# How I created a fully featured COMBINE archive

Martin Scharm

July 21, 2015

This is a small report on how I created the fully featured COMBINE archive github.com/SemsProject/CombineArchiveShowCase to demonstrate the power of the COMBINE archive approach. The developed COMBINE archive basically consists of (i) publication, (ii) model, (iii) simulation description, and (iv) simulation results. The data encoded in the archive has shown that the developed simulation experiment is able to reproduce the results of a published study.

## 1. Introduction

The steadily increasing size and complexity of models and derived data poses the challenge of sharing reproducible results [SWP+14b]. Unfortunately, problems with replication and reproducibility are quite common [PSA11, IAB+09, BE12]. Therefore, researchers became aware of the importance of reproducibility [SNTH13] and there is an increasing demand for means to transfer simulation studies, ensuring reproducibility [CVB+12, BRG+10, BAM+14]. Several projects and initiatives already deal with the problem of reproducibility, such as the Reproducibility Initiative[1] or FAIRDOM[2].
Within the past decade the Computational Modeling in Biology Network (COMBINE) developed several standard formats to encode the different aspects of a simulation study: SBML [HFS+03] and CellML [CLN+03] to encode the biological networks; SBGN [LNHM+09] to encode their visual representation; SED-ML [WAB+11] to encode the simulation recipes; NuML[3] and SBRML [DSPM10] to encode numerical data and simulation results. These markup languages allow for encoding single parts of a simulation study in an exchangeable format. However, a simulation study may consist of multiple files, and the very model might have been decomposed into various modules. Dagmar: It would be nice to have a

[1] reproducibilityinitiative.org
[2] fair-dom.org
[3] github.com/numl/numl

concrete example here, eg. the simulation study by XYZ has the following files (and describe in detail) - this would be motivating the work a lot better than simply the following sentence. Thus, it was still challenging to transfer reproducible research results.
To close this gap, the COMBINE community developed the COMBINE archive: A single file that aggregates all the information necessary for a modelling and simulation experiment in biology [BAM+14]. The skeleton of a COMBINE archive consists of a manifest and a meta data file, specified by the Open Modeling EXchange format (OMEX). When bundled in a COMBINE archive, essentially a .zip container, simulation experiments can be encoded in a reproducible manner.
To demonstrate the capabilities of a COMBINE archive I created a demo archive and made it available from our GitHub project[4]. In the following I describe what I did to build that archive. Dagmar: Would it be useful to name and shame provenance here?

# 2. Methods

I devide the process of developing a fully featured COMBINE archive into three major parts. First I decide for a simulation study to encode in an archive; then I create the initial archive; third I enrich the archive with information I found on the internet and with data that I generated on my own. These steps are described in the following.

## 2.1. Deciding for a Simulation Study

I developed 3 criteria to decide for a simulation study:

**Criteria 1: Open Access.** As I want to share the demo archive openly, I need to find an experiment that offers as much open data as possible. Obviously and unfortunately, the publication, which is an essential document for the documentation of experiments, is the bottleneck. Dagmar: I would argue that the actual bottleneck may be the data itself. However, it may not be necessary for the archive – depending on whether you want it to contain full provenance, in which case you'd also need to know where all the model parameters come from etc... and then the data become a much larger bottleneck than the publication, I think. Dagmar: Also, it would be nice to link to a study on percentage of open access publication here, if you find one. Therefore, I have to find a simulation experiment which was published using an open access license.

**Criteria 2: Much data already available in standard formats.** As I do not want to encode the model myself, I will only consider studies which are already available as SBML and CellML models. Ideally, these models are already curated, which increases my trust in the encoding.

[4] github.com/SemsProject/CombineArchiveShowCase

**Criteria 3: I have ideally already dealt with that publication.** As I will need to work with the study I need to understand it. It usually takes a lot of time to dig into a new field, so I prefer studies that I already investigated. However, this is just a soft criteria to decrease my workload.

**Searching for a matching study.** Dagmar: I suggest making this paragraph a section - Otherwise people may get confused due to ¿ 3 paragraphs from three criteria... Searching for a study was harder than I thought. I failed to encode my search criteria in a format consumable by the search engines of current databases. I ended up asking a common search engine to look for names of open access journals at the websites of the databases and eventually `site:models.cellml.org "Molecular Systems Biology"`[5] resulted in a model that is available from the CellML model repository (Calzone, Thieffry, Tyson, Novak, 2007[6]) [LL+08] and from the BioModels Database (BIOMD0000000144[7]) [LDR+10].

