nature genetics

Crowdsourcing human mutations

The first Human Variome microattribution review shows that data citation and publication credit can work as incentives for systematic curation of gene variant and phenotype data. Analysis of the formal assertions in both databases and journal articles argues for better separation of data structures from narrative so that they can better support one another to communicate meaning.

In a previous editorial (*What is the Human Variome Project? Nat. Genet.* **39**, 423, 2007), we suggested that the global project of collecting all the phenotypically interesting variants of the human genome could use microattribution, a form of data citation by journals and databases whereby submitters and curators receive citation credit that is linked to the unique identifiers of each report of a gene variant, allele frequency, population origin or tabular phenotypic data. On page 295 of this issue, George Patrinos and colleagues now show that a process of community annotation of rare and common variants influencing hemoglobin levels can bring unpublished variants into the public domain. Credit for contributors operates at two timescales to incentivize the process, with immediate microattribution provided by the database and periodic publication credit provided by participating journals and preprint servers. Following the curation process, their systematic analysis generates new knowledge and new hypotheses for the field as well as many corroborating reports of variants that support existing observations. The microattribution process has a further purpose. It reports the level of activity across the genome contributed by collectors of genotype and phenotype information so that resource decisions can be made on the needs for extra curation or timely community re-review of a locus, trait or pathway.

Examining the same datasets, Barend Mons and colleagues (pp. 281–283) comment on the relative difficulty of extracting and coding into XML large numbers of semantic multiples taking the form [subject] [predicate] [object] (e.g., [gene variant] [has] [frequency]), together with provenance information such as the database or publication context and contributors and contributor role identifiers. Because of the difficulty of extracting unambiguous identifiers and the logical connections between them from narrative articles, they recommend that it is easier and more accurate to create semantic assertions by publishing in databases first, with complementary journal articles to follow.

The current culture of measuring citation of entire articles alone tends to favor the articles describing techniques and resources or long-standing hypotheses with lasting predictive value. It may even contribute to the fixation of initially tentative hypotheses by repetition rather than by replication. The prospect of having citation metrics for semantic assertions is exciting because it might now be possible to recognize and reward efficient hypothesis construction, testing and rapid falsification. These complementary skills are particularly valuable to researchers when they navigate new conceptual territory.

The two articles in this issue taken together are something of a wake-up call for publishers who, ever since the web became the platform of choice, have been creating ever more baroque decoration of journal articles with hypertext links pointing anywhere but toward the heart of the science the authors are relating. If this practice is not actually making the articles richer in meaning (and it can certainly clutter the reading experience to the point where many readers take the article offline to avoid distraction), then it would be better to separate the semantic content into databases and tables of formal multiples, construct our hypotheses and results from those and weave a sparser narrative for what remains.

The problem of identifying which of the hordes of well-formed unambiguous assertions are meaningful to geneticists seems smaller than the problem of finding such assertions buried in print. It is time both to say what we mean and mean what we say. ■