
            Whether hold-one-out cross-validation will be used to select the best k value
          </dc:description>
          <ot:paramScope rdf:datatype="http://www.w3.org/2001/XMLSchema#string">optional</ot:paramScope>
          <ot:paramValue rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean">false</ot:paramValue>
        </ot:Parameter>
      </ot:parameters>
      <dc:title rdf:datatype="http://www.w3.org/2001/XMLSchema#string">kNNclassification</dc:title>
      <dc:creator>tobias.girschick@in.tum.de</dc:creator>
```


OpenTox

k-NN Classification

Mozilla Firefox

Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe

http://opentox.informatik.tu-muenchen.de:8080/OpenTox-dev/algorithm/kNNclassification

Meistbesuchte Seiten | Erste Schritte | Aktuelle Nachrichten | news.ORF.at | sport.ORF.at | SPIEGEL ONLINE - Nac... | derStandard.at | sueddeutsche.de Topt... | SPIEGEL ONLINE - Uni...

http://opentox.inf.../kNNclassification


```
  - <ot:parameters>
    - <ot:Parameter>
        <dc:title rdf:datatype="http://www.w3.org/2001/XMLSchema#string">windowSize</dc:title>
      - <dc:description rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
          Gets the maximum number of instances allowed in the training pool. The addition of new instances above this value will result in old instances being removed. A value of 0 signifies no limit to the number of training instances.
        </dc:description>
        <ot:paramScope rdf:datatype="http://www.w3.org/2001/XMLSchema#string">optional</ot:paramScope>
        <ot:paramValue rdf:datatype="http://www.w3.org/2001/XMLSchema#int">0</ot:paramValue>
      </ot:Parameter>
    </ot:parameters>
    <dc:date rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">Sun Sep 19 08:26:31 CEST 2010</dc:date>
  - <ot:parameters>
    - <ot:Parameter>
        <dc:title rdf:datatype="http://www.w3.org/2001/XMLSchema#string">dataset_uri</dc:title>
        <dc:description rdf:datatype="http://www.w3.org/2001/XMLSchema#string">URI to the dataset to be used</dc:description>
        <ot:paramScope rdf:datatype="http://www.w3.org/2001/XMLSchema#string">mandatory</ot:paramScope>
        <ot:paramValue rdf:datatype="http://www.w3.org/2001/XMLSchema#string"/>
      </ot:Parameter>
    </ot:parameters>
  - <ot:parameters>
    - <ot:Parameter>
        <dc:title rdf:datatype="http://www.w3.org/2001/XMLSchema#string">dataset_service</dc:title>
        <dc:description rdf:datatype="http://www.w3.org/2001/XMLSchema#string">URI to the dataset service to be used</dc:description>
        <ot:paramScope rdf:datatype="http://www.w3.org/2001/XMLSchema#string">optional</ot:paramScope>
        <ot:paramValue rdf:datatype="http://www.w3.org/2001/XMLSchema#string">http://apps.ideaconsult.net:8080/ambit2/dataset</ot:paramValue>
      </ot:Parameter>
    </ot:parameters>
  </ot:Algorithm>
</rdf:RDF>
```

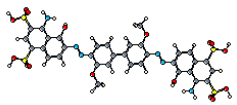
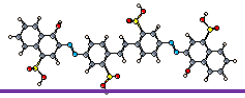
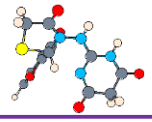
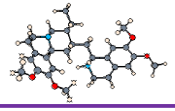
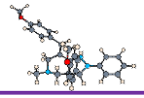
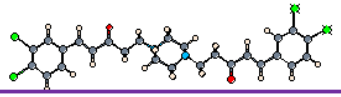

OpenTox

Development of Novel Algorithms

- Substructure / fragment generation algorithms („product line“)
  - FreeTreeMiner
  - BBRCs (backbone refinement classes)
  - LastPM (latent structure pattern mining)
- Structural clustering and local models
- Fast conditional density estimation for QSAR:
  - quantifying uncertainty in QSAR, confidence intervals, …

OpenTox

Substructure/Fragment Generation

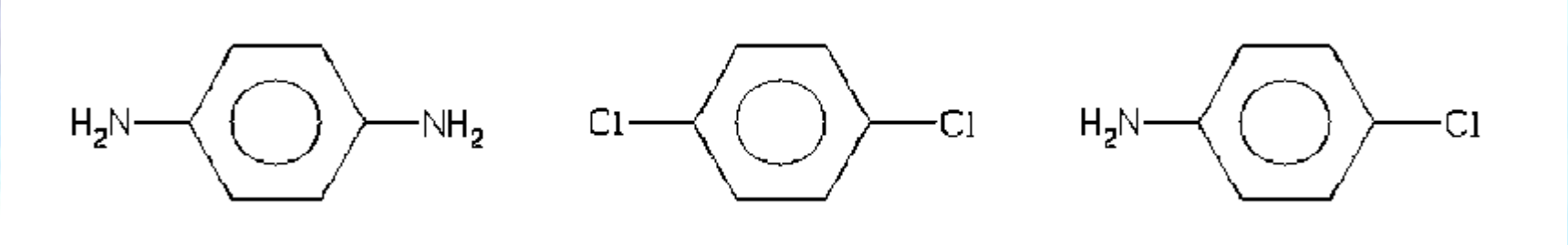
$c_{A,1}$		CA
$c_{A,2}$		CA
$c_{A,3}$		CA
…	…	…
$c_{I,1}$		CI
$c_{I,2}$		CI
$c_{I,3}$		CI
…	…	…

CA: confirmed active
CI: confirmed inactive

- First step: computation of *descriptors*
- Computed physico-chemical properties?
- Predefined functional groups?
- …?

