ACS SyntheticBiology

pubs.acs.org/synthbio Technical Note

Experimental Data Connector (XDC): Integrating the Capture of Experimental Data and Metadata Using Standard Formats and Digital Repositories

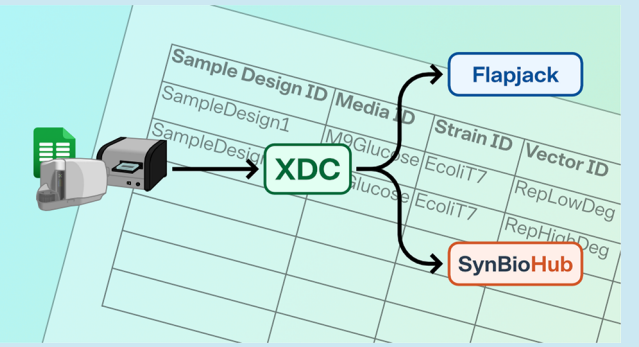
Published as part of the ACS Synthetic Biology virtual special issue "IWBDA 2022".

Sai P. Samineni,[‡] Gonzalo Vidal,[‡] Carolus Vitalis, Guillermo Yáñez Feliú, Timothy J. Rudge, Chris J. Myers, and Jeanet Mante*

Cite This: *ACS Synth. Biol.* 2023, 12, 1364–1370 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Accelerating the development of synthetic biology applications requires reproducible experimental findings. Different standards and repositories exist to exchange experimental data and metadata. However, the associated software tools often do not support a uniform data capture, encoding, and exchange of information. A connection between digital repositories is required to prevent siloing and loss of information. To this end, we developed the Experimental Data Connector (XDC). It captures experimental data and related metadata by encoding it in standard formats and storing the converted data in digital repositories. Experimental data is then uploaded to Flapjack and the metadata to SynBioHub in a consistent manner linking these repositories. This produces complete connected experimental data sets that are exchangeable. The information is captured using a single template Excel Workbook, which can be integrated into existing experimental workflow automation processes and semiautomated capture of results.



KEYWORDS: *experimental data, metadata, Excel, SBOL, SynBioHub, Flapjack*

INTRODUCTION

Synthetic biology (SynBio) aims to engineer biological systems in a predictable and reproducible way. To this end, software tools are needed to aid researchers in iterating the *design-build-test-learn* (DBTL) cycle. *Experimental workflow automation* (EWA) is a critical aspect of developing synthetic biology applications. In the field of SynBio, EWA promotes automation by integrating software tools to support assembly planning, protocol management, and digital data storage with well-defined sets of instructions, specific designs, materials, and machines. Through this set of software tools, EWA captures the underlying information needed to build and test genetic circuits in a reproducible manner. The use of software is essential to manage the development of complex SynBio applications that remain relevant across diverse operational contexts. Digital repositories are used to increase the reproducibility of DBTL iterations. There are several types of digital repositories, including ones focused on metadata and ones focused on experimental data. They overlap in some ways; however, they are generally not connected.

Digital repositories that support the build and test stages of EWA include SynBioHub[1] and Flapjack.[2] SynBioHub is a web application that provides a repository for DNA sequences, biological parts and devices, strains, experimental setup information, and other metadata (https://synbiohub.org). The information in SynBioHub is stored using the *Synthetic Biology Open Language* (SBOL).[3−5] SBOL is a free and open-source standard for the representation and electronic exchange of information on the structural and functional aspects of biological designs. Flapjack is a data management system that enables researchers to store, visualize, analyze, and share genetic circuit experimental data, including measurement data and corresponding metadata (http://flapjack.rudge-lab.org/). It stores its data using the Flapjack data model that has been optimized for experimental data. Flapjack does not use the SBOL data model as it does not natively support measurement data (though this is supported via extensions). Flapjack plans to move to future versions of the SBOL data model that will

