QsarDB – not just open data, but also open predictive models in chemistry and related sciences

Uko Maran

Institute of chemistry, University of Tartu

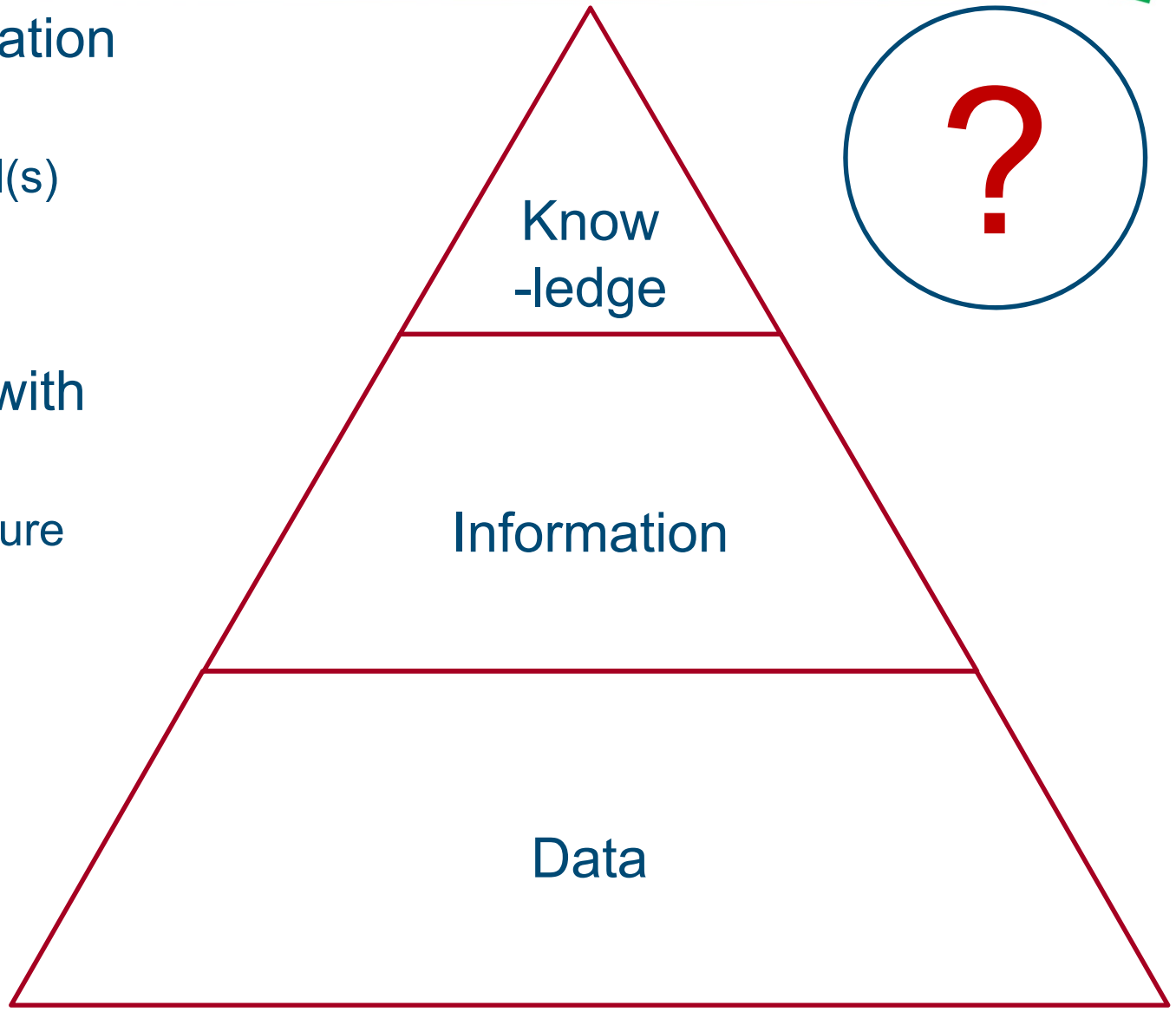
Estonia

uko.maran@ut.ee

www.qsardb.org

Data pyramide: data › information › knowledge

- Pieces of information are ordered …
  - In form of model(s)
- Data in content with other data …
  - molecular structure
  - Annotations
- Measurements, calculations, …

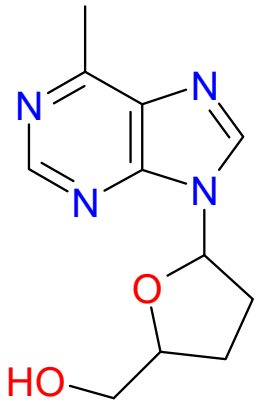


What is our data … (QSAR)

Quantitative

- Continous
- Discrete

Structure



$\Psi=\Psi(r,R)$

Activity

- *Physical*
  - t_B , η, n_D, …
- *Chemical*
  - pKa, *log*k, …
- *Spectroscopic*
  - δ_H,ν_{max}, …
- *Thermodynamic*
  - ΔH_f, c_v, …
- *Biomedical*
  - IC_{50}, LD_{50}, …
- Etc.

