Development 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

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

Search Opentox RDF - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://apps.ideaconsult.net:8080/ontology/query/Endpoints

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

Search Opentox RDF

Features Algorithms Models Endpoints

Import RDF data into Ontology service

URL

SUBMIT

Ontology service 17969 triples

SPARQL

```
PREFIX ot:<http://www.opentox.org/api/1.1#>
PREFIX otee:<http://www.opentox.org/echaEndpoints.owl#>
        select ?Endpoints ?title ?id
        where {
        ?Endpoints rdfs:subClassOf otee:Endpoints.
    OPTIONAL {?Endpoints dc:title ?title}.
    OPTIONAL {?Endpoints dc:identifier ?id}.
        }
```

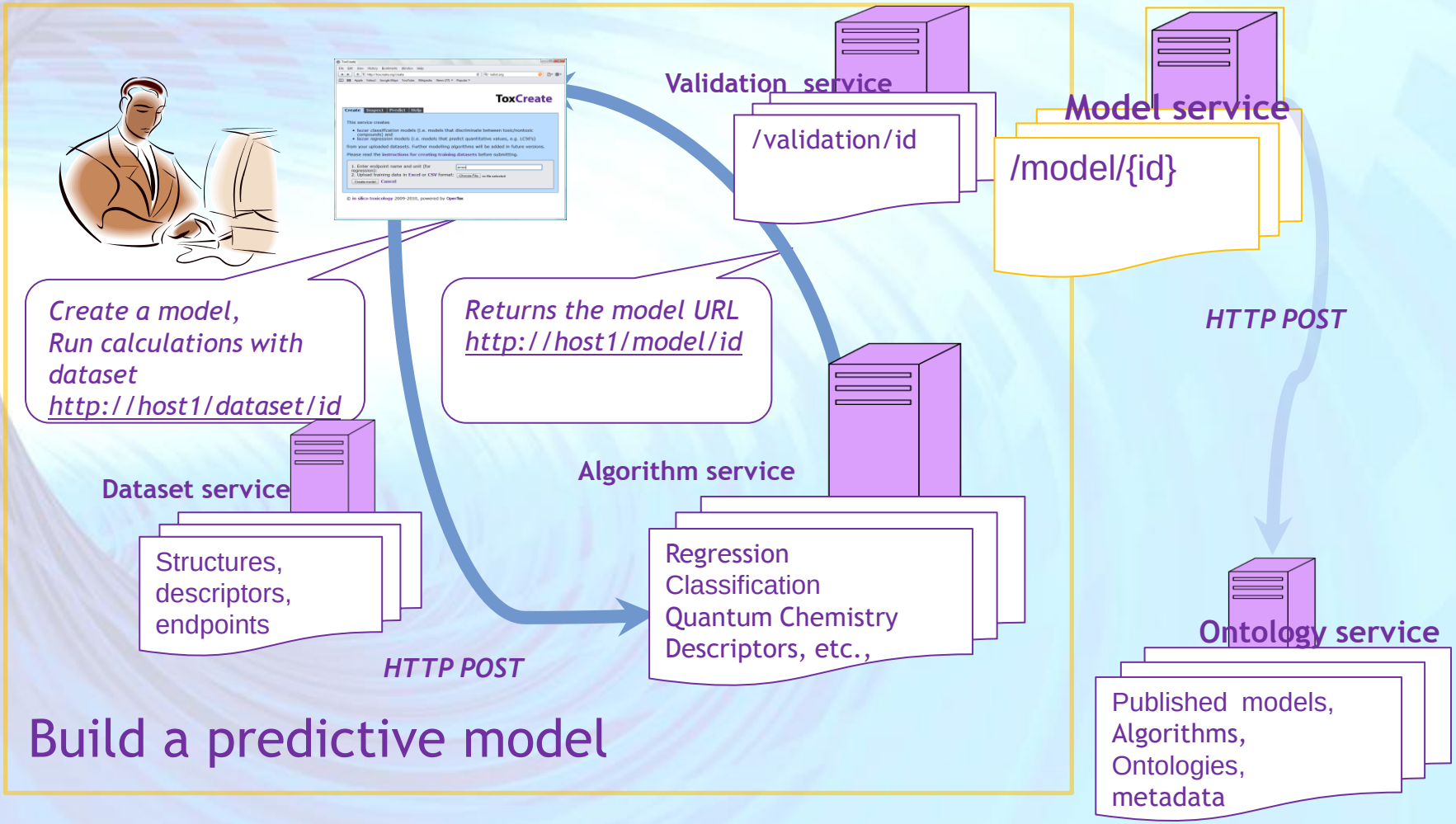
Submit Query

Results [found in 1 ms]

Endpoints	title	id
http://www.opentox.org/echaEndpoints.owl#PhysicoChemicalEffects	'Physicochemical effects '^^http://www.w3.org/2001/XMLSchema#string	'1'^^http://www.w3.org/2001/XMLSchema#string
http://www.opentox.org/echaEndpoints.owl#ToxicoKinetics	'Toxicokinetics '^^http://www.w3.org/2001/XMLSchema#string	'5'^^http://www.w3.org/2001/XMLSchema#string
http://www.opentox.org/echaEndpoints.owl#EcotoxicEffects	'Ecotoxic effects'^^http://www.w3.org/2001/XMLSchema#string	'3'^^http://www.w3.org/2001/XMLSchema#string
http://www.opentox.org/echaEndpoints.owl#EnvironmentalFateParameters	'Environmental fate parameters '^^http://www.w3.org/2001/XMLSchema#string	'2'^^http://www.w3.org/2001/XMLSchema#string
http://www.opentox.org/echaEndpoints.owl#HumanHealthEffects	'Human health effects'^^http://www.w3.org/2001/XMLSchema#string	'4'^^http://www.w3.org/2001/XMLSchema#string

