An adventure in open research data remix, licensing and provenance: case study of the Open Access Article Processing Charges (OA APC) Longitudinal Study

Proposed presentation for the RDA 10th Plenary, Montréal, Sept. 20, 2017

Presenter: Heather Morrison, Associate Professor, University of Ottawa School of Information Studies & Principal Investigator of the *Sustaining the Knowledge Commons (SKC)*, funded by Canada's Social Sciences and Humanities Research Council (SSHRC), Insight Development Grant (2014-2016), Insight Grant (2016 - 2017)

Abstract

This presentation will focus on analysis of emerging issues in the creation of a remixed dataset derived from diverse sources. The OA APC dataset (Morrison et al, 2017b) << doi:10.5683/SP/KC2NBV>>, includes data hand-created by the SKC research team, publisher price lists that are captured through screen scrape or from an excel or PDF list, the DOAJ metadata set, and data contributed by other researchers. The dataset is available as open data, and the documentation has been peer-reviewed and published in a new type of open access journal. Specific issues to be covered include licensing (why a decision was made not to license the data) and the importance of understanding provenance to avoid errors in downstream data research.

Licensing

The statement for data use, from Morrison et al. (2017a), states:

> **4. Using These Data (Licensing)**
>
> This dataset is derived from several sources, including the DOAJ metadata (which has its own license terms posted on the DOAJ website), other data screen-scraped from DOAJ, factual data gathered from publisher's websites, 2015 data provided by Walt Crawford, 2010 data provided by Solomon and Björk, and our team's analysis. If you are making use of our dataset as a whole, please cite: Morrison, H.; Brutus, W.; Dumais-DesRosiers, M.; Kakou, T.L.; Laprade, K.; Merhi, S.; Salhab, J.; Volkanova, V. & Wheatley, S. Open access article processing charges longitudinal study 2016 dataset [http://dx.doi.org/10.5683/SP/KC2NBV]. If you are drawing from the other sources, please cite the other sources. There is no license for the dataset as a whole, as individual elements are derived from different sources, which may have their own terms. When posting your own dataset, please include at minimum the journal title and ISSN as these are key matching points for merging together different datasets.

A bit of elaboration may be helpful to explain the decision not to license the dataset. The dataset provided to the SKC team by Solomon and Björk (2012) was never published as open data by the original authors; this was researcher-to-researcher sharing of data. Walt Crawford's (2016) data is open data. The dataset includes a substantial portion of the Elsevier APC price list, all of their fully open access journals, derived from a PDF on the Elsevier web site. Permission to use this list was not sought, on the basis that publishers do not have the right to refuse permission to conduct research on publisher prices, thanks to the willingness of Barschall and the APS to fight for this through the courts in four continents over more than a decade (Lustig, 2001). My right as an academic to conduct research on prices does not mean that I have the right to grant blanket downstream rights for commercial and derivative rights.

The full DOAJ metadata set is included, to allow for correlational studies such as whether a particular characteristic of a journal is correlated with ongoing journal activity, tendency to use APC or amount of the APC for those that do charge. The DOAJ metadata is licensed under the Creative Commons Attribution - Sharealike (CC-BY-SA) license. According to this license, my work as a derivative of the

DOAJ metadata must be released under the same license, which according to my analysis would not be appropriate. For me, this is not a reason to refuse to release the dataset as I am confident that DOAJ was not aiming to create a barrier to the re-use of the data and to date has not objected to this use of the DOAJ metadata. If I did use a CC-BY-SA license on the data, I would in effect be releasing other data such as the Elsevier APCs under this license. These are just a few of a great many data sources, hence this is the tip of the iceberg with respect to the complexity involved.

Provenance

In a re-mixed dataset created for research purposes, the data elements are derived from multiple sources. Simply crunching this data without a full understanding of its provenance could result in substantive errors. For example, it is very useful to compare the APC data gathered by Solomon and Björk (2012) in 2010 with the SKC team's 2016 data; this is the main purpose of this longitudinal study. However, to interpret the data it is essential to understand that Solomon and Björk conducted a random sample of journals listed in DOAJ that were known to charge APCs, and that they estimated a per-article cost where a per-page cost model was used, while in 2016 the SKC data is a full sample of DOAJ journals whether they charge APCs or not, includes some journals not listed in the DOAJ at the time (journals removed from DOAJ during the longitudinal study or included on publisher's price list but not DOAJ), treats per-article and per-page pricing as two separate models, and in some cases uses the APC data provided by Crawford (2016) or DOAJ. Crawford also uses a slightly different data-gathering method, estimating APC cost and converting to USD where the SKC team reports pricing in original currency. A straightforward comparison of 2010 and 2016 data in this dataset without full understanding of what might look like apples and apples but actually is apples and oranges, would have a strong likelihood of arriving at false conclusions. For example, in 2016 a publisher might have had both APC charging and non-charging journals; if this publisher was included in the 2010 sample, all journals would have APCs because of the sampling limitations. One might come to a conclusion that the publisher had changed their business model, without understanding that this is likely an artefact of the difference in sampling methods. To achieve the full potential of open sharing of data, I argue that this is an issue that requires careful attention.

References

Crawford, W. (2016). *Gold Open Access Journals 2011–2016.* Retrieved April 1, 2017 from http://walt.lishost.org/2016/05/gold-open-access-journals-2011-2015-its-here/

Lustig, H. (2001). The APS in an age of litigation. *Physics and Society Newsletter*, 29(1).

Morrison, H.; Brutus, W.; Dumais-Desrosiers, M.; Kakou, Tanoh L.; Laprade, K.; Merhi, S.; Ouerghi, A.; Salhab, J.; Volkanova, V.; Wheatley, S. (2017a). Open Access Article Processing Charges (OA APC) Longitudinal Study 2016 Dataset. *Data* 2, no. 2: 13.

Morrison, H., ; Brutus, W.; Dumais-Desrosier, M.; Laprade, K.; Merhi, S.; Ouerghi, A.; Salhab, J.; Volkanova, V.; Wheatley, S., (2017b), Open access article processing charges 2016. doi:10.5683/SP/KC2NBV, *Scholars Portal Dataverse*, V3.

Solomon, D.J.; Björk, B.C. (2012) A study of open access journals using article processing charges. *J. Am. Soc. Inf. Sci. Technol. 63*, 1485–1495.