The cost to bring a new drug to market in 2007 will exceed $1.3B, and it will have taken 16 years to move from inception to market. Pharmaceutical research and development costs will exceed $35B in 2007, up from $3B in 1980. Productivity, measured in terms of NCE or NME approvals per year, will remain flat or will decrease—a trend that has been consistent since 1996. Pharmaceutical companies must evaluate many diverse opportunities to decrease R&D costs and increase productivity and profitability. Informatics organizations have risen to the challenge of decreasing operating costs while delivering increased business value through a variety of innovative technologies discussed in this review.

Strategies to support drug discovery through integration of systems and data

Chris L. Waller, Ajay Shah and Matthias Nolte

Chemistry Informatics, Pfizer Global Research and Development, Pfizer, Eastern Point Road, Groton, CT 06340, United States

Much progress has been made over the past several years to provide technologies for the integration of drug discovery software applications and the underlying data bits. Integration at the application layer has focused primarily on developing and delivering applications that support specific workflows within the drug discovery arena. A fine balance between creating behemoth applications and providing business value must be maintained. Heterogeneous data sources have typically been integrated at the data level in an effort to provide a more holistic view of the data packages supporting key decision points. This review will highlight past attempts, current status, and potential future directions for systems and data integration strategies in support of drug discovery efforts.

Introduction

No review on drug discovery would be complete without a brief review of the pressures that shape the industry. It has been reported that an estimated 16 years and an excess of $1.3B will be required to bring a new drug to the market in 2007 [1]. Many factors have contributed to this trend that has produced an increase in the cost and time and a decrease in the profit associated with drug discovery and development (Figure 1). Among the well-known external factors are (1) increased scrutiny of drug safety assessment that increases dramatically the cost incurred during the clinical stages, (2) generic competition, and (3) re-importation [2]. While these factors are largely outside the direct control of pharmaceutical companies, there are a number of factors that are internal to the business and can be used as levers in reducing the overall time and cost to bring a drug to the market. Many companies have invested heavily in a wide variety of technology solutions in order to increase productivity and efficiency.

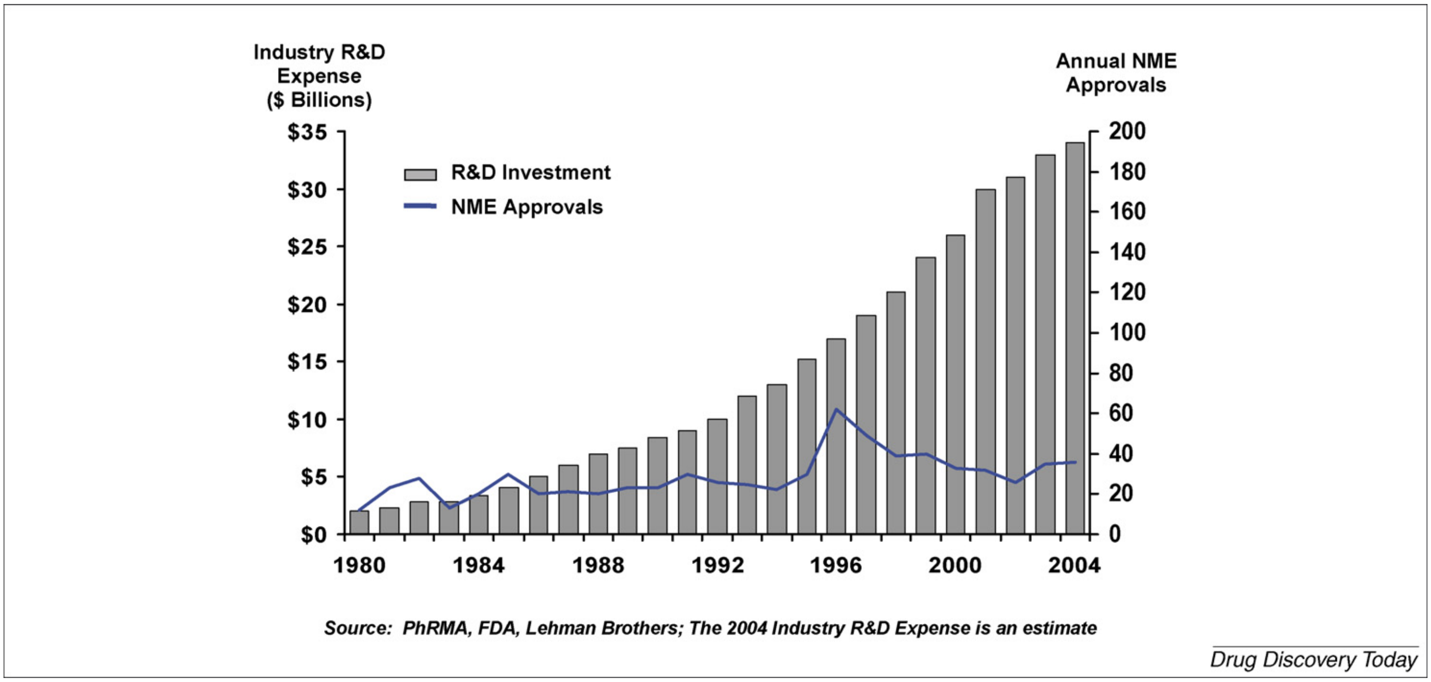
To date, these investments have not yielded results on par with the level of investment (Figure 1) and that has left most large pharmaceutical companies looking for additional cost saving opportunities, either in the core drug discovery business units (chemistry, biology) or in the functional support lines (informatics). A recent IBM report speculates that a 5% improvement in productivity at the early research stage can save $9.9 million per New Chemical Entity (NCE), but more significantly, can reduce the discovery time by 3.3 months [3]. This time difference can determine whether the NCE results in a drug or loses in its competition.

