Nano-Publication in the e-science era

Barend Mons[1,2,3] and Jan Velterop[1,2],

[1] Concept Web Alliance, [2] Netherlands BioInformatics Centre, [3]Leiden University Medical Center.
barend.mons@nbic.nl, velterop@conceptweballiance.org

Abstract. The rate of data production in the Life Sciences has now reached such proportions that to consider it irresponsible to fund data generation without proper concomitant funding and infrastructure for storing, analyzing and exchanging the information and knowledge contained in, and extracted from, those data, is not an exaggerated position any longer. The chasm between data production and data handling has become so wide, that many data go unnoticed or at least run the risk of relative obscurity, fail to reveal the information contained in the data set or remains inaccessible due to ambiguity, or financial or legal toll-barriers. As a result, inconsistency, ambiguity and redundancy of data and information on the Web are becoming impediments to the performance of comprehensive information extraction and analysis. This paper attempts a stepwise explanation of the use of richly annotated RDF-statements as carriers of unambiguous, meta-analyzed information in the form of traceable nano-publications.

Keywords: Semantic web, rich RDF-triples, disambiguation, publication.

1 Introduction

This paper is paradoxical: it is a paper in classical format that seems to make a plea for the ending of precisely such textual classical publication. For two reasons, this is only seemingly a contradiction: a) a paper like this is a plea, not a research paper, and therefore relies on verbal reasoning more than a presentation of research results usually does, so it is a full paper and not a set of nano-publications; and b) full papers may not be suitable any longer for efficient dissemination and exchange of knowledge, but they are suitable, perhaps even essential, for the detailed record. The point is made that sets of nano-publications are more suitable to the presentation of the relationships between research data and efficient exchange of knowledge than traditional papers.

To avoid additional redundancy, the scope and acceleration of the information abundance in biomedical research will not be addressed in this paper as such. Suffice it to say that the feeling that we are drowning in information is widespread and that we often feel that we have no satisfactory mechanisms in place to make sense of the data generated at such a daunting speed [1, 2]. Some pharmaceutical companies are apparently seriously considering refraining from performing any further GWA (genome-wide association) studies (also referred to as WGA – whole genome

association – studies, to drive home the point that disambiguation is needed) as the world is likely to produce many more data than these companies will ever be able to analyze with currently available methods (personal communication).

The dawn of the Semantic Web Era has brought a first wave of reduction of ambiguity in the Web structure as terms and other tokens are increasingly mapped to shared identifiers for the concepts they denote. Initiatives like Linked Open Data [3] have gone a long way to connect Web resources at the 'concept' level, rather than at the term level, as done by Google and other word based systems. However, the redundancy of factual information in the Web is still very substantial. In practice, it does not help a current biologist much to know instantly that there are 800 data sources for each gene in a list from his last micro-array experiment, all containing relevant information.

Classical publication on paper, even when converted to electronic formats, has not even begun to seriously exploit the possibilities that Web Publishing, even in its current, still early stage of development, has opened up. Yet most available so-called electronic publications are mere analogues of the paper versions, and often only in PDF. Terms are rarely, if ever, mapped to unambiguous concepts and, together with the habitual repetition of factual statements in each consecutive paper for the sole purpose of human readability, analysing scientific information with computers can currently not be considered in any way close to its potential. As computers will likely play an ever more important role as our reading devices in the (near) future, it is incumbent upon the research community to start making all text and database records truly computer-readable.

Computers can deal extremely efficiently with structured data. Unfortunately, people seem to dislike structured data entry, as evidenced by their reluctance to do it, and that is where the central problem of classical publishing arguably lies.

Here we develop a stepwise approach to data interoperability across language barriers, jargon, database formats, and eventually, ambiguity and redundancy. The basic principle is: natural guidance of human authors to structure their data in such a way that computers understand them. It should be clear that the 'semantic web' as we know it, is only a first step as it does not address as yet the *a priori* disambiguation of language and data records and it does not (yet) solve the redundancy problem. A meta-analyzed semantic web may go a long way to solve these major scholarly communication problems in the 'terabyte-per-experiment' phase of science, particularly life science.

2 Steps to be Taken

2.1 The First Step: from Terms to Concepts

In order to understand the problem of using many tokens to refer to the same concepts, the Ogden Triangle offers a good guide [see figure 1]. The concept is the

(essentially non-lingual) Unit of Thought. Tokens are all terms or identifiers used to refer to a concept, and many concepts have an 'object' in the material world (a specific person for instance), while many are only intellectual concepts and can be intellectually or physically experienced, but not be measured or touched, such as 'love'.

