LINKING THE SILOS
DATA AND PREDICTIVE MODELS
INTEGRATION IN TOXICOLOGY

NINA JELIAZKOVA
IdeaConsult Ltd.
Sofia, Bulgaria
www.ideaconsult.net

15TH INTERNATIONAL QSAR WORKSHOP
TALLINN, ESTONIA
LAST DECADE CHANGES

Biology, Bioinformatics
- Data, human genome

Internet
- Online databases
- Online collaboration
- Crowd Sourcing

Open access, open source
- Database management systems
- Machine learning and statistics
- Cheminformatics

Williams J M et al. Brief Bioinform 2010;11:598-609
Although biological databases are only a fraction of all available bioinformatics software resources, their rise is representative of the overall growth of these resources, and is concurrent with the number of base pairs released in GenBank.
OPEN SOURCE CHEMINFORMATICS

A. Dalke EuroQSAR 2008 poster

DATABASES

- 2001: Binding DB
- 2003 Ligand Depot
- 2004: ZINC database
- 2004: PDBbind database
- 2004: PubChem
- 2004: sc-PDB
- 2004: Binding MOAD
- 2005: DrugBank
- 2006: Chemical Structure Lookup Service (CSLS)
- 2007: ChemSpider
- 2008/2009: ChEMBL
- 2009: Chemical Identifier Resolver (CIR)
- 2003: First Nucleic Acids Research (NAR) annual Database issue
  - >1000 databases published per year since 2008;
  - >1300 in 2011

Marc Nicklaus, 5th Meeting on U.S. Government Chemical Databases and Open Chemistry, 2011
http://cactus.nci.nih.gov/presentations/meeting-08-2011/meeting-2011-08-25.html
EASY ACCESS

Increased activity

Technology advances, more products

What was revolutionary few decades ago with establishing the software for chemicals registration (e.g. CAS database) is a routine nowadays.

 It could be a Computer Science Master project
 or a research project …
EU PROJECTS (AN INCOMPLETE LIST)

- EU FP6/ FP7: 2-FUN, ACROPOLIS, ACuteTox, AQUATERRA, ARTEMIS, CADASTER, CAESAR, carcinoGENOMICS, CONTAMED, ESNATS, GENESIS, HEIMSTA, INVITROHEART, LIINTOP, MENTRANS, MODELKEY, NOMIRACLE, OSIRIS, PREDICT-IV, OpenTox, ReProTect, RISKBASE, RISKCYCLE, Sens-it-iv, VITROCELLOMICS

- Innovative Medicines Initiative – 30 ongoing projects, including eTox, OpenPhacts

- Safety Evaluation Ultimately Replacing Animal Testing (SEURAT) cluster – 6 research projects
SCIENTIFIC DATABASES IMPLEMENTATION

- Identify the data model and functionality
- Translate the data model into a database schema
- Implement the database and user interface functionality
- (Optionally) provide libraries or expose (some) of the functionality as web services

**Advantages**
- Use one’s favourite technology and jump directly into implementation
- Attract end-users with nice GUI relatively quickly
- Relatively easy to persuade funding organisations this will be a useful resource

**Disadvantages**
- Proliferation of incompatible resources, providing similar functionality, but incompatible programming interface
- Difficult to extract and collate data automatically

This is not the only nor the best way!
EASY ACCESS, HENCE MANY NEW SYSTEMS

SOFTWARE SYSTEMS

Life sciences software
- and toxicology in particular

Live in their own world
- Mostly developed independently
- Compatibility is rarely perceived as a primary design goal.

Lack of communication and common goals

THE SILO EFFECT

Silo storage system
- Designed to store one single type of grain.

