Why is Biomedical Research Not More Like Airbnb?

Vivien Bonazzi PhD and Philip E. Bourne PhD, FACMI. Last Update 112716

The thesis presented here, and first introduced at the International Data Forum[1] in Denver in 2016, is that biomedical research would be conducted more efficiently if the trusted exchange of services were more integrated, ideally into fewer platforms than exist today, if they exist at all. While much simpler than biomedical research, much can be learnt from internet-based platforms, like Airbnb that have emerged in the last few years.

At first glance, the idea that biomedical research has any relationship to an accommodation rental service probably seems absurd. Let us explain.

As we work through a strategy for improved data management and sustainability as part of our work with the National Institutes of Health (NIH), one of us (PEB) was also in the process of making an apartment available through Airbnb. While having a number of satisfactory rental experiences with Airbnb, he had never been a host (the person renting) before. It further exemplified what one experiences as a renter - it succeeds because it is a relationship built upon *trust*. Using Google drive or Github is also a relationship built on trust, but not to the same degree. The host (renter) trusts that the accommodation is going to be as advertised; the host trusts that you are not going to trash their property. Host and rentee trust Airbnb to manage the transaction. The software platform upon which Airbnb is based makes every effort to gather as much data on both renters and hosts as to determine a basis for trust. The platform is easy to use, and transactions are inexpensive. This is not to say the service is perfect. Issues have arisen concerning how Airbnb can change neighborhoods in areas with high tourist potential[2] or indeed of claims of racial bias by hosts[3]. Nevertheless, something is working, since as of February 2016, Airbnb had 60 million users searching 1.5 million listings in 191 countries, with an average of 500,000 stays per night[4]. All leading to a valuation of US$25bn. So what does this have to do with biomedical research?

Airbnb supports a trusted and arguably democratic service between providers and consumers of that service. Consider biomedical preclinical and clinical research, where the trusted service involves the exchange of papers, data, software, reagents and so on. An author publishes a paper having applied rules of scientific conduct and is a supplier. That paper is read by consumers based on how much they likely trust the work based on what they know about the authors, what journal it was published in, which speaks to the presumed quality of the review,

---

[1] http://www.internationaldataweek.org/International-data-forum
[2] https://www.linkedin.com/pulse/more-shocking-news-airbnb-colin-shaw
[3] Edelman, Benhamin (10 May 2014). "Digital Discrimination: The Case of Airbnb.com" (PDF). Harvard Business School. Retrieved1 October 2014
[4] https://en.wikipedia.org/wiki/Airbnb

who else cited it and what trusted bloggers and other social media contributors have had to say about it. The same applies to data. Knowing what laboratory the data comes from and having sufficient information about the methods used instills a sense of trust, as does a curated data repository that is well managed. A similar trust-based argument can be made for other resources that can be exchanged, including people. A principal investigator will hire a postdoctoral fellow based on the trust they have that the individual will contribute to the laboratory. That trust comes from not only what they have published, but where they have published and what their references, themselves trusted investigators (or not), have had to say.

From the perspective of managing the trusted relationship, what is different between Airbnb and biomedical research? Airbnb operates through a single platform through which services are exchanged and transactions conducted; biomedical research has multiple discrete and poorly connected platforms (Fig. 1).

