

Improving the Management of Computational Models

storage – retrieval & ranking – version control

**MARTIN SCHARM^{1,2}, RON HENKEL¹, DAGMAR WALTEMATH¹,
OLAF WOLKENHAUER^{1,3}**

¹Department of Systems Biology & Bioinformatics, University of Rostock

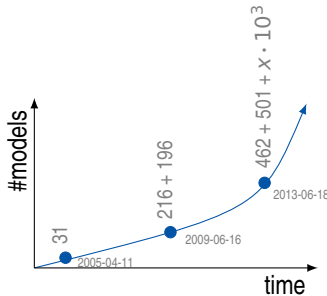
²Department of Computer Science, University of Oxford

³Stellenbosch Institute for Advanced Study, South Africa

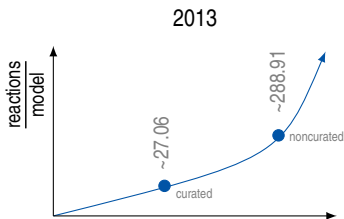
<http://sems.uni-rostock.de>

European Bioinformatics Institute
UK 2013

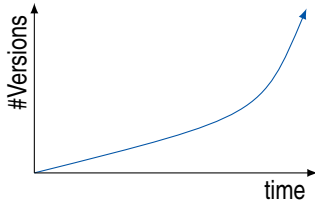




- Number of models is steadily increasing
- Models tend to get more complex
- Continuous development



- Number of models is steadily increasing
- Models tend to get more complex
- Continuous development



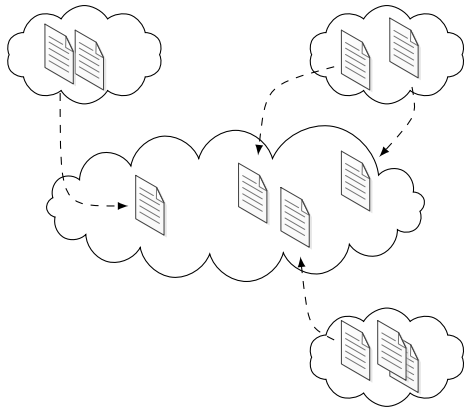
- Number of models is steadily increasing
- Models tend to get more complex
- Continuous development



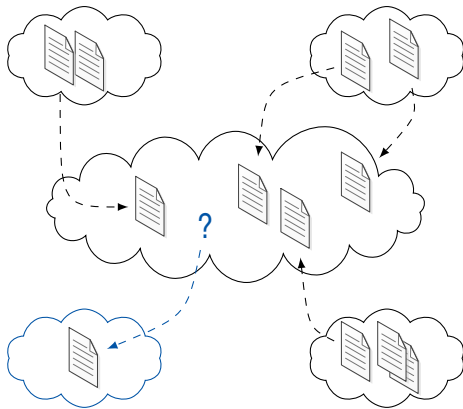
- Storage & Availability
- Search & Retrieval
- Version Control



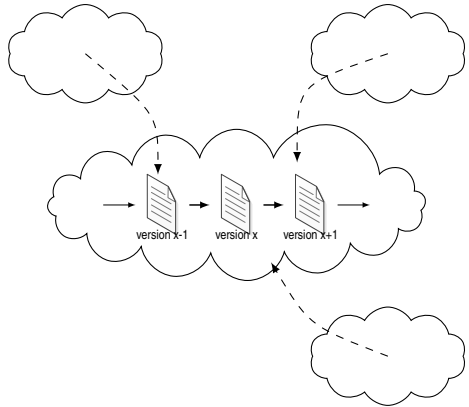
- **Storage & Availability**
- Search & Retrieval
- Version Control



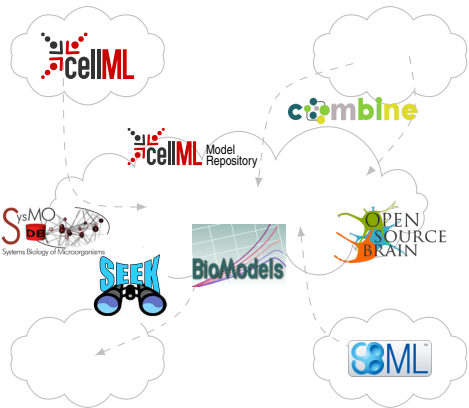
- Storage & Availability
- Search & Retrieval
- Version Control

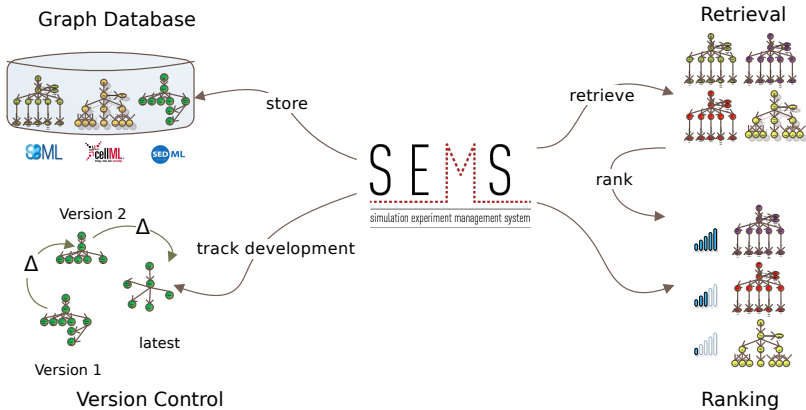


- Storage & Availability
- Search & Retrieval
- Version Control



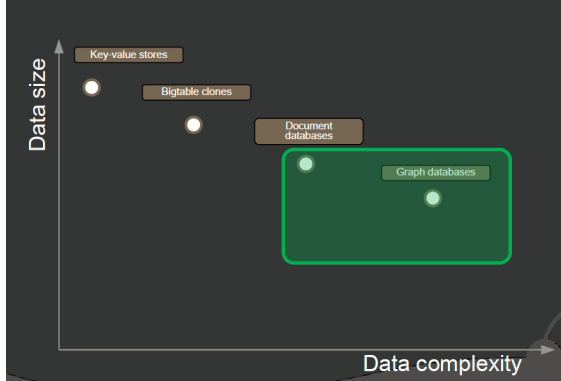
Several projects develop
methods to support the
community





