Credit where credit is due

One of the defining features of scholarly writing is careful citation of the source from which a fact, opinion or method is drawn. There are several different reasons for doing this (see the older blog, *Look Ma, no giants!*) but one moral reason is simply to give credit to the person who did the work. The mechanism for giving credit has, until very recently, been straightforward - one writes the fact and then cites the source a publication like this (Frankenstein, V, 1818) or perhaps like this[1]. At the end of the paper or book, all of the references are listed in alphabetical or numerical order, with all authors, date of work, title of work, title of journal or book, volume and page numbers being written out in full. Nowadays, a digital object identifier (doi) may also be given, for a paper that is available in an electronic version of a research journal or for a e-book. The point is that a reader should be able to go back to the original work.

This system worked very well in the era of paper publications, and works well today for citations of e-paper and e-books because they work just the same way. But in many fields of current research, research papers and books are no longer the main source of published facts; indeed they may not even be the major source. A great deal of biological information, especially the kind that consists of a mass of details, such as gene sequences, protein structures, drug activities, patient blood analyses etc., exists not in the form of individual research publications but as entries in large searchable databases. This creates a problem for citation. A common way of addressing the problem is for the main creators of a database to write a peer-reviewed paper that describes the database, and for everyone who uses the database to cite that paper. That approach "works" in the sense that it seems to obey all the rules and fits the format of journals, but it brings with it a serious problem. Usually, the people who create and run databases are specialists in programming and data curation; they are not the people who generate the biological data itself. Thus someone who uses the pharmacological database run from this lab - www.guidetopharmacology.org - to identify the target for a drug and cites our latest paper describing the database will in fact be crediting mainly a bunch of curators and programmers with the discovery, and not the medicinal chemists who actually discovered the drug-target relationship. This may get us a vast number of citations (see *Mene, mene, tekel... (Part 2)*) but it is hardly fair on the wet-lab discoverers.

Apportioning credit through citation is not just a matter of courtesy. The number of times a scientist's work is cited is often used as a metric to measure their effectiveness and can influence

appointments or promotions: I am not defending this practice here, but simply reporting it. Thus a scientist who spends most of her time generating data that goes into a large database, rather than into conventional research journal publications, will seem on a citation count to have achieved very little. Worse, even if someone did try to cite that scientist's contribution by making a citation to the URL of the precise database location in which her data lay, the unconventional citation would not be understood and 'counted' by the automated systems that generate reports on how often every scientist's work gets cited (again, I am not defending the use of such systems, just reporting their existence).

So, something new has to be done. And, thanks to some original thinking by my friend and colleague Peter Buneman, something has been done. Peter has spent a very long time thinking about both the structures of databases (see his Wikipedia entry, in 'Links' below) and also the uses to which they are put, and has recently devoted a lot of attention to the matter of citation and attribution of credit. He came up with an interesting proposal to solve the problems described above, and approached me and my colleagues who work on the Guide to Pharmacology database to ask whether we would be willing to work with him to use this database to demonstrate the solution. Of course, we said yes!

Peter proposed that, if automated systems that measure scientific impact can only cope with things that look like papers, there will have to be things that look like papers that will give correct credit. In association with Edinburgh University Library, we therefore created a new electronic "journal", called "*IUPHAR/ BPS Guide to Pharmacology CITE*". This is not a normal journal to which anyone can send articles; instead, it publishes only 'articles' placed there by database curators. These 'articles' are really abstracts of data entries, in which all contributors to an entry in the database are listed (with their permission) as authors, there is a title that describes the database entry and that always ends "...in the IUPHAR/ BPS Guide to Pharmacology Database", and there is an abstract to describe the entry. Then there is a very brief main text of the 'article', that is the same for all articles and essentially explains that what is being read is intended to be read by computers rather than humans, and briefly explains why the 'journal' and its 'articles' exist, and directs any human readers to the database itself. The 'article' then finishes with a list of all relevant database links and also with references to all (conventional) articles relevant to the database entry. The critical point is that the journal is a real journal from a properly respected publisher (University of Edinburgh) and its 'articles' have exactly the structure that automated readers of academic articles expect.

People wanting to cite a specific entry in the database can cite one of these articles; anyone who uses the database will be told how if they click a "how to cite this page" link. The articles themselves are picked up properly by 'citation scrapers' such as Scholar Google, which means that the authors of papers cited by the database entries are now properly credited, and also the authors of the database entries are properly credited too. Brownie points thus accrue, in the automated assessment systems, to the people who deserve them. Contributing to the database is no longer a less career-enhancing activity (in terms of these citation metrics) than writing papers.

There are some complications I have not discussed above. For example, when the same group of people contribute to database entries on a whole family of molecules, we write the 'articles' for the family not the individual entries, and there are subtleties about how decisions like this are made. We describe the (mathematical) logic employed in this sort of decision in full in a peer-reviewed paper that has just been accepted for the (conventional) journal *Database* (see links).

This is in a sense just an experiment. So far, it applies to only one database, although in principle it could apply to almost any similar database, with a suitable new 'journal' to carry the citations. Indeed, if anyone reading this is running a database and looking for a similar solution, please do contact me: we may well be able to help.

Links

Wikipedia page on Peter Buneman - https://en.wikipedia.org/wiki/Peter_Buneman

Our article in Database: