Knowledge is like Love….it multiplies when shared !

They are getting too powerful …… Let's give them Bioinformaticians

Summary of the inaugural lecture by Prof.dr. Barend Mons at the acceptance ceremony of the first Chair in Biosemantics, established at the Leiden University Medical Centre on behalf of the Netherlands Bioinformatics Centre. 25th of February, 2013 at 16.15 in the Grand Auditorium of Leiden University.

Mr. Chairman,

We are at, or already over the verge of a completely new way of doing science, also referred to as eScience. **Biosemantics** is an integral part of eScience and indispensable to understand the complexity of biology. Not because biology is so much more complex than in the past, but simply because we uncover layer after layer of the complexity that was always invisibly there. Now the complexity of life begins to scare us to death. With the advent of high-throughput technologies in the life sciences we have had a crisis in scientific communication evolving in stealth mode, which is now hitting with full force: an unmanageable amount of relevant data.
The use of scientific literature and databases is rapidly shifting from 'exploratory' reading to 'confirmational reading' to check elements of the hypotheses that massive data analyses throw at us. This requires a very different role for scholarly communication and this is not the classical scientific article.

Brief History
The very short (one decade) history of Biosemantics feels very much like a tough sailing trip. The scepsis in biology around text mining, artifical intelligence and bioinformatics in general is like a strong headwind when trying to 'sell' the added value of *in silico* knowledge discovery. Unless one has a very strong money-powered boat, one has to sail 'against the wind'. From a helicopter the journey of the past ten years may be judged as 'going all over the place', but at closer observation, a regression line can be see that steadily brought us closer to the goal:

All information as openly as possible available, traceable and understandable for everyone, including computers.

During the first seven years of the fledgling Biosemantics group, first at the EMC in Rotterdam and as of 2008 in close collaboration with the new group at the LUMC, our research was mostly a matter of 'digging to build a strong foundation' We mainly published about the painful attempts to reconstruct something from narrative literature and databases that was at least *more or less* understandable for computers. Only in 2007 our first 'scary' article appeared, demonstrating that (re-) assignment of functions to proteins could be more efficiently done with the help of computers than with human experts.

Ever since, approximately half of our research output is still on methodological improvements in biodata mining, disambiguation, and associative reasoning. However, the other half, and growing, is on actual knowledge discovery processes and new biological findings, only later to be confirmed in the laboratory.

However, in case you might think the demonstration of actual *in silico* knowledge discovery examples would silence the criticism, you are mistaken. In the review process of almost every article we submit, there is minimally one reviewer who writes a variant of the conservative and elitist statement: *'I cannot accept that the computer is more clever than I am'*. That is actually the last thing we suggest. We just demonstrate that the computer is much better than we are in systematically sifting through 22 million articles and thousands of databases.

During this first decade of Biosemantics we gradually developed, with colleagues in Amsterdam and later internationally, the notion of a fluid, dynamic conceptual space where associative reasoning by computers could be complemented with Description Logic type reasoning. We called that the 'Concept Web'.
In 2009, under the auspices of NBIC a group of thought leaders met in New York to found the Concept Web Alliance, which became an Open Think Tank addressing the formidable technical challenges associated with Big Data Science and the communication and stewardship of all these valuable and re-usable data.

The brainchaild of the Concept Web Alliance is the notion of 'nanopublication'. Today, only a few years later, nanopublications are widely studied and accepted as a way to deal with massive data sets, interoperability of data, data citation and crowdsourcing. A major project in the Innovative Medicines Initiative with over 30 partners from the pharmaceutical industry and the public sector is implementing nanopublications at its core and this year a meeting will take place gathering most organizations already working with nanopublications in Amsterdam to investigate the need for a 'Nanorepublic' to support the crucial elements of the nanopublication ecosystem.

The period in which 'web publishing' apparently meant 'putting dead PDF's on the Internet' is as good as over. However, the so called 'article of the furture', in fact not more than a mishappened Christmas tree of hyperlinks in the text that was originally meant for reading, is *not the way to go*. The mistake is that publishers are apparently as of yet still unable to think enough out-of-the-box to depart from the 'article' as the principle unit of scientific cmmunication and migrate towards entirely new and computer enabled ways of scholarly communication. The brief period in which journals started to accept 'supplementary data' linked to classical articles is also coming to an end. The first journals are already abandoning this policy. First, because reviewers hardly ever come around to review these data sets. Second, because the data are frequently at the proverbial 'laptop of the PhD student' and therefore untraceable after a few years and third, because most data are in exotic, undoubtedly brilliant, but unreadable formats reinvented on a daily basis by bioinformaticians. On top of all of this, the archaic system used by universities and funding agencies to judge scientific output based on the article as the smallest unit of scientific communication is rapidly becoming one of the single most inhibitory factors in eScience.

Especially in biology, we already feel the pain of the widening gap between our ability to generate data and our inability to do something useful with them. There is a very critical paradox in modern life sciences: The massive datasets we produce are very valuable '*in principle'*, they can be re-used by an entire generation of scientists '*in principle'*. However, and paradoxically, there is not a *single* incentive mechanism to promote and award sharing of these valuable data. In fact the way science funding and evaluation works, precludes the trading of data, resulting in the infamous 'Data Graveyards'.

Now, science is in its essence an ‘Egosystem’ and especially older scientists have a tendency to keep their data abreast as their main fuel on their dreamed journey towards the Nobel Price. If we ever want to make open data sharing the norm (and if we don’t, science will come to a grinding halt) we urgently need to change the reward system in Science. Nanopublications give at least the technical basis for ‘altmetrics’ supporting data sharing, and they also have bearing on some of the social aspects, although not all.

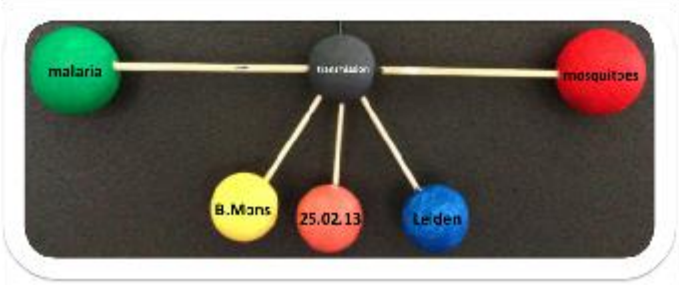
eScience, here defined as ‘science that can not be done without computers’, *needs a computer-friendly way of scholarly communication.* We need to accept that computers are now our most important technicians outside the wet lab. So let’s start to feed them information that they can easily consume and not human language full of rethoric, aphorisms, synonyms and jargon. In fact eScience is in dire need for ‘Social Machines’. These are environments where continuous ‘Mind-Machine Interaction’ is enabled. In biology, not only 22 million articles have already been published to date, but every 40 seconds a new one is produced. The average data set in high-throughput biology yield hundreds of thousands, frequenty even hundreds of millions, and in some cases already billions of known and novel associations between concepts such as genes, proteins, metabolites, cells, tissues, oranisms and their phenotypes, including disease. In the near future data sets for personalized –omics will routinely exceed a billion data points.
In biological Social Machines therefore, millions of new assertions need to be processed per second to enable effective routing, annotation and validation by millions of people. Once such Social Machines will come up to full speed, which I predict will be in the very near future, the way we conduct scientific research and knowledge discovery will change fundamentally.

For the foreseeable future it is likely that people will make the final decision on whether they believe a given computer-generated scientific assertion or hypothesis. However, I predict that we will see a steep increase in the contribution of not only biological but also computational high-throughput streaming systems to the body of hypotheses in biology.

In order to frame Biosemantics in biological terminology, I will upset some people by introducing another –ome. I here define everything we have already explicitly asserted in the literature and databases as the **Explicitome.**
The Explicitome of current biology is estimated to be in the order of 10^{14} meaningful assertions (representing biological claims). Tomorrow, the Explicitome will have grown with a few hundred thousand new associations.
There is no computer cloud or brain that can efficiently handle this amount of information and therefore intelligent data reduction steps are crucial in order to feed Social Machines only with the information they cannot ignore if they need to reason over the entire body of information on a daily basis. The good news is that the vast majority of assertions in the Explicitome are literal citations (or repetitions if you will) of assertions that were already done before.

