

Excel-SBOL Converter: Creating SBOL from Excel Templates and Vice Versa

Jeanet Mante

jet@mante.net

University of Colorado Boulder

Isabel Pöttsch

imp31@cam.ac.uk

University of Cambridge

Julian Abam

Julian.Abam@colorado.edu

University of Colorado Boulder

Jacob Beal

jakebeal@ieee.org

Raytheon BBN Technologies

Chris J. Myers

chris.myers@colorado.edu

University of Colorado Boulder

1 INTRODUCTION

Synthetic biology is bringing together engineers and biologists [10]. Associated with this interdisciplinary movement is the need for reusable tools that supplement the current understanding of genetic sequences. To satisfy this need, Synthetic Biology communities across the world have developed tools and ontologies to help describe their unique semantic annotations [1, 3–9, 13, 14, 17–19, 22]. Shared representations for data and metadata, grounded in well-defined ontology terms, can help reduce confusion when sharing materials between practitioners.[20]. The Synthetic Biology Open Language (SBOL) [5] is one of the approaches that has been developed to address this challenge. SBOL provides a standardized format for the electronic exchange of information on the structural and functional aspect of biological designs, supporting use of engineering principles of abstraction, modularity, and standardization in synthetic biology. Many tools have been created that work with SBOL, including the SynBioHub repository software for storing and sharing designs [12].

Using formal representations such as SBOL, however, typically requires either a thorough understanding of these standards or a suite of tools developed in concurrence with the ontologies [11]. Unfortunately, this poses a significant barrier to use for scientists not trained to work with such abstractions. One approach to lowering this barrier was demonstrated in the Systems Biology for Micro-Organisms (SysMO) consortium [2]. In SysMO, the MicroArray Gene Expression Markup Language (Mage-ML) was set up as an XML schema [15], and users were expected to submit data to the SysMO Assets Catalogue (called SEEK) in XML format in order to publish work. To allow the use of the Mage-ML language without having to understand XML, the RightField tool was created [21], an ontology annotation and information management application that can add constrained ontology term selection to Excel spreadsheets. This tool enables administrators to create templates with controlled vocabularies, such that the scientists utilizing the tool would never actually see

the raw RightField, only the more familiar Excel spreadsheet interface.

Similarly, users of SBOL and SynBioHub have faced a steep learning curve for understanding the underlying ontology: as assessed in [16], “For successful use and interpretation of metadata presented in SynBioHub, the semantic annotation process should be biologist-friendly and hide the underlying RDF predicates.” Accordingly, the Excel-SBOL Converter presented here has been designed to provide a simple way for users to generate SBOL data without needing a detailed understanding of the underlying ontology and associated technologies. The converter provides a simple way for users to manage data by allowing users to download SBOL into Excel templates and submit Excel templates for conversion into SBOL.

2 RESULTS

The Excel-SBOL Converter enables the round trip conversion of SBOL2 into Excel Templates and Excel Templates to SBOL2, as shown in Figure 1. This metadata conversion allows researchers the flexibility of working in Excel without losing the benefits of SBOL2 and maintains inter-operability with the SBOL2 tool suite. The converter also facilitates understanding of the metadata relationships, management, and structure within the SBOL2 XML file. The converter has been implemented and published as two Python packages, Excel2SBOL and SBOL2Excel, under a free and open license and published to pypi for easy pip installation and integration with other projects. Specific benefits of each of the packages are described below.

SBOL2Excel

The SBOL2Excel package is designed in a modular way to allow further expansion. The package’s design includes an initial step from SBOL2 to a dictionary structure. This function could be swapped out with other functions to enable the conversion of other formats such as SBOL3 and GenBank to the Excel Templates.

