Sage

Internationalization Workstream

Report – April 2010

Mission and First Results

The creation of a disease biology commons faces a unique set of challenges as it moves from a single-country specific effort to an international one. Laws governing data, databases, and data products differ widely across regions, as does the capacity of clinical subjects and scientists in developing nations to participate in a contributor network.

At the beginning of the Sage Commons Congress planning, it became clear that the project lacked a meaningful understanding of the scope and scale of the issues posed by the international nature of the proposed work. Yet it was clear that the project posed an enormous potential asset to the international community, and that these issues needed to be understood sooner rather than later. To begin developing this understanding, a workstream project was created to find and interview key scientists and thought leaders about the challenges of internationalization, the experiences gained in spaces like immunology and infectious disease dealing with broad international communities of clinical practice, and develop a questionnaire for international participation and research.

The workstream group performed a series of interviews with key scientists from the international community. It became clear that even though we knew the internationalization issue was complex, it is even more complex than we thought. The questionnaire we had hypothesized to create is more like a PhD thesis than something that can be quickly crafted and released. Instead, we present this short report to summarize the key issues that emerged in the research, and plan to continue the conversations we have begun. We will be actively engaging key individuals and institutions to assist the Sage Commons going forward on international issues, and expect the questionnaire to re-emerge after additional consultation and research with the community. We call on interested parties to annotate this document with experiences, solutions, and new ideas.

These key individuals will also be identified during the Sage Commons Congress and through the network of Sage participants and volunteers, and can take roles as project workstream collaborators or interviewees. After the Congress we aim to keep an ongoing conversation with those identified experts, complemented by a literature review of the related topics and challenges.

The mission of the International Project is to provide a broad understating of international challenges and perspectives and possibly a series of design recommendations, which will provide input to the Sage Commons expansion over the coming year. Our goal is not to provide answers, but instead to generate a conversation among a global group of experts before the Congress, and to begin sketching a map of the complexity of a global, peer-production system for scientific data.

Sage: Globally Coherent Data Sets

The international issues around clinical data frequently revolve around the privacy rights of the human individuals being studied. Sage's "globally coherent data sets" may, or may not, be subject to these rights, and may or may not be subject to various contracts imposed by the institutions involved in collecting the data.

Sage data sets are composed of three core kinds of data. First, a genotype is captured at a high level of resolution (such as 1500 SNPs). Second, RNA is isolated and expression levels are measured via microarray. Third, clinical measurements such as plasma cholesterol, HDL, LDL, and so forth are measured from samples collected from the individual. Some data sets are based on mice and therefore face no barriers of privacy or contract (simply requiring a lightweight approach dealing with copyright and database rights), while data sets related to humans may require more attention, especially if there is enough metadata about the subjects to allow re-identification of individuals.

Key Themes Regarding International Data Collection and Distribution

The first key theme related to the Sage Commons concept is, not surprisingly, the constraints created by the use of human subjects as the source of data. These constraints exist at multiple levels of hierarchy, from the individual's right to privacy and/or anonymity to the obligations posed by institutional review boards (IRB), in both developed and developing nations, to observe strict regulations and pre-existing clinical study designs. Individuals participating in clinical studies have privacy rights that govern who can access and study their data, and how that data can be used, which can have the side effect of making complex data integration and commons-based access impossible.

The nature of the privacy rights differ nation to nation, and within nations, from institution to institution, making standardized access to data sets generated in disparate contexts very difficult to achieve "post hoc" or to design for in new clinical studies. These privacy rules exist due to the necessity of respect of the prior consent granted when that human subject took part in a specific clinical trial, to guarantee the idea of control over personal information, and to secure people against acts of discrimination.

For instance, UK data protection laws question if it is legitimate to even collect data that comes in randomly from around the world (such as in a Sage contributor network) as opposed to the accepted, legitimized practice in which institutions design a partnership with institutions for pre-arranged data collection. There was real interest in the group interviewed in the creation and promulgation of standardized, or even semi-standardized, text for inclusion in IRB agreements that would promote data access to anonymized information, though there was also doubt about the power to truly anonymize data about humans. One scientist interviewed noted that "agreeing to policy up front is more important than figuring it out retrospectively", which sums up this theme quite well.

The second key theme that emerged is the complex set of relationships between scientists and institutions across national boundaries. Some of this complexity is related to the national policies

that govern some scientists (such as the distaste many non-US scientists, institutions, and governments share for the data access potential imposed by the US Patriot Act) but most appears to be driven by competitive instinct, negotiating leverage, and perceptions of potential to exploit human subjects by scientists in the developed world. Several interviewed felt this was the most important issue in the space, and gave examples inspired by the experience of collecting tissue samples in developing nations and then attempting to "Repatriate" the data back to those nations and institutions.

One standout example is that in one case, a developed nations ethics committee prevented the sending back of analyzed data to the very scientists who collected tissues in the developing country that hosted the study and sent tissues to the developed nations institution for analysis. The committee required irreversible anonymization of data, with no possibility for any data being linked back to data in the countries where it was collected, so that it might help with conditions on the ground. In this case, even when all scientists were in agreement and wanted to share data, a solution took two years.

A third theme that emerged is that many complex data sharing projects are governed by multiple principles, ranging from the ethical to the legal and scientific. These principles are not always going to agree with each other and can create "orthogonal problems with data interoperability and usability" - interviewees pointed out the need to have consultative groups of different stakeholders to give legitimacy to decisions in complex situations in which principles conflicted. In these cases, success depends on reaching agreement as to where the jurisdiction of the principles lay, and which principles should prevail in a specific situation. Success was noted in cases with formal but lightweight process using standardized, pre-negotiated terms and documents creating a paper trail of ownership, responsibility, and confirmation.

A fourth theme was the potential for unexpected problems. Clinical data and public health data, it was noted, could be used as weapons if those data have negative economic externalities. The impact of asserting an epidemic of cholera in a major Indian city was an example, as was the impact of asserting the existence of a genetic variation associated with a long-term chronic disease. The first would impact tourism, spending, and more, while the second would impact life insurance and potential job discrimination (neither of which are covered under existing genetic information non-discrimination legislation in the United States). Hub-and-spoke privacy networks were suggested as ways to mitigate this risk. In this regard, and within the USA, efforts around genetic non-discrimination legislation have been enacted to protect individuals[1].

A fifth theme was the importance of clear marking of rights associated with data – a transparency and certainty problem. Several scientists noted they were "definitely happy to work with data that is actually in the public domain" but needed to know explicitly about the rights associated. The scientists were quite sensitive about the legal issues - "if the terms are clear enough then we can download the data...if [they are] not clear, we might ask the provider the reality of the terms."

Concluding Thoughts

[1] http://www.nhgri.nih.gov/10002077

It is clear that internationalization of the Sage Commons presents a remarkable set of challenges. The role of human subjects joins the "normal" complexity of international scientific research, the mixture of legal, norms-based and intellectual property constraints, and the significant potential of economic impact of clinical and public health information. But there is at least the early sign-post of a potential solution, which is the application of standard and pre-negotiated terms to stud-ies that are designed, from the start, to be "integrable" into something like a disease commons.