Finding the critical path: applying the semantic web to drug discovery and development

There is much happening in drug development these days: new target classes, increasing costs, new forms of high throughput screening, and the promise of personalised medicine. But of all the issues, the one that now takes centre stage (and directly impacts the aforementioned ones) is drug safety. It accounts for 70% of late stage drug failures, and is estimated to contribute to 80% of the overall cost of drug development, a whopping $0.8 billion per first in class drug1. There is no obvious sign of this coming under control, and appears to be the manifestation of needing to focus on new targets, the economics of pharmaceuticals, market pressures and changing public opinion. The industry's R&D arms needs to confront this challenge immediately by becoming not only more efficient but smarter as well.

By Dr Eric Neumann

There is no shortage of intelligent, industrious minds in the pharmaceutical industry, so how can one improve on this talented base? It should be obvious to the reader that employing smart people has no guarantee that the organisation as a whole is utilising all of the scientific and social knowledge efficiently. Although meetings and teleconferences are (part of the daily repertoire) the common tools of the trade, some forms of knowledge are not handled well by such forums; much of the dynamics remains tacit at the end of the day. And even though much consideration is given to document management systems, the utility is dependent more on how effectively one finds a useful piece of knowledge when it is needed most. With the ubiquity of internal websites, e-mails and attached PowerPoint presentations, is there anything left that can be used to improve the knowledge environment substantially and over a wide range? I hope to show you that there is a lot that can be improved, as long as one is willing to address the meaning in information content, more precisely known as semantics.

Semantics is fundamentally not an information technologies issue – though it can help information technologies, as will be discussed below. It originates out of the need for groups of individuals to work together towards common goals (aka Communities of Practice), who must agree upon a set of meanings around terminologies, concepts, relations and actions they will be using. This is often not so straightforward, and a lot of

Figure 1
Basic elements of an RDF triple: things connected by properties to their things throughout the web

confusion arises before people realise whether they are talking about the same or different things. Coming to terms within a group over a set of meanings is referred to in knowledge management as 'Negotiation of Meaning'[2]. Since our use of information technologies permeates throughout all our communication (e-mail), analytics (computational and statistical tools), documents (content management), data storage (databases), and management (network administration) systems, we need to ensure that the agreed upon set of human semantics is consistent and well-supported by these information systems we have in place. This is a kind of community (or enterprise) digital harmonisation, and it empowers the users of information to define for themselves how to organise, share and retrieve information in order to best suit their needs. It is not identical to the much discussed 'data integration' issue, but as will be shown below it has a direct bearing on it.

It is worth remarking that artificial barriers have arisen between information systems and knowledge practices, as in scientific research. The reasons are historic in that a set of technologies (relational tables, search engines, word indexing) took advantage of encoding easily definable structures of data (text, numbers, lists), but were often not extended to handling open structures consisting of conceptual relations, such as assertions, refutations, conditional assumptions and hypotheses (ie, statements). Databases became very efficient for searches on cleanly chunked entities, yet were brittle to adding new relations and new data-types. As a consequence, scientists find it easier to dump information into spreadsheets or presentation slides than

Figure 2
Multiple ontologies and data models can be linked together by common objects and bridge concepts. Here several public databases are linked by their semantics and the Drug-Target Ontology, such that instance data can be connected together (http://www.w3.org/2005/04/swls/BioDash/Demo/ls-ont-1.0.rdf)