Information silo
- Rigid design
- No easy exchange of information
- No integration with other systems
THE NEED OF INTEGRATION


- Calculations, Descriptors, Statistics, Models
- Data (chemical structures, properties, predictions)

Approaches toward integration:
- Workflow management systems
- Standalone container applications (chassis)
- Web applications
- Web services, web mashups

WORKFLOW MANAGEMENT SYSTEMS

Commercial
- Accelrys Pipeline Pilot

Open source:
- Kepler, KNIME, Taverna, Triana

Workflow repositories
- MyExperiment
  http://www.myexperiment.org/

Advantages
- Flexibility
- Reproducibility

Disadvantages
- Overhead: 30% of the tasks defined and run in workflows are related to data conversion, rather than data analysis.
- Convenience:
  - Domain experts often prefer simple user interfaces or software with predefined functionality;
  - Power users prefer scripting languages, rather than GUls or graphical workflow builders with their specific constraints.
- Compatibility: Nodes are not transferable between workflow engines
INTEGRATION PLATFORMS

CONTAINER APPLICATIONS

 OECD Toolbox
   Windows only
   OECD / ECHA funded

 Bioclipse
   Multiplatform, Java based; standard OSGI interface for modules
   Open Source

WEB TOOLS

 ChemBench
 ChemMine
 Collaborative Drug Discovery (CDD)
 OCHEM
 OpenTox

CHEMBENCH

Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. Journal of computer-aided molecular design.


**Jeliazkova, N., et. al. (2011). AMBIT RESTful web services: an implementation of the OpenTox application programming interface. *Journal of cheminformatics*, 3, 18.**
MORE WEB PROJECTS

OPENTOX ALGORITHMS

Uniform interface: (OpenTox web services API)

- Descriptor calculation, feature selection;
- Classification and regression algorithms;
- Rule based algorithms;
- Applicability domain algorithms;
- Visualization, similarity and substructure queries;
- Composite algorithms (workflows);
- Structure optimization, metabolite generation, tautomer generation, etc.

An algorithm /ˈælɡərɪðəm/ is a step-by-step procedure for calculations. Algorithms are used for calculation, data processing, and automated reasoning.
UNIFORM APPROACH TO DATA PROCESSING

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

Once we have a set of uniform building blocks, we can build new tools.

http://myhost1.com/dataset/trainingset1

http://myhost2.com/algorithm/{myalgorithm}

http://myhost3.com/dataset/results
UNIFORM APPROACH TO MODELS BUILDING

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

Once we have a set of uniform building blocks, we can build new tools.
Model management is obtained as a side effect
**WEB APPLICATIONS**

- [http://ToxPredict.org](http://ToxPredict.org) – aggregates remote predictions
- OpenTox Wrapper for OCHEM/CADASTER web services
- QMRF database
- Applicability domain used by CADASTER web site

**BIOCLIPSE-OPENTOX**

*Willighagen E., Jeliazkova N., Hardy B., Grafstrom R., Spjuth O.,* Computational toxicology using the OpenTox application programming interface and Bioclipse, *BMC Research Notes* 2011 4 (1), 487
HOW TO CREATE/PUBLISH A MODEL

Upload the training set and rebuild the model;
- Example: Upload the Open Melting Point dataset and run the linear regression or random forest algorithm

Develop an OpenTox API compatible solutions, allowing to train and run predictive models;
- Example: Lazar models, OpenTox partner models (BG, DE, CH, GR, RU)

Use thin wrappers for third-party models, and exposing them through the compatible web service API.
- OCHEM models are exposed as OpenTox API compatible models using this approach

All models become potentially visible to client applications (ToxPredict, Bioclipse), subject to access rights.
Uploading a dataset makes it structure and similarity searchable; access rights could be controlled. An unique URI is assigned. This is a web service as well.
LOAD DATA, BUILD MODEL, APPLY MODEL

---

**OpenTox Model REST API**

- Query compounds
- Chemical compounds
- Datasets
- Algorithms
- Models
- Features
- Templates
- Similarity
- Substructure
- Ontology
- service
- ToxPredict
- Depiction
- Reactions
- Tautomers
- Help

---

**ambit**

- [Search](#)