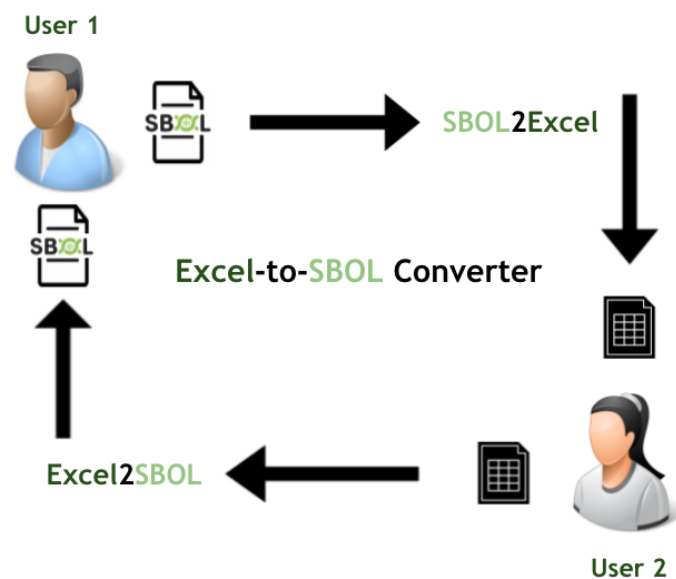


Figure 1: Round trip journey of the Excel-to-SBOL Converter: User 1 converts an SBOL file into an Excel spreadsheet using the SBOL2Excel library, and shares that spreadsheet with User 2, who then converts the Excel spreadsheet back into an SBOL file using the Excel2SBOL library. The SBOL file that User 2 returns to User 1 is exactly the same as the SBOL file that User 1 originally had.

Another key feature is that any arbitrary Component-Definition properties can be converted into Excel columns. However, some common types of properties are given additional post processing to make them more user friendly. SBOL2Excel operates in five general steps:

- (1) An SBOL2 document is read (with property-specific post-processing) and converted to a dictionary.
- (2) The dictionary is converted to a pandas dataframe.
- (3) The dataframe’s columns are reordered based on the default Excel Template.
- (4) Extraneous columns are dropped from the dataframe.
- (5) The dataframe is output into a formatted Excel Spreadsheet.

Figure 2 shows a high level representation of SBOL2Excel’s modular architecture.

Excel2SBOL

This package is able to parse and add information from arbitrary columns into component definitions as annotations. Specific columns are added as a particular type of annotation based on a column conversion table that is accessible in the Excel Template. The package is also able to deal with different templates as long as the Template parameters are

added to the modules config file. SBOL2Excel operates in 3 general steps:

- (1) Parse the different sections of the template (Overview Information, Design Description, and Part Table).
- (2) Parse the table indicating how columns are converted.
- (3) For every row, create a component definition and add attributes based on each column in the parts table.

Figure 3 is a high level representation of Excel2SBOL’s modular architecture.

3 DISCUSSION

The next steps in the development of the Excel-SBOL Converter are:

- Making the incorporation of ontologies with Excel Templates easier and less hard-coded, perhaps by relying on RightField Templates. Integrating RightField may be useful in preparing ontologies for templating.
- Making the benefits of adherence to community standards clearer to users to create greater uptake.
- Establishing a core set of part metadata which can then be enforced via the Excel Templates.
- Making both packages SBOL3 Compliant. Enabling this feature would allow users to perform conversions from SBOL documents (perhaps generated by SynBioHub) that contain SBOL2, SBOL3, or both.
- Creating SynBioHub plugins for both packages to integrate their functionality with SynBioHub. This would enable users to perform conversions without directly interacting with Python code.
- Expanding the range of top-level SBOL types that can be converted (e.g. Activities, Composite Components, and Modules).

ACKNOWLEDGEMENTS

JA, JM, and CM are supported by the National Science Foundation under Grant No. 1939892. JM is additionally supported by a Dean’s Graduate Assistantship at the University of Colorado Boulder. IP was supported by the Google Summer of Code. JB is partially supported by AFRL and DARPA contract FA8750-17-C-0184. This document does not contain technology or technical data controlled under either U.S. International Traffic in Arms Regulation or U.S. Export Administration Regulations. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies.