Relationship

- Regression
- Decision Tree
- Neural Networks
- Random Forest
- Support vector machine
- k-Nearest neighbors
- Ensemble
- Etc.

Activity = *f* (structure)

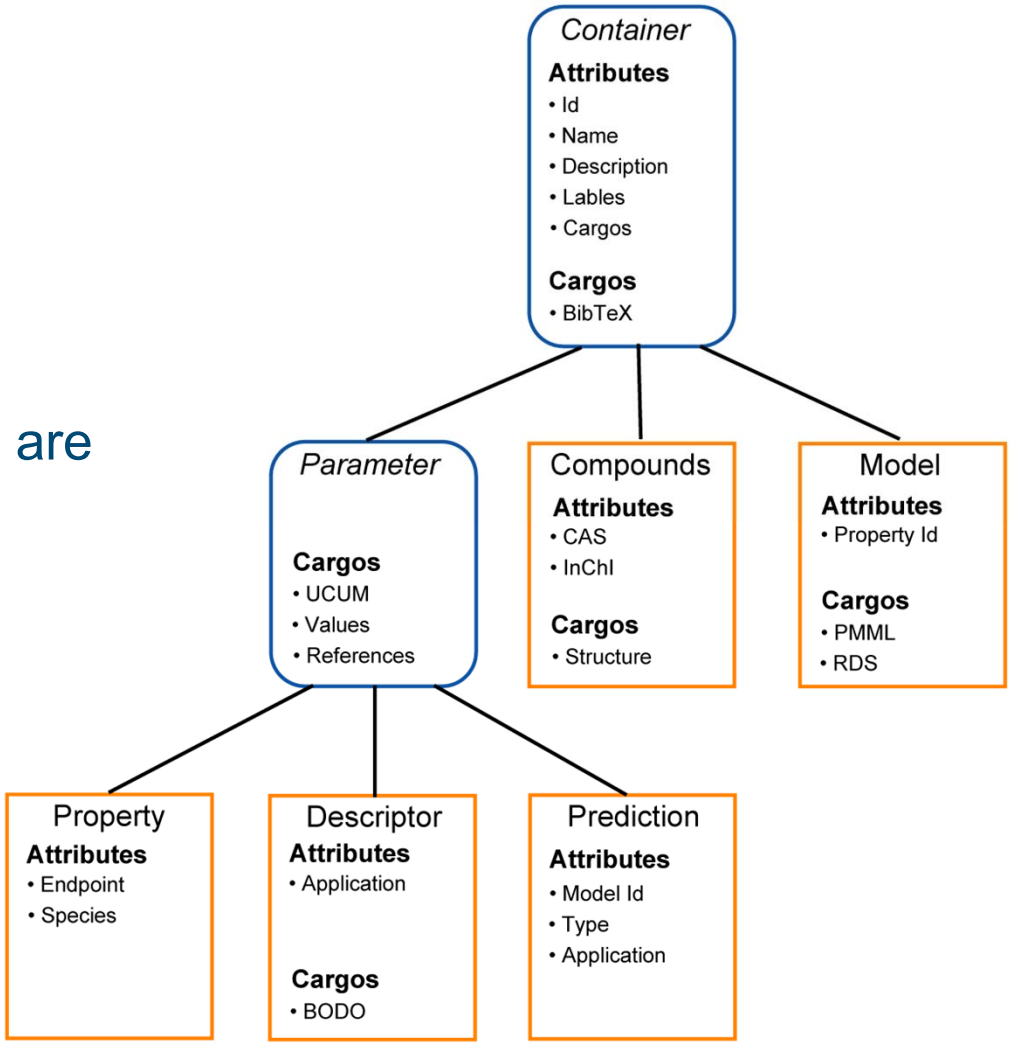
~~Excel?, PDF?, …~~

Four major components

- Data format
  - QSAR model archive (ie. small database - QDB)
- Smart Repository
  - collection of archives
- Tools for QDB archive creation
  - Command line
  - Graphical User Interface (QDB Editor)
- Web Services
  - http://qsardb.org/repository/service/predictor/10967/104/models/rf?CCO

