Sharing intermediate datasets from systematic reviews: connecting the long tail of data

This manuscript (permalink) was automatically generated from lubianat/curation@e68de81 on June 1, 2021.

Authors

- **Tiago Lubiana**
  0000-0003-2473-2313 · lubianat · lubianat
  School of Pharmaceutical Sciences, University of São Paulo; Ronin Institute · Funded by Grant #2019/26284-1 from the São Paulo Research Foundation (FAPESP).

- **Olavo Bohrer Amaral**
  0000-0002-4299-8978
  Institute of Medical Biochemistry Leopoldo De Meis, Federal University of Rio de Janeiro

- **Kleber Neves**
  0000-0001-9519-4909 · KleberNeves
  Institute of Medical Biochemistry Leopoldo De Meis, Federal University of Rio de Janeiro

Abstract

To conduct a systematic review, researches spend hours of work on extracting data from and about the individual papers that are screened for the synthesis. This step often leads to the production of spreadsheets of structured data and metadata about scientific publications, which are the basis for later steps in the project. These datasets will often remain unavailable, hidden in some local folder, despite the time invested to produce them and the potential value it has for the scientific community. We argue that, with a little planning, making these intermediate datasets available to the public is a simple but valuable step for the scientific community, which also allows the recognition of the work done.

Draft

Whenever we systematically search and record information from the literature, we generate curation data. Curation data comes in many flavors: a group of gene expression datasets related to a given disease, a set of articles united by the same methodology, a list of entities (species, drugs, microRNAs, etc.) and extracted information about them. In systematic reviews, the systematicity is unavoidable - the data in these tables are used when articles are considered for inclusion. However, the data collection achieved after many hours of work are almost always relegated to forgotten folders or lost supplementary tables.

While we often ignore the value of curation as data, curation tables are the distilled products of several days of highly specialized work. They can serve as evidence for organizing claims, both in our minds and on knowledge bases; thus, they count as "data" [1]. Data sharing of raw and processed data is maturing as a core part of science [2], and curation data deserves to be included in the movement.

Connecting the "long tail" of small datasets is a valuable prospect that extends beyond raw and processed data, too [3]. Some large biocuration projects, such as UniProt and Gene Ontology, have became core tools of modern life sciences. Systematic curations, albeit smaller, are high-quality structured perspectives of experts and can complement the ecosystem of bio-knowledge bases. With the growth of collaborative knowledge graphs, like Wikidata [4] and BIO2RDF [5], small curations become first-class citizens in the knowledge exchange framework, adding value for users and recognition for curators.

Take an example from our own work on what we mean by "curation data". One of the authors works in a large project whose goal is to perform direct replications of biomedical experiments published by Brazilian scientists [6]. To select the experiments to be replicated, a large systematic review of the last 20 years of biomedical literature was conducted, with a focus on finding articles which made use of specific experimental methods. As a result, we obtained information about dozens of articles - about the methods used in the articles (e.g. RT-PCR or cell viability assay) and experimental models (e.g. cell line or species) -, information that was extracted and checked by two separate biomedical researchers. In the course of the project, we used these data to select the sample of experiments to be replicated in the end - that was the goal of this data collection - and the information regarding the experiments that were replicated will be made public when the project is finished. However, despite its agregate value, most of the data obtained in this step remained hidden in some abandoned folder.

These are dozens of hours of specialized work, with a concrete result - the intermediate dataset containing the metadata about the articles - which remained unusable. We believe that the tools exist to overcome this with little effort and make these small intermediate datasets usable, unlocking their potential. We advocate for sharing small (or not so small) curated datasets made by every researcher. We argue that this would benefit both the researchers' careers and, collectively, provide a new core knowledge resource for life scientists.

First of all, your work has value: other researchers will love to see your curation, even if it is not perfect. It could save them many hours of work. Searching and harmonizing many small datasets that are relevant to your research is no small feat. For instance, in comparative neuroanatomy, a researcher has published a [tutorial] (<https://dieterlukas.github.io/data.html>) enumerating the multiple steps one might have to follow to find and gather datasets.

Second, your work will be findable by anyone who wants to work with the data. For instance, Google Datasets makes it very easy to find the work (see Box). Others will find it, and you will find it. It makes science a more communal and shared endeavor, it helps bring in scientists who otherwise could not

participate [7]. Also, most likely, it is you (or someone in your research group) that will be the one who will try to use the dataset again in the future. Say, in 5 years time, when you need it, you will be able to find your curated dataset on Google, instead of spending hours sifting through old e-mails, Google Drive, Dropbox, or even hard drives. You are your most likely future collaborator.