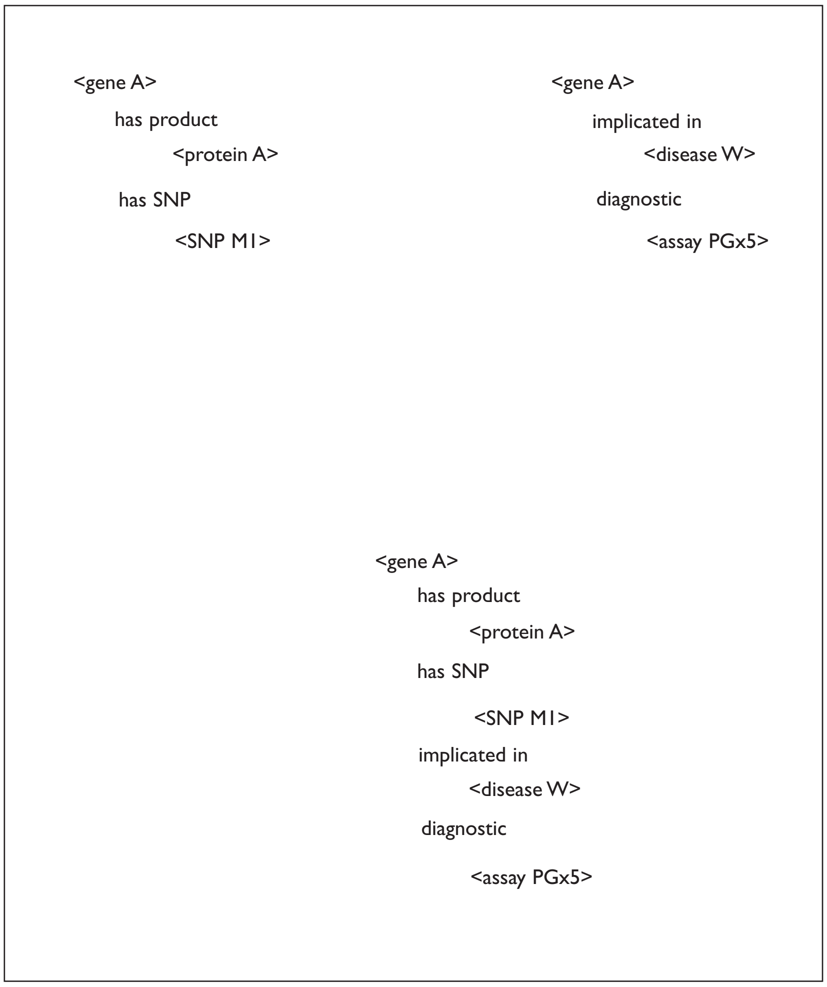

Figure 3
Two RDF documents containing the same subjects, when both retrieved, merge to form one common graph for the subject

to define data models that would allow the information to reside securely within databases. The result is that more real-world, context-specific knowledge is stored in personal laptops as Excel and Powerpoint documents and are not available to the rest of the organisation, forming 'knowledge cul-de-sacs'. They often become lost when people change positions or jobs, and the cost estimates are believed to be in the billions of dollars. Even if these documents are uploaded into document management systems the meaning behind the content is still inaccessible to search engines – searching is applied only to words.

So what would the incorporation of semantics into information systems get us? By explicitly utilising semantics whenever information is exchanged and stored, we are able to concisely capture object relations, categorise by concept, leverage experiential knowledge and represent theories and hypotheses as statements about things. Here is an example:

A researcher performs a search query for all recent papers that investigate the range of toxicity of a class of kinase targeting compounds. Had the researcher used a querying engine that supported semantics, she would not have had to read through dozens of abstracts. Nonetheless, once the papers are read, the researcher decides to categorise them along three dimensions: kinase group (target) specificity, tissue toxicity and mode of drug clearance. This can be done by adding semantic tags to the articles and storing them in some form of content management system. Once categorised this way, other researchers can locate these papers by querying along any combination of these three semantic dimensions. In addition, any extracted information from the research articles (through the application of text mining), can be also converted into semantics statements, and associated directly with the papers as well.

However, with a system able to take advantage of semantics, one can go even further: any new dimensions can be added by researchers without redesigning the database or rewriting code. These new dimensions are immediately available to others to use as well. If one earlier document overlaps with the kinase class of a new document, and the new one is tagged by tissue toxicity, the former document will become (indirectly) associated with this tissue toxicity once a similarity-link is made between the documents.

If all scientific and clinical information could be catalogued and inter-related this way using a web-based approach, much more relevant (eg, does this paper describe a better biomarker for cardiotoxicity?) and context-specific (eg, does this pertain to pharmacokinetics?) information could be made available throughout the web or across an enterprise's intranet. Such an approach is indeed being developed and standards already exist for use by applications across the information space.

What is the Semantic Web?

The Semantic Web (SW) (http://www.w3.org/2001/sw/) is a model for the web proposed by Tim Berners-Lee, inventor of the web and current director of the World Wide Web Consortium (W3C) (http://www.w3.org), to create a web that will support the universal exchange of all information through the incorporation of machine-readable meaning or semantics. The semantics can be directly linked to documents and data on the web[3,4], which can then be used by search engines, data organisers and knowledge management systems. W3C is helping direct Semantic Web activities through the definition of standards, mark-up languages and key applications.

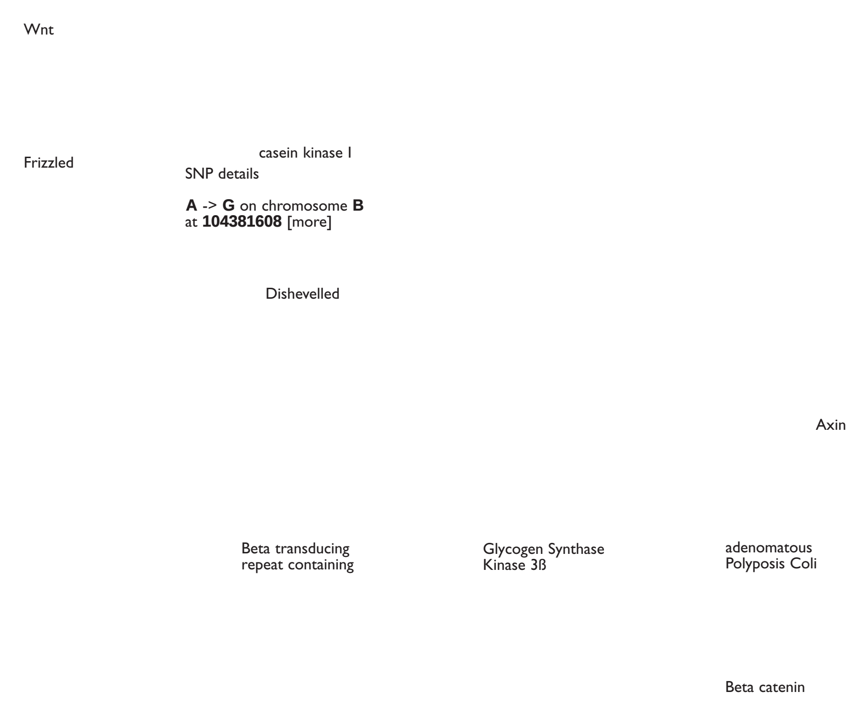

Figure 4
Visualisation in BioDash of composite knowledge from three domains: molecular pathways (BioPAX, cascade graph), pharmacological agents (drug-target model, compound squares), and SNPs (DB-SNP, purple dashes). The integrated graph is an RDF-OWL representation, and can be shared over the web

The foundation standards are the Resource Description Framework (RDF), a special form of XML (http://www.w3.org/XML) for describing objects and relations across the web, and the Web Ontology Language (OWL, based on RDF as well) for specifying ontologies (systems for defining concepts and relations). The OWL-defined ontologies are used to specify all types of objects and how they can or should relate to one another. These objects and their relations using real data can be automatically represented and interchanged as RDF, providing an open-standard for tying together databases and knowledge repositories. As an additional effect, RDF-OWL supports true information integration independent of where the information is located (see **Figure 2**).

In order for this system to work throughout the web, there needs to be a way of identifying every entity and relation uniquely in the web. The key ingredient that makes this possible is the Universal Resource Identifier or URI, for each entity referenced from a document (aside from literal text and numbers). For example, a gene would have a URI specifying it uniquely as a subject and where it could be found (similar to the URLs current browsers use). However, the gene is explicitly typed as 'Gene' and linked to it are a set of properties and relations (predicate-object) composed of URIs as well. Together the URIs form a system of subject-verb-objects statements, or triples, that define relations between things (**Figure 1**). The set of all triples creates a graph structure (node-edge-node) that can hold much more valuable information than simple web content, and more accessible to powerful query tools than Google™. Several applications of semantic web have been proposed for the life sciences[5]. As a basic example relating to molecular biology, consider the statement 'gene A has product protein B'. In SW triples form it would appear as:

<gene A> <has product> <protein B>

This states that subject 'A' of type gene has a protein product 'B'. If a set of such statements are created and published through the web, they can be aggregated by anyone who retrieves them, even if they exist in different documents or databases (see **Figure 3**). To avoid confusion if term 'A' means something different to another group, a namespace is prepended to make it unique on the web; for example, 'nlm:A' for 'A as defined by the National Library of Medicine'.

