



DRUG DISCOVERY
TODAY
TECHNOLOGIES

Editors-in-Chief

Kelvin Lam – Pfizer, Inc., USA

Henk Timmerman – Vrije Universiteit, The Netherlands

Knowledge management

Ontologies in drug discovery

Stephen P. Gardner

CTO BioWisdom Ltd, Harston Mill, Harston, Cambridge, UK CB2 5GG

Knowledge is a fundamental driver of innovation in drug discovery. Although investment in the automated generation of data has led to swelling databases, these have not been harnessed effectively to meet the challenges of real-world business. It has been too difficult to make the necessary connections between pieces of knowledge in a range of systems towards the bigger picture and to formulate and execute a response to this. The new generation of ontology-based semantic technologies is changing the way that information is represented and used, building a new power system for R&D, based on the generation, distribution and application of knowledge across the organisation.

Introduction: knowledge-led drug discovery: vision or reality?

To paraphrase Sir Francis Bacon [1], in the context of drug discovery, knowledge is power. It is a vital service that should be flowing out of the walls of pharmaceutical companies, as fundamental to powering innovation as electricity is for powering the instruments and lab equipment. However, just like electricity, knowledge can be unproductive and even dangerous if mishandled. Unfocused, raw energy is expensive and destructive – 60 years ago, the Hiroshima bomb delivered an explosive yield of approximately 15 kt in a blast that devastated an area of 5 square miles. This same energy harnessed properly could power the whole of New York City for about 3 h (18GW-hours) [2], long enough for the city to generate US\$ 200 million of gross metropolitan product (http://www.usmayors.org/metroeconomies/1004/metroeconomies_1004.pdf).

E-mail address: S.P. Gardner (steve.gardner@biowisdom.com)

Section Editor:

Manuel Peitsch – Novartis, Basel, Switzerland

In the past decade, the pharma industry had its own explosion of data, investing billions of dollars in new automated technologies. Whereas this has generated a smokescreen of rapid action and technological development, R&D productivity has continued to decline as the investments in systems to harness and focus that data on real business problems have lagged way behind. Far from being knowledge-led, drug discovery scientists often feel powerless in the face of large sets of increasingly complex and disconnected data, with no means of pulling together the big picture that would generate the insight to answer their questions. Backed into this corner, discovery science risks becoming something of a data-driven crystal ball, generating large sets of data and looking within them for enlightenment in the swirling latent patterns that might be revealed.

Without transparent, unfettered access to diverse knowledge from inside and outside the organisation and the tools to pull it all together, complex decisions simply cannot be made in a fully informed fashion. Yet the pace of business insists that crucial investment decisions are made rapidly and continuously. Harnessed properly, the knowledge of a company's assets could help pump its pipeline full of profitable new compounds but if managed badly, they could lead to the wrong decisions being taken, which could cost billions if those products fail in trials or, worse still, after they reach the market.

Gathering and applying knowledge in drug discovery

There are several elements required to be able to exploit information effectively inside a large corporation, especially one dealing with such diverse and complex information as a pharmaceutical company. First, information has to be gen-

erated, by transforming it from its diverse primary sources into a medium that can be harnessed. Second, the infrastructure has to be in place to distribute the knowledge on demand to scientific and business users throughout the company (and possibly its collaborators). Third, the applications of the user have to be capable of taking advantage of the diverse set of information, so that it can be put to work for real business benefit.

Different informatics technologies have addressed these different phases but few bridge all three. One of the emerging technologies that have the potential to provide a coordinated solution across all three is the new breed of semantic technologies driven not by the structure or format of data but by the meaning (semantics) of those data. One of the primary development drivers of standards for these technologies is the Semantic Web [3], a project to help develop the next generation of Worldwide Web technologies, managed by W3C (<http://www.w3.org/Consortium>).

The semantics are embodied in descriptions of the concepts in the application domain, the relationships between them and their properties. This description of the domain is called an ontology. In truth, the definition of an ontology can be hard to pin down as the term is often loosely applied to dictionaries, thesauri and taxonomies as well as to much more complex networks of related information [4].

Ontologies were originally used by Greek philosophers such as Aristotle [5] to name the things they saw in the universe and the relationships between them, so that they could reason about the nature of the universe. Computer scientists have extended the principle and attempted to use ontology-based models to reason computationally about the domain, for example, to predict the optimal strategy for deploying air and ground resources in airline traffic management or space planning [6]. Biologists have likewise created nomenclatures and vocabularies based on ontologies to manage genomic and pathway integration [7–9].

