Collaborative Terminology: SBOL Project Dictionary

Jacob Beal, Daniel Sumorok, Bryan Bartley, Tramy Nguyen* Raytheon BBN Technologies, Cambridge MA

jakebeal@ieee.org

1 INTRODUCTION

Sharing information about biological experiments between researchers is often challenging. Reagents, strains, and genetic constructs are often given "shorthand" names that are ambiguous (e.g., "ara" for L-arabinose), differ between researchers (e.g., "L-arab" vs. "Arabinose") or are unknown outside of a particular group (e.g., "plasmid 37"). Likewise, the particular combinations used in each sample of an experiment are often expressed in variable personal shorthands, often accidentally omitting important details.

Humans can sometimes infer sufficient information to interpret such informal documentation of experiment designs, but this is typically an ad-hoc, challenging, and errorprone process, not particularly susceptible to automation. At the same time, precise and unambiguous specifications of both elements and their combinations can be expressed in machine-readable representations such as SBOL [2], but making use of these tools is difficult for many researchers. Common machine-readable terminology also typically needs to be agreed upon in advance, which is often onerous or impossible given the ongoing evolution of terms in research projects—indeed, some studying the philosophy of science argue that the ability to define terminology is a good marker for the conclusion of a scientific investigation! [3]

The SBOL Project Dictionary helps bridge this gap with a simple spreadsheet-based interface that allows researchers to collaboratively and incrementally construct a shared terminology. This interface provides a structured format of tabs and columns for researchers to link lab-specific names to shared names, aliases, and canonical definitions using SBOL. Software tooling can then access this set of relations at any time in order to translate metadata terms into an evolving common vocabulary, thereby supporting simple post-hoc integration and debugging across collaborators.

2 ARCHITECTURE

Figure 1 shows the architecture used to implement the SBOL Project Dictionary. This implementation is based on two key



Figure 1: Architecture of SBOL Project Dictionary: the dictionary is held in a Google Sheet, which is maintained and synchronized with canonical storage of names, aliases, and definitions in SynBioHub. Users interact either directly or through tools using a Dictionary Writer API, and receive email notification of integration problems to resolve.

pre-existing tools: Google Sheets, which provides a collaborative spreadsheet editing interface and API for software tools, and SynBioHub [4], a database server for sharing synthetic biology information encoded in SBOL [2]. These are linked by configuring a Dictionary Maintainer service to connect to a particular Google Sheet and SBOL collection in SynBio-Hub. Periodically, the Dictionary Maintainer updates and synchronizes. First, it ensures the Google Sheet is configured to follow a specified format, including protecting regions that should not be user-editable. It then validates all of the information in the SBOL Project Dictionary and synchronizes with SynBioHub, creating SBOL objects as needed to store new dictionary entries, importing SBOL links when matched by a new entry in the dictionary, and reporting errors to the user via a status column in the spreadsheet.

In particular, the SBOL Project Dictionary provides and maintains spreadsheet columns for the following:

- A single common name, which is the researchers' current consensus on a human-friendly term for each entry, e.g., "Synthetic Complete Media"
- Aliases, which are alternative terms shared between researchers, e.g., "SC Media", "Synthetic Complete". There may be many such aliases per entry.

^{*}This work was supported by the Air Force Research Laboratory (AFRL) and DARPA under contract FA875017CO184. 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 AFRL and DARPA.