---

**Table:**

<table>
<thead>
<tr>
<th>Title</th>
<th>Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>d8d397c7.6e4f4733-a970.d778641b981f,weka.classifiers.functions.LinearRegression</td>
<td>LR</td>
</tr>
</tbody>
</table>

**Model:**

- Enter dataset or compound URI to apply the model
- Dataset URI: [http://ambit.uni-plodiv.bg:8080/qmrfd/data/dataset/R0053](http://ambit.uni-plodiv.bg:8080/qmrfd/data/dataset/R0053)
- Training algorithm: LR
- Variables: Independent, Dependent, Predicted
- Model content: [application/x-javaserialized-object]

---

**Graph:**

- X-axis: Ave °C
- Y-axis: Ave °C
- Scatter plot with data points

---

6/20/2012

**15TH INTERNATIONAL QSAR WORKSHOP**

**TALLINN, ESTONIA**
# Melting Point Models Comparison

## Table: Melting Point Data

<table>
<thead>
<tr>
<th>Compound</th>
<th>Low °C</th>
<th>High °C</th>
<th>Average °C</th>
<th>Melting Point Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comp 1</td>
<td>-80</td>
<td>-80</td>
<td>-80</td>
<td>40.68</td>
</tr>
<tr>
<td>Comp 2</td>
<td>-100</td>
<td>-100</td>
<td>-100</td>
<td>40.68</td>
</tr>
<tr>
<td>Comp 3</td>
<td>-10.5</td>
<td>-95</td>
<td>-97.33</td>
<td>46.68</td>
</tr>
<tr>
<td>Comp 4</td>
<td>-100</td>
<td>-95</td>
<td>-97.33</td>
<td>46.68</td>
</tr>
<tr>
<td>Comp 5</td>
<td>-100</td>
<td>-95</td>
<td>-97.33</td>
<td>46.68</td>
</tr>
</tbody>
</table>

### Chemical Structures

![Chemical Structure 1](image1)

![Chemical Structure 2](image2)

![Chemical Structure 3](image3)
**EXAMPLE: REPDOSE DATASETS**

<table>
<thead>
<tr>
<th>Study Reliability</th>
<th>Study Reliability</th>
<th>h Effect</th>
<th>Study Reliability</th>
<th>Study Reliability</th>
<th>h Effect</th>
<th>h Effect</th>
<th>Study Reliability</th>
<th>Study Reliability</th>
<th>h Effect</th>
<th>Study Reliability</th>
<th>Study Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
<td>1.012</td>
<td>A</td>
<td>0.973</td>
<td>0.88</td>
<td>0.88</td>
<td>1.12</td>
<td>0.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>B</td>
<td>17.083</td>
<td>volume</td>
<td>2.937</td>
<td>A</td>
<td>0.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Similarity search results**
PUBLISH/SHARE A MODEL

Upload the training set and rebuild the model;
- Use existing algorithms (descriptors, statistics), under existing license. Host on existing servers.

Develop an OpenTox API compatible solutions, allowing to train and run predictive models;
- Guaranteed exact reproduction of the model!
- Any license (closed or open source). Host anywhere.

Use thin wrappers for third-party models, and exposing them through the compatible web service API.
- Guaranteed exact reproduction of the model!
- Any license (closed or open source). Host anywhere.

Any OpenTox resource could be assigned restricted access
PUBLISH/SHARE A MODEL

The advantage of a clean underlying API
- Compounds, properties, dataset, algorithms and models

Any kind of processing
- Could be formalized as algorithm or a model
- Applicability domain is an algorithm
- Structural alerts are models! Models of a human expertise
- We don’t need a separate database to handle structural alerts
- Just publish the structural alerts as models and apply the models to your dataset!
DATASET COMPARISON USING OPENTOX ALGORITHMS AND MODELS

ECHA Preregistration list

- Alert for SNA
- Alert for Michael Acceptor
- Alert for Acyl Transfer agent
- No DNA binding alerts

PubChem

- Alert for SNA
- Alert for Michael Acceptor
- Alert for Acyl Transfer agent
- No DNA binding alerts
MODELS FOUND IN THE LITERATURE: AMBIGUOUS!