Not only is aggregation useful for end-users, but it is also a powerful interface for transmitting information between other machines tasks and web services. Since the relations are included in the transmitted information explicitly, SW supports what is called semantic interoperability[6] defined as "the ability to act upon information with a consistent understanding of the meaning of that information". This allows programs not familiar with all data types to be able to seek out more information in order to handle them properly. For example, if a clinical statistics tool does not immediately know how to handle 'microarray data', there is enough semantic information provided in the data bundle that it can decide whether to ignore the data or call other services that have specific knowledge of how to handle microarray data.

Semantic structures make it possible to go beyond basic data handling and begin to encode protocol prerequisites and actions as a series of business rules. Rules are defined using SW so that programs can handle and apply them. They can serve many important functions including:

- Model workflow and assure all the required data are collected at a particular point in time.
- Ensure protocol compliance.
- Present information effectively to decision bodies, and capture their decisions.
- Guarantee correct access only to parts of information based on agreed rights (group permission).
- Co-ordinate activities within alliances and establish an IP audit trail.
- Define and enforce of legal policies and prerequisites.

SW is being defined to address the need to define and follow policies in a network-based world (http://www.cs.umbc.edu/swpw/). It therefore seems reasonable that SW could support many of the broader protocol issues for drug discovery and clinical development that just a data standard cannot.

Semantic Web meets drug development

In drug development, there are many possible applications of SW[5], since each research domain (biology, chemistry, pharmacology, clinical) has its own set of semantics, as do the business processes (target validation, therapeutic strategy, compound progression, NDA submissions) that direct the research activities. Semantic definitions serve as 'smart glue' across all enterprise activities and provide a co-ordination framework for the synchronisation of information systems. Since most of the processes reside in different parts of an organisation (or silos), the semantic web model may offer a practical solution to bridging the different areas of basic research, drug development and clinical trials, while keeping the management of knowledge local to each group.

Researchers can use the SW framework at a personal level to annotate findings and record decisions for use by both humans (eg, report generation) and programs (knowledge mining), which consequently enable companies to derive more knowledge from the vast amount of recorded and interpreted analyses. Such structured knowledge would be accessible to inference rules, able to help scientists create new insights.

Listed here are some possible areas within drug R&D that could benefit from SW:

- Chemical information and knowledge
  Inter-connect commercial and internal chemical knowledge.
  Connect all bioassays and HTS across all targets.
  More wide-ranging definition of drugability.
  Improve managing and interpreting of ADME/Tox analyses.
- More usable insight from animal studies
  Facilitate the capture of results to build a knowledge-base of all *in vivo* toxic responses.
  More comprehensive validation of animal (disease) models.
  Find and index predictive signals in animals and how they map to humans.
- Identify, Evaluate and Manage Biomarkers
  Disease characterisation, progression and sub-typing.
  Associate biomarkers to mechanisms and pathways.
  Assemble existing data on the association of markers to clinical outcomes and their performance in intervention trials.
  Identify clinical trials under development in which data gaps and uncertainties could be addressed.
- Drug Safety
  Pharmacovigilance.
  Compilation of toxicological signatures from 'related' studies (eg, hepatoxicity).
  Micro-dosing studies.
- Application to clinical designs
  Use of genotypic information for patient recruitment.
  Bayesian trial designs.
  Post-launch surveillance.

Moreover, the Semantic Web has the potential to strongly support FDA's Critical Path Initiative[1], since this initiative relies on stakeholder organisations to track and utilise an increasing network of