To refine the definition somewhat, a concept is the smallest, unambiguous unit of thought. The addition of 'smallest and unambiguous' may seem overkill and implicit in the general definition of a concept, but it should be emphasized that for proper scientific reference concepts should be defined to such a level of granularity that they are really unambiguous in the minds of all researchers working in a certain domain. This means for example that when two iso-forms of a given protein are discovered, both the general protein and the two iso-forms should be treated as separate concepts. Also, different languages capture different numbers of concepts with the same homonym. For instance, in Dutch only one word is known for the classical Greek concepts of έρως (eros), φιλία (philia) and αγάπη (agape). Although each may be translated into Dutch as 'liefde' (love), they denote clearly distinct concepts. Unless we remove the ambiguity of the word 'liefde' (love) in Dutch, we will never be able to express the richness of information in classical Greek.

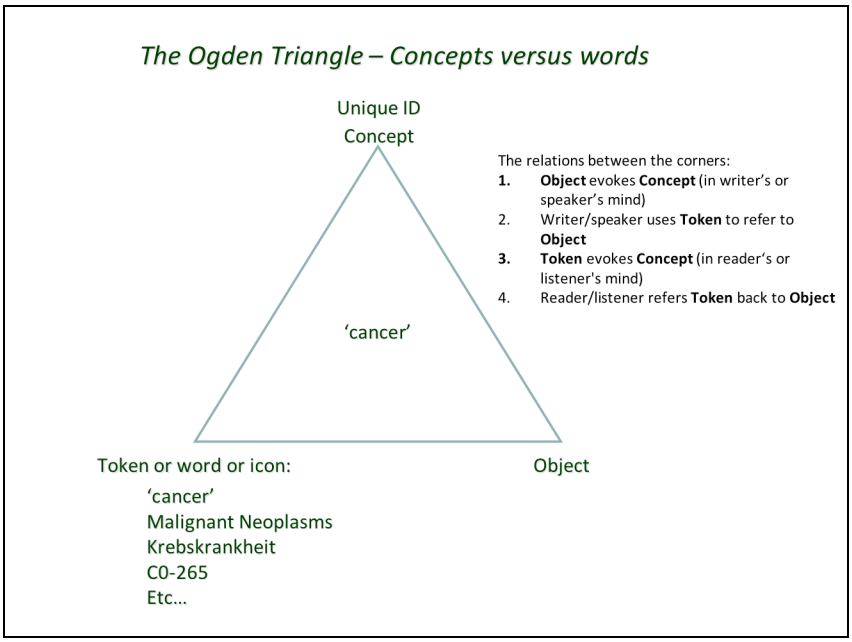

Fig. 1. Ogden Triangle: the relationship between the unit of thought, the tokens referring to it and the object in reality.

Biomedical science is plagued by ambiguity and classical publishing has not been able to ameliorate this [4]. And for good reasons: readability for humans actually

increases when the same concept is denoted in text by various different synonyms (the often used quasi-rhetorical rule of aesthetics is even to avoid using the same word twice in a sentence). The widespread use of acronyms has exacerbated homonym problems in the scientific literature, a problem we now try to alleviate to a degree by text-mining, disambiguation and more recently, structured digital abstracts [5]. Ideally however, each concept denoted in web-text or databases by terms or identifiers (together referred to here as 'tokens') should be unambiguously mapped to a universally resolvable concept reference, which represents the unambiguously defined unit of thought. This can be done *a priori* (capture reference numbers up front during the editing and review process) or *a posteriori* (text- and data mining, sometimes combined with human curation). The former approach is only attempted in embryonic form to date, including by publishers who request authors to use accepted identifiers and symbols for genes and proteins and the emerging 'structured digital abstract approaches' about which we will say more later on.