You will also be rewarded in other ways. A table in Zenodo with a DOI is citable, which means you get recognition for your valuable work. Empirical research also suggests that publications whose datasets are open are cited more often [8]. Also, openness is increasingly recognized as an essential aspect of scientific work that should be recognized and rewarded [9]. Publishers and funders recognize the importance of open data and are moving in that direction with their policies [10].

Lastly, if you connect your data to Wikidata (the sister project of Wikipedia for data), the community benefits from an integrated knowledge graph. For instance, it enables powerful queries via the SPARQL query system. It enables the use of the Scholia platform (https://scholia.toolforge.org) to visualize the topics you have curated. It makes it visible for everyone to improve academic search engines. Although Wikidata is not yet in widespread use for academic purposes [11], it has a lot of potential for research - especially for biocuration and organized reviews [4]. Wikidata is a gateway to fancy ways to integrate knowledge - like structured reviews [12] and 5-star linked open data (https://5stardata.info/en/) - that is acessible without coding skills, and with tutorials in tens of languages (https://www.wikidata.org/wiki/Wikidata:Introduction).

Planning makes all the difference to our argument here. Publicizing these intermediate datasets is a lot of value for little work - but it's little work only if there is a plan. There is a trade-off between the time spent planning the data collection and the time it takes to publicize the datasets at the end of the process. Coming back to our example, the referred intermediate table is now available (https://zenodo.org/record/4737506). However, we did not plan for public release of those data, it was an afterthought. The consequence is that it took many hours of work to gather the data from various files and spreadsheets and then to double check and standardize the information. Much of this work could have been avoided had we followed an agreed standard format and had a data management and preservation plan established from the start.

Thousands of systematic reviews are published each year [13], and a large part of the curation performed in the course of these reviews remains invisible, unusable, and unrecognized. Systematic reviews could be designed from the start to have their curated datasets in a format that makes it easy to find and use them later. We argued that this is an excellent opportunity for open science, where there is a lot to be gained from a small additional effort. This large amount of hidden small curated datasets, if distributed and combined, could have a massive impact on the flow of scientific information in the life sciences.

How to connect your curation (BOX)

Our proposal tries to balance the cost of making these intermediate datasets available with the value for the researchers who performed the data curation and for the scientific community at large [14]. Hence our focus on systematic reviews, where the data curation is already a necessary step and the data provenance - sources used and methods of curation - will likely be already documented in the published article. Because the structured dataset is a side effect of the research project, the only added step is making the dataset available - researchers likely already have a structured table (say, in CSV or Excel format) or it can be easily exported from other tools, like database management systems (such as Microsoft Access or Libre Office Base). In any case, the same steps could be followed for other structured datasets, not only those coming from systematic reviews. While we give general pointers here, we develop more detailed step-by-step tutorials, which can be found here: **LINK**.

1. Make it findable: nowadays, there are many possibilities for making datasets available. We recommend Zenodo(https://zenodo.org/), which will make the dataset citable, with its own DOI. As we mentioned above, the curation of these datasets is specialized work, it is a contribution to the scientific community. Being citable makes it easy for this work to be recognized. Another benefit is that Zenodo datasets are automatically indexed by data-specific search engines, such as Google Datasets, which makes it findable by interested researchers.

2. Make it reusable: Describe datasets with enough information to facilitate reuse. A simple solution is to have a data dictionary available together with the dataset. A data dictionary is essentially a thorough description of what is contained in each column of the dataset. On a similar note, whenever possible, use unique identifiers (e.g., identify scientific articles by their DOI, instead of a full citation). [15]

3. Make it interoperable: if you feel comfortable, an extra-step is to add the data to open knowledge repositories, such as Wikidata (https://wikidata.org), while referencing your publicly available table. This makes your curation available in a standard format, and immediately integrates it with data from many other sources.

Author Contribution Statement

TL, OBA and KN conceived the idea. TL and KN wrote the initial draft. TL, OBA and KN made critical revisions to the manuscript.

References

1. **What Counts as Scientific Data? A Relational Framework.**
   Sabina Leonelli
   Philosophy of Science (2015-12-01) https://www.wikidata.org/wiki/Q31044579
   DOI: 10.1086/684083

2. **Sharing biological data: why, when, and how**
   Samantha L. Wilson, Gregory P. Way, Wout Bittremieux, Jean-Paul Armache, Melissa A. Haendel, Michael M. Hoffman
   FEBS Letters (2021-04-01) https://www.wikidata.org/wiki/Q106498418
   DOI: 10.1002/1873-3468.14067

3. **Big data from small data: data-sharing in the “long tail” of neuroscience**
   Adam R. Ferguson, Jessica L. Nielson, Melissa H. Cragin, Anita Bandrowski, Maryann E. Martone
   Nature Neuroscience (2014-11-01) https://www.wikidata.org/wiki/Q24790499
   DOI: 10.1038/nn.3838

