Disease Models & Mechanisms 2, 521-525 (2009) doi:10.1242/dmm.004630
Published by The Company of Biologists 2009

EDITORIAL

Shining a light on dark data

Chris Patil[1] and Vivian Siegel[2]

Increasing the publication of dark data will require us to acknowledge and alter the incentives for multiple agents of action: journals, citation indexes, funding agencies, academic institutions and, not least, the researchers themselves

Dark data

Many scientists spend much of their time doing work that doesn't get published. Put another way: our system of scientific communication implies that we don't consider most of what we do to be worth sharing. The ratio of published text to performed experiments is infuriatingly low.

To the extent that the knowledge isn't disseminated, within the scientific community or the culture as a whole, in practical terms it doesn't exist. The time and money spent to produce such data are essentially wasted.

Wouldn't it be useful, both to the scientific community or the wider world, to increase the publication of negative results? Wouldn't it make more sense to squeeze more (disseminated) knowledge out of the resources we use? In other words, should we not make an effort to increase our society's return on its investment? And, if so, how should we do that?

Before we attempt to answer these questions, we will examine the sort of work that goes into 'dark time', i.e. the time that we spend doing *something* that doesn't map directly onto a figure, text or other element of a publication.

The black hole

The exploration of the unknown is hazardous. We don't know where the science will lead and we sometimes end up walking down blind alleys. These dead ends appear at several distinct conceptual scales. In order of increasing scientific value, they follow this conventional perspective:

Failures: experiments that simply do not yield interpretable results and send us back to the drawing board.

Negative results: experiments that work according to design but fail to support the hypothesis or motivation underlying the work.

Orphan results: experiments that yield positive results but fall beneath the threshold of significance for publication in a form deemed worthy of the effort, e.g. a successful rotation project that another student never picks up.

Abandoned results: perfectly good experiments that don't fit into the 'story' of a study that is ultimately published. In some cases, abandonment is the consequence of complexity: a zealous scientist performing 'one too many experiments' and ending up in a rough part of the Petri dish, now finding themselves unable to elegantly explain the entirety of the data. In that position, it is tempting to simply defer analysis of the unforeseen complexity until 'the next paper'. Another sort of abandonment – removal of data because of journal-imposed limitations on manuscript length – has fallen out of fashion in the era of supplemental materials (see supplementary material S1).

Scooped papers: completed projects that meet the criteria for publication but appear to simply confirm the work of studies that have already been published. These papers invariably include nonredundant information, but because journal editors are obsessively vigilant about rejecting papers that fall below a threshold of 'novelty', these papers become unpublishable in practical terms.

[1]Chris Patil is at the Buck Institute for Age Research, Novato, California, USA
[2]Vivian Siegel is at the Department of Medicine, Vanderbilt University School of Medicine, Nashville, Tennessee, USA
(e-mails: cpatil@buckinstitute.org; dmmeditor@biologists.com)

That giant sucking sound

Now that we have identified the material inside the black hole of dark data, we can begin to ask what forces drew it there. What makes the unpublishable unpublishable?

In the case of abandoned results, we could blame our cultural assumptions about what ought to go in a paper. Scientific papers are not historical records of the scientific process; rather, they are ahistorical texts designed to maximize their chances of acceptance by the editors and reviewers of high-impact journals. Much of their structure, and many decisions about what to include, owe themselves to authors' attempts to second-guess reviewers and editors. 'How should we spin this? How can we package that? This doesn't help the story. That raises more questions and the reviewers will just ask for more experiments.' (If you have never heard sentences like these, please tell us how scientists write papers on your happy planet.)

We are authors one day and reviewers the next. Although we lament the difficulty of packaging our results into a coherent narrative or story, we perpetrate the same abuse on our colleagues when the tables are turned. You don't like how the sausages are made? Then get out of the kitchen – after all, everyone else is doing it and presumably that means you should do it the same way. The system isn't obviously beneficial to everyone, but it is nonetheless self-reinforcing.