<http://sems.uni-rostock.de/>

NOSQL data models



Taken from: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.174.8830&rep=rep1&type=pdf>

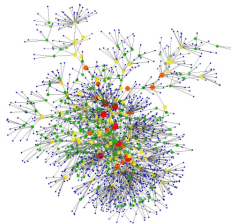
Model & Simulation



- Constituent names
- XML encoded representation



Annotation & Ontologies



- Biochemical background
- Synonyms



A model's network



- Model structure
- Aggregation values
- Model related data

Tyson 1991 - SBML



Model

Entities

Annotations

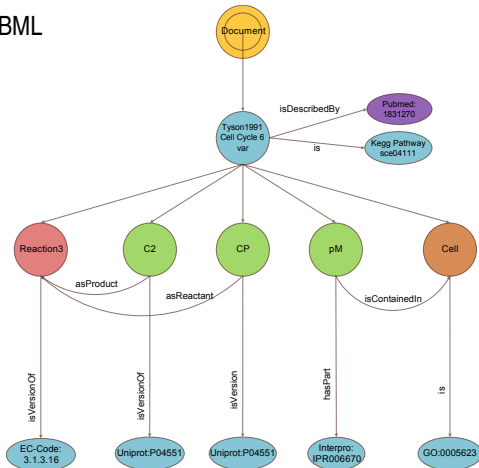


fig. courtesy R. Henkel

Tyson 1991 - CellML



Model

Component

Variable

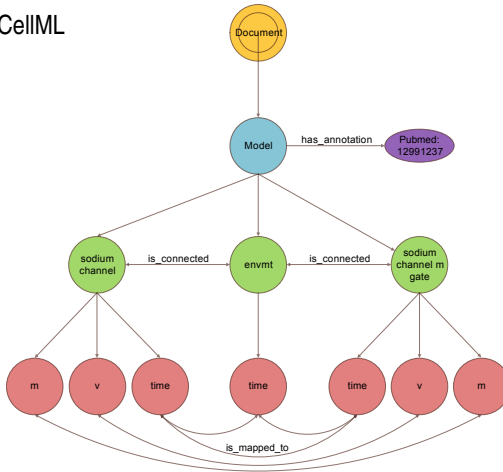


fig. courtesy R. Henkel

Tyson 1991 - SED-ML



Simulation

Entity

Variable

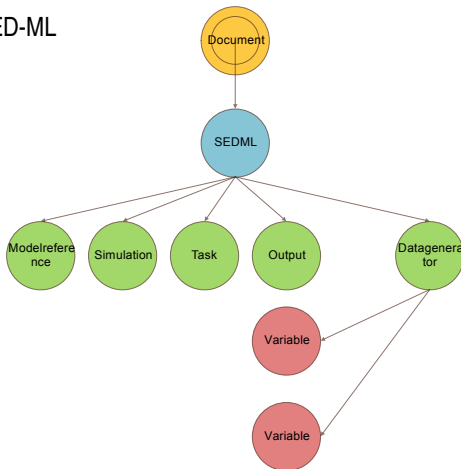
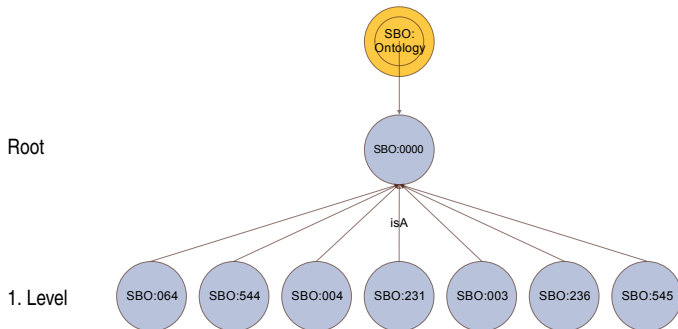


fig. courtesy R. Henkel

Systems Biology Ontology



All Ontologies available in OWL can be added easily

fig. courtesy R. Henkel

Models



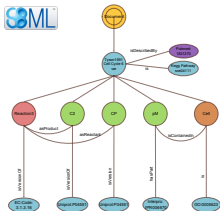
Model related data



the Gene Ontology



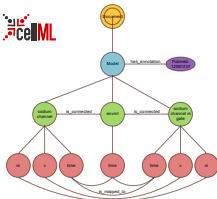
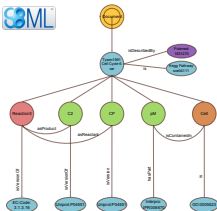
Models



Model related data



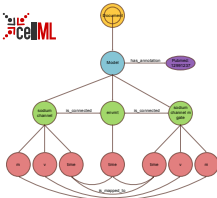
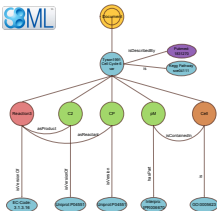
Models