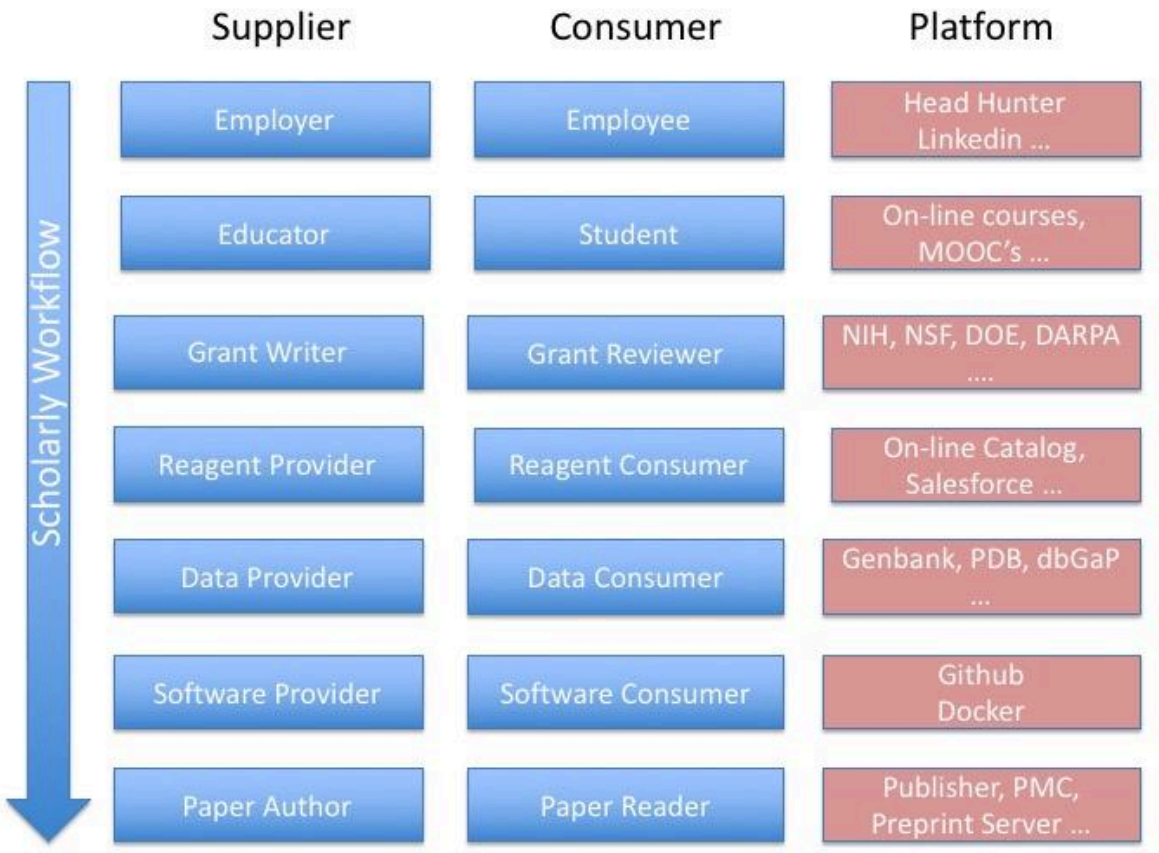

Fig. 1 Discrete Suppliers and Consumers and Associated Platforms Supporting the Scholarly Workflow. While presented as a linear path for simplicity, in reality, it is more a network of interactions.

Of course, the services being exchanged by Airbnb are much simpler than those of biomedical research. They are also direct services between two distinct parties, whereas biomedical

research involves many participants in the exchange. Moreover, Airbnb was born digital, whereas biomedical research has a long analog history, and for many stakeholders, the manuscript still represents the medium through which to communicate research - journals are *the* platform unifying contemporary research. Thus, cultural change is the biggest impediment to a presumed value proposition. Airbnb overcame this need for a cultural change presumably by immediately offering a platform and associated trusted service that was a win for all stakeholders. Can the same happen in biomedical research? Probably not since there are technical barriers greater than Airbnb faced, most notably the sheer size of the software development undertaking to create a single robust platform that is modular and customizable to different stakeholder needs. Nevertheless, platforms exist, and their adoption continues to grow. Therefore, we have seen some aggregation between the rows shown in Figure 1, but as of yet, a single platform does not encompass all trusted relationships between consumer and supplier.

Consider examples where the merger of the rows in Figure 1 - the traditional boundaries of research - into a single platform would advance the biomedical research enterprise. Realistically a single platform may be unobtainable as a result of remaining different business practices, governance models and technologies. In that case interoperability may be more realistic, but we should aspire to what appears to be a single platform, even if it is not.

Publishers provide an exchange of services between an author and readers; data resources provide an exchange of services between a data provider and a data consumer; software developers provide software to a user community. Think about these exchanges from the point of view of first the transaction that takes place, and second, the platform used to make the exchange. The transactions are similar, but the platforms are not, as one of us wrote eleven years ago[5]. Regrettably, not enough has changed in those preceding eleven years, even though the technology to support that change is available. What is still lacking is broad adoption through new business models and incentive structures. Consider the nature of these exchanges.

The authors submit a paper to a journal they trust via an editor who seeks reviewers who pass judgement on the paper, usually without review of any data or software used to generate the results and the conclusions presented. Assuming the paper is perceived to make a worthy contribution, one or more iterations occur between authors, reviewers and editor before the paper is published and appears online. That paper is then indexed by PubMed, Google Scholar, etc. so that it can be found and the findings disseminated. There are a small number of commercial platforms used by publishers to process manuscripts, garner reviews, and make available the final article in a small number of formats, notably HTML, PDF and XML (particularly JATS). The form of the output - abstract, introduction, material and methods and so on, is consistent across the corpus, but the end product is disjoint from the data, software, materials etc. used to support the findings. This was a natural consequence of an analog world,

[5] Bourne P (2005) Will a Biological Database Be Different from a Biological Journal? PLoS Comput Biol 1(3): e34 http://dx.doi.org/10.1371/journal.pcbi.0010034

where these were physically disparate entities, but it makes no sense in a digital world, where even physical entities, e.g., reagents, instruments, have a digital signature which uniquely defines and describes the resource.