Nanopublications

The name nanopublication was coined for a computer readable, single meaningful assertion, including its rich provenance. Just like any classical publication, a nanopublication has authors, a time date stamp, a unique reference etc. The guidelines for nanopublications as now developed by the CWA recommend keeping the assertion in a nanopublication 'as small as possible to be a self-contained scientific claim'. The smallest thinkable associative assertion between two concepts simply states in a Subject>Predicate>Object triple how two concepts relate to each other, for instance Protein A > interacts with > Protein B. In practice, many nanopublications need more than one triple in the assertional part. Also provenance and context (such as for instance the conditions under which a reaction takes place) can be formatted as triples in RDF or related computer readable languages.



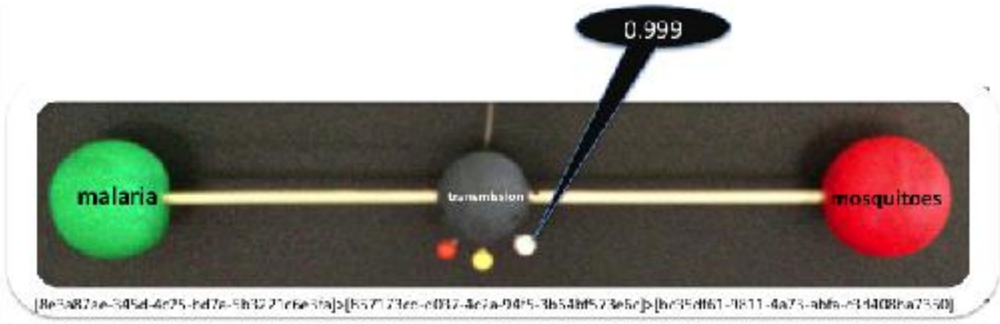
Altogether these triples form a small graph that represents the Nanopublication in computer readable format which can be represented to human users in their own language of choice at the same time. In order to create high quality nanopublications, each concept in the graph should be mapped to a unique reference (identifier) that unambiguously defines which concept is referred to. This major effort, which requires rich ontologies, Identity Mapping Services and many other software workflows and standards is an international effort by default and takes place in the context of the CWA and several international projects at the moment.

Based on the (combination of) unique identifiers, each assertion in a nanopublication can be easily recognized, and very importantly, identical assertions with only different provenance can be mapped and 'treated as citations of the oldest version of the assertion'. Here I take my old example of the assertion [malaria] > [is transmitted by] > [Mosquitoes]. Obviously, in human language we understand that [malaria] > [is transmitted by] > [Culicidae] (in fact a more precise statement) and [malaria]> [wordt overgenbracht door] > [muskieten] (Dutch) and [Le Paludisme]> [est transmis par]> [des moustiques] (French) are in fact asserting the same claim. In RDF these three concepts can be referred to with multiple URL's as well. 'Under the hood' we should be able to map the basic statement to one 'cardinal assertion' with three UUID's that never change, because they are non-semantic (have no meaning whatsoever) and belong to the 'community'. Now all the lingual assertions made in this paragraph are recognized by computers as the basic assertion:

[8e3a87ae-345d-4c25-bd7a-5b3221c6e3fa]>[857173cd-d037-4c2a-94f5-3b64bf573e6c]>[bc35df61-9811-4a73-abfa-c3d408ba7350]

Computers are excellent in recognizing these unique and identical sequences and will therefore promptly recognize every new assertion that enters the system with the same three UUID's as identical. Also 'nearly identical' can be accommodated because we also have a nanopublication that states that [malaria] [is caused by] [Plasmodium spp.] Very simple inferencing can now result in the mapping of the assertion [Plasmodium spp.]> [are transmitted by] > [Culicidae] as one more supporting nanopublication for the 'cardinal assertion' [malaria] > [is transmitted by] > [Mosquitoes].
In the current Explicitome the assertion with the UUID combination [8e3a87ae-345d-4c25-bd7a-5b3221c6e3fa]>[857173cd-d037-4c2a-94f5-3b64bf573e6c]>[bc35df61-9811-4a73-abfa-c3d408ba7350] is present at least 10,000 times, harvested from literature and databases. The Cardinal Assertion [malaria] > [is transmitted by] > [Mosquitoes] can now be endowed with a very high 'evidence

factor' and one reference to all supporting (> 10,000) nanopublications stored elsewhere. This is obviousy an enormous data reduction, accompanied in fact with a gain in information (the 'trustworthiness' of the cardinal assertion has been calculated and can now be used for reasoning purposes.
The result will be that the very fact that in this lecture, I have repeated this assertion numerous times in different languages should not 'wake up' any Social Machine as I have not changed the position of the concepts malaria and mosquitoes in the Concept Web in any way.


