Intent Parser: a tool for codifying experiment design

Tramy Nguyen¹, Nicholas Walczak¹, Jacob Beal¹, Daniel Sumorok¹, Mark Weston^{2*} ¹Raytheon BBN Technologies, ²Netrias, Cambridge MA tramy.t.nguyen@raytheon.com,nicholas.walczak@raytheon.com

1 INTRODUCTION

Many biological experiments are described in text documents, such as laboratory notebooks, capturing information such as the purpose, execution, and results of an experiment. In such descriptions, however, authors typically present information in a highly personal and idiosyncratic manner, at varying levels of detail and often omitting critical information. Consequently, this lack of consistency lead to a variety of issues commonly encountered when attempting to compare experimental reports created by different authors or to build upon those results in new work. Humans can sometimes infer sufficient information to interpret such informal documentation of experiment designs, but this is typically an ad-hoc, challenging, and error-prone 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 investigators.

A "middle-ground" approach combining both accessibility and representational precision, however, has been known at least as far back as Winograd's SHRDLU system [4], using machine feedback and prompting to shape human input into a semi-structured form that can be readily interpreted and checked by machines. We have applied this approach to develop Intent Parser, a tool that combines a word-processing interface with structured tables and assisted linking to definitions to provide a simple interface for incremental codification of experiment designs. Use of this tool can help synthetic biology collaborations by reducing the time and skills required to produce precise experiment designs, enabling automatic checking for errors and ambiguities, and simplifying interpretation of experimental data.

2 INTENT PARSER ARCHITECTURE

Fundamentally, the Intent Parser acts as a link between a user-friendly document editor (in this case Google Docs) and a repository of formal definitions (in this case SynBio-Hub [3]). By linking these and adhering to certain document

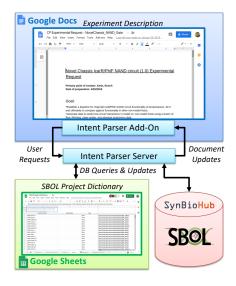


Figure 1: Architecture of Intent Parser: a Google Docs addon sends user requests to a "back-end" server that interacts with databases of definitions (SynBioHub) and terminology (SBOL Project Dictionary). Results return to the addon, which offers the user choices and alters the document.

conventions, the tool allows users to generate unambiguous machine-readable descriptions of experiment design. In our implementation, we have chosen to use Google Docs for the editor, SynBioHub [3] as the repository (in combination with the SBOL Project Dictionary interface [1]), and to output experiment design specifications in JSON.

Intent Parser is implemented as an "add-on" to Google Docs using the Google Docs API to implement the architecture shown in Figure 1. Once installed, this add-on provides a file menu of operations that can be performed on any Google Doc. Users conceiving of an experiment write up an experiment description in a Google Doc and invoke requests around two workflows: 1) grounding document text with links to definitions in SynBioHub [3], and 2) defining, validating, and exporting experiment requests that make use of those definitions. These requests can be made at any time, supporting an incremental and collaborative approach to experiment design.

When the user makes a request in the add-on, an HTTP request is sent to the Intent Parser Server, which then parses the document and returns an HTTP response with the result back to the add-on. The server is a backend that carries out requests by interacting with SynBioHub [3], which provides

^{*}Supported by 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 AFRL and DARPA.

IWBDA 2020, August 2020, Online

Nguyen, Walczak, et al.

01000 010000 01000 01000 <t< th=""><th></th><th>py_of_NovelChassis Request Bacillus WT Characterization ☆ 🖿 nsert Format Tools Add-ons Help <u>Last edit was made on January 28 by Tramy I</u></th><th>Nguyen</th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th>🗏 🚊 Share</th></t<>		py_of_NovelChassis Request Bacillus WT Characterization ☆ 🖿 nsert Format Tools Add-ons Help <u>Last edit was made on January 28 by Tramy I</u>	Nguyen									🗏 🚊 Share
Image: Section of the performed? By who? Add Image: Section of the performed? By who? Image: Section of the performed? By who? Image: Section of the performed? By who? Image: Section of the performed? By who? Image: Section of the performed? By who? Image: Section of the performed? By who? Image: Section of the performed? By who? Image: Section of the performed? By who? Image: Section of the performed? Section of the performed? Image: Section of the performed? By who? Image: Section of the performed? Section of the performed? Image: Section of the performed? By who? Image: Section of the performed? Image: Section of the performed? Section of the performed? Image: Section of the performed? Image: Section of the performed? Section of the performed? Image: Section of the performed? Image: Section of the performed? Image: Section of the performed? Image: Section of the performed? Image: Section of the performed? Image: Section of the performed? Image: Section of the performed? Image: Section of the performed? Image: Section of the performed? Image: Section of the performed? Image: Section of the performed? Image: Section of the performed? Image: Section of the performed? Image: Section of the performed? Image: Section of the performed? <td colspan="11"></td> <td>Apps Script application X</td>												Apps Script application X
Inter large response Inter suggestion Inter suggestion Inter respin												Add link to Glucose?
Other suggestions Other suggestions Other suggestions Other suggestions Importance (C-Shatocat) Importance (C-Shatocat)		Where might it be performed? By who?										No to All Never Link
Dextrose (D-Glucose) measurement file-type replicate stains timepoint temperature samples 0.000 </td <td>- - </td> <td>Lab: Ginkgo</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>Other suggestions:</td>	- - 	Lab: Ginkgo										Other suggestions:
C Implify/hub Add angly set/L O No Status				file-type	replicate	strains	timepoint	temperature	samples			
International control (11/4) (12/7) (13/3) (14/2) (13/3) (14/2) (13/3) (14/2) (14/3) (14/3) (14/2) (14/3)	-											Dextrose (D-Glucose) Link All
1009,1091,1174,1287,1391,1422,1504,1597,1670,1782,1835, ER Marburg In Print Prin		1.009, 1.091, 1.174, 1.257, 1.339, 1.422, 1.504, 1.587, 1.670, 1.752, 1.835,	RNA_SEQ		3		6 hours	30 celsius	72			Dextrose (D-Glucose)
What data as you expecting to collectivesive? Is there an ETL process in blace and who is resconsible? Air there expected results, and I so what are they? State any assertions about the dependence/independence of Link		1.009, 1.091, 1.174, 1.257, 1.339, 1.422, 1.504, 1.587, 1.670, 1.752, 1.835,		PLAIN, CSV	3		6, 7, 8 hours	30 celsius	216			(a.k.a. M9 Glucose
mini una un you capecing to concentent la trata una mini la reaponante. Pre una capecia reana, allo il so mini ale uney sia dalla adoli ile dependencente en		u. Mihat data ara unu experting to collect/raceks2 is there an ETL process in place	and who is respon	sible2 Are there	avpacted results	and if on what are they? Sta	te anv accertions also	ut the dependence	Sordeneordence of			SC+Glucose+Adenin Link ₽ Link All
measurements. If possible, estimate how much data expected.		measurements. If possible, estimate how much data expected.	and who is respon	auto i Ale Inere	expected (CSUIIS,	unu a so mina dre tiley i Sid	ic any assertions abo	ar are acpendence	anacycrachice of		Q	SC+Glucose+Adenin Link

Figure 2: Screenshot of Intent Parser in action on a document from the DARPA SD2 program, showing a measurement table with reagents linked to definitions in SynBioHub [3]. The navigation panel on the right suggests links to add, in this case a link for the term "Glucose" (document location not shown), providing both a best guess and a number of potential alternatives.

information about referenced elements in SBOL format [2], and with the SBOL Project Dictionary [1], which provides a spreadsheet interface that tracks shorthand and lab-specific names. Figure 2 shows an example of linking information found on SynBioHub to names and terminologies on the Google Doc.

Experiment designs are based on tables following a specific format, for which templates can be generated from an add-on menu item. In this abstract, these tables are referred as Intent Parser tables. Validation and export requests are sent to the Intent Parser Server to validate the contents of Intent Parser tables parsed from the Google Doc. Validation follows a predefined data schema that checks for the required information, using SynBioHub and the SBOL Project Dictionary to validate that all terms extracted from the table are properly grounded. From these, the server generates both reports on validity and JSON representations of experiment design.

3 CASE STUDY: DARPA SD2 PROGRAM

The DARPA Synergistic Discovery and Design (SD2) program aims to accelerate scientific discovery by machineassisted integration of experimental design, build, test, and learning, and is testing these aims via a collaboration of over 100 researchers across over a dozen organizations. In SD2, Intent Parser is used by both data scientists and experimentalists to define and request experiments via Google Docs.

Stakeholders including data scientists, subject matter experts, and experimental labs were consulted to help determine a format that was sufficiently general to specify experiment designs spanning across multiple challenge problems, protocols, laboratories, and experiment designs. The information recorded in these experiment requests includes the name of the lab to execute the experiment, which measurements are to be taken and at what time points, amounts of reagents, strains, and media to be used in each sample, as well as experimental conditions and parameters such as culturing temperature. As users describe the experiment, they also check its validity and required number of samples with Intent Parser. Finally, when when the experiment design is validated and all parties are satisfied, the users can request that the experiment be executed.

The generality of the approach is demonstrated by the breadth of usage in this program: during a four month period, 19 SD2 users from various organizations generated 34 experimental requests in multiple different areas of investigation, resulting in a total of 16,876 experimental samples executed using three different protocols and collecting data from a variety of instruments. Because these experiments are generated systematically with grounded definitions, meta-data assignment and analysis has been greatly simplified and accelerated, helping enable faster analysis and more effective sharing of data and analyses across the SD2 program.

4 CONTRIBUTIONS

Intent Parser provides a user-friendly process for describing experiments, grounding narrative design descriptions in links to a SynBioHub repository, and generating and validating specifications for wet-lab experiments. The positive experiences of users in the SD2 program suggest this approach has value, and should continue to be elaborated. Potential future directions include improving integration and UI, increasing scope of descriptions, and using natural language processing to extract additional semantic content from prose.

This tool is actively developed at SD2E's GitHub repository https://github.com/SD2E/experimental-intent-parser.

REFERENCES

- BEAL, J., ET AL. Collaborative terminology: SBOL project dictionary. Submitted to 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] MCLAUGHLIN, J. A., ET AL. SynBioHub: A standards-enabled design repository for synthetic biology. ACS Synthetic Biology 7, 2 (2018), 682–688. PMID: 29316788.
- [4] WINOGRAD, T. Procedures as a representation for data in a computer program for understanding natural language. Tech. rep., Massachusetts Institute of Technology, Project MAC, 1971.