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 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).