0.999
malaria
transmission
mosquitoes
[8e3a87ae-345d-4c25-bd7a-5b3221c6e3fa]>[857173cd-d037-4c2a-94f5-3b54bf573e6c]>[bc35d161-1811-4a73-abfa-c3440Bba7350]

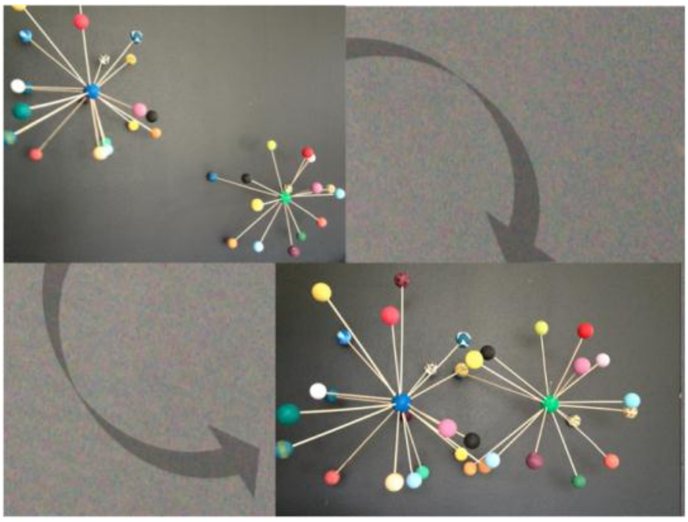
However, if somewhere in the world someone would publish right now the statement that [malaria] > [is transmitted by] > [Sandflies] all smartphones of malariologists should ring immediately and they should express their fierce doubts about this claim. The system will either detect that the assertion
[8e3a87ae-345d-4c25-bd7a-5b3221c6e3fa]>[857173cd-d037-4c2a-94f5-3b64bf573e6c]>[68e081bc-edcf-4926-8a60-5347cafb162a]
is novel to the Explicitome, or that the provenance states that this is a previously contested statement and proven errenuous.

Using this similarity principle at the UUID level, we have now reduced a space of 10^{14} individual nanopublished assertions to approximately 10^{11} Cardinal assertions, each with a dynamically (and transparently) computed 'Evidence Factor'. This Cardinal Assertion Store is where applications for crowdsourcing, and reasoning will go. The provenance of each Cardinal Assertions will allow a one-click exopansion to see all supporting nanopublications and their sources.

The Knowlet.
Now that we have these 10^{11} Cardinal assertions, it becomes very easy for computers to 'index' all Cardinal Assertions based on their 'subject' (the first UUID)
All Cardinal Assertions starting with 8e3a87ae-345d-4c25-bd7a-5b3221c6e3fa
can now be clustered into what we call a 'Knowlet'. In the picture, the green 'subject' ball in the centre represents the concept Malaria, and all other 'object' balls represent concepts like Africa, mosquitoes, children, chloroquine, Plasmodium, fever, red bloodcells etc.
A 'real Knowlet frequently contains many thousands of concepts (see below). The sobering but also comforting fact is that we published on

about less than 3 million concepts in the current Explicitome and therefore we have now ‘zipped’ the Explicitome from 10^{14} nanopublications to 10^{11} Cardinal assertions to 3 x 10^{6} Knowlets (a 10^{11} reduction). Obviously, although not depicted here, between each of the coloured balls in the Knowlet, we know the ‘predicate’ and each connection has a ‘value’ attached based on the Evidence factor as described before or any custom algorith a researcher wishes to implement on the nanopublication space.

