A conversation with Joshua Greenberg on 03/13/13

Participants

- Joshua M. Greenberg — Program Director, Digital Information Technology, The Alfred P. Sloan Foundation
- Alexander Berger — Senior Research Analyst, GiveWell

Note: This set of notes was compiled by GiveWell and gives an overview of the major points made by Joshua Greenberg.

Summary

Joshua Greenberg is the director of the Sloan Foundation's Digital Information Technology program. GiveWell spoke with him as part of our investigation of opportunities to improve scientific research. We discussed the organizations and projects that the Sloan Foundation's Digital Information Technology program is funding, and other issues related to scientific communication.

The Sloan Foundation's Digital Information Technologies program

The Sloan Foundation's Digital Information Technologies funds projects in two areas:

1. Data and computational research, which is focused on how you turn data into knowledge, from both a workforce and technical perspective. On the technical side, this includes tracking provenance, versioning, and logging of research. On the workforce side, this includes asking who does that, how they're funded, etc.
2. Dissemination and sharing. As more and more scholarly materials show up on the internet, whether in open access journals or alternative kinds of products, it's important to try to better link together disparate products and provide filters for users.

In this context, the Sloan Foundation serves as an early-stage investor to early stage public good projects, hoping that they'll be successful and receive funding from others later on.

The production of software for the analysis of data and efforts around stewardship and archiving of data are less of an immediate priority for Sloan, as both are perceived to have significant existing attention from the funding community.

Overall, the program aims to maximize the efficiency and trustedness of academic research.

Sloan Foundation Digital Information Technology program grantees

Some grantees include:

- iPython notebook — a browser based lab notebook for scientists who do computational research to document their work. This had been an open source volunteer project until recently, when its creators started working on it full time.
- RunMyCode — A website designed for researchers to upload their code, and for other researchers to run it, in order to reproduce the results from their papers.
- Dataverse Network — An organization working on a project to make data sharing and data preservation an intrinsic part of the publication process.
- Knowledge Infrastructures: UCLA — A project to study data management practices at the beginning and at the end of large-scale scientific projects. This project's goal is to determine what practices researchers can adopt at the beginning of the project in order to ensure that they'll be able to deliver the data from their research later on.
- ImpactStory — A service that aggregates alternative metrics of scientists' research outputs.
- PressForward — A platform for overlay journals, which aims to generate automated metrics to be used in conjunction with peer review and editorial filtering.
- Hypothes.is — a platform for users to annotate web site content, and to read the annotations of other uses.
- *Beyond the PDF* conferences

The value of encouraging data and code sharing

It's currently the case that there's a high barrier to scientists sharing their data and code, because they often haven't organized it in a way that's easy for others to understand and because there aren't substantial incentives for doing that additional data curation. If there were norms of

- Carefully documenting one's data and code as one produces them
- Organizing one's data and code so that they can be used outside of the context in which they were originally created

then it would become much easier for researchers to share data and code. Creating software to help researchers with these things is a potentially promising philanthropic area.

The practices above would help the researchers themselves. It's often the case that

- Scientists forget how to interpret the data that they've collected.

- The person in a lab who knows how to interpret the data later leaves the lab, so that the people in the lab don't know how to interpret the data.

If the data and code are well documented, then researchers will be able to use their own data and code even if they don't remember how to interpret them.

Researchers have various reasons for not sharing their data and code, but the difficulty of sharing it in a public context is often the easiest explanation for not doing so. If it became easier to share, then researchers might feel more pressure to share, because the technical excuse would cease to be credible.

If researchers shared their data more, this would improve researchers' care and the soundness of their research, and so improve the trustworthiness of science.

Other foundations supporting technology to improve academic communication

- The Moore Foundation. Chris Mentzel is the program officer who works in this area.
- The Arnold Foundation. Stuart Buck is the program officer for this area. The Arnold Foundation recently awarded a large grant to the Open Science Framework, which is a project aimed at creating software to support scientists' workflow, and better align scientific practice with scientific values, for example, by increasing data sharing.
- The Mozilla Foundation. Sloan has actually given them a grant to hire a program lead for open science. Mozilla has played a catalytic role in the open source community, and has done compelling work, for example, on digital tools within the field of journalism.
- The Kauffman Foundation. Sam Arbesman is a scholar there who's interested in the intersection of entrepreneurship and academic research.
- The Mellon Foundation (which works in the humanities rather than the sciences). The Scholarly Communication program, led by Don Waters and Helen Cullyer, is involved in similar issues.

People for GiveWell to talk to

Aside from the foundations listed above:

- Kaitlin Thaney: The manager of external partnerships at Digital Science. Digital Science is a subsidiary of Macmillan that aims to serve scientists' needs for computer technology. Kaitlin Thaney previously worked on Science Commons at Creative Commons.
- Kathleen Fitzpatrick: a former professor of English at Pomona, now head of Scholarly Communication at the Modern Language Association. She wrote a

book titled *Planned Obsolescence* arguing that the filtering of scholarly materials should take the form of post-publication peer review rather than pre-publication peer review.

All GiveWell conversations are available at http://www.givewell.org/conversations