**The final study.** The study I chose was published by Calzone *et. al.* in Molecular Systems Biology. They propose a dynamical model for the molecular events underlying rapid, synchronous, syncytial nuclear division cycles in *Drosophila* embryos [CTTN07]. Dagmar: I would include a bit more information on the study here, or a sketch of the SBGN map (which presumably exists), to keep people's attention. In an earlier study dealing with the cell cycle I already touched that publication, so it was the perfect study for the demo archive project.

## 2.2. Creating an initial COMBINE archive

I created an initial version of the COMBINE archive using M2CAT [SW15]. The web interface at m2cat.sems.uni-rostock.de searches in MASYMOS, a graph database to retrieve links to models published in open databases [HWW15]. The search for `Calzone` resulted in two matches, one of them representing the model in the CellML model repository. In addition to searching in a graph database, M2CAT retrieves files that correspond to a simulation study. For this it includes other sources, such as open model repositories, and bundles them in COMBINE archives using the library of the CombineArchive Toolkit [SWP+14a]. Moreover, the web interface provides a link to conveniently explore the generated archive in the CombineArchiveWeb application [SWP+14b].
In case of the Calzone *et. al.* study, the files of the CellML model repository were cloned and aggregated in a new COMBINE archive. Additionally, the resulting archive and its files were annotated with all the meta data available from the GIT project of the CellML model repositories. Specifically, the archive was annotated with a description informing that it was generated by M2CAT. The files that were cloned from the CellML model repository were annotated with creators, contributors, and modification times as available from the corresponding GIT project (`git log`), see Figure 1. Thus, the initial version of the COMBINE archive was obtained automatically.

[5] duckduckgo.com/?q=site%3Amodels.cellml.org+%22Molecular+Systems+Biology%22
[6] models.cellml.org/exposure/1a3f36d015121d5596565fe7d9afb332
[7] www.ebi.ac.uk/biomodels-main/BIOMD0000000144


**Figure 1: The CombineArchiveWeb application showing the archive as it was exported from M2CAT.** The files were cloned from the CellML model repository and immediately annotated with meta data. Thus, information such that Catherine Lloyd is the creator of the model file or the creation date are not lost.

## 2.3. Extending the COMBINE archive

As the initial version only contains the model encoded in CellML together with some figures (all files exclusively from the CellML model repository) I needed to extend the archive manually. To organize the files I developed a file structure in the COMBINE archive containing the following directories:

- `model/`: files that encode and visualise the biological system
- `experiment/`: files that encode the *in silico* setup of the experiment
- `documentation/`: files that describe and document the model and/or experiment
- `result/`: files that result from running the experiment

The CombineArchiveWeb application made it very easy to get rid of the unrelated `.DS_Store` file. It also helped with moving all other files into the model directory, as they all encoded the model. Other directories are to be filled in the next sections.

### 2.3.1. Retrieving Data and Information from other Services

The first part of extending the archive dealt with a search for resources related to this study on the internet. I was able to retrieve a journal publication and a model encoded in SBML.

**The article** is usually the central object of a research study. As I said earlier, Calzone *et. al.* published their findings in Molecular Systems Biology. So I went to the corresponding website[8] and downloaded the article and supplementary information. As both describe the simulation experiment, I uploaded the files to the `documentation/` directory in the CombineArchiveWeb application. The web interface decently added meta data listing me as the creator of the file, which, unfortunately, is in this case incorrect. Thus, I modified the entries to attribute the real creators and to state when and where the files were downloaded. Additionally, I added modification dates to the meta data of these files by copying them from the website of the article and the meta data of the PDF files. The CombineArchiveWeb application provides a nice interface to modify the meta data without any necessity of touching or knowing about XML or RDF. However, in the background it nevertheless created an RDF/XML tree to describe the article in a machine readable format, see section A.2. This XML tree was then added as a subtree to the meta data file of the archive.

**The model** encoded in SBML format was already available from the BioModels Database, so I went there to retrieve the file `BIOMD0000000144.xml`[9] (SBML level 2 version 1). I then uploaded the SBML file to the `model/` directory in the CombineArchiveWeb application and the meta data was modified to attribute the correct authors, curators, and

[8] msb.embopress.org/content/3/1/131
[9] www.ebi.ac.uk/biomodels-main/download?mid=BIOMD0000000144

contributors, according to the BioModels Database website[10] and as stated in the model document.