Data provision and use follows a similar process *yet is completely disjoint from the publishing process in most research domains*. The data generator, under the auspices of the research funder and the data generator's institution, both with conditions for data sharing, provides the data to an identified trusted repository. That submission is reviewed, either automatically or manually by biocurators, or in combination. At which point the submission is accepted or returned to the data producers for improvement. Once accepted, these data are provided with a unique identifier and in a format typically defined by the field of researchers who generate this type of data. So far, the process for the exchange of data is identical to the exchange of knowledge embodied in a paper. As such, it makes no sense to separate the two, particularly when a serious understanding of the paper, as for example, in reproducing the results, requires that the data and analytical methods, likely embodied in software, be accessible. Huge amounts of time are spent by researchers essentially putting back together the various pieces of the research process that were taken apart in the name of analog-based dissemination.

The senselessness does not end there. Whereas at least research papers are presented somewhat uniformly, data are not. In fact, the means by which data infrastructure is funded, typically as part of a research endeavor, fosters environments which seek to be as different to each other as possible so as to highlight their advantages. The end result is a multitude of different user interfaces, access methods and general lack of uniformity. Contrast this to Airbnb where a single platform manages all transactions, for example, local taxes and payment from renter to host are all are handled uniformly. To anticipate that all funders would use a single platform for transactions between suppliers (the agency) and the consumer (awardees) is unrealistic, yet working together to create a more uniform supplier side as it relates to data and other resources is where progress can be made. For example, the Food and Drug Administration (FDA) has taken steps to make things more uniform -- all trials that FDA regulates have to sit on a CDISC platform. This will make it easy for FDA to compare trial results to each other. Similarly the NIH is encouraging more uniform use of common data elements across the Institutes and Centers.

A final act of inefficiency occurs when important data that was available in digital form, but presumably had no obvious data repository as a home, is later manually extracted from the paper and placed into a digital repository at significant cost. The value of a platform that integrates data, software and research articles in a way that is trusted by suppliers and consumers should be obvious.

While beyond the scope here, the change that platform adoption, in the way described above, could have on scholarly communication is enormous. Suddenly the unit or currency does not remain the artificial boundaries imposed by journal article page limitations and the amount of

supplemental data that can be included, but becomes a matter of attributable units of data and associated narrative that could change how we conduct and report scientific output.

So far, we have focused on the relationship between data and publications, yet other layers of the research enterprise, as embodied in Figure 1, can be described in a way that points towards the need for greater integration. Software is increasingly available through open platforms like Github, which is commendable, but again usually disjoint from the research to which it was applied, although at least links to software are beginning to appear in research articles and elsewhere.

The commercial world has its own platform technology broadly referred to as Customer Relationship Management (CRM), exemplified by products like Salesforce. CRM's manage the interactions between a customer and a supplier and have some limited use in academia, but there is nothing substantive for managing that trusted interaction that happens around, say, the exchange of antibodies, exchange of best practices for a given experimental protocol and many other exchanges that form the basis of biomedical research. Such exchange is ad hoc. You discover reference to an antibody in a product catalog or reading someone's paper. Best practices are only exchanged years after you discover through the literature or a conference that a laboratory is adept at a procedure and you send a student to learn it. Moreover, CRM's are disjoint from other aspects of the research enterprise, such as publishing platforms, data archives, software archives and so on.

In summary, there is not currently a widely adopted single platform or highly integrated platforms for the exchange of services in biomedical research. Either there is a platform per service, or limited set of services, or no platform at all. Why have we not done better and what are the impediments today?

Surrounding the analog model of biomedical research is a strong tradition and a set of business practices that drive the enterprise. Change to such an ecosystem generally comes slowly. From a business point of view, such a change can involve a very significant investment without a clear understanding of the value of that change to the business in question. Only when the customers see the clear value of making a change is it likely to drive that business into making change. This is antithetical to the attributed Henry Ford approach of, "*If I had asked the people what they wanted, they would have said faster horses*." Or Steve Jobs for that matter, "*A lot of times, people don't know what they want until you show it to them."* These giants had two advantages in overcoming the inability of the stakeholders to see the value - technology and money. The production line for the Ford Model T is an example of technology and, well, Apple has a valuation greater than the GNP of 2/3rds of the world's countries, which speaks to having money.