4. **Wikidata as a knowledge graph for the life sciences**
   Andra Waagmeester, Gregory Stupp, Sebastian Burgstaller-Muehlbacher, Benjamin M. Good, Malachi Griffith, Obi Griffith, Kristina Hanspers, Henning Hermjakob, Toby Hudson, Kevin Hybiske, … Andrew I. Su
   eLife (2020-03-17) https://www.wikidata.org/wiki/Q87830400
   DOI: 10.7554/elife.52614

5. **Bio2RDF: Towards a mashup to build bioinformatics knowledge systems**
   François Belleau, Marc-Alexandre Nolin, Nicole Tourigny, Philippe Rigault, Jean Morissette
   Journal of Biomedical Informatics (2008-10-01) https://www.wikidata.org/wiki/Q27921271
   DOI: 10.1016/j.jbi.2008.03.004

6. **The Brazilian Reproducibility Initiative**
   Ana P. Wasilewska-Sampaio, Olavo Bohrer Amaral, Kleber Neves, Ana P. Wasilewska-Sampaio, Clarissa F. D. Carneiro, Olavo Bohrer Amaral, Clarissa F. D. Carneiro
   eLife (2019-02-05) https://www.wikidata.org/wiki/Q61799268
   DOI: 10.7554/elife.41602

7. **Improving data access democratizes and diversifies science**
   Abhishek Nagaraj, Esther Shears, Mathijs de Vaan
   Proceedings of the National Academy of Sciences of the United States of America (2020-09-08)
   https://www.wikidata.org/wiki/Q99233710
   DOI: 10.1073/pnas.2001682117

8. **The citation advantage of linking publications to research data**
   Giovanni Colavizza, Iain Hrynaszkiewicz, Isla Staden, Kirstie J. Whitaker, Barbara McGillivray
   PLOS ONE (2020-04-22) https://www.wikidata.org/wiki/Q93150448
   DOI: 10.1371/journal.pone.0230416

9. **Assessing scientists for hiring, promotion, and tenure.**
   David Moher, Florian Naudet, Ioana A. Cristea, Frank Miedema, John P. A. Ioannidis, Steven N. Goodman

PLOS Biology (2018-03-29) https://www.wikidata.org/wiki/Q52622119
DOI: 10.1371/journal.pbio.2004089

10. **A data citation roadmap for scientific publishers**
    Helena Cousijn, Amye Kenall, Emma Ganley, Melissa Harrison, David Kernohan, Thomas Lemberger, Fiona Murphy, Patrick Polischuk, Simone Taylor, Maryann E. Martone, Timothy W. Clark
    Scientific Data (2018-11-20) https://www.wikidata.org/wiki/Q59134700
    DOI: 10.1038/sdata.2018.259

11. **A systematic literature review on Wikidata**
    Marçal Mora-Cantallops, Salvador Sánchez-Alonso, Elena García-Barriocanal
    Data Technologies and Applications (2019-08-20) https://www.wikidata.org/wiki/Q66724305
    DOI: 10.1108/dta-12-2018-0110

12. **Structured reviews for data and knowledge-driven research**
    Núria Queralt Rosinach, Gregory Stupp, Tong Shu Li, Michael Mayers, Maureen E. Hoatlin, Matthew Might, Benjamin M. Good, Andrew I. Su
    Database (2020-01-01) https://www.wikidata.org/wiki/Q91866899
    DOI: 10.1093/database/baaa015

13. **Are systematic reviews and meta-analyses still useful research? We are not sure.**
    Morten Hylander Møller, John P. A. Ioannidis, Michael Darmon
    Intensive Care Medicine (2018-04-16) https://www.wikidata.org/wiki/Q52584125
    DOI: 10.1007/s00134-017-5039-y

14. **When are researchers willing to share their data? - Impacts of values and uncertainty on open data in academia**
    Stefan Stieglitz, Konstantin Wilms, Milad Mirbabaie, Lennart Hofeditz, Bela Brenger, Ania López, Stephanie Rehwald
    PLOS ONE (2020-07-01) https://www.wikidata.org/wiki/Q96836757
    DOI: 10.1371/journal.pone.0234172

15. **Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data**
    Julie A. McMurry, Nick Juty, Niklas Blomberg, Tony Burdett, Tom Conlin, Nathalie Conte, Melanie Courtot, John Deck, Michel Dumontier, Donal K. Fellows, … Helen Parkinson
    PLOS Biology (2017-06-29) https://www.wikidata.org/wiki/Q33037209
    DOI: 10.1371/journal.pbio.2001414