Many groups and initiatives have treated interoperability almost as synonymous with 'defining standards'. Obviously it is a truism that if everyone were strictly to adhere to standards and structured data entry, data would become interoperable and computable. However, this approach does not take into account that: a) at the very moment that the community understands the need for standards in a given domain, multiple standards are likely to be developed and subsequently everyone is likely to start to defend their own standard, and b) the nigh ubiquitous human character trait that makes us, consciously or subconsciously (or on occasion just lazily) ignore standards. We strongly believe in the process of 'bottom up standard emergence', a process by which useful and intuitive standards emerge from joint community action. Therefore we have proposed in the Concept Web Alliance (CWA) [6] to develop systems that, instead of choosing, or developing standards, will take an approach that aims to accommodate all standards developed so far. This means that we first need to map all tokens to the relevant concepts in Life Sciences, and that we can subsequently accept all non-ambiguous identifiers denoting these concepts as long as they are properly mapped to a universal reference in a public environment, which is 'owned' and governed by the user community. The same is true for the next step: creating interoperable statements.

2.2 The Second Step: from Concepts to Statements

Essentially, each smallest insight (as opposed to smallest unit of thought) in exact sciences is a 'triple' of three concepts. However, we will argue here that three concepts are usually not enough to make a statement clear enough to be always placed and used in the correct context. First of all (hence step 1) the three concepts in the triple should be indeed unambiguously defined, and therefore terms and even sometimes identifiers will not suffice as tokens for the constituting concepts unless they are absolutely unambiguous and correctly mapped. It is therefore important to know for any n-gram (term consisting of one to ***n*** words/tokens) whether it is ambiguous (denoting more than one concept) or not. We will address this issue in more detail when we describe the Concept Wiki below.

Once we have unambiguously defined the three constituting concepts of a statement, a form for interoperability of statements should be found.

The central format of choice in the CWA so far is to exchange statements in the form of richly annotated RDF triples [7]. The annotation will be described in the next session. Here we wish to emphasize that also the choice for RDF is pragmatic, not dogmatic. If partners wish to express the statements in, for instance, XML format, they can still be translated, even on-the-fly, into a format that can be processed by all other tools using the same data. Figure two depicts a number of examples of triples of concepts forming a statement. It is clear from these examples that the notion of a concept applies to many more units of thought than the 'classical types of biomedical concepts, such as genes, diseases, drugs et cetera. Each person and each of the over 18 million scientific articles are regarded as a concept.

```
<rdf:Description
rdf:about="http://www.nbic.nl/cwa/relation/C0035820#C0060383#1240641052059">
<cwa:typeRelation rdf:resource="http://www.nbic.nl/cwa#cooccurrence"/>
<cwa:strength>0.0625</cwa:strength>
<cwa:has_query>limb girdle</cwa:has_query>
<cwa:discovered_by rdf:resource="http://www.nbic.nl/cwa#TripleMiner"/>
<cwa:timestamp>1240641052059</cwa:timestamp>
<cwa:annotation rdf resource="http://www.uniprot.org/uniprot/flnc_human">
</rdf:Description>(free text?)

<rdf:Description
rdf:about="http://www.nbic.nl/cwa/relation/C0035820#C0060383#1240641052059">
<cwa:typeRelation rdf:resource="http://www.nbic.nl/cwa#cooccurrence"/>
<cwa:strength>0.0625</cwa:strength>
<cwa:has_query>limb girdle</cwa:has_query>
<cwa:annotated_by rdf:resource="http://people.conceptwiki.org/index.php/Concept:85094810"/>
<cwa:timestamp>1240641052059</cwa:timestamp>
<cwa:annotation rdf resource="http://www.uniprot.org/uniprot/flnc_human">
</rdf:Description>(free text?)
```

Fig. 2. The first example is a triple describing the connection between two concepts as mined by a custom designed triple miner, from Uniprot. The second triuple has been annotatied by a person referred to as concept 85094810 as the opaque reference number. When following this URL, the user will find out that the person annotating this triple was Prof. Johan den Dunnen, a top expert in the two concepts referred to in the triple.

Based on current knowledge and ontologies, we estimate that the initial set of concepts in the life sciences includes more than 3 million 'classical' biomedical concepts, close to 2 million unique author names (in Pubmed alone), over 18 million articles (with a DOI) and around 20 million small molecules. With the additions expected in the years to come we predict that a Concept Wiki as described below will contain at least 50 million unique concepts and many more terms. If the ratio terms/concepts is estimated as roughly the same as in UMLS [8], we anticipate more than 200 million terms in English alone. With the addition of more and more languages, the synonyms of these terms in minimally 25 main languages should be mapped to the concepts, leading to an estimated $5x10^8$ tokens (terms and identifiers). As statements (in the form of concept triples) are composed of concepts and not terms we still 'only' deal with 50 million odd concepts. The number of 'realised' triples, that is to say, the representation of what we, collectively, have stated so far in biomedical research history is estimated currently to be around 10^{14} (personal communication F van Harmelen, and see [9]).