OpenTox

Substructure/Fragment Generation



- *Path* patterns
- Minimum frequency = 2
- Just *most specific patterns* used here

x1 $=_{def}$ N-c:c:c:c:c:c

x2 $=_{def}$ Cl-c:c:c:c:c:c

x1	*x2*	*Class*
true	false	-
false	true	+
true	true	+

Substructure/Fragment Generation

Statistical learning schemes like SVMs are very good at combining substructures as features into (Q)SAR models

$$f(\mathbf{x}) = \mathrm{sgn}\left(\sum_{j}^{m} \gamma_j x_j + c\right) =$$

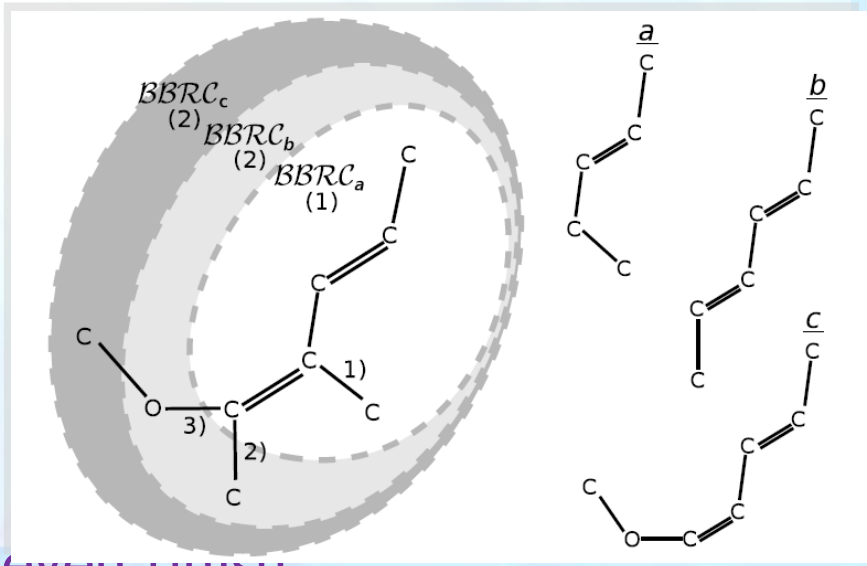
sgn(+1.63 * ‘c:c:c:c:c:c:c:c:c‘(x)
+1.44 * ‘C-Cl‘(x)
+1.32 * ‘C-C-C-C-N-C‘(x)
+1.31 * ‘C-C-C-O‘(x)
+0.95 * ‘C-C=C‘(x)
+0.87 * ‘c:c:c:c:c:n‘(x)
+0.82 * ‘C-C-C-C=C‘(x)
+0.82 * ‘C-C-C-N-C‘(x)
+0.80 * 'c:c:c-C=O‘(x)
+0.78 * ‘C-N-C‘(x)
+...

-1.48 * ‘Cl-C-Cl‘ (x)
-1.45 * ‘C-C-C=C-C’(x)
-1.01 * ‘C-N-c:c‘(x)
-1.01 * ‘C-N-c:c:c‘(x)
-0.95 * ‘C-C‘(x)
-0.95 * ‘C-C-N-C‘(x)
-0.94 * ‘C-O-C=O‘(x)
-0.94 * 'c:c:c:c:c:c-S‘(x)
-0.94 * 'c:c:c:c:c-S‘(x)
-0.94 * 'c:c:c:c-S‘(x)
- ...)

OpenTox

Scalability: A New, Practical Class of Substructures

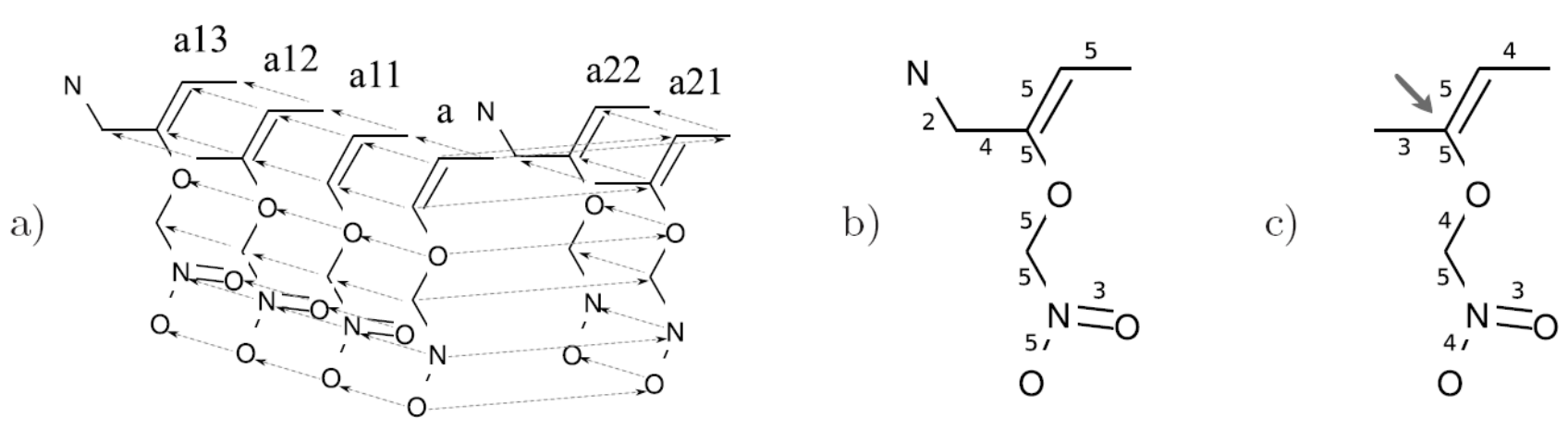
- A new, practical class of substructures: *backbone refinement classes (BBRC), i.e., trees sharing a common backbone*
- Then pick the most significant representative from this class
- > 23,000 compounds from NCI Yeast Anticancer Drug Screen data: BBRC representatives computed in 4m52s, other approaches did not even finish
- 87,264 possible; producing reasonable coverage of structures



OpenTox

Latent Structure Pattern Mining

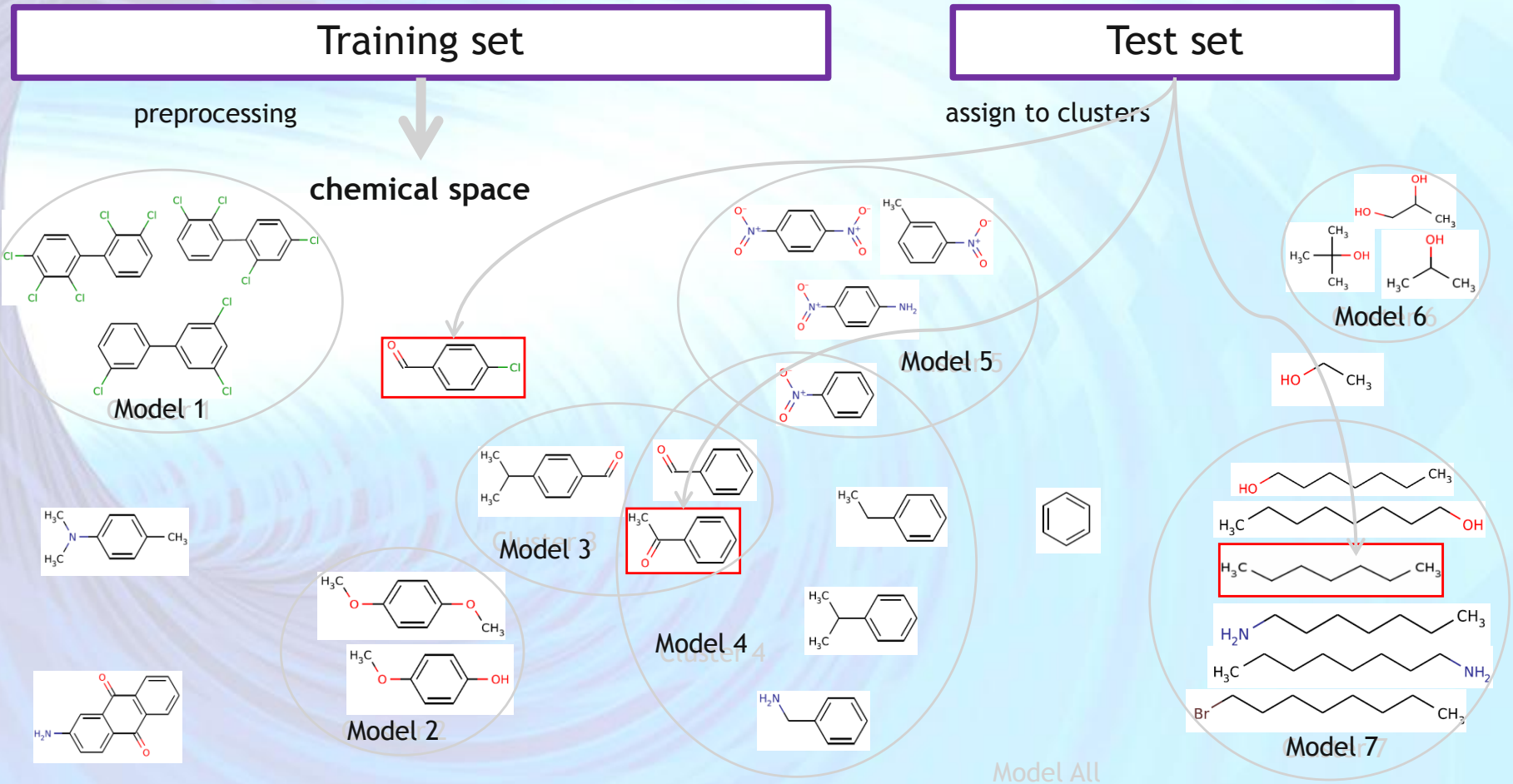
- Automatically discovering structural alerts
- 3 steps: (a) align, (b) stack, (c) compress
- Results for: blood-brain barrier, bioavailability, …



A. Maunz *et al.* (2010) in: *Proc. ECML/PKDD 2010.*

OpenTox

Structural Clustering and Local Models

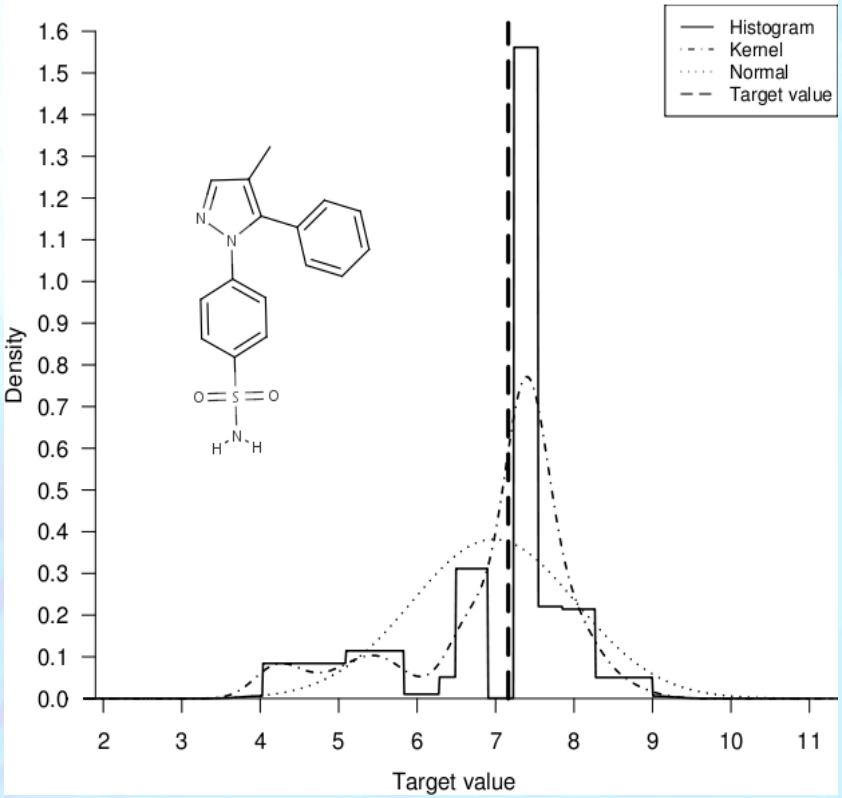


M. Seeland *et al.* (2010) in: *Proc. ECML/PKDD 2010.*
F. Buchwald *et al.* (2010) in: *Proc. EuroQSAR 2010.*

OpenTox

Fast Conditional Density Estimation for QSARs

- Prediction of distribution of activities
  - not point estimates
  - quantifying uncertainty
- Doing it fast…
- …using general purpose machine learning as plug-in