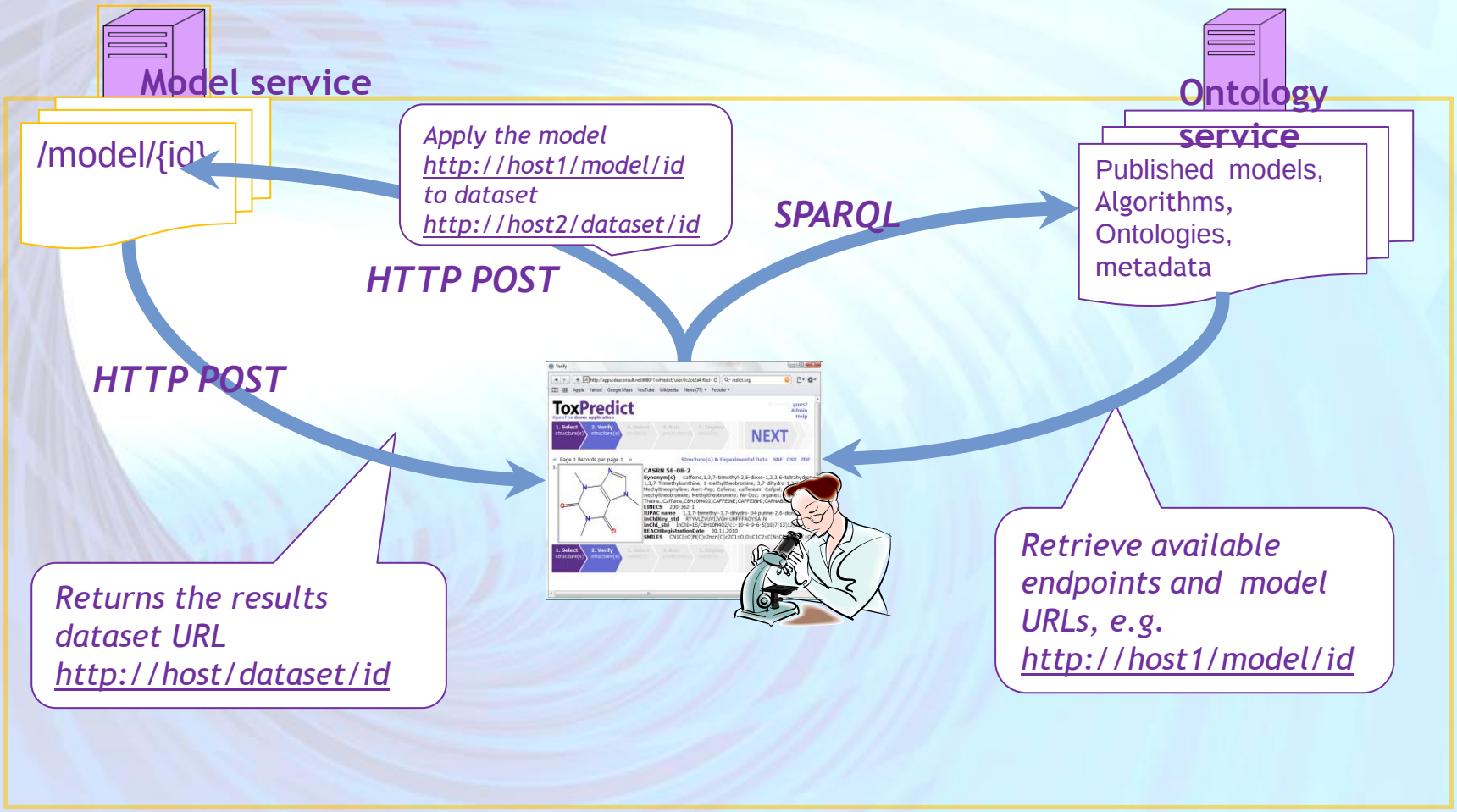
OpenTox

Build a predictive model



OpenTox

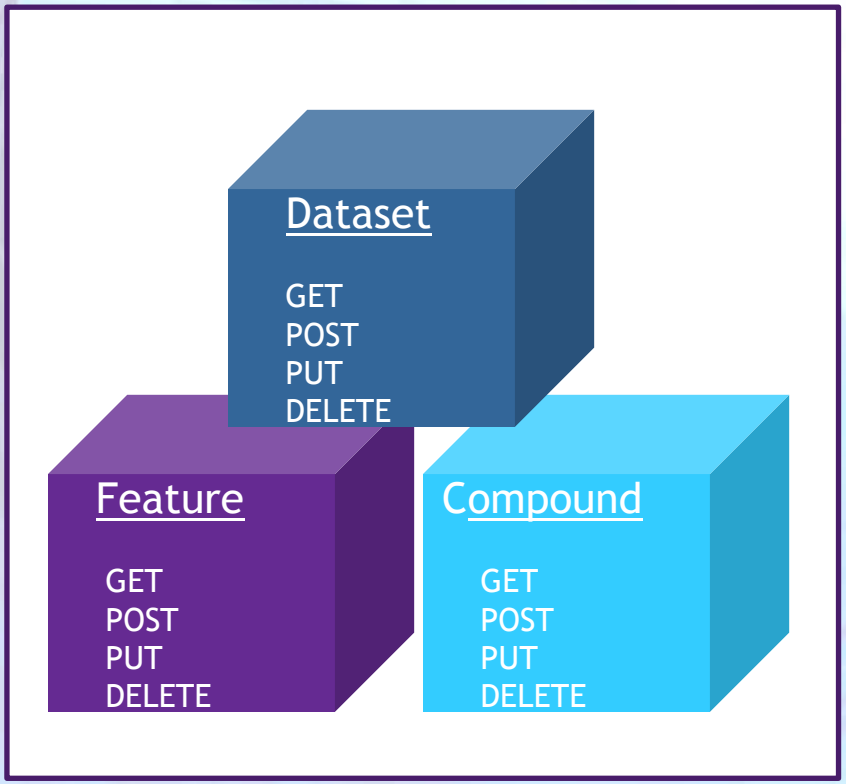
Apply predictive models

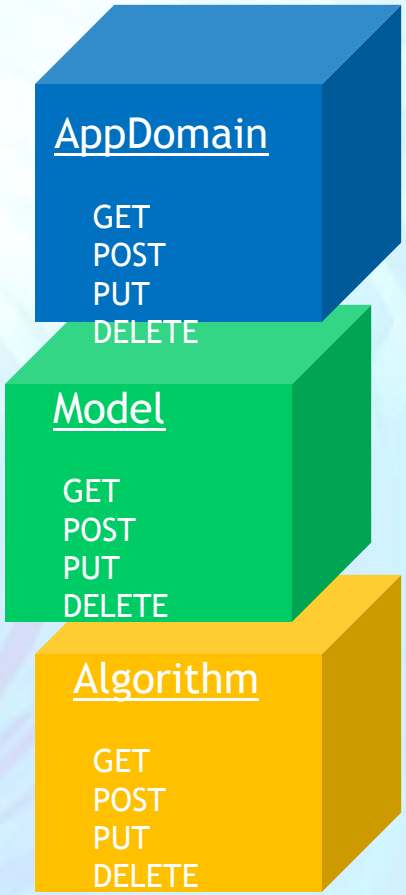


OpenTox

OpenTox API (Application Programming Interface)