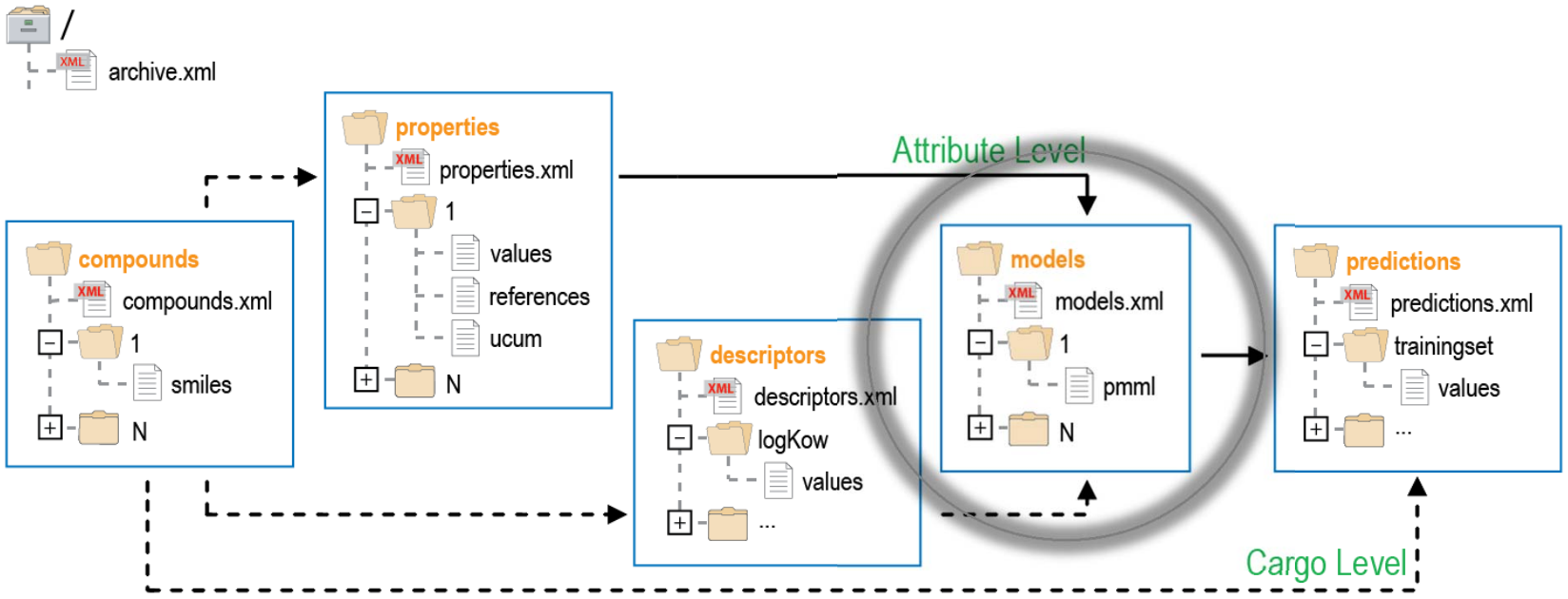
QsarDB data schema

- QsarDB vocabulary
  - Container
  - Attribute
  - Cargo
- Only primary data
  - Secondary data (R^2, etc.) are calculated on demand



Villu Ruusmann, Sulev Sild, Uko Maran*,
QSAR DataBank - an approach for the digital organization and archiving of QSAR model information.
Journal of Cheminformatics, **2014**, *6*:25.

QsarDB data format (container relationships)



