ObjTables: structured spreadsheets that promote data quality, reuse, and integration

Jonathan R. Karr[1,2*], Wolfram Liebermeister[3], Arthur P. Goldberg[1,2], John A. P. Sekar[1,2] & Bilal Shaikh[1,2]

August 6, 2020

[1]Icahn Institute for Data Science and Genomic Technology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA.

[2]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA.

[3]Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France.

*Corresponding author: Jonathan Karr (karr@mssm.edu)

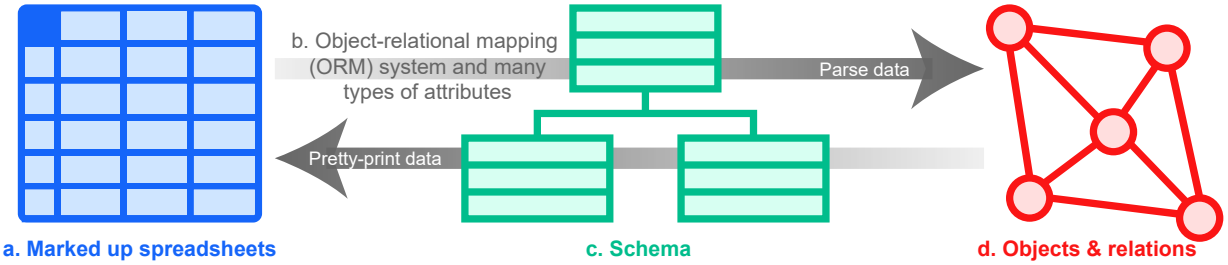
Many scientific problems require integrating multiple types and sources of data.[1] As one of the most common media for scientific data, supplementary spreadsheets associated with articles are a key resource.[2] Spreadsheets are popular because they are easy for authors to write with software such as Excel, LibreOffice, and Google Sheets and easy for journals to archive.

However, spreadsheets are often difficult for other investigators to reuse.[2–4] First, spreadsheets frequently capture information ad hoc because spreadsheets only support a few data types, and there are no universal conventions for encoding multi-dimensional data and metadata into spreadsheets. Second, the ad hoc nature of spreadsheets fosters errors.[5,6] Due to their popularity, the limited reusability of spreadsheets hinders a wide range of research. For example, this inhibits meta-analyses of multiple studies, comparative analyses of multiple organisms, and integrative research such as whole-cell modeling.[7]

To facilitate comparative and integrative research, we encourage authors to write their data and metadata into spreadsheets systematically and thoroughly check their spreadsheets for errors. Several fields of science have begun to develop standards for rigorously encoding domain-specific data into spreadsheets such as ISA-Tab for experimental studies[8] (Section 4 of the Supplementary Information). However, it remains challenging to reuse most scientific spreadsheets.

We developed the *ObjTables* toolkit (https://objtables.org) to both help authors create high-quality spreadsheets and help other researchers reuse them. *ObjTables* facilitates data quality control and reuse by combining spreadsheets (Fig. 1a) with rigorous schemas (Fig. 1c) that describe types of objects represented by a spreadsheet and software tools for using schemas to systematically error check, compare, and compose data and translate spreadsheets to computational objects suitable for analysis with tools such as Python (Fig. 1b,d).

Using a dataset of two genes and their splice variants as an example, Fig. 1 illustrates how *ObjTables* assists researchers. First, authors use programs such as Excel to write data (e.g., genes and splice variants) to one or more worksheets (Fig. 1f,g). *ObjTables* supports the Office Open Spreadsheet XML (XLSX) format,[9,10] as well as sets of comma- and tab-separated values (CSV, TSV) files. Second, authors define an additional schema worksheet that describes the types of objects in their worksheets (e.g., genes and splice variants) and their attributes (e.g., id, location) and relationships (variants of each gene; Fig. 1h). *ObjTables* supports numerous types of attributes for



e. '!!_Table of contents' worksheet

!!!ObjTables objTablesVersion='1.0.0' author='John Doe' date='2020-05-01'		
!!ObjTables type='TableOfContents'		
!Worksheet	**!Description**	**!Objects**
Genes	Genes in the genome	2
Transcript variants	Splice variants expressed from the genome	4

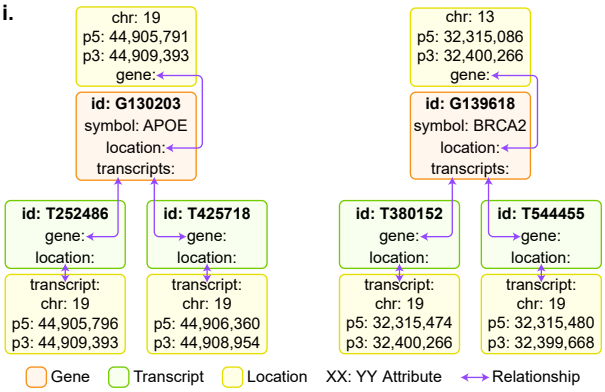
f. '!!Genes' worksheet

!!ObjTables type='Data' class='Gene'				
			!Location	
!Id	**!Symbol**	**!Chromosome**	**!5'**	**!3'**
ENSG00000130203	APOE	19	44,905,791	44,909,393
ENSG00000139618	BRCA2	13	32,315,086	32,400,266

g. '!!Transcript variants' worksheet

!!ObjTables type='Data' class='Transcript'				
			!Location	
!Id	**!Gene**	**!Chr...**	**!5'**	**!3'**
ENST00000252486.9	ENSG00000130203	19	44,905,796	44,909,393
ENST00000425718.1	ENSG00000130203	19	44,906,360	44,908,954
ENST00000380152.7	ENSG00000139618	13	32,315,474	32,400,266
ENST00000544455.5	ENSG00000139618	13	32,315,480	32,399,668

i.



h. '!!_Schema' worksheet

!!ObjTables type='Schema'				
!Name	**!Type**	**!Parent**	**!Format**	**!Verbose name**
Gene	**Class**		**row**	**Gene**
id	Attribute	Gene	String(primary=True, unique=True)	Id
symbol	Attribute	Gene	String	Symbol
location	Attribute	Gene	OneToOne('Location', related_name='genes')	Location
Transcript	**Class**		**row**	**Transcript**
id	Attribute	Transcript	String(primary=True, unique=True)	Id
gene	Attribute	Transcript	ManyToOne('Gene', related_name='transcripts')	Gene
location	Attribute	Transcript	OneToOne('Location', related_name='transcripts')	Location
Location	**Class**		**multiple_cells**	**Location**
chromosome	Attribute	Location	String	Chromosome
five_prime	Attribute	Location	PositiveInteger(primary=True, unique=True)	5'
three_prime	Attribute	Location	PositiveInteger	3'

Fig. 1. Overview of how *ObjTables* helps authors create high-quality spreadsheets and how *ObjTables* helps other investigators reuse them. This figure uses a dataset of human genes and their splice variants as an example. First, authors use programs such as Excel to write their data (e.g., genes and splice variants) to worksheets (**a**, **f**, **g**). Second, authors define an additional schema worksheet that describes the types of objects in their spreadsheets (e.g., genes, splice variants), their attributes (e.g., id, location), and their relationships (variants of each gene; **c**, **h**). Third, authors use markup syntax (e.g., '!!Gene', '!Id', **a**, **f**, **g**) to indicate the types of data in each worksheet and column. Fourth, authors use the *ObjTables* software to error check their data (**d**). Finally, other investigators use the *ObjTables* software to compare and compose spreadsheets and map spreadsheets to high-level data structures (**d**, **i**).

scientific information such as mathematical expressions and chemical equations. Third, authors use markup syntax that begins with exclamation points to indicate the type of object and attribute in each worksheet and column (e.g., '!!Gene', '!Id'; Fig. 1f,g). Fourth, authors use the *ObjTables* software to error check their data. Finally, other researchers use the *ObjTables* software to systematically compare and compose spreadsheets, as well as translate spreadsheets to high-level data structures (Fig. 1i). The *ObjTables* software is available as a web application, command-line program, web service, and Python package.

To make data readable by people and machines, *ObjTables* supports common spreadsheet layouts such as grouped columns with bi-level headings; grammars for encoding information such as reaction equations into strings in cells; and transposed worksheets that encode records into columns rather than rows. The *ObjTables* software can also pretty-print spreadsheets by highlighting column headings, embedding descriptions of each column into notes on their headings, and creating a table of contents worksheet (Fig. 1e). To help researchers develop spreadsheets collaboratively, *ObjTables* also provides tools for tracking versions of schemas and migrating spreadsheets to new versions of their schemas. *ObjTables* can also export datasets to JSON and YAML.

ObjTables is openly available at https://objtables.org. Section 1 of the Supplementary Information and the *ObjTables* website contain examples, a tutorial, and documentation.

Going forward, we aim to make *ObjTables* more powerful and easier to use by developing attributes for additional data types, supporting additional layouts, developing libraries for additional programming languages, developing a repository for schemas, and developing a graphical user interface (Section 5 of the Supplementary Information).