2.3 The Third Step: Annotation of Statements with Context and Provenance

It is not enough to store statements just in the form of their basic components, three concepts in a specific sequence, indicating *subject* > *predicate* > *object*. It is obvious that a statement only 'makes sense' in a given context. The context is in fact defined by another set of concepts. If a dogmatic triple approach were chosen, each connection would again be a triple and the triple store representing biomedical knowledge in RDF would explode. Without pre-empting the conclusions of the CWA working group on triple structure [10] we here reflect earlier discussions in the CWA that led to the approach of 'richly annotated triples', a term in fact standing for disambiguated, non-redundant statements in proper context and with proper provenance. Statements should be treated as the smallest building blocks of ontologies, and also as the principle building blocks of pathways, semantic networks, and 'on-screen hypotheses' in e-science. Methods to format, store, browse and reason with RDF statements are being discussed in specific CWA working groups [11] but are outside the scope of this paper.

Most statements are conditional. A statement such as `malaria > is transmitted by > mosquitoes` although 'as true as it comes' in science, is still conditional, since it is unidirectional, as it is clearly not true that '`malaria > transmits > mosquitoes`'. The statement '`DMD < > interacts with < > SNT1`' is an example of a truly 'symmetrical' or bidirectional triple. In both cases, the (sometimes ambiguous) terms in the triple are represented in the RDF version as universal references to the concepts. Daughter-concepts such as *`Plasmodium falciparum`* (> `is form of` > `malaria`) and *`Anopheles Gambiae`* (> `is species of Culicidae` > `is species of mosquito transmitting malaria`) can be ontologically mapped to the parent concept, so that the textual statement: '*Plasmodium falciparum* is transmitted by *Anopheles gambiae'* can be treated as another instance of the general statement 'malaria is transmitted by mosquitoes'.

Many statements are also only 'true' or 'relevant' under certain conditions. Not just physical conditions, such as a given PH, but also, for instance, true only for a given species, or in a certain tissue, or only if a protein is truncated because of a mutation in the gene (now causing a disease). These 'conditions' can be annotated to the statement in the form of conditional concepts. There is in principle no limit to the conditional annotations of any given triple statement. It is, however, crucial that the annotations are also made with unambiguous concepts, so that reasoning, indexing, sorting and clustering of statements can be performed, based on their basic three constituting concepts as well as on their annotation concepts.

It is no-doubt possible to commit the 'sin of exceptionalism' [12] and find statements that cannot be expressed in the proposed format, but we argue that – assuming that we can get the format rich enough – virtually every insight in the exact sciences, and probably even in the humanities, can be captured as a richly annotated RDF statement and begin to form an element of ontology building or reasoning.

Provenance is included here in the context of a statement. Typical provenance information includes (`typewriter font` = concept): `person` who made the statement, `source` from which the statement was mined (e.g. `UniProt` or

`Journal X`), `date` on which the statement was made (time-date-stamp), '`ownership`' (see below for nuance on copyright issues), and most importantly: `status`. Status can include: `community`, `authority`, `peer-reviewed`, `curated`, `disputed`, `retracted`, `hypothetical`, `observational`, `repetitive`, *et cetera*.

2.4 The Fourth Step: Treating Richly Annotated Statements as Nano-Publications

In a scientific context, publications are only publications if they are citeable and appropriate credit is given to the authors. There is no intrinsic reason why such publications need necessarily be full-length papers. Published contributions to science can be as short as single statements that interpret data, and yet be valuable to scientific progress and understanding. If and when such contributions could be properly attributed and credited, the incentive to publish them would increase, and with that quite conceivably the speed of dissemination of useful research results. We distinguish the following types of statements that would be suitable for what we call 'nano'-publications:

Curated Statements (Essentially Annotations). Some statements represent 'facts', Obviously, any fact in science only remains a fact until progressive insights may prove the statement wrong, but curated triples (such as curated protein-protein interactions in Uniprot) are 'as true as it gets' in science, meaning that they are conform current scientific insight. Usually, these 'curated statements' are seen as the typical building blocks of ontologies and more simple thesauri. Examples are for instance that '`breast cancer > is a form of > cancer`' and '`DMD < > interacts` with `< > SNT1`'. In the case of curated statements, usually, such statements can be found in formalized databases such as OMIM (Gene-disease), UNiProt (protein and protein-protein interaction) or GO (gene-protein-function). Curated triple statements should ideally have provenance data associated about both the originator of the triple (usually the first co-occurrence of the two 'telomeric' concepts) and the curator(s), as both should receive credit.