Figure 5
Gene Expression Heatmap Lens in BioDash of RDF-formatted microarray data. All cells, columns and rows are web objects that can connect to any other web objects on the internet (see **Table 1**)

knowledge for better decision-making on issues regarding drug safety and efficacy. The phrase 'from Bench to Bedside' implies the use of a lot of knowledge and policies in some uninterrupted form. Simply placing documents in eRooms or document management systems and bookmarking web pages is insufficient to achieve more effective management of internal and external knowledge, policy compliance (21 CFR Part 11), or study findings in any given area. As a specific case, the semantic web could be used in drug safety to enable the meaningful and relevant comparisons between preclinical and clinical toxic studies. Relations within analysed data sets such as 'corresponds to', 'is correlated with', 'has same response pattern as', and 'is an indicator for' can be directly used to link across study data using RDF-OWL. In addition, data would reside not as a simple table of values, but as typed multi-valued entities (**Figure 5**) linked by multiple semantic relations to probes, conditions, and diseases (eg, using ICD10). This form has the advantage that it can support annotations and alert tags at an individual measured value level, as well as possible correlations between studies and species (see **Table 1**). Not all analyses are simply statistical tests; it is just as important to make association between known drug metabolism pathways and evidence of excreted metabolites.

In addition to linking sets of data in meaningful ways, mechanisms of action can be represented and associated using combinations of standards such as BioPAX for pathways (www.biopax.org[8]) and CDISC for clinical observations. Combinations like these are straightforward using RDF-OWL and would permit scientists to directly associate analytical results of clinical studies with proposed causal models without the need to develop additional standards (**Figure 4**). Results from animal studies involving other molecular analytes (eg, metabolomics profiles) can be semantically aligned with human phase 1 studies, to assess and select predictive-tox biomarkers that work for both species. This would go a long way towards making better use of preclinical and clinical knowledge. As described in the FDA report:

"The concept of model-based drug development, in which pharmaco-statistical models of drug efficacy and safety are developed from preclinical and available clinical data, offers an important approach to improving drug development knowledge management and development decision making".

Considering the heterogeneous kinds of data and different experimental designs, capturing all required information will require a semantics-based approach, and SW ensures that this will work between information systems as well.

Safety Lenses

Using a Semantic Web approach, it is possible to aggregate an enormous amount of knowledge available throughout the web (as well as any intranet accessible database) on a specific subject. This could produce large data structures (graphs) that have one central subject hub, and literally thousands of relations directly tied to it, and many millions more that are indirect. The question then becomes, how does one focus on the relevant set of relations on a subject without becoming inundated with information? One solution being developed to address this are sets of filters known as semantic lenses, which perform a similar function to what style-sheets do to improve the format of regular web-pages, only here they work with relational information and knowledge.

Lenses are associated with specific kinds of information types (see **Table 2**), but different lenses can be constructed for use by different professionals to view the same information selectively

(See **Figures** 4 and 5). For example, consider the aggregate knowledge assembled around a specific inherited disease. A geneticist would most likely want to know the distribution of specific alleles in the population that could contribute to the disease, while a molecular biologist might be more interested in the specific variants of the gene and the mechanism by which they promote a pathology progression. Clinicians of course would like to understand the impact on health, and how to genetically diagnose the chance of the disease ahead of time. Lenses can also be directed to aggregate additional information from specified sources (including scientific papers) and combine this with the original data bundle.

Lenses often do not require extra compiled code to be written, since they are invocations of predefined modules (similar to the stylesheets used by HTML pages). The lenses typically reside in next-generation browsers called semantic browsers, but could be part of future query engines as well. An excellent example of such a semantic browser is the Haystack project at MIT[9], which has been used to create the BioDash semantic drug discovery prototype (http://www.w3.org/2005/04/swls/BioDash/Demo/). Lenses can easily be downloaded, shared and modified over the Internet, allowing others to 'see' and 'handle' the information the same way their colleagues intended them to. Lenses enable multi-team collaborations to work on projects requiring distributed sources of information and knowledge.

Clinical trials require a lot of planning and investment, and involve a series of complex steps and multiple players. Stakeholders such as the clinics, sponsors (typically pharmaceuticals), CROs, government regulators and the reviewers, all need to co-ordinate their activities (from a data perspective), which involves data generation, capture, transfer, analysis, interpretation, preparation (of application), reviewing and responding. The same data needs to be exchanged while preserving its integrity and relations, and the subsequent analyses and interpretations performed on them need to be robustly associated as well. A special kind of lens could be defined specifically to analyse and score preclinical and clinical studies using biomarker measurements, associate additional background information from related cases and include possible mechanisms of toxicity. These 'Safety Lenses' could be constructed for different classes of toxic responses, such as hepatotoxicity, neural toxicity, HERG toxicity, nephrotoxicity and genotoxocity. Each would take advantage of the semantics that need to be considered for each of the areas. In parallel, pharmacogenomic standards and validated biomarker profiles recommended by the FDA could be placed on secure websites to be used automatically within the Safety Lenses by the clinical sponsors. The results of these specialised analyses would also be available to be incorporated into the NDA itself.