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





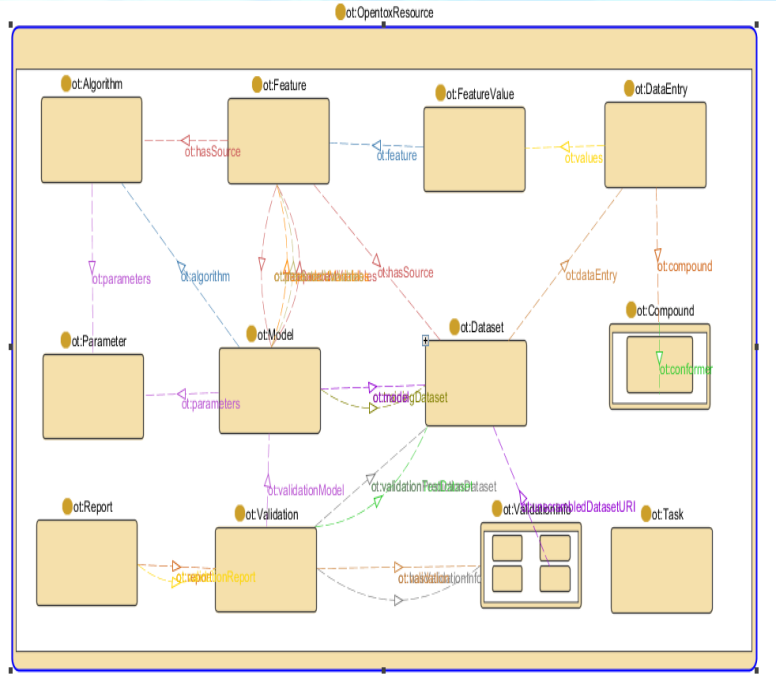
OpenTox

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
```

Compound

GET
POST
PUT
DELETE

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}
```

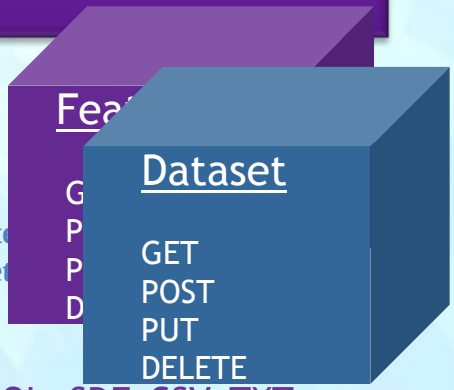
OpenTox

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)

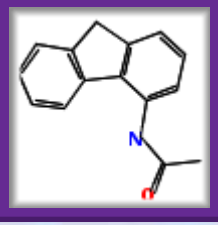
- **The dataset consists of data entries.**
- Each entry is associated with exactly **one chemical compound**, identified by its URI and available via OpenTox Compound service API;
- One and the same compound can be associated with multiple dataset entries;
- Every "column" is associated with a **Feature,**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

Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://apps.ideaconsult.net:8080/ToxPredict/user/admin/report/Stats?header=TRUE

cebs effects

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

http://apps.ideac...Stats?header=TRUE

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 Receptor 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	0	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

Find: issca Next Previous Highlight all Match case Phrase not found

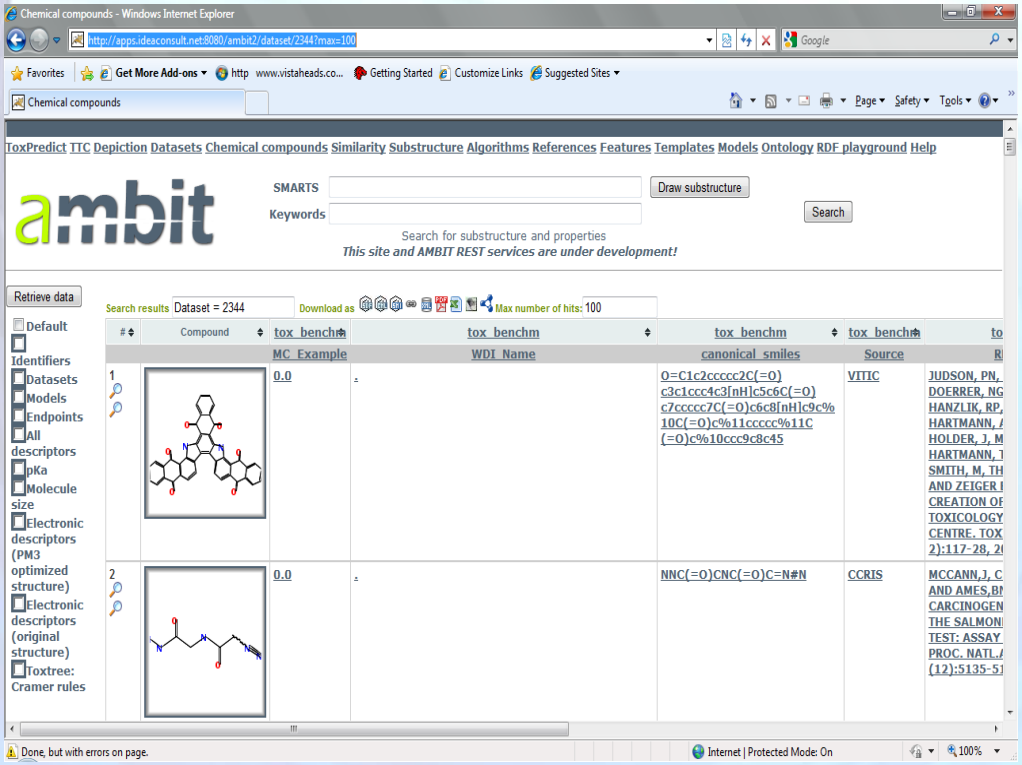
Done

Search results Dataset = 112.Dataset Download as Max number of hits: 100

Compound	ToxCast_To CHR_Mouse_Ureter_2_PreneoplasticLesion	ToxCast_To CHR_Mouse_Nose_1_AnyLesion	ToxCast_To CHR_Rat_Trachea_3_NeoplasticLesion	ToxCast_To CHR_Mouse_Pit
1	1000000.0	1000000.0	1000000.0	1000000.0
2	1000000.0	1000000.0	1000000.0	1000000.0
3	NA	NA	1000000.0	NA

