

Nanopublications*: the future of coping with information overload.

Jan Velterop

Concept Web Alliance – Academic Concept Knowledge Limited (ACKnowledge)

Scientific information is for the most part cumulative. Successive generations each have more information than ever upon which to further build their knowledge and understanding. However, having access to a relentlessly growing body of information clearly has its problems. The notion of information overload is well known.

In a recent article in the British Medical Journal, Alan Fraser and Frank Dunstan¹ call for new strategies to deal with information overload. They observe that a new entrant in the field of diagnostic imaging in cardiology would need a total of some 19 years and 202 days to read all the relevant literature, assuming he or she were to devote their entire normal working time to reading (5 articles an hour, 8 hours a day, 5 days a week, 50 weeks a year). That, clearly, is practically impossible. It does mean, however, that nobody can have a comprehensive overview of the available information about a certain topic. Not only that, but the cherished idea of shared knowledge between scientists is also an illusion. Fraser and Dunstan calculate the chance that a cardiac imaging specialist on a somewhat more realistic diet of one paper a day will read a particular paper is only 1 in 8.9. And the chance that a colleague elsewhere will read the same paper in a given year as only 1 in 79.

This bleak picture is not exclusive to medical science. Areas such as genomics, proteomics, chemistry, and likely many others suffer the same fate. In everyday reality it means that conclusions are drawn from, and decisions made on the basis of, very incomplete information. As Fraser and Dunstan argue, it makes evidence-based medicine nigh impossible. In their phrase, “being expert means knowing and publicly acknowledging the limits of your ignorance”, by which they presumably mean the ‘limits of your knowledge’.

Staying in the medical science sphere, in an interesting blog post, Annabel Bentley² draws attention to a paper by Hilda Bastian, Paul Glasziou, and Iain Chalmers³ on the difficulty, or rather, impossibility, of keeping up with 75 new published trials and 11 new systematic reviews every day.

Although information overload is not an entirely new phenomenon, the scale it has now reached has no precedent. It makes dealing sensibly with that information – even with just the most relevant bits – entirely unmanageable if new methods to ingest the essence, or sophisticated filtering systems, do not become available.

¹ Fraser A G and Dunstan F D (2010) On the Impossibility of being expert. BMJ 2010; 341:c6815 doi: 10.1136/bmj.c6815

² Annabel Bentley (Blog Post) <http://blogs.bmj.com/bmj/2010/11/03/annabel-bentley-information-overload-are-you-waving-or-drowning/>

³ Bastian H, Glasziou P, Chalmers I (2010) Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up? PLoS Med 7(9): e1000326. doi:10.1371/journal.pmed.1000326

A solution that is often put forward is to limit the publication of new information. Bastian et al. mention that “unnecessary trials need to be reduced.” Well, yes, they need to be, and not just because they add to the information overload. But it is hard to see how just reducing ‘unnecessary trials’ (supposing we could make that judgment) would solve the problem. Reducing publication of research results, or – worse – reducing the generation of data, cannot be a solution. That would be against the very nature of science.

What is likely to happen, though, is a separation of the traditional purposes of publishing, which have hitherto been intertwined: keeping the record on the one hand, and disseminating knowledge on the other. For record keeping (I like to call it “keeping the minutes of science”⁴), reducing the number of articles published is not needed, as long as the record is there, securely curated, and relatively easily accessible, so that when and if desired, the ‘minutes’ can be consulted. Neither is reducing the number of articles necessary for knowledge dissemination, as long as methods can be developed to ingest the overwhelming amount of knowledge contained in them in such a way that the essence is captured as well as any mutual interdependence and cohesion between knowledge elements.

In a position paper Barend Mons and I⁵ discuss ‘nanopublications’⁶ as a method with which to extract assertions from published text as well as data collections, and subsequently using them to build an overall picture of the state and the dynamic development of knowledge in a given field. In this method, the extracted (“text-mined”) assertions are typically triples, of the form ‘object–predicate–subject’. In order to become a ‘nanopublication’, an assertion is given metadata such as author(s), date stamp, journal, doi, non-journal source, form of peer-review (or none), condition under which it is valid (e.g. at room temperature), *et cetera*. This makes the nanopublication citable and provides the connection back to the full-text publication (the ‘minutes’) from which it comes. The terminological inconsistency of a collection of nanopublications is almost always substantial, often vast. For that reason, nanopublications are subsequently semantically ‘normalised’, so that the meaning of the underlying concepts denoted by the object–predicate–subject triples is disambiguated.

Furthermore, most assertions are made many times, resulting in much redundancy. That means that the number of actual assertions being made in a collection of nanopublications is in reality considerably less. An average of even a thousand times less is estimated on the basis of a cursory analysis of the body of some eight million PubMed abstracts⁷. When all the redundancy is stripped away, the remaining assertions, which retain the information on all its constituent nanopublications, we call ‘cardinal assertions’. Those cardinal assertions (CAs) are all disambiguated triples and each concept in a given CA is present in other CAs as well. CAs, therefore, form a multi-dimensional web of information, over which one can reason, and which can be ‘projected’ in a number of ways to reveal an overview of

⁴ Velterop J (1995) “Keeping the Minutes of Science” in: Proceedings of Electronic Libraries and Visual Information Research (ELVIRA) Conference, Aslib, London, No. 2.