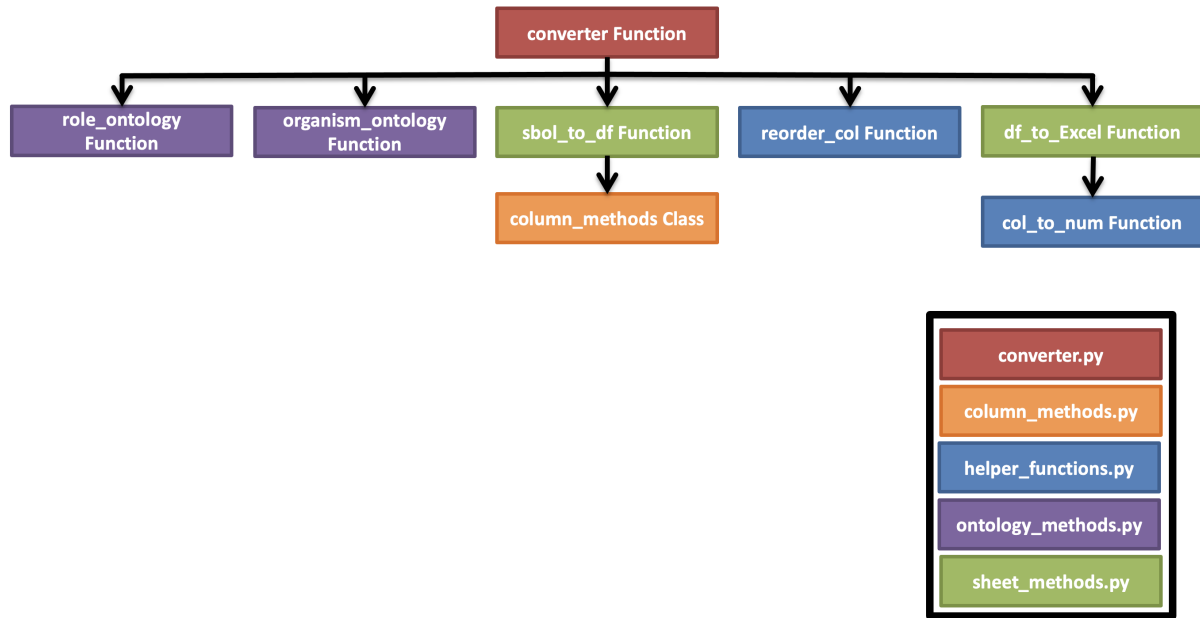


Figure 2: The module architecture of the SBOL2Excel Python Package. Note that the colors of the functions indicate which Python module they can be found in, as indicated in the key on the bottom right.

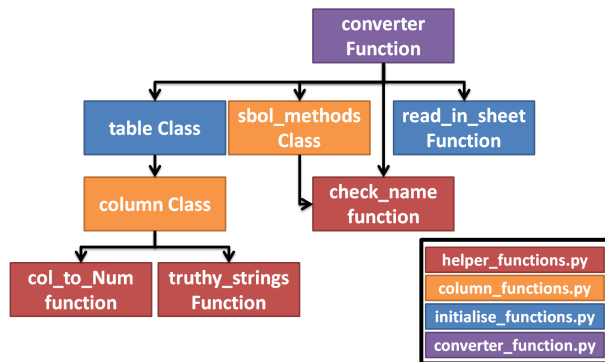


Figure 3: The module architecture of the Excel2SBOL Python Package. Note that the colors of the functions indicate which Python module they can be found in, as indicated in the key on the bottom right.

REFERENCES

[1] BERGMANN, F. T., ADAMS, R., MOODIE, S., COOPER, J., GLONT, M., GOLEBIEWSKI, M., HUCKA, M., LAIBE, C., MILLER, A. K., NICKERSON, D. P., AND ET AL. Combine archive and omex format: one file to share all information to reproduce a modeling project. *BMC bioinformatics* 15 (Dec 2014), 369.

[2] BOOTH, I. R. Sysmo: back to the future. *Nature Reviews Microbiology* 5, 8 (Aug 2007), 566–566.