- Then use histogram estimator, Normal estimator, Kernel estimator



F. Buchwald *et al.* (2010) in: *Proc. AAAI 2010.*

Summary

- Algorithms: descriptor calculation, feature selection, classification (SAR) and regression (QSAR), …
- Simple API for algorithms
- Development of useful novel algorithms:
  - substructure generation, structural clustering, local models, fast conditional density estimation, multi-label classification, …

OpenTox

Development and Use of Predictive Toxicology Applications

An OpenTox Workshop

19 Sep 2010, Rhodes, Greece

Validation

presented by Haralambos Sarimveis
(National Technical University of Athens, Greece)

OpenTox

SEVENTH FRAMEWORK PROGRAMME

Use of QSARs under REACH (Annex IX)

- Acceptance of QSAR results - BOTH positive and negative results will be accepted if
  - Models have been validated
  - Models are adequately documented and meet acceptance criteria for a given application- “fit for purpose” concept

OpenTox

SEVENTH FRAMEWORK PROGRAMME

Compelling Needs of Users

Integrated Testing

in silico

in vitro

TTC

Read Across

Category Formation

REACH Reporting (QPRF, QMRF)

Applicability Domain

Validation

Human Data

Communicated to OpenTox 2009 by Grace Patlewicz (Du Pont)

OpenTox

OpenTox Framework - Standards

Validation

Algorithm Validation

- common best practices such as k-fold cross validation, leave-one-out, scrambling

QSAR Validation (Model Validation)

- OECD Principles www.oecd.org/dataoecd/33/37/37849783.pdf
- QSAR Model Reporting Format (QMRF) qsardb.jrc.it/qmrf/help.html
- QSAR Prediction Reporting Format (QPRF) ecb.jrc.it/qsar/qsar-tools/qrf/QPRF_version_1.1.pdf

Reports

REACH

- Guidance on Information Requirements and Chemical Safety Assessment

Part F

- Chemicals Safety Report
- Appendix Part F guidance.echa.europa.eu/guidance_en.htm

OpenTox

OECD - The Organisation for Economic Cooperation and Development

- Intergovernmental Organisation grouping 30 industrialised countries, aiming to: support sustainable economic growth, boost employment, raise living standards, maintain financial stability, assist other countries' economic development, contribute to growth in world trade.
