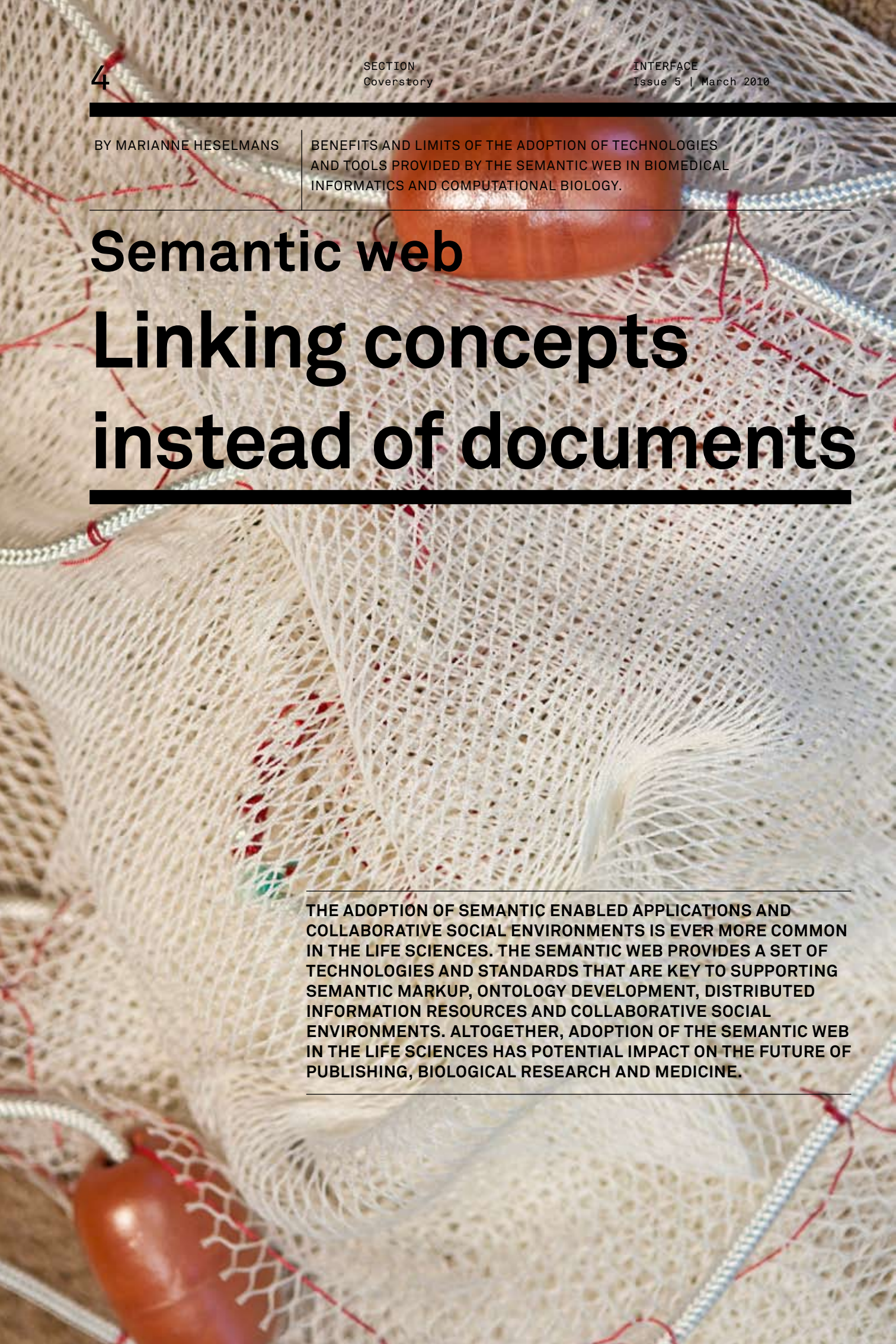
BY MARIANNE HESELMANS

BENEFITS AND LIMITS OF THE ADOPTION OF TECHNOLOGIES AND TOOLS PROVIDED BY THE SEMANTIC WEB IN BIOMEDICAL INFORMATICS AND COMPUTATIONAL BIOLOGY.