[3] CANNON, R. C., GLEESON, P., CROOK, S., GANAPATHY, G., MARIN, B., PIASINI, E., AND SILVER, R. A. Lems: a language for expressing complex biological models in concise and hierarchical form and its use in

underpinning neuroml 2. *Frontiers in Neuroinformatics* 8 (2014), 79.

[4] FAEDER, J. R., BLINOV, M. L., AND HLAVACEK, W. S. *Rule-Based Modeling of Biochemical Systems with BioNetGen*, vol. 500 of *Methods in Molecular Biology*. Humana Press, 2009, p. 113–167.

[5] GALDZICKI, M., CLANCY, K. P., OBERORTNER, E., POCOOCK, M., QUINN, J. Y., RODRIGUEZ, C. A., ROEHNER, N., WILSON, M. L., ADAM, L., ANDERSON, J. C., AND ET AL. The synthetic biology open language (sbol) provides a community standard for communicating designs in synthetic biology. *Nature Biotechnology* 32, 6 (Jun 2014), 545–550.

[6] GENNARI, J. H., NEAL, M. L., GALDZICKI, M., AND COOK, D. L. Multiple ontologies in action: composite annotations for biosimulation models. *Journal of Biomedical Informatics* 44, 1 (Feb 2011), 146–154.

[7] GLEESON, P., CROOK, S., CANNON, R. C., HINES, M. L., BILLINGS, G. O., FARINELLA, M., MORSE, T. M., DAVISON, A. P., RAY, S., BHALLA, U. S., AND ET AL. Neuroml: A language for describing data driven models of neurons and networks with a high degree of biological detail. *PLoS Computational Biology* 6, 6 (Jun 2010), e1000815.

[8] HUCKA, M., FINNEY, A., SAURO, H. M., BOLOURI, H., DOYLE, J. C., KITANO, H., ARKIN, A. P., BORNSTEIN, B. J., BRAY, D., CORNISH-BOWDEN, A., AND ET AL. The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics (Oxford, England)* 19, 4 (Mar 2003), 524–531.

[9] HUCKA, M., NICKERSON, D. P., BADER, G. D., BERGMANN, F. T., COOPER, J., DEMIR, E., GARNY, A., GOLEBIEWSKI, M., MYERS, C. J., SCHREIBER, F., AND ET AL. Promoting coordinated development of community-based information standards for modeling in biology: The combine initiative. *Frontiers in Bioengineering and Biotechnology* 3 (2015), 19.

[10] KHALIL, A. S., AND COLLINS, J. J. Synthetic biology: applications come of age. *Nature Reviews Genetics* 11, 5 (2010), 367–379.

[11] MACCAGNAN, A., RIVA, M., FELTRIN, E., SIMIONATI, B., VARDANEGA, T., VALLE, G., AND CANNATA, N. Combining ontologies and workflows

319 to design formal protocols for biological laboratories. *Automated*
 320 *Experimentation* 2, 1 (Apr 2010), 3.

321 [12] McLAUGHLIN, J. A., MYERS, C. J., ZUNDEL, Z., MISIRLI, G., ZHANG,
 322 M., OFITERU, I. D., GOÑI-MORENO, A., AND WIPAT, A. Synbiohub:
 323 A standards-enabled design repository for synthetic biology. *ACS*
 324 *Synthetic Biology* 7, 2 (Feb 2018), 682–688.

325 [13] NICKERSON, D., ATALAG, K., DE BONO, B., GEIGER, J., GOBLE, C., HOLL-
 326 MANN, S., LONIEN, J., MÜLLER, W., REGIERER, B., STANFORD, N. J., AND
 327 ET AL. The human physiome: how standards, software and innovative
 328 service infrastructures are providing the building blocks to make it
 329 achievable. *Interface Focus* 6, 2 (Apr 2016), 20150103.

330 [14] NOVÈRE, N. L., HUCKA, M., MI, H., MOODIE, S., SCHREIBER, F., SOROKIN,
 331 A., DEMIR, E., WEGNER, K., ALADJEM, M. I., WIMALARATNE, S. M., AND
 332 ET AL. The systems biology graphical notation. *Nature Biotechnology*
 333 27, 8 (Aug 2009), 735–741.