- The OECD works on global issues in different areas, such as economy, society, governance, development, finance, innovation, sustainability.
- In November 2004, the OECD member countries agreed on the principles for validating (Q)SAR models for their use in regulatory assessment of chemical safety. The agreed principles provide member countries with basis for evaluating regulatory applicability of (Q)SAR models and will contribute to their enhanced use for more efficient assessment of chemical safety.

OpenTox

	OECD Principle	To facilitate the consideration of a (Q)SAR model for regulatory purposes, it should be associated with the following information::
1	Defined Endpoint	Ensure clarity in the endpoint being predicted by a given model, since a given endpoint could be determined by different experimental protocols and under different experimental conditions.
2	Unambiguous Algorithm	Ensure transparency in the model algorithm that generates predictions of an endpoint from information on chemical structure and/or physicochemical properties. The issue of reproducibility of the predictions is covered by this Principle.
3	Defined Applicability Domain	(Q)SARs are inevitably associated with limitations in terms of the types of chemical structures, physicochemical properties and mechanisms of action for which the models can generate reliable predictions
4	Goodness-of-fit, robustness and predictivity	Internal performance of a model (as represented by goodness-of-fit and robustness) and the predictivity of a model (as determined by external validation).
5	Mechanistic interpretation (if possible)	The intent of Principle 5 is not to reject models that have no apparent mechanistic basis, but to ensure that some consideration is given to the possibility of a mechanistic association between the descriptors used in a model and the endpoint being predicted.

OpenTox

	OECD Principle	OpenTox addresses Validation Principles by...
1	Defined Endpoint	providing a unified source of well defined and documented toxicity data with a common vocabulary
2	Unambiguous Algorithm	providing transparent access to well documented models and algorithms as well as to the source code
3	Defined Applicability Domain	integrating tools for the determination of applicability domains during the validation of prediction models
4	Goodness-of-fit, robustness and predictivity	providing scientifically sound validation routines for the determination of errors and confidences
5	Mechanistic interpretation (if possible)	integrating tools for the prediction of toxicological mechanisms and the recording of opinions and analysis in reports

OpenTox