Finally, by including narratives and interpretations, lenses can be used to generate reports and help prepare submission dossiers. Regulatory groups could also develop a set of clinical lenses (such as Safety Lenses) to be used by the reviewers, yet these could also be made available to the clinical sponsors to help them understand how reviewers are envisioning their results. At the same time, since the full information set should be available to everyone, any possibility to hide or manufacture information that might mislead a reviewer is strongly prevented. The necessary set of semantic lenses could be defined within a larger healthcare initiative such as HL7, using approaches that are not as protracted as current standards implementation methodologies.

Summary

One would not be amiss to assume that the complexity of clinical data will increase in response to drug safety con7ety1 Tf10 w as the need to innovate. The latest guidelines from the FDA (http://www.fda.gov/cber/gdlns/pharmdtasub.htm) propose the use of diagnostics for pharmacogenomic1 Tf10 w as data from biomarkers and images. Traditional approaches for incorporating new data standards will most likely not be able to keep up with the required changes, since the time from standard specification development to implementation usually is three to five years (see www.omg.org and standards.ieee.org). It is becoming evident that what is required is an adaptive, extensible model that can be used by a community to define functional domain standards that support semantic interoperability and the addition of new forms of information as soon as they are defined and agreed upon. This is where the SW can offer some practical solutions to standards development and adoption.

Finally, the knowledge collected from all completed clinical trials and submitted NDAs would already be in a form (assuming SW format) for use in a comprehensive drug development knowledgebase, the value of which would be highly prized by any drug company. Quantitative information on subject responses for many classes of compounds around specific

References

1 Innovation or Stagnation, Challenge and Opportunity on the Critical Path for New Products. FDA-Report March 2004.

2 Wenger, E (1998). Communities of Practice, Cambridge University Press.

3 Berners-Lee, T, Hendler, J and Lassila, O (2001). The Semantic Web, Scientific American, May (2001).

4 Hendler, J. Science and the Semantic Web. Science Vol 299 24 (2003).

5 Neumann, E. Perspectives: "A Semantic Web for Life Sciences: Are we there yet?" Science-STKE, Issue: 283, p 23, 2005.

6 Pollock, Jeffrey T, Hodgson, Ralph. Adaptive Information: Improving Business Through Semantic Interoperability, Grid Computing, and Enterprise Integration. Wiley Press (2004).

7 Boston Consulting Group. "A Revolution in R&D: How Genomics and Genetics Will Affect Drug Development Costs and Times." PAREXEL's Pharmaceutical R&D Statistical Sourcebook 2002/2003.

8 Luciano, JS .PAX of mind for pathway researchers. Drug Discov Today. 2005 Jul 1;10(13):937-42.

9 Quan, D and Karger, D. How to make a Semantic Browser. WWW 2004, May 17–22, 2004 ACM 1-58113-844-X/04/0005., New York, New York, USA.

10 Kashyap, V. The UMLS Semantic Network and the Semantic Web. AMIA Annu Symp Proc. 2003;:351-5.

11 Goebel, G, Leitner, KL, Pfeiffer, K. Use of semantic web technologies in medicine and health care. Medinfo. 2004;2004(CD):1618.

12 Murray-Rust, P, Rzepa, HS, Tyrrell, SM, Zhang, Y. Representation and use of chemistry in the global electronic age. Org Biomol Chem. 2004 Nov 21;2(22):3192-203. Epub Oct 22 (2004).

Continued on page 33

(and even multiple) diseases applications would be available for use by future discovery projects. This would almost certainly have a major positive impact on innovations and drug development, and would clearly demonstrate return of investment of specific approaches.

