Project I)

Transparency through the Leadership of Publishers

Myles Axton
Editor of *Nature Genetics*

What matters about research papers is utility to others: you cannot publish thirty papers in a year on one dataset, so your reputation is made in the labs of others. This observation is the basis for journals mandating that data, analytical decisions, methods and materials be accessible not only to referees but to readers. With the incentive of utility, journals can partner with and multiply community initiatives such as data commoning, contest challenges, community curation initiatives and open peer review pipelines. In return, we expect to receive publications not only of great novelty and conceptual advance but publications that reflect the analytical decisions and datasets more closely and therefore have greater resource value and reproducibility.

In the Scientific Data initiative at NPG, we are addressing the timing, formatting and attribution of data by introducing versionable, citable data descriptors. These “User guides to the dataset” complement research publications and can lay out data use conditions, caveats and community standards for analysis and re-use, aid data formatting, recruit collaborators and publishers and act as advance publicity for research publications. The descriptors complement community database accessions by providing extensive details of freezes of multidimensional datasets, metadata in standard (for example ISA-Tab) format, tabular data exchange formats and metrics on standardized provenance such as funding and contributor roles.

Access to papers and data are the very minimum requirement for transparency. Editors deal daily with standard methodologies even when handling experimental designs of considerable novelty and even new methodologies face commonly recurrent issues of overfitting, statistical power, randomization, protection from bias, significance thresholds, false discovery rates, and correction for multiple hypothesis testing. Reporting of elements of methodology and analytical design will need to be improved across journals by ensuring that referees and readers alike have access to standard details of statistical and experimental design. Detailed protocols can be published in our Protocols Exchange. Provenance needs particular attention if journals are to incentivize good collaborative and stewardship practices. This means recording the granularity of author contributions within consortia, including roles of curators and clinicians and proper acknowledgement of funding sources.

We would like to partner with and incentivize demonstration projects that have research utility and new biological insights. In particular, we have commissioned a small number of community annotation projects on genes and pathways (microattribution reviews http://www.nature.com/ng/journal/v43/n4/full/ng.785.html) using the Leiden Open Variation Database (LOVD) and provenanced RDF nanopublications.

We are also open to the idea of exploring the granularity, precision and degree of openness of supervised peer review of datasets, software and analytical pipelines on the Synapse platform (http://www.nature.com/ng/journal/v44/n2/full/ng.1089.html). We see the progressive improvement of research by peer review and the incentive of journal publication as a natural complement to transparent community platforms and open innovation challenges. We can use what we do best to improve one another’s work.

We are also searching for an answer as to how fast to plunge into the semantic publishing future. Everything should have a unique accession number to reduce redundancy and increase machine reasonability of scientific output. But is this aim better served by editing manuscripts for simplicity and natural language search or by building in semantic precision to our databases and journals?

What do we want from the community?

Community standards

Over to You

I- Transparency through the Leadership of Publishers

G	Organization	Last name	First
I	Nature	Aschheim	Kathy
I-Lead	Nature Genetics	Axton	Myles
I-Anchor	Sage Bionetworks	Bot	Brian
I	PLOS	Chodacki	John
I	UCSC	Hospodor	Andy
I-Anchor	Sage Bionetworks	Huang	Erich
I	Harvard Medical School	McCray	Alexa
I	NIH	McEwan	Jean
I	Moore	Mentzel	Chris
I	O'Reilly Media	Oram	Andy
I	NSF	Strawn	George
I	Pipsqueak Productions, LLC / Ushahidi	Werby	Olga
I	George Mason University	Yang	Tony

Publishers Enable Transparency
Myles Axton, Nature Publishing Group

Project Overview- Major Themes

Citable data descriptors
Data metrics
Data management plans refereed in grants
Versioning
Freezes from databases
Harmonization of metadata and formatting

Community annotation
Continuous curation and publication
Microattribution
Nanopublication
Genome browser (UCSC) and Synapse enabled for commissioning (deadlines, contacts)

Pipeline peer review
Synapse enabled for recording referee comments
Analysis papers associated with Challenges
Harmonizing pipeline refereeing tools across publishers

Potential alignment with existing Commons' approaches

List examples from other efforts that could be applied to this project.
Examples around Governance, Incentives, Platform, etc.

Data descriptors
TCGA Pan-cancer analyses

Community annotation
BRIDGE (patient driven microattribution reviews)

Pipeline peer review
Challenges
TCGA Pan-cancer analyses
All projects in Synapse

Unmet needs and issues

Potential commons approaches needed but not yet built and remaining issues to be highlighted

Project management tools

External interface for Synapse

- Deadlines
- Corresponding author
- Journal commissioning commitments
- CrossMark (leave to publishers?)

Pipeline peer review

Export format

- Tabular view of pipeline (Cf. phylogenetic trees, is there a programming convention – eg. tabs?)

Cloning and franchising Synapse (assume all publishers will need one)

Career metrics

Pilot projects to recognize data activities as well as publications (Vivo, deans)

Bioresource impact factor/ escrow anonymity/ research subject and support group track use

1-year vision for the future of this project

1) Nature's Scientific Data to commission data descriptors from TCGA and Sage.
2) Nature Genetics to peer review on Synapse an Analysis paper arising from a Challenge.
3) BRIDGE to interface with Genetic Alliance and Human Variome Project to propose a Microattribution Review for Nature Genetics and/or Human Mutation using Synapse as the platform (entire Fanconi Anemia pathway FANCA-J might be a possibility).
4) Nature to consider whether either Research Journals or Scientific Data will need their own Synapse or GitHub to maintain community expectations of granular and transparent peer review as well as transparency and reproducibility of the data resources and code.
5) Long term view: will databases become publication platforms? Will old publishers adapt or be replaced?