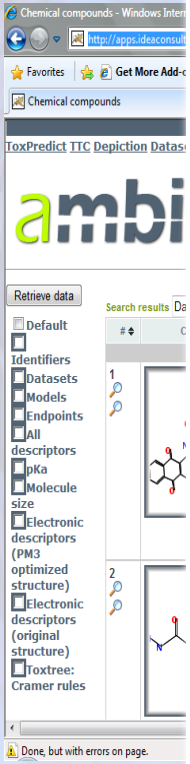
Example: mutagenicity dataset

- 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

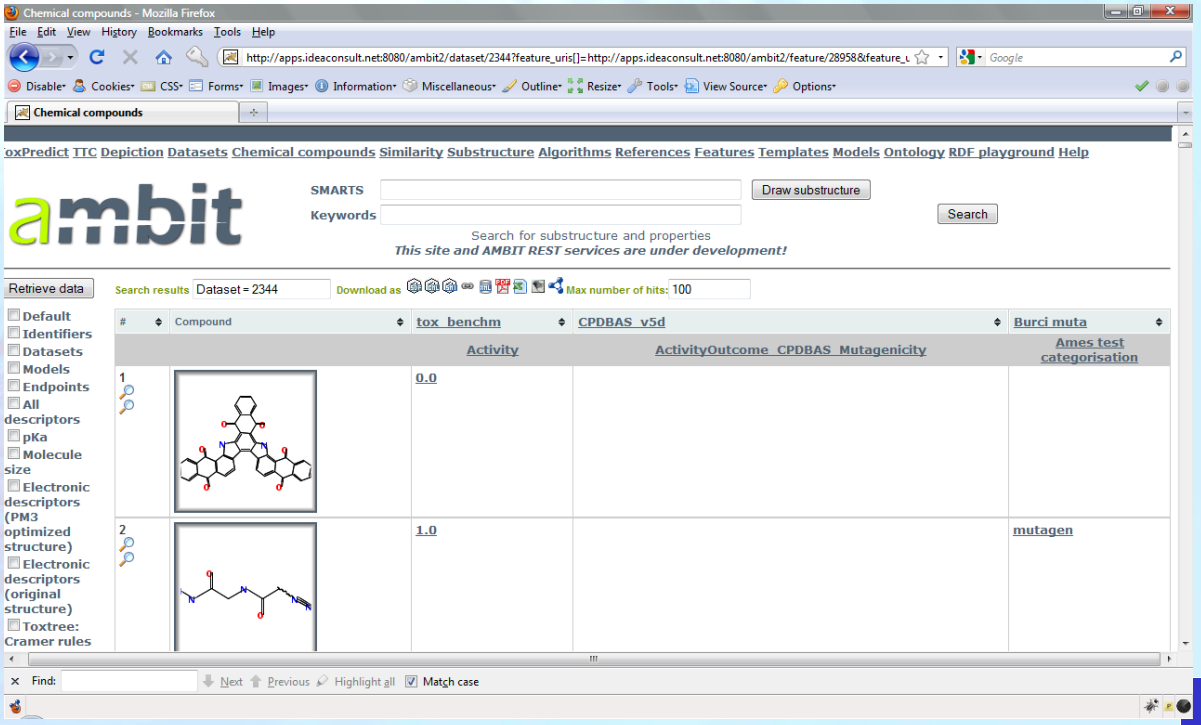
http://apps.ideaconsult.net:8080/ambit2/feature/28958

Feature

Find	Name				
	SAL		...endpoints.owl#Mutagenicity	ISSCAN_v3a_1153_19Sept08.1222179139.sdf	NO
	ActivityOutcome_CPDBAS_Mutagenicity		http://www.opentox.org /echaEndpoints.owl#Mutagenicity	CPDBAS_v5d_1547_20Nov2008.sdf	YES
	ActivityOutcome_CPDBAS_Mutagenicity		http://www.opentox.org	ambit2_75288.sdf	NO
	ActivityOutcome_CPDBAS_Mutagenicity				
	ActivityOutcome_CPDBAS_Mutagenicity				
	Mutagenic Activity in TA100 (3=active; 1=inactive)		...opentox.org /echaEndpoints.owl#Mutagenicity	qsar6train.csv	YES
	Ames test categorisation		http://www.opentox.org /echaEndpoints.owl#Mutagenicity	Burci mutagenicity dataset.sdf	YES
	Mutagenic Activity in TA100 (3=active; 1=inactive)		http://www.opentox.org	270.sdf	NO
	Mutagenic Activity in TA100 (3=ac 1=inactive)				
	Prediction feature for http://apps.ideaconsult.net:8080 /ambit2/feature/26701 endpoint prediction			/OpenTox-dev/model/10MOpenToxModel_j48_30	
	Activity		http://www.opentox.org /echaEndpoints.owl#Mutagenicity	tox_benchmark_N6512.sdf	NO

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.

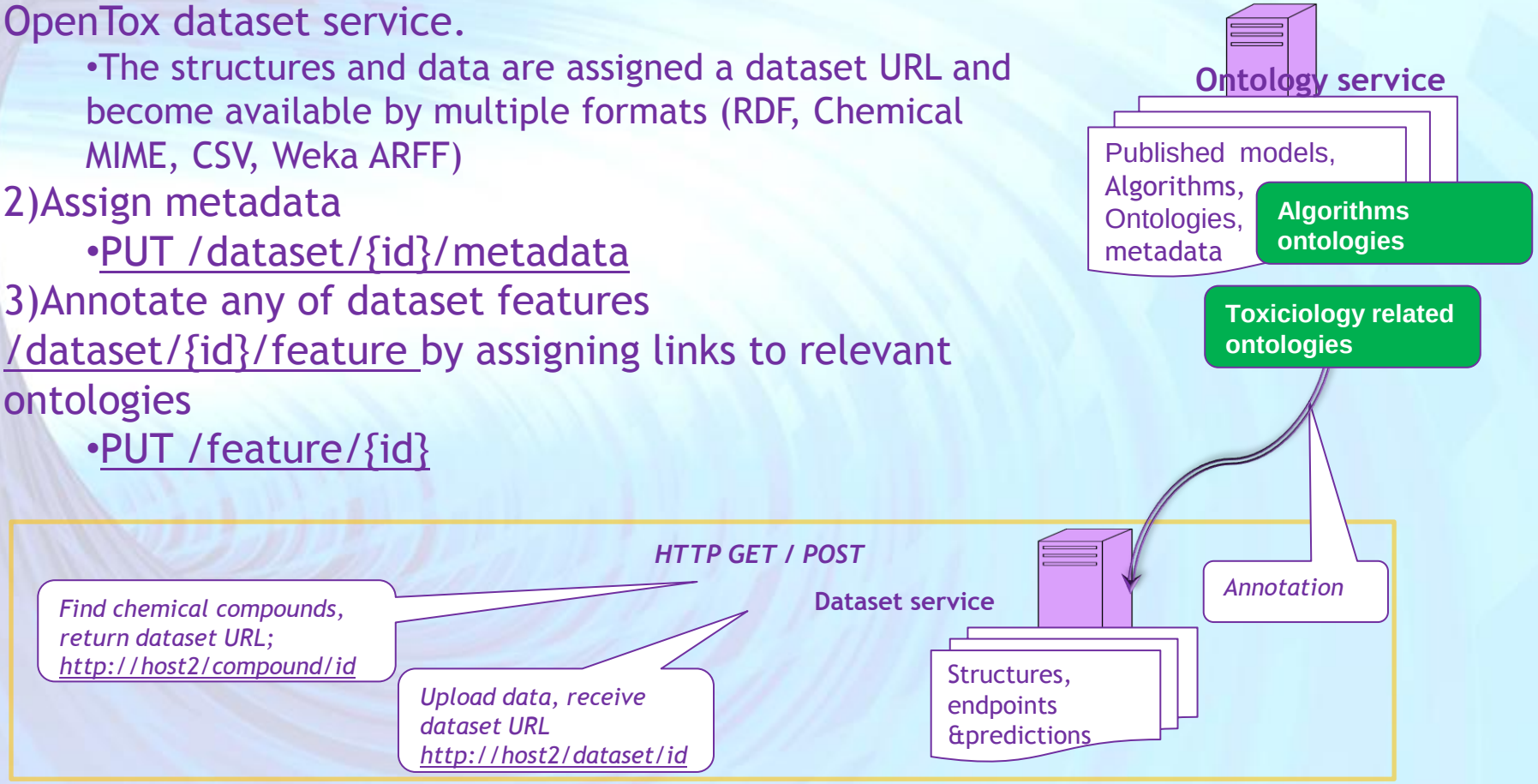
- The structures and data are assigned a dataset URL and become available by multiple formats (RDF, Chemical MIME, CSV, Weka ARFF)