Cramer rules – textual description

- Q29. Readily Hydrolysed
- How to implement – abiotic / biotic?
  - Context: oral toxicity
  - Yet there are different implementations
- Knowledge of the model internals is essential; compare predictions!

Substructure alerts

- Have come a long way since the practice of publishing only textual description.
- Mostly SMARTS
- Custom languages in some of the tools

No established procedure to extend SMARTS or SMILES – thus all the custom extensions.
SMARTS

Multiple implementations
- Not entirely compatible
- No real standard, except the Daylight web page

Steep learning curve (even for experts)
- Q: Does SMARTS query [CH2][NH2] match [CH2][NH2+]?
- A: YES. Any attribute which is not specified in the SMARTS is not tested. So if you do not mention formal charge for an atom, any charge is allowed. The same is true for any other query attribute.

Validate published structural alerts!
- Publishing structural alerts should be no different than publishing QSAR models and require validation dataset

THE EXPERT KNOWLEDGE IS AMBIGUOUS

Mode of actions

Adverse outcome pathways
- Effectopedia http://www.qsari.org/index.php/software/100-effectopedia – an attempt to gather the expert knowledge
  - Hopefully will not become yet another silo
  - Could be useful, if adopting the approach biology and bioinformatics are using to capture similar information

Ontologies; semantic annotation
- Knowledge representation with strong ground in logic and computer science
- Allow automatic reasoning
- Many computer frameworks
- Huge amount of biology/bioinformatic resources

Ontology:
A formal, shared conceptualization of a domain
Chemical entities of biological interest

http://www.ebi.ac.uk/chebi/
The Open Biological and Biomedical Ontologies

The OBO Foundry is a collaborative experiment involving developers of science-based ontologies who are establishing a set of principles for ontology development with the goal of creating a suite of orthogonal interoperable reference ontologies in the biomedical domain. The groups developing ontologies who have expressed an interest in this goal are listed below, followed by other related efforts in this domain.

In addition to a listing of OBO ontologies, this site also provides a statement of the OBO Foundry principles, discussion fora, technical infrastructure, and other services to facilitate ontology development. We welcome feedback and encourage participation.

Click any column header to sort the table by that column. The link to the term request trackers for the listed ontologies.

### OBO Foundry ontologies

<table>
<thead>
<tr>
<th>Title</th>
<th>Domain</th>
<th>Prefix</th>
<th>File</th>
<th>Last changed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biological process</td>
<td>biological process</td>
<td>GO</td>
<td>gene_ontology_edit.obo</td>
<td>2012/06/18</td>
</tr>
<tr>
<td>Cellular component</td>
<td>anatomy</td>
<td>GO</td>
<td>gene_ontology_edit.obo</td>
<td>2012/06/18</td>
</tr>
<tr>
<td>Chemical entities of biological interest</td>
<td>biochemistry</td>
<td>CHEBI</td>
<td>chibi.obo</td>
<td>2012/06/02</td>
</tr>
<tr>
<td>Molecular function</td>
<td>biological function</td>
<td>GO</td>
<td>gene_ontology_edit.obo</td>
<td>2012/06/18</td>
</tr>
<tr>
<td>Phenotypic quality</td>
<td>phenotype</td>
<td>PATO</td>
<td>quality.obo</td>
<td></td>
</tr>
<tr>
<td>Protein Ontology (PRO)</td>
<td>proteins</td>
<td>PR</td>
<td>pro.obo</td>
<td></td>
</tr>
<tr>
<td>Xenopus anatomy and development</td>
<td>anatomy</td>
<td>XAO</td>
<td>xenopus_anatomy.obo</td>
<td>2012/02/17</td>
</tr>
<tr>
<td>Zebrafish anatomy and development</td>
<td>anatomy</td>
<td>ZFA</td>
<td>zebrafish_anatomy.obo</td>
<td>2012/06/15</td>
</tr>
</tbody>
</table>

