

# SBML2SMW: bridging System Biology with semantic web technologies for biomedical knowledge acquisition and hypothesis elicitation

Tobias Mathäβ†, Peter Haase‡, Hiroaki Kitano\*, Luca Toldo†

‡fluid Operations GmbH, Walldorf, Germany; \*The Systems Biology Institute, Japan; Sony Computer Science Laboratories, Japan; Okinawa Institute of Science and Technology, Japan; †Merck KGaA, Darmstadt, Germany;

## ABSTRACT

**Motivation:** Generation of biomedical hypothesis is a very important task in the pharmaceutical industry since it serves the whole drug development pipeline: which are the physiopathological mechanisms underlying a disease, which biomarkers shall be measured in a clinical trial in order then to be able to do proper data mining and then deliver best service to the patients, which unexpected events could occur if one inhibits a certain molecular pathway, which could be new indications for a given compound. These are only few of the many questions which require discovery of hidden links. In this work we describe our experiences and a small tool we developed and make freely available which bridges in bidirectional way semantic wiki and system biology technologies. This CellDesigner plugin therefore enable easy share and reuse of knowledge. Source code available at <http://code.google.com/p/sbml2smw/>

## 1 INTRODUCTION

The invention of a new drug therapy is a high risk knowledge intensive process which lasts often a dozen of years and involves a large number of people from different organisations. Different kinds of knowledge are required(1): **Marketing** knowledge is needed to put a business in perspective and deliver business plans, **Medical** knowledge is needed to specify physiopathological processes involved in the disease; **Genetic** knowledge is needed to identify biomarkers to stratify the patients; **Biological** knowledge is needed to identify pathways and molecular entities to specifically “target” by limiting unexpected adverse events; **Chemical** knowledge is needed to identify the appropriate scaffolds to exploit; and many more. The same variety of knowledge types is encountered when one has the task of proposing new indication for a given compound (e.g. Cladribine for Multiple Sclerosis, although originally developed for Hairy Cell Leukaemia)

Although semantic technologies have since long passed the academic stage and are well exploited in the industry (e.g. planning of elevator cablings; aviation industry; soft-

ware and IT service industry; etc) the pharmaceutical industry still is exploiting them only in “vertical” scenarios and almost only as terminology resources and not as knowledge modelling resources (e.g. Gene Ontology in target discovery; MeDDRA terminology for coding adverse events).

More recently, several vendors have approached the pharmaceutical industry exploiting very large knowledge networks and offer them either as technology (e.g. GeneStruct, BioWisdom SOFIA suite, Cellucidate Rule Studio), or as knowledge repositories (e.g. Ingenuity Pathway Assistant; GeneGO MetaCore; Biobase Knowledge Library; GVK Bioscience Biomarker knowledge base), or exploit them within consultancy services (e.g. BMSystems; Life Biosystems GmbH; BioWisdom).

The Semantic Media Wiki(2) technology allows very flexible creation of knowledge bases, and collaborative sharing for knowledge, and it comes at no costs. Several implementations of wiki exists in biology (e.g. wikigenes, wikiproteins, wikipathways) having the aim of creating a “scientific wiki” having an emphasis more on genes or on protein or on “networks”. At the same time, Payao system(3) enables more systematic community-based annotation and curation with SBML and SBGN compliance. The Project HALO is showing how low-cost highly-scalable modeling of basic scientific knowledge in biology could be appropriately handled with Semantic MediaWiki (SMW).

System biology scientists have the purpose of qualitatively or quantitative modelling the biology and eventually physiology of human beings, and for this reason acquire knowledge in form of networks, which then are studied quali-quantitatively, therefore would very much need using the SMW, through their tool of choice, CellDesigner(4). However, at the moment no “bridge” was possible to exploit these 2 technologies together.

In this work we describe SBML2SMW: a small plugin which combines freely available state-of-the-art software from System Biology and open source semantic wiki technology in order to:

- a) enable semantically enriched, distributed, biomedical knowledge acquisition
- b) share and reuse knowledge networks

†To whom correspondence should be addressed.

in the context of the pharmaceutical biomedical hypothesis services.