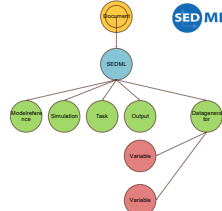
Model related data



Models



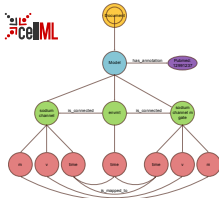
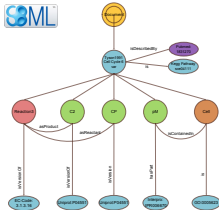
Model related data



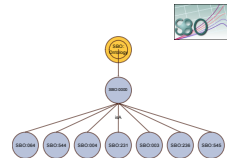
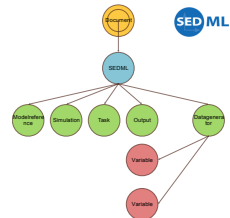
the Gene Ontology



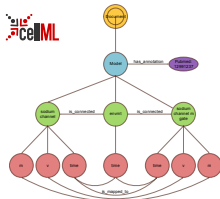
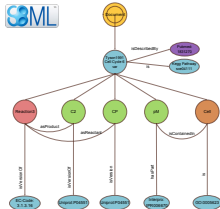
Models



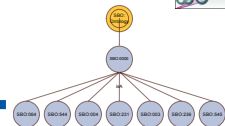
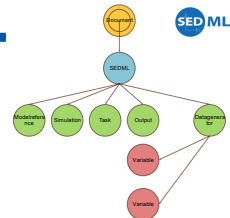
Model related data



Models

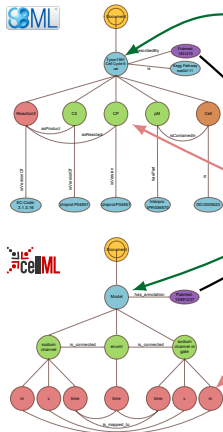


Model related data

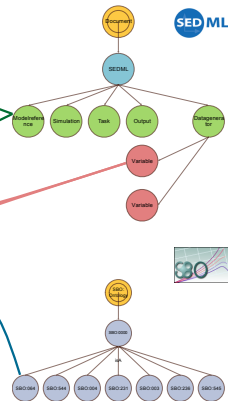


Neo4j
the world's leading
graph database

Models

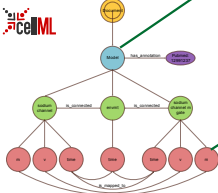
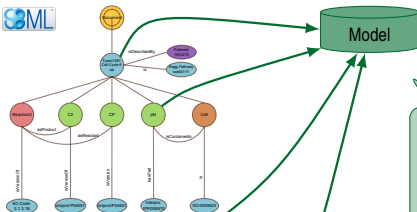


Model related data



Henkel et al. 2013, Graph-based storage of computational biology models and associated simulations, in preparation

Models

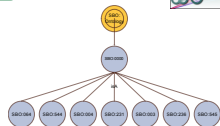
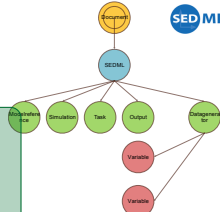


Index

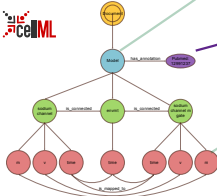
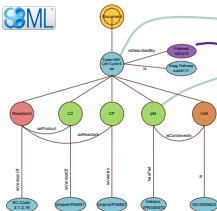


Id
Name
Component
Variable
Species
Reaction
Compartment

Model related data



Models

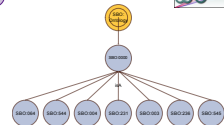
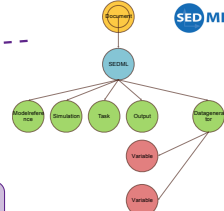


