## Structured digital abstract makes text mining easy

SIR — Your Editorial "The database revolution" (*Nature* **445,** 229–230; 2007) highlighted the difficulty in maintaining a stable information architecture for biology — in terms of both funding it consistently and evolving a common format.

In addition to the suggestions you made, we urge journals to take the lead in making articles suitable for digital parsing and text mining by providing a structured digital abstract (M. R. Seringhaus & M. B. Gerstein *BMC Bioinformatics* **8,** 17; 2007).

The distinction between journals and databases is blurring. The results published in journal articles of new structures, genome sequences and microarray experiments are automatically deposited to large databases, while the articles themselves in these disciplines are largely accessed in electronic form via PubMed queries. In the future, the text of articles will be systematically mined by computer programs, allowing interrelation of journal text with the vast repository of knowledge stored in databases. But making these interconnections now is challenging. With few exceptions, the facts published in journals are not in a format easily parsed by computer: in particular, text mining has difficulties linking names to database objects, and identifying key findings from the language of a paper.

The structured abstract would act as a gateway for text-mining engines to access an article, much as the traditional abstract now does for readers. The structured abstract consists of three main elements. First is a translation table or 'cast of characters', which lists all named genes, proteins, metabolites or other objects in the article, and relates their human-readable names to precise database identifiers. Second is a list of the main results described in simple ontologies using controlled vocabulary — for example, interactions ('protein A binds to protein B'), phenotypes ('mutation C suppresses deletion D'), and protein modifications ('protein E is phosphorylated at residue F by protein kinase G'). Third is standard evidence codes for how the results were obtained — for example, 'affinity purification' or 'mass spectrometry'. Thus the structured abstract is not only a synopsis of the results but is readily computer-readable.

Such digital summaries could be produced by authors and editors as part of the editorial process, subject to peer-review and copy editing. They could be published on journals' websites, using semantic web standards such as XML and OWL, and indexed by central repositories for fast look-up.

Adoption of the structured abstract would require action by scientists and editors to establish formats and vocabularies, as was done for Gene Ontology (*Nature Genet.* **25,** 25–29; 2000). Early incorporation by a few journals or a single community — for example, yeast researchers — could provide a prototype before it enters widespread use.

**Mark Gerstein*†, Michael Seringhaus†, Stanley Fields‡**
*Program in Computational Biology and Bioinformatics, Yale University, PO Box 208114, New Haven, Connecticut 06520-8114, USA
†Department of Molecular Biophysics and Biochemistry, Yale University, PO Box 208114, New Haven, Connecticut 06520-8114, USA
‡Howard Hughes Medical Institute and Departments of Genome Sciences and Medicine, University of Washington, Box 355065, Seattle, Washington 98195, USA

**Readers are welcome to comment at http://blogs.nature.com/nautilus/2007/05/making_names_and_descriptions.html**

## Human reference sequence makes sense of names

SIR — Most journals, including *Nature*, require authors to annotate a new entity (a gene, protein or loci, for example) with references to a standard database. However, journals do not require references to standard databases for discoveries of functions or diseases associated with previously defined genes. Since most genes have more than one name, and many gene names refer to more than one gene, the choice of a name without reference to a common or standard database can inhibit the integration of results from transcriptomics, population studies or comparative genomics.

In this post-genomic era, researchers have to be able to make associations among many genes, which requires being able to correctly identify a gene and all its synonyms. The most obvious way to ensure this would be for journals to insist that genes in a publication should be identified with reference to the Human RefSeq (see www.ncbi.nlm.nih.gov/RefSeq). In this way, genomic analyses are more likely to identify genes of common interest.

**Douglas L. Crawford**
Marine Genomics, Rosenstiel School of Marine Sciences and Atmospheric Sciences, University of Miami, 4600 Rickenbacker Causeway, Miami, Florida 33149, USA

## Codes must be updated so that names are known to all

SIR — Sandra Knapp and colleagues, in their Commentary article "Spreading the word" (*Nature* **446,** 261–262; 2007), stop short of urging the radical steps required to effectively transform nomenclature and access to plant and animal names.

Some important and necessary steps have been made towards opening access to existing literature, by efforts such as AnimalBase (www.animalbase.de), Cornell University's Core Historical Literature of Agriculture (chla.library.cornell.edu) and the Biodiversity Heritage Library (bhl.si.edu). But the name-access problem remains, and there is no excuse for enlarging it with each passing year.

Immediate and mandatory registration of names should be adopted as an emergency measure by the International Commission on Zoological Nomenclature (ICZN) and by the International Code of Botanical Nomenclature (ICBN). It is irresponsible, in a world so dependent upon reliable information, to permit 25,000 new names to be introduced each year, with no requirement for them to be universally known and accessible. A registry such as the proposed ZooBank (A. Polaszek *et al. Nature* **437,** 477; 2005) can only ensure that names 'available' under the codes are truly available.

We would strongly oppose any measure that was prohibitive or that imposed censorship.

We urge the relevant botanical and zoological bodies to make three immediate, decisive amendments to the codes. First, require such registration before a name is formally available for use. Second, require full text descriptions of species to be deposited by publishers or authors in a central, publicly open 'bank', free of charge, such as will be provided by ZooBank for zoological names (A. Polaszek *et al. Bull. Zool. Nom.* **62,** 210–220; 2005). And third, require electronic publications to include a 'hot' link to these banks of names and descriptions. This will ensure precision in reference to names.

At the same time, we would urge those bodies to work with publishers to institute an electronic counter that notes every e-publication that mentions, or links to, a scientific name. In this way, each reference to a species would count as the equivalent of a citation, and circumvent the serious problems imposed upon taxonomy by current citation indices such as the impact factor (F. T. Krell, *Nature* **415,** 957; 2002).

**Quentin D. Wheeler*, Frank T. Krell†**
*International Institute for Species Exploration, Arizona State University, PO Box 876505, Tempe, Arizona 85287-6505, USA
†Department of Zoology, Denver Museum of Nature and Science, 2001 Colorado Boulevard, Denver, Colorado 80205-5798, USA

**Contributions to Correspondence may be submitted to correspondence@nature.com. They should be no longer than 500 words, and ideally shorter. Published contributions are edited.**