Information challenges for scientific publishing

Arne Babenhauserheide

August 28, 2015

Scientific publishing has come a long way since it's beginning, and it's principles have allowed it to scale up from a few hundred active scientists worldwide to conferences with tens of thousands of people for a given topic. But in the last few years it hit its limits. It becomes harder each year to keep up with the amount of new papers being published and even scientists from similar fields repeatedly reinvent the same methods. To scale further and to continue to connect the scientific community, it must adapt to make it easier to get an understanding of the current state of science and keep up to date with new findings.

To grow from these challenges, scientific publishing needs to

- make it easier to get and stay up to date with several fields,
- foster reproducible research,
- add incentives for reproduction studies,
- introduce propagating corrections and
- reduce the pressure to publish which intensifies all existing problems.

Contents

1 The Good

Before I start with my critique of scientific publishing, I want to show where it really shines. This will put its shortcomings in the proper perspective and also serve as a reminder about which methods are proven by time. In this part I will focus on the aspects of scientific publishing which help dealing with a huge amount of information.

I will also contrast these aspects to ordinary websites, because these have become the standard information medium for non-scientists, yet they took up technology much faster than scientific publishing, which allowed some non-scientific publications to get on par with scientific publications in many aspects and even surpass them in a few.

1.1 Different levels of content

Scientific publications are expected to have a title, keywords, an abstract, an introduction and conclusions - in addition to any other content they have. This makes it easy for readers to choose how deep they want to delve into the topic of the paper.

- The title and keywords allow readers to decide whether the paper could be important to their own interests.
- The abstract gives a short takehome message: Just reading the abstract allows remembering later that there was a publication which might be useful for the question at hand.
- The introduction gives the necessary information to gain a rough understanding of the paper, even if it's not about ones own speciality.
- The conclusions provide the results of the publication: If you only read the abstract, the introduction and the conclusions, you can already reason about the impact of the research on your own work.

All this taken together creates a medium where every reader can decide how much information he or she wants to ingest. This allows priorizing a specific field while still getting a rough understanding of the larger developments happening in similar topics.

Where websites typically only provide one or two representations of any given topic - often title plus teaser and the main text - scientific publications provide several layers of information which are all useful on their own.

1.2 Referencing other works

While the internet allowed ordinary publications to catch up a lot via hyperlinks (though these are still mostly used by hobby-writers and not so much by big newspapers), scientific publication still holds the gold standard for referencing other works in a robust way.

They include the title, the author, the journal, the date of publication and a link. Even if the journal dies and the DOI system breaks, a paper can still be found in third party databases like university libraries.

In the internet however, links regularly break, even those referenced in court cases. So here the web still has a lot to learn from the tried and true practices of scientific publishing.

(in the meantime, if you're a blogger yourself, please preserve your links (german original))

1.3 Summary

The different levels of information and the robust references create a system which managed to sustain its quality during a growth in the number of researchers and publications by several orders of magnitude.

These two topics aren't the only strengths of scientific publishing (which for example also include the peer review process in which a trusted editor asks people from the same field to provide high-quality feedback), but they are the most important strengths for the topics in which the next part identifies challenges that need to be resolved to preserve the integrity of scientific publications and avoid and reduce the fragmentation of science by keeping researchers connected with current work from other groups.

2 The Challenges

2.1 Core Questions

The gist of the challenge of scientific publishing can be summarized in two questions:

- "What's the expected reading for scientists?"
- "How do you know that you can trust this paper?"

Journals are already trying to tackle both of these, but the current steps fall far short of solving the problem.

2.2 Expected reading for scientists

> Suddenly you realize that there is a group of scientists in Korea who also work in your field.

This actually happened: I shared a paper with experts in the field who did not even know that the group doing the research existed.

The problem behind this experience is that the number of scientists increased more than a hundred fold (at EGU more than 15000 people met, and that's only for earth sciences), but scientific publishing still works similar to how it worked when there were only a few hundred (communicating) scientists worldwide. And the pressure to publish as much as possible intensifies the problem a lot.