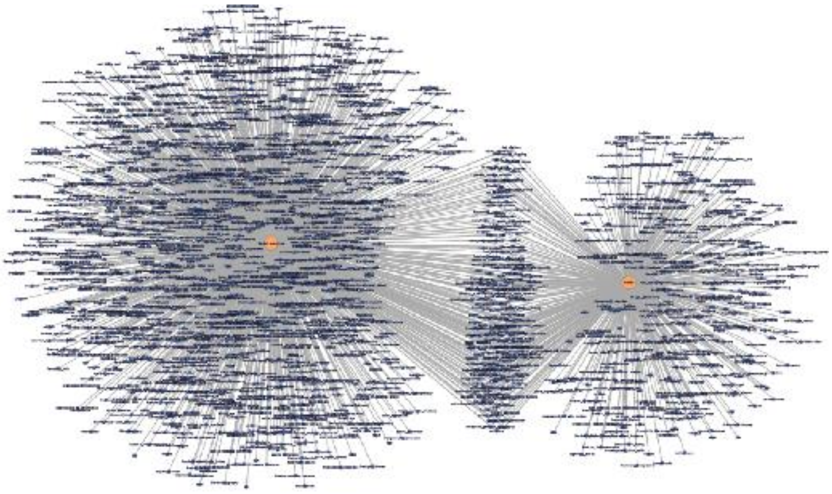
Knowlet Discovery



!

Each Knowlet has an ‘attraction’ to each other Knowlet with similar concepts in its ‘Object Cloud’, while some knowlets have very little to do with each other and may even repulse each other, some others may move closer and close in the Concept Web, pushed and pulled by many other knowlets in the space. The Knowlet of Malaria may move so close to the Knowlets of the cancer drug Tegafur that, even though they have never been mentioned together before, the Social Machine will hypothesise a potential anti-malaria role for the drug. In the figure below we depict a ‘real life’ example of a new gene suggested for Huntington disease showing how many concept real Knolwets contain and how many may cause the attraction.

eScience needs different complementary levels of reasoning. Think of the metaphor of the helicopter view. One would never see the abbarent growth pattern in a cornfield caused by the remains of a Roman fortress when walking in the midst of the field. However, after spotting the pattern from the helicopter, one needs to land, take a shovel and dig to find the ruines. Next step would be the laboratory experiments to demonstrate the age of the stones before the conclusion can be drawn that indeed the pattern observed revealed a Roman fortress. Knowlets enable the helicopter view. With for instance Description Logics the immediate surroundings of the new associations can be explored (compared to the shovel), whilst final comfirmation of causal biological relationships in the wet Lab will follow.



All this is already feasible in practice or at least will be in the very near future.
On the page where this text was downloaded from, you will soon find the first links to practical applications and Smartphone applications.

Science has always been a more or less elitist endeavour. When the printing press was invented contemporary scientists warned that ‘common people’ would now have access to the sacred scientific knowledge. When the Internet became a commonplace tool, and even more scary, crowdsourcing sites like Wikipedia and Social networking came up scientists rallied to ‘raise red flags’ about the validity of such semi-scientific communication. The idea to embrace all these opportunities to bring a ‘million minds’ into science was initially received as swearing in church. Now we are following up on the Nature Genetics Editorial accompanying our Value of Data paper in the same issue (see links) called ‘Crowdsourcing Human Mutations. No matter how scary it may sound to traditional lab-contained biologists, computers as well as millions of interested lay people (including patients) will increasingly influence our scientific progress.
The wisdom of the crowd, as well as computers will generate myriads of hypotheses. These will be far too numerous to all be tested in the laboratory.
So either we accept Hypothesis Graveyards as the successor of Data Graveyards, or we adapt and use the tremendous power of a million minds and a trillion CPU’s to advance biology to the layers of complexity we now need to try and understand. With appstores, micropayments, twitter and citable nanopublications now becoming mainstream, we will soon see the fundamental changes in the way we conduct mobile social sciences.

Soon enough, millions of nanopublications will be routed every hour to millions of scientists for comments, hypothesizing, error detection and other forms of curation and annotation. Until further notice, well fitting computer-generated hypothesis will serve as ‘conditional truth’, until the assertion gains so much importance that solid experimental elucidation of the underlying biology is warranted. With well over 300,000 nanopublished high confidence associations between genes, proteins and diseases from LUMC alone we will make a headstart with science in Social Machines this year.