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}

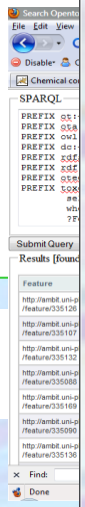


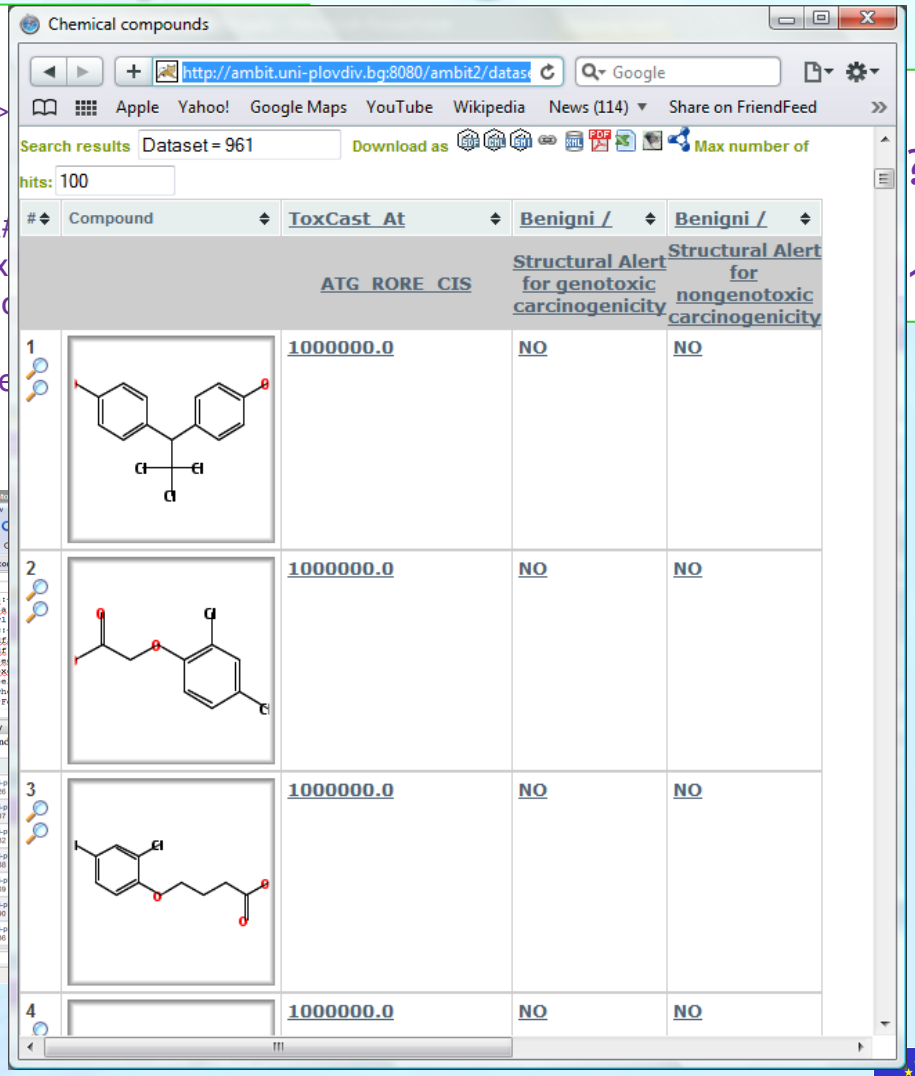
OpenTox

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.c
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