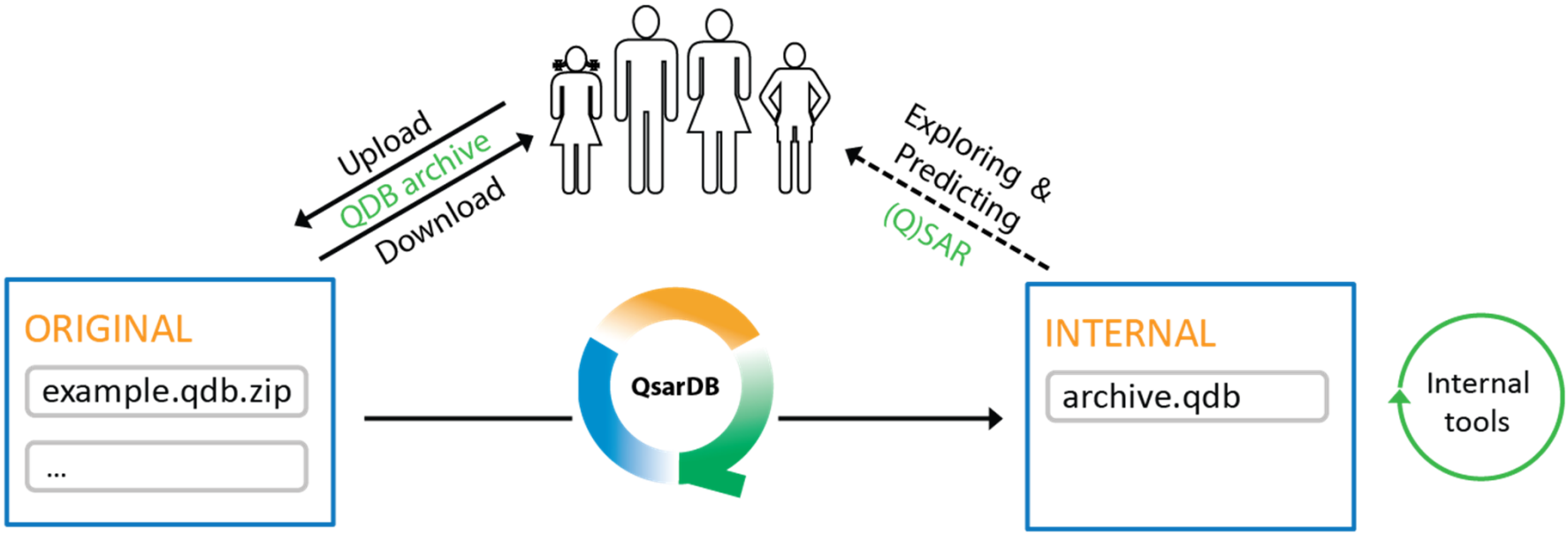
Villu Ruusmann, Sulev Sild, Uko Maran*,
QSAR DataBank - an approach for the digital organization and archiving of QSAR model information.
Journal of Cheminformatics, **2014**, *6*:25.

Electronic representation of predictive models

- PMML
  - Open standard for representing data mining models in XML format
  - PMML covers the following topics
    - Data preprocessing described through data dictionary, mining schema, transformations
    - Model representation
    - Post-processing (e.g. scaling model outputs)
- Other options are possible:
  - For example RDS data format (R native model representation mechanism)

- PMML format supports:
  - Association rules
  - Cluster models
  - Neural network
  - Regression
  - k-Nearest neighbors
  - Random forest
  - Tree models
  - Support-vector machines
  - Ensemble models (all of the above)
- http://qsardb.org/resources#pmml

The lifecycle of QDB archive (repository)



Villu Ruusmann, Sulev Sild, Uko Maran*
QSAR DataBank repository: open and linked qualitative and quantitative structure–activity relationship models
Journal of Cheminformatics, **2015**, 7:32.

Metadata & provenance information

Dublin Core	
Element name	**Description**
dc.date.accessioned	Date DSpace takes possession of item.
dc.date.available	Date or date range item became available to the public.
dc.date.issued	Date of publication or distribution.
dc.description.provenance	The history of custody of the item since its creation, including any changes successive custodians made to it.
dc.identifier.uri	Uniform Resource Identifier
dc.publisher	Entity responsible for publication, distribution, or imprint.
dc.rights.uri	References terms governing use and reproduction.
dc.rights	Terms governing use and reproduction.
dc.title	Title statement/title proper.

QsarDB	
Element name	**Description**
qdb.descriptor.application	Descriptor calculation software
qdb.model.qmrf	QMRF reference
qdb.model.type	Model type
qdb.prediction.application	Modeling software
qdb.property.endpoint	QMRF endpoint
qdb.property.species	Species

