The Data Ark: An attempt to retrieve, preserve, and liberate important scientific data

2018-07-31

### Brief summary

*We’ve just published a paper reporting our efforts to retrieve raw data from 111 highly-cited psychology and psychiatry papers and make them available to the scientific community in a repository called the Data Ark. This post is intended to complement the paper and summarise the key ideas behind the Data Ark and the findings of our initial attempts to populate it. We are hoping to start a conversation about data stewardship, so let us know your thoughts in the comments below.*

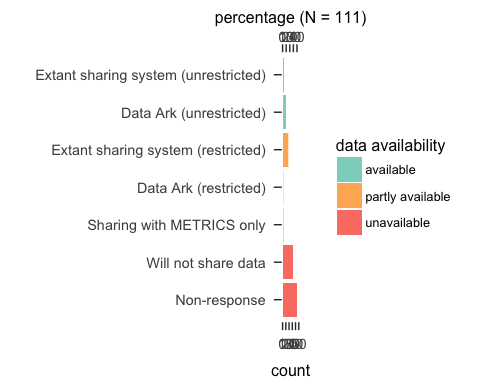
The Ptolemies of Egypt were seized by bibliomania. They established the Great Library of Alexandria with the lofty ambition to gather and preserve all the world’s knowledge under one roof. Agents of the king were dispatched to the bustling marketplaces of Athens, Syracuse, and Rhodes in search of valuable papyri. When ships sailed into Alexandria harbour they were searched and any books on board were confiscated[[1]](#footnote-22). Eventually, a large collection of influential ancient works was amassed, covering all manner of subjects from mathematics to poetry, penned by such luminaries as Aristotle and Euripides. And then, torched by fire, jolted by earthquakes, and ravaged by warfare, the Great Library crumbled into ruin, and innumerable ancient works were lost[[2]](#footnote-23).

The demise of the Great Library of Alexandria.

The demise of the Great Library of Alexandria.

Today, the demise of the Great Library of Alexandria is often regarded as a cautionary tale: a catastrophic loss of knowledge that we would not wish to see repeated. And yet, the gradual attrition of the scholarly record in the modern era is widely tolerated. The scientific literature consists of articles reporting findings and making claims, but it is rare to see the research data that substantiates them[[3]](#footnote-25). After articles have been published, the fragile system that temporarily kept the data alive starts to atrophy: hard drives fail, e-mail addresses stop functioning, colleagues move on, and before long, the data is practically no longer retrievable[[4]](#footnote-28). Requesting data directly from authors is typically not successful[[5]](#footnote-30). There was no fire, or earthquake, or war to draw our attention to this state of affairs, but it could be that most of the data generated by humanity’s previous scientific endeavors has now been irrecoverably lost.

What can be done? Many fantastic efforts are underway to increase data availability for future studies. For example, we [recently found](https://dx.doi.org/10.31222/osf.io/39cfb), that a mandatory open data policy introduced at the journal *Cognition* was highly effective at increasing access to data. But these efforts will not address the fact that data remains unavailable for the majority of the extant published literature. This was the motivation for our recent effort to create the Data Ark: an open online repository where data retrieved from influential scientific studies could be preserved and liberated for verification and re-use by the research community. We began by contacting the authors of 111 of the most highly-cited articles published in psychology and psychiatry between 2006 and 2016, and asked if they would be willing to share the corresponding raw data publicly in the Data Ark. We also provided the option of suggesting that certain access restrictions be put in place, for example in the case of sensitive data. If authors were unwilling to share, we asked for the reason(s) why so we could better understand barriers to data sharing.



Responses to our data sharing request in an effort to populate the Data Ark. You can see a breakdown of responses based on field (psychology vs. psychiatry) and time period (2006-2011 vs. 2014-2016) in the paper.

The main findings are shown in Figure @ref(fig:data-ark-fig). As you can see, we were only able to retrieve 10 out of the 111 target data sets for unrestricted sharing in the Data Ark (you can [find them here](https://osf.io/view/DataArk/)). Note that in 22 cases data were already being shared via some existing system, but only 5 of these allowed unrestricted access. Of course data sharing is not always straightforward and can be constrained by legal or ethical obligations. We discuss additional findings related to these issues in our paper, such as the access restrictions for existing sharing schemes and the reasons provided by authors who responded that they would not share. However, in the many cases where we received no response, it was impossible to learn about these potential barriers, and for the scientific community to judge if there are appropriate reasons not to share. These data are essentially lost.

Our initial efforts to populate the Data Ark have highlighted just how challenging it can be to retrieve data from published studies. We had hoped that particularly influential studies would yield a higher success rate. Given the influence of the associated articles, it should be a priority that these data are preserved and liberated to enable independent verification and re-use by the scientific community. It may be that data requests are more likely to be successful in the context of a specific research project and tangible benefits (e.g., authorship on a paper) for original authors are more obvious. For example, data retrieval seems to have been highly successful in [a recent effort to establish an empirical common ground](https://orca.cf.ac.uk/109813/1/Morey.%20Benchmarks%20for%20model.pdf) for the purposes of computational modeling in the domain of short-term and working memory.

Unfortunately, we do not know what substantive scientific projects of the future would bear fruit if only they could be pollinated by the scientific data of today. That is why it is essential to ensure the preservation of important data while we still have the chance to retrieve them. Our findings clearly raise questions about the current norms of data stewardship. Is it appropriate to maintain an ad-hoc system whereby researchers who originally generated data retain responsibility for their preservation and govern their access? Would journals, data repositories, and/or ethics boards be more impartial custodians of data? The modern scholarly record is not under threat from a great fire or flood, but action is needed to ensure that important data does not become lost at sea. Storage in a Data Ark, an independent repository that will continue to float above the murky depths, may enable long-term preservation and accessibility of the evidence that underlies the scientific literature.

We are keen to hear our colleagues thoughts’ on this endeavour. Do you think the Data Ark is a worthwhile initiative and if so, what could be done to improve the retrieval rate?

1. It appears that the books were taken to the Great Library to be copied. The owners were later reimbursed with the copies, whereas the originals were purloined for the library’s collection. Cheeky. [↑](#footnote-ref-22)
2. Naturally there’s considerable uncertainty about historical events when resources are scarce and potentially unreliable. [Delia (1992)](http://dx.doi.org/10.2307/2165947) discusses various accounts of the library’s demise. [↑](#footnote-ref-23)
3. See for example [Alsheikh-Ali et al. 2011](https://dx.doi.org/10.1371/journal.pone.0024357) and [Iqbal et al. 2016](https://dx.doi.org/10.1371/journal.pbio.1002333) [↑](#footnote-ref-25)
4. See [Vines et al. (2014)](https://dx.doi.org/10.1016/j.cub.2013.11.014) [↑](#footnote-ref-28)
5. See for example [Vanpaemel et al. 2015](http://doi.org/10.1525/collabra.13) and [Wicherts et al. 2006](https://doi.org/10.1037/0003-066X.61.7.726) [↑](#footnote-ref-30)