Observational Statements (Co-expression, Co-occurrence, Statistical). Many factual statements, including the well-established fact that '`malaria > is transmitted by > mosquitoes`' do not have a 'curated instance' somewhere, in many cases simply because there is no database dedicated to this class of triples. One of the goals of community annotation [4] is to 'elevate' as many factual statements in the current biomedical literature from 'observational, usually mined by co-occurrence-based methods, to 'curated'. However, there are more sources for 'observational' connections between concepts than the literature. A prime example concerns data regarding co-expression of genes originating from large numbers of differential expression experiments around the world. The expression profiles of such experiments are increasingly shared with the global research community in databases such as GEO [13] and Array-Express [14]. If two genes are consistently correlated in

their expression pattern without a clear biological explanation found as yet, their co-expression pattern is the basis for an observational triple of the class *'expression correlation'*. Without trying to be exhaustive here, one more example could be a statistical correlation between a *locus* or a *genomic region*, with a given genetic disorder. Obviously, the more observational triples can be elevated to the status of consolidated, curated statements with proper annotations detailing context, conditions and provenance, the richer biomedical ontologies of established knowledge will become.

Hypothetical Statements (Inferred by Established and Published Algorithms). A third, probably most intriguing, category of triples may be what we call 'hypothetical' triples. These concept combinations have been inferred from text or data mining or from direct reasoning with existing triples to generate new, hitherto non-observed triples that are likely to represent undiscovered statements with high probability to be 'true'. Esoteric as this may sound, the Biosemantics Group in The Netherlands has, in a recent paper, predicted many unknown protein-protein interactions to be 'real' even if the two proteins in the triple do not have co-occurrence in the discoverable literature [15]. The paper contains evidence that some of the predicted interactions could be confirmed in the wet-lab, to the surprise of the experts working on these proteins for many years. Once such triples – properly annotated with the *algorithm* used for prediction, statistical likelihood (with a threshold) and provenance – are collected in a central triple store, they can become a rich source for *in silico* knowledge discovery without expensive wet-lab experiments up front.

In terms of publication, it is conceivable that in-text semantic support tools as shown in the on-line version of the paper mentioned can reveal predictions even during the typing process of a new scientific paper. In fact this example is so close to reality that recently, a novel paper sent for review to one of our collaborators independently reported a new protein-protein interaction contained as prediction in our recent paper. In the proposed situation, where triples of all categories described above are treated as nano-publications, the hypothetical triples would be citeable and the authors of the paper could be credited for the prediction. Obviously the authors confirming the protein-protein interaction in reality would still get the credits for their wet lab experiments.

2.5 The Fifth Step: Removing Redundancy, Meta-analyzing Web-Statements (Raw Triples to Refined Triples)

It may be obvious, particularly for people familiar with the Semantic Web and initiatives like Linked Open Data [3] and the 'shared names initiative' of the Semantic Web Health Care and Life Sciences (HCLS) Interest Group [16] that in principle, with proper concept mapping, the ambiguity currently crippling e-science can be dealt with. Much work still has to be done, but there are no major intellectual hurdles left, as will be argued in the practical section of this paper.

However, unambiguous data linking is not enough; it is not at all useful for biological researchers to be presented with the evidence that for the 100 genes emerging from their high-throughput studies there is an average of 100 papers and 20 database records containing additional information on each of these genes, simply because it is impossible to read 10,000 records. The good news is that a major part of the information in those records, once converted to universal triples, appears to be redundant. Research in text mining and information retrieval has shown that repetition of statements in scientific publications in the broadest sense has some merit (likelihood of being reproducible increases) [17].

Beyond a certain number, further repetition of 'established' facts is good for linear human reading, but it is not useful for computer assisted *in silico* discovery processes. Even pure copying or re-annotation of, for instance, protein–protein interactions in IntAct to UniProt has merit for the likelihood that an experimentally observed interaction actually represents a biologically meaningful interactive process. The fear that 'blind copying' of statements of earlier discoveries in scientific papers in fact makes us 'standing on the shoulders of bias' rather than of giants, falls outside the scope of this paper, but this phenomenon could be very well studied once redundancy of statements is properly documented.

