

ONTOLOGIES

semantic networks of pharmaceutical knowledge

In the current climate making information more transparent to internal R&D staff and to external audiences such as regulators and consumers is crucial to business effectiveness. A new generation of semantic technologies capable of generating a consistent view of knowledge derived from multiple diverse data sources is maturing rapidly. These networks of knowledge, called ontologies, can act as a reusable substrate for a number of real business applications.

By Dr Stephen P. Gardner

The pharmaceutical industry has massive issues with information transparency throughout its entire range of operations. Internally, getting access to all available relevant information is obviously crucial to support R&D innovation, but many other core functions have issues as well. Product litigation groups have to understand and correlate the full range of knowledge in the internal and public domains to build defensive strategies in response to personal injury suits and external market events such as the withdrawal of a competitor's compound in the same class as one of their drugs. Other functions such as biomarker identification, evaluation of in-licensing opportunities, and generation of novel mechanism of action hypotheses all require the correlation of complex multi-disciplinary information. Externally, demonstrating the same information transparency is crucial to communications with regulators, analysts and patients in an increasingly risk averse and litigious market. Regulatory authorities are continually demanding more detailed evidence for the efficacy and safety of compounds. Analysts, shareholders and insurers are demanding to have a much clearer view of the

full range of risks that a compound may incur before committing to investment and protection decisions. Patient groups are demanding more stringent testing and safety criteria and a much fuller disclosure of information at all stages of approval and marketing.

The risks are real, immediate and potentially huge. At the end of September 2004 Merck voluntarily withdrew its rheumatoid arthritis drug Vioxx from the market when one of its long-term studies showed that normal doses could increase the risk of heart attack and stroke after 18 months of treatment. Even though Merck withdrew Vioxx voluntarily and the FDA voted narrowly to restore the product with a black box warning, the potential cost to Merck could run into tens of billions of dollars. Merck's market capitalisation dropped more than \$30 billion, it stands to lose a significant portion of the expected revenue stream of more than \$2.5 billion per year from its portfolio, and estimates of the cost of settling potential litigation have exceeded \$20 billion. Merck is not alone in having high profile and high cost failures; other manufacturers' compounds such as Bextra, Iressa and Tysabri have

also been withdrawn recently. In this climate, the reputation of pharma has been called into question publicly by advocacy groups, analysts and even an Attorney General and, partly as a consequence of this, the budgets set aside by many pharma companies just to cover their legal fees associated with product litigation have risen to several hundreds of millions of dollars per year. The implicit foundation of trust that has characterised clinical practice has been seriously eroded in the public mind, and this is already creating additional hurdles for the future approval of compounds, and their subsequent sale and prescription, especially via large, litigation sensitive healthcare insurers such as HMOs.

The suspicion fuelling much of this mistrust is that pharmaceutical companies know much more than they tell the regulators, consumers, clinicians or investors. This has been compounded in some cases by the subsequent discovery of data that, had they been known, could have contributed to altering an approval or labelling decision. While this is very unfortunate, it is hardly surprising, and with fuller public disclosures of clinical trials data in the future it is likely to happen more frequently. The scale of a large pharma company is a difficult thing to leverage when it involves knowing everything that the company could reasonably be expected to know given all of the information it generates. Pharmaceutical companies have tens or even hundreds of thousands of employees, often with R&D budgets running into multiple billions of dollars, and generating tens of terabytes of data per year. Their compounds are given to millions of people each with a complex genetic make-up and a vast array of competing and conflicting environmental factors that guarantee some low-frequency adverse effects even with the safest of compounds. Making sure that everyone who needs to know has full access to all of the relevant information and can see the proper signals without them being swamped by the huge sea of irrelevant data is a very serious informatics challenge.

How much should I know?

The traditional difficulty facing groups trying to make information more accessible across the business is that the information is fragmented in dispersed, disparate silos throughout the company. Add to this the huge volumes of information generated by an increasingly automated R&D process, the traditional information-protective, expert-led culture of pharmaceutical R&D, and the unpredictability of biological systems and disease processes and the pharmaceutical enterprise infor-

