Simons Foundation Draft

“If I have seen further it is by [standing on the shoulders of giants](http://en.wikipedia.org/wiki/Standing_on_the_shoulders_of_giants).” Isaac Newton 1676

This famous quote embodies our goal. National agencies support scientific research. The findings are published in papers, some derived data is dumped into national databases, but often the detailed experimental results holding the raw data are lost. Further the programs that produce the derived data from the raw data are also lost. Imagine a world in which a data repository existed that held the raw (or nearly raw) experimental results, metadata about the meaning of those results (experimenters, location, instruments, conditions etc.), the code to produce derived results, and a method to query for the data and code of a particular paper or across papers. Imagine further that it is possible to run that code on different platforms at different times, possibly on different data or using different parameters. Thus the repository would include not only the raw or derived data, but also the code, scripts, specifications used to process/analyze the data. Querying would enable fusing different data sources and software and executing that software on the data.

Achieving this requires overcoming three obstacles: social (how do we get researchers to share the data and code they took so long to get, especially in a form that others can use), technical (how do we design the metadata formats so that they are meaningful to each subdiscipline and how do we query that data, query the program code, and combine the two), and economic (where to house all of this on some kind of heterogeneous cloud infrastructure, implement the query system, and ensure that the data is queryable by as wide a community as possible). We deal with each in turn.

Social – many journals, especially the best ones, require that data be provided with papers. Thus, for such data at least, one could with the cooperation of journals achieve compliance. Unfortunately, this data is often in a format that only the researchers can read, so in addition to the stick of demanding compliance for an article to be published, we can add the carrot of funding a way to make the data and code available (perhaps paying a few hundred dollars per journal article) based on a short form grant proposal (one page). A human or automated verifier would ensure that the data conformed to the fields required for that subdiscipline using [a](http://www.relaxng.org)a) semantic web format [Juliana, is there one you like?], so that the data of this paper could be queried independently or with data and code from other papers.

Technical – every data type has some common metadata elements such as the experimenters involved, the lab or labs, the instruments (e.g. the features of Dataverse http://thedata.org/book/features, the independent variables, and the dependent ones, but subdisciplines will have additional attributes, e.g. the metadata for microarray experiments differ from mass spec experiments. So the semantic web schema for each kind of experiment would have to be established by domain experts who would act as paid consultants for the project and receive credit as contributors. This goes beyond the query single article interface offered by Dataverse. Once the schema is established, we would then need to develop a way to allow users to both browse and query the data. For this we would use [Juliana, what is a system we could use to query]

Economic – The national agencies see their role as producing science, not really archiving it or making it accessible except through a few databases of highly-refined data (e.g. the results of a sequencing effort). The Simons foundation could fill the gap of making accessible less refined or even raw data, derived data, and the code to go from raw to derive. The amounts of data are vast – potentially 10s of terabytes per day, but the metadata should be fairly small. For this data to be useful, the metadata must be in a common format and make sense to each subdiscipline. For this, the effort would require substantial computing and data infrastructure, domain experts to create the discipline schemas, funds for scientists to make their data accessible through the semantic web, and a team to build and maintain the query system and the system that ensures that code provided can be run on new platforms.