In any case, treating RDF statements as nano-publications and properly acknowledging and crediting them, will require this analysis and where necessary the removal of undue repetition and redundancy. An illustrative example again: when the community annotation paper [4] was published in genome Biology [28 May 2008], the number of co-occurrences between malaria and mosquitoes in PubMed was 5018 [4]. About 14 months later, the co-occurrence of malaria and mosquitoes (just in abstracts) is 6470. Assuming that the majority of the 1452 new co-occurrences repeat the statement '`malaria > is transmitted by > mosquitoes`' in some form, the fact will not change. However, it is illustrative that the most recent PubMed entry about malaria and mosquitoes at the day that this section of the paper was written, states in its first sentence: "Despite their importance as malaria vectors, little is known of the bionomic of An. nili and An. Moucheti" [18]. It is therefore important to note, although it is ontologically known, that *Anopheles nili* and *Anopheles moucheti* are mosquitoes of the genus Anopheles, which is an important genus in terms of malaria transmission. The fact that both species may play a role in malaria transmission is only implicit in this abstract. Please also note that the token *An. nili* is not a preferred term to refer to this species of Anopheles. If the specific triple: '`Anopheles nili > transmits > malaria`' is known in the triple store, and we know that this is in fact another instance of the more generic statement that malaria is transmitted by mosquitoes, an alert on this triple, which would be superfluous, could be avoided. However, in case this should be the first co-occurrence between *Anopheles nili* and malaria, an alert to all malaria-interested scientists would be justified and most likely welcomed.

With more and more 'grey literature' being made available on the Web, not only in, for instance, Wikipedia, but also in patient blogs**,** and a plethora of web sites about health related subjects, it is increasingly important to be able to detect undue repetition, such as mere parroting, but also to detect 'new co-occurrences' at the earliest possible time. New co-occurrences may represent new statements. New statements may range from major scientific discoveries to complete nonsense. It is not

very difficult to reference the triple store to find out whether two concepts have ever been mentioned in the same sentence before, and it is also not very complicated to detect the 'stress' a certain statement may introduce in a semantic concept map. However, apart from statements that are ontologically illogical, like a *human gene* being *expressed* in *wings*, it is very difficult to judge whether a statement is wrong or misleading as opposed to a novel finding. Therefore it is crucial for nano-publication in the form of single statements to allow for *a posteriori* annotation of RDF statements.

The meta-analysis of individual RDF statements to remove redundancy, create concept maps, cluster meaningful statements and include observational triples as well as hypothetical triples in such meta-analyses, would lead to a growing, dynamic concept web which should be very easy to access, browse and analyse. Nano-publication of new triples in all three categories should lead to real time alerts to scientists who have indicated that they are interested in one of the concepts in the statement or in closely related areas of this 'concept web'. With appropriate recognition and traceability of the statements this could enable an entirely different way of scholarly communication, much more adapted to the current rate of data production.

3 Practicalities

3.1 The Concept Wiki

The Concept Wiki contains concepts as 'units of thought'. Those are differentiated from 'tokens', which can be the words or expressions in language that describe and refer to concepts (linguistic tokens), but also the various identifiers that refer to the concept in, for instance, databases (numeric or alphanumeric tokens). For example, the concept of a certain specific malignant skin lesion is described by the linguistic token 'Melanoma' in English (in this case quite a few other languages also use the same word), by the alphanumeric token DOID:1909 in the human disease ontology, by the alphanumeric token NCI/C0025202 in NCIT, and quite likely by other linguistic tokens (words) in other languages and alphanumeric tokens (identifiers) in other ontologies and databases.

In the Concept Wiki (www.conceptwiki.org), concepts and their various tokens are associated with one another so that interoperability and mapping between different identifier systems and languages is facilitated.

The Concept Wiki will contain, for each concept, an anchor page with a random unique numeric reference number. This page will contain, *inter alia*, the following information:

- Originating ontology or ontologies (also indicates domain, by implication)
- Preferred term in each of those ontologies
- Synonyms in English and links to synonyms in languages other than English (e.g. in the current OmegaWiki, www.omegawiki.org).