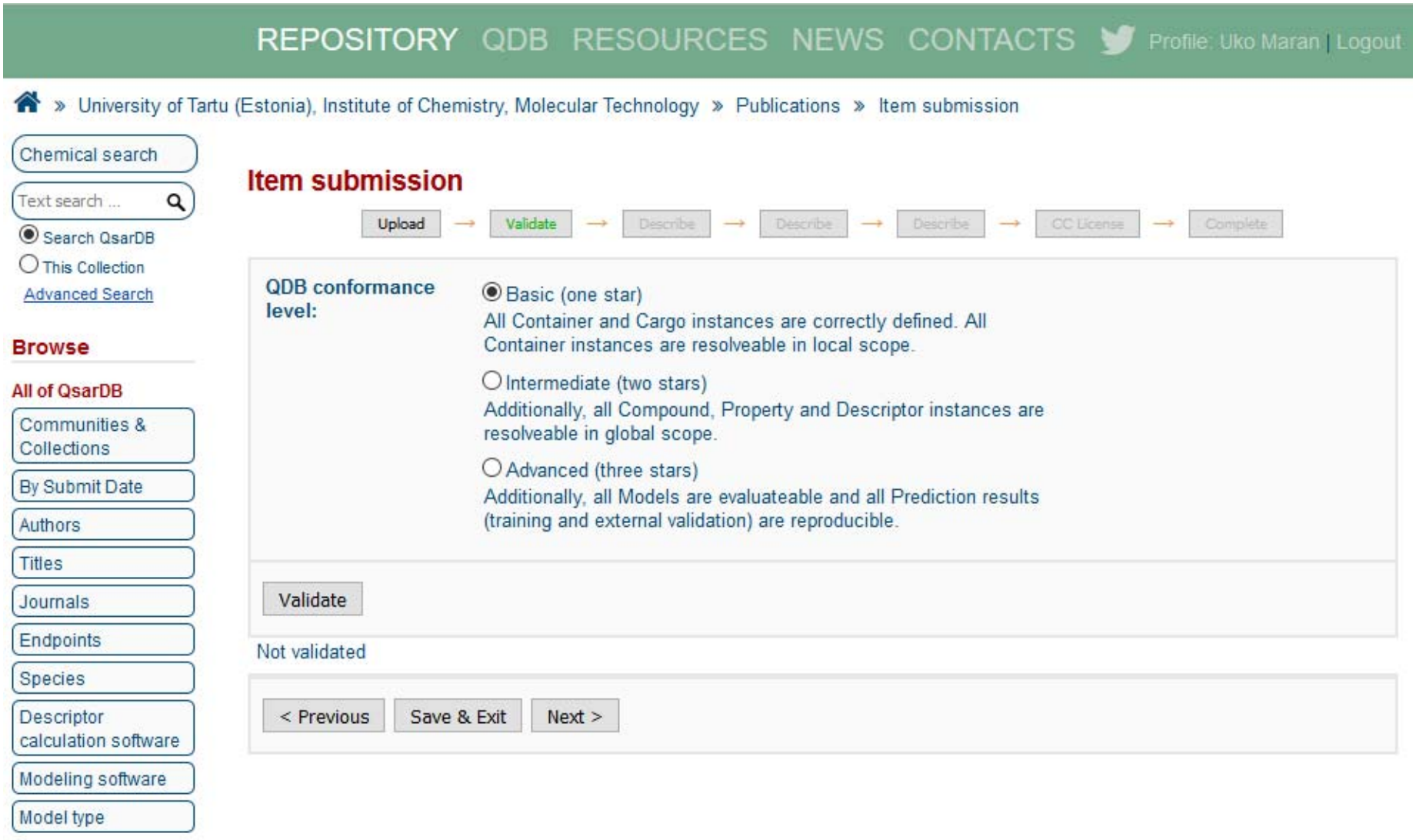
Bibliography (BibTeX)	
Element name	**Description**
bibtex.entry	The BibTeX entry type
bibtex.entry.address	Publisher's address
bibtex.entry.author	The name(s) of the author(s)
bibtex.entry.booktitle	The title of the book
bibtex.entry.doi	The Digital Object Identifier (DOI)
bibtex.entry.editor	The name(s) of the editor(s)
bibtex.entry.eprint	A specification of an electronic publication, often a preprint or a technical report
bibtex.entry.journal	The journal or magazine
bibtex.entry.month	The month of publication
bibtex.entry.note	Miscellaneous extra information
bibtex.entry.number	The "(issue) number" of a journal, magazine, or tech-report, if applicable.
bibtex.entry.organization	The conference sponsor
bibtex.entry.pages	Page numbers, separated either by commas or double-hyphens
bibtex.entry.publisher	The publisher's name
bibtex.entry.series	The series of books the book was published in
bibtex.entry.title	The title of the work
bibtex.entry.url	The WWW address
bibtex.entry.volume	The volume of a journal or multi-volume book
bibtex.entry.year	The year of publication

Villu Ruusmann, Sulev Sild, Uko Maran*, *Journal of Cheminformatics*, **2015**, 7:32.

Archive uploading policy

- Must have (scientific) publication
- (or have otherwise practical value …)

Quality control of the QDB archive file



Repository: Persistent digital identifiers

- Handle service (HDL): May 1, 2012
- DOI support: August 21, 2014

http://hdl.handle.net/10967/106 → http://qsardb.org/repository/handle/10967/106

Chemosphere 96 (2014) 23–32

Contents lists available at SciVerse ScienceDirect

Chemosphere

journal homepage: www.elsevier.com/locate/chemosphere

Measurement of baseline toxicity and QSAR analysis of 50 non-polar and 58 polar narcotic chemicals for the alga *Pseudokirchneriella subcapitata*

Villem Aruoja [a,*], Maikki Moosus [b], Anne Kahru [a], Mariliis Sihtmäe [a], Uko Maran [b,*]

[a] Laboratory of Environmental Toxicology, National Institute of Chemical Physics and Biophysics, Akadeemia tee 23, Tallinn 12618, Estonia
[b] Institute of Chemistry, University of Tartu, Ravila 14a, Tartu 50411, Estonia

HIGHLIGHTS

- REACH-relevant algal toxicity data were obtained for 50 nonpolar narcotic chemicals.
- Most of the tested compounds so far lacked published algal growth inhibition values.
- Toxicity of non-polar narcotic compounds correlated with hydrophobicity: $R^2 = 0.95$.
- MLR QSAR model was derived for non-polar and polar narcotic compounds: $R^2 = 0.92$.
- The Verhaar classification of non-polar narcotics appears to apply for algae.

