

Friction in the scholarly workflow: Obstacles and opportunities

John Sack

Founding Director, HighWire Press

E-mail: sack@highwire.org

Abstract. Is disruptive change coming to the scholarly world? Friction has been mounting in the scholarly workflow as well as scholarly publishing. After an introductory clarification of the workflow versus publishing, I go on to examine telltales of scholarly friction, which can be inferred from the over 600 small fixes that have been invented in the past years to keep the system going. Next, I turn to the 12 most annoying friction points, Finally, I examine the rising likelihood of disruptive change and the key role that machine learning will play.

Keywords: Disruptive change, scholarly publishing, scholarly workflow, machine learning

The scholarly workflow as well as scholarly publishing are described and experienced as a well-organized pattern that has endured over decades. That both patterns be obvious and repeatable has been important to the scientific enterprise because it supports the universal aspiration, reproducibility of results, evaluation of new knowledge claims, generation of new hypotheses, and so on. The workflow has also included peer review as a deliberate friction point in the form of, for example, thesis disputation, manuscript evaluation, and promotion review.

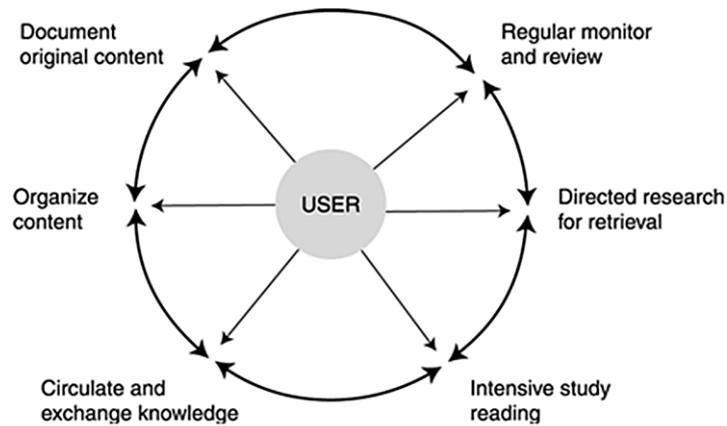
Yet, and this is my focus in this article, friction between scholarly workflows and publishing seems to be escalating rapidly as evident from the multitude of innovations that have sprung up to patch over a misalignment of practices, bridge a gap, circumnavigate a rough patch and so on. The misalignment or “friction” increases as yesterday’s scholarly practices rub up against today’s consumer electronic communication practices and the scholarly possibilities they reveal. The underlying question that I also address is how and how much scholarly workflows and publishing are changing, and whether this change will be continuous or become disruptive.

As I will show, there are a lot of friction “telltales”. Yet, the contours of an altered, more communicative and conversationalist system are still hazy.

1. The changing scholarly workflow

The scholarly workflow as a repeatable pattern allows the researcher to get things done, and “done” is a way that others can connect with and accredit as research. A typical representation of the workflow would include a number of steps or core practices. As Fig. 1 on the core practices shows, at the turn of the century – just before “e-journals” were ubiquitous – the academic writing and reading workflow was described by the following six elements:

1. Regular monitoring and review of the literature.
2. Directed research for retrieval of scientific information.

Six Core Information Practices

Source: Institute for the Future

Fig. 1. The scholarly workflow wheel.

3. Intensive study reading.
4. Circulation and exchange of new knowledge.
5. Organization of new content.
6. Documentation of original content.

Some ten years later the Internet and electronic publishing had become ubiquitous – we had dropped the “e” in “e-journals” as redundant – but the workflow had not changed significantly. To be sure, new tools had emerged, e.g. portable annotation supported intensive study reading; or integrated literature management supported the organization and monitoring of content; and both of these could be shared more or less in real time with local or distant colleagues, but the workflow continued much as before.