Goodness-of-fit, robustness and predictivity

- OpenTox is developing unified and objective validation routines for model and algorithm developers and for external (Q)SAR programs, including procedures for validation with artificial test sets (e.g. n-fold cross-validation, leave-one-out, simple training/test set splits, bootstrapping, Y-scrambling).
- An important goal is to integrate statistical tests for the comparison of (Q)SAR models under consideration, a versioned database to store validation results and their history, and tools for the inspection of the toxicological plausibility of (Q)SAR predictions.

OpenTox

Implemented validation algorithms

Classification methods

- Number of correctly classified instances
- Number of incorrectly classified instances
- weighted_area_under_roc
- f_measure
- num_false_positives, negatives
- num_true_positives, negatives
- sensitivity
- specificity
- Classification confusion matrix

Regression methods

- root_mean_squared_error
- mean_absolute_error
- sum_squared_error
- r_square
- correlation_coefficient

http://www.opentox.org/data/documents/development/validation/validation-statistics

OpenTox

Validation API

Description	Method	URI	Parameters	Result	Status codes
Get all validations	GET	/	-	List of validation URIs	200,404
Retrieves a validation representation	GET	/{id}	-	Validation representation in one of the supported MIME types	200,404
Validates a model on a test dataset	POST	/	model_uri test_dataset_uri test_target_dataset_uri (default = test_dataset_uri)	Validation URI or Task URI	200,400,404,500
Builds a model on a training dataset and validates it on a test dataset	POST	/	algorithm_uri prediction_feature algorithm_params (string, default="") training_dataset_uri test_dataset_uri test_target_dataset_uri (default = test_dataset_uri) y_scramble (boolean, default=false) y_scramble_seed (integer, default=1)	Validation URI or Task URI	200,400,404,500
Splits a dataset into training and test dataset according to a certain ratio, and performs a validation	POST	/training_test_split	algorithm_uri prediction_feature algorithm_params (string, default="") dataset_uri split_ratio(float, default=0.66) random_seed(integer, default=1) y_scramble (boolean, default=false) y_scramble_seed (integer, default=1)	Validation URI or Task URI	200,400,404,500
OPTIONAL: *Performs a bootstrap validation*	*POST*	*/bootstrap*	*algorithm_uri* *prediction_feature* *dataset_params (string, default="")* *dataset_uri* *bootstrap_percentage(float, default=0.66)* *random_seed(integer, default=1)* *y_scramble (boolean, default=false)* *y_scramble_seed (integer, default=1)*	*Validation URI or Task URI*	*200,400,404,500*
