DRUG DISCOVERY TODAY BIOSILICO

Ontologies and semantic data integration

Stephen P. Gardner

The increased generation of data in the pharmaceutical R&D process has failed to generate the expected returns in terms of enhanced productivity and pipelines. The inability of existing integration strategies to organize and apply the available knowledge to the range of real scientific and business issues is impacting on not only productivity but also transparency of information in crucial safety and regulatory applications. The new range of semantic technologies based on ontologies enables the proper integration of knowledge in a way that is reusable by several applications across businesses, from discovery to corporate affairs.

The pharmaceutical landscape in 2005

Our information-saturated society produced more raw data between 1999 and 2002 than in the rest of human history [1]. The pharmaceutical industry has been transformed during the past ten years by the adoption of high-throughout technologies such as the human genome initiatives, combinatorial chemistry, uHTS and automated ADME. Unfortunately, the promise of these technologies has largely failed to be borne out in real-world productivity [2]. The generation of all of this information has guaranteed neither its accessibility to the scientist making decisions at the bench nor that the scientist can put those data into their proper context by comparing them with other relevant information. Too often, the data generated by the automated technologies gather in vast silos that are impressive in scale but limited in usefulness to the organization. With different user interfaces, file formats, database systems, operating systems and data semantics, each of these repositories becomes an isolated island of data in the sea of risk and uncertainty that underpins drug discovery, development and safety surveillance. Trapped in these silos, this knowledge is not visible to the rest of the organization (which does not know where to look), nor can it be used as context for making future business decisions.

The big business challenges – establishing and monitoring the safety profile of a compound, differentiating it effectively from competitive compounds in the same therapeutic class, finding alternative indications and defining strategies for leapfrogging generic competition – all rely heavily on integrating a broad range of information in a more meaningful way than the current industry norm. This, in turn, requires a rethink of the value of the underlying information and the ways in which it is managed.

Improving information transparency

In the current pharmaceutical environment, safety issues have become central to not only the removal of compounds with potential liabilities from the pipeline but also the effective differentiation of compounds in the market place. Efficacy alone has never been the definitive metric of the competitive potential of a drug, except in areas of unmet medical need; however, post-Vioxx®, it will be increasingly difficult for compounds without a superior safety profile to be

Stephen P. Gardner
BioWisdom,
Harston Mill,
Harston,
Cambridge,
UK, CB2 5GG
e-mail: steve.gardner@biowisdom.com

1359-6446/05/$ – see front matter ©2005 Elsevier Ltd. All rights reserved. PII: S1359-6446(05)03504-X

accepted into formularies and onto the approved drug lists of the key health maintenance organizations (HMOs) and other prescribing bodies. The rate at which compounds are accepted onto the approved lists of HMOs has already slowed dramatically from weeks to years [as detailed by Karen Katen, the Pfizer (http://www.pfizer.com) Chief Financial Officer, at the Pfizer Q3 2004 Analysts Meeting Webcast], and health insurers are showing major aversion to risk when questions about safety remain [3]. Overall, the pharmaceutical industry is under more pressure today than it has ever been [4,5], and the role of the regulatory authorities has come under much closer public scrutiny than before [6,7].

Much of the underlying mistrust is caused by the unmet desire of the regulators, consumers and analysts to know more about a compound, to know it more quickly and to be able to interpret the information better. For marketed compounds, the whole apparatus of safety-information gathering, integration and analysis has fallen into question, largely because of a failure to keep up with the technological progress made in the past ten years. The current state of the art in adverse-event reporting has only recently replaced paper submission of documents with electronic formats, and even these are largely unstructured and poorly suited to information retrieval. Collecting data in electronic form is only the first step in an extremely long process.

Representing knowledge

Much of the lack of whole-process productivity can be attributed to the inefficient use of information and to difficulties in making knowledge held in distributed data silos visible inside large multidisciplinary organizations. The most promising solution to the problem is data integration. The promise is that, if all of the discrete information can be integrated together so that the islands are connected, a much greater body of knowledge can be presented to a researcher, and better, faster and more well-informed decisions can be taken.

However, data integration is not without its own challenges and pitfalls. Many knowledge management (KM) and data-integration strategies have had limited success because of patchy implementation, incomplete rollout and their inherent technological constraints. One of the biggest technical risks, which contributes to more than half of the KM project failures [8,9], is the scale of the resource required to integrate source data at the beginning of each project. This data integration is usually performed piecemeal in data warehouses and other static repositories and is rarely reusable between projects. Each new project has to perform its own data integration from scratch. This can become such an all-encompassing technical challenge that the project is delayed and misses key functionality.

Traditional data-integration techniques

There are many ways in which data have been integrated, at least four of which have been attempted on a large scale in pharmaceutical R&D.