	- The East view insert Format De	ita roois	Auu	ons help Anonanges saved online				
ŝ	→ → → → 100% → \$ % .00	0 123∓	Arial	- 11 - B <i>I</i> - A		G2 🖬 🛋 🝸 🖌 Σ -		^
λ.								
	۸	• C		D	• G •	L L	M	• 0
1			Ŧ					SD2 Dictionary (1.4.7) Updated 20
2	Common Name	Туре		SynBioHub URI	Transcriptic UID	Definition URI / CHEBI ID	Stub Object?	Status
1	SC+Oleate+Adenine	Media	-	https://hub.sd2e.org/user/sd2e/design/culture_m	Modified M9 Media		NO	Updated 2020-02-20 22:48:32: Tra
2	SD+Glucose+Adenine+Leucine+Isoleucine	Media	-	https://hub.sd2e.org/user/sd2e/design/culture_m	edia_6/1		NO	Updated 2020-03-11 22:39:58
3	Yeast_Extract_Peptone_Adenine_Dextrose (a.	k Media	-	https://hub.sd2e.org/user/sd2e/design/culture_m	srs1b8cybgdd7gz		NO	Updated 2020-03-11 22:39:58
4	Synthetic_Complete	Media	-	https://hub.sd2e.org/user/sd2e/design/culture_m	grs1b4uj7zbdy6m,sc_media,SC Media		NO	Updated 2020-03-11 22:39:58
5	LB Broth (Miller)	Media	*	https://hub.sd2e.org/user/sd2e/design/LB_Broth/	1	https://www.thermofisher.com/us/en/home/life-scie/	NO	Updated 2020-03-11 22:39:58
16	LB_Cm50	Media	*	https://hub.sd2e.org/user/sd2e/design/LB_Cm50/	1		NO	Updated 2020-03-11 22:39:58
17	DO Supplement -His/-Leu/-Trp/-Ura	Media	*	https://hub.sd2e.org/user/sd2e/design/CAT_6304	25/1	http://www.clontech.com/	NO	Updated 2020-03-11 22:39:58
18	Thermo Scientific Remel Yeast Nitrogen Base v	v Media	*	https://hub.sd2e.org/user/sd2e/design/CAT_R459	942/1	http://www.fishersci.com/	NO	Updated 2020-03-11 22:39:58
19	BD Bacto Yeast Extract BD Biosciences	Media	¥	https://hub.sd2e.org/user/sd2e/design/CAT_9000	0_726/1	http://us.vwr.com/	NO	Updated 2020-03-11 22:39:58
20	BD Bacto Dehydrated Culture Media Additive_	Media	¥	https://hub.sd2e.org/user/sd2e/design/CAT_DF01	23_17_3/1	http://www.fishersci.com/	NO	Updated 2020-03-11 22:39:58
21	M9 Glucose CAA (a.k.a. M9 Glucose Stock)	Media	Ŧ	https://hub.sd2e.org/user/sd2e/design/M9_gluco	s rs1apwazmvzbqy	https://www.thomassci.com/Laboratory-Supplies/M	NO	Updated 2020-03-11 22:39:58
22	M9 Media Salts	Media	¥	https://hub.sd2e.org/user/sd2e/design/teknova_N	/1902/1	https://www.teknova.com/5X-MINIMAL-SALTS	NO	Updated 2020-03-11 22:39:58
23	LUDOX-S40	Solution	Ŧ	https://hub.sd2e.org/user/sd2e/design/ludox_S40	y rs1b6z2vgatkq7		NO	Updated 2020-03-11 22:39:58
24	SYTOX Red	Stain	Ŧ	https://hub.sd2e.org/user/sd2e/design/CAT_S348	5 rs1bga86uq52tz		NO	Updated 2020-03-11 22:39:58
25	Phosphate Buffered Saline	Buffer	Ŧ	https://hub.sd2e.org/user/sd2e/design/pbs/1	rs1dbqpedruj98j		NO	Updated 2020-03-11 22:39:58
26	IPTG	CHEBI	Ŧ	https://hub.sd2e.org/user/sd2e/design/IPTG/1	rs18vwgfgxq597, IPTG	http://identifiers.org/chebi/CHEBI:61448	NO	Updated 2020-03-11 22:39:58
27	L-arabinose	CHEBI	Ŧ	https://hub.sd2e.org/user/sd2e/design/Larabinose	rs1apwddkptp5q, Arabinose	http://identifiers.org/chebi/CHEBI:30849	NO	Updated 2020-03-11 22:39:58
28	aTc	CHEBI	Ŧ	https://hub.sd2e.org/user/sd2e/design/aTc/1		http://identifiers.org/chebi/CHEBI:17146	NO	Updated 2020-03-11 22:39:58
29	ddH2O (sterile ultra-pure water)	CHEBI	Ŧ	https://hub.sd2e.org/user/sd2e/design/ddH2O/1		http://identifiers.org/chebi/CHEBI:15377	NO	Updated 2020-03-11 22:39:58