In a field like Physics of the Atmosphere, hundreds of papers are published every month. Even the reading list filtered by interest which I get per E-Mail every week contains several tens of papers per journal. And when I started to dive into my research field at the beginning of my PhD, a huge challenge was to get the basic information. It's easy to find very detailed information, but getting the current state of scientific knowledge for a given field takes a lot of effort, especially if you don't start in a group working on the same topic. So how should scientists keep a general knowledge of the broader field, if it's already hard to get into one given field?

The current answers are review papers and books. Good review papers allow understanding a core topic of a given scientific community within a few days. A nice example is Data assimilation: making sense of Earth Observation. A book gives a good overview of a given field, but it requires a hefty time investment. So how do you keep a general understanding of other fields? How can we avoid reinventing the wheel again and again, just in different contexts?

A simple idea to achieve this would be to create a hierarchy of quarterly overviews:

- STEM/MINT and social sciences.
- A broad field (like atmospheric physics).
- A specific subgroup (about 100 scientists).

With every overview including two aspects:

- The state of scientific knowledge.
- Core changes since the last overview.

The core changes would be suggested reading for all scientists in the given field, while the state of scientific knowledge would allow people to get up to speed in a given field, or to understand something interesting, and provide a path to the more detailed reviews and papers.

Assuming that on average 2-3 broad fields and subgroups are interesting to a scientist, this would allow keeping up to date with scientific development by reading **one overview paper per month**, and it would allow getting a broad understanding of many fields by reading **the overview of an additional field every quarter**.

These structured overviews would reconnect science.

To support the creation of the overviews, we might need more dedicated, paid overview writers.

Part of this job is currently done by publications like Annual Reviews, Physik-Journal (german) and Scientific American (in order of decreasing specialization), and awareness of the need to reconnect science could make it possible to extend these and similar to make it easier to acquire and keep a good understanding of the current state of science.

2.3 Trustworthy research

The second big question is: "How do you know that you can trust this paper?" To be able to trust the results shown in any paper, there are two aspects:

1. It must be possible to reproduce the results independently, and
2. The prior assumptions of the research have to be correct.

2.3.1 Reproducible research

The first problem can be tackled by requiring scientists to share the data and programs they analyzed, so others can reproduce the results (plots, table content and so on) with as little effort as possible. Ideally the paper should use something like autotools and org mode (german original) to create a distribution package which allows others to reproduce the paper straight from the data and ensures that the data in the package actually suffices to generate the results. This would ensure that papers provide all the small details which might not seem worthy of publication on their own but can be essential to reproduce the results with a new experimental setup.

Aside from making it possible for others to reproduce your work, this also makes it easy to go back years later answer the question:

- „How exactly did I create the publication?“

The minimal requirements for a system for reproducible research are:

- Create diagrams and tables directly from the data
- Include required data and scripts (as much as allowed)
- Automate creating the publication and checking whether it fulfills the first two requirements

That data and scripts should be under Open Access licenses for this to work should be self-evident. It is about enabling easy reproduction, and that requires building upon the previous work.

Basic reproduction of the results would then be as simple as calling

```
./configure; make
```

An example for such a system are the GNU automake which provides a `make distcheck` command to verify that the released data suffices to create the publication. If you want to give this a try, have a look at Going from a simple Makefile to Autotools.

2.3.2 Incentives

The main challenge for such reproducibility is not technical, however. It is the competition forced upon scientists by the need to apply for external funding. If you release your scripts and data, you cannot monopolize them to apply for followup funding. On the other hand, publishing the scripts and data can help get more visibility and citations. To create incentives for publishing everything used in the research, there also need to be incentives for publishing reproduction studies.

For the publishing scientist, people who use the research provide references. If other scientists in the same field reproduce research locally, that encourages followup research which might reference the original scientist, but it is a game of luck whether other scientists will actually use and reference the published data and scripts or just use it as inspiration. Or just ignore it, because they have to focus on doing work they can publish to make it into the next round of funding. As such the incentive to create research which is easy to reproduce would rise a lot, if reproduction studies could be published more easily, because every reproduction publication would provide a reference. When we want more reproduction of research, skillfull reproduction has to provide value for scientists in its own right.