### 2.3.2. Generating more Data

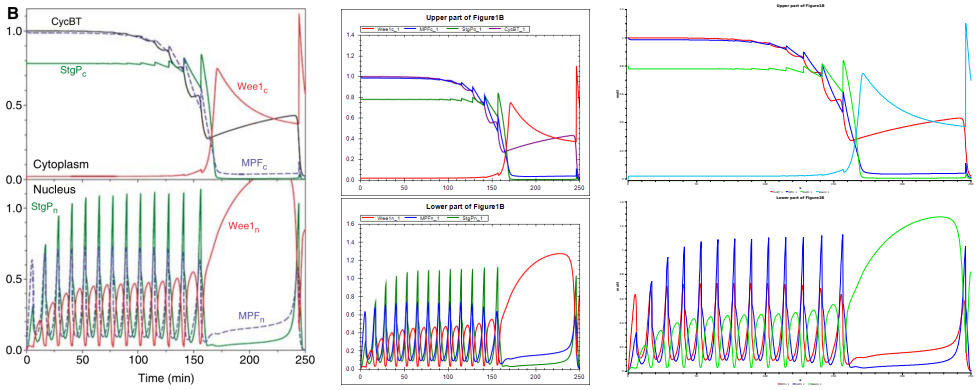
The model files and the journal article are obviously not sufficient to reproduce the experiment's results. However, I was not able to find any other information on the internet. Thus, I generated more data myself.

**The simulation description** is essential to run the experiment. It defines the environment and the output of the *in silico* execution. Creating an initial simulation description was easy using the SED-ML Web Tools[11] (SWT). I just needed to upload the model and to tick some boxes and the SWT created an initial SED-ML script which runs a time course simulation on the model. Dagmar: I guess "tick some boxes" demands some further details in order to be reproducible... Dagmar: Also, add KiSAO ID here for more transparency on the used algorithm In the end it produces a graph for the concentration of every species and the value of every parameter and combines everything in a large table. So I ended up with 66 plots and a huge table. I compared the graphs with those shown in the paper and sensed a match. As I considered this a good start I uploaded the SED-ML `Calzone2007-simulation-all-figures.xml` script to the `experiment/` directory in the CombineArchiveWeb application.
To prove that the simulation setup reproduces the results described in [CTTN07] I developed another SED-ML script `Calzone2007-simulation-figure-1B.xml` which generates Figure 1B of the publication. This script was developed using the *Edit Script* functionality of the SWT and is based on the initially generated script. It produces both graphs shown on the right-hand side of Figure 1B in `documentation/Calzone2007.pdf`. Finally, I uploaded it to the `experiment/` directory in the CombineArchiveWeb application.

**The simulation results** show the effect of running the model under the environment defined in the simulation description. To simulate the study I ran the experiment defined in `Calzone2007-simulation-figure-1B.xml` at the SWT and using the stand-alone software program COPASI [HSG+06]. The plots generated by both tools prove that the developed *in silico* experiment reproduce the figure shown in the publication, cmp. Figure 2. The SWT, see Figure 2(b), as well as COPASI, see Figure 2(c), were able to reproduce Figure 1B of the publication, see Figure 2(a).
The figures produced by the SWT and COPASI were uploaded to the `result/` directory in the CombineArchiveWeb application and meta data were added accordingly. Dagmar: Which versions of the software?

**The visualisation** displays the biological system. There was a figure available from the CellML model repository, but it can also be encoded in a standard format. I used SBGN-ED [CKS10] to load the SBML model, which immediately presented a layout of the

[10] www.ebi.ac.uk/biomodels-main/BIOMD0000000144
[11] bqfbergmann.dyndns.org/SED-ML_Web_Tools




(a) Figure 1B of the Publication (b) SED-ML Web Tools (c) COPASI

**Figure 2: Comparison of simulation results.** The figure shows the simulation results included in the publication (2(a)) with those results generated by the SWT (2(b)) and COPASI (2(c)) using the SED-ML script `Calzone2007-simulation-figure-1B.xml`. It confirms that the results described in the publication could be reproduced.