Rule-based links

Rule-based links (e.g. SRS from Lion Bioscience [10]) comprise the simplest integration strategy. This strategy is based on the fact that many data sources share names for the same gene, protein or chemical, or have explicit cross-references to other databases in their annotations. If, for example, the accession-code field in a GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) record is ‘X56494’ and the database-reference field in the SwissProt (http://www.ebi.ac.uk/swissprot/) record ‘P14618’ also contains ‘X56494’, the two records can be considered ‘equivalent’ or, at least, ‘related’.

Data warehouses

Data warehouses (e.g. Atlas [11]) use specialized database schemas to abstract and store a copy of data from several sources, and enable those data to be queried through a single query. A central fact table that holds only the key pieces of information for each concept is constructed, and any further details and properties are stored in satellite dimension tables to prevent them from affecting the performance of the key business questions that the warehouse is designed around.

Ad hoc *query optimizers*

Ad hoc query optimizers (e.g. Discovery Link [12]) are systems that attempt to find the optimal way of phrasing a question when the data that answer the question might be spread across multiple tables or databases. The user asks a question in a single query interface. The system then devises a strategy for querying the various source databases and it might test query fragments to decide the best way to formulate the query for optimal performance.

Federated middleware frameworks

Federated middleware frameworks (e.g. GRIDS [13–15]) are systems that employ the most advanced integration strategies. They attempt to connect multiple applications and user interfaces to multiple data sources, regardless of the format, type or structure of the underlying data. They require the development of a common representation (or model) of the data contained in the underlying data sources. By enforcing a contract between components so that a given type of data will always be presented in a certain form, middleware systems can achieve great flexibility and are the most effective technique for integrating data, applications and processes in complex enterprise applications.

All of these techniques have strengths and weaknesses. Systems founded on rule-based links suffer from one of the fundamental limitations of integration systems: the combinatorial explosion of connections between data sources. Much less effort is required to develop and support the smaller number of connections between data sources if a common format in the centre is used instead of connecting every possible combination of data sources (Figure 1).

Furthermore, rule-based links systems do not contain semantic knowledge beyond the simple rule, usually of the form 'this value in this field matches that value in that field'. This is an extremely simplistic starting point that could fail to capture the detail and richness of relationships that might exist between concepts coming from multiple sources of evidence, and that is prone to missing valid connections and including invalid connections.

Data warehouses are notoriously difficult to build, expensive to maintain and inflexible to changes in the questions that can be asked. This is largely because they require a copy to be made of data from all of the underlying data sources in a synchronized extraction, transformation and loading (ETL) process. Data not extracted into the warehouse cannot be queried conveniently, and changing the data that are selected involves considerable redesign work. This places a large upfront design burden on the warehouse schema and the ETL process. Fundamentally, data warehouses and data marts are designed to answer a set of specific questions repeatedly as new data become available: something that is typically more useful to a retailer than a drug-discovery application. Data warehouses do not, themselves, embody any semantic knowledge and, for example, the onus of knowing that two compounds with different names are structurally related is on the user composing the questions.

Ad hoc query optimization systems suffer from the same problems created by a lack of semantic integration. However, some also compound this problem by having a limitation that the namespace (the names of columns in the databases) must be non-overlapping. This often requires a structural change to the source databases, with the consequent propagation of disruptive alterations to the applications that access data from those databases.

Federated middleware systems attempt to model the types of data held within the underlying data sources (e.g. gene sequences and expression profiles) and provide common ways of moving those types of data between components and processes. Their success at this depends largely on the semantic richness and granularity of the model that they employ.

Because most of the data-integration projects undertaken within pharmaceutical R&D have lacked true semantic integration, in many cases they have left a legacy of another super-silo of partially linked information that is no more accessible than the original data and that provides no wider a context in which new data can be viewed. In most cases, the integration is static, and new questions and contexts cannot be accommodated easily. In some cases, costly integration projects have been abandoned owing to lack of tangible results.

A more effective approach is to focus explicitly on the representation of knowledge rather than just its management. If a highly descriptive semantic representation of the available knowledge could be built, it could be reused to power a variety of business applications, without the need for repeated bespoke integration exercises. New knowledge gathered from different sources can build upon current knowledge because it all exists in a semantically consistent framework.