Index

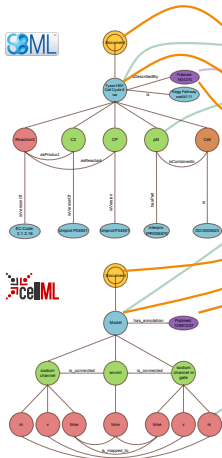


Title
Journal
Abstract
Authors

Model related data



Models

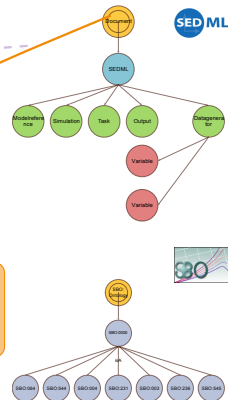


Index



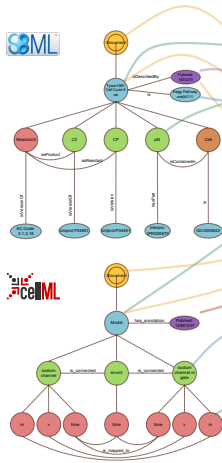
First name
Last name
Organization
Email

Model related data

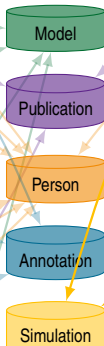




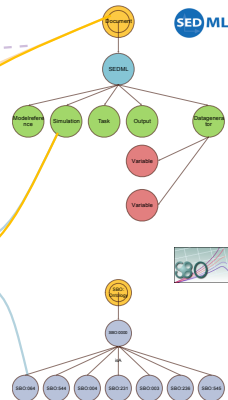
Models



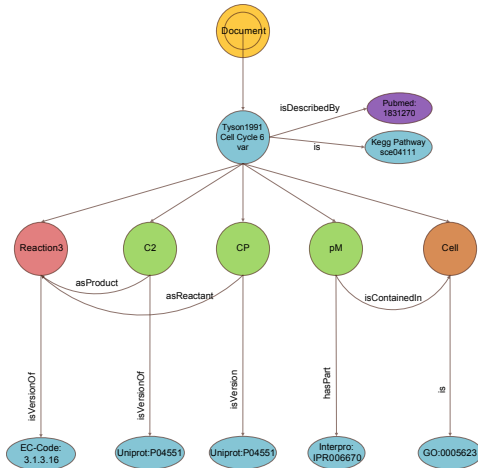
Index



Model related data

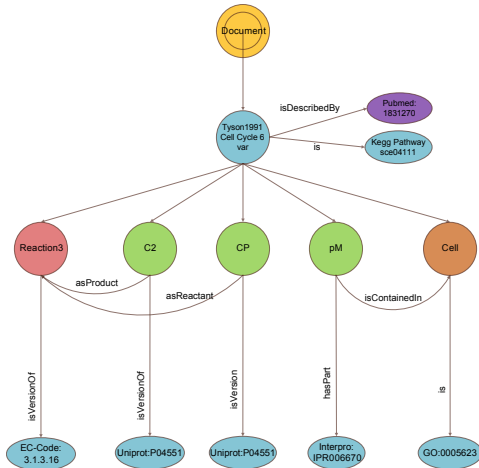


Id
Name



Show me models by
Tyson describing the cell
cycle and have cdc2!

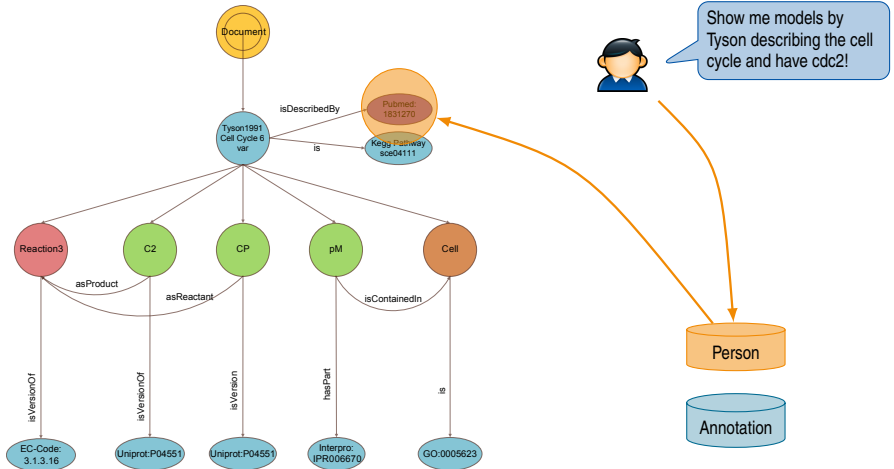


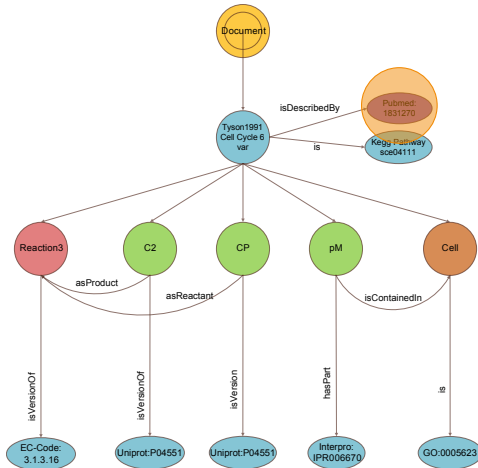


Show me models by
Tyson describing the cell
cycle and have cdc2!