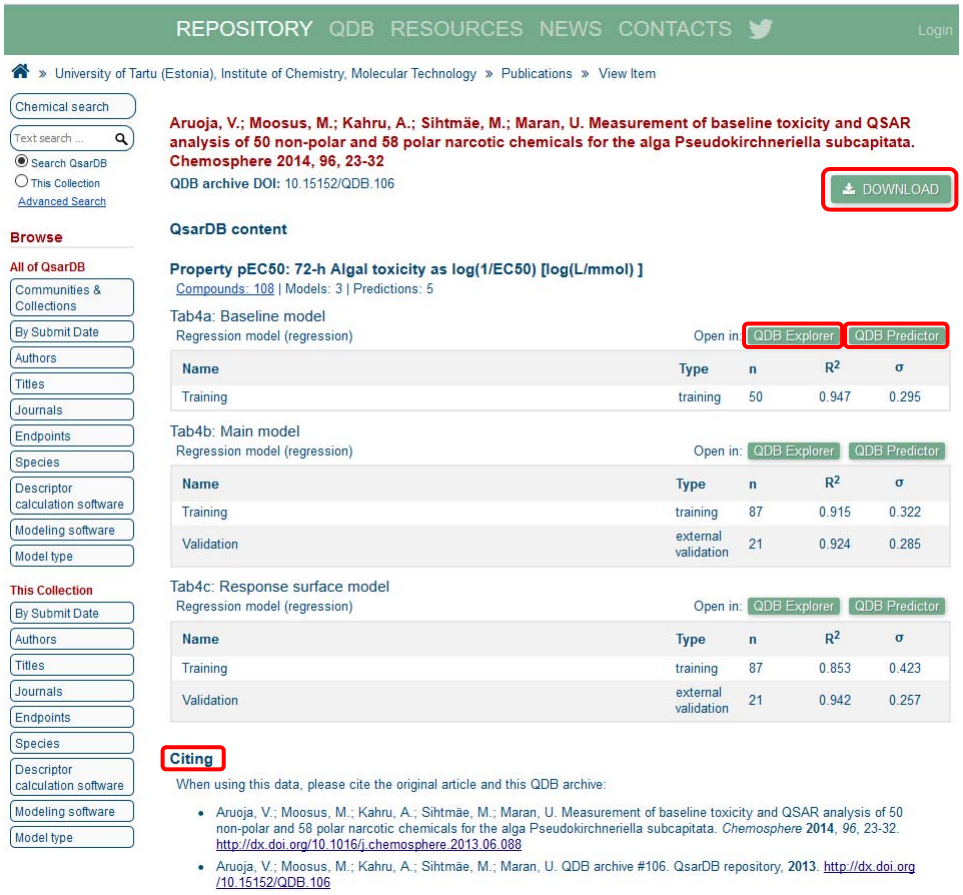
ARTICLE INFO

Article history:
Received 20 March 2013
Received in revised form 28 June 2013
Accepted 30 June 2013
Available online 26 July 2013

Keywords:
REACH
Baseline toxicity
QSAR
Non-polar narcosis
Algae
Pseudokirchneriella subcapitata

ABSTRACT

In this paper a set of homogenous experimental algal toxicity data was measured for 50 non-polar narcotic chemicals using the alga *Pseudokirchneriella subcapitata* in a closed test with a growth rate endpoint. Most of the tested compounds are high volume industrial chemicals that so far lacked published REACH-compliant algal growth inhibition values. The test protocol fulfilled the criteria set forth in the OECD guideline 201 and had the same sensitivity as the open test which allowed direct comparison of toxicity values. Baseline QSAR model for non-polar narcotic compounds was established and compared with previous analogous models. Multi-linear QSAR model was derived for the non-polar and 58 previously tested polar (anilines and phenols) narcotic compounds modulating hydrophobicity, molecular size, electronic and molecular stability effects coded in the molecular descriptors. Descriptors in the model were analyzed and applicability domain was assessed providing further guidelines for the *in silico* prediction purposes in decision support while performing risk assessment. QSAR models in the manuscript are available online through QsarDB repository for exploring and prediction services (http://hdl.handle.net/10967/106).