Deletes a validation.	DELETE	/{id}	-	-	200,404

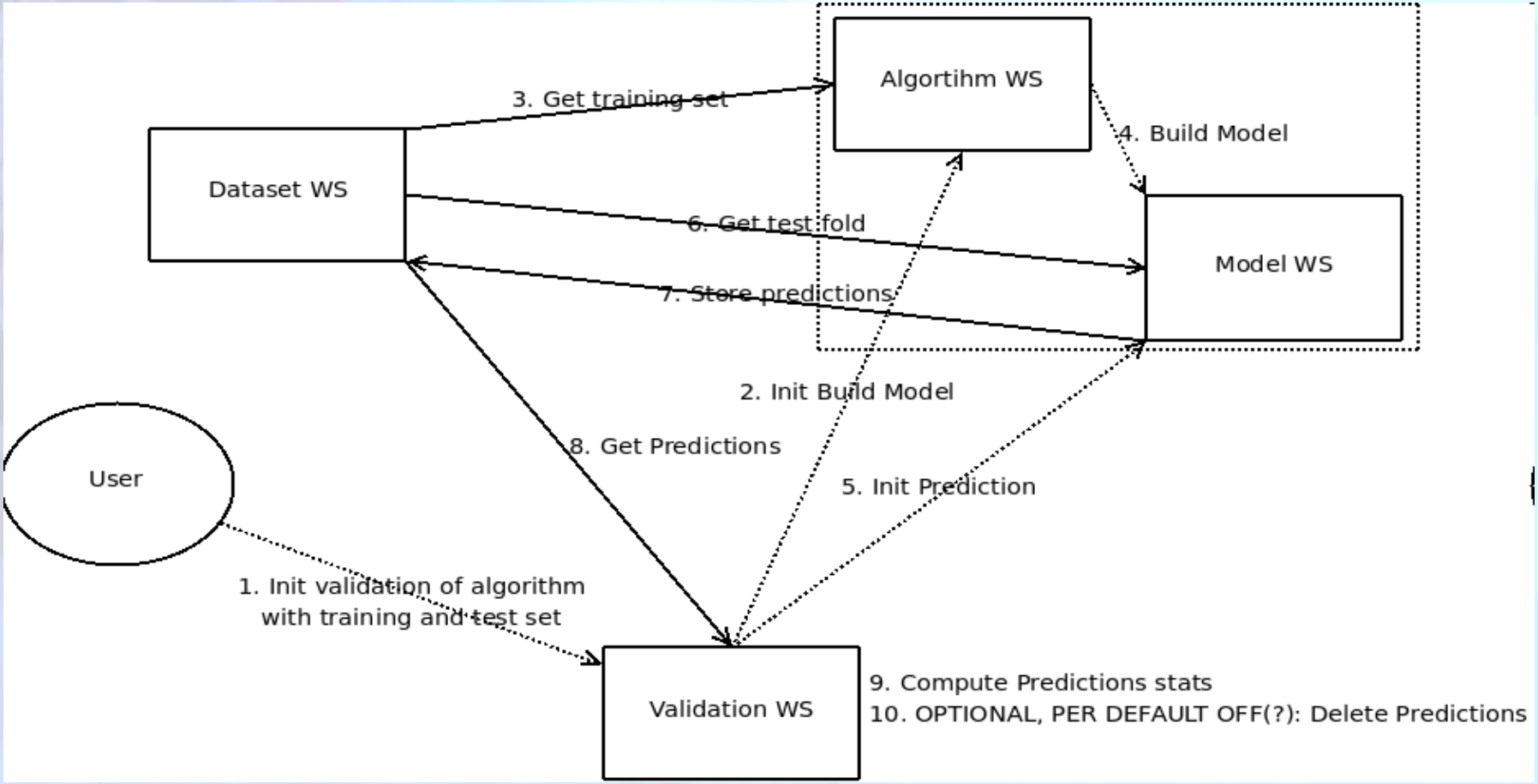
OpenTox

Cross-Validation API

Description	Method	URI	Parameters	Result	Status codes
Get all cross-validations	GET	/crossvalidation	-	List of crossvalidation URIs	200,404
Retrieves a cross-validation representation	GET	/crossvalidation/{id}	-	Cross-Validation in one of the supported MIME types	200,404
Returns all (k) validations that belong to a crossvalidation	GET	/crossvalidation /{id}/validations	-	List of validation URIs	200,404
Performs a k-fold cross-validation.	POST	/crossvalidation	algorithm_uri prediction_feature algorithm_params (string, default="") num_folds (integer, default=10) random_seed (integer, default=1) stratified (boolean, default=true) y_scramble (boolean, default=false) y_scramble_seed (integer, default=1)	Cross-Validation URI or Task URI	200,400,404,500
Performs a leave-one-out cross-validation.	POST	/crossvalidation/loo	algorithm_uri prediction_feature algorithm_params (string, default="") y_scramble (boolean, default=false) y_scramble_seed (integer, default=1)	Cross-Validation URI or Task URI	200,400,404,500
Deletes a cross-validation.	DELETE	/crossvalidation/{id}	-	-	200,404

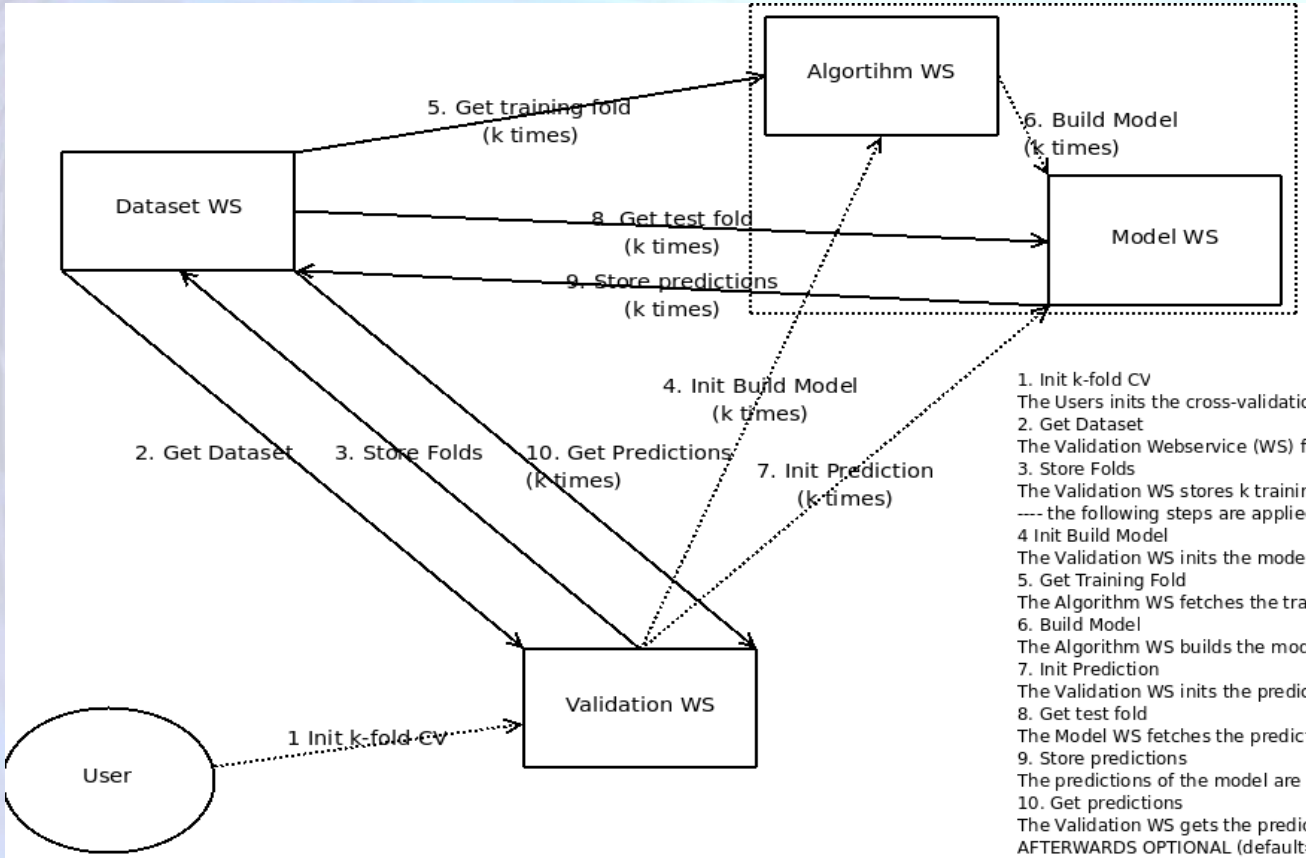
OpenTox

Validation WorkFlow



OpenTox

Cross-Validation WorkFlow



1. Init k-fold CV
The Users inits the cross-validation, params are alogrihtm, dataset, num-folds(=k), etc..
2. Get Dataset
The Validation Webservice (WS) fetches the dataset and splits it into different folds
3. Store Folds
The Validation WS stores k training and k test-datasets in the Dataset WS
---- the following steps are applied k times ----
4 Init Build Model
The Validation WS inits the model building
5. Get Training Fold
The Algorithm WS fetches the training fold
6. Build Model
The Algorithm WS builds the model, which is available at the Model WS
7. Init Prediction
The Validation WS inits the prediction
8. Get test fold
The Model WS fetches the prediction dataset, and predicts the compounds
9. Store predictions
The predictions of the model are stored in a dataset
10. Get predictions
The Validation WS gets the predictions and computes the validation stats
AFTERWARDS OPTIONAL (default=off):
The Validation WS deletes the predicition, training and test datasets

Applicability domain/confidence in prediction

- A definition of AD which is also used by the OECD is the following: "The applicability domain of a (Q)SAR model is the response and chemical structure space in which the model makes predictions with a given reliability."