334 [15] SPELLMAN, P. T., MILLER, M., STEWART, J., TROUP, C., SARKANS, U.,
 335 CHERVITZ, S., BERNHART, D., SHERLOCK, G., BALL, C., LEPAGE, M.,
 336 AND ET AL. Design and implementation of microarray gene expres-
 337 sion markup language (mage-ml). *Genome Biology* 3, 9 (Aug 2002),
 338 research0046.1.

339 [16] URQUIZA-GARCÍA, U., ZIELIŃSKI, T., AND MILLAR, A. J. Better research
 340 by efficient sharing: evaluation of free management platforms for
 341 synthetic biology designs. *Synthetic Biology* 4, 1 (2019), ysz016.

342 [17] WALTEMATH, D., ADAMS, R., BERGMANN, F. T., HUCKA, M., KOLPAKOV,
 343 F., MILLER, A. K., MORARU, I. I., NICKERSON, D., SAHLE, S., SNOEP, J. L.,
 344 AND ET AL. Reproducible computational biology experiments with
 345 sed-ml—the simulation experiment description markup language. *BMC*
 346 *systems biology* 5 (Dec 2011), 198.

347 [18] WITTIG, U., KANIA, R., GOLEBIEWSKI, M., REY, M., SHI, L., JONG, L.,
 348 ALGAA, E., WEIDEMANN, A., SAUER-DANZWITZ, H., MIR, S., AND ET AL.
 349 Sabio-rk—database for biochemical reaction kinetics. *Nucleic Acids*
 350 *Research* 40, Database issue (Jan 2012), D790–796.

351 [19] WOLSTENCROFT, K., KREBS, O., SNOEP, J. L., STANFORD, N. J., BACALL,
 352 F., GOLEBIEWSKI, M., KUZYAKIV, R., NGUYEN, Q., OWEN, S., SOILAND-
 353 REYES, S., AND ET AL. Fairdomhub: a repository and collaboration
 354 environment for sharing systems biology research. *Nucleic Acids*
 355 *Research* 45, D1 (Jan 2017), D404–D407.

356 [20] WOLSTENCROFT, K., OWEN, S., HORRIDGE, M., JUPP, S., KREBS, O., SNOEP,
 357 J., DU PREEZ, F., MUELLER, W., STEVENS, R., AND GOBLE, C. Stealthy
 358 annotation of experimental biology by spreadsheets. *Concurrency and*
 359 *Computation: Practice and Experience* 25, 4 (2013), 467–480.

360 [21] WOLSTENCROFT, K., OWEN, S., HORRIDGE, M., KREBS, O., MUELLER,
 361 W., SNOEP, J. L., DU PREEZ, F., AND GOBLE, C. Rightfield: embedding
 362 ontology annotation in spreadsheets. *Bioinformatics* 27, 14 (2011),
 363 2021–2022.

364 [22] YU, T., LLOYD, C. M., NICKERSON, D. P., COOLING, M. T., MILLER, A. K.,
 365 GARNY, A., TERKILDSEN, J. R., LAWSON, J., BRITTON, R. D., HUNTER,
 366 P. J., AND ET AL. The physiome model repository 2. *Bioinformatics*
 367 (Oxford, England) 27, 5 (Mar 2011), 743–744.

368 372
 369 373
 370 374
 371 375
 372 376
 373 377
 374 378
 375 379
 376 380
 377 381
 378 382
 379 383
 380 384
 381 385
 382 386
 383 387
 384 388
 385 389
 386 390
 387 391
 388 392
 389 393
 390 394
 391 395
 392 396
 393 397
 394 398
 395 399
 396 400
 397 401
 398 402
 399 403
 400 404
 401 405
 402 406
 403 407
 404 408
 405 409
 406 410
 407 411
 408 412
 409 413
 410 414
 411 415
 412 416
 413 417
 414 418
 415 419
 416 420
 417 421
 418 422
 419 423
 420 424