Tim Clark's Statement of Use Cases

There are two possible use cases for authors citing data, from the publisher point of view.

(1) The author is presenting a set of conclusions based on HIS/HER OWN DATA, not yet archived.

Here the dataset is OUTPUT of the authors' research process and is NEW to the archives.

==> Publisher says: go get an accession number from XYZ database, we'll check the data is there, and then we'll cite it.
==> Publisher needs:

- a list of databases they are comfortable recommending to authors;
- some sort of labels or decision tree to pick the right one(s);
- a default for when nothing else fits;
- a schema supporting data citation e.g., JATS;
- a translation from the schema into their journal reference style.

See: Nature scientific data & F1000 Research workflows. http://bit.ly/1Pty5CB

Typically this means the data is produced directly from an experiment or set of observations.

(2) The author used the data already in some archive, as INPUT to analysis or meta-analysis

Here the dataset is INPUT of the authors' research process and is ALREADY IN the archives.

==> Publisher requests accession numbers for datasets used as input; and if there is an OUTPUT dataset as well, request that be archived to, and have a NEW accession number.

==> Publisher needs:

- only the accessions for the INPUT data
- same things as above for any new OUTPUT data produced

Typically this means the data is produced by meta-analysis of pre-existing data.

[PROPOSAL]

I have proposed and emphasized that the FIRST USE CASE is the one we want to focus on initially. That is because it is the driver to get data into archives in some reasonably uniform way. It is also the simplest use case. I would think doing the second use case is unrealistic for a one-year project. Could easily be a follow-on if we ask for a second year after a successful first year.

Tim

From an email exchange between Tim, Maryann Martone, and Joan Starr on November 6, 2015