- Furthermore, OECD advises that the AD principle should be applied in a model-specific manner. Thus, every model should be associated with its own AD derived not only on the chemicals in the training set but also on the descriptors and (statistical) approach used to develop the model. Ideally, the AD should be defined and documented by the model developer.
- Related to the concept of an AD is the concept of confidence in predictions inherent in some learning algorithm, so that the predictive model itself provides estimation of applicability domain. For example, classification algorithms do not only provide a categorical class label, but also a probability with which the class is predicted. The main difference is that the confidence is only known when the model is already applied, whereas the applicability domain is defined on the input space directly.

Implemented applicability domain algorithms

A. The predictive model itself provides estimation of applicability domain

- Lazar

B. Applicability domain is estimated by a procedure , separate from the predictive model

- PCA ranges
- Euclidean distance
- Cityblock distance
- Mahalanobis distance
- Nonparametric density estimation
- Leverage
- Fingerprints, Tanimoto distance

OpenTox

Future Work

- Validation with a test set that is completely unknown to the model developer is certainly the gold standard in this area, because there is no way to cheat voluntarily or involuntarily (e.g. by "optimizing" model parameters for a specific test set).
- OpenTox will provide facilities to access confidential (inhouse) data. For validation purposes we will provide a facility to test (Q)SAR models remotely against confidential datasets without getting access to the actual entries of the database to ensure security and confidentiality of proprietary data.
- Confidential validation data will be sought from external sources including members of the advisory board.
- OpenTox already provides facilities to protect confidential information located at URIs. Two tasks are involved here:
  - Authentication: Confirming the identity of the user requesting access
  - Authorisation: Granting the confirmed identity access according to a set of restrictions described in policies

OpenTox

Development and Use of Predictive Toxicology Applications

An OpenTox Workshop

19 Sep 2010, Rhodes, Greece

Reporting

presented by Andreas Karwath
(Albert-Ludwigs-Universität, Freiburg, Germany)

OpenTox

SEVENTH FRAMEWORK PROGRAMME

Compelling Needs of Users

Integrated Testing

in silico

in vitro

TTC

Read Across

Category Formation

REACH Reporting (QPRF, QMRF)

Applicability Domain

Validation

Human Data

