nature genetics

Data producers deserve citation credit

Datasets released to public databases in advance of (or with) research publications should be given digital object identifiers to allow databases and journals to give quantitative citation credit to the data producers and curators.

Data are always generated within a context. They are generated for a particular purpose, under particular hypotheses and, above all, by a particular group of producers. Funding bodies and journals insist upon broad data release to maximize the resource value of the information. But data still remain someone's life work to be bartered in an economy of knowledge production. The value of research publications is currently acknowledged by citation. If this practice of citation is extended to datasets, these datasets and their producers will be properly recognized. Some data stimulate research by other users and some datasets remain unused commodities. Knowledge of which datasets are useful will help with redirecting funding and research efforts.

Recent opinions on prepublication data release (*Nature* **461**, 168–170, 2009) and on access to materials at time of publication (*Nature* **461**, 171–173, 2009) proposed citation of unique data identifiers as a way to give appropriate credit to data producers. This idea was first raised in this journal by Anne Cambon-Thomsen (*Nat. Genet.* **34**, 25–26, 2003), who emphasized the ethical necessity for citation and tracking of unique resource identifiers. Citation of database entries is now widely seen as appropriate for large resource projects as well as for distributed projects such as collecting gene variants from clinics. More recently, we have discussed citation credit in a number of editorials (*Nat. Genet.* **39**, 931, 2007) and blog posts.

Generators of large resource projects such as the Wellcome Trust Case Control Consortium (https://www.wtccc.org.uk/), the 1000 Genomes (http://www.1000genomes.org/) and the International Cancer Genome Consortium (http://www.icgc.org/) stand to receive many citations of their papers by users of the data they are releasing. Inevitably, any group adopting very early data release will have discussed imposing use conditions that might limit competing publications until publication of the first global analyses by the consortium. We predict that any use restrictions will be counterproductive because they will limit citation and utility of the authors' own work and will be unenforceable by journals. In practice, data producers often have plans for their own data well beyond the period of any embargo, and they also expect data users to continue to avoid competing with them on declared or obvious large-scale analyses. One problem with imposing use restrictions specifying arbitrary units of time and data is that data production evolves in magnitude and type; by the time a large project publishes its analysis, the state of the art is usually a larger or different type of dataset. Indeed, some large resource projects may never be associated with high-impact papers, but if data are citable from the time of release, the datasets can become citation classics in their own right.

What form should citable data identifiers take? They must work with existing unique resource identifier conventions and with the existing well-funded stable repositories used by research communities. However, these identifiers are not just for locating data but are for stably identifying the data units and versions with particular data producers, curators, funders and affiliations in a citable form. Because publications are currently the main source of scientific credit and because publishers have already developed citable digital object identifiers (DOI), it would seem to be their opportunity to grasp or to fumble. We propose citing DOIs that tag a combination of repository, database, accession, version, contributor and funder.

Of course, precise citation of all research output represents the bare minimum of respect for colleagues and competitors. This journal also endorses communication between data producers and data users. Whereas it is impossible for journals to restrict the use of data already in the public domain, we can show evidence of communication between producers and users to referees. Many funders of large resource projects now require a data release policy and plan for global analysis by the data producers. These parts of the successfully refereed grant should be published as a 'marker paper' or deposited in a citable preprint archive such as *Nature Precedings*. At very least, the details of the producers' work and intents should be available to users in a citable form in the database holding the data. Data users can submit an email demonstrating that they have contacted the data producers with their plan for use of the data and showing that they have read the producers' data release policy, conditions and plan for analysis.