There is something fascinating about science

Based on a paper entitled 'Beyond open access' presented at the 32nd UKSG Conference, Torquay, March/April 2009

The increasing overabundance of scientific information exposes the need for techniques to represent this information in forms that make it possible to make optimal use of the knowledge it contains. Expressing it in the form of semantic conceptual triples enables 'navigating' existing knowledge as well as transforming any traditional scientific text into an efficient gateway to more relevant information beyond what is contained in the text itself.

JAN VELTEROP
Knewco and Concept Web Alliance

"There is something fascinating about science. One gets such wholesale returns of conjecture out of such a trifling investment of fact."

(Mark Twain, Life on the Mississippi)

Mark Twain was very witty. But he lived in a different age, and his witticism would certainly not apply to modern science. The problem of today is precisely the opposite: far too few returns in terms of usable knowledge out of such overwhelming investment of fact. That being the case, we are running up against two problems. The first is that a lot of information is deeply hidden, and the second that information is increasingly and overwhelmingly abundant, so that 'connecting the dots' is ever more difficult, because there are so many dots to connect. This is rather a paradoxical situation. Wouldn't we actually increase the information overabundance were we to find all the relevant information that exists? Of course we would. However, limiting information, or stopping to gather facts, cannot be a solution. We must approach it from the other side, by finding ways to deal with the overabundance of information. We must think of the situation not as information overload, but perhaps as something like 'organizational underload' of information; as having a lack of sufficient conceptual structure. We have to do this, because the amount of relevant information is not going to diminish. Instead, its abundance and availability is bound to increase, quite possibly exponentially. Not only in terms of the peer-reviewed literature, but also in terms of raw data, and of informal literature, such as science blogs, wikis and the like. A kind of Moore's law (after Gordon Moore, the co-founder of Intel)[1] seems to be going on, with the amount of scientific information doubling every so often (Katy Börner quotes J Scholtz in a 2006 chapter[2] as reporting a doubling of the volume of data every 18 months in some areas, and that was in 2000).

When *Homo sapiens* was at the beginning of his evolutionary development, he didn't really have any other use for water than to drink it, and perhaps swim in it. Large bodies of water, such as lakes or seas, formed effectively the end of his range. He had to wait until a bright spark developed rafts and boats, and then he started using water also to navigate and go to places hitherto unreachable. The rest is history. Empires were built that way and the world was conquered. My sense is that we are at the beginning of a similar point in our evolutionary development with regard to information. Now, we mainly take it in by reading and consulting databases (the equivalent of drinking in the water analogy), and we haven't got the means yet to 'navigate' the existing information effectively. We are making good progress with searching. But searching, although often called 'navigating', isn't the same and has its own drawbacks. After all, if you want to search, you must already know what for, and you have to formulate your search

argument. Finding what you didn't even know you should be searching for – a form of serendipity if you wish – is hardly given a chance, even though serendipitous findings are often the stuff of scientific breakthroughs. True navigation is different. It is about using information to 'carry' you from one place to the next. It is about discovering that connecting information that seems far apart, semantically speaking, can lead to insights otherwise extremely difficult to attain.

In order to be able to navigate 'oceans of information' we must first release information from its silos and make it compatible. One of the ways in which that can be done is by breaking information down into its smallest elements: semantic triples. Briefly speaking, triples consist of three elements (as the name implies): a source concept, a target concept and a relationship between the two. The general format is this: ***<concept1><relationship><concept2>***, for instance, ***<this article><is published in><the journal Serials>***. This example also makes clear that the relationship can be directional (though isn't always). The triple ***<this article><is published in><the journal Serials>*** is not the same as ***<the journal Serials><is published in><this article>***, should this triple be a valid statement in the first place (in the event, it is a nonsensical one). A non-directional triple is, for instance, ***<triple><co-occurs in the same sentence with><statement>***, which would be valid in either direction. The relationship between the concepts in the latter example is much more vague of course than the relationship in the first example. Though I think I made triples look simple, it isn't quite as simple as this. In order to make vague triples, any triples, more meaningful, you need to qualify them. Triples have attributes, such as a provenance, date and time, conditions (e.g. is true in certain circumstances only), identifiers for the concepts, and potentially many others. We can, for instance, indicate where the sentence occurs from which the triple was mined, who wrote it, when, etc. If the same conceptual triple occurs in the literature often enough, and is written down by enough different people, then there may well be added significance in the numbers. Another reason to qualify triples and to add provenance labels, date labels, digital identifiers and the like to them, is the desirability of crediting, or at least acknowledging, the author (and journal, publisher, database, etc.), as the whole social fabric of scientific knowledge is dependent on acknowledgement.

I mentioned 'semantic triples' above, but the examples I gave do not yet demonstrate what semantic means in this context. The triple ***<this article><is published in><the journal Serials>*** could also be written as ***<this paper><appears in><Serials>***. The meaning would be exactly the same. As word triples, they are different, because they use different words, but as semantic triples they are identical, because they mean the same. In order to get semantic triples, disambiguation of the words needs to occur. Synonyms and homonyms need to be resolved. Synonyms are relatively easy to deal with. The difficulty lies in disambiguating homonyms. They require careful analysis of the context. 'Paper' could mean 'article', but it could also mean 'newspaper' or the material on which either are printed. 'Serials' could mean the journal, or periodicals in general, and quite possibly other things as well. Disambiguating homonyms is sometimes nigh impossible with just automated contextual analysis. Imagine an article about turkey farming in Turkey. Or a piece describing the little statuette of a jaguar on the bonnet of a Jaguar. In cases like this, human intervention is needed. If one wants to be able to get a reasonable overview of a field of knowledge via the use of triples, one needs a reasonable degree of disambiguation and conversion into semantic triples in order to draw any conclusions at all, though there is always likely to remain some fuzziness in the results. Some fuzziness doesn't perhaps matter too much as long as one is aware of the fact that reasoning with triples is just a tool, albeit an important tool, to discover new knowledge, to identify what the most promising areas of research are, with the most likely chance of finding significant insights, and that it is not the new knowledge itself.