REPOSITORY QDB RESOURCES NEWS CONTACTS Login

» University of Tartu (Estonia), Institute of Chemistry, Molecular Technology » Publications » View Item

Chemical search
Text search ...
Search QsarDB
This Collection
Advanced Search

Browse
All of QsarDB
Communities & Collections
By Submit Date
Authors
Titles
Journals
Endpoints
Species
Descriptor calculation software
Modeling software
Model type

This Collection
By Submit Date
Authors
Titles
Journals
Endpoints
Species
Descriptor calculation software
Modeling software
Model type

Aruoja, V.; Moosus, M.; Kahru, A.; Sihtmäe, M.; Maran, U. Measurement of baseline toxicity and QSAR analysis of 50 non-polar and 58 polar narcotic chemicals for the alga Pseudokirchneriella subcapitata. Chemosphere 2014, 96, 23-32

QDB archive DOI: 10.15152/QDB.106

DOWNLOAD

QsarDB content

Property pEC50: 72-h Algal toxicity as log(1/EC50) [log(L/mmol)]

Compounds: 108 | Models: 3 | Predictions: 5

Tab4a: Baseline model
Regression model (regression) Open in: QDB Explorer QDB Predictor

Name	Type	n	R^2	σ
Training	training	50	0.947	0.295

Tab4b: Main model
Regression model (regression) Open in: QDB Explorer QDB Predictor

Name	Type	n	R^2	σ
Training	training	87	0.915	0.322
Validation	external validation	21	0.924	0.285

Tab4c: Response surface model
Regression model (regression) Open in: QDB Explorer QDB Predictor

Name	Type	n	R^2	σ
Training	training	87	0.853	0.423
Validation	external validation	21	0.942	0.257

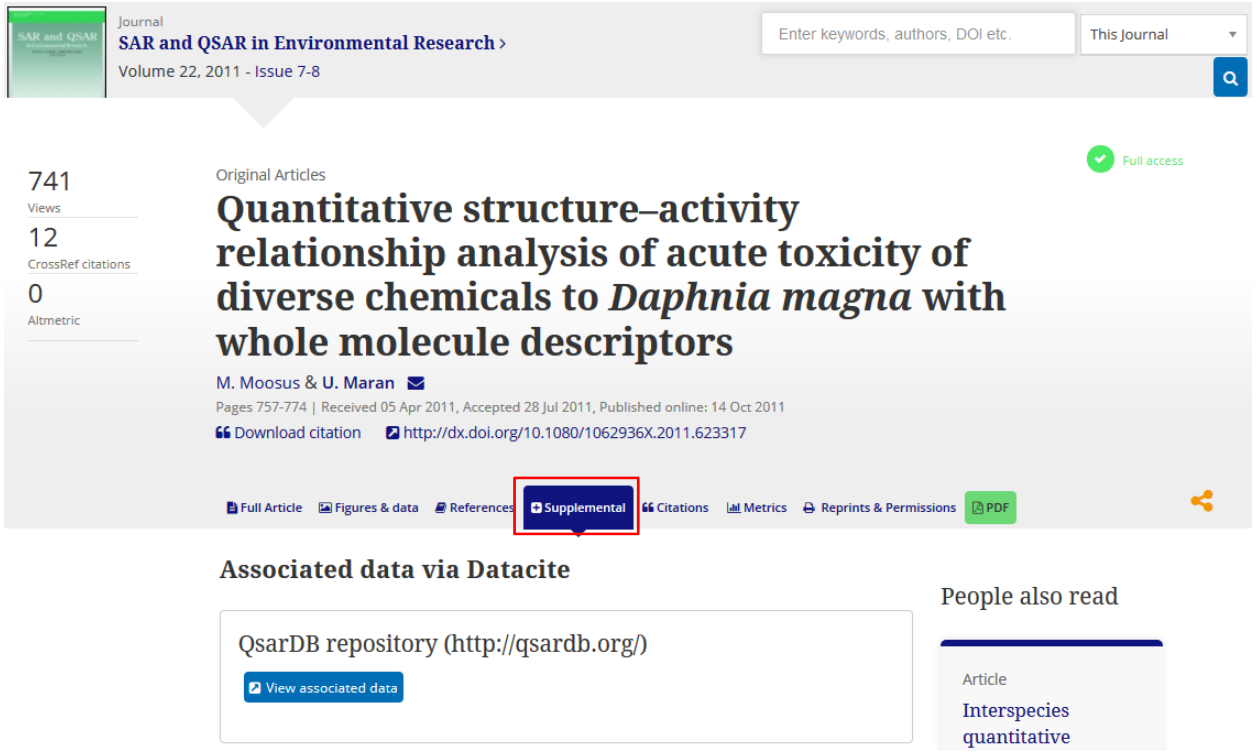
Citing

When using this data, please cite the original article and this QDB archive:

- Aruoja, V.; Moosus, M.; Kahru, A.; Sihtmäe, M.; Maran, U. Measurement of baseline toxicity and QSAR analysis of 50 non-polar and 58 polar narcotic chemicals for the alga Pseudokirchneriella subcapitata. *Chemosphere* **2014**, *96*, 23-32. http://dx.doi.org/10.1016/j.chemosphere.2013.06.088
- Aruoja, V.; Moosus, M.; Kahru, A.; Sihtmäe, M.; Maran, U. QDB archive #106. QsarDB repository, **2013**. http://dx.doi.org/10.15152/QDB.106

Citing: advanced level

Create QDB archive → Upload QDB archive → Add article DOI to metadata → WAIT...



QsarDB in numbers & Access stat.

- QDB archives
  - Total number [186]
  - With models [154]
  - With data only [32]
- Endpoints
  - Physical Chemical Properties [16]
  - Environmental fate parameters [8]
  - Ecotoxic effects [7]
  - Human health effects [12]
  - Toxicokinetics [3]
  - Other [34]
- Species [32]
- Models [**452**]
- Most popular models
  - 2006 Liu,H.: #124 [3096]
  - 2013 Aruoja,V.: #106 [2978]
  - ONS: melting point #104 [2572]
- Model Types [12][archives/models]
  - Decision tree (*classif.*) [3/3]
  - k-Nearest neighbors (*regr.*) [2/2]
  - k-Nearest neighbors (*classif.*) [1/1]
  - k-Nearest neighbors ensemble (*regr.*) [1/1]
  - Neural network (*regr.*) [5/7]
  - Neural network ensemble (*regr.*) [1/3]
  - Random forest (*classif.*) [1/3]
  - Random forest (*regr.*) [13/13]
  - Regression model (*classif.*) [5/5]
  - Regression model (*regr.*) [**133/407**]
  - Regression model ensemble (*regr.*) [4/4]
  - Support vector machine (*regr.*) [1/3]
- Web service statistics (predictions)
  - Over 7.5 million (since 2012-07)
- More than 50 scientific journals

Did not have time for this

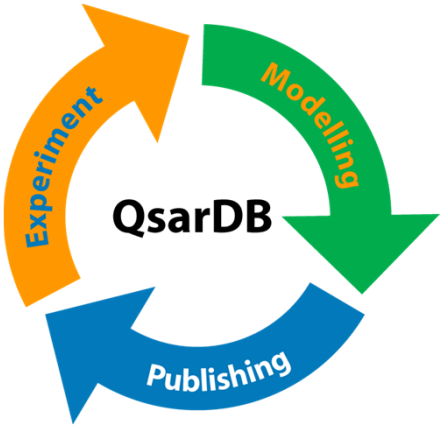
- QDB Explorer
- QDB Predictor
- Full text search
- Chemical structure and substructure search
- Linking to other databases
- Metadata for (Q)SAR-s
- QsarDB tools for making QDB archives
- Web services
- Communities and Collections
- …

Concluding Phrases

- Scientific data is by definition open, when published, … ?
- Scientific data is side product of research and also starting point for new research.
- Scientists in fact share data …
- Look for best practices around … (new wheels?)
- Ice-braking should be done on the root level …
- Content aware domain specific vs. general repositories!
- Easy availability and accessibility … transparency …
- Interactive and smart solutions keep data alive …
- How to guarantee sustainable development of infrastructure?
- EU is loosing research diversity momentum in funding policies …
- Open data and open publishing movement – changes business model …

Thank you!

- University of Tartu
  - Sulev Sild, Dr
  - Villu Ruusmann, MSc
  - Geven Piir, Dr
  - Kalev Takkis, Dr
  - Alfonso T. Garcia-Sosa, Dr
  - Mare Oja, MSc
- Tallinn University of Technology
  - Andre Lomaka, Dr
  - Iiris Kahn, Dr
  - Elmar Laigna, BSc
  - Priit Ahte, MSc



www.qsardb.org

Contact: **uko.maran@ut.ee**

Funding:

- EU FP6 Chemomentum Project (IST-5-033437: years 2006 – 2009);
- Estonian Science Foundation (Grants – 5805: years 2004-2008; 7709: years 2009-2011);
- Estonian Ministry for Education and Research (Grants – SF0182644Bs04: years 2004 – 2008; SF0140031Bs09: years 2009 – 2014; IUT34-14: 2015-2020)
- European Union, Regional Development Fund (3.2.1201.13-0021: years 2013-2015).