The focus I put on reproducibility does not mean that errors in publications are widespread. There are some fields with problems – for example research on new medicines, where there is lots of pressure to have a positive result, since that is required to sell a new product – but most scientific publications are sound, even where there are incentives to cut corners. Most scientists value their scientific integrity more than money, the review process works pretty well at catching inaccuracies, and the penalty for being caught red handed is severe.

However if there are no easy means to reproduce a given result, sincere errors are hard to detect, and it might take years until they show up. Requiring better reproducibility would make this much easier. Where full source data cannot be shared, it is often possible to provide example data, so this is a problem of process and legalities, not of practical feasability.

2.3.3 Propagating corrections

The second problem however is harder: What happens if a problem does go undetected. Papers usually cite other papers to provide references to the foundation they build upon, but when a paper has to be corrected, only that paper is changed, even though the correction affects all papers which cited it. This destabilizes the foundation of science, which is made worse by the sheer volume of publications: a new paper contesting the existing one will be missed by most people. If a (relevant) error in even a single publication goes undetected, it can turn up in many more publications which build upon the research.

To fix this, the journals could explicitly propagate the correction: When a publication contradicts a previous publication, the journal marks the previous publication as contested. If the authors of the previous publication support the claim, the publication is marked as corrected and all works which cited it are marked as unstable. Since the journals usually know in which part of the publication the corrected paper was cited (it's in the latex source), they could highlight the impacted parts and then check whether the correction affects the core message of the new publication.

A common example which shows the two different cases are results referenced in the introduction. Often these provide a background which motivates the relevance of the research. But some are used as basic assumption for the rest of the paper. In the first case, a correction of the cited paper is inconsequential for the citing paper. The contesting need not be propagated to other papers using the results from the citing paper. In the second case, however, the correction might invalidate the foundation of the citing paper which casts doubt on its results and needs to be propagated to all papers which reference them.

Marking papers as contested could easily be accomplished by creating corresponding microformats: When publishing a paper which corrects an earlier paper, add a link to the earlier paper which says "A corrects B" (marked in microformat syntax to make it machine readable). As second step inform the journal which published the earlier paper. The journal then markes the paper as "contested by A". Then it asks the authors of the earlier paper for comment. If they agree that they were corrected, the earlier paper gets marked as "corrected by A". If they do not agree that the earlier paper was corrected, the paper gets marked as "B contests A". That way journals could routinely scan research cited in the papers they provide to ensure that all the assumptions used in the papers are solid - which would allow them to provide additional value to their readers: Show the last time, all references were checked to ensure that they weren't contested - and if a reference is contested, check whether its correction impacts the core message of the research.

It would strengthen the role of journals as guardians for the integrity of scientific publication.

2.4 Summary

With the current state of scientific publishing, it is hard to keep a general knowledge of related fields, which leads to repeatedly reinventing the same methods in different contexts. Also errors which make it through the review-process and persist until they are referenced by other publications can persist even though they might be corrected in the original publication.

These challenges can be addressed by periodic overviews at different levels of specialization, reporting on both the state and the changes of scientific knowledge and methods, more support for reproducible research and reproduction studies and propagating corrections to papers into those which reference them.

3 Conclusions

3.1 Conclusions

Many aspects of scientific publishing are unmatched even with all the new development in the web, but the rising number of publications per year creates new challenges.

To meet these challenges, structured overviews and high-level updates to the current state of the art could help reconnecting different fields of science, and reproducible research, incentives for reproduction studies and propagating corrections to papers could ensure that published results stay trustworthy with the growing number of active scientists.

There are already journals and organizations who try to fill the role of reconnecting science, so I am confident, that these problems will be addressed with time. I hope that this article can contribute by providing an overview of the challenges and a clear vision of questions which need new and improved answers with the growing number of scientists and publications:

- "What's the expected reading for scientists?"
- "How do you know that you can trust this paper?"