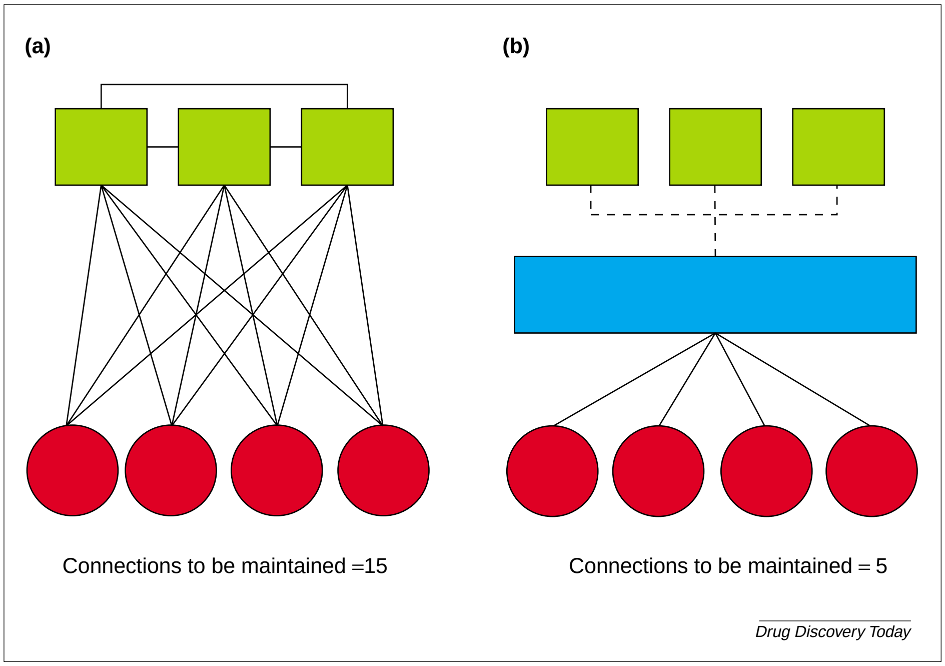

(a)
(b)
Connections to be maintained =15
Connections to be maintained = 5
Drug Discovery Today

FIGURE 1

Differences in connections to be maintained using fully wired and middle-layer connection strategies. The maintenance of connections between data sources to facilitate integration of their content is expensive.The number of connections required to integrate data fully from multiple sources can be substantially reduced from the connection clutter in **(a)** by using a 'middle layer' **(b)** rather than direct connections between sources.

The role of semantics

The key to being able to integrate information in a reusable way is the use of semantics, which describe the meaning of a word or concept. First, semantics are used to ensure that two concepts, which might appear in different databases in different forms with different names, can be described as truly equivalent (i.e. they describe the same object). This can be obscured in large databases when two records that might have the same name actually describe two different concepts in two different contexts (e.g. 'COLD' could mean 'lack of heat', 'chronic obstructive lung disorder' or the common cold). More frequently in biology, a concept has many different names during the course of its existence, of which some might be synonymous (e.g. 'hypertension' and 'high blood pressure') and others might be only closely related (e.g. 'Viagra', 'UK92480' and 'sildenafil citrate'). The ability to distinguish among these synonyms, homonyms and related terms is essential when integrating data from