Is there a Ford, an Apple, or a philanthropist for that matter, willing to make the investment in developing a platform for biomedical research that embraces all the needed exchanges? It would seem the answer is no. Certainly in academia there is neither the incentives - it won't

bring you tenure - nor long term funding to undertake such an endeavor. Rather, we have a situation where government funders, foundations, the private sector, have each invested in a small piece of what is needed (Fig. 2).

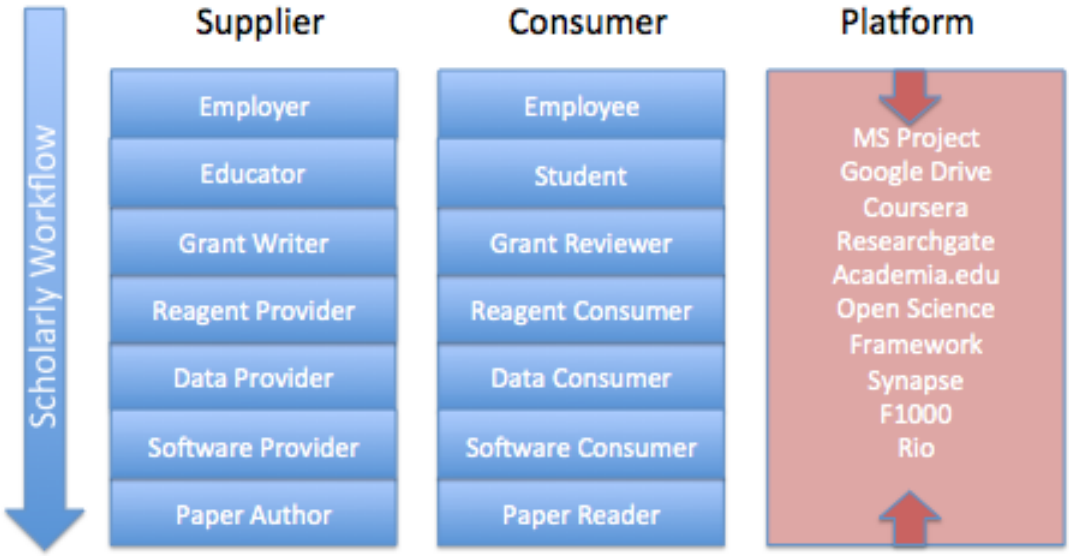

Fig. 2 Example Platforms Currently Supporting Biomedical Research

Figure 2 by no means covers all platforms in use in biomedical research, and each of these example platforms addresses a varying amount of the biomedical research enterprise. For example, Google drive enables file sharing in a generic way and thus facilitates the exchange of data, software, and knowledge but does not directly support actions on these particular objects of research. Synapse[6], in a specialized way, and the Open Science Framework (OSF)[7] in a more generic way, both support the exchange of these research objects and actions upon them. Researchgate[8] supports the exchange of knowledge through papers and social networking between people. Coursera[9] supports a learning environment but no other forms of exchange. Each is a piece of the puzzle, but not a broad solution. OSF comes closest to being a generic exchange platform with the ability to plugin what the community requires as the system evolves, thus fostering a uniform digital ecosystem.

What would be the advantage of having a more complete trusted exchange platform across the research enterprise? Consider a couple of hypothetical use cases.

[6] https://www.synapse.org/
[7] https://osf.io/
[8] https://www.researchgate.net/home
[9] https://www.coursera.org/

An obvious example is the long talked about executable paper. Rather than the paper being an advertisement for the research[10] it becomes an actionable entity. Computational components of the analysis can immediately be rerun, data associated with the study explored, and modifications made to the experimental protocol and the experiments rerun, perhaps on large amounts of data in a cloud environment. As such, this truly builds on the original study. This of course happens now, as any graduate student could attest, but in an arduous and time consuming and hence costly way. You download what you can by way of data and software associated with the paper and attempt to recreate and build upon the experiment. We attempted to quantify the added cost associated with these necessary actions for one in silico experiment[11] and came up with 280 hours of added labor 3 years after the original work was carried out. Now, 6 years after the original experiment was conducted, the cost of reproducing or building on the work would be much higher, if indeed it was possible at all as data and software atrophy away from a project that no longer has grant support.