Particularly sad is that a 'story' may not reflect how some readers engage the literature. Papers are written for two audiences: the scientists who will use the data for their own research and the newcomers to the field. The former are primarily looking for information presented in the results and methods sections. Such a reader is unlikely to scan through a paper and lament, 'There is clear evidence for a physical interaction between protein A and protein B, which is why I pulled up the paper on PubMed, but I have real trouble with the plot and there are major third act problems, so I can't use it'. By contrast, the newcomers will focus on the introduction and the interpretive discussion. These readers truly benefit from a ripping yarn that keeps them engaged and interested in seeing how the next experiment will turn out. Thus, journal articles are research reports wrapped in literature reviews.

Both types of writing are useful, in different contexts. The standard model for research papers can indeed create nice stories. But what about results that don't support an otherwise coherent narrative?

In these cases, the problem isn't that the data are unpublishable in any journal, but that they are unlikely to be published by journals that boost the reputation of the author. There are journals that are willing to take scooped papers or results of limited significance, but authors might not consider the undertaking to be worth the effort. Time spent publishing small papers is time not spent developing big ones.

Thus, ironically, some data are unpublishable precisely because of our desire to publish – but to do so in a specific manner, in journals of our choosing.

What's the point of it all?

Now that we have discussed a few reasons that results go unpublished, let us ponder whether it would be useful to increase the dissemination of dark data.

For orphaned and abandoned results, as well as the nonredundant parts of scooped papers, the most obvious argument in favor of publishing is that someone might find them useful. Thanks to the limitations on inter-lab transparency and the sheer size of the scientific enterprise, individuals have very little idea whether a given piece of data will be useful, and to whom. If we bury them in our notebooks, as we do now, we will never know. If, on the other hand, we publish the little tidbits that didn't fit into our big pies, readers who are suitably empowered with the appropriate search technology

could find valuable information and interpret it from their own unique perspectives. Publishing small-scale results could be the catalyst for major progress; one person's trash could be someone else's treasure.

In the case of negative results, publication could help prevent duplication of effort. Chances are that a hypothesis that seems worth testing to me will also seem worth testing to my five closest colleagues (and competitors). Knowing that an obvious line of investigation will yield no fruit, a scientist could save months or years of effort by sharing that information.

As for failures: we learn facts and obtain ideas by reading papers, but we learn our craft by performing experiments and watching others work. Sometimes, we learn most from our failures. Disseminating information about failures would help us improve the practice of our trade.

These justifications all invoke some benefit to the scientific community. But science already works well enough to generate an unmanageably large literature – it's not clear whether any improvement is necessary, especially if such an 'improvement' would itself increase the size of the literature.

Beyond that, publishing takes time for the author, and it can be distracting. Is it worth it to spend time on small pieces, which are individually unlikely to make or break a career, when the trade-off is that one would have less time to make headway on larger-scale studies?

Thus, although the arguments in favor of small-unit publishing all seem to revolve around benefits to the community, the costs of generating these small units would fall on individual authors. If the community is to reap the benefits, then the costs to the individual authors must be driven to zero – or associated with some reward.

Defying gravity: how can we provide incentives for small-unit publication?

The simplest incentive for alternative styles of publication would be to co-opt the existing publishing templates: create journals for the publication of negative, orphaned and abandoned results that are barred entry into conventional scientific media. These journals could compete in the same way that 'real' journals do: by impact factor, article-level metrics, etc. If the published pieces are useful, people will cite them and link to them. Authors can put these pieces on their resumes and everyone is happy.

However, this incentive might prove insufficient. Why sully a CV with papers from the 'Journal of Failed Experiments'? Don't we want our colleagues (and especially our competitors) to believe that we succeed at every undertaking? Besides, it could be argued that the very mission of these papers makes them unlikely to be cited: if they stop people from following an unproductive path, there should be no papers written to refer to them. Beyond that, journals often limit the number of citations in a paper; references to papers about what didn't work in the past are likely to be the first to go.