Informatics imperatives

In the mid 1990s consultants at Accenture surveyed the informatics landscape in the pharmaceutical sector, in an effort to understand better the value proposition. The results of this survey suggested that there was, in general, a consistent lack of forward-looking strategies. Specifically, it was suggested that in order to be successful, it was imperative that informatics/information technology groups do the following: (1) establish an integrated informatics strategy that defines and prioritizes activities on the basis of the impact on discovery (business) goals, (2) drive informatics with an integrated view of process, people requirements, and impact, (3) implement an integrated data and application architecture to support the needs of individuals and functional and therapeutic areas via integration, (4) build integrated decision support solutions to support the business process, and (5) focus on generating, integrating and delivering knowledge to the entire organization that enables rapid and consistent analysis and decision-making on the basis of common goals and values.

Corresponding author: Waller, C.L. (chris_l_waller@pfizer.com)

1359-6446/06/$ - see front matter © 2007 Published by Elsevier Ltd. doi:10.1016/j.drudis.2007.06.007



Source: PhRMA, FDA, Lehman Brothers; The 2004 Industry R&D Expense is an estimate

FIGURE 1

Annual research and development expenses for the pharmaceutical industry and productivity measured as the number of new molecular entities during the period 1980–2004 [1].

Evolutionary pressure on informatics systems through advances in science and technology

The late 1980s and early 1990s witnessed the emergence of combinatorial/parallel chemistry and high-throughput screening (HTS) as mainstream technologies utilized in the drug discovery process. Before this, the typical drug discovery program would involve designing, synthesizing, and evaluating hundreds or maybe thousands of candidates. Computational models used for the design of new compounds were routinely created using data sets of fewer than 100 compounds and the techniques were optimized for this low-throughput scenario. As HTS and combinatorial chemistry emerged, it was quickly determined that many of the techniques that were used to design new compounds, such as quantitative structure–activity relationship (QSAR) modeling and structure-based drug design techniques, did not scale to accommodate the dramatic increase in input. New techniques, such as binary QSAR [4] and rapid docking and scoring [5,6], were developed as derivatives of the older methods to meet the emerging needs of the shifting drug design paradigm, with the goal of adding decision-making data points earlier within the discovery workflow. Additional informatics capabilities that were developed to support the evolving drug discovery process included (1) *in silico* library design and analysis [7], (2) large-scale data management techniques to support analysis of the human genome [8], (3) receptor/target characterization techniques [9], (4) similarity searching techniques in 2D and 3D structure databases [10], and (5) the creation of conformationally rich databases to search for pharmacophores [11].

In addition to the novel computational methods that were developed, entirely new techniques were designed and implemented to capitalize on the growing amount of information stored in the databases, which were being created as a result of the new drug discovery paradigms. Data warehousing and mining techniques were introduced to allow better and effective analysis of millions of data points [12]. Information exploitation became a popular term, if not a real activity, in this time frame as well. Finally, the field of knowledge management [13] was popularized in an attempt to capitalize on the shared learnings of others in our journeys to discover the next blockbuster drug, which might already be hidden in our vast databases. In total, these techniques were developed to support our ability to transform data into knowledge. For instance, laboratory notebooks in a pharmaceutical company may describe hundreds of thousands of reactions. Informatics organizations are increasingly asked, ‘Is there a way to digest these data and convert them into useful knowledge that can guide future synthetic efforts’?

The vast amount of connected data available to the typical drug discovery scientist allowed the formulation of questions and hypotheses that were previously unapproachable. Questions like

- I wonder if there are similar compounds to compound A?
- Compound A has biological activity against Target 1, are there other compounds with this activity?
- Compound A has biological activity against Target 1, are there other targets for this compound?
- I have Target 1, I wonder if there are similar targets? and
- Do similar compounds to Compound A show similar activities in similar targets?

became commonplace and put evolutionary pressure on information systems to provide integration of internal proprietary and external public-domain repositories of chemical and biological data. Additional ‘data streams’ of simulated or calculated data could also be integrated to enhance the information being delivered to the scientists in an effort to answer the ‘Holy Grail’ question—Can I get all the information from all the sources for a given compound?

Having all these data available at one's fingertips introduces new challenges to the scientist, specifically with respect to the relevance of available data to the problem of interest. On the surface, it appears that the relative identity of the compounds, targets, or diseases, and so on, under examination could be used to provide an indication of the strength of the relationship, ranging from precisely similar to precisely dissimilar. But, does similarity correlate to relevance? Perhaps, if we can learn anything from the Google page-ranking model in which an ingenious method for extracting and ranking the most relevant information from the vast amount of 'stuff' available on the net is by using a so-called 'secret formula'. Can this paradigm be modified to work in a drug discovery setting? As a first approximation, it appears that an improved representation of all the screening results for a compound, as well as inclusion of synthetic feasibility and synthetic pathway information could be facilitated by using on-the-fly search clustering techniques [14].

The need for data and application integration

In modern pharmaceutical companies, research is generally decentralized, with teams working in various geographical locations and varied therapeutic areas. The data generated by these teams and the accumulated