Received: December 14, 2022
Published: March 30, 2023

https://doi.org/10.1021/acssynbio.2c00669
ACS Synth. Biol. 2023, 12, 1364−1370

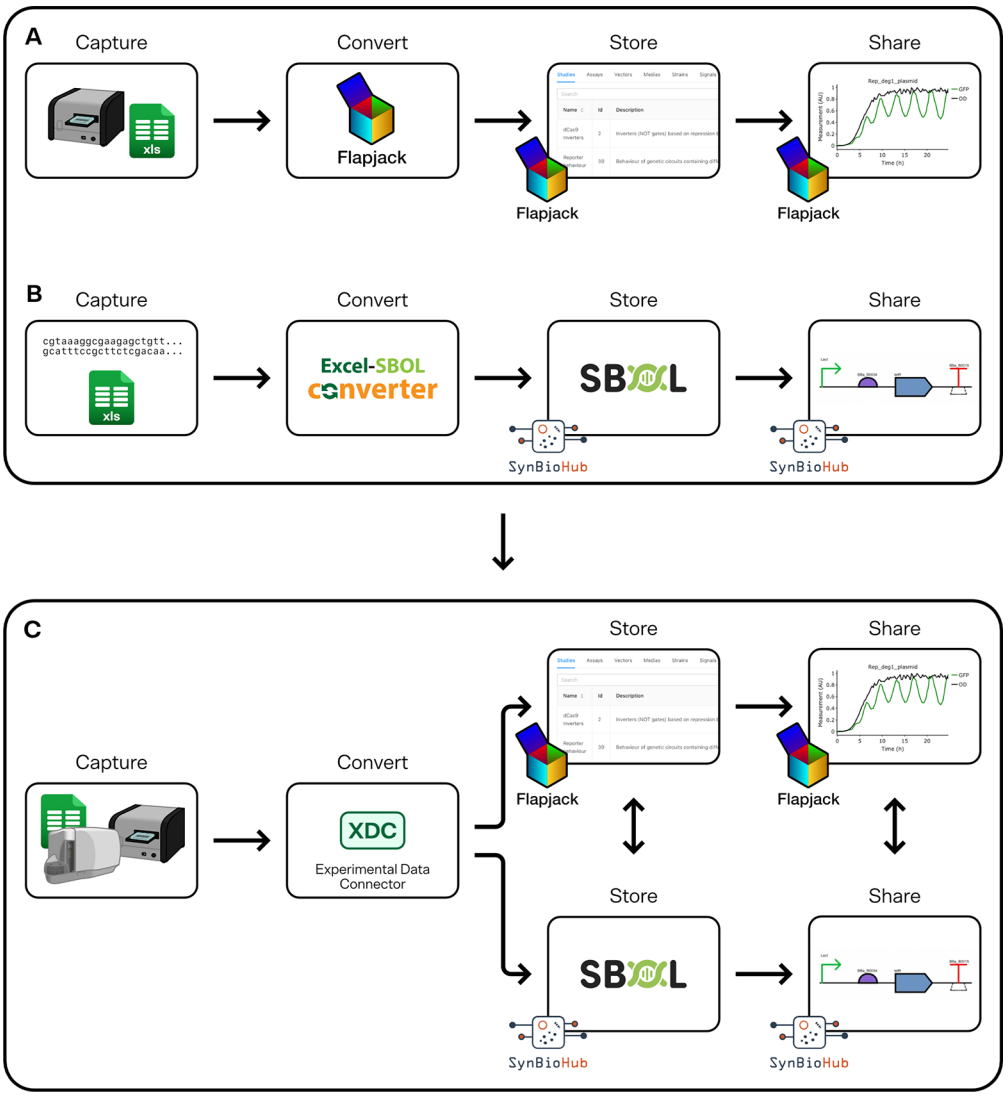


Figure 1. Diagram of the existing workflows (A and B) and the new workflow (C) for capturing, encoding, and connecting experimental data. (A) Workflow for storing experimental data in Flapjack. The data is captured from a plate reader, which delivers the readings in an Excel (XLS) file, which is modified to incorporate the missing metadata necessary to be uploaded to Flapjack; this is then converted to the Flapjack data model and uploaded to the platform through the frontend and stored on the server. Once the data is stored, it can be viewed or shared. (B) Workflow for storing data in SynBioHub. Genetic information is captured either through sequences obtained from text, GenBank, or by using an Excel file. This file is then converted to an SBOL file by using the Excel-to-SBOL Converter,[6] which stores metadata in SynBioHub and subsequently is shared. (C) Workflow using the XDC. Data can be obtained from various sources, including a plate reader and fluorescence cytometer, which are then captured by a template Excel workbook. These sheets are converted using the XDC, which generates an SBOL file with the experimental measurement data, uploads the relevant information to Flapjack with links to the SBOL objects, and uploads the SBOL representation of the metadata with links to Flapjack experimental data in SynBioHub.

incorporate measurement data, but in the meantime, if the XDC is used, the SBOL representation of the metadata can be accessed via SynBioHub. SynBioHub and Flapjack are accessible by API frameworks to facilitate development and integration with the wider computational synthetic biology environment.