More recently, interconnected workflows rather than individual procedures have become more prominent. Scholars increasingly communicate with colleagues and other interested parties all the time: before, during, and after any task. This continuous interconnection is beginning to reshape scholarly workflows as well as publishing, and I think it is hard to overestimate the long-term impact of this digital and communicative transformation. Just as “always on” and “always reachable” network technologies have transformed consumer culture, the connected scholarly workflow will alter how we work together.

2. Scholarly telltales of friction

Friction is a distraction from the task at hand. We pause and need to change course, patch, repair, and so on. This happens because the next and obvious step is not facilitated by the system, or at least not facilitated well any longer. Friction may cause delay only, but it can have more serious consequences affecting the scientific enterprise.

When friction becomes a problem across groups and for the system, then this is also an opportunity for innovation. A notable project at Utrecht University has been tracking innovations in scholarly communication. By the latest count, over 600 innovations have been identified. Figure 2, the circle of

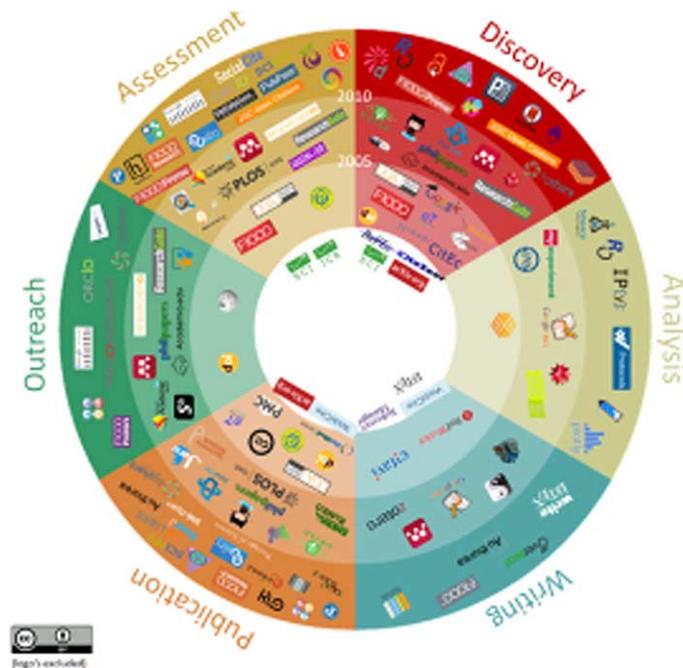


Fig. 2. The circle of innovations in scholarly communication.

innovations,¹ indicates that the sheer volume of innovation leads to the escalation of friction, not least because solutions are piecemeal only.

No scholar, department, or institution can keep track of these innovations, which also impacts their ability to integrate many of these meaningfully. In sum, we may expect that piecemeal innovation at a high volume adds to the burden – like having to assemble a car from a kit of car-parts, when what you want is *get* somewhere, not build a car.

Piecemeal innovation adds pressure to calls for change as it becomes clear that the workflow is no longer smooth, and increasingly broken. Current innovation provides only ‘fixes’ but no lasting and feasible rearrangement of the system.

3. Annoying friction points

Below I have noted twelve of the most annoying friction points in scholarly practice.

1. Manuscript submissions (Assessment): Publishers and journals have different rules and formats for submission, requiring authors to prepare every manuscript individually, particularly also re-submission to a second journal. The process is clearly burdensome and inefficient, but ‘format-neutral’ submission solutions are only now emerging – e.g. with Genetics Society of America, Rockefeller University Press, EMBO Press – despite the simplicity of the solution.
2. Barriers to dissemination (Assessment): The formal process of reviewing before publishing slows down the dissemination of new knowledge claims. Noticeably, the expert community trusts itself

¹Jeroen Bosman and Bianca Kramer at Utrecht University Library have been leading on this project; see <https://101innovations.wordpress.com/>.

over the review system, as is evident from the proliferation of preprint servers and content sharing networks. While physics and economics have had a preprint server for over 20 years, the same thing is only recently rising in life sciences (bioRxiv) and chemistry (ChemRxiv), with other fields likely to follow. The uptake of preprint servers may be the biggest change in scholarly communication since journals went online in the mid-1990s.