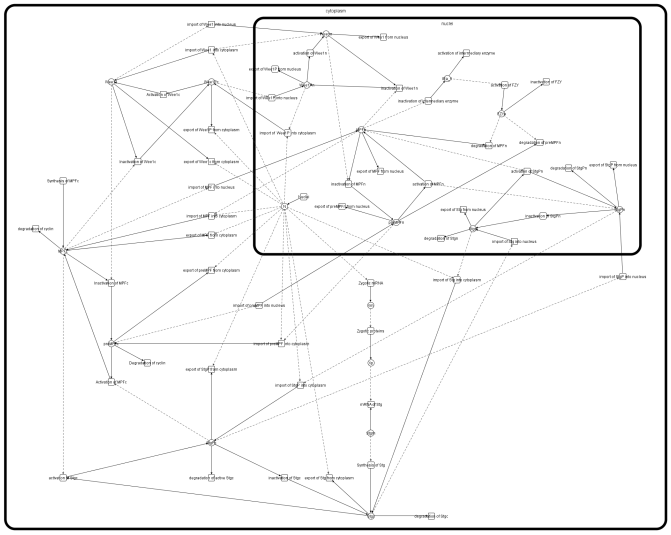


**Figure 3: SBGN-compliant figure**. This figure shows the reaction network encoded in the model document.

reaction network encoded in the model. After applying built-in graph layout algorithms from VANTED [JKS06] and a bit of manual layouting I obtained the network, as shown in Figure 3. Due to an error in SBGN-ED, I was not able to export the figure in SBGN-ML layout, but uploaded it as GraphML [BEH+02], GML[12], PNG image, and PDF to the `model/sbgn` directory in the CombineArchiveWeb application.

# 3. Results

## 3.1. The COMBINE archive

During my work I kept the COMBINE archive in a GIT repository for version control. The latest version of the archive can be found at GitHub[13], the latest archive can be downloaded from our website[14]. At the time of writing this report[15], the archive consists of 25 files organized in four directories (cmp. Section 2.3), including the `manifest.xml` and `metadata.rdf` (the skeleton of the archive, see Section 1) and a `README.md` file for the GitHub repository. The manifest listing all ingredients of the COMBINE archive is attached in Section A.1.
The archive basically consists of 4 modules: (i) the publication stored in the `documentation/` directory, (ii) the model of the biological system encoded in standardised formats stored in the `model/` directory, (iii) the simulation description encoded in SED-ML format and stored in the `experiment/` directory, (iv) the simulation results in form of graphs stored in the `result/` directory. The files in these directories were either retrieved from other websites (publication, SBML model, CellML model) or generated especially for this archive (SED-ML scripts, simulation results, SBGN map). The goal of this archive is to encode a reproducible simulation study. Figure 2 has shown that the developed study is able to reproduce graphs shown in the corresponding publication.

## 3.2. Reproducible Simulation Study

The developed study can easily be reproduced using the SWT, which provides a neat API. As the archive can be obtained from scripts.sems.uni-rostock.de/getshowcase.php you can import the archive by passing the link to the archive as the `url` parameter to the SWT. In other words, open the following URL in your browser:

`bqfbergmann.dyndns.org/SED-ML_Web_Tools/Home/SimulateUrl?url=http://scripts.sems.uni-rostock.de/getshowcase.php`

The SWT will immediately present you the results as shown in Figure 4. Dagmar: Does not sound too magic right now - maybe you can describe what is happening behind the scences - and what is the advantage over state-of-the-art handling of CAs? The SWT API also allows you to run the simulation encoded in the demo archive without visiting a website. You can simply send the archive using an HTTP POST request and obtain the simulation results, which are again encoded in a COMBINE archive:

[12] www.fim.uni-passau.de/index.php?id=17297&L=1
[13] github.com/SemsProject/CombineArchiveShowCase
[14] scripts.sems.uni-rostock.de/getshowcase.php
[15] latest git commit: 2e946ce1adfd05d16350c30176e29546301603a2 – 2015-06-11

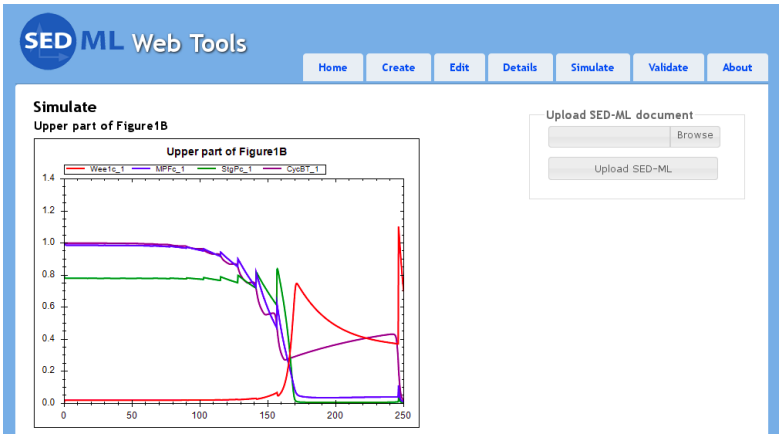



**Figure 4: Reproduced Results.** The results were reproduced by uploading the COMBINE archive to the SED-ML Web Tools.

**Obtain the simulation results using the API of the SED-ML Web Tolls**

```
wget -O demoarchive.omex http://scripts.sems.uni-rostock.de/getshowcase.php
curl -L -F file=@demoarchive.omex
  ↪  http://sysbioapps.dyndns.org/SED-ML_Web_Tools/Home/SimulatePostArchive >
  ↪  simulation-results.omex
```

You will then find the simulation results in the file `simulation-results.omex`. To explore the obtained archive it can, for example, be uploaded to the CombineArchiveWeb application.