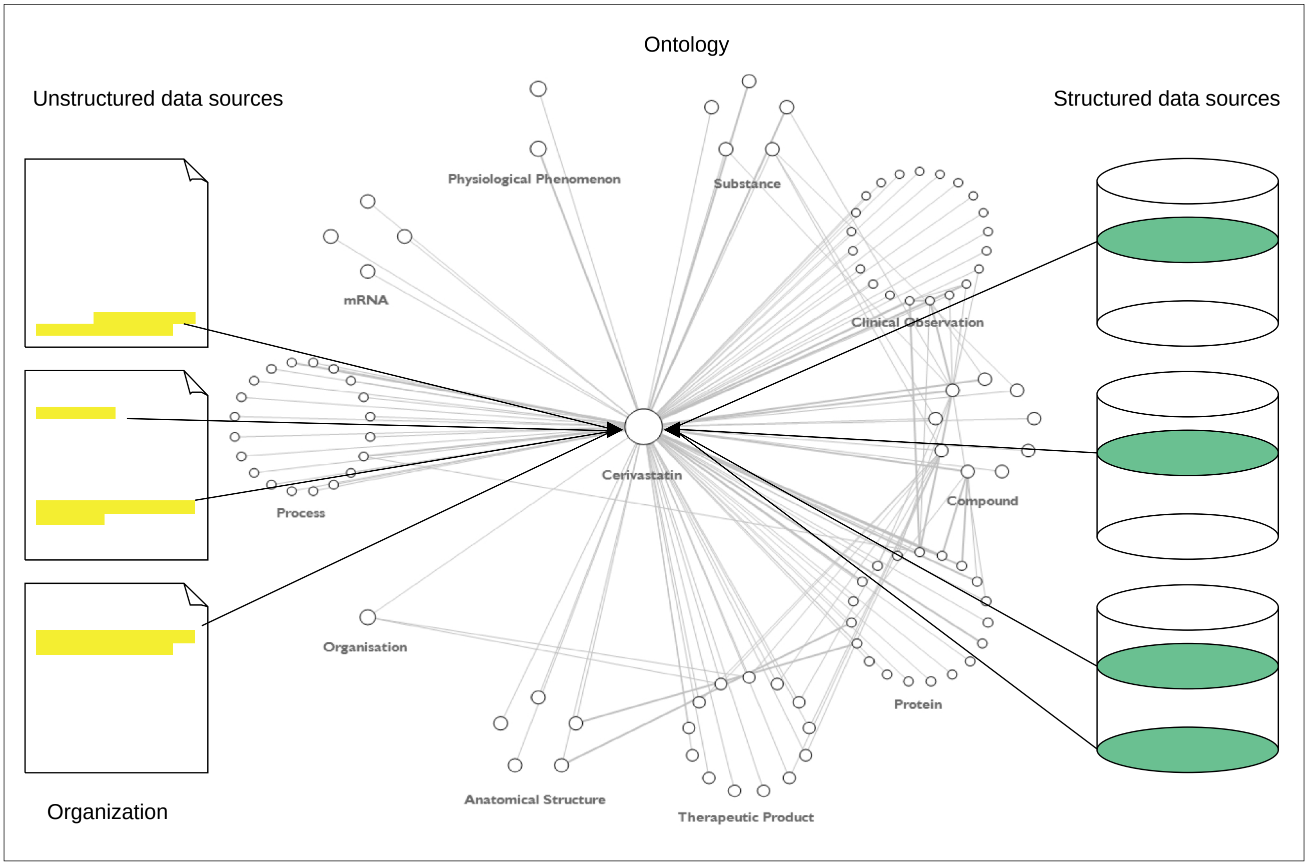


FIGURE 2

Granular semantic data integration using an ontology. An ontology can be used as a semantic middle layer to map references semantically to the same concepts among multiple data sources. In the example shown, the concept 'Cerivastatin' in the ontology is used as a consistent marker to connect data from multiple structured and unstructured data sources that refer to 'Cerivastatin,' 'Baycol' or other equivalent terms.

different repositories. Second, semantics describe the specific form of the relationship that exists between concepts (rather than just co-occurrence in text or lexical equivalence of a label). This enables a more complete and fully descriptive representation of all of the available information, showing what things interact with and what role they might have in a given context.

Semantic technologies, ontologies and beyond

A truly semantic representation of knowledge that is inherently more scalable and flexible, and better able to support multiple existing and future business applications has recently become available at an appropriate scale. This representation is ontology [16], which is a discipline founded in philosophy and computer science that is at the heart of the new wave of semantic technologies such as Semantic Web [17].

An ontology contains a representation of all of the concepts present in a domain and all of the relationships between them. These associations between concepts are captured in the form of assertions that relate two concepts by a given relationship. These triplets (variously described as concept–relationship–concept, subject–predicate–object or assertions) are the building blocks of most ontology formats, including the open Semantic Web standards of RDF and OWL (http://www.w3.org/TR/owl-ref/). In addition to commercial suites, several open-source tools have been developed to support the application of such standards [18,19].

These assertions can, in their simplest form, use IS-A relationships (e.g. 5HT2B IS-A SEROTONIN RECEPTOR) that, when aggregated, build into a hierarchy or taxonomy. The taxonomies of concepts (and relationships) can be extremely useful in their own right, especially when the concepts are annotated with properties such as synonyms. This enables users to specify high-level 'family' concepts such as GPCR when performing a search or selecting data for analysis.

True ontologies, however, have a broad range of relationships among concepts such as IS-EXPRESSED-IN, BINDS-TO, HAS-AFFINITY-FOR and IS-USED-FOR-TREATMENT-OF [20]. These relationships also include all known synonyms. This enables, for example, all of the many English variant forms of the BINDS-TO relationship between proteins and compounds to be used to build a complete and detailed picture of the interactions around a protein or protein family.

