Myles Axton PhD
Editor, Nature Genetics
Nature Publishing Group
New York NY

Preservation, Discoverability, and Access

(1) Metrics and reporting reduction. If public data can be shown to produce measurable results, then scale up the promotion of data access policies and implementation. If trusting people to maximize their gain from data generated with federal funds works better, let them do it and get government out of the way.

(2) Make the granularity of attribution of data to individuals, grants and teams easily discoverable. Datasets are often submitted by one informatics professional or grantholder but are the product of a team effort. Perhaps patenting could be streamlined to protect data registered in publicly accessible repositories (either as IP or precompetitive).

(3) Researchers in each of the fields will always play up the differences. In practice the most generic data storage with the fullest and most standardized metadata will always be best. Re-use of data from one field to the next sometimes requires handshaking workshops to identify formatting and quality issues and to set standards and formats: (http://www.nature.com/ng/journal/v42/n1/full/ng0110-1.html)

(4) The effort to develop metrics should precede any attempts to enforce data access. The game may not be worth the candle. Public datasets may be only trial runs or burn-in for the datasets and technologies that follow.

(5) Use unique citable identifiers (UUID, DOI, URI, ORCID) for individuals, roles, grants, datasets, samples, departments, institutes, funders and projects. Have unique identifiers for everything that can be shared and an access plan for everything. Where possible the unique identifier should be citable whether or not the data are accessible. Use successful simple templates for data management plans in a journal, database or public repository rather than reinvent them from scratch in a private repository.

(6) There is a metrics requirement here, especially in reallocating resources for IT infrastructure and curation to those projects that justify it. Funding for consortia and for the publishers that present their data and publications to the public would represent a step in funding open access in business models that differ from the current one where the author pays article charges and funders subsidise the publisher. Transparently accounted curation services operated by publishers are a possible alternative to publicly-funded bodies such as NLM (http://www.nature.com/ng/journal/v43/n5/pdf/ng.827.pdf).

(7) Funding agencies will always try compliance approaches first. These are deadening and turn research reporting into a cheating game. Standardization (10, 11) and metrics (1,2,4,5,6) may be more helpful. Rewarding data sharing consortia or defined communties with extra funding for existing grants that are still live - in response to high re-use in substantial secondary publications by other data users – should be tested to see if it will encourage pre-publication data sharing. Minimization of reporting for grantees can be done by engaging publishers who already deposit papers in PubMed Central to help with reporting standards.

(8) The Million Veterans Project should be given help to overcome institutional barriers to become a national cohort for healthcare research and translational improvement via the Veterans' administration. An open interface converting self-reported experiences to medical ontology modeled on Patientslikeme.com would help with recruitment and coordination.

(9) Microattribution (http://www.nature.com/ng/journal/v41/n10/full/ng1009-1045.html) based on ORCID is a central tenet of the drive to data citation via attribution credit. The Datacite inititative is another example that may be useful. I think it is a mistake to have PubMed as the central reputation server

(http://www.nature.com/ng/journal/v41/n4/full/ng0409-383.html), rather standard attribution formats should be used openly with each provider (journal, database, institute, researcher) offering to display attribution credit for the items it holds.

(10) Format datasets in a restricted set of interoperable formats (http://www.nature.com/ng/journal/v43/n1/full/ng0111-1.html)and standardize metadata that contains field-specific reporting standards (Example: http://isatab.sourceforge.net/tools.html).

(11) MIAME: GEO made deposition easy, ArrayExpress made formatting and compliance part of deposition at the price of deterring submissions. The existence of standards and their enforcement does not have the desired result and other incentives are needed (http://www.nature.com/ng/journal/v41/n2/full/ng0209-135.html).
GWAS: A user community, funder (NHGRI) and Nature Genetics decided that replication and correction for multiple testing and stratification would make the technique mor robust to false positives (http://www.nature.com/nature/journal/v447/n7145/full/447655a.html).
ORCID: Thomson Reuter was persuaded Researcher ID would work better if shared
PDF: a proprietary format from Adobe can be replaced by HTML5

(12) Engage with Datacite and with international publishing intitatives (CrossMark from CrossRef) and publishers who get the point (Nature, BMC and PLoS).

(13) Universal versioned DOIs or other persistent granular electronic identifiers. We also need a convention on bidirectional linking as well as technology to make it easy.

Myles Axton, Ph.D.
Editor
Nature Genetics
http://www.nature.com/naturegenetics