3. **Missing media (Writing):** Researchers increasingly create complex and interactive media objects – e.g., interactive simulations that accept input from a reader and respond to that input, as with javascript or similar languages – that do not fit into ‘flat’ publishing containers, and in any case have to be multiple-versioned for HTML and PDF. Researchers then must post (rather than publish) these objects to either local servers or to vertical services intended for them. In either case, the permanent accessibility of these objects in such locations (aka “link rot”) is a serious issue, with links to non-archival objects “rotting” (failing) at a rate of greater than 20% per year. That is, every year after publication, an additional 20% of the links fail, so that after a few years, most – and eventually all – will fail to resolve.
4. **Supplementary information (Analysis):** Many articles come with supplementary information on experimental methods and data, but researchers have difficulties using it because the data is locked in a separate PDF; thus re-use is cumbersome. Further, article context for this information must be restored with much effort: essentially the reader has to re-integrate the supplemental information back into the article narrative. Storage of the supplements is often a challenge, beyond storing the PDF of the article itself. This is a further instance where the flat publishing containers are a problem.
5. **‘Grey’ literature (Outreach):** Authors explain their work in helpful ways across a range of online venues: blog posts, meeting presentations, university press releases, video abstracts, journal clubs, etc. Typically, these venues will link in to publication sites, but the publication sites do not link to this “grey” literature, thus the journal site becomes a publication silo – a “walled garden” was the phrase early in Web development – rather than a hub for outreach. Editors and publishers feel responsible for what appears on a journal site, but can’t possibly review authors’ works beyond the article itself.²
6. **Collaborative annotation (Reading):** Annotation typically is personal, based on the PDF, and available in proprietary formats only. Yet, digital science is all about collaboration, and critical notes help scholars advance their work. Annotation is not yet part of or linked from the article, though solutions from hypothes.is and PaperHive suggest that standards-based tools are on the way.
7. **Online reading (Reading):** Publishers’ web design has become brand-heavy. Furthermore, much reading resembles ‘distracted driving’: There is way too much diverse information on any website or article page, hampering readers’ awareness and comprehension. This contributes to researchers’ continued preference for the PDF.
8. **Keeping up with the literature (Reading):** Literature alert systems are popular but do not reduce the time spent to stay abreast of a field’s literature. The costs of the traditional ‘just in case’ approach of identifying, scanning and filing relevant literature have become too high for the individual scholar. We must now resort to a ‘just in time’ approach of updating literature awareness as we write for publication, which means that less and less scholars have a genuine overview of the relevant literature.

²See: “Google’s Magic Carpet: The Value of Grey Literature” at <https://blog.highwire.org/2015/09/21/googles-magic-carpet/>.

9. Literature search (Discovery): The rise of interdisciplinary and translational research requires the exploration of new literatures, with vocabularies unfamiliar to researchers new to the domain. However, despite search engines, any search is far from straightforward as you perform a ‘search-term hopscotch’ to identify the relevant articles.³ Solutions may come from concept-based search engines, such as Yewno, rather than from traditional keyword-based search engines. Similarly, following references is another time consuming activity as one typically has to scan the whole of a referenced article to find the relevant bit buried inside. Another example is the lack of book-chapter indexing in search engines, making it difficult to identify the relevant information in scholarly books: you can get to the book, but not the relevant chapter, much less the relevant paragraph.
10. Scholarly-image indexing (Discovery): For scholarly research, figures and images play an increasingly important role. In fact, many now focus on the figures and then refer to the text for narrative context. However, if you were to hope that a search for figures will lead you to the right articles – much less the right figures – you will be searching in vain. We have no viable scholarly-image indexing.
11. Off-campus access (Discovery): Being off-campus is an excellent opportunity for discovery and reading, but to gain access to the version of record requires following a tedious sign-in process every time and sometimes multiple times, to use a campus of employer’s proxy server. Typically, it is faster to get access to a free author-version, so scholars will make do with that version, even if it lacks the imprimatur of peer review. Or they will access an article via Sci-Hub, bypassing the elaborate sign-in procedure.
12. Mobile access (Discovery): The proportion of mobile use of Google Scholar is only one-fifth to one-tenth the mobile use of Google Web Search. What is the barrier to scholarly discovery on mobile devices? Some would say there is no real interest in reading articles on mobile devices. While this is probably true, this explanation ignores the very significant use case for scanning, finding and filtering research content to identify what to read. And this is what mobile is great for – except in scholarly search. Mobile scholarly search is held back by the slow response of web sites to show an abstract, which is necessary for scanning literature beyond the title. We can fix this.