Semantic web

Linking concepts instead of documents



THE ADOPTION OF SEMANTIC ENABLED APPLICATIONS AND COLLABORATIVE SOCIAL ENVIRONMENTS IS EVER MORE COMMON IN THE LIFE SCIENCES. THE SEMANTIC WEB PROVIDES A SET OF TECHNOLOGIES AND STANDARDS THAT ARE KEY TO SUPPORTING SEMANTIC MARKUP, ONTOLOGY DEVELOPMENT, DISTRIBUTED INFORMATION RESOURCES AND COLLABORATIVE SOCIAL ENVIRONMENTS. ALTOGETHER, ADOPTION OF THE SEMANTIC WEB IN THE LIFE SCIENCES HAS POTENTIAL IMPACT ON THE FUTURE OF PUBLISHING, BIOLOGICAL RESEARCH AND MEDICINE.

How to get more useful answers on search queries such as *Which genes are involved in this disease?* Designing a semantic web with billions of concepts and proven relationships between these concepts may offer the solution. NBIC focuses on building tools and developing technologies to support a semantic web.

Suppose you want to know which genes are involved in the so called Sjögren-Larsson syndrome, a genetically inherited disease often characterised by thickened skin, spasticity and mental retardation. If one uses current search engines such as Google, the answer will be quite frustrating. After all, Google does not really understand this question: the software can only link all documents on the web containing the letter orders SJÖGREN-LARSSON SYNDROME and GENES. So you will receive hundreds of documents ('hits') including documents about, for instance, an American artist John Sjögren-Larsson who is selling a painting called 'Genes', or documents from a town Sjögren where students can follow a course 'genes and genetic inheritance'.
Getting more useful answers to questions such as *Which enzymes are involved in this metabolic pathway?* or *What side-effects could this drug have?* is an important goal for the life scientists and computer technologists who are now confronted with millions of documents and scientific articles and thousands of protein and gene databanks.
"The current worldwide web only consists of documents and pictures," explains Frank van Harmelen, professor at the Computer Science department of the Vrije Universiteit Amsterdam. "But the search engines cannot relate and interpret this information. So now we are developing a second web with information meant for engines."

SECOND WEB This second web is called semantic web. Last year NBIC initiated an international forum called Concept Web Alliance (CWA) to cooperate in building up such a semantic web with the emphasis on free knowledge sharing. Barend Mons of NBIC is the scientific co-ordinator for this alliance. Other groups – like W3C Semantic Web Health Care and Life Sciences Interest Group (W3C-HCLSIG) – already started working on similar goals a few years ago. In these groups, life scientists along with computer technologists cooperate in developing 'semantic' search software. And even more important, they put data on the semantic web in such a way that other similar applications can make use of it.
The semantic web is first of all a web of concepts, each with an identifier (a number). At the moment, most genes, proteins and diseases – information about which is stored in many different databases around the world – also have many different names. How can a search engine recognise that experts use the words cancer, kanker, tumour and melanoma for the same disease? Or that gene A in databank X is the same as gene B in databank Y? Linking all the different names to one concept (identifier) makes a solution to this searching problem possible. Participating in a group like the W3C-HCLSIG and the CWA, one can put his or her word for cancer under the concept 'cancer' (and thus the identifier). Asking which genes are involved in cancer, the questioner can be sure that the search engine has used all the known information in the world about cancer and the involved genes.
"Every existing database can continue to use its own name for a specific disease, gene or protein. The only thing the owners have to do is link their names to our concepts or numbers in such a way that search engines can understand which names are synonyms," explains Frank van Harmelen, who is an active member of the Concept Web Alliance. "The semantic web gives you a way to deal with all these different names," adds Scott Marshall, assistant professor of Bioinformatics at the Leiden University Medical Centre and co-chair of W3C HCLSIG.

SETS OF TRIPLES The so called 'triples' are the second characteristic of the semantic web. A triple describes one specific relationship between two (other) concepts or, in the words of Scott Marshall: a subject predicate object ('node-edge-node'), each of which can be represented by a so called URI. For instance: Sjögren-Larsson syndrome (concept) is caused by (specific relationship) ALDH3A2 mutation (concept). Or: ALDH3A2 encodes FALDH. Or: FALDH deficiency causes fatty alcohol accumulation.
These three examples of triples come from the Peroxisome Knowledge Base, one of the first being developed as part of the BioExpert project (www.bioexpert.nl).
In this recently launched, free database, sets of triples (concepts and their relationships) are organised and published as 'concept maps' that describe different aspects of peroxisome biochemistry and related diseases such as Sjögren-Larsson syndrome. "We can integrate data from sources as diverse as literature, pathway databases, protein databases, gene databases and of course the experts themselves in our BioExpert system," according to project leader and bioinformatician Andrew Gibson from the University of Amsterdam. 1000 concepts have already been identified and described with concept maps. The concept map on the website shows at least 14 coupled triples, and the knowledge base contains over 150 concept maps.
"The challenge is finding all the millions of known relationships between the involved concepts," explains Barend Mons. Several groups around the world have developed software programs that can recognise triples in scientific articles and databases. A triple that has been found in many documents and a triple that is elaborated by a vested research group will receive a higher status for reliability. For instance, if the software finds a sentence that means: ALDH3A2 encodes FALDH 500 times in all documents and databases, this relationship has been identified as a 'fact' – more than just a hypotheses. But the programs are not perfect. "So 'expert-mining' is also important," says Gibson. He and other computer technologists talk a lot with the (molecular) biologists and medical researchers in the peroxisome field to find out which concepts and names different experts use and how reliable specific relation-