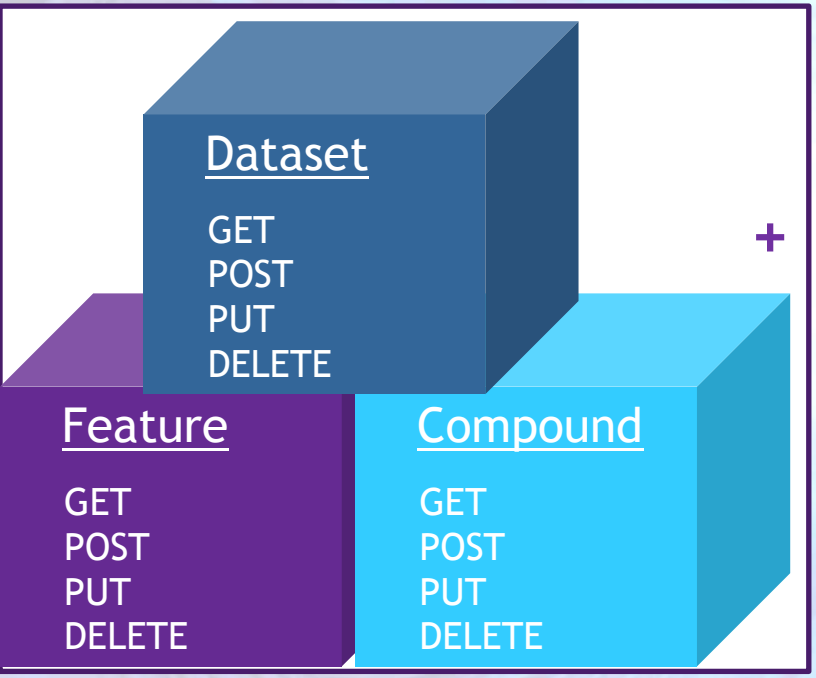
126

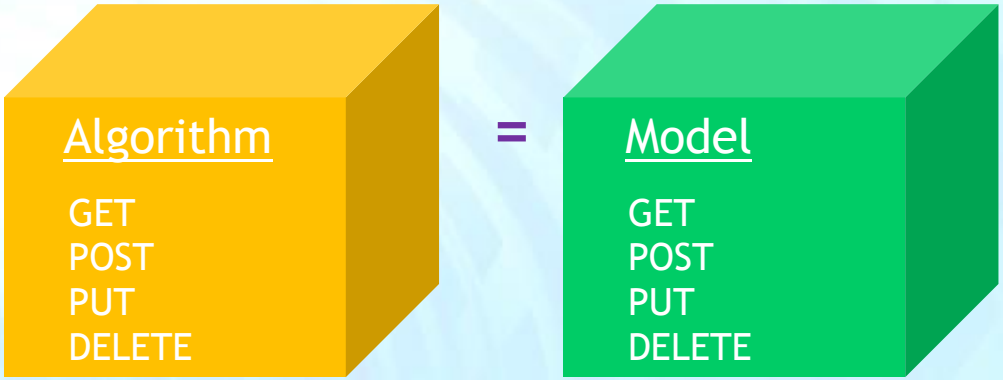
Query an OpenTox ontology service at http://ambit.uni-plovdiv.bg:8082/ontology

OpenTox

OpenTox dataset : create a model

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





http://myhost.com/algorithm/neuralnetwork

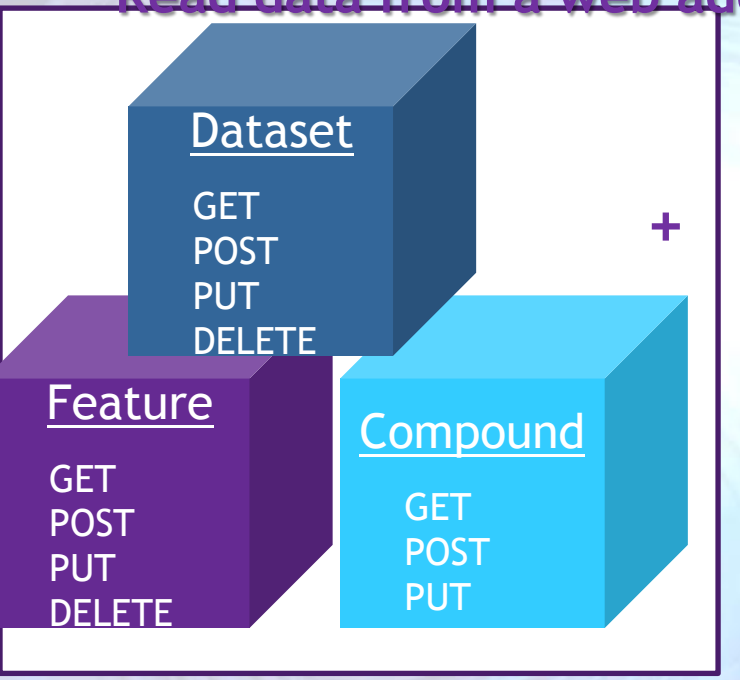
http://myhost.com/dataset/trainingset1

http://myhost.com/model/predictivemodel1

OpenTox

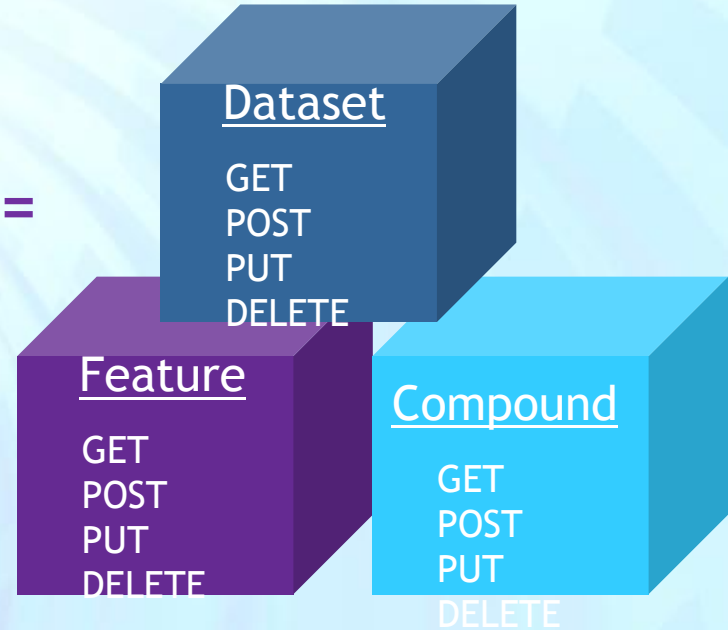
OpenTox dataset : descriptor calculation