Realizing the full potential of *ObjTables* as a platform for data reuse will require community adoption. To encourage community participation, *ObjTables* is an open project, and we invite researchers to try *ObjTables*, request help via GitHub issues, provide input via pull requests, or join the *ObjTables* team. Long-term, we aim to push the community to publish reusable spreadsheets by lobbying journals and data repositories to require reusable formats such as *ObjTables*.

While *ObjTables* requires additional effort from authors, we believe that the benefits to other investigators of higher-quality and more reusable data are worth the effort. We also believe that authors who use *ObjTables* will be rewarded with more citations. Due to the popularity of spreadsheets, we believe that *ObjTables* can increase the reusability of a substantial fraction of new scientific data. Once adoption of *ObjTables* reaches a critical mass within a field, *ObjTables* could also enable unprecedented secondary analyses such as meta-analyses of data reported by multiple investigators, comparative analyses of multiple organisms, and multi-dimensional analyses of large systems. For example, we are using *ObjTables* to merge kinetic and thermodynamic information about *Escherichia coli* into a more predictive genome-scale model of its metabolism (Section 3.1 of the Supplementary Information).

By making it easier to develop new formats for new types of data, we believe that *ObjTables* can also accelerate emerging fields of science. For example, we have used *ObjTables* to develop WC-Lang, a format for whole-cell models that has become an essential tool for collaboratively developing models (Section 3.2 of the Supplementary Information).

Availability

ObjTables is freely and openly available under the MIT license. The web application is available at https://objtables.org/app. The web service is available at https://objtables.org/api. The command-line program and Python package are available from PyPI at https://pypi.org/project/obj-tables. The source code is available at https://github.com/KarrLab/obj_tables. Examples, tutorials, and documentation are available at https://objtables.org.

Acknowledgements

We thank Yin Hoon Chew, Paul Lang, Timo Lubitz, Elad Noor, and the Center for Reproducible Biomedical Modeling for thoughtful feedback. This work was supported by National Institutes of Health grants R35GM119771 and P41EB023912 and National Science Foundation grant 1649014 to J.R.K, German Research Foundation grant LI 1676/2-2 to W.L, and the Icahn Institute of Data Science and Genomic Technology.

Author contributions

J.K., W.L. and A.G. conceived of the project. J.K. and W.L. designed the formats. J.K., A.G., J.S., and B.S. implemented the software. J.K. and W.L. developed the case studies. J.K. wrote the manuscript. All of the authors contributed to and approved this manuscript.

Competing interests

The authors declare no competing interests.

Supplementary information

Supplementary information
Descriptions of the *ObjTables* format for schemas, markup syntax for spreadsheets, attributes, and software for validating, comparing, and composing spreadsheets; case studies of using *ObjTables* to represent, quality control, and compose metabolomic data and develop a format for whole-cell models; comparison of *ObjTables* with other tools; and future directions for the development of *ObjTables*.

Supplementary dataset 1
XLSX version of the spreadsheet and schema in Fig. 1e–h.

Supplementary dataset 2
CSV version of the spreadsheet and schema in Fig. 1e–h.

Supplementary dataset 3
TSV version of the spreadsheet and schema in Fig. 1e–h.

References

1. Gurevitch, J., Koricheva, J., Nakagawa, S. & Stewart, G. Meta-analysis and the science of research synthesis. *Nature* **555**, 175–182 (2018).

2. Greenbaum, D., Rozowsky, J., Stodden, V. & Gerstein, M. Structuring supplemental materials in support of reproducibility. *Genome Biol.* **18**, 64 (2017).

3. Pop, M. & Salzberg, S. L. Use and mis-use of supplementary material in science publications. *BMC Bioinformatics* **16**, 237 (2015).

4. Cheifet, B. Making supplemental information more accessible. *Genome Biol.* **19** (2018).

5. Powell, S. G., Baker, K. R. & Lawson, B. A critical review of the literature on spreadsheet errors. *Decis. Support Syst.* **46**, 128–138 (2008).

6. Ziemann, M., Eren, Y. & El-Osta, A. Gene name errors are widespread in the scientific literature. *Genome Biol.* **17**, 177 (2016).

7. Goldberg, A. P. *et al.* Emerging whole-cell modeling principles and methods. *Curr. Opin. Biotechnol.* **51**, 97–102 (2018).

8. Sansone, S.-A. *et al.* The first RSBI (ISA-TAB) workshop: "can a simple format work for complex studies?". *OMICS* **12**, 143–149 (2008).

9. Ecma International. Standard ECMA-376: Office Open XML file formats. https://www.ecma-international.org/publications/standards/Ecma-376.htm (2016).

10. International Organization for Standardization. ISO/IEC 29500-1:2016: Information technology – Document description and processing languages — Office Open XML file formats. https://www.iso.org/standard/71691.html (2016).