What does this all mean for publishing? There are currently at least two areas in which publishers can take advantage of the semantic triple methodology. Firstly, there is the actual publishing of triples. The material that is being published, in journals, books and databases, can be turned into triples, as long as it is available in electronic form, of course. Such triple collections can be valuable complementary formats to enable the research community to potentially get more out of the published knowledge. The currently prevailing formats, print, PDF, and even HTML, still require the material to be read. Of course, PDF and HTML make distribution easier and cheaper than print, and HTML also increases findability enormously, especially in

combination with metadata, but the actual usage of published information remains, on the whole, limited to traditional methods of taking the knowledge in at source. Meta-analyses are being carried out in some areas, but are usually text-based, within narrow fields of research, and their usefulness suffers from the major drawback that is presented by widespread ambiguity in most texts. Were the literature available in the form of triples, however, in particular semantically disambiguated triples, it could be put to much wider use. Not only for meta-analyses, but also for the purpose of pointing to – and linking to – topics of research that are more promising than others, or topics that might easily be missed if the literature is only taken in by reading, and for generating new hypotheses. Properly constructed triples that adhere to common models, such as rich RDF (Resource Description Framework) and best practices, and that are disambiguated, delivered and packaged in a convenient way, are potentially worth significant amounts.

Secondly, the ability and high precision of matching concepts and triples is also very useful in an early manifestation of this semantic triple methodology, helping readers of scientific articles to find new and less obvious information related to what they are studying. The technology already exists to semantically index pages on-the-fly, and then recognize the concepts in the text. Not just keywords, but concepts, so that, for example, the word 'cancer' is recognized as the concept 'malignant neoplasm', which is the preferred term in the Unified Medical Language System (UMLS). These pages can subsequently be made to highlight, in one form or other, the concepts recognized, and the concepts, when clicked upon, be made to show a whole host of other information relevant to it. The obvious example is showing any synonyms (as with cancer and malignant neoplasm – but really greatly helpful in the case of, for instance, proteins and genes, which often have many synonyms). And then definitions could be given, equivalents in other languages, links to literature (books, journal articles), links to experts or just other researchers on the topic, links to highly specific laboratory materials (such as antibodies or cell lines or fungus cultures), deep into the suppliers' catalogues, thus saving the researcher a lot of trouble.

The semantic approach also makes it possible to link to other related concepts (including the nature of the relationship) and, interestingly for publishers, other places in their portfolios where the concept in question or related concepts occur. Apart from convenience for the user, this also increases the likelihood that the user stays within the publisher's site longer, which must be good from a publisher's point of view. All these instances mentioned reintroduce a measure of serendipity in the scientific literature, a way of stumbling upon connections that are not, and would never be, obvious. And yet, as already mentioned above, serendipitous findings are the stuff of unexpected insights, even breakthroughs, in science. Serendipity and the chance of finding what you didn't even realize you were looking for is lost, to an extent, with the advent of search functionalities, which require that you at least formulate the search argument. This nudges you into the direction of seeking out more of what you already know: a form of homophily, if you wish – the 'birds of a feather' syndrome – which is impeding out-of-the-box thinking, or at least making it more difficult. Semantic concept technology may not quite be the search engine that presents you with exactly the opposite of what you are looking for – though that might perhaps be extremely interesting and beneficial for research – but it certainly has the potential to take you into areas where you would not normally go for discovering knowledge.

There are other reasons why a publisher might deploy this technology. The embedding of highlights and links, which can be made invisible until the reader moves the cursor over them, makes it possible to have a wealth of information at hand without disturbing the user's reading experience and without the need to leave the page and re-type the concept or keyword in the search box of some other site. More knowledge is 'served up', as it were, in the page the user is already reading, increasing the information density of a scientific text without it becoming unreadable, effectively transforming traditional texts into efficient gateways to other relevant information.

The examples mentioned are just the beginning. The over-abundance of information will increasingly force us to find ways around having to read all relevant scientific articles. Paradoxically, the role that publishers can – and ought to – play is one of helping researchers who go out of their way to avoid reading scientific articles, to do just that: getting the essence of scientific information without having to read too many full articles.

References

1. ftp://download.intel.com/museum/Moores_Law/Articles-Press_Releases/Gordon_Moore_1965_Article.pdf (Accessed 2 June 2009)
2. Börner, K, Semantic Association Networks: Using Semantic Web Technology to Improve Scholarly Knowledge and Expertise Management. In: *Visualizing the Semantic Web*, Ed. Geroimenko, V and Chen, C, 2006, Springer Verlag, 2nd Edition, chapter 11, 183–198. Available at: http://ella.slis.indiana.edu/~katy/paper/04-chapter11.pdf

NB The author has been unable to locate Scholtz's original article, but Börner cites:

'Scholtz, J. 2000. DARPA/ITO Information Management Program Background' as her source.

■ **Jan Velterop**
Knewco and Concept Web Alliance
Contact address:
'Walden' 9 Benfleet Close
Cobham
Surrey, KT11 2NR, UK
E-mail: velterop@knewco.com