The current workflow to upload data to SynBioHub and Flapjack is a manual, time-consuming process that requires new data imports for each new experimental context and leads to data that is not linked. This is partially due to the fact that Flapjack's data model lacks a field to implement this connection and its frontend interface has not yet implemented a bulk upload functionality. Thus, having more than one assay implies uploading each Excel file individually to SynBioHub and Flapjack. Furthermore, plate readers from different brands produce different outputs requiring a different Flapjack parser for each one. This creates difficulties for the tool.

Here we present a new software tool, the experimental data connector (XDC), for experimental data capture using the Excel-to-SBOL Converter,[6] Flapjack,[2] and SynBioHub.[7] The XDC uses an Excel template to capture both experimental data and metadata and upload it to Flapjack and SynBioHub, respectively. The XDC links data stored in SynBioHub and Flapjack. Thus, experimental metadata can be found in SynBioHub with links on the page to the Flapjack IDs of the associated measurement data. In the future, the measurement data in Flapjack will contain URLs that link to the associated metadata found in SynBioHub. The uploaded data may thus be accessed either from the Flapjack web interface or the SynBioHub web interface, though neither portal contains the

https://doi.org/10.1021/acssynbio.2c00669
ACS Synth. Biol. 2023, 12, 1364−1370

A

File Upload

SynBioHub

Username

Password

Submission Collection Name

This must be characters a-z,A-Z, or 0-9.
It may not contain spaces and must start with a letter.

Overwrite Collections of the same name?

Flapjack

Username

Password

Excel File

Choose File No file chosen

Submit

B

Studies

Study ID	Study Name
Study 1	Reporter behavior

Assays

Assay ID	Assay Name	Study ID
Assay 1	GFP plate 1	Study 1
Assay 2	GFP plate 2	Study 1

Sample Designs

Sample Design ID	Media ID	Strain ID	Vector ID	Supplement ID
Sample Design 1	M9Glucose	E.coli T7	RepLowDeg	—
Sample Design 2	M9Glucose	E.coli T7	RepHighDeg	—

Samples

Sample ID	Assay ID	Sample Design ID	Row	Column
Sample 1	Assay 1	Sample Design 1	1	1
Sample 2	Assay 1	Sample Design 2	1	2
Sample 3	Assay 2	Sample Design 1	1	1
Sample 4	Assay 2	Sample Design 2	1	2

Measurements

Measurement ID	Sample ID	Signal ID	Time	Value
Measurement 1	Sample 1	OD	0	0.03500000
⋮				
Measurement 800	Sample 4	GFP	24.75	0.95729347

Figure 2. (A) The XDC front end. The XDC uses the information collected here to create the linked uploads to Flapjack and SynBioHub. (B) XDC Case Study. Showing the design of a case study experiment comprising a study with measurements every 15 min over a period of 25 h with two repeats. For each repeat, there is one plate per repeat with one assay per plate. Each plate has two samples per assay for a total of four samples at two different sample design setups. Measuring each sample at each of the 100 time points results in 800 data points. Given this experimental setup, five different sheets are used to describe and connect information for Studies, Assays, Sample Designs, Samples, and Measurements. Note the way the sheets are linked via IDs. For example, assays identify the study they are part of using the "Study ID" column and similarly, samples to assay with "Assay ID", and measurements to sample with "Sample ID".

full uploaded data set. This allows users to retrieve data from either repository and share it with others in a standard manner, improving reproducibility and collaboration.

RESULTS

The XDC, illustrated in Figure 1, provides a user-friendly workflow to capture, encode, and connect experimental data with its related metadata leveraging the established digital repositories Flapjack and SynBioHub. The existing steps involve uploading data to Flapjack and SynBioHub as separate workflows. To store experimental data, it has to be collected in a formatted spreadsheet and uploaded to Flapjack using the frontend (Figure 1A). Then, to store the related metadata, it is necessary to capture it from a text, GenBank, or SBOL file; then, the file is later separately converted to an SBOL file as part of the SynBioHub upload process (Figure 1B). The proposed workflow streamlines this process by aggregating all the captured data into a template Excel workbook, which in turn is processed by the XDC to generate an SBOL file and upload the data to Flapjack and SynBioHub simultaneously, linking the data between the platforms in the process (Figure 1C). The XDC can be accessed via a user interface shown in Figure 2A (https://xdc.synbiohub.org/upload).