# 4. Discussion

The complexity of research results poses a challenge to successful transfer – in a way that others are able to reproduce. More specifically, the following problems occur when sharing research results. Dagmar: I suggest to put problems into paragraphs

**All relevant files need to be shipped.** Simulation studies usually consist of multiple relevant files, which are all necessary to run the experiment. Thus, to reproduce a study it is essential to have access to all these files.

**Users need to know the type and meaning of these files.** The files should be encoded in standardized formats, so that users get a chance (i) to grasp what is encoded inside, (ii) to understand how to use them, and (iii) to execute them in a software tool of their choice. Especially, if the tool used to create the files is not available (anymore).

**There must be a manual on how to use the files.** For example, the model might rely on modules and the simulation description might expect this model in at a certain location, cmp. A.3. Thus, if the files are not arranged properly, the study cannot be rerun.

**It must be clear whom to consult in case of problems.** Users need to know who is/was responsible for a certain part of the simulation study. Creators and contributors should be attributed appropriately.

As these requirements are crucial yet error-prone the COMBINE archive approach gives researches a hand.
A COMBINE archive is basically a container that aggregates all necessary files for a simulation study. The archive maintains a manifest listing its contents (files and their formats) and a some meta data describing the files and the archive itself. Tools supporting COMBINE archives are able to understand its ingredients and can, for example, run the encoded experiment just as its creator intended run it. Thus, users do not need to understand each of these files in detail. They do not even need to know that the model is decomposed, or that it is encoded in, e.g., the SBML format. Users just need the archive which provides all necessary information to rerun the experiment.
However, tool support for COMBINE archives is still very limited. The experiment encoded in the developed archive is able to reproduce the figure published in the corresponding publication, see Figure 2. However, the SWT are, as far as I know, the only tool able to consume the developed archive and run the experiment. That's reasonable, as the standard of COMBINE archives is still very young. I hope that the archive developed in this study helps tool developers to implement support for COMBINE archives.
During the creation of the demo archive I also spotted some shortcomings of the corresponding standard. I was missing means to encode the provenance of the files in a machine readable format. To document, for example, that the file `/model/BIOMD0000000144.xml` was retrieved from the BioModels Database I had to add a Dublin Core [WKLW98] description:

*"This model was downloaded from the BioModels Database Jun 11th, 2015."*

The information in this text is very important yet not understandable for machines. A standardized way for encoding such provenance information, for example using the PROV Ontology[16], would be a significant enhancement. In addition, dependency links could be included to the meta data to, for example, link the simulation description and the model, or to encode which simulation description produced a certain figure. Thus, there is still space for improvements.
The COMBINE archive approach, however, is already a solid format for sharing reproducible models and *in-silico* experiments with collaborators and public databases. In the future I expect to see more tools able to read and export COMBINE archives. It would support researchers in sharing and distributing their hard-earned scientific results.

[16] www.w3.org/TR/prov-o/

# 5. Acknowledgements

# References


[BAM+14] F. T. Bergmann, R. Adams, S. Moodie, J. Cooper, M. Glont, M. Golebiewski, M. Hucka, C. Laibe, A. K. Miller, D. P. Nickerson, B. G. Olivier, N. Rodriguez, H. M. Sauro, M. Scharm, S. Soiland-Reyes, D. Waltemath, F. Yvon, and N. Le Novere. COMBINE archive and OMEX format: one file to share all information to reproduce a modeling project. *BMC Bioinformatics*, 15(1):369, 2014.

[BE12] C. G. Begley and L. M. Ellis. Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533, Mar 2012.

[BEH+02] Ulrik Brandes, Markus Eiglsperger, Ivan Herman, Michael Himsolt, and M Scott Marshall. Graphml progress report structural layer proposal. In *Graph Drawing*, pages 501–512. Springer, 2002.

[BRG+10] Sean Bechhofer, David De Roure, Matthew Gamble, Carole Goble, and Iain Buchan. Research objects: Towards exchange and reuse of digital knowledge. *The Future of the Web for Collaborative Science*, 2010.