Ontologies can, however, be much more complex than simple hierarchies used to taxonomically organise a domain or provide a dictionary of certain types of concepts. They can significantly enhance all three aspects of effective information handling: generation, distribution and application.

Generating knowledge

One of the major problems in creating dictionaries, taxonomies or ontologies is the scope of the domain, the rate of knowledge acquisition and the limited time and resources available. It is simply very difficult in a world where knowledge is changing rapidly to rely on human editors to maintain an up-to-date and accurate reflection of the state of the domain [10]. As far back as 1694, Leibniz was building calculating machines to automate the task of managing categories of concepts describing an ever more complex world

(<http://www.scienceandsociety.co.uk/results.asp?image=10211220&wwwflag=2&imagepos=7>) [11]. This is not so much of a problem with structured information sources, which can be parsed relatively easily in most cases but is a serious problem with information from free textual sources.

Mining information from free text sources such as Medline, patent documents, regulatory submissions and preclinical reports is crucial to the population of any comprehensive R&D knowledge resource. These resources contain not just simple data but also crucial interpretations, inferences and opinions derived from experimentation. The difficulty comes in extracting the useful information from the text and transforming it into a semantically consistent form. Keyword searches are effectively useless at this and tools such as Natural Language Processing (NLP) systems, although powerful, struggle with accurately identifying the complex nomenclature in the life sciences domain [12–14]. Solutions involving the use of an ontology as a domain guide can provide an effective way of optimising the population of specific relations. For example, if a corpus was to be searched for references to a pattern such as '*neuroleptic compounds interacting with GPCRs*,' it would obviously be advantageous to have not only all the names of all the neuroleptics and GPCRs and all of their synonyms *a priori* but also all of the forms that the interaction between compounds and proteins might take, for example, *binds*, *blocks*, *inhibits*, among others. A comprehensive ontology will provide this level of domain knowledge and can significantly improve the accuracy and productivity of NLP-based text mining in biomedical literature [15].

Over and above their contribution as an input to optimise text mining, ontologies also provide a repository for new information generated during structured data parsing or text mining. By bringing together information from a wide variety of resources into one ontology, they can be connected, managed and distributed as a coherent dataset. This ensures that information that would traditionally get lost or otherwise become invisible over time (for example, an emailed copy of a PowerPoint presentation or Excel spreadsheet) will instead remain visible. It also allows knowledge from multiple sources inside and outside an organisation to be semantically normalised so that it can be compared and distributed as a single consistent set of information [16].

Distributing knowledge

Several standards have been developed for the delivery of diverse information in ontologies to different applications. The predominant standards in ontology development are those defined by W3C as part of their Semantic Web efforts. In particular, the Resource Data Framework (RDF) (<http://www.w3.org/RDF/>), RDF Schema (<http://www.w3.org/TR/rdf-schema/>) and Web Ontology Language (OWL) (<http://www.w3.org/2004/OWL/>) formats are important. These standards provide a framework in which information from several

disparate sources can be abstracted, represented and distributed to applications. The motivation of the standards is, however, to enable better machine understanding of web-based information resources, not necessarily to facilitate scientific discovery. This point causes some tensions as will be discussed later.

The fundamental knowledge representation framework is provided by RDF, a language (based on XML) that uses an abstract graph-based model to represent information in triplets of the form <subject> <predicate> <object>. RDF allows the expression of simple statements (sometimes called assertions) about concepts using named properties and values, for example, a triplet might be <'Ontologies in Drug Discovery'><author><'Steve Gardner'>. Graphs are made up of a collection of triplets, which might overlap, so that all papers with Steve Gardner as author could easily be extracted. The concepts in RDF (such as the paper and author names) are identified uniquely using Uniform Resource Identifier references (URIs) (<http://www.isi.edu/in-notes/rfc2396.txt>), so that there is no ambiguity if different authors, for example, share the same name.

RDF Schema extends the capabilities provided by RDF and allows an ontology creator to define a specific vocabulary for their domain ontology. This allows them to describe specific types of resources (such as all proteins, GPCRs or authors) and their properties and use them as types, classes and sets when interacting with the ontology. Whereas an RDF representation is very abstract, the definitions of vocabularies allowed by RDF Schema provide much of the domain-specific semantics that will go into the ontology and make the ontology much more useful to domain specialists.