The template Excel workbook is designed to collect a wide range of experimental SynBio data (Figure 2B). It was developed as an extension of the Excel-SBOL master template.[6] The workbook is agnostic to the equipment that generated the data allowing the use of different plate readers, flow cytometers, or any other equipment from a variety of vendors. An example of plate reader output data going into the template is shown in Figure 3. There is also a macro-enabled template that automates this process; however, it is not equipment agnostic. In the macro enabled template macros and functions are added to automatically pull information from Synergy HTX plate reader outputs (Algorithm 3). The workbook encodes all the data fields to standardized formats using SBOL2[8] and the Flapjack data model.[2] If necessary, the Excel-SBOL mapping can be modified by changing the "column_definitions" sheet, and the Excel-Flapjack data mapping can be modified by changing the "FlapjackCols" sheet.

The XDC provides an improved workflow for capturing, encoding, and representing experimental data and metadata. An example of this workflow could be as follows:

1. A researcher designs an experiment to compare gene expression in repressilators with different degradation rates.
2. The researcher designs two repressilators, one with the transcription factors and GFP reporter without degradation tags and another with a degradation tag like AAV on all those proteins.
3. The researcher builds the necessary plasmids, transforms them into a chassis, and tests them by measuring OD and fluorescence on a plate reader.
4. The plate reader data is entered into the template workbook (either manually or in a semiautomated way), see Figure 3.
5. Researchers can relate experimental measurement data to its metadata, including the conditions of the experimental study, how the DNA was built, the parts used to create the DNA, and the original developer of the parts, see Figure 2B. This Excel template workbook is provided to