### OBO Foundry candidate ontologies and other ontologies of interest

<table>
<thead>
<tr>
<th>Title</th>
<th>Domain</th>
<th>Prefix</th>
<th>File</th>
<th>Last changed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adverse Event Reporting Ontology</td>
<td>health</td>
<td>AERO</td>
<td>AAO_v1_edit.obo</td>
<td></td>
</tr>
<tr>
<td>Amphibian gross anatomy</td>
<td>anatomy</td>
<td>AAO</td>
<td>amphibian_gross_anatomy</td>
<td></td>
</tr>
<tr>
<td>Amphibian taxonomy</td>
<td>anatomy</td>
<td>ATO</td>
<td>amphibian_taxonomy.obo</td>
<td></td>
</tr>
<tr>
<td>Anatomical Entity Ontology</td>
<td>anatomy</td>
<td>AEO</td>
<td>aeo.obo</td>
<td>2012/06/01</td>
</tr>
<tr>
<td>Acompmate phenotype ontology</td>
<td>phenotype</td>
<td>APO</td>
<td>acompmate_phenotype.obo</td>
<td>2012/03/01</td>
</tr>
<tr>
<td>Basic Formal Ontology</td>
<td>upper</td>
<td>BFO</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Bilatera anatomy</td>
<td>anatomy</td>
<td>BILIA</td>
<td>bilatera_mrc.obo</td>
<td></td>
</tr>
<tr>
<td>Biological imaging methods</td>
<td>experiments</td>
<td>PRIN</td>
<td>image.obo</td>
<td></td>
</tr>
<tr>
<td>BRENDA tissue / enzyme source</td>
<td>anatomy</td>
<td>BTO</td>
<td>BredaTissueOB0</td>
<td></td>
</tr>
<tr>
<td>Caenorhabditis development</td>
<td>anatomy</td>
<td>WBSi</td>
<td>worm_development.obo</td>
<td></td>
</tr>
</tbody>
</table>
Search BioPortal for “hepatocellular necrosis”
WIKI PATHWAYS

Welcome to WikiPathways BETA

WikiPathways is an open, public platform dedicated to the curation of biological pathways by and for the scientific community. More about WikiPathways...

Finding Pathways

Search

You can search by:
- Pathway name (Apoptosis)
- Gene or protein name (p53)
- Any page content (cancer)

Browse

Browse by species and category

Contributing New Pathways

Create

Create a new pathway page

Suggest

Add a pathway to the wish list

Sample Pathway Pages

Today's Featured Pathway

Signal Transduction of S1P Receptor (Homo sapiens)

Latest discussions

14 June 2012
- Great start! (1) by Alexander Pico

23 May 2012
- Good start! (1) by Alexander Pico

27 April 2012
- Maybe include effect on stem cell lineage and mechanical effects (1) by Tyler Davis
  - more...
LINKED OPEN DATA CLOUD

http://www.w3.org/DesignIssues/LinkedData.html
SPARQL: http://linkedlifedata.com/sparql
BIOINFORMATICS

- Proliferation of databases
- Compatibility is an issue
- Many (relatively recent) open data initiatives / standardisation efforts.
- **ISA-TAB**: Susanna-Assunta Sansone et. al., Toward interoperable bioscience data, Nature Genetics 44, 121–126 (2012)

[http://isa-tools.org](http://isa-tools.org)
CHEMINFORMATICS

- Historically, the cheminformatics world has been driven by **de facto** standards, developed and proposed by different vendors.
- Examples: SDF, MOL, SMILES, PDB
- **No agreed way to modify or extend the formats!**
- A number of initiatives (relatively recent), have adopted open standardisation procedures
- Examples: InChI, CML, BlueObelisk initiatives, ToXML
- No requirements for independent interoperable implementations so far
## DATA STANDARDS (LACK OF)