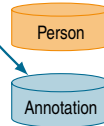
Person

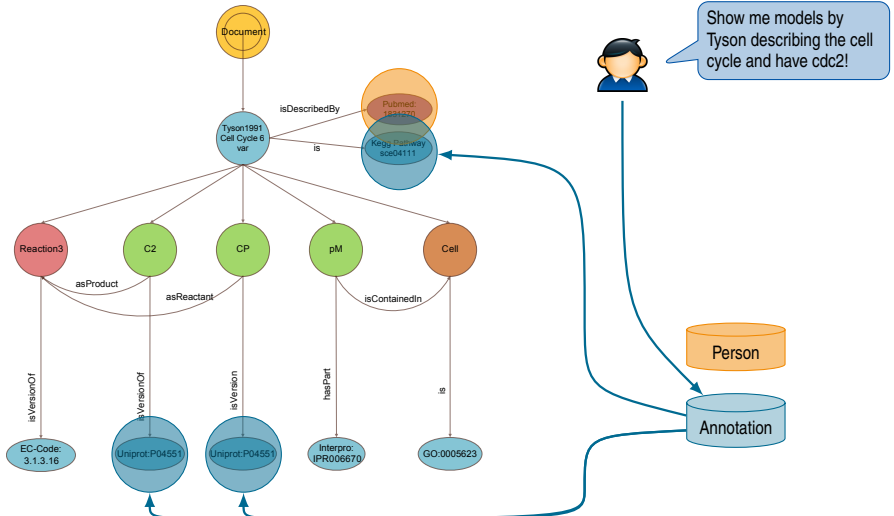
Annotation

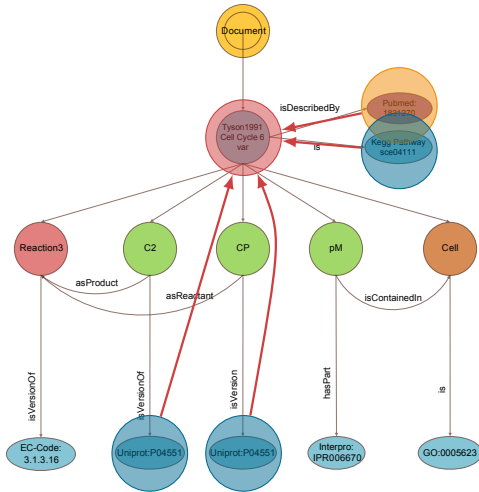




Show me models by
Tyson describing the cell
cycle and have cdc2!

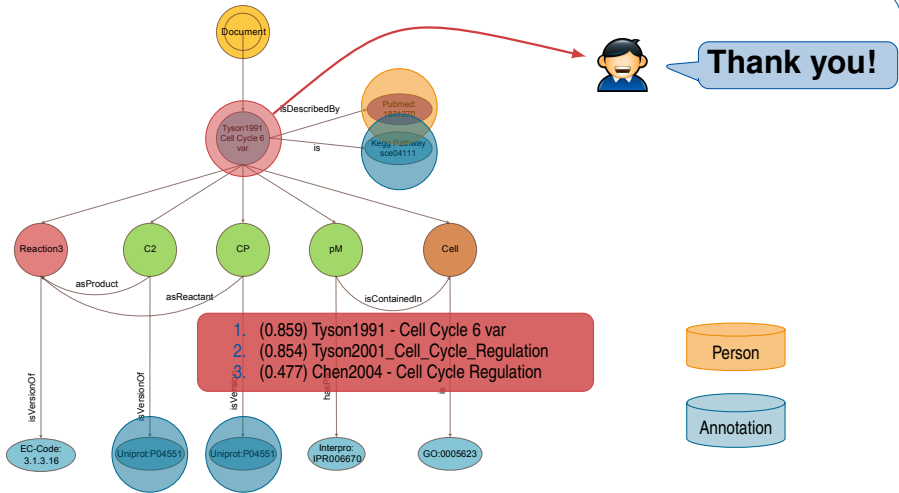


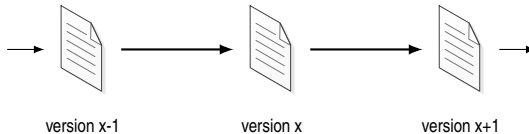


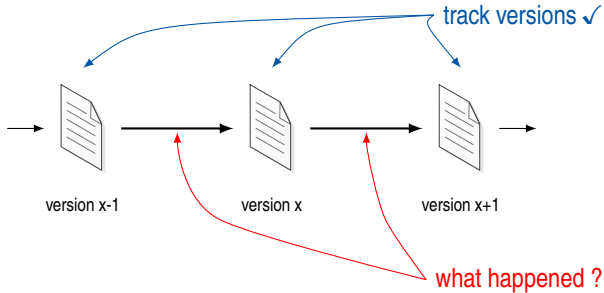


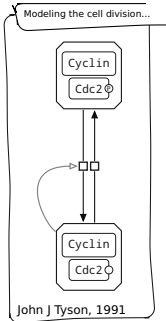
Show me models by
Tyson describing the cell
cycle and have cdc2!