[CKS10] Tobias Czauderna, Christian Klukas, and Falk Schreiber. Editing, validating and translating of sbgn maps. *Bioinformatics*, 26(18):2340–2341, 2010.

[CLN+03] Autumn A. Cuellar, Catherine M. Lloyd, Poul F. Nielsen, David P. Bullivant, David P. Nickerson, and Peter J. Hunter. An overview of CellML 1.1, a biological model description language. *SIMULATION*, 79(12):740–747, 2003.

[CTTN07] Laurence Calzone, Denis Thieffry, John J Tyson, and Bela Novak. Dynamical modeling of syncytial mitotic cycles in *Drosophila* embryos. *Molecular systems biology*, 3:131, 2007.

[CVB+12] Óscar Corcho, Daniel Garijo Verdejo, K Belhajjame, Jun Zhao, Paolo Missier, David Newman, Raúl Palma, Sean Bechhofer, Esteban García Cuesta, José Manuel Gómez-Pérez, et al. Workflow-centric research objects: First class citizens in scholarly discourse. 2012.

[DSPM10] Joseph O. Dada, Irena Spasić, Norman W. Paton, and Pedro Mendes. SBRML: a markup language for associating systems biology data with models. *Bioinformatics*, 26(7):932–938, 2010.

[HFS+03] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, , the rest of the SBML Forum:, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J.-H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Novère, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson,

P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang. The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003.

[HSG+06] Stefan Hoops, Sven Sahle, Ralph Gauges, Christine Lee, Jürgen Pahle, Natalia Simus, Mudita Singhal, Liang Xu, Pedro Mendes, and Ursula Kummer. COPASI—a complex pathway simulator. *Bioinformatics*, 22(24):3067–3074, 2006.

[HWW15] Ron Henkel, Olaf Wolkenhauer, and Dagmar Waltemath. Combining computational models, semantic annotations and simulation experiments in a graph database. *Database*, 2015, 2015.

[IAB+09] J. P. Ioannidis, D. B. Allison, C. A. Ball, I. Coulibaly, X. Cui, A. C. Culhane, M. Falchi, C. Furlanello, L. Game, G. Jurman, J. Mangion, T. Mehta, M. Nitzberg, G. P. Page, E. Petretto, and V. van Noort. Repeatability of published microarray gene expression analyses. *Nat. Genet.*, 41(2):149–155, Feb 2009.

[JKS06] Bjorn Junker, Christian Klukas, and Falk Schreiber. Vanted: A system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics*, 7(1):109, 2006.

[LDR+10] C. Li, M. Donizelli, N. Rodriguez, H. Dharuri, L. Endler, V. Chelliah, L. Li, E. He, A. Henry, M. I. Stefan, J. L. Snoep, M. Hucka, N. Le Novere, and C. Laibe. BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Syst Biol*, 4:92, 2010.

[LL+08] C.M. Lloyd, J.R. Lawson, et al. The CellML model repository. *Bioinformatics*, 24(18):2122–2123, 2008.

[LNHM+09] N. Le Novere, M. Hucka, H. Mi, S. Moodie, F. Schreiber, A. Sorokin, E. Demir, K. Wegner, M. I. Aladjem, S. M. Wimalaratne, F. T. Bergman, R. Gauges, P. Ghazal, H. Kawaji, L. Li, Y. Matsuoka, A. Villeger, S. E. Boyd, L. Calzone, M. Courtot, U. Dogrusoz, T. C. Freeman, A. Funahashi, S. Ghosh, A. Jouraku, S. Kim, F. Kolpakov, A. Luna, S. Sahle, E. Schmidt, S. Watterson, G. Wu, I. Goryanin, D. B. Kell, C. Sander, H. Sauro, J. L. Snoep, K. Kohn, and H. Kitano. The Systems Biology Graphical Notation. *Nat. Biotechnol.*, 27(8):735–741, Aug 2009.

[PSA11] F. Prinz, T. Schlange, and K. Asadullah. Believe it or not: how much can we rely on published data on potential drug targets? *Nat Rev Drug Discov*, 10(9):712, Sep 2011.

[SNTH13] Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, and Eivind Hovig. Ten simple rules for reproducible computational research. *PLoS Comput Biol*, 9(10):e1003285, 10 2013.

[SW15] Martin Scharm and Dagmar Waltemath. Extracting reproducible simulation studies from model repositories using the CombineArchive Toolkit. *PeerJ PrePrints*, 3:e792v1, 2015.