A slightly less obvious example is the relationship between the products of research and the people who stand to gain the most from those products and from each other. At present, that relationship is of limited value and disjoint. You may receive alerts of work from a particular laboratory, mostly through the papers they publish. In our experience these are usually inaccurate, so you end up with unwanted interruptions and certainly for those of value that laboratory has no way of knowing you are interested in their work. Tweets and other social media provide a blanket awareness but there is no accurate, ubiquitous, trusted point-to-point interaction. That interaction can come from matching an accurate and current profile of the scientist wanting to be alerted to metadata, and/or the accurate extraction of semantic content, from the paper, Github, wikis, Twitter, etc. The technical wherewithal to make this happen exists, but accessibility on the same trusted platform is prerequisite and does not exist in a broad context. Broadly speaking, the problem is the social contract needed to make this a reality.

A social contract can potentially develop over time, provided the platform is in place and parties to the contract have a measure of trust in each other. Even then incentives are needed. In the case of Airbnb, the platform was enough of an incentive, even though other similar services existed. Why Airbnb would become the dominant platform has been discussed most notably by Sangeet Paul Choudry[12]. In the case of biomedical research, there is already a way to conduct business as usual, so further incentives are needed.

Scientists themselves are incentivizing others. For example, the work of Ben Best as Duke University using cloud-based data services with R code executed directly from Github[13] or those

---

[10] Jonathan Buckheit and David Donoho, paraphrasing Jon Claerbout; J. Buckheit et al., About Wavelab, tech. report EFSNSF-491, Dept of Statistics, Stanford Univ., 1995.
[11] Garijo et al 2013 PLOS One, 8(11):e80278; http://dx.doi.org/10.1371/journal.pone.0080278.
[12] http://platformthinkinglabs.com/
[13] https://marinebon.github.io/analysis/obis_biodiv.html

posting their Jupyter[14] notebooks. This is not enough to incentivize the community in a significant and timely way.

Funders and publishers are uniquely positioned to provide those incentives. It is not clear that traditional publishers will provide those incentives - they have too much to lose from their current business practices. Newer non-traditional publishers have the desire to change the status quo, but not the resources and/or level of trust from the community to make a difference.

Thus it falls to funders to change the system. Even then, there are processes and procedures that one can think of as the equivalent of business practices that are not easily changed to enable the wholesale adoption of a biomedical platform. At NIH, we are taking an agile approach and creating the NIH Commons[15] as our platform. By Commons we mean an open and inclusive data storage and compute environment which facilitates the discovery, execution and reuse of digital research products. The Commons constantly builds upon itself as new research products are generated and assessed by the broad community of users. By agile, we mean running a series of experiments to evaluate the value of a large scale effort. Those first experiments use public infrastructure, notably clouds, to support the platform, which is focusing on computing on large datasets. Initially, there is not a focus on people, educational materials, digital signatures for reagents and the like. What can we evaluate from such a limited Commons experiment? An obvious potential benefit is there is no other option that the cloud for computing on large datasets. Less obvious is the sharing of data, software and computational protocols implied by the Commons. In principle, as new data and methods are generated from the initial experiments they too are accessible on the NIH Commons platform. For this to work, users must be able to find the content on the platform, and they must be able to trust it (or not). To be findable, content must be indexed, and that requires that each component of the research be uniquely identifiable and resolvable. Various experiments underway through the NIH Big Data to Knowledge (BD2K) program, notably DataMed[16], provide indexing tools, the metadata and associated standards to make this workable. Access and usage more broadly of the research components resident on the platform at least provide some metric for the trust that users have in those components, and more elaborate incentives through the platform can be envisaged.

Airbnb incentives suppliers and consumers to provide the best service through alerts, comparisons to those providing similar services and so on. A similar model can be imagined for research where, for example, comparative usage metrics and recommendation engines for data and software can speak to the relative value of the service.

We hope to have some sense of the value of the NIH Commons as a platform in a year or two based on pilots that are underway or soon to start. Success would see the services offered extend into additional layers of the scholarly workflow (Figure 2). Just as Airbnb has created a

[14] http://jupyter.org/
[15] https://datascience.nih.gov/commons
[16] https://datamed.org/

new and frequently rewarding way to travel, we hope for a more cost-effective and productive way to perform biomedical research.

Acknowledgements: Thanks to Bill Barnett, Sky Bristol, Daniel Mietchen, Michael Lauer, Melissa Haendel and Michel Dumontier for useful insights into improving this perspective.