https://doi.org/10.1021/acssynbio.2c00669
ACS Synth. Biol. 2023, 12, 1364−1370

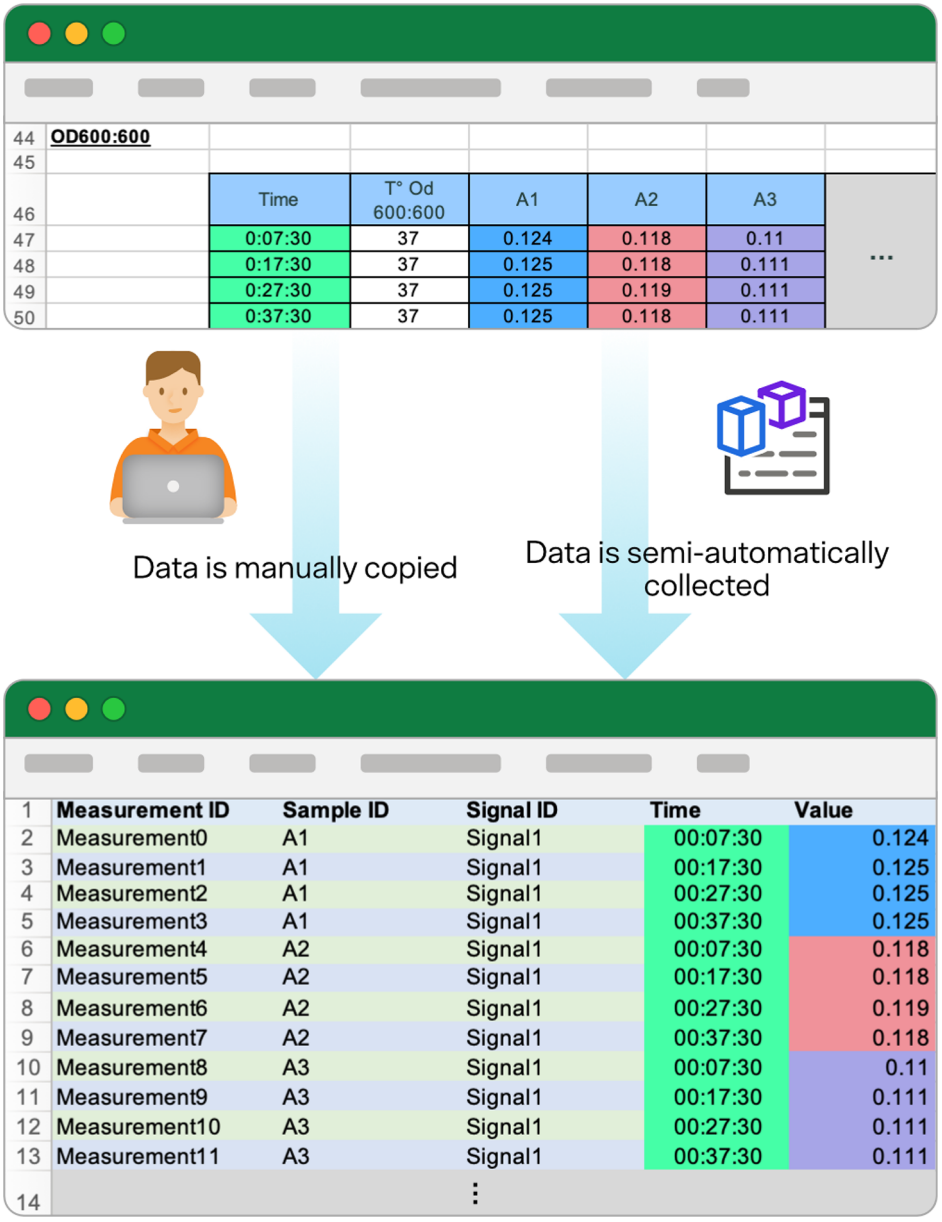


Figure 3. Workflows for converting plate template reader data into the XDC template data. The top shows an example plate reader output file and the bottom shows the XDC measurement sheet. The columns have been highlighted to show how the data from one relates to data in the other. In the **Manual workflow**, the researcher takes the data from the exported file from the plate reader and copies the data into the XDC spreadsheet; this is a versatile approach since the researcher is in charge of making the data correlations regardless of the input file. In the **Semiautomated** approach, a macro combined with cell formulas takes data from the plate reader file and automatically enters it into the XDC template (Algorithm 3). To enable this, both files (the plate reader export and the XDC sheet) must be in the same folder. This allows fewer user interactions and a more straightforward modification than using Excel's functions; however, it is currently limited to Synergy HTX spreadsheet files.

researchers in the XDC repository (https://github.com/SynBioDex/Xperimental-Data-Connector-Server).

6. The Excel workbook is uploaded to the XDC server, see Figure 2A. The XDC creates a complete linked data set stored across Flapjack and SynBioHub. Data can then be accessed via either web portal.

This workflow enables the analysis of the measurement data in Flapjack to be more reproducible with the capture of the study setup, sample designs, and other design data. In this way, the user is able to both identify the implementations of DNA parts employed in the study and access the underlying experimental data results through links that may be developed for measurement and analysis data in Flapjack.

DISCUSSION

The XDC provides a semiautomated workflow to assist researchers with the transition from build and test to the learn stage. This workflow captures data using a template Excel workbook then a software engine creates, encodes, and uploads experimental data and metadata in a standardized format. The XDC enables the use of digital repositories, Flapjack and SynBioHub, to reproduce experimental findings and improve data sharing. The user interface manages the data encoding and conversion to provide access to a wide range of researchers, particularly those with limited coding skills. Finally, the tool supports the use of the SBOL standard to capture build and test experimental data in machine readable formats without prior

https://doi.org/10.1021/acssynbio.2c00669
ACS Synth. Biol. 2023, 12, 1364−1370

Algorithm 1: Experimental Data Connector

Input: Filled Excel Template, Flapjack Username, Flapjack Password, SynBioHub Username, SynBioHub Password, SynBioHub Collection Name

Create Temporary Directory
Read in Excel file and use Excel-SBOL Converter to create an SBOL File with placeholder Flapjack IDs
Log in to SynBioHub
Pull user graph URI
Create a dictionary of object names to SBOL URIs using the appropriate graph URI and Collection Name
Upload data, including SBOL URIs to Flapjack using Excel-Flapjack Converter and save returned Flapjack IDs
Update SBOL File with new Flapjack IDs
Upload SBOL File to SynBioHub