- Each language is also a concept and will also have a unique numeric reference number, but in the ConceptWiki anchor pages the languages will be shown in ISO 639-3 (e.g. ENG for English; NLD for Dutch; ZHO for Chinese [Zhōngwén]) and also the language name in English, if it exists, and its native name, where possible, for convenience.
- Mapping to concept IDs in any of the ontologies in which the concept is included (e.g. [1234567890] [UMLS-ID] [CO14897]). "UMLS-ID" is also a concept and will also have a unique numeric reference number, but in the ConceptWiki Anchor pages the mnemonic term for such identifiers will be shown for convenience.
- Other functional, structural, and physical information, where relevant
- Conceptual and terminological information
- Reference information
- Tags (such as semantic type of the concept, domains in which it is relevant, each again concepts by themselves)

Each Concept Anchor page will have a URI that incorporates the unique numeric reference number, e.g. http://conceptwiki/123909473890 (exact URI format not yet established). We intend to prepopulate the Concept Wiki with the more than 3 million 'classical' biomedical concepts, millions of chemical concepts and close to 2 million unique author names mined from PubMed, as an initial step to reach the critical mass needed to make the Concept Wiki useful. We also intend to place the Concept Wiki in the public domain, indicated by the Creative Commons so-called 'CC Zero Waiver + SC Norms', indicating that copyrights are waived, but that adherence to scientific community norms regarding attribution and citation are expected (but crucially, not laid down as a contractual obligation). In this way, the community can be regarded to 'own' the Concept Wiki and the Concept Reference Numbers in it.

3.2 New Ways of 'Valuing' Scientific Contributions

As said above, citeability and credit to authors are of prime importance to the way the scientific publishing system works. Annotated statements, as described earlier, are both citeable and credit the authors. This is the case whether or not they are contained in a regular peer-reviewed journal article or in other media, such as curated databases, and even in informal publications or databases, where subsequent annotations may perform the function of peer-review.

The annotations themselves, in turn, can also be credited to those who contribute them and be citeable, which opens up the possibility that those who are not in the position to have their papers published in prestigious journals – for instance because they live and work in countries that do not quite have the research infrastructure to facilitate top level science – can still build up a public record of their contributions to science. Especially for scientists in the developing world this may be a welcome addition to the possibilities they have for sharing their knowledge and insights in a structural way.

3.3 The Role of Traditional Publishers, Institutional Repositories, Libraries and Funding Agencies

While arguing that research results should be available in the form of nano-publications, we are emphatically not saying that traditional, classical papers should not be published any longer. But their role is now chiefly for the official record, the "minutes of science" [19], and not so much as the principle medium for the exchange of scientific results. That exchange, which increasingly needs the assistance of computers to be done properly and comprehensively, is best done with machine-readable, semantically consistent nano-publications.

One should not consider classical publications and nano-publications to be two entirely different things. Classical papers are full of statements, and therefore contain nano-publications. It is just that they are not semantically coded in a way so that they are recognised as such. Traditional publishers and repositories should have their material semantically coded. Not just new material, but it should also be done retrospectively for all the content that is in electronic format. The technology exists, and it is not expensive to have it done. Bearing in mind that each of these nano-publications can be linked to or from other publications or web sites, they are in effect citeable items and can contribute to the visibility of a paper and the journal it is published in. Services that provide science metrics, such as Thomson/Reuters' Web of Science and Elsevier's Scopus, would do well to incorporate these citations into their analyses and rankings.

Should publishers be reluctant or unwilling to semantically code the content they publish, all is not lost. The technology exists to provide Web browsers with the functionality for users to identify meaningful statements – nano-publications – and annotate them. Libraries could have such browser plug-ins installed throughout their computer networks, and so contribute to an increase in the efficiency and value of knowledge exchange. In this case, of course, what is being identified and annotated is purely up to the users, and publishers lose control.

Authors and their funders should start requesting and expecting the papers that they have written and funded to be semantically coded when published. The efforts are so small and the benefits so great. But the greatest impact should come from funders re-adjusting their current focus which often is mainly on data generation, even when much of that data is deeply sub-optimally usable because it cannot properly be analyzed, shared or used to build further research upon. The funders' attention to proper storage and availability of data generated with their financial support, in widely usable formats, is urgently called for. Even if the amounts set aside to make data much more interoperable are minute, if seen per data entry, the cumulative amount would have the potential to make the infrastructure possible to discover the knowledge contained in these data much more efficient and effective.

