

PROFILES IN OPEN: MICHAEL GOTTLIEB



MICHAEL GOTTLIEB served for 13 years as the Deputy Director of Science at the Foundation for the National Institutes of Health (FNIH). In this capacity, he served as the principal investigator for the Etiology, Risk Factors, and Interactions of Enteric Infections and Malnutrition and the Consequences for Child Health and Development (MAL-ED) study. Supported by a grant from the Bill & Melinda Gates Foundation, MAL-ED explored how the interaction among a variety of factors – including environment, nutrition, public health, and local medical issues – influenced physical and cognitive childhood developments across eight locations in Africa, Asia, and South America. MAL-ED required the compilation and analysis of a wealth of complex data. Because of the gravity of the issues under examination, the MAL-ED team felt an urgency to share these data quickly and widely with qualified researchers around the world. Dr. Gottlieb’s perspective on open science issues has been further influenced by his work as a grant program manager at both the National Institutes of Health and the Gates Foundation.

How did you begin to engage in open science practices?

My interest in open science dates back to the 1990’s, when I served as the program officer for a number of NIH and National Institute of Allergy and Infectious Diseases supported parasite genome sequencing projects. For example, I managed a project on *plasmodium falciparum*, which causes malaria. This was a distributed project, one of the first of its kind. There was a need for common mechanisms to share pathogen data among the investigators at the three participating centers. To facilitate this, The University of Pennsylvania developed PlasmoDB to house the data. The traditional view among researchers was that data should be closely held through the publication life cycle. However, with this pathogen project, the participating researchers agreed that there was greater value in sharing this data to help identify genes, drug targets, and parasite invasion paths. PlasmoDB has since evolved into a component of EuPathDB, a structure warehousing data for eukaryotic pathogens, sequence data, expression data, and a range of other information. It is not just a repository to house the data. It includes tools to identify sequence similarities and annotate specific sequences. In this sense, it is a good demonstration of how open science can work. It is a resource for the community to share data at an early stage and render it usable. EuPathDB has helped accelerate eukaryotic research, pushed it into new directions, and allowed a wide range of investigators to contribute to the problem space.

What did your funder ask of you with respect making your research open?

The Bill & Melinda Gates Foundation funded my research into malnutrition and disease in the developing world. We were asked to consider, from the outset of the project, the best way to make our data quickly available. We developed mechanisms to share our data while still respecting the rights of the individual study participants and the local communities.

How did you feel about that?

Investigators in studies like this have some concerns about open data sharing, particularly scientists in the developing world. Every country has a different set of data sharing requirements and it can be difficult to align these interests. We needed to move carefully to respect these concerns, while simultaneously pushing to make our data widely available.

How did you make your research outputs available?

In the course of the malnutrition study, we often received requests for data and samples. We established ad hoc mechanisms for early data sharing. Eventually, the Gates Foundation worked with the University of Pennsylvania to develop the Clinical Epidemiology Database (ClinEpiDB). ClinEpiDB is geared to accommodate population-based epidemiological studies, which are incredibly complex and can include many thousands of clinical observations and human subjects. ClinEpiDB offers researchers

standardized processes for accessing and exploring complex clinical data. This makes the data not just open, but usable.

How did making your research outputs available impact further exploration of this topic?

We found that sharing our data was actually a good way to extend our work. Not only did it allow other researchers to explore and build off of it, but that follow-on work, in turn, encouraged us to talk with these other groups about new approaches and ideas. For us specifically, having access to shared data and shared samples led to the testing of new hypotheses about other environmental factors impacting growth and development. In our field more generally, the parasite and microbial communities have been totally transformed by data sharing. It has allowed researchers to analyze more data, and also to collaborate on common interests like functional analysis of gene sets.

Did making your work more open lead to subsequent analysis and debate about your findings? If so, how does this experience impact your attitude toward open sharing?

The malnutrition project was incredibly complex, so making our data open created additional debate about methods and findings. Ultimately, conflicting results and interpretations point the way to future research, which is good for the field. That's the way science works.

What advice would you give to other researchers who are contemplating making their work more open?

In general, make a decision early in a project about how to share your data. Work with your funder to make sure it will best benefit the field. For larger projects with multiple stakeholders, get early alignment among investigators. Think about how the data will be supported in the long run. Quality control is also critical. For a shared dataset to be useful to others, it must be accompanied by a set of precise definitions of the datatypes included in the set. This goes for both primary data (e.g., "length") and for derived data (e.g., a "length for age Z-score"). Further, there should be an indication of what steps have been taken to verify the quality and standardization of the data especially when data are collected at different sites by different investigators using similar but not identical equipment.

What would you like to tell funders who are thinking about embedding open science principles into their grants?

Funding agencies need to evolve to understand the best mechanisms to share data in their given field. There is a cost to preserving and sharing data that funders should strive to support. Also, different jurisdictions and funders have different policy requirements. To the extent that these can be harmonized, it will make it easier for researchers to know what to do and how to comply.

Additional Resources

Profiles in Open are a service of the Open Research Funders Group (ORFG). The ORFG is a partnership of funding organizations committed to the open sharing of research outputs. Visit our website (www.orfg.org) for more resources including:

- **"Open 101" Tip Sheets**, designed to help specific audiences understand the benefits of open science
- The **"HowOpenIsIt?" Guide to Research Funder Policies**, created to help philanthropic organizations develop open policies consistent with their values
- **The ORFG Curated Reading List**, containing a wealth of scholarly research and real-world case studies that demonstrate the myriad ways in which open access and open data benefit researchers and society alike