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



+

=



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

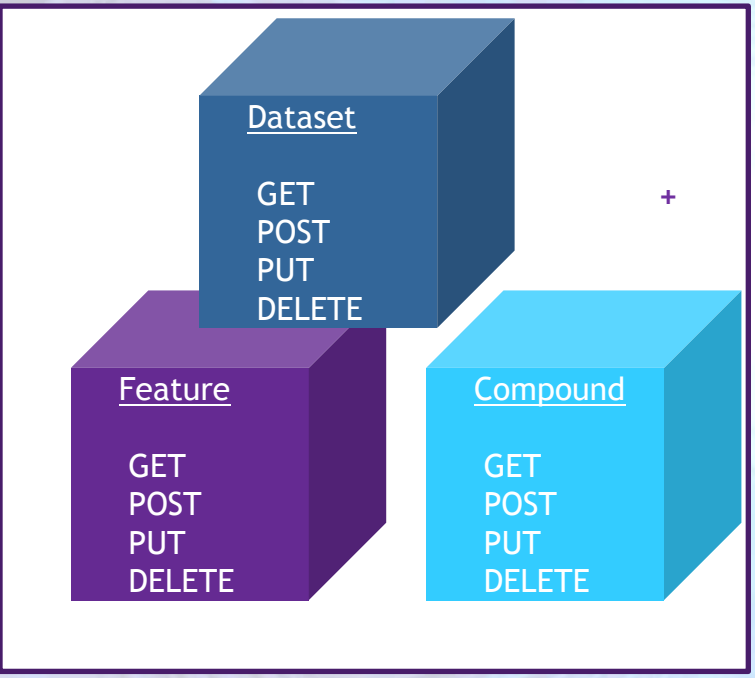
http://myhost.com/dataset/trainingset1

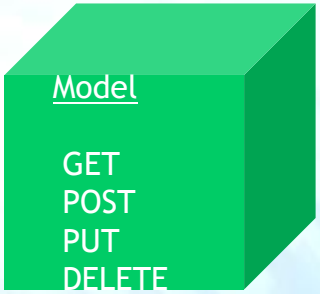
http://myhost.com/dataset/results

OpenTox

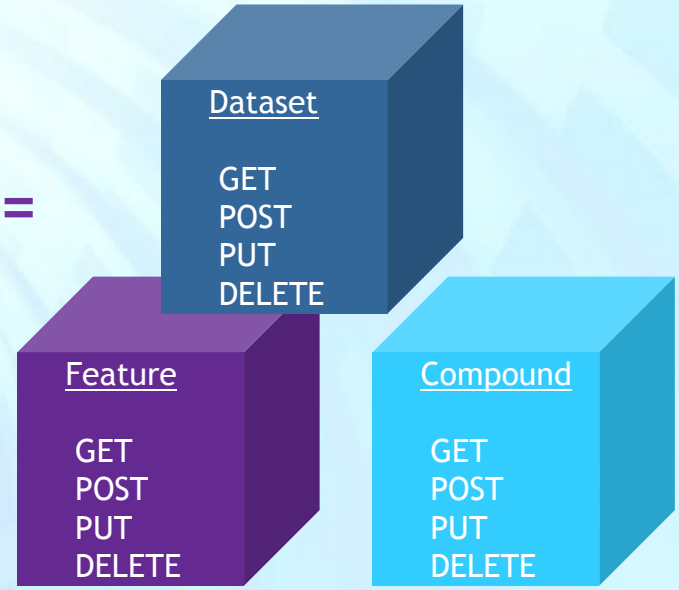
OpenTox dataset: apply a model

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





=



http://myhost.com/model/predictivemodel1

http://myhost.com/dataset/id1

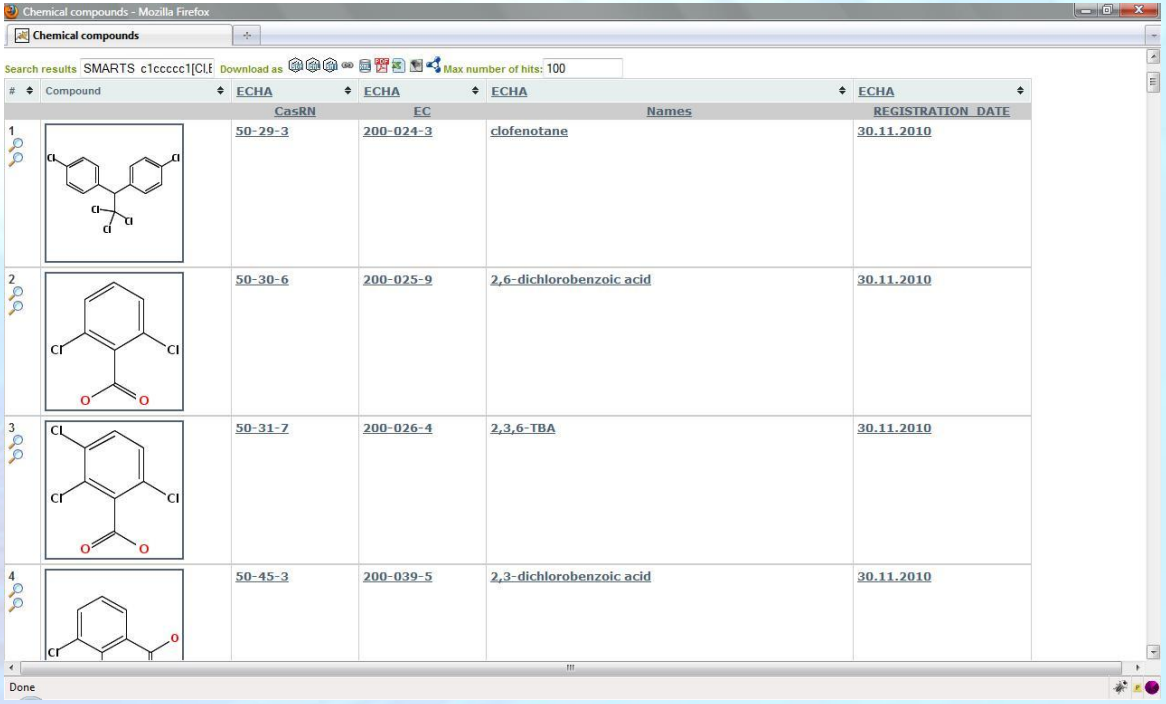
http://myhost.com/dataset/results1

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]