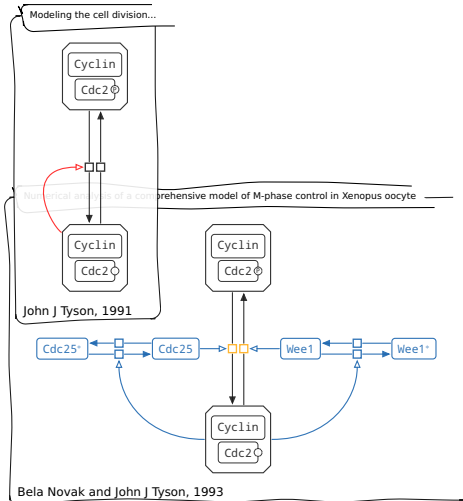


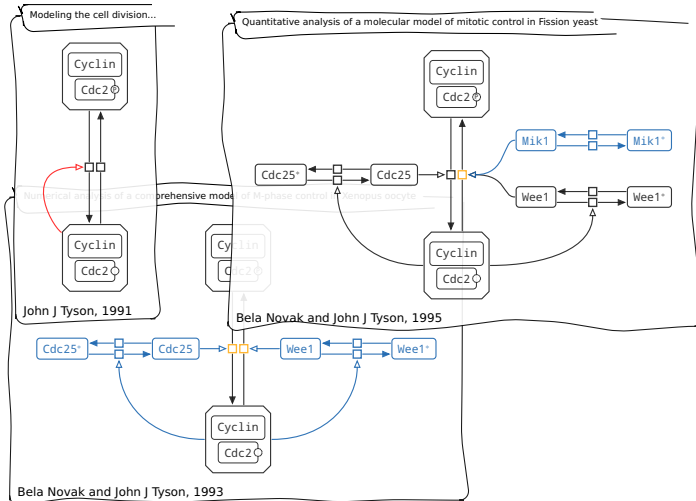


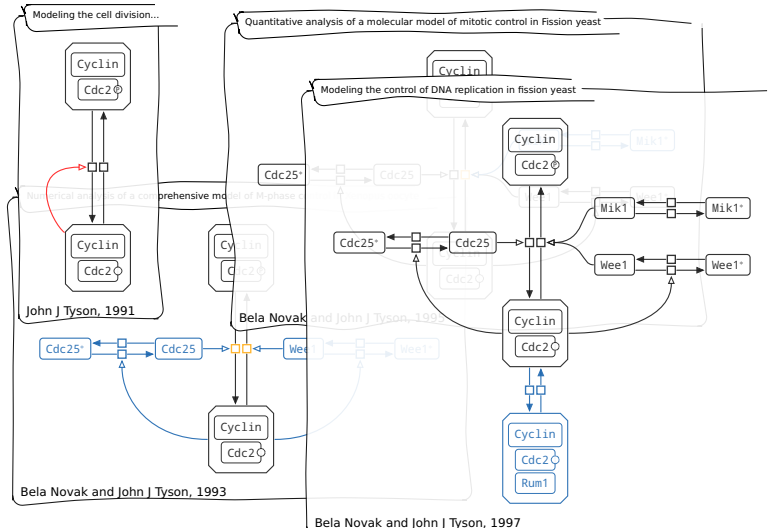


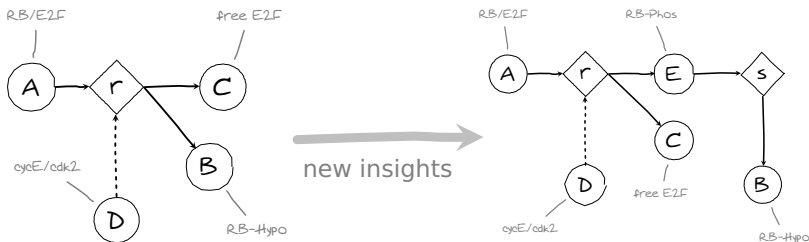




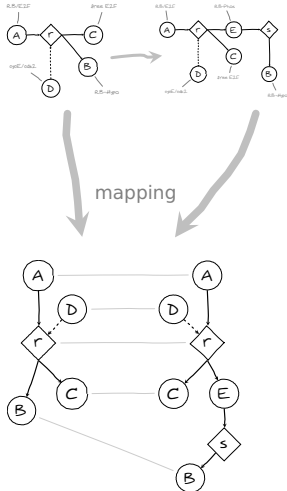






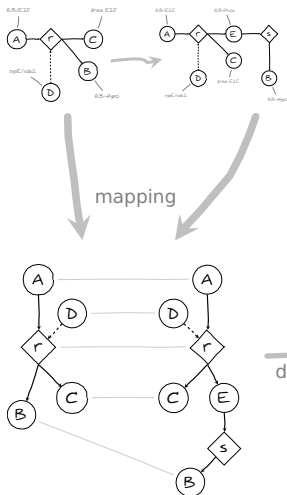


Waltemath *et al.*: Improving the reuse of computational models through version control. *Bioinformatics* (2013) 29(6): 742-728;



Biochemical Model Version Control System

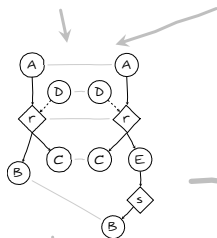
- compares models encoded in standardized formats (currently:  and )
- maps hierarchically structured content



Biochemical Model Version Control System

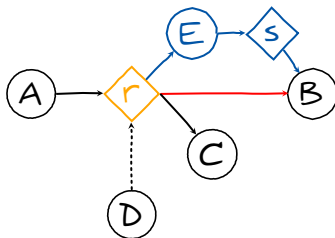
- compares models encoded in standardized formats (currently:  and )
- maps hierarchically structured content
- constructs a diff (in XML format)
- is able to interpret this diff

```
<XML>
Diff
├── moves
│   └── product of r: C
├── deletes
│   └── product of r: B
├── inserts
│   ├── species: E
│   ├── product of r: E
│   └── reaction s
</XML>
```

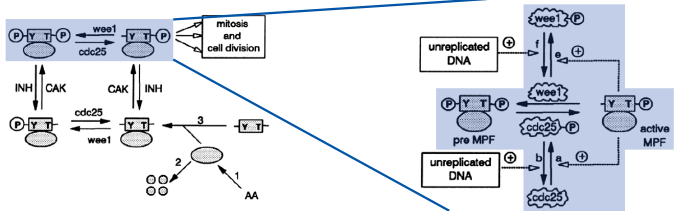


```
<XML>
Diff
├── moves
│   └── product of r: C
├── deletes
│   └── product of r: B
├── inserts
│   ├── species: E
│   ├── product of r: E
│   └── reaction s
</XML>
```

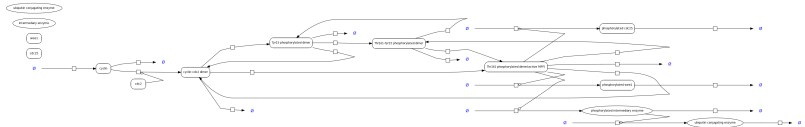
- calls BiVeS to construct the diff
- displays the result in various formats
 - the XML diff
 - a reaction network highlighting the changes using 
 - a human readable report



Novak & Tyson '93: Numerical analysis of a comprehensive model of M-phase control in *Xenopus* oocyte extracts and intact embryos



Biomodels Database [BIOMD000000107](https://identifiers.org/Biomodels/BIOMD000000107):



Fixed the model → How to communicate the changes?