## 2 SEMANTIC WIKI: NOT ONLY FRONTEND

Wikis have been rapidly established as collaborative tool for sharing information, reducing the burdening of community creation to simply filling boxes with content. Semantic wikis are wiki “with an underlying model of the knowledge described in its pages”. Their current exploitation in biology has increased rapidly (NETTAB2010 conference) however their intrinsic formal power has been rarely reported in depth in the biomedical domain.

Semantic Media Wiki (SMW) is a free extension of MediaWiki that adds semantic annotations allowing therefore the wiki to function as a collaborative database with Semantic Web- tagged content. The use of SMW for R&D in Pharma has already been reported e.g. for the purpose of self-service portal to compile reports for drug lots(5), and in the “Pfizerpedia”: an internal tool used in Pfizer for tracking project, people and knowledge focused on patent information(6). In our work, we use SMW as core biomedical knowledge base, performing both knowledge acquisition and hypothesis generation. Therefore, we do not use it only as “front-end” for user content, but also as “back-end” for storing and managing relations and entities that have been entered in the knowledge base by whatever mean (e.g. using the SMW import mechanisms and/or the CellDesigner “client”)

## 3 CELLDISIGNER: NOT ONLY CELL BIOLOGY

CellDesigner is the state-of-the-art structured diagram editor for drawing gene-regulatory and biochemical networks. Its intuitive user-interface helps drawing diagrams in rich graphical representation with personalized design. Networks are drawn based on a state transition diagram, proposed by Kitano and recent version comply with SBGN Process Description Diagram(7). Designed as a stand-alone tool, this powerful software is however network-aware, and therefore can connect to several major databases (DBGET, SGD, iHOP, Genome Network Platform, PubMed, Entrez Gene, SABIO-RK) and retrieve models from BioModels.net. The internal representation format that CellDesigner uses is the standard Systems Biology Markup Language (SBML) with CellDesigner specific annotation section to retain specific information needed for layout and other special features, and it has direct integration to the powerful Systems Biology Workbench (SBW) for performing quantitative simulations of cellular networks. Beyond system biology, CellDesigner can also be used for modelling system physiology(8) and it is projecting along those lines that we are exploiting it for improving our formal understanding on

physiopathological processes and rationalise drug discovery(1).

Although CellDesigner is publicly available, it is not open source, however through extension tag specification and through the API it can be very much extended without the need of changing the main source code.

In this work, we report the extension of CellDesigner with the plugin “SBML2SMW” that enables a bidirectional exploitation of the semantic content of the SMW, and allows the use of CellDesigner as front-end for entering data in SMW.

## 4 SBML ONTOLOGY: ENABLING KNOWLEDGE REUSE

To bridge between the internal data representation of the CellDesigner and the Semantic MediaWiki, we have developed an ontology for the representation of the SBML knowledge models to be stored and reused.

The ontology has been modeled in OWL. It is intentionally kept small and concise, covering exactly those aspects of SBML models that are intended to be reused by others.

We identified the types of entities, all relations between these entities and all the plain information necessary to reconstruct the relevant parts of SBML models. A graphical representation of the resulting ontology can be reviewed figure 1.

At the core of the ontology are the following four classes: Species, Reaction or Modification. We will describe these classes and the properties associated to these classes in further detail:

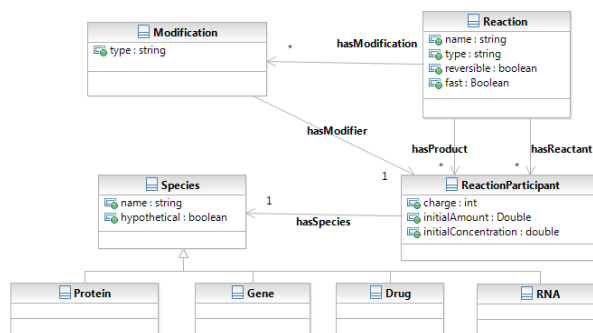


Figure 1: Overview of the SBML-ontology

- **Species:** A species is the top level type of all chemical compounds in a model. The species class has several subclasses, like Protein, Gene or Ion Channel, which allows a very detailed modeling and classification of these species.