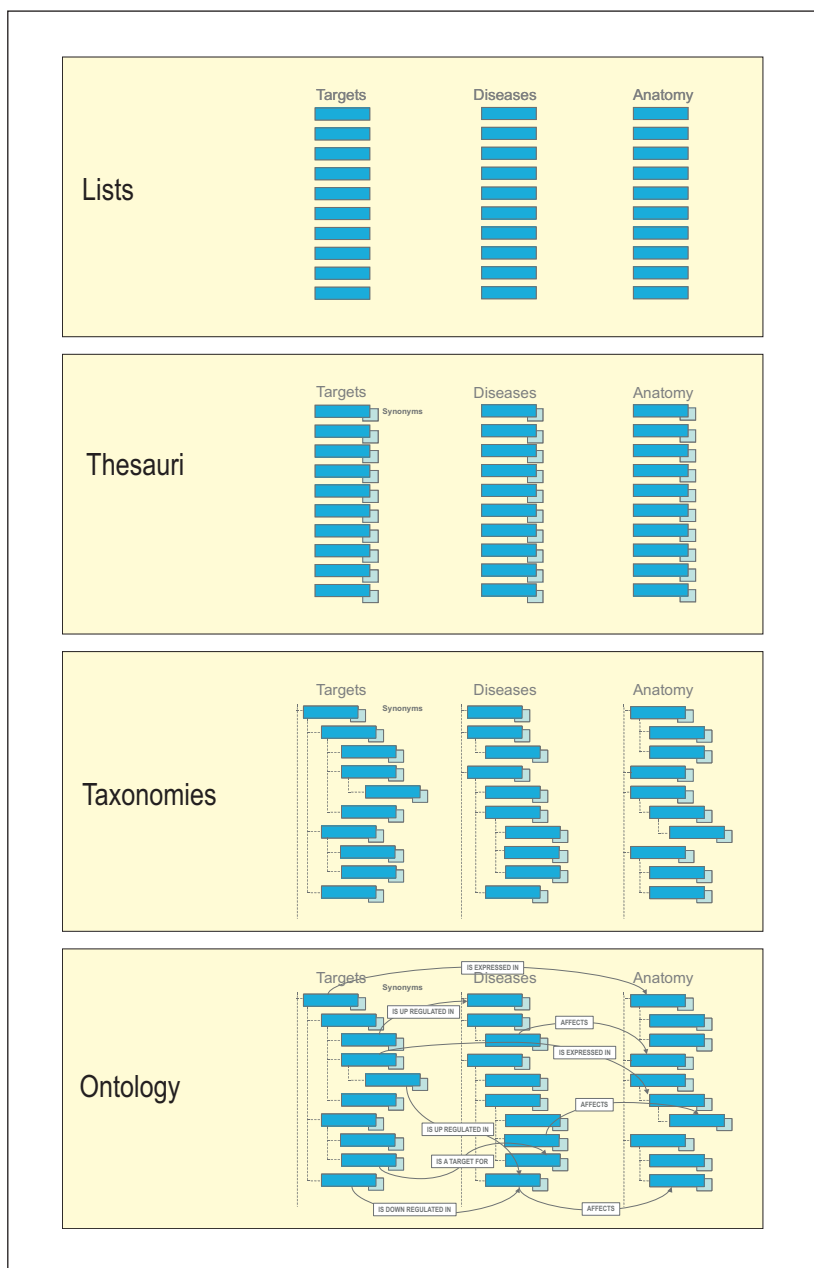


Figure 1
The progression of data representation from simple lists, to thesauri, to taxonomies and then multi-relational ontologies

matics problem becomes complex as well as merely complicated.

The state-of-the-art is, however, changing with the introduction of a new range of semantic technologies. What it should be possible to infer or deduce from a given set of source information is changing as semantic systems that not only collect, but also ‘understand’, the information begin to be deployed. This has profound implications across the pharmaceutical industry and beyond. All pharmaceutical companies have access to the same public domain data sources and a similar set of subscription sources with which they augment their

internal R&D and post-market surveillance data. The regulatory authorities not only have the same basic public data but also the various information for multiple compounds in a given class from a variety of different companies. Any one of these players who can make new connections between pieces of knowledge and use those to spot new trends, generate new hypotheses and draw new conclusions has potentially a very significant advantage over its competitors.

Why semantics help understand what you know

Semantic technologies draw their power from the flexibility with which they can represent any information in a semantically consistent framework so that information from many diverse sources can be compared, connected and viewed from any perspective. Traditional informatics systems in contrast are constrained by the syntax (structure) of data, and only represent information in the forms that are explicitly specified by the designers of the database schemas. This means that the systems contain fixed and limited semantics (meanings) that describe only a very narrow perspective of what a concept (eg a protein or drug) is and what it is doing in various circumstances in the body in response to varying events. These traditional database systems tend to focus on the concepts (the genes, proteins, compounds etc) and the properties that they have in a given circumstance (eg location, molecular weight, therapeutic class etc). In these traditional systems, the underlying meaning of a given concept label is uncertain, so that the same concept called by two different names in two different sources (eg myocardial infarction and heart attack) cannot readily be connected, but two concepts that share the same name, but which are fundamentally different, will be considered to be identical (eg COLD could be Chronic Obstructive Lung Disorder, environmental hypothermia, or a rhinoviral disease).

