Conecting the biocuration longtrail

This manuscript (permalink) was automatically generated from lubianat/curation@e70ac01 on April 12, 2021.

Authors

- **Kleber Neves**
  0000-0001-9519-4909 · KleberNeves
  Department of Something, University of Whatever · Funded by Grant XXXXXXXX

- **Tiago Lubiana**
  XXXX-XXXX-XXXX-XXXX · lubianat · lubianat
  Department of Something, University of Whatever; Department of Whatever, University of Something · Funded by Grant #2019/26284-1 from the São Paulo Research Foundation (FAPESP).

Abstract

Draft

When starting a new line of work, researchers search for sources of information to become familiar with the relevant knowledge. Online search has fundamentally taken the place of book digging on libraries: the curation of the literature, which once included physical cards and folders [1] is carried today in various ways, which are peculiar for every researcher. While this is somewhat true for all research projects, it is made explicit for narrative and systematic reviews.

Even though each person has their organizational quirks, a common feature of these curation processes is the generation of intermediate products containing organized, tabulated information. From basic topic-tagging to the extraction of various types of information, these intermediate products provide the basis for the research process: they influence the mental models, serve as raw material for planning experiments, and are indispensable when writing introductions and discussions. In systematic reviews, this is an unavoidable step - the data in these tables are used when articles are considered for inclusion. It is often the case that information is collected for many articles, even if only a tiny fraction of the studies end up being included. However, these data, produced after many hours of focused, specialized work, are almost always relegated to forgotten folders.

These intermediate tables are a rich source of organized information about the published literature. They are the distilled product of several days of highly-specialized work. Moreover, they are already organized (the scientists involved needed to organize it for their projects!), which means they could easily be made available to the community.

Most scientists are already producing these small datasets - most of which, currently, are not made public. When pooled together, this might amount to enormous quantities of high-quality data curated by people with the relevant local-expertise [2] [2]. Massive biocuration projects, such as UniProt and Gene Ontology, are core tools of modern life sciences and cover a fraction of any researcher's topics of interest. There is potential for open-knowledge tools such as WikiData to incorporate all of these small datasets in an integrated manner and eventually play a similar role for more topics.

Take an example from our own work on what we mean by "intermediate tables". One of the authors works in a large project whose goal is to perform direct replications of biomedical experiments published by Brazilian scientists [3]. 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, most of the data obtained in this step remained hidden in some abandoned folder.

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 [4]. 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 [5]. Also, openness is increasingly recognized as an essential aspect of scientific work that should be recognized and rewarded [6]. Publishers and funders recognize the importance of open data and are moving in that direction with their policies [1].

Lastly, if you connect your data to Wikidata, 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 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 [7], it has a lot of potential for research - especially for biocuration and organized reviews [8]. Wikidata is a gateway to fancy ways to integrate knowledge - like structured reviews [9] - that is acessible without coding skills, and with tutorials in tens of languages (https://www.wikidata.org/wiki/Wikidata:Introduction).

Thousands of systematic reviews are published each year [10], and a large part, if not most, 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 share 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 combined and distributed, could make a massive impact on the flow of scientific information in the life sciences.

How to connect your curation (BOX)

`Tentei colocar essa parte no frame do FAIR, mas não sei se usar DOIs tá em reusable ou interoperable. De qualquer forma, acho que isso aqui devia ser uma caixinha no artigo, sabe? Separado do texto principal, como um "passo a passo sugerido pra fazer isso com os seus datasets". Por isso eu trouxe essa seção pro final e adicionei uns trechos redundantes com o "Why", caso alguém resolva ler só a caixinha.`

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 [11]. 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). 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, 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).

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

References

1. **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

2. **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

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

4. **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

5. **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

6. **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

7. **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

8. **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

9. **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

10. **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

11. **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