- **ReactionParticipant** (Anm: evtl SpeciesAlias, aligned to CellDesigner-name): Since the same Species (e.g. a certain protein) can be associated to several reactions, a further abstraction layer between the Species itself and the reaction it contributes to has to be introduced. A ReactionParticipant links a species to a certain reaction and assigns it its specific initial amount or initial concentration and other reaction-specific properties.
- **Reaction**: The reaction class links a set of species, its reactants, with another set of species, its products. For each reaction, further information like the fact if it is reversible or if any modifications to the reaction are in play, can be modeled.
- **Modification**: A reaction optionally can have one or more associated modifications, e.g. if this reaction demands the presence of a catalyst. Such a modification associates a set of modifiers to the modified reaction and contains information about the type of this modification.

## 5 SBML2SMW: BIDIRECTIONAL PLUGIN

SBML2SMW enables users of CellDesigner to persist arbitrary information from a graphical model in CellDesigner and to make it available for other users having access to the underlying Semantic MediaWiki. This makes it possible to reuse the facts from any model stored by any user in any other CellDesigner model. To achieve this, we use a species-centric mapping from models stored in SBML to the ontology-based representation described in section 4. Using this representation, we are not only able to load complete models saved by a user back into CellDesigner, but we also can load context-dependent information, e.g. all the reactions in the database a certain species participates on. The elements from the ontology-centric representation are then mapped to a SMW representation. Therefore, all the entities are mapped to SMW pages, the links between these entities are stored as semantic links between these pages. This way, we preserve the whole semantics of the SBML-ontology-instance.

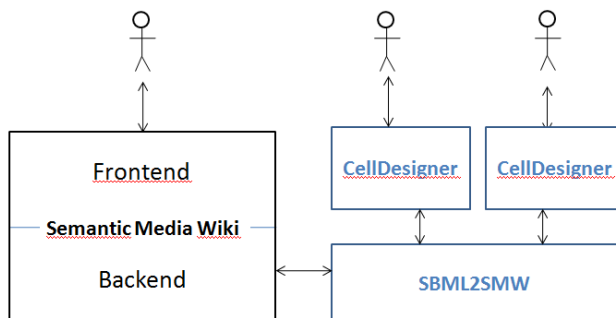


Figure 2: Conceptual Architecture

The users of this plugin should not only be able to use the store and load functionalities of the plugin completely transparently, but they also should be given the opportunity to add and correct information directly in the wiki.

Figure 2 shows the different ways of interaction our implementation supports. The SBML2SMW-plugin operates directly on the SMW backend, but data manipulation is also possible via the Webbrowser-based SMW frontend.

We now want to take a closer look on the functionality of the plugin from the user’s perspective:

**Store:** The user first designs his model in CellDesigner by drawing all the important species, interlinking them to model the reactions taking place with their participation, setting the intended properties for all the species and adding modifications to the reactions. Next, he decides which parts of the whole model are worth storing into SMW and sharing with his colleagues. He selects these parts (i.e. the species and reactions) in CellDesigner, starts the SBML2SMW plugin and uses the store function. The selected parts of the model are extracted from the model, translated in a OWL representation and written to the SMW.

**Load:** If another user starts to create a model in CellDesigner, he now has the ability to add a Species, select it in CellDesigner, start the plugin and “expand” this species. The plugin accesses the SMW store, finds the selected species and retrieves all the stored information, i.e. all the reactions having this species as a reactant or a product. The plugin loads all these reactions, with all their other reactants and products and renders them into the CellDesigner window, together with all the associations between them.

**Edit/Delete:** Change or delete operations are not supported by using the plugin. We decided that if information was stored by any user, it has its legitimation to remain in the database, even if it has been modified in a certain model by any user. Deletion of facts has to be performed directly on the corresponding pages in the wiki.

Figure 3 shows the store and load operations in CellDesigner. Figure 3a) shows a model designed by user A, he decided to store a Protein X and a State Transition Reaction transforming it into another Protein Y.