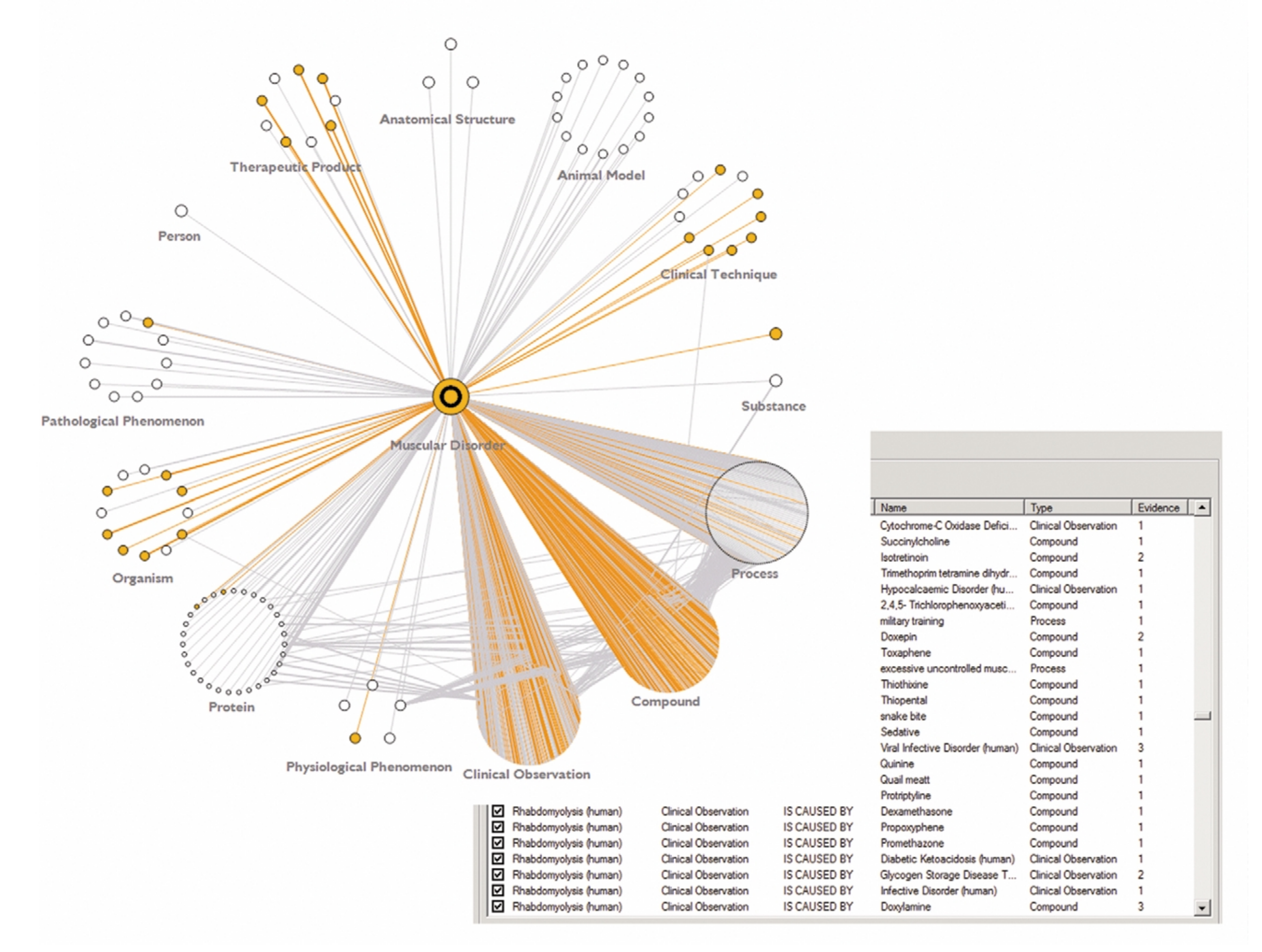
Although ontologies have been providing a metaphysical description of the world since Aristotle and his Categories [21], they have recently started to be applied at scale for the computational representation of domain knowledge [22,23]. Ontologies are currently being applied to not only pioneering research projects [24–27] but also real-world enterprise-scale projects in major pharmaceutical companies [28].

An example of such a project is the use of ontology to guide a natural language processing (NLP)-based text-mining tool in the discovery and analysis of cofactors of androgen and liver X α and β receptors from a corpus of >7500 documents. At a level of quality equivalent to an expert scientist, the ontology-based system improved the search productivity by a minimum factor of ten [29]. The ongoing value of this type of productivity enhancement alone, regardless of all other applications, has been estimated by a major pharmaceutical company to be worth tens of millions of US dollars (presentation by William Hayes at InfoTechPharma, March 2005, London).

Ontologies as a data-integration enabler

The map of concepts and relationships in an ontology provides a crucial enabling resource for true semantic data integration. An ontology enables information from one resource to be mapped accurately at an extremely granular level to information from another source. In the example shown in Figure 2, multiple instances of a concept (or its synonyms) in different structured or unstructured data sources can be mapped to a specific ontology concept and, therefore, the data in those original sources can be integrated semantically.

The ontology provides the common vocabulary for the data integration – showing the preferred names for a concept, and the synonyms and properties associated with it. This enables forward-looking integration by collecting data using names that are already well understood rather than ones that might not be shared widely throughout the organization. This makes the assimilation of new data easier and quicker, and facilitates communication between groups [30]. Organizing data integration around the ontology provides the middle layer that makes data integration more efficient – reducing the cost, maintenance and risk of the project. Furthermore, because the ontology can be grown over time as new data become available, new