Communicated to OpenTox 2009 by Grace Patlewicz (Du Pont)

OpenTox

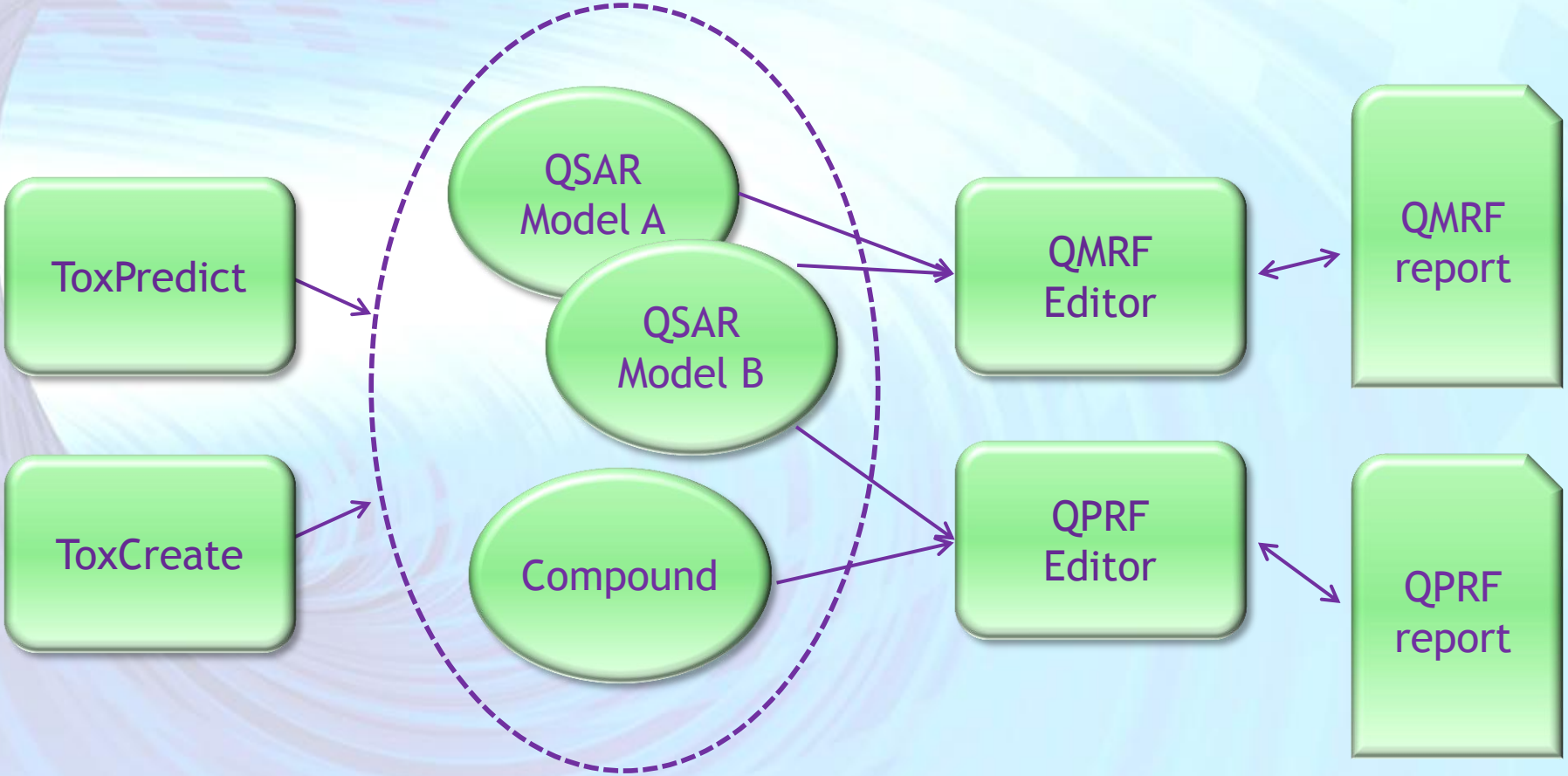
REACH reporting formats

QMRF and QPRF :

- What are they?
  - harmonized templates for summarizing and reporting key information on (Q)SAR models and predictions generated by these models
- Why is it important in OpenTox?
  - QMRF and QPRF are expected to be the communication tool between industry and the authorities under REACH.

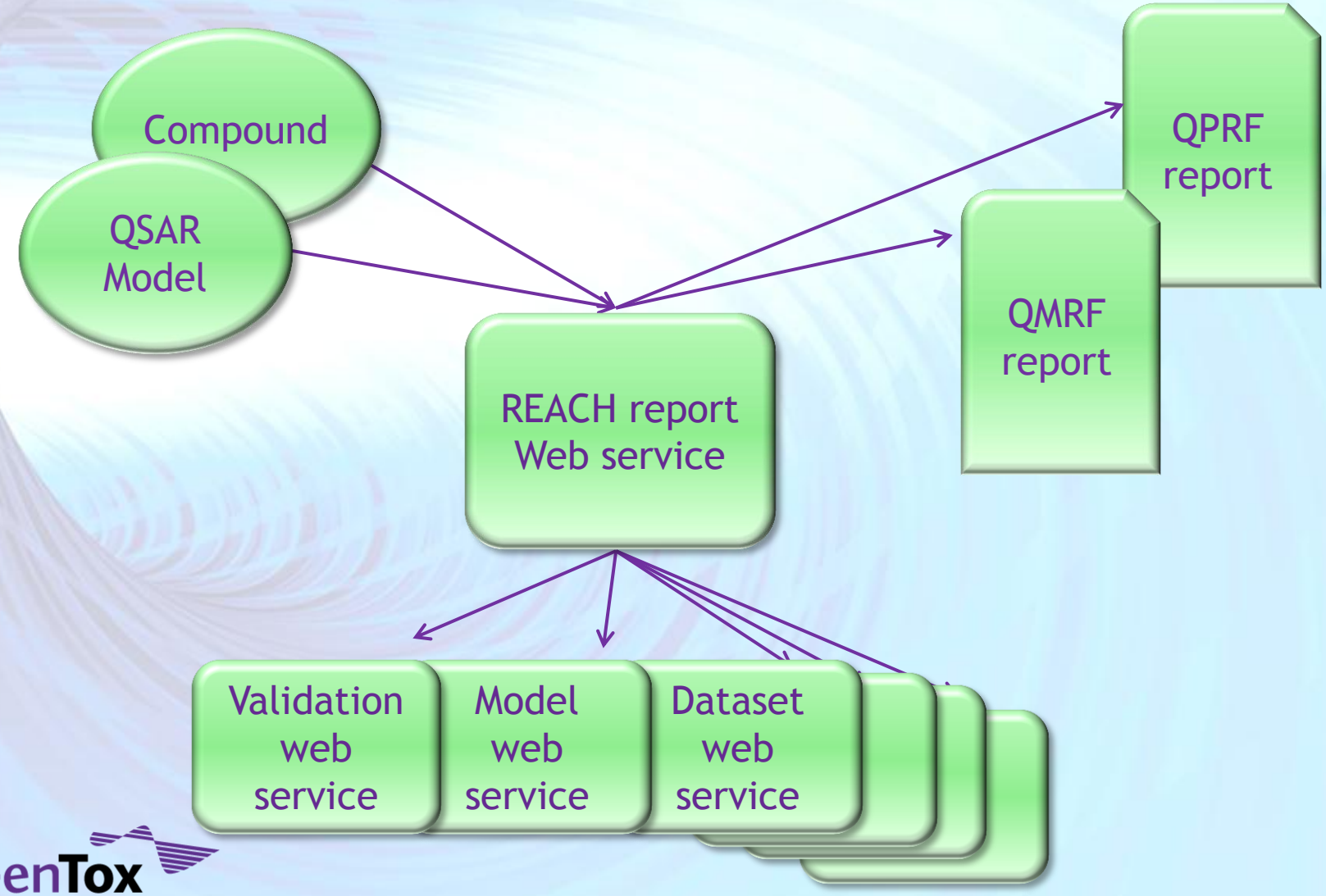
OpenTox

User Perspective



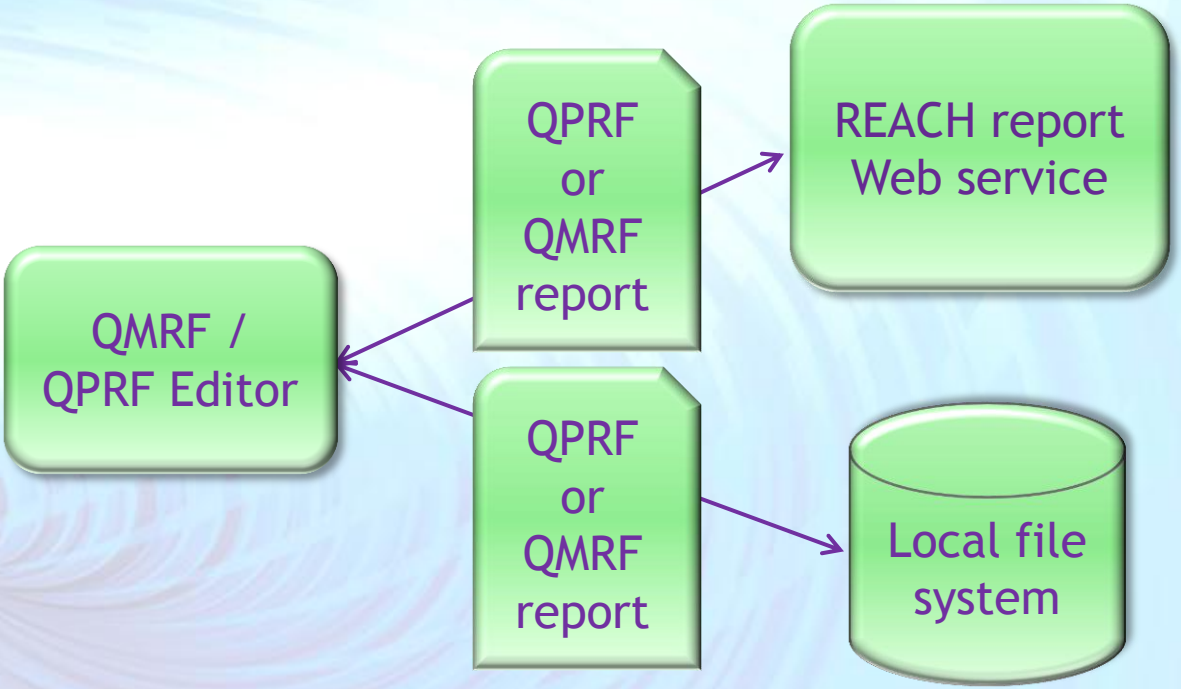
OpenTox

Creating Reports



OpenTox

Storing and Editing Reports



OpenTox

Implemented Reporting Services - QMRF

Report web service:

- Automatic generation of reports, including:
  - Meta-information (creation date, algorithm, model endpoint, …)
  - Model training data
  - Validation results (cross-validation, bootstrapping, …)
  - Prediction results on external test-data
- Reports can be downloaded/uploaded/deleted

OpenTox

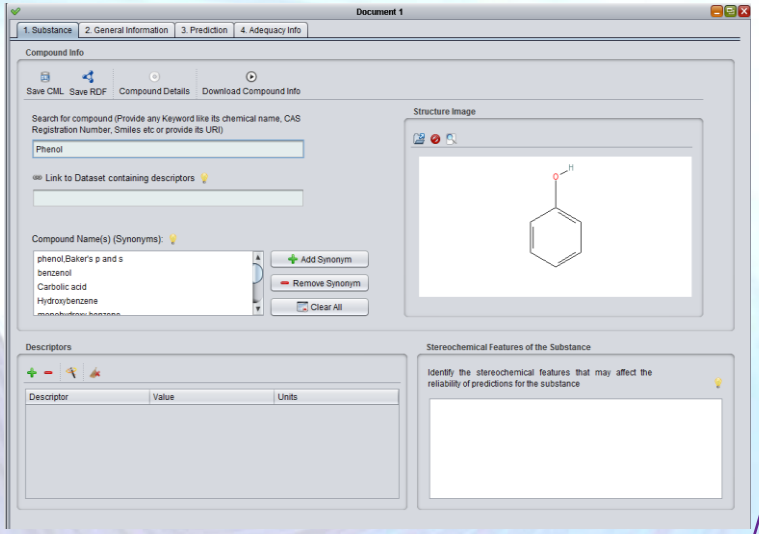
Implemented Reporting Services – QMRF [2]

QMRF Editor:

- Based on existing and EU approved implementation (see http://qsardb.jrc.it/qmrf)
- Comprehensive functionalities (edit/store reports, export to pdf)
- Extended to communicate with web service (download and upload reports)
- Embedded into ToxCreate

Implemented Reporting Services – QPRF

QPRF Editor (Q-Edit):



...Using AMBIT services

- Version 0.1.3 (alpha), heading towards the first beta version
- PDF Creation fully supported
- Compound lookup services facilitate users to find the compound they are looking for
- Similarity Search using the Q-Edit GUI
- Reports are (locally) stored in a binary format (RDF is under development)
- Available for download from http://github.com/alphaville/Q-edit

OpenTox

Future Work

QMRF Report Services and Editor:

- Include more automatically generated information:
  - Detailed description of model and algorithm
  - Related models
  - Authors
  - …
- Enable Authentication

QPRF Report Services and Editor:

- Establish web services for QPRF
- Enable Authentication

OpenTox

Final words...

For more information, visit

www.opentox.org

Contact Project Coordinator:

barry.hardy@douglasconnect.com

+41 61 851 0170

We welcome your involvement!

OpenTox - An Open Source Predictive Toxicology Framework, www.opentox.org, is funded under the EU Seventh Framework Program: HEALTH-2007-1.3-3 Promotion, development, validation, acceptance and implementation of QSARs (Quantitative Structure-Activity Relationships) for toxicology, Project Reference Number Health-F5-2008-200787 (2008-2011).

OpenTox

SEVENTH FRAMEWORK PROGRAMME