Unix diff?

```
1c1
< <?xml version='1.0' encoding='UTF-8' standalone='no'?>
---
> <?xml version="1.0" encoding="UTF-8" standalone="no"?>
```

287 lines...

Fixed the model → How to communicate the changes?

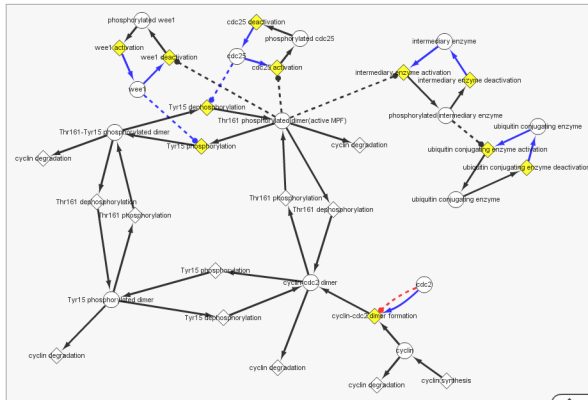
BiVeS Diff?

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<bives type="fullDiff">
  <update/>
  <delete>
    <node id="1" oldChildNo="1" oldParent="/listOfReactions[1]/reaction[1]" />
    <attribute id="2" name="rdf:resource" oldPath="/listOfReactions[1]/reaction[1]" />
    <node id="5" oldChildNo="4" oldParent="/listOfReactions[1]/reaction[1]" />
    <node id="6" oldChildNo="1" oldParent="/listOfReactions[1]/reaction[1]" />
    <attribute id="7" name="species" oldPath="/listOfReactions[1]/reaction[1]" />
  </delete>
  <insert>
    <attribute id="3" name="metaid" newPath="/listOfReactions[1]/reaction[1]" />
    <attribute id="4" name="metaid" newPath="/listOfReactions[1]/reaction[1]" />
    <attribute id="9" name="metaid" newPath="/listOfReactions[1]/reaction[1]" />
    <node id="10" newChildNo="1" newParent="/listOfReactions[1]/reaction[1]" />
  </insert>
</bives>
```

much smarter, but still 24 lines..

Fixed the model → How to communicate the changes?

BiVeS Diff: Highlighted chemical reaction network



Fixed the model → How to communicate the changes?

BiVeS Diff: Report

SBML Differences

- Both documents have same Level/Version: **L2V1**

Reactions

R19 (wee1 deactivation)

- wee1 → wee1_p
- Modifiers: dimer_p(unknown);

R17 (cdc25 activation)

- cdc25 → cdc25_p
- Modifiers: dimer_p(unknown);

R18 (cdc25 deactivation)

- cdc25_p → cdc25

R13 (Tyr15 dephosphorylation)

- p_dimer_p → dimer_p
- Modifiers: cdc25(unknown)

R12 (Tyr15 phosphorylation)

- dimer_p → p_dimer_p
- Modifiers: wee1(unknown)

R24 (ubiquitin conjugating enzyme deactivation)

- UbE_star → UbE

R23 (ubiquitin conjugating enzyme activation)

- UbE → UbE_star
- Modifiers: IE_p(unknown);

R22 (intermediary enzyme deactivation)

- IE_p → IE

R21 (intermediary enzyme activation)

- IE → IE_p
- Modifiers: dimer_p(unknown);

R20 (wee1 activation)

- wee1_p → wee1

R3 (cyclin-cdc2 dimer formation)

- cyclin + cdc2 → dimer
- Modifiers: cdc2(unknown)



Biomodels Database

963	# Models
10193	# Versions
9148	# prod. Diffs

CellML Model Repository

Biomodels Database

CellML Model Repository

963	# Models
10193	# Versions
9148	# prod. Diffs

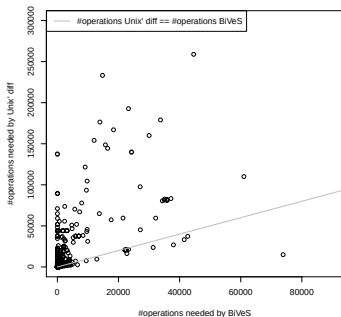
BiVeS	Unix' Diff	
2233	1890	# empty Diffs
434.825	1977.210	Ø #ops

Biomodels Database

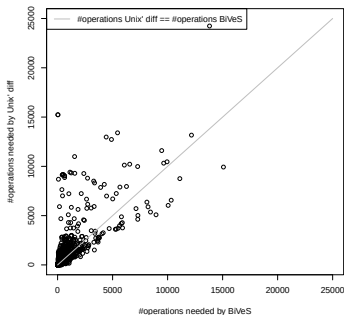
CellML Model Repository

	963	# Models	1589	
	10193	# Versions	12878	
	9148	# prod. Diffs	12936	
BiVeS	Unix' Diff		BiVeS	Unix' Diff
2233	1890	# empty Diffs	10113	10043
434.825	1977.210	Ø #ops	349.5282	480.249
6915 non-empty diffs		following	2823 non-empty diffs	

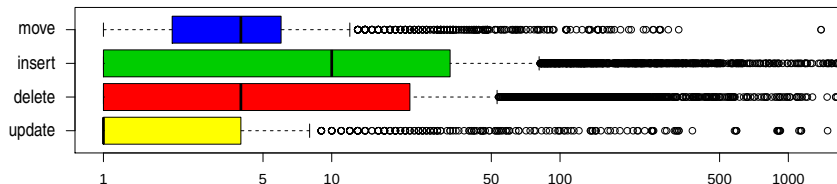
Biomodels Database



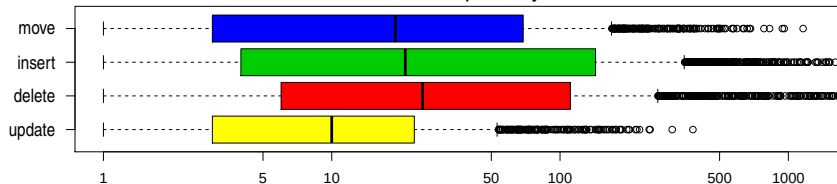
CellML Model Repository



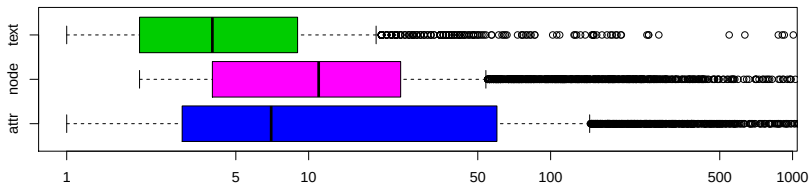
Biomodels Database



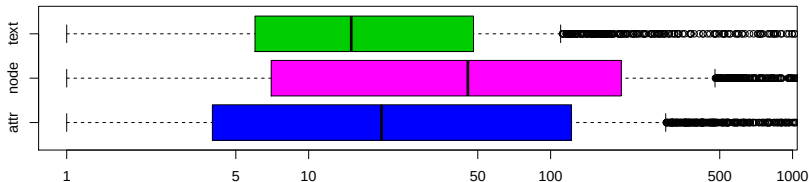
CellML Model Repository



Biomodels Database



CellML Model Repository





- SEMS develops methods for improved model management
 - Storage
 - Retrieval & Ranking
 - Version Control
- ... and implements these methods in tools that are
 - modularly designed
 - open source
 - easy to integrate

- Support for further standards
 - PharmML
 - CombineArchive
 - SBGN
 - respect further ontologies in storage/search
- Improved version control
 - mapping w/ respect to biological meaning
 - scoring schemes for differences
 - filter for changes
- Improved availability and analyzability of data

Thank you for your attention!

SEMS group

Dagmar Waltemath

Ron Henkel

Martin Peters

Markus Wolfien

Rebekka Alm

Olaf Wolkenhauer



@SemsProject

<http://sems.uni-rostock.de>



jvm

network

cmd