As perceived by the FDA[1], "the applied sciences needed for medical product development have not

```
add { <Diabetes-1-dataset-8-2>
          rdf:type          ls:ClinicalTrial ;

          dc:title          "Diabetes Phase I Study" ;
          ls:period "1/2/02-3/4/03" ;
          ls:experimentalist "Thomas Parker" ;
          ls:targetSystem :Human ;
          ls:design :Hu_MetaDiab_3_2_Protocol ;
          ls:litref :MetabolicSyndromeRef ;
          ls:litref :GLUT_1_Expr_Ref ;
          ls:litref :DT2_Genetics_Ref ;
          ls:litref :ster-CoA_desaturase_Ref ;

          ls:rowProperties (ls:observHub ls:probeHub) ;
          ls:rowTypes    ( :LDL :Diastolic_BP :Systolic_BP :HR :potassium :haemoglobin
:creatinine :CRP :Weight :PSA gsk:CaseinK gsk:DVL gsk:Axin gsk:GBP gsk:APC gsk:GSK3beta
gsk:Catenin gsk:BTCP gsk:WNT8b gsk:Friz ) ;
          ls:colProperties (ls:subjectHub ls:sampleHub) ;

ls:indivCell ${ rdf:type ls:ClinObs_Cell;          ls:observHub :LDL              ; ls:subjectHub
          :HS41273      ; ls:mg_ml         "163.353"        } ;
ls:indivCell ${ rdf:type ls:ClinObs_Cell;          ls:observHub :Diastolic_BP     ; ls:subjectHub
          :HS41273      ; ls:mmHg          "137.041"        } ;
ls:indivCell ${ rdf:type ls:ClinObs_Cell;          ls:observHub :Systolic_BP      ; ls:subjectHub
          :HS41273      ; ls:mmHg          "81.319"         } ;
ls:indivCell ${ rdf:type ls:ClinObs_Cell;          ls:observHub :HR               ; ls:subjectHub
          :HS41273      ; ls:bps           "67.281"         } ;
ls:indivCell ${ rdf:type ls:ClinObs_Cell;          ls:observHub :haemoglobin      ; ls:subjectHub
          :HS41273      ; ls:mg_ml         "2.681"          } ;
ls:indivCell ${ rdf:type ls:ClinObs_Cell;          ls:observHub :creatinine       ; ls:subjectHub
          :HS41273      ; ls:mg_ml         "1.826"          } ;
ls:indivCell ${ rdf:type ls:ClinObs_Cell;          ls:observHub :CRP              ; ls:subjectHub
          :HS41273      ; ls:ug_ml         "0.041"          } ;
ls:indivCell ${ rdf:type ls:ClinObs_Cell;          ls:observHub :Weight           ; ls:subjectHub
          :HS41273      ; ls:kg            "62.157"         } ;
ls:indivCell ${ rdf:type ls:ClinObs_Cell;          ls:observHub :PSA              ; ls:subjectHub
          :HS41273      ; ls:ng_ml         "28.277"         } ;
ls:indivCell ${ rdf:type ls:GE_Cell;               ls:probeHub gsk:CaseinK        ; ls:sampleHub
          :HS41273_L1   ; ls:GE_Expected_Ratio    "0.857"   } ;
ls:indivCell ${ rdf:type ls:GE_Cell;               ls:probeHub gsk:DVL            ; ls:sampleHub
          :HS41273_L1   ; ls:GE_Expected_Ratio    "1.084"   } ;
ls:indivCell ${ rdf:type ls:GE_Cell;               ls:probeHub gsk:Axin           ; ls:sampleHub
          :HS41273_L1   ; ls:GE_Expected_Ratio    "0.785"   } ;
ls:indivCell ${ rdf:type ls:GE_Cell;               ls:probeHub gsk:APC            ; ls:sampleHub
          :HS41273_L1   ; ls:GE_Expected_Ratio    "1.135"   } ;
ls:indivCell ${ rdf:type ls:GE_Cell;               ls:probeHub gsk:GSK3beta       ; ls:sampleHub
          :HS41273_L1   ; ls:GE_Expected_Ratio    "1.118"   } ;
ls:indivCell ${ rdf:type ls:GE_Cell;               ls:probeHub gsk:Catenin        ; ls:sampleHub
          :HS41273_L1   ; ls:GE_Expected_Ratio    "1.094"   } ;
ls:indivCell ${ rdf:type ls:GE_Cell;               ls:probeHub gsk:BTCP           ; ls:sampleHub
          :HS41273_L1   ; ls:GE_Expected_Ratio    "1.025"   } ;
ls:indivCell ${ rdf:type ls:GE_Cell;               ls:probeHub gsk:WNT8b          ; ls:sampleHub
          :HS41273_L1   ; ls:GE_Expected_Ratio    "1.487"   } ;
ls:indivCell ${ rdf:type ls:GE_Cell;               ls:probeHub gsk:Friz           ; ls:sampleHub
          :HS41273_L1   ; ls:GE_Expected_Ratio    "1.079"   } ;
```

Table 1: Clinical data with serological and microarray data combined, containing observation relations (orange), probe relations (green), subject relations (blue), and sample relations (red)

Table 2: Example how specific lenses are associated with ToxicoGenomic data using RDF. This association can be made separate from the data definitions and at any time

```
<fda:ToxGenomicData>          <hs:hasLens>     <fda:GenericToxGxLens>
          <hs:hasLens>          <fda:HepatoToxLens>
          <hs:hasLens>          <fda:CNSToxLens>  .
```

kept pace with the tremendous advances in the basic sciences". Much of this hinges on using information more consistently and effectively in development, so that we are learning from the failures as well as the successes. A small 10% enhancement in predicting failures before trials can save close to $100 million in development costs for each drug. Managing data as well as the insights gained will require using something better than a spreadsheet table.

The Semantic Web is more than just about data management or the use of ontologies. It is being developed to support the definition and application of a broad range of policies, rules, domain-specific features and workspaces for all types of end users. Much still needs to be developed in order for the full vision to be realised, but it is critical for domain specific users to participate and manage their own issues accordingly. Semantics approaches are being explored in several areas already: medical language systems[10], healthcare management[11], chemistry[12], cancer research[13] and clinical trial management[14].

To facilitate the co-ordination of activities in life science research, the Healthcare and Life Sciences Interest Group (HCLSIG) is being formally initiated at W3C (http://www.w3.org/2005/05/swlsig-charter) to bring the requirements of the scientists into close proximity with the semantic technologies community. In October 2004, a workshop was held at MIT to assess some of the critical needs (http://www.w3.org/2004/10/swls-workshop-report.html). A follow-up meeting is planned for later this year. We invite members of the life science and healthcare communities to participate together and begin getting the most out of the Semantic Web. DDW

Dr Eric K. Neumann *is founder of the Clinical Semantics Consulting Group and is co-chair of W3C's Healthcare and Life Science Interest Group (HCLSIG), focusing on domain applications of Semantic Web Technologies. Recently, he led the development of BioDASH, a prototype drug discovery dashboard based on Semantic Web technologies. He previously was Global Head of Knowledge Management for Scientific and Medical Affairs within Sanofi-Aventis, covering all of its R&D needs. Dr Neumann is an expert in knowledge-based methods of working in the pharmaceutical industry, which has been his interest for the past 14 years. Prior to Aventis, Dr Neumann was at Beyond Genomics, a biopharmaceutical company based in Waltham, Massachusetts, which was founded to discover and develop new drugs by exploiting unique technologies and the knowledge generated from the '-Omics' revolution. He is also the co-founder of Genstruct, a Cambridge-based company that applies Knowledge Assembly and Molecular Epistemics to disease elucidation. Dr Neumann has also served as VP of Life Science and Informatics at 3rd Millennium, Director of Research at NetGenics (now LION), a company that built integrated informatics solutions, and a Senior Scientist at Bolt, Beranek & Newman an R&D technologies company. Dr Neumann has a bachelor's degree from the Massachusetts Institute of Technology, in Cambridge, Massachusetts, and a PhD in neurobiology and developmental genetics from Case Western Reserve University.*

Continued from page 32

[13] De Coronado, S, Haber, MW, Sioutos, N, Tuttle, MS, Wright, LW. NCI Thesaurus: Using Science-based Terminology to Integrate Cancer Research Results. Medinfo ; 2004:33-7 (2004).

[14] Kamel Boulos, MN, Roudsari, AV, Carson, ER. A dynamic problem to knowledge linking Semantic Web service based on clinical codes. Med Inform Internet Med. Sep;27(3):127-37 (2002).