A

Sheet Name	Column Name	Flapjack Name
Supplement	Supplement Name	name
	Chemical ID	chemical
	Concentration	concentration
Chemical	Chemical Name	name
	Chemical Description	description
		pubchemid
Vector	Vector Name	name
	DNA ID	dnas
Strain	Strain Name	name
	Strain Description	description
Media	Media Name	name
	Media Description	description
Signal	Signal Name	name
	Signal Description	description
	Signal Color	color
DNA	DNA Name	name
		sboluri
Measurement	Signal ID	signal
	Value	value
	Sample ID	sample
	Time	time
Sample	Row	row
	Column	col
	Supplement Name	supplement
	Vector ID	vector
	Strain ID	strain
	Media ID	media
	Assay ID	assay
Assay	Assay Name	name
	Machine	machine
	Description	description
	Study ID	study
	Temperature	temperature
Study	Study Name	name
	Description	description
	DOI	DOI
		public

B

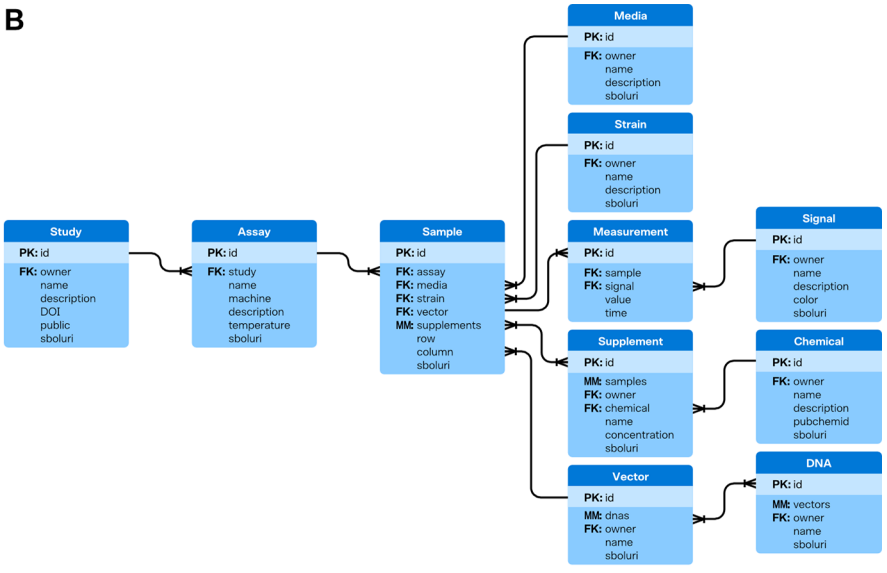


Figure 4. (A) FlapjackCols sheet. The Connector uses this sheet to determine how data should be uploaded to Flapjack. It provides the Sheet Name, Column Name, and the Flapjack property name it should be turned into. This allows greater flexibility in the column names while maintaining the Flapjack data model. (B) The Flapjack data model. Arrows show how data types connect to each other. For example, studies contain assays.

knowledge of SBOL. The column definition sheet in the Excel workbook supports the creation of new workbook versions that could encode the data into future version of SBOL[5] and convert data to future versions of Flapjack or other repositories.

In the future, we plan to expand the linkage between SynBioHub and Flapjack repositories by making the Flapjack IDs link to Flapjack viewing pages. Additionally, we intend to integrate the XDC more fully with SynBioHub by incorporating it in an authorization plugin enabled submit plugin to improve user identification (for more information on plugins, see ref 9); this would also allow submission to different instances of SynBioHub and Flapjack (at the moment the instance configurations are not accessible to the user).[10] We envision a more connected and automated DBTL cycle in research laboratories and industrial applications. However, more tools and workflows are needed to connect the DBTL cycle fully. Automation tools like liquid handling robots and lab management systems will automatically capture relevant metadata. Automated workflows will coexist with spreadsheet metadata and experimental data capture to improve the reproducibility of experimental workflows.