In Figure 3b) we see how another user B added the Protein X to an empty model, and expanded this protein. The before stored State Transition Reaction is retrieved from the database and added to B’s model

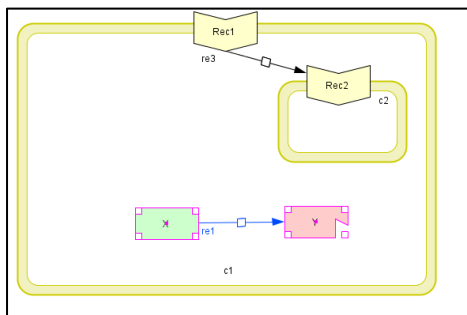


Figure 3a) CellDesigner model, X and Y selected to be saved

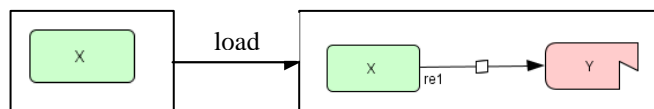


Figure 3b) before and after “load” on Protein X

## 6 CONCLUSIONS AND NEXT STEPS

In this work a free enabling technology is described, which extends the leading *free* system biology and system physiology CellDesigner platform by exploiting bi-directionally the leading *open source* Semantic Media Wiki technology. Knowledge can now be imported in the Semantic Media Wiki and made automatically available to the system modelers, and as-well they can share each of the relations they discover themselves (e.g. reading literature, or based on experiments or other observations), without having to interact with the SMW platform.

All the knowledge that is encoded in the CellDesigner pathways is now fully exposed to the SMW; and as well the whole knowledge contained in the SMW is now exposed to CellDesigner. This minor technological bridging thus removes any possible “knowledge gap” and maximises knowledge reuse both for an individual and for an organisation.

The technology we are making now freely available could be exploited through academic partners and have a public server sharing then properly formatted knowledge, thus seamlessly moving from a “wiki of science” to a “wiki of scientific knowledge”.

In spite of the appealing scenario that the work here reported is offering, we are aware that this report is at the moment only a development report. We are currently exploiting the technology here reported, and are confident in few months to be able to report on a detailed quantitative benefit/exploitation results.

## ACKNOWLEDGEMENTS

We would like to thank Lina Yup Sonderegger (Merck Serono) for reviewing the manuscript, Abdul Mateen Rajput

for extensive use of the tool and debugging it in real-case applications.

## REFERENCES

### LINKS:

- [http://semantic-mediawiki.org/wiki/Semantic\\_MediaWiki](http://semantic-mediawiki.org/wiki/Semantic_MediaWiki)  
<http://sbw.kgi.edu/>  
<http://www.projecthalo.com/>  
<http://www.wikigenes.org/>  
<http://www.wiki-proteins.org/>  
<http://www.nlm.nih.gov/research/umls/>  
<http://www.cellucidate.com/>  
<http://www.nettab.org/2010/>

### ARTICLES

1. Kant CS, Ibberson MR, Scheer A. Building a disease knowledge environment to lay the foundations for in silico drug discovery and translational medicine. *Expert Opinion on Drug Discovery*. 2010;5(2):117-22.
2. ; [cited]; Available from: <http://www.mediawiki.org/wiki/MediaWiki>.
3. Matsuoka Y, Ghosh S, Kikuchi N, Kitano H. Payao: a community platform for SBML pathway model curation. *Bioinformatics*. 2010 May 15, 2010;26(10):1381-3.
4. Funahashi A, Matsuoka Y, Jouraku A, Morohashi M, Kikuchi N, Kitano H. CellDesigner 3.5: A Versatile Modeling Tool for Biochemical Networks. *Proceedings of the IEEE*. 2008;96(8):1254-65.
5. OntopriseGmbH. An SMW+-based self-service portal for R&D in Pharma. Karlsruhe; 2010 [updated 2010; cited 2010 15-06-2010]; Available from: [http://smwforum.ontoprise.com/smwforum/index.php/R%26D\\_portal\\_in\\_pharma\\_industry](http://smwforum.ontoprise.com/smwforum/index.php/R%26D_portal_in_pharma_industry).
6. Walsh D, Berridge A, Gardner B, editors. *Pfizer-patents Semantic MediaWiki - The How, What, When, Who and Why of patents*. The International Conference for Science & Business Information; 2009 18-21 October 2009; Spain. Infonortics.
7. Noverre NL, Hucka M, Mi H, Moodie S, Schreiber F, Sorokin A, et al. The Systems Biology Graphical Notation. *Nat Biotech*. 2009;27(8):735-41.
8. Kitano H. Grand challenges in systems physiology. *Frontiers in Systems Physiology*. [Original Research Article]. 2010 2010-May-07;1.