OWL is designed to provide computational reasoning (also called inferencing) capabilities across ontologies. These could in theory be used to infer, for example, that if a marketed compound had a specific side effect associated with a particular receptor and that receptor was targeted by a therapy for a different disease, that the original marketed compound might potentially have an alternate use for the second disease. There are different levels of the OWL language, termed Full, DL and Lite, which have different capabilities and different levels of support for computational reasoning. All OWL dialects are built as a vocabulary extension of RDF, in such a way that any RDF graph will form an OWL Full ontology. OWL Full has minimal support for reasoning but maximum support for fully expressing any knowledge in any form. Other more restrictive OWL sublanguages include OWL DL (Description Logic), which provides a language subset that supports Description Logic-based computational reasoning, and OWL Lite, which is a further restricted subset of OWL DL. The discriminating feature of OWL DL and OWL Lite is that the relationships particularly should be constrained to a logically discriminable subset that can be guaranteed to be computationally tractable (decidable) using a

reasoning engine. This effectively means reducing the amount of information contained in the ontology to those pieces that fit the limitations of the expressiveness of the language. These reductions will be governed largely by the needs of the reasoning engine and not by the end-user.

It is, therefore, important to consider which standards are the most appropriate to use. RDF, RDF Schema and OWL Full are obvious choices but the reasoning-based standards have limitations based on the applications that they are designed to support. Whereas computational reasoning based on ontologies is an interesting computer science problem and has been successful in several problem areas, for example [17,18], it is not as immediately applicable to the wider life science domain. The ideal characteristics of a system to which computational reasoning can successfully be applied are that it should be comprehensively described, well bounded, relatively small (<100 K concepts) and above all deterministic in nature (that is actions have defined and reproducible consequences) [19]. Life science is obviously almost the exact opposite; our scientific understanding of metabolism, disease, pharmacology and the network of mutually compensating effects that any one stimulus might have in the body is far from complete, our nomenclatures overlap and intersect awkwardly, biology is in few cases truly deterministic and we are continually observing and describing new concepts and relationships.

At the same time, there are hundreds of real-world business applications inside pharma that are desperate for the delivery of integrated sets of all available knowledge so that human reasoning (i.e. expert scientists and business people) can make decisions in a more informed manner. When developing ontologies for applications in life sciences, it is therefore important to choose the level at which the standards represent sufficient information to power the real-world applications the knowledge will be used to support, rather than necessarily following a series of standards developed for computer science applications that are better suited to other domains.

Applying knowledge

Ontologies function as an atlas, describing the information content of the primary data sources that underpin them. They provide a convenient schema to organise and deliver information from many disparate sources to a range of different business applications, in the same way that the map in a GPS system might be used for finding a route between cities, locating a nearby restaurant or avoiding speed cameras on the roads. The single consistent representation of knowledge can be delivered to several different applications because of the abstract way that knowledge is held in the ontology.

There are several drug discovery applications for which ontologies are providing a knowledge substrate. This is timely, as the job of drug discovery has become more chal-

lenging recently owing to the increasing risk aversion of the regulatory authorities, prescribing bodies, insurers and patients. It has become crucial to establish the comparative safety profile of a new compound relative to its class competitors much earlier in the development process, and the discovery of early-stage biomarkers for adverse events is much more important than it was even 2 years ago. This sensitivity is particularly prevalent in first-in-class medicines, where an acceptable baseline of adverse events has not yet been established, and the regulatory bodies have begun to demand full mechanistic explanations of any unusual events seen during trials. This might, in turn, bring into play peripheral factors such as accepted assays that do not actually directly measure the level of a biomarker, or whose relationship to a toxicity is at best poorly understood (e.g. phototoxicity or liver injury). In this environment, it is crucial to be armed with all of the available information and tools to help find the relevant connections between nuggets of knowledge.

Many of the tools that are already in place inside a pharmaceutical company's R&D environment can benefit from being connected to the network of knowledge embodied inside an ontology. Some of these underlying technologies that have been enhanced by ontologies are listed below:

Data mining and analysis: Although very good data mining and analysis packages such as Spotfire™, SAS™ or R are in widespread use, the Achilles heel is getting sufficient high-quality information to work from in a timely fashion. The gathering and transformation of systematic multidimensional data into a form that can be analysed can take several

times as long as the actual analysis. Because it is so difficult, this data preparation phase is often curtailed and simpler, less systematic analyses are performed instead. This has the obvious downside that fewer data are considered and important trends or correlations are less likely to emerge.