```
import de.uni-rostock.sems.bives.api.SBMLDiff;

[...]

SBMLDiff differ = new SBMLDiff (sbmlFileA, sbmlFileB);
differ.mapTrees ();
String graph = differ.getCRNGraphML ();

[...]
```

jvm

network

cmd

```
curl -d '{
  "get":
  [
    "documentType",
    "xmlDiff"
  ],
  "files":
  {
    "versionA": "http://your.db/path/to/versionA.sbml",
    "versionB": "http://your.db/path/to/versionA.sbml"
  }
}' http://bives.server.tld
```

jvm

network

cmd

```
java -jar BiVeS.jar path/to/versionA path/to/versionB
```



```
git diff 88fealcddf b64477d742 model.file
```

delete a single line in an XML:

```
<variable name="T" units="kelvin" initial_value="295" public_interface="out"/>
```

for Unix' diff it is 1 operation. Bives detects 5 operations (delete node `variable` and 4 attributes).

Worse if multiple nodes on the same line:

```
<para><emphasis><ulink url="caffeine.zip">Caffeine</ulink></emphasis> - Caffeine-induced Ca2+ transients from 18 i
```

And worst if whole document only 1 line → max. 2 operations for the Unix diff tool.

Single file containing models and model related data:

- simulation experiment descriptions in SED-ML
- graphical representations in SBGN-ML

necessary for the description of a model and all associated data and procedures
→ simplifies exchange

Proposal by Nicolas Le Novère, Frank Bergmann, and Richard Adams

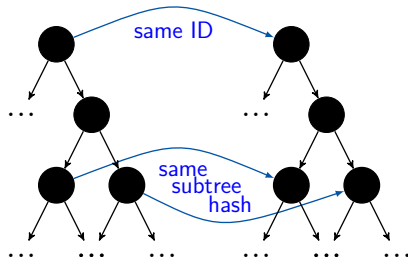
<http://co.mbine.org/documents/archive>

We developed a JAVA library to create/read/manipulate CombineArchives

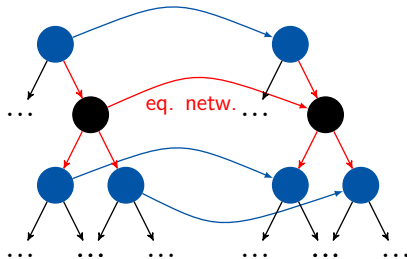
<http://sems.uni-rostock.de/projects/combinearchive/>

→ database import/export, difference detection between whole archives

- input: 2 XML documents
- map unambiguous nodes/subtrees
- propagate mapping up/down
- exploit further knowledge of biological systems



- input: 2 XML documents
- map unambiguous nodes/subtrees
- propagate mapping up/down
- exploit further knowledge of biological systems



- input: 2 XML documents
- map unambiguous nodes/subtrees
- propagate mapping up/down
- exploit further knowledge of biological systems