				Name	Type	Evidence
				Cytochrome-C Oxidase Defici...	Clinical Observation	1
				Succinylcholine	Compound	1
				Isotretinoin	Compound	2
				Trimethoprim tetramine dihydr...	Compound	1
				Hypocalcaemic Disorder (hu...	Clinical Observation	1
				2,4,5- Trichlorophenoxyaceti...	Compound	1
				military training	Process	1
				Doxepin	Compound	2
				Toxaphene	Compound	1
				excessive uncontrolled musc...	Process	1
				Thiothixine	Compound	1
				Thiopental	Compound	1
				snake bite	Compound	1
				Sedative	Compound	1
				Viral Infective Disorder (human)	Clinical Observation	3
				Quinine	Compound	1
				Quail meatt	Compound	1
				Protriptyline	Compound	1
☑	Rhabdomyolysis (human)	Clinical Observation	IS CAUSED BY	Dexamethasone	Compound	1
☑	Rhabdomyolysis (human)	Clinical Observation	IS CAUSED BY	Propoxyphene	Compound	1
☑	Rhabdomyolysis (human)	Clinical Observation	IS CAUSED BY	Promethazone	Compound	1
☑	Rhabdomyolysis (human)	Clinical Observation	IS CAUSED BY	Diabetic Ketoacidosis (human)	Clinical Observation	1
☑	Rhabdomyolysis (human)	Clinical Observation	IS CAUSED BY	Glycogen Storage Disease T...	Clinical Observation	2
☑	Rhabdomyolysis (human)	Clinical Observation	IS CAUSED BY	Infective Disorder (human)	Clinical Observation	1
☑	Rhabdomyolysis (human)	Clinical Observation	IS CAUSED BY	Doxylamine	Compound	3

FIGURE 3

Causes of skeletal muscle toxicity highlighted in an ontology. Graphical and tabular views showing the causes of skeletal muscle toxicity integrated in a semantically consistent way from >45 structured and unstructured data sources. The graph shows all of the knowledge gathered about muscle disorders from the data sources. Each node represents a different concept, and each line (edge) represents a specific relationship among the nodes. The 'CAUSES' relationships are highlighted in orange. Image reproduced, with permission, from the BioWisdom sofiaBrowser™.

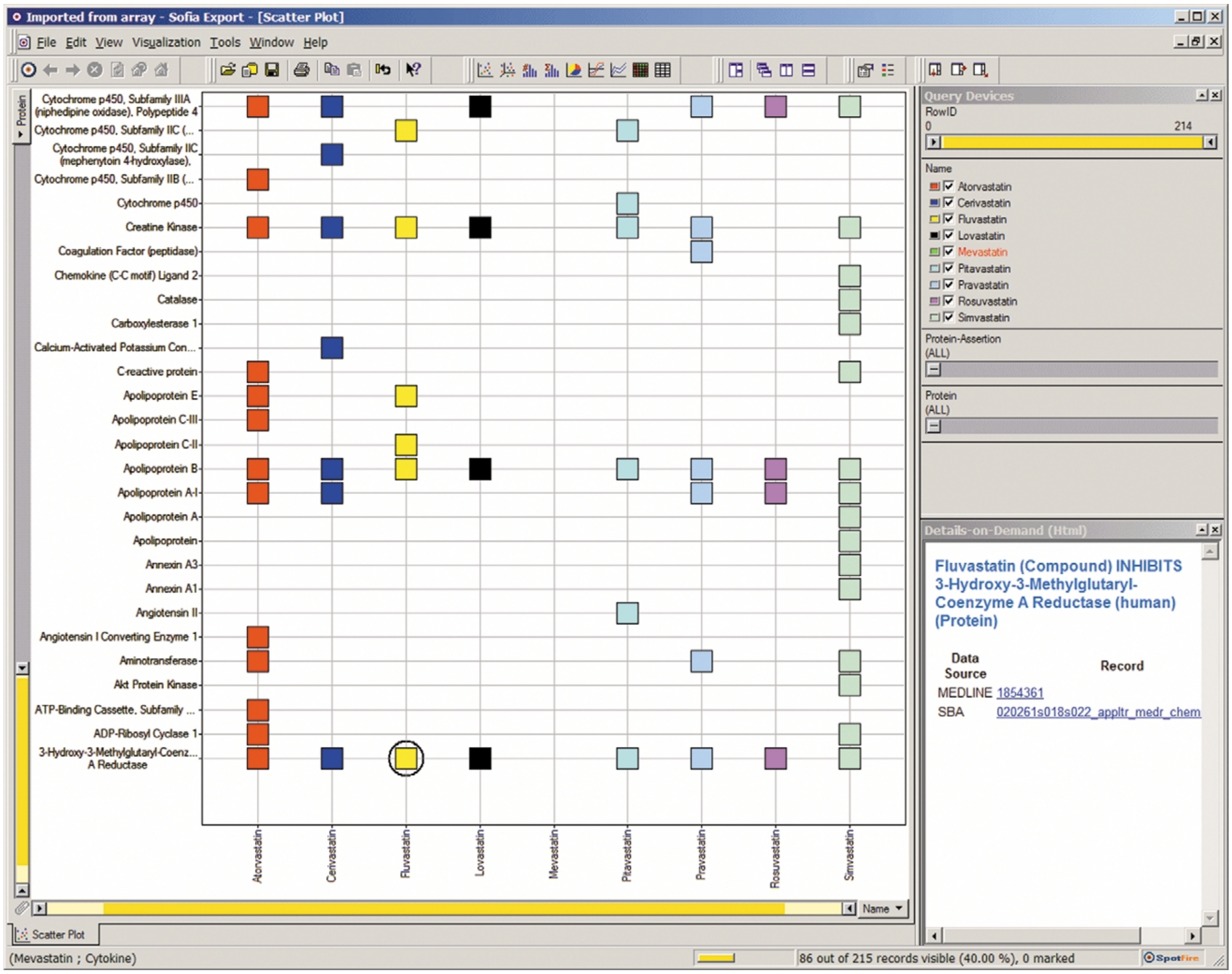
links are continually being made and new knowledge assimilated in the ontology.