3.4 Community Ownership

Whilst in principle nano-publications extracted from classical papers would be subject to copyright, this could in practice only be used to ensure proper acknowledgement. Putting up payment or legal barriers to access would not be tenable. Imagine the information *"this statement made by author X published in article Y in journal Z"* being put behind tollgates. That would be the same as putting the information that *"this book was written by author X and is published by publisher A"* behind tollgates. Publishers are, presumably, wiser than that. Nano-publications that are rich semantic triples are in essence references, and wide and open availability of references to the content they publish is what most publishers crave. Nano-publications are therefore necessarily open access. And this open access is actually beneficial not just to scientists, but to publishers as well.

Acknowledgements

We would like all colleagues participating in the Concept Web Alliance (CWA) for discussions and insights leading to this consolidated view. However, we take sole responsibility for any statement made in this paper, and it does not necessarily represent the view of any of the Concept Web Alliance partners. We thank NBIC, LUMC and the Bill Melton Foundation for the early funding of the CWA.

References

1. Dennis, C., Biology databases: Information overload, Nature 417 (2002) doi:10.1038/417014a
2. Stokstad, E., Information Overload Hampers Biology Reforms, Science (2001): Vol. 293, no. 5535, p. 1609
3. Linked Data, http://linkeddata.org (accessed on 21 September 2009)
4. Mons B, Ashburner M, Chichester C, van Mulligen E, Weeber M, den Dunnen J, van Ommen GJ, Musen M, Cockerill M, Hermjakob H, Mons A, Packer A, Pacheco R, Lewis S, Berkeley A, Melton W, Barris N, Wales J, Meijssen G, Moeller E, Roes PJ, Borner K, Bairoch A., Calling on a million minds for community annotation in WikiProteins, Genome Biology 2008; 9(5):R89 (2008)
5. Gerstein, M., Seringhaus, M., Fields, S., Structured digital abstract makes text mining easy, Nature, Vol. 447 (2007)
6. Conceptweblog, Concept Web Alliance Declaration, http://conceptweblog.wordpress.com/declaration (accessed on 21 September 2009)
7. W3C RDF Core Working Group, http://www.w3.org/RDF (accessed on 21 September 2009)
8. National Library of Medicine, Unified Medical Language System, http://www.nlm.nih.gov/research/umls (accessed on 21 September 2009)

9. Van Harmelen, F., Slideset: LarKC the large knowledge collider, http://www.slideshare.net/Frank.van.Harmelen/larkc-the-large-knowledge-collider (accessed on 21 September 2009)
10. CWA working group 2.6: triple model, http://www.myexperiment.org/groups/192 (accessed on 21 September 2009; access free, registration required)
11. Conceptweblog, Concept Web Alliance Groups, http://conceptweblog.wordpress.com/groups/ (accessed on 21 September 2009)
12. Goble, C., Slideset: The seven deadly sins of bioinformatics, http://www.slideshare.net/dullhunk/the-seven-deadly-sins-of-bioinformatics (accessed on 21 September 2009)
13. NCBI GEO (Gene Expression Omnibus), http://www.ncbi.nlm.nih.gov/geo/ (accessed on 21 September 2009)
14. EMBL EBI Array Express, http://www.ebi.ac.uk/microarray-as/ae/ (accessed on 21 September 2009)
15. Van Haagen, H., et.al (in press)
16. W3C Semantic Web Health and Life Sciences Interest Group, http://www.w3.org/blog/hcls?cat=85 (accessed on 21 September 2009)
17. Spence, D.P., Owens, K.C., Lexical co-occurrence and association strength, Journal of Psycholinguistic Research, Vol. 19, no. 5, pp. 317-330 (1990), doi: 10.1007/BF01074363
18. Antonio-Nkondjio, C., Ndo, C., Costantini, C., Awono-Ambene, P., Fontenille, D., Simard, F., Distribution and larval habitat characterization of Anopheles moucheti, Anopheles nili, and other malaria vectors in river networks of southern Cameroon, Acta Tropica (2009), doi: 10.1016/j.actatropica.2009.08.009 (Corrected proof, available online since 13 August 2009)
19. Velterop, J., Keeping the Minutes of Science, in: Proceedings of Electronic Libraries and Visual Information Research (ELVIRA) Conference, Aslib, London, No. 2 (1995)