<table>
<thead>
<tr>
<th>ID</th>
<th>CAS</th>
<th>NAME</th>
<th>WA</th>
<th>Mv</th>
<th>H-073</th>
<th>nCb-</th>
<th>MAXDP</th>
<th>nN</th>
<th>Log1/LC50 Exp</th>
<th>Y-Pred.</th>
<th>Hat</th>
</tr>
</thead>
<tbody>
<tr>
<td>73</td>
<td>00091-66-7</td>
<td>N,N-Diethylaniline</td>
<td></td>
<td>0.58</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.034</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>0.333</td>
<td></td>
<td>3.959</td>
<td>3.51</td>
<td></td>
</tr>
</tbody>
</table>

OpenBabel11300911173D

```
26 26 0 0 0 0 0 0 0 0 0999 V2000
  0.2812  0.8575 -0.6609 N 0 0 0 0 0 0
-0.3126  2.0484 -1.0695 C 0 0 0 0 0 0
```

SDF files come in many different flavours
DATA STANDARDS (LACK OF)

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

   ActivityOutcome
active
unspecified/blank
inactive

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

   Canc
3 = carcinogen;
2 = equivocal;
1 = noncarcinogen

QMRF
NAME 0616091047
56 55 0 0 0 0 0 0 0 0 0999 V2000
0.0000 0.0000 0.0000 C 0 0 0 0 0 0 0
0 0 0 0
1.4940 0.0000 0.0000 C 0 0 0 0 0 0 0
0 0 0 0
.... [output skipped]
20 55 1 0 0 0 0
20 56 1 0 0 0 0
M END

> <Eye irritation SP pred.>
4.380000114440918

> <CAS>
30399-84-9

> <Type>
Testing
$$$$

No common meaning of SDF fields!
Each tool provides means to map into its internal semantic.
Every data field is a resource with unique URI and metadata assigned (not only name!)
DATA CURATION

“JUST structure-name validation is a long, torturous, iterative task.”


Approaches:
• Manual, crowdsourced
• A “standard” workflow - these may differ across toolkits and models!
• Compare with PubChem
• Compare as many sources as possible

“obviously, cheminformaticians must only use correct chemical structures and biological activities in their studies”
OPENTOX DATABASE QUALITY ASSURANCE

<table>
<thead>
<tr>
<th>Automatic Classification</th>
<th>Initial Quality Label Assigned</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consensus</td>
<td>OK</td>
</tr>
<tr>
<td>Majority</td>
<td>Probably OK for the structure that belongs to the majority</td>
</tr>
<tr>
<td></td>
<td>Probably ERROR for the structure(s) that belong(s) to the minority</td>
</tr>
<tr>
<td>Ambiguous</td>
<td>Unknown (multiple sources)</td>
</tr>
<tr>
<td>Unconfirmed</td>
<td>Unknown (single source)</td>
</tr>
</tbody>
</table>
QMRF DATABASE : NOT ERROR FREE!
CHEMICAL STRUCTURE REPRESENTATION

IT DEPENDS!

This is a chemical structure, represented by SMILES c1ccc2c(c1)ccc(=O)o2. Please enter the number of aromatic bonds.

- 0: 21%
- 6: 43%
- 10: 7%
- 11: 21%
- 12: 7%

This is a chemical structure, represented by SMILES O2C1=CC=CC=C1C=C2=O. Please enter the number of aromatic bonds.

- 0: 14%
- 6: 57%
- 10: 14%
- 11: 14%
- 12: 0%

http://tinyurl.com/smilesquiz
CHEMICAL STRUCTURE REPRESENTATION

IT DEPENDS!

This is a chemical structure, represented by SMILES c1ccc2c(c1)ccc(=O)o2. Please enter the number of aromatic bonds.

- 0: 21%
- 6: 43%
- 10: 7%
- 11: 21%
- 12: 7%

This is a chemical structure, represented by SMILES O2C1=CC=CC1C=CC2=O. Please enter the number of aromatic bonds.