Figure 2: Screenshot of the SBOL Project Dictionary as deployed in the DARPA SD2 program showing key entry columns, including name, type, canonical SynBioHub URI, laboratory UIDs, grounding definition, and status.

- A URL for an entry in SynBioHub, which is guaranteed never to change once an entry has been created.
- Type information (e.g., RNA, DNA, cell strain, media).
- Lab-specific identifiers for each distinct set of collaborators, such as personal shorthands, obsolete terms, or laboratory information management systems (LIMS) unique identifiers (which are generally not humaninterpretable). Each lab gets its own column, and there may be many identifiers for each item in each lab.
- Links to a canonical definition, e.g., to curated databases such as ChEBI, NCBI taxonomy, or UniProt, or to the suppliers of complex reagents.
- Status, time last updated, and errors such as type mismatches or detection of duplicate identifiers.

All of these except for the SynBioHub URL and status can be freely edited by researchers at any time, reflecting the ongoing evolution of their private and shared vocabularies (note that ensuring a term's definition remains coherent over time is not generally machine-checkable, and is thus left to the researchers sharing the document). The SBOL Project Dictionary provides different tabs for six main types of vocabulary items—genetic constructs, strains, proteins, reagents, collections, and attributes—with drop-down menus supporting sub-types on each tab as needed.

Users can create and edit entries directly by browsing to the Google Sheet or via software tools they set up to interact with a provided Dictionary Writer API, written in Python. Once entries have been established, other tools can then use the definitions in SynBioHub for integration and display of metadata, mapping between private names, common names, and detailed canonical definitions as needed. When such mappings fail, the SBOL Project Dictionary also provides a "Mapping Failures" tab into which such problems can be reported. Periodically, the dictionary scans this tab and emails the provided contacts for a laboratory to let them know when there are missing entries or errors in entry columns that they are responsible for, such that they can update those entries and resolve mapping failures. This feedback closes the loop to enable not only incremental and post-hoc integration but also incremental and post-hoc error resolution

3 CASE STUDY: DARPA SD2 PROGRAM

In DARPA's Synergistic Discovery and Design (SD2) program, the SBOL Project Dictionary forms a key component in integrating a design-build-test-learn round trip between experimentalists, laboratory automation, and data analysts [1]. Over a period of approximately a year and a half, the SD2 deployment of the SBOL Project Dictionary has been used to curate collaborative terminology for more than 1000 terms: 521 genetic constructs, 304 strains, 54 proteins, 89 reagents, 7 collections, and 62 attributes, a sampling of which are shown in the screenshot in Figure 2. A large fraction of these have been entered and updated incrementally by hand by many different participating researchers, while others have been uploaded using automation-particularly entries for laboratory LIMS identifiers. The entries in the SD2 dictionary have been used to support the integration of data and metadata for dozens of experiments spanning four performing laboratories and five different working groups of researchers, each involving different organisms, goals, and technologies.

REFERENCES

- BRYCE, D., ET AL. Round-trip: An automated pipeline for experimental design, execution, and analysis. In *IWBDA* (2020).
- [2] COX, R. S., ET AL. Synthetic Biology Open Language (SBOL) version 2.2.
 0. Journal of integrative bioinformatics 15, 1 (2018).
- [3] LATOUR, B., AND WOOLGAR, S. Laboratory Life: The Construction of Scientific Facts. Princeton University Press, 2013.
- [4] MCLAUGHLIN, J. A., ET AL. SynBioHub: A standards-enabled design repository for synthetic biology. ACS Syn. Bio. 7, 2 (2018), 682–688.