Ontologies enable real business applications

Semantic data integration can be done at enterprise scale for real business applications. Figure 3 illustrates a semantically consistent view of safety data highlighting all of the reported causes of rhabdomyolysis (skeletal muscle toxicity). These data have been integrated from 45 data sources, including existing vocabularies (e.g. UMLS, GO and ICD), structured data (e.g. genomics, proteomics, HTS and chemical structure) and unstructured data [PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed), patents and the FDA (http://www.fda.gov/) Summary Basis for Approval documents].

Because an ontology contains a wide range of semantically consistent information, it provides a powerful environment in which to perform systematic data mining – asking and answering questions with the full authority of all of the available data sources in a fraction of the time taken to read the first 100 headers of a keyword search. This type of analysis can be used in applications as diverse as biomarker discovery, alternate indications discovery, in-licensing, market differentiation and predictive toxicology. Figure 4 shows such a systematic analysis of the statin family, illustrating the different protein interactions that the various marketed statin compounds have been reported to display. This type of analysis could lead to testable hypotheses about the pathogenesis of the toxicity and to an understanding of why the different statins have different reported rates of incidence of a specific side effect.

Obviously, creating detailed ontologies on a large scale is a new and challenging discipline. PubMed alone has >15 000 000 abstracts from which a broad range of information can be extracted, and this is only one of, potentially, hundreds of relevant data sources. Undertaken manually and without a strong commitment to process, quality assurance (QA) and ontological structure, ontology building



Fluvastatin (Compound) INHIBITS 3-Hydroxy-3-Methylglutaryl-Coenzyme A Reductase (human) (Protein)

Data Source	Record
MEDLINE	1854361
SBA	020261s018s022_appltr_medr_chem

(Mevastatin ; Cytokine)

86 out of 215 records visible (40.00 %), 0 marked

FIGURE 4

Systematic analysis of statins showing their respective interactions with various proteins. Image shows the various protein-binding profiles of the different members of the statin family, derived from an ontology that integrated 45 data sources and that was displayed in the Spotfire DecisionSite™. This enables immediate comparison of binding profiles, and the generation and subsequent validation of hypotheses about the pharmacological basis of the differences in the observed rates of side-effect incidence among different members of the statin family.

can consume major resources (similar to those currently spent in building taxonomies and thesauri manually) and might lead to poor-quality ontologies [21,30]. Several ontology-building projects have failed to achieve the scalability required to represent information at an enterprise scale.

However, the construction of ontologies can be automated more readily than that of taxonomies. High-quality ontologies can be created relatively quickly using a broad range of workflow, automation, text mining, NLP, inferencing and pattern-matching algorithms, a strong QA-based methodology and process, and with a strong domain ontology to guide the discovery and creation of new ontology content.

Concluding remarks

Ontologies provide a highly dynamic and flexible map of the information contained in the data sources within a domain. Because ontologies enable true semantic integration across the data sources that they represent, it is possible not only to draw wider conclusions from the data but also to look at the data from several distinct perspectives relevant to the specific job being undertaken. The generation of ontologies to represent data from several underlying data sources is a precursor to and an important enabler of semantic data integration. Ontologies make data integration more efficient and more detailed, and reduce the risk associated with the continual redevelopment of project-specific integration strategies.

Ontologies form an atlas of all of the knowledge of an organization as embodied in its databases, licensed data sources and personal observations of its scientists. The ontology grows as new data sources are added to it, becoming a core corporate asset rather than another new super-silo for data. The ontology can underpin a range of applications that delivers the right information at the right time to make better-informed decisions throughout the lifecycle of drug discovery, development, marketing, sales and post-market surveillance. In terms of safety, identifying new patterns of toxic response, understanding the molecular basis of adverse events and formulating a response to events in the marketplace such as the Vioxx® withdrawal depends on the delivery of just such a set of consistent and systematic information.