The new semantic technologies tend to focus on the connections between concepts, so that the way that they interact and the role that they play can be represented in some detail. When a new concept is added to a semantic network, it is defined not only by the properties and other information that was entered alongside its name, but also by the connections that it has to other concepts in other contexts. This has the added advantage that as time goes on, old information remains visible and new information becomes associated with a concept. For example, during its development phase the compound UK92480 underwent a number of

internal efficacy and safety pre-clinical studies for which there are a variety of reports, databases and spreadsheets. As it progressed to IND, sildenafil citrate underwent a number of carefully orchestrated and documented clinical trials, generating reams of new clinical information. Once approved, Viagra has generated even more sets of marketing, legal, post-market surveillance and sales information that also exists in multiple databases inside Pfizer, the regulatory authorities, HMOs, health insurers and other bodies.

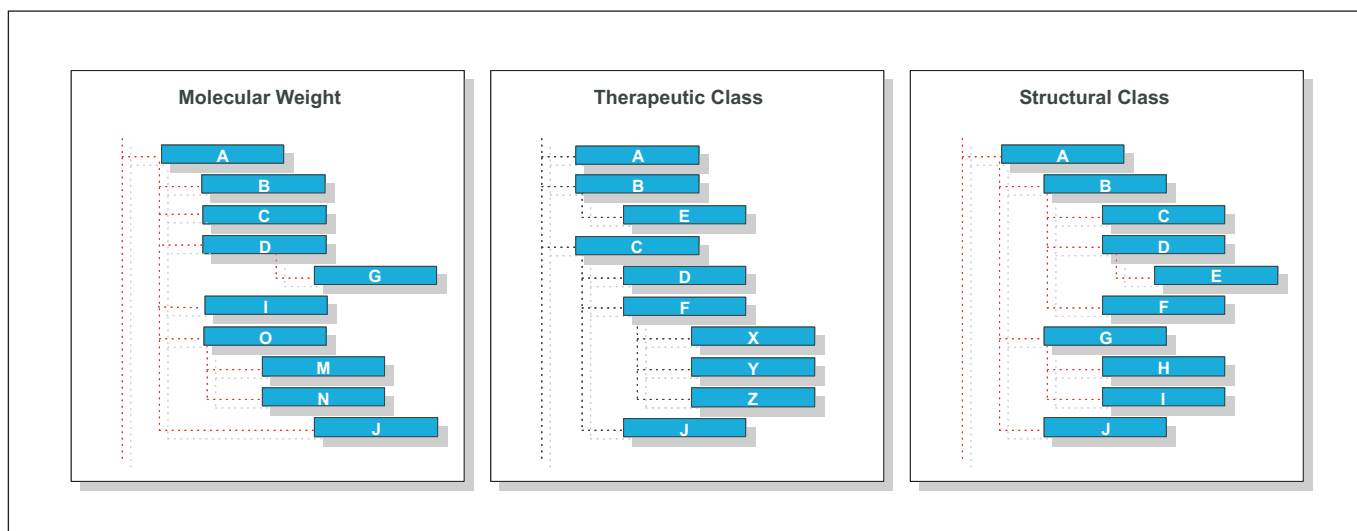
In total there may be hundreds of different data sources containing different types of information generated over the course of a drug's life cycle. Some of these will be structured sources such as relational databases, warehouses, and existing taxonomies/thesauri. Others will be unstructured, such as pre-clinical reports, scientific literature, patents, and regulatory filings. In a semantic network some or all of this information, depending on the application, might be associated with the main concept (for example the chemical structure of sildenafil).

This is not to suggest that UK92480, sildenafil citrate and Viagra are considered to be the same thing in a semantic network. While a discovery project manager might not care about the differences between these concepts, and may simply wish to view any information relating to potential toxicities collected at any stage of development or marketing, a product litigation specialist or chemist would see them as fundamentally different concepts. The semantic network provides the flexibility to represent them as distinct entities, but be able to view the context around the whole set of information.

Semantic technologies: going beyond the taxonomy

While it may appear that semantic networks (often called ontologies) are radically different from traditional systems, they do share a common heritage. As the name implies, a semantic network represents information in a more complete and interconnected fashion than a traditional relational database, data warehouse or the hierarchical taxonomies used in most file systems. The differences between the various types of data representation are shown in **Figure 1**.

The simplest form of knowledge representation is a list of all of the available values for a given subject (eg the names of all of the genes in the human genome or all of the tissues in a body). A list is usually an alphabetically sorted catalogue of the names of all known concepts in a given field with



no implicit or explicit relationships between them other than that they are all of the same type. Lists are very useful for simple applications such as indexed keyword searching, and controlled vocabularies, where they are used to constrain and accelerate the entry of accurate information into a system, avoiding the potential for mis-spellings and typographical errors. Many relational database schemas are simply lists of concepts with some associated properties for each of the concepts.

Thesauri are derived from lists but have one important additional component. Thesauri store synonyms for the entries in the list alongside the list entry. A synonym is a concept that is identical or very similar to the entry in the list, but which has a different name. Synonyms are very useful in improving the accuracy and completeness of keyword searching. By expanding a search to include all the synonyms for a given concept, results for all of its related synonymous terms can also be returned. This means that a single search for 'myocardial infarction' would not miss articles simply because they used the alternate term 'heart attack'.

Taxonomies build on thesauri by adding relationships between the concepts to provide a parent-child type organisation to the lists. The relationships are of the form IS-A, eg 5HT1A IS-A GPCR or Anorexia IS-A Eating Disorder. Other similar relationships, IS-A-PART-OF, CONTAINS, or HAS-A may also be used in taxonomies. When taken together these relationships allow the creation of a taxonomic hierarchy of all the concepts for a given subject. This provides a familiar tree-like structure that organises and classifies information and gives an immediate visual clue as to the

likely function or nature of a given concept through its position in the hierarchy.

Taxonomies however are largely static, and represent information from one perspective only. For example it is possible to imagine several different ways of organising a taxonomy of compounds, for example by therapeutic class, structural class, or molecular weight as shown in Figure 2. Building and maintaining each of those alternate taxonomies is time-consuming and expensive, becoming ever more so as the taxonomies grow. Worse still, as anyone who tried to find specific information in a large set of e-mail documents, photos or PowerPoint slides organised on their hard disk knows, taxonomies also tend to become harder to maintain and use as they get bigger. It gets harder to find information if you can't remember exactly which category it was put in, and the more categories there are, the harder it becomes to decide which one is right for any new piece of information.

Ontologies are in many practical ways the exact opposite of taxonomies. Rather than struggling to remember whether a photo was filed under 'Holidays', 'Caribbean', 'Sunset' or 'Family', any one of those pieces of information can be used to retrieve the photo as it is connected to all of those categories at the same time. Ontologies provide a true semantic network, where having more information in the network means more connections and therefore more directions and perspectives by which relevant information can be found.

Ontologies build further on the taxonomies to include much richer and more descriptive relationships between concepts, eg p53

Figure 2
Alternate taxonomic organisations of the same concepts according to various criteria

IS-UPREGULATED-IN Breast Cancer, which connects a concept from the Targets list and one from the Diseases list using a specific relationship. By connecting all of the knowledge derived from variety of primary sources in this detailed fashion, ontologies can synthesise a very rich and powerful map of all of the information known about a whole domain. This is akin to marrying a street map of a town with a bus timetable, a website of pubs, a cinema listing and a *Yellow Pages* directory listing of restaurants when planning a night out on the town. This synthesis of data from individual silos is something we usually do in our heads, reconciling slight differences in nomenclature and data representation to form a coherent plan of action. When the more complex data available to pharmaceutical companies are represented in semantic networks, they provide exactly the same type of knowledge substrate that can be 'understood' by computer systems. These in turn allow for a new range of much more powerful business applications that allow joined up thinking about real pharma business problems.

Ontology-enabled technologies and business applications

A wide variety of existing applications and technologies that are already deployed inside pharmaceutical companies can benefit from accessing semantic networks and ontologies. Some of these technologies and applications are listed below:

Market differentiation of products

Technologies such as data mining (eg Spotfire, SAS or other statistical analysis application) benefit from having all of the relevant data available for analysis in a consistent form. Rather than spending weeks to get the data into shape to answer a specific question, ontologies can provide the substrate for systematic analysis and correlation of all of the available data. Any of the concepts, properties or relationships from any of the data sources mined when compiling the ontology are automatically available for use in analyses. In turn this allows rapid iterations to ask and answer further questions as the results of one analysis leads to new questions. This can be used for example to differentiate compounds within a therapeutic class on the basis of profiles of receptor binding or the invocation or inhibition of specific pathways, and to generate new hypotheses for the observed variations in reported rates of side-effects between members of that class.

Identifying and evaluating in-licensing opportunities

In-licensing relies on being aware of reports of new compounds or targets that match a specific profile and become available for licensing. It then involves being able to evaluate the structures in the light of known biological, safety, adverse event, licensing, competitive and patent data. This data may be spread across many very diverse sources. Ontologies can correlate all of this contextual information and even connect it to chemical structure data. This facilitates a very broad range of analyses of the roles that particular structural components play in various scenarios and allows subsequent evidence-based licensing decisions to be made.

Biomarker discovery

A hypothesis of what an ideal biomarker for a given condition might look like can be formulated, for example proteins expressed only in a given tissue, released into a bodily fluid, involved in the function of the tissue, and in the early-stage processes of the effect under study, and with a good signal:noise ratio. Evidence for those biomarkers that fulfil all of these criteria can then be examined using the ontology, which in turn is representing information from and searching across tens or hundreds of primary sources. This can lead quickly to a very credible set of evidence-based candidates for further experimental validation. If one stage of the definition of the ideal biomarker is found to be too specific or too loose this can be amended and the process re-run in minutes.

Obstacles to be overcome

Obviously since we are not yet all living in a semantically driven world, building ontologies and semantic networks is not a trivial exercise. Getting access to all of the information sources, reconciling the systems, formats, and syntax of the information, mining information out of structured and free-text sources, and normalising the semantic content of the resulting information are all complex and challenging problems. To perform this accurately, reproducibly and at the enterprise scale needed requires the combination of a number of state-of-the-art technologies and knowledge representation standards that are themselves evolving in capability. Bringing together knowledge in a way that is useful for multiple applications, that ask different questions from diverse perspectives, places a very high burden on the up-front design of the knowledge representation. This requires not only a

solid appreciation of semantic technologies, but also of the business problems to be solved.

At the same time, a technology that can present a comprehensive and unbiased representation of the current state of knowledge from hundreds of different sources is culturally challenging. Researchers are used to being experts in their field, whose opinion is unchallenged and does not need to be supported by evidence. This expertise can over time become bias and even dogma, which does not always reflect the changing nature of knowledge as new information becomes available.

Even the way in which information is used will be profoundly changed. We have become accustomed to being data rich and information poor. We can generate huge volumes of data about very specific topics, coming from microarray or other automated technologies. We also have instant access to the full range of scientific literature. It has become easy to generate or retrieve a set of data and allow the interesting patterns within it to guide our discovery strategy. The current keyword interfaces such as PubMed or Google for example generate tens of thousands of hits, most of which are only peripherally related to the topic under study. We often simply browse as many as we can conveniently, read the abstracts of a few dozen and then read four or five papers, convincing ourselves that this is systematic analysis. Again semantic technologies are very different. They understand the context in which a particular word should occur, the concepts that it should be linked to and the properties that it should have. This requires much more thought about the construction of the questions, which go back to being hypothesis driven, formulating an idea and testing it by looking at the very specific and relevant results that are returned.

Conclusions

Notwithstanding these obstacles, semantic technologies are maturing quickly, and a number of standards setting bodies such as W3C are formalising the representational standards in initiatives such as the Semantic Web to allow for interchange between ontologies and other semantic resources. There are also a growing number of semantic applications that can take advantage of information expressed in these standards. At the same time initiatives such as the various GRIDS projects, and a number of vendors are beginning to apply semantic networks to a wide range of pharmaceutical business problems such as the ones described above with some notable successes. These applications are much more real and pressing than the simple collection of knowledge for its own sake,

and are often well beyond the scope of existing bioinformatics, data integration or knowledge management technologies.

The state-of-the-art in knowledge management is undergoing a period of rapid change as the influence and reach of semantic technologies develops. Applications of these technologies are already becoming available and their impact will be felt much more quickly than many observers might expect. Not least they will in the short-medium term help to redefine what is considered to be best practise for making information available internally for decision-making, and for transparent communications to external audiences such as regulators, analysts and ultimately consumers. **DDW**

Dr Steve Gardner has 18 years' experience in building innovative informatics technologies, products and companies in the life science domain. Steve was the founder and CEO of Synomics Ltd, CTO of Viaken Systems, Senior Product Manager at Oxford Molecular and Worldwide Director of Research Informatics for Astra AB. He is currently CTO for BioWisdom where he is responsible for the development of leading ontology curation, management and delivery technologies.