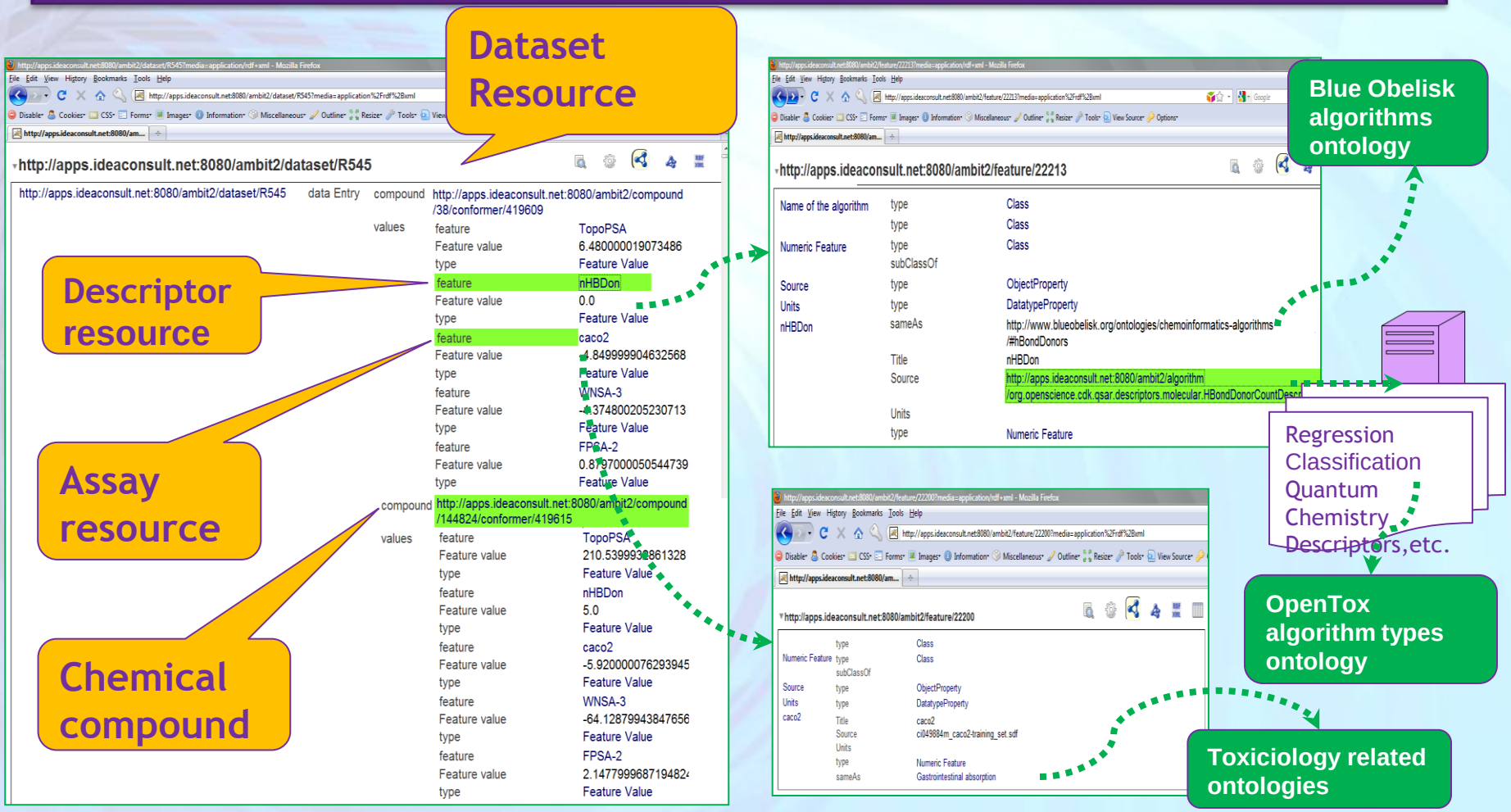


Datasets : Structure and quality labels

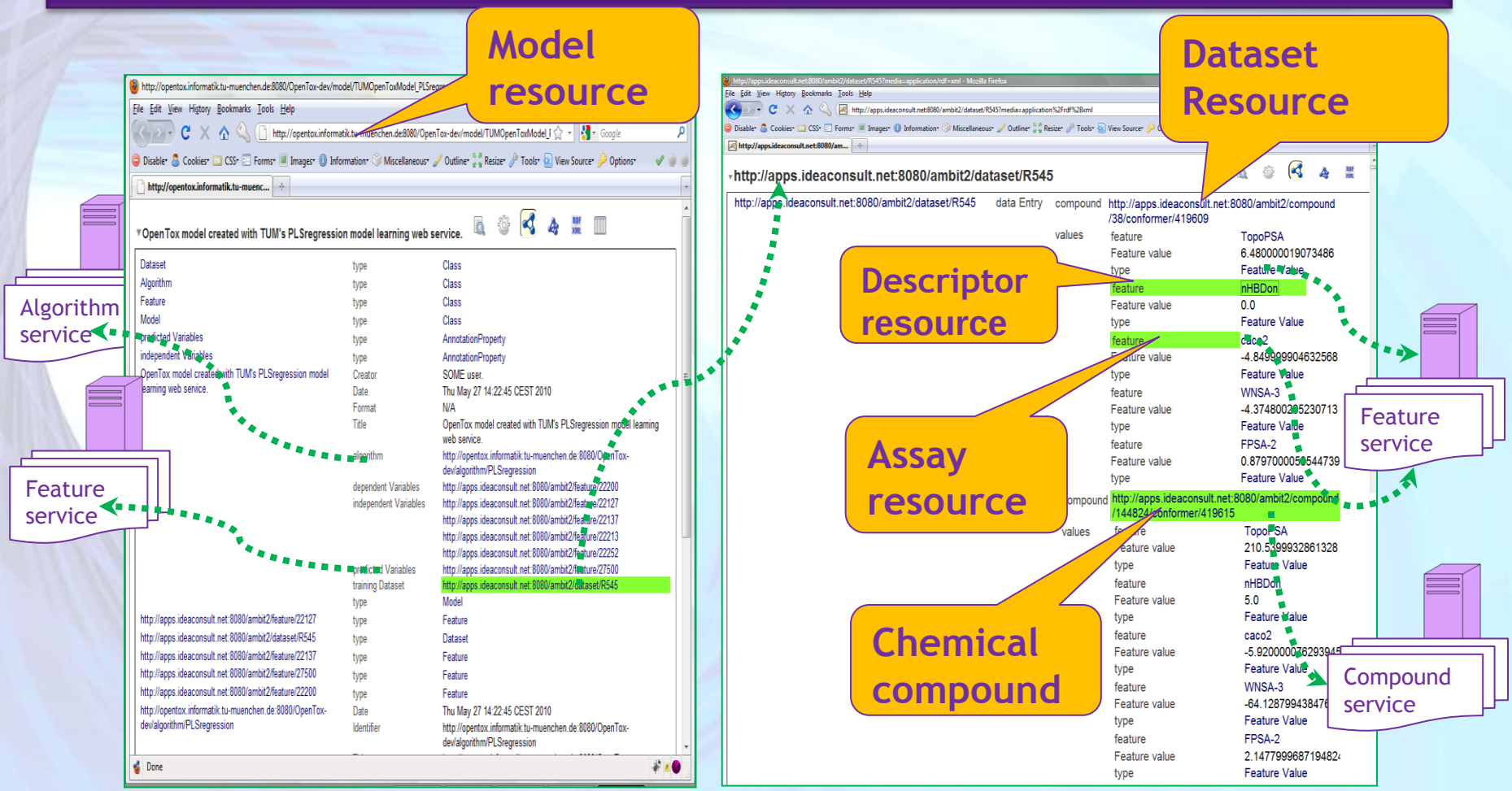
Dataset	OK	Probably OK	Probably ERROR	Unknown	Probably ERROR%
ECHA list of pre-registered substances	N/A	N/A	N/A	N/A	N/A
Chemical Identifier Resolver	67779	5314	3638	3471	4.75%
ChemIDplus	64802	7986	921	1745	1.24%
ChemDraw	17918	1147	502	478	2.57%
JRC PRS list	61332	4833	4022	2880	5.83%
ISSCAN	931	50	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



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



Thank you!