While the advent of data sharing plan submission requirements at the NIH and the NSF is a welcome development, encouraging the reuse of scientific data needs far more policy intervention.

First, Standards should be developed that can be used to grade data sharing plans, so that grant review panels can know both whether or not a specific data sharing plan is satisfactory and so that for any given call for submissions the reviewers have a sense of how important data sharing is versus the scientific goals of the project. Second, data sharing plans should be made public alongside the notices of awards and contact information for the principal investigators, so that both taxpayers and scientists know what promises were made and how to contact a scientist and ask for data under the plan approved.

Third, tracking should be possible to begin to estimate compliance: annual grant review forms should contain fields where the researcher is obliged to place URLs to data shared under the plan (or if left blank, explain why), for example. It should also be easy to create a data request system in which those asking for data send a copy of their request to the grants database, which can then be cross-referenced against the review forms to provide at least a rough estimate of compliance. And fourth, scientists with a record of subpar execution against data sharing plans should be downgraded in their applications for new funding. Taken together, these four elements create an incentive structure that would significantly increase the incentive for scientists to provide public access to the digital data resulting from federally funded research.

In tandem, the funding agencies might develop financial models for the preservation of these digital data in much the same way that models exist for estimating overhead and other baseline costs as a percentage of the grant. This could fund not only new library services and jobs in the research enterprise but also serve as a non dilutive funding source for a new breed of data science startup companies focused on preservation, governance, querying, integration, and access to digital data.

However, we should be careful not to treat data as property by default. Intellectual property is a useful frame through which to view creative works and inventions in science, as well as to protect valuable “marks” and secrets. But in the United States at least, data is typically in the public domain already, and therefore the extension of intellectual property rights to it would represent a vast expansion of rights in a space where there is zero empirical evidence that it is needed.

Typically data is treated more as a secret, which is at odds with the public nature of the idea of data access, and the obstacles to data sharing are less legal than they are professional and economic. The ugly reality is that sharing data represents a net economic loss in the eyes of many researchers: it takes time and effort to make the data useful to third parties (through annotation and metadata) and that is time that could be spent exploiting the data to make new discoveries. On top of this, there is a twin incentive problem. Scientists see no benefit to sharing data and are not punished if they fail to share data, while there is a pervasive fear that other scientists will “scoop” them if their data are available before being fully explored. This creates a collective action problem that can be overcome most easily by clear funder policy as enumerated above: data sharing plan mandates with transparency, accountability, tracking, and impact on future funding.

One policy action that would be very welcome would be an unambiguous signal that publicly funded science data is in the public domain worldwide, not just in the United States. This could be accomplished either through the use of a copyright waiver, such as the Creative Commons Zero tool, or through other means. But it is vital to make it unambiguous and clear when and where data are free to reuse, because applying conditions imported from creative works and inventions to a class of information that is fundamentally far less like "property" can have serious unintended consequences. Easily imaginable consequences include vast cascades of attribution requirements, so that a query to 40,000 data sets requires 40,000 attributions – every time – or worse, the poisoning of data for use in job creation by small companies who wish to build atop data as a platform or infrastructure.

The intellectual property status of data does differ across the scholarly disciplines and its own status in how far it's been processed. Some sciences rely on inherently copyrightable "containers" for data, from field books to recordings to photographs. And raw data converted to beautiful information by visualizations will touch on copyright. Policy should be flexible enough to account for this, but start with a default bias that public domain data is the most reusable, while providing "opt-out" capacity for data and disciplines where the public domain is simply not the best solution.

There is an obvious problem with this set of policy recommendations. They rely on money to work. We do not yet know the true costs of storing digital data over the same time frames that we store the scholarly literature. As our capacity to generate data explodes, we must invest at the same time in our capacity to steward it. Research projects into large data information science should be a priority, with specific attention paid to when and where it is possible to compress data, move data to secure "cold storage", jettison data (either because it is duplicative, or because it can be regenerated again later) , and more. We do not have the sociotechnical infrastructure required to answer questions of data stewardship with any authority, and we must create it on the fly at the same moment that the data creation burden is hitting exponential heights.

Solving these stewardship problems might be best achieved through a coalition of research institutions, the library community, publishers, and funders. Taken together these groups already heavily regulate the daily life of a federally funded scientist. It is a small extension to imagine leveraging that regulatory power to provide new services to the scientist – a university and its library might keep an archive of standard data sharing plans, standard budget items to implement, which together would take the guesswork out of filing and operating a data sharing plan. Even better would be a federal program to certify a small number of such plans for each discipline.

Missing from the set of stakeholders mentioned in the RFI is, notably, the business community, both the large scientific companies and the vast potential of startup firms. In an ideal world, the stewardship conversation will bring in actors from those industries, from pharma to venture capital, as we are missing an entire professional class of data stewards and data engineers (not just data scientists) who could serve the needs of the research enterprise while creating stable. Even better, because the data stewards must be close to the researchers to serve them, these jobs are less likely to move offshore. An investment in small business grants, job training (and retraining) vouchers, and the creation of

community college pedagogy for data stewardship functions could go a long way towards stimulating the emergence of this professional class.

In order to stimulate the interaction among these stakeholders and the emergence of a new class of data stewardship jobs, agencies could take additional steps to stimulate use of data. Contests are one obvious route, where a prize is posted in return for solving a problem (or simply for coming up with innovative ideas and/or applications that run on government data). Another route is the expansion of SBIR grants to create a track focused specifically on data startups, which lower the risk of company formation and job creation as well as creating non-dilutive funding sources for entrepreneurs.

A route that is vital, but less obvious, is investment in and commitment to the emergence of standards that enable interoperability of, and thus reuse of, digital data. Standards lie at the heart of the Internet and the World Wide Web, and together lower the cost of failure to such a low point that companies built on the web and the internet can begin in garages. Such is not the case in the sciences. And it will not spontaneously emerge, even if data flow onto the web. As long as those data are in a tower of babel of formats, incoherent names, and might move about every day, they will be a slippery surface on which to build value and create jobs. Federal policy could call for a standard method for providing names and descriptions both for digital data and for the entities represented in digital data, like the proposed standard of the Shared Names project at http://sharedname.org .

Standards also make it far easier to provide credit back to scientists who make data available, as well as increasing the odds that a user gets enough value from data to decide to give credit back. Embracing a standard identifier system for data posters will make it easier to link back unambiguously to a researcher as well as to make it easier for grant review committees and universities to receive a full picture of a scientist's impact, not just their publication list.

Standards for Interoperability, Re-Use and Re-Purposing

About me:

I am a Senior Fellow at the Kauffman Foundation, the Group D Commons Leader at Sage Bionetworks, and a Research Fellow at Lybba. I've worked at Harvard Law School, MIT's Computer Science and Artificial Intelligence Laboratory, the World Wide Web Consortium, the US House of Representatives, and Creative Commons. I also started a bioinformatics company called Incellico, which is now part of Selventa. I sit on the Board of Directors for Sage Bionetworks, iCommons, and 1DegreeBio, as well as the Advisory Board for Boundless Learning and Genomera. I have been creating and funding jobs since 1999.