With this transparent view of the whole landscape of knowledge available to the organization, a pharmaceutical company can begin to redefine its best practice for handling information: making informed scientific and business decisions, and communicating with its key constituents – the regulators, analysts and customers.

References

1 Lyman, P. and Varian, H.R. (2003) How Much Information? 2003 (http://www.sims.berkeley.edu/research/projects/how-much-info-2003)

2 Anon. (2001) High Performance Drug Discovery Accenture (http://www.accenture.com/xdoc/en/industries/hls/pharma/hpdd.pdf)

3 Appleby, J. (2004) What do you Believe When Drug Messages Conflict? *USA Today* December (http://www.usatoday.com/money/industries/health/drugs/2004-12-26-crestor-cover_x.htm)

4 Simmons, J. (2004) Pharma's annus horribilis: a world of hurt. *Fortune* 27 December

5 Baum, R. (2004) Outlook for pharmaceuticals. *Chem. Eng.* News 82, 3

6 Leaf, C. (2004) Why the FDA keeps blowing it. *Fortune* 29 November

7 Abbasi, K. (2004) Is drug regulation failing? BMJ DOI: 10.1136/bmj.329.7477.0-g (www.bmj.com)

8 Cutter Consortium (2002) Corporate use of data warehousing and enterprise analytic technologies (http://www.the-infoshop.com/study/cu11950_data_warehousing.html)

9 Friedman, T. (2005) Data integration forms the technology foundation of EIM (http://www.gartner.com/research/spotlight/asset_120116_895.jsp), Gartner Group

10 Etzold, T. *et al.* (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.* 266, 114–128

11 Shah, S.P. *et al.* (2005) Atlas – a data warehouse for integrative bioinformatics. *BMC Bioinformatics* 6, 34–49

12 Haas, L.M. *et al.* (2001) DiscoveryLink: a system for integrated access to life sciences data sources. *IBM Systems Journal* 40, 489–511

13 Stevens, R.D. *et al.* (2003) myGrid: personalised bioinformatics on the information grid. *Bioinformatics* 19 (Suppl. 1), i302–i304

14 HealthGrid Association and Cisco Systems (2004) HealthGrid – a summary (http://whitepaper.healthgrid.org)

15 Foster, I. (2003) The grid: computing without bounds. *Sci. Am.* 288, 78–85

16 Sowa, J., ed. (2000) *Knowledge Representation Logical, Philosophical, and Computational Foundations*, Brooks/Cole

17 Berners-Lee, T. *et al.* (2001) The Semantic Web. *Sci. Am.* 284, 34–43

18 Noy, N.F. *et al.* (2003) Protege-2000: an open-source ontology-development and knowledge-acquisition environment. *AMIA Annu. Symp. Proc.* 2003, 953

19 Lambrix, P. *et al.* (2003) Evaluation of ontology development tools for bioinformatics. *Bioinformatics* 19, 1564–1571

20 Smith, B. *et al.* (2005) Relations in biomedical ontologies. *Genome Biol.* 6, R46

21 Schulze-Kremer, S. (2002) Ontologies for molecular biology and bioinformatics. *In Silico Biol.* 2, 179–193

22 Rosse, C. and Mejino, J.L.V., Jr (2003) A reference ontology for biomedical informatics: the foundational model of anatomy. *J. Biomed. Inform.* 36, 478–500

23 *The Gene Ontology Consortium* (2001) Creating the gene ontology resource: design and implementation. *Genome Res.* 11, 1425–1433

24 Bader, G.D. *et al.* (2004) *BioPax:* an OWL early adopter perspective (http://lists.w3.org/Archives/Public/public-swls-ws/2004Sep/att-0063/w3c-position-paper_BioPAX_new.pdf)

25 Hartel, F. *et al.* (2004) OWL/RDF/LSID utilization in NCI cancer research infrastructure (http://lists.w3.org/Archives/Public/public-swls-ws/2004Sep/att-0039/W3_Position_Paper.htm)

26 Murray-Rust, P. *et al.* (2004) Representation and use of chemistry in the global electronic age. *Org. Biomol. Chem.* 2, 3192–3203

27 Pisanelli, D.M. *et al.* (2004) Coping with medical polysemy in the Semantic Web: the role of ontologies. *Medinfo* 11, 416–419

28 Forsberg, K. and Andersson, B. (2004) We believe RDF/OWL and LSID can help Pharma out from the InfoMaze! (http://lists.w3.org/Archives/Public/public-swls-ws/2004Sep/att-0049/We_believe_RDF__OWL_and_LSID_can_help_Pharma_out_from_the_InfoMaze__2004-09-15.htm)

29 Milward, D. *et al.* (2004) Ontology-based interactive information extraction from scientific abstracts. *Comp. Funct. Genomics* 6, 67–71

30 Eilbeck, K. *et al.* (2005) The sequence ontology: a tool for the unification of genome annotations. *Genome Biol.* 6, R44