METHODS

The XDC backend and frontend are written in Python. The pseudocode description of the algorithm is shown in Algorithm 1. The use of the "column_definitions" sheet is explained in the

https://doi.org/10.1021/acssynbio.2c00669
ACS Synth. Biol. 2023, 12, 1364−1370

Algorithm 2: Excel-to-Flapjack

Input: Filled Excel Template, Flapjack Username, Flapjack Password

Sign in to Flapjack
Read in Excel file **for** Object *in* [Chemical, DNA, Supplement, Vector, Strain, Media, Signal, Study, Assay, Sample, Measurement] *do*
  Read in the FlapjackCols Sheet with the sheet name of the Object
  Convert the table to a dictionary
  Read in the Sheet with the name of the Object
  Set the 'Object ID' column as the index
  Drop any columns not used by Flapjack and rename the ones that are in the FlapjackCols Sheet
  Turn the table into a dictionary
  Change any 'lookup columns' to Flapjack IDs rather than names
  Upload the dictionary information to Flapjack
  Add Flapjack IDs to a dictionary to allow further processing of 'lookup columns'

Algorithm 3: Excel Plate Reader Input Automation

Place both spreadsheets in the same directory
Macro updates the directory information in cell formulas
Measurement IDs are created using the starting value found in K6 and then incrementing
Time values are added using cell referencing to column B (starting at row 47)
Measurement values are found using the `VLOOKUP`, the time value in the XDC sheet and the column number specified in K2 (e.g. 3 means column E).

Excel-to-SBOL previous work.[6] The "FlapjackCols" is shown in Figure 4A and takes column names and changes them to Flapjack data object names.

The XDC creates Flapjack objects using the excel2flapjack module (https://github.com/SynBioDex/Excel-to-Flapjack) developed in this work. The pseudocode of the excel2flapjack module is explained in Algorithm 2. The Connector creates SBOL objects using the existing Excel-to-SBOL library (https://github.com/SynBioDex/Excel-to-SBOL).

To allow the connections between SynBioHub and Flapjack, SynBioHub URI fields were added to the Flapjack data model for every object but Measurement, which is the experimental data. pyFlapjack (https://github.com/RudgeLab/pyFlapjack) was also improved in this work by extending the Create function to support overwrite prevention during the bulk upload of data. Simulated data for test and examples were generated using LOICA.[11]

The macro enabled XDC template uses a combination of macros and cell formulas as explained in Algorithm 3.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssynbio.2c00669.

Excel Template for Use with the Experimental Data Connector (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Jeanet Mante − *University of Colorado Boulder, Boulder, Colorado 80309, United States;* orcid.org/0000-0002-1450-5638; Email: jv@mante.net

Authors

Sai P. Samineni − *University of Colorado Boulder, Boulder, Colorado 80309, United States*

Gonzalo Vidal − *Interdisciplinary Computing and Complex Biosystems, School of Computing, Newcastle University, Newcastle upon Tyne NE1 7RU, United Kingdom;* orcid.org/0000-0003-3543-520X

Carolus Vitalis − *Interdisciplinary Computing and Complex Biosystems, School of Computing, Newcastle University, Newcastle upon Tyne NE1 7RU, United Kingdom;* orcid.org/0000-0003-3867-0395

Guillermo Yáñez Feliú − *Interdisciplinary Computing and Complex Biosystems, School of Computing, Newcastle University, Newcastle upon Tyne NE1 7RU, United Kingdom*

Timothy J. Rudge − *Interdisciplinary Computing and Complex Biosystems, School of Computing, Newcastle University, Newcastle upon Tyne NE1 7RU, United Kingdom;* orcid.org/0000-0001-9446-9958

Chris J. Myers − *University of Colorado Boulder, Boulder, Colorado 80309, United States;* orcid.org/0000-0002-8762-8444

Complete contact information is available at:

https://pubs.acs.org/10.1021/acssynbio.2c00669

Author Contributions

‡SPS and GV contributed equally to this research.

Author Contributions

All authors contributed to the writing of this manuscript. SPS developed the experimental data conversion workflows supported by JM and GV. JM wrote the XDC library and front end. JM, SPS, and GV created the Excel-to-Flapjack library. CV developed the semiautomated measurement input supported by SPS and GV. GV, GYF, and CV provided support on the Flapjack API. SPS and JM created the workbook template. GV wrote the code to produce simulated data used in the Excel workbook case study. CJM and TJR supervised the entire project and team.

Notes

The authors declare no competing financial interest.