- 0: 14%
- 6: 57%
- 10: 14%
- 11: 14%
- 12: 0%

http://tinyurl.com/smilesquiz
PROLIFERATION OF NEW TOOLS THAT ARE RARELY ABLE TO TALK TO EACH OTHER.

- Chemistry & Biology software and databases may continue to live in their own worlds, unless we want data shared and tools interoperable.
- Interoperability / standards may affect business models.

THE SILO EFFECT

Silo storage system
- Designed to store one single type of grain.

Information silo
- Rigid design
- No easy exchange of information
- No integration with other systems
THE INTERNET & INTERNET STANDARDS

Internet Engineering Task Force (IETF) working groups have the responsibility for developing and reviewing specifications intended as Internet Standards. The process starts by publishing a Request for Comments (RFC) – the goal is peer review or to convey new concepts or information.

IETF accepts some RFCs as Internet standards via its three step standardisation process. If an RFC is labelled as a Proposed Standard, it needs to be implemented by at least two independent and interoperable implementations, further reviewed and after correction becomes a Draft Standard.

With a sufficient level of technical maturity, a Draft Standard can become an Internet Standard. Organisations such as the World Wide Web consortium and OASIS support collaborations of open standards for software interoperability.

The existence of the Internet itself, based on compatible hardware and software components and services is a demonstration of the opportunities offered by collaborative innovation, flexibility, interoperability, cost effectiveness and freedom of action.

An unique example of what society can achieve by adopting common standards
**SCIENTIFIC DATABASES IMPLEMENTATION**

- Identify the data model and functionality
- Translate the data model into a database schema
- Implement the database and user interface functionality
- (Optionally) provide libraries or expose (some) of the functionality as web services

**Advantages**

- Use one’s favourite technology and jump directly into implementation
- Attract end-users with nice GUI relatively quickly
- Relatively easy to persuade funding organisations this will be a useful resource

**Disadvantages**

- Proliferation of incompatible resources, providing similar functionality, but incompatible programming interface
- Difficult to extract and collate data automatically

This is not the only nor the best way!
DON’T LIVE IN THE PAST

“Twenty to thirty years ago, most applications were written to solve a particular problem and were bound to a single database. The application was the only way data got into and out of the database.

Today, data is much more distributed and data consistency, particularly in the face of extreme scale, poses some very interesting challenges”

http://www.zdnet.com/blog/microsoft/microsoft-big-brains-dave-campbell/1749
A COMMON API FIRST, MULTIPLE INDEPENDENT INTEROPERABLE IMPLEMENTATIONS LATER

Advantages:
- Compatibility! Facilitates collation of distributed resources!
- Avoid proliferation of incompatible resources (this however only makes sense if the API is adopted beyond a single implementation)
- Easy to develop multiple GUI applications, once the API/library functionality is in place

Disadvantages: Think first, then implement 😌.
- GUI comes last
- Harder to persuade funding organisations (because reviewers usually look for nice GUIs)

Integration means compatibility and interaction; NOT necessarily storing everything on a single place.
DECENTRALIZED INFORMATION INTEGRATION

- Distributed, yet sufficiently interoperable model for information access.
- The future convergence between cheminformatics and bioinformatics databases poses new challenges to the management and analysis of large data sets.
- Evolution towards the right mix of flexibility, performance, scalability, interoperability, sets of unique features offered, friendly user interfaces, programmatic access for advanced users, platform independence, results reproducibility, curation and crowdsourcing utilities, collaborative sharing and secure access.

Interoperability is a key.
WHY DOES INTEROPERABILITY MATTER

Facilitates:
- Data, models and prediction results comparison
- A key to decision making

No need to wait until the perfect standard emerges
- Annotate and link the sources
- The least powerful standard wins!

There will always be new databases and tools
- Let them talk to each other
THANK YOU!

Acknowledgments:

OpenTox

OpenTox REST API
http://opentox.org/dev/apis
Download AMBIT Implementation of OpenTox API and launch your OpenTox service
http://ambit.sourceforge.net