4. The future of friction, and of the scholarly workflow

Scholars as consumers – like the rest of us – have experienced discontinuous, technology-driven change in using phones, watching videos, listening to music, and getting the news. Our consumer experiences shape our expectations with regard to the possibilities for our online work. And the scholarly workflow of the future will increasingly be guided by our consumer expectations.

If we can get “same day” delivery of a physical toy or a pharmaceutical, why can’t we get same-day publication of an information object like a research article? There are a number of reasons, but all of them are under pressure from friction.

I believe it important to make the effort to ‘see around the corner’; to anticipate what discontinuous change may lie ahead. We must remember that more of the same, e.g. Moore’s law about the doubling of the number of transistors in a densely integrated circuit approximately every two years, means that the ‘knee of curve’ will take us around the corner, e.g. the emergence of the smartphone. The smartphone is a device used less and less as a telephone. Yesterday a phone was something you talked into to converse

³‘Hopscotch’ refers to the more or less creative and cumbersome process of identifying which keywords actually will give you relevant search results.

with another person. But today's phone is a Swiss Army knife for communication and action – and not even predominantly with people.

Today's metaphor of the scholarly web is that it is like a library: full of documents to read and to write. Yet, tomorrow's metaphor is likely to be different, and the question is how much different will it be?

Consumers already see the web as a place to *do* things, not just read about them. Even libraries at universities are changing to places where you do things, not just borrow and read documents. The scholarly web will evolve this same way, as workflow goes beyond engagement with the literature and integrates literature into the overall work of the researcher to discover and communicate.

The scholarly publishing model transferred easily to the Web 1.0 as a one-way broadcasting system. Yet, two-way and *n*-way communication models are becoming increasingly prevalent in scholarship, but their convincing implementation in scholarly communication is still missing. In the meantime, friction continues to build up, since the equivalent “Web 2.0” technologies have been popular in the consumer space for over a decade. And younger researchers expect to communicate their results, not just publish them.

The frictions outlined above also point to the rise of supporting technologies and actionable knowledge. Machine learning is a major theme, and it is not just about reading the literature. Machines increasingly understand text and data, act on it, and contextualize and personalize the results. We see this in Google Web Search's personalized results, and in our use of GPS apps for navigation on highways and in unfamiliar cities. But we don't see personalization in Google Scholar results – and intentionally this is the case.

Searching for the scholarly-communication-equivalent of ‘precision medicine’ is a useful way of understanding what machine learning might yield. Hitherto, drugs have been developed to match and treat a condition in a large population. However, with ‘precision medicine’ you match a drug to a mutation by personalizing it to your genetic signature, leading to a much more effective treatment. The equivalent may be called ‘precision scholarship’. It may lead to a supporting machine knowing which topics you personally are interested in and finding the relevant texts and data, presenting them with annotations and highlights. In this sense the machine will not only read the literature, but it will *write* the literature that you, specifically, must know about. While this is certainly the realm of machine learning, it is hard to imagine it taking longer than, say, autonomous vehicles. The friction of keeping up, of finding and filtering, of awareness and comprehension, are just too significant in scientific fields.

The information retrieval system has been around since the 1960s, and the library metaphor has enabled its prolongation on the Internet. Obviously, the system is not broken yet, or at least not completely broken. Nevertheless, friction is escalating, and piecemeal innovation brings temporary relief only. That is why discontinuous and disruptive change will be coming to scholarly communication and practice too.