Links with further information regarding the Experimental Data Connector. Server GitHub: https://github.com/SynBioDex/Xperimental-Data-Connector-Server. Package GitHub: https://github.com/SynBioDex/Xperimental-Data-Connector. Excel-to-Flapjack GitHub: https://github.com/SynBioDex/Excel-to-Flapjack. Usable Instance: https://xdc.synbiohub.org/upload. Excel Template: https://github.com/SynBioDex/Xperimental-Data-Connector-Server/blob/main/xperimental_data_connector_v001.xlsx. Macro Enabled Excel Template: https://github.com/SynBioDex/Xperimental-Data-Connector-Server/blob/main/xperimental_data_connector_macro_enabled_v001.xlsm.

■ ACKNOWLEDGMENTS

JM, SPS, and CM are supported by the National Science Foundation under Grant Nos. 1939892 and 2231864. The Experimental Data Connector is run on a Microsoft Azure Server provided by Microsoft Research. GV, CV, GYF, and TJR are supported by the Newcastle University School of Computing. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies.

■ REFERENCES

(1) McLaughlin, J. A.; Myers, C. J.; Zundel, Z.; Mısırlı, G.; Zhang, M.; Ofiteru, I. D.; Goñi-Moreno, A.; Wipat, A. SynBioHub: A Standards-Enabled Design Repository for Synthetic Biology. *ACS Synth. Biol.* **2018**, *7*, 682−688.

(2) Yáñez Feliú, G.; Earle Gómez, B.; Codoceo Berrocal, V.; Muñoz Silva, M.; Nuñez, I. N.; Matute, T. F.; Arce Medina, A.; Vidal, G.; Vidal Céspedes, C.; Dahlin, J.; Federici, F.; Rudge, T. J. Flapjack: Data Management and Analysis for Genetic Circuit Characterization. *ACS Synth. Biol.* **2021**, *10*, 183−191.

(3) Galdzicki, M.; Clancy, K. P.; Oberortner, E.; Pocock, M.; Quinn, J. Y.; Rodriguez, C. A.; Roehner, N.; Wilson, M. L.; Adam, L.; Anderson, J. C.; et al. The Synthetic Biology Open Language (SBOL) provides a community standard for communicating designs in synthetic biology. *Nature biotechnology* **2014**, *32*, 545−550.

(4) Roehner, N.; Beal, J.; Clancy, K.; Bartley, B.; Misirli, G.; Grunberg, R.; Oberortner, E.; Pocock, M.; Bissell, M.; Madsen, C.; et al. Sharing structure and function in biological design with SBOL 2.0. *ACS synthetic biology* **2016**, *5*, 498−506.

(5) McLaughlin, J. A.; Beal, J.; Mısırlı, G.; Grünberg, R.; Bartley, B. A.; Scott-Brown, J.; Vaidyanathan, P.; Fontanarrosa, P.; Oberortner, E.; Wipat, A.; et al. The synthetic biology open language (SBOL) version 3: simplified data exchange for bioengineering. *Front. Bioeng. Biotechnol.* **2020**, *8*, 1009.

(6) Mante, J.; Abam, J.; Samineni, S. P.; Pötzsch, I. M.; Beal, J.; Myers, C. J. Excel-SBOL Converter: Creating SBOL from Excel Templates and Vice Versa. *ACS Synth. Biol.* **2023**, *12*, 340.

(7) McLaughlin, J. A.; Myers, C. J.; Zundel, Z.; Mısırlı, G.; Zhang, M.; Ofiteru, I. D.; Goni-Moreno, A.; Wipat, A. SynBioHub: a standards-enabled design repository for synthetic biology. *ACS synthetic biology* **2018**, *7*, 682−688.

(8) Galdzicki, M.; et al. The Synthetic Biology Open Language (SBOL) provides a community standard for communicating designs in synthetic biology. *Nat. Biotechnol.* **2014**, *32*, 545−550.

(9) Mante, J. V. Promotion of Data Reuse in Synthetic Biology. Ph.D. thesis, CU Boulder, Boulder, CO, 2022.

(10) Mante, J.; Zundel, Z.; Myers, C. Extending SynBioHub's Functionality with Plugins. *ACS Synth. Biol.* **2020**, *9*, 1216−1220.

(11) Vidal, G.; Vidal-Céspedes, C.; Rudge, T. J. LOICA: Integrating Models with Data for Genetic Network Design Automation. *ACS Synth. Biol.* **2022**, *11*, 1984−1990.