⁵ http://www.linkpdf.com/ebook-viewer.php?url=http://www.nbic.nl/uploads/media/Nano-Publication_BarendMons-JanVelterop.pdf

⁶ Groth P, Gibson A, Velterop J (2010) The anatomy of a nanopublication. Information Services & Use, Vol. 30, pp. 51–56, doi: 10.3233/ISU-2010-061

⁷ Barend Mons, personal communication

existing knowledge and, if viewed over a period of time, where that knowledge is subject to the most, or fastest, changes.

Areas in which the available knowledge is subject to the most or the fastest change are likely to be the most interesting to a scientist, and worthy of focussing one's attention. Because CAs all refer back to their constituent nanopublications, and each nanopublication to the publication it has been extracted from, prioritising one's reading of any full papers from the literature can be based on their actual information content rather than on subjective filters such as the impact factors of the journals in which the articles are published. In other words, it allows for choosing on the basis of comprehensiveness rather than perceived notions of journal quality.

An added advantage is that the pitfalls of journal bias, a consequence of the journal system and journal-specific peer review ("does this article fit in this journal?"), can be avoided.

Though at this stage it is envisioned that nanopublications are mainly being extracted from publications, there is no reason why they shouldn't be contributed to the collective information pool via other routes, such as, for instance, databases such as the LOVD⁸, or scientific blogs. Blogs of patient societies, particularly those devoted to rare or 'orphan' diseases, are an important source of epidemiological information. Their status in terms of peer review (*in casu*: none) will be known, and in any reasoning or overview they can be taken for the value the user wishes to attribute to them. This may develop into an important source for information on negative results, as those are all too seldom published in journals.

What is needed for this vision to become reality?

The vision is clear, and is beginning to be implemented. In March 2011, in the so-called "OpenPHACTS" (Open Pharmacological Concept Triple Store) project, one of the projects under the aegis of IMI, the European Innovative Medicines Initiative⁹, the 22 core project partners (who come from universities as well as the private sector – mainly the pharmaceutical industry in Europe) will embark on a 3-year programme to apply the semantic technology and methodology described to information of importance to the pharmaceutical industry, focussing on chemical, biological, and medical information, and bioinformatics, in order to create a so-called Open Pharmacological Space. The potential of *in silico* models and *in silico* reasoning, greatly enhanced by the availability of large bodies of CAs, is expected to significantly accelerate drug discovery and underlying drug-oriented research.

Initially, nanopublications will be mined from open resources (such as PubMed abstracts and SwissProt database) and resources specifically made available for the OpenPHACTS project by the various partners. It is the intention, however, to convince publishers that it is in their interest to append properly formatted nanopublications to the articles they publish, in a similarly open and freely accessible way as now generally is the case for article abstracts. There are two main reasons to be optimistic about publishers endorsing this development: (1) those who publish medical, biological, chemical, and pharmaceutical journals will want to satisfy their important customers in the pharmaceutical industry (the

⁸ LOVD – Leiden Open Variation Database. <http://www.lovd.nl/2.0/>

⁹ IMI – Innovative Medicines Initiative. <http://www.imi.europa.eu/>

majority of pharmaceutical companies in Europe are OpenPHACTS partners), and (2) nanopublications are in essence references to full articles, enabling 'deep' citation, i.e. not just to the articles in question, but to actual assertions in those articles, thus considerably helping citation rates – and therefore impact factors – of the participating journals.

Is this finally the creation of the 'semantic web'?

Perhaps, but it is prudent to be cautious about that. The benefits of the technology have to be understood properly. If I may use the metaphor of archeologists making aerial surveys of a terrain before they start digging in order to ensure that they do so in those places that are most likely to yield results and not randomly, it is something like that that the semantic technology delivers. Exact answers are not to be expected in most cases, but strong or even very strong hints as to what should be worth testing or which papers are worth reading in full can be expected. Accuracy and comprehensiveness may have to be traded off to a degree. That said, having a set of tools to be able reliably to make choices from the *mer à boire* that scientific publications and associated information sources have become, is major progress and will save much time and effort.

There is also still an area of concern that hasn't really been addressed much yet in the context of using this semantic technology to deal with large amounts of information. That is the propensity of us human beings to recognise patterns. Given the right tools, it is relatively easy to recognise patterns in large amounts of information that lead us to discover connections that are difficult to discern otherwise. But it is also easy to recognise patterns that are artifacts of our imagination only and that do not represent underlying realities. A helicopter view does come with a lot of noise, or if you wish, a view from a hot air balloon with a lot of hot air. Perhaps mathematical concepts such as randomised matrix theory can help, calculating the patterns that may occur purely by chance, so that whichever patterns that remain are more likely to be based on real phenomena.

**The work on nanopublications was initiated and is led by NBIC, the Netherlands Bioinformatics Centre, a constituent member of CWA.*