[SWP+14a] Martin Scharm, Florian Wendland, Martin Peters, Markus Wolfien, Tom Theile, and Dagmar Waltemath. The CombineArchive Toolkit – facilitating the transfer of research results. *PeerJ PrePrints*, 2:e514v1, 2014.

[SWP+14b] Martin Scharm, Florian Wendland, Martin Peters, Markus Wolfien, Tom Theile, and Dagmar Waltemath. The CombineArchiveWeb application – A web-based tool to handle files associated with modelling results. In *Proceedings of the 7th International Workshop on Semantic Web Applications and Tools for Life Sciences, Berlin, Germany, December 9-11, 2014.*, 2014.

[WAB+11] Dagmar Waltemath, Richard Adams, Frank Bergmann, Michael Hucka, Fedor Kolpakov, Andrew Miller, Ion Moraru, David Nickerson, Sven Sahle, Jacky Snoep, and Nicolas Le Novere. Reproducible computational biology experiments with SED-ML - the simulation experiment description markup language. *BMC Systems Biology*, 5(1):198, 2011.

[WKLW98] Stuart Weibel, John Kunze, Carl Lagoze, and Misha Wolf. Dublin core metadata for resource discovery. Technical report, 1998.

# A. Appendix

## A.1. The Mainfest

**The `manifest.xml` of the final COMBINE archive**

```
<?xml version="1.0" encoding="UTF-8"?>
<omexManifest xmlns="http://identifiers.org/combine.specifications/omex-manifest">
  <content location="." format="http://identifiers.org/combine.specifications/omex" />
  <content location="./manifest.xml" format="http://identifiers.org/combine.specifications/omex-manifest" />
  <content location="./README.md" format="http://purl.org/NET/mediatypes/text/x-markdown"/>
  <content location="./model/BIOMD0000000144.xml" format="http://identifiers.org/combine.specifications/sbml.level-2.version-1" />
  <content location="./model/calzone_2007.ai" format="http://purl.org/NET/mediatypes/application/illustrator" />
  <content location="./model/calzone_2007.png" format="http://purl.org/NET/mediatypes/image/png" />
  <content location="./model/calzone_2007.svg" format="http://purl.org/NET/mediatypes/image/svg+xml" />
  <content location="./model/calzone_thieffry_tyson_novak_2007.cellml" format="http://identifiers.org/combine.specifications/cellml" />
  <content location="./model/sbgn/Calzone2007.gml" format="http://purl.org/NET/mediatypes/text/plain" />
```

```
  <content location="./model/sbgn/Calzone2007.graphml"
   ↪  format="http://purl.org/NET/mediatypes/application/xml" />
  <content location="./model/sbgn/Calzone2007.png"
   ↪  format="http://purl.org/NET/mediatypes/image/png" />
  <content location="./model/sbgn/Calzone2007.pdf"
   ↪  format="http://purl.org/NET/mediatypes/application/pdf" />
  <content location="./experiment/Calzone2007-simulation-all-figures.xml"
   ↪  format="http://identifiers.org/combine.specifications/sed-ml.level-1.version-1"
   ↪  />
  <content location="./experiment/Calzone2007-simulation-figure-1B.xml"
   ↪  format="http://identifiers.org/combine.specifications/sed-ml.level-1.version-1"
   ↪  master="true"/>
  <content location="./documentation/Calzone2007.pdf"
   ↪  format="http://purl.org/NET/mediatypes/application/pdf" />
  <content location="./documentation/Calzone2007-supplementary-material.pdf"
   ↪  format="http://purl.org/NET/mediatypes/application/pdf" />
  <content location="./result/Fig1B-top-webtools.png"
   ↪  format="http://purl.org/NET/mediatypes/image/png" />
  <content location="./result/Fig1B-bottom-webtools.png"
   ↪  format="http://purl.org/NET/mediatypes/image/png" />
  <content location="./result/Fig1B-top-COPASI.svg"
   ↪  format="http://purl.org/NET/mediatypes/image/svg+xml" />
  <content location="./result/Fig1B-bottom-COPASI.svg"
   ↪  format="http://purl.org/NET/mediatypes/image/svg+xml" />
  <content location="./metadata.rdf"
   ↪  format="http://identifiers.org/combine.specifications/omex-metadata" />
</omexManifest>
```

## A.2. RDF/XML meta data snippet

**Meta Data of the Publication** `Calzone2007.pdf`

```
<rdf:Description rdf:about="/documentation/Calzone2007.pdf">
  <dcterms:description>Article published in Mol Syst Biol. 2007; 3: 131. DOI:
   ↪  10.1038/msb4100171, downloaded June 11th, 2015 . It describes the models
   ↪  /model/calzone_thieffry_tyson_novak_2007.cellml and
   ↪  /model/BIOMD0000000144.xml .</dcterms:description>
  <dcterms:creator>
    <rdf:Bag>
      <rdf:li rdf:parseType="Resource">
        <vCard:n rdf:parseType="Resource">
          <vCard:family-name>Calzone</vCard:family-name>
          <vCard:given-name>Laurence</vCard:given-name>
        </vCard:n>
        <vCard:org rdf:parseType="Resource">
          <vCard:organization-name>Molecular Network Dynamics Research Group of
           ↪  Hungarian Academy of Sciences and Budapest University of Technology
           ↪  and Economics, Budapest</vCard:organization-name>
        </vCard:org>
      </rdf:li>
    </rdf:Bag>
  </dcterms:creator>
  <dcterms:creator>
```

```
    <rdf:Bag>
      <rdf:li rdf:parseType="Resource">
        <vCard:n rdf:parseType="Resource">
          <vCard:family-name>Thieffry</vCard:family-name>
          <vCard:given-name>Denis</vCard:given-name>
        </vCard:n>
        <vCard:org rdf:parseType="Resource">
          <vCard:organization-name>Université de la Méditerranée, Campus
          ↪  Scientifique de Luminy, Case 928, Marseille,
          ↪  France</vCard:organization-name>
        </vCard:org>
      </rdf:li>
    </rdf:Bag>
  </dcterms:creator>
  <dcterms:creator>
    <rdf:Bag>
      <rdf:li rdf:parseType="Resource">
        <vCard:n rdf:parseType="Resource">
          <vCard:family-name>Tyson</vCard:family-name>
          <vCard:given-name>John J</vCard:given-name>
        </vCard:n>
        <vCard:org rdf:parseType="Resource">
          <vCard:organization-name>Department of Biological Sciences, Virginia
          ↪  Polytechnic Institute and State University, Blacksburg, VA,
          ↪  USA</vCard:organization-name>
        </vCard:org>
      </rdf:li>
    </rdf:Bag>
  </dcterms:creator>
  <dcterms:creator>
    <rdf:Bag>
      <rdf:li rdf:parseType="Resource">
        <vCard:n rdf:parseType="Resource">
          <vCard:family-name>Novak</vCard:family-name>
          <vCard:given-name>Bela</vCard:given-name>
        </vCard:n>
        <vCard:email>bela.novak@bioch.ox.ac.uk</vCard:email>
        <vCard:org rdf:parseType="Resource">
          <vCard:organization-name>Molecular Network Dynamics Research Group of
          ↪  Hungarian Academy of Sciences and Budapest University of Technology
          ↪  and Economics, Budapest</vCard:organization-name>
        </vCard:org>
      </rdf:li>
    </rdf:Bag>
  </dcterms:creator>
  <dcterms:created rdf:parseType="Resource">
    <dcterms:W3CDTF>2007-01-08T00:00:00Z</dcterms:W3CDTF>
  </dcterms:created>
  <dcterms:modified rdf:parseType="Resource">
    <dcterms:W3CDTF>2007-01-08T00:00:00Z</dcterms:W3CDTF>
  </dcterms:modified>
  <dcterms:modified rdf:parseType="Resource">
    <dcterms:W3CDTF>2007-06-22T00:00:00Z</dcterms:W3CDTF>
  </dcterms:modified>
  <dcterms:modified rdf:parseType="Resource">
```

```
66      <dcterms:W3CDTF>2007-07-24T08:58:57Z</dcterms:W3CDTF>
67    </dcterms:modified>
68    <dcterms:modified rdf:parseType="Resource">
69      <dcterms:W3CDTF>2014-03-26T11:22:15Z</dcterms:W3CDTF>
70    </dcterms:modified>
71  </rdf:Description>
```

## A.3. Snippet of a Simulation Description

**Snippet of** `experiment/Calzone2007-simulation-figure-1B.xml` **referring to the model** `/model/BIOMD0000000144.xml`

```
1  <?xml version="1.0" encoding="utf-8"?>
2  <sedML level="1" version="1" xmlns="http://sed-ml.org/">
3    [...]
4    <listOfModels>
5      <model id="model1" language="urn:sedml:language:sbml"
         ↪  source="../model/BIOMD0000000144.xml" />
6    </listOfModels>
7    [...]
8  </sedML>
```