ships are. “Experts from the whole world can add their knowledge to the map.”

KNOWLEDGE BASE Frank van Harmelen has used the new software to find all triples hidden in the protein database Uniprot. This protein database alone resulted in 1.6 billions triples – facts about proteins. “As a rough estimate, we expect at least 20 billion triples in the life sciences,” says Van Harmelen. A few years ago, Scott Marshall joined the W3C HCLS interest group, where a knowledge base for Alzheimer disease and its biochemistry was eventually built. The knowledge base already contained 350 million triples in 2008. Now he is assisting in building up an even more challenging map: to connect a wide spectrum of laboratory research with clinical research and practices (translational medicine = ‘bench to bedside’). Marshall explains: “Many databases in the drug development pipeline remain isolated, including the ones for genes, proteins, metabolic pathways, drugs research, clinical studies and patient dossiers. There may be millions of concepts, but we are now defining the core-concepts – this will be less than a million.”
Another goal is developing so-called inference machines that can make meaningful connections between all these triples. Machines that can make new triples from given triples without interference of the human mind. For instance: one of the triples in a map is ‘Molecule X has been used in medicine Y’. Another triple is ‘Mister Z has an allergy to the molecule X’. The triple ‘Mister Z shouldn’t use Medicine Y’ is not in the map. However, an inference machine with built-in rules such as ‘in case of the relationship allergy, find the medicines which have the compound’ could find this meaningful relationship.
Frank van Harmelen’s research group has written a program for an inference machine. This machine has recently analysed the 1.6 billion triples in the protein database Uniprot. Many unexpected interactions between proteins were noted. But one of the things that has yet to be checked is how reliable or meaningful these new relationships are. Van Harmelen: “Some people say: the facts are correct, the rules are correct and so the new triples are also correct. Others fear that that too many facts and rules are not correct enough to give reliable new conclusions.”

OPENNESS IS CRUCIAL Openness about the data and software is a crucial success factor. Therefore, an ‘Open collaborative environment to jointly address the challenges associated with high volume data production (....),’ is the mission of the Concept Web Alliance, according to its Declaration of May 8th 2009. Van Harmelen assures us that the CWA will continue in this effort. However, CWA is only one of several semantic web initiatives and life scientists groups in the world. So in general, he cannot yet foresee how open other maps will be. It is conceivable that maps belonging to companies will be closed.
“The W3C HCLS Interest Group is more or less open,” explains W3C co-chair Scott Marshall. W3C consists of about 100 participants including participants from six pharmaceutical companies.
W3C fees vary from about 1000 euro annually to 65,000 euro, depending on the annual revenues, type and location of headquarters. In addition, active membership means participating in the (bi)weekly phone meetings, joining the discussions on the mailing list and, if possible, participating in the face to face meetings. However, a scientist can also participate as an ‘Invited Expert’. “We are quite liberal about this,” says Marshall, “because we want to broaden our initiative.” But certainly, after a while the invited guest is expected to decide on his or her membership, like in a badminton club.
So money *and* time can be restrictive factors in completing concept maps. Marshall: “You have to realise that many professionals are putting hours and hours into the concept maps at the moment. Many fit in the work in the evenings and at the weekends without getting paid for it.”

ACKNOWLEDGEMENTS
We thank the interviewees for their contributions to the discussion on usefulness and the need for constructing a Semantic Web to manage life sciences information.
- Andrew Gibson: project leader for the BioExpert system and at the University of Amsterdam.
- Frank van Harmelen: professor at the Computer Science department of the Vrije Universiteit, Amsterdam.
- Scott Marshall: assistant professor of Bioinformatics at the Leiden University Medical Centre and co-chair of W3C HCLSIG.
- Barend Mons: scientific co-ordinator of Concept Web Alliance (CWA), leader of NBIC BioAssist Programme and member of NBIC Management Team.