(Journal editors could eliminate this objection by releasing the constraint on reference limits or by creating another kind of searchable acknowledgment where authors comprehensively record all input that informed a body of work. Scientists could record these sorts of contributions as 'community activities' on their CVs; academic institutions and other employers could fuel the system by making such activity an asset in considerations of promotion.)

If the criticisms were entirely true, however, then there would be neither journals willing to publish 'difficult' data nor authors to fill their pages. This turns out not to be the case. Although a comprehensive enumeration of every available journal is beyond the scope of this editorial, here are two examples: the web-only Journal of Negative Results in BioMedicine is 'ready to receive manuscripts on all aspects of

unexpected, controversial, provocative and/or negative results/conclusions in the context of current tenets, providing scientists and physicians with responsible and balanced information to support informed experimental and clinical decisions' (from http://www.jnrbm.com/). New articles are posted almost monthly, and the journal has a measurable (although modest) impact factor. A broader effort is underway at the All Results Journals ('Because all your results are good results'; http://www.arjournals.com/ojs/), which are currently seeking to recover and publish negative or 'failed' studies in chemistry, biology, physics and nanotechnology.

Given that the venues for publication can and do exist, perhaps we might encourage their use by tweaking the current incentive structure. One approach would be to mandate reporting of all data. Funding agencies might require that all results be reported somewhere, as a matter of accounting time and resources spent on research. Journals might insist that researchers log experiments before they are performed, analogous to the clinical trials registry, to ensure that results aren't reported selectively. However, such methods would not be without their own perils: publication requirements would force funding agencies to develop even larger bureaucracies to guarantee compliance. Moreover, it is unlikely that scientists would welcome any 'reform' that causes the routine conduct of research to more closely resemble the bloated and sluggish process of clinical trials.

Instead of the heavy hand of regulation, a lighter touch might be more appropriate, such as reducing the time required to publish more of our data. Accomplishing this goal would require us to rethink the limitations on the structure of research papers. As we pointed out above, modern research papers are burdened by the necessity of reaching two audiences – specialists within the field, who focus on data, and outsiders, who focus on context and interpretation. The specialist benefits most from the increased dissemination of dark data; therefore, why not liberate the research report from its review-like wrapper, perhaps even from the main text altogether? We could strip the paper down to its minimal components: the methods, the data and enough well-chosen keywords to enable the manuscript to come up in response to relevant searches.

The new standard for expedited publishing of dark data could be designed with devoted search technologies in mind. We can imagine software tools that allow a scientist to describe the experiment they wish to perform – in a confidential manner that prevents their plans from falling into the hands of their competitors – and that search the database of these stripped-down research reports, returning papers in which related experiments have already been tried (perhaps unsuccessfully).

The logic of minimizing the effort required to publish data – beyond the initial effort required to do the work in the first place – could be taken one step further. Why not take advantage of the record keeping that we are already doing, i.e. our notebooks? We are increasingly keeping scientific records in electronic form; it would be straightforward to wrap our notebook pages describing an orphan result with a bit of searchable text, generate a web page, and submit the whole thing to a database. The act of conducting research would thus become practically synonymous with the act of disseminating the resulting knowledge. Along the way, we would have to spend some energy improving the records that we keep in order to ensure that our notebooks were more accessible to outside readers and less like the quirky private diaries they often become.

Increasing the publication of dark data will require us to acknowledge and alter the incentives for multiple agents of action: journals, citation indexes, funding agencies, academic institutions and, not least, the researchers themselves. At present, these

stakeholders' individual goals sometimes place them at odds with one another. With aligned incentives, however, the dissemination of knowledge holds the promise of improving the efficiency, and therefore ultimately the efficacy, of the scientific enterprise.

This article is freely accessible online from the date of publication.

SUPPLEMENTARY MATERIAL

Supplementary material for this article is available at http://dmm.biologists.org/lookup/suppl/doi:10.1242/dmm.004630/-/DC1