Ontologies can shortcut this information gathering phase and provide a semantically consistent and comprehensive substrate for data mining and analysis. Because the ontology already contains all of the important concepts and properties from all of the different primary data sources and because this has already been semantically normalised, information can be directly compared in the analysis system. An example of multidimensional information exported from an ontology into the Spotfire DecisionSite™ software is shown in Fig. 1. **Integration of chemical information:** The seamless integration of chemical information with biological and business data is difficult to achieve except in very specific, often single vendor, systems. This has limited the development of several applications where chemical data need to be correlated freely with a range of other information from different sources. Example applications that could take advantage of this would include extended SAR analysis including biological data, chemical clustering based on affected target and pathway information, freedom to operate searches in patent literature or the selection of in-licensing candidates based on correlations to known toxicities. Ontologies again provide a mechanism to represent chemical structure and property data alongside all of the other contextual data, whether it is biological or business oriented in nature. Fig. 2 shows a chemical substructure search using CambridgeSoft's ChemOffice™

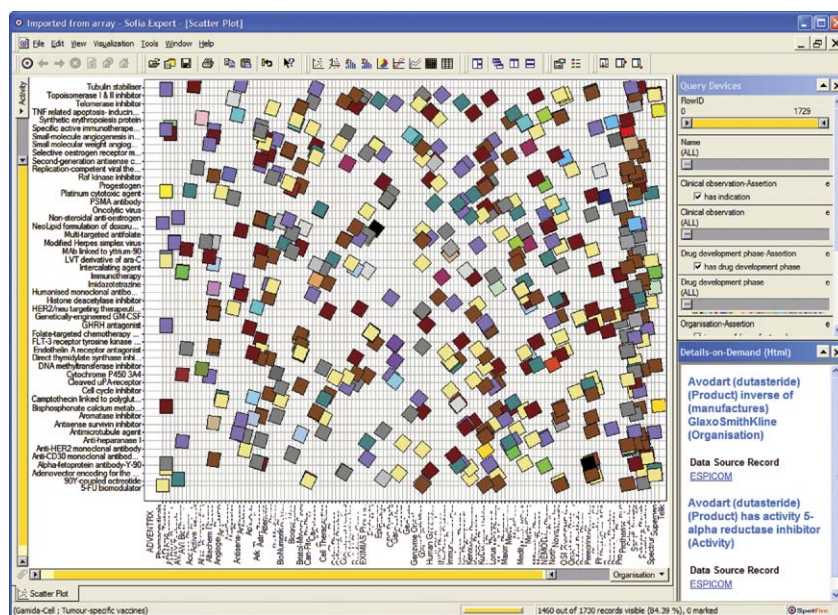


Figure 1. Systematic analysis of anticancer therapeutic pipelines. Image shows all known anticancer therapeutics in development or marketed correlated with manufacturers and activities. The colours of the squares represent the development phase information of the compounds.

Figure 2. Chemical substructure search driving contextual information retrieval. Image shows a substructure search for a fluorobenzene moiety, that has returned Atorvastatin as a hit. This, in turn, has retrieved information regarding the targets, processes, disease, among others, associated with Atorvastatin from a multirelational ontology derived from 45 structured and unstructured data sources.

Figure 3. Large-scale document navigation. Image shows a set of 14 Summary Basis for Approval (SBA) documents from the FDA regarding Rosuvastatin. The ontology has been used to tag references to Rosuvastatin in specific contexts. In the case highlighted, this is showing all references to the different Rosuvastatin-induced events mentioned within the 1200 page corpus.

tool driving the retrieval of contextual information from BioWisdom's SofiaTM ontology.

Document analysis and submission: Navigating a complex document structure and finding the correct context of all references to a particular concept can be daunting in the life science domain where primary documents such as regulatory submissions, Summary Basis for Approval (SBA) documents or patents are typically unstructured, poor quality, image only PDF documents. Ontologies coupled with text tagging tools such as Corpora's Jump!TM as shown in Fig. 3 can provide a navigation framework around a document that makes finding information in large documents fast and intuitive.

Many business applications such as biomarker discovery, market differentiation, in-licensing and safety assessment can take advantage of these and other ontology-powered technologies to bring more knowledge into the decision-making process.

Concluding remarks

Although the technological development around ontologies and their application to pharmaceutical business problems is very recent, they are beginning to live up to their potential and are already demonstrating that they can deliver knowledge inside an enterprise pharmaceutical environment [17,20].

Some of the benefits that they offer might be disruptive to established patterns of behaviour inside the discovery function. For example, the notion that experts can automatically maintain their currency in a domain simply by working in it for years is sometimes misplaced, especially with the pace of change in modern research. Experience suggests that without a prolonged effort, expertise gradually gives way to bias and finally, in some cases, to dogma. This is difficult to recognise but can be a major obstacle to new hypotheses and ideas that are essential to innovation. Dogma exists in many forms, not simply individual but also ingrained in years of laboratory or clinical practice. Often molecules are typecast, with the site in which they were discovered being imbued with a mythical unquestionable status as the definition of their function. When they turn up years later being expressed in different tissues and possibly playing a different role, it comes as a surprise that assays that are used routinely might have different interpretations.

Our ability to continually innovate is predicated on the ability to see new information and patterns and reinterpret past experience. Ontology is one tool that has the potential to provide the objective, comprehensive knowledge substrate that allows continuous researching and reappraisal of hypotheses against new information as it becomes available to challenge old ideology and identify a range of new opportunities.

References

- 1 Bacon, F. (1587) *Meditationes Sacrae*, De Haeresibus
- 2 Huber, P. and Mills, M.P. (2005) *The Bottomless Well*. Basic Books
- 3 Berners-Lee, T. et al. (2001) The semantic web. *Sci. Am.* 284, 34–43
- 4 Sowa, J., ed. (2000) *Knowledge Representation Logical, Philosophical and Computational Foundations*, Brooks/Cole
- 5 Aristotle (2004) *The Categories*, BookSurge Classics
- 6 Rajpathak, D. and Motta, E. (2004) An ontological formalization of the planning task. *Proceedings of the International Conference on Formal Ontologies in Information Systems (FOIS'04)*, Torino, Italy, Nov. 4–6, 2004 (http://kmi.open.ac.uk/people/dnyanesh/Publications/FOIS-2004/Fois2004_camera-ready-final.pdf)
- 7 The Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res.* 11, 1425–1433
- 8 Williams, J. and Andersen, W. (2003) Bringing ontology to the genome ontology. *Comp. Funct. Genomics* 4, 90–93
- 9 Luciano, J.S. (2005) PAX of mind for pathway researchers. *Drug Discov. Today* 10, 937–942
- 10 Gardner, S.P. (Summer 2005) Ontologies: semantic networks of pharmaceutical knowledge. *Drug Discov. World* 78–83
- 11 Leibniz, G.W. (1965) *Monadology and Other Philosophical Essays* (Schrecker, P. and Schrecker, A.M., eds and Trans.), Bobbs-Merrill Co.
- 12 Thomas, J. et al. (2000) Automatic extraction of protein interactions from scientific abstracts, *Pac. Symp. Biocomput.* 541–552 (<http://helix-web.stanford.edu/psb00/thomas.pdf>)
- 13 Melton, G.B. and Hripcsak, G. (2005) Automated detection of adverse events using natural language processing of discharge summaries. *J. Am. Med. Inform. Assoc.* 4, 448–457
- 14 Uramoto, N. et al. (2004) A text mining system for knowledge discovery from biomedical systems. *IBM Syst. J.* 43, 516–533
- 15 Milward, D. et al. (2004) Ontology based interactive information extraction from scientific abstracts. *Comp. Funct. Genomics* 6, 67–71
- 16 Gardner, S.P. (2005) Ontologies and semantic data integration. *Drug Discov. Today* 10, 1001–1007
- 17 Stojanovic, L. et al. (2004) The role of ontologies in autonomic computing systems. *IBM Syst. J.* 43, 598–616
- 18 Rector, A. (2003) Medical informatics. In *Chapter 13 of the Description Logic Handbook* (Baader, F. et al. eds), pp. 406–426, Cambridge University Press
- 19 Rector, A. (2003) Modularisation of domain ontologies implemented in description logics and related formalisms including OWL. *Proceedings of K-CAP'03*, 23–25 October, 2003, Sanibel Island, FL, USA, pp. 121–128
- 20 Torr-Brown, S. (2005) Advances in knowledge management for pharmaceutical research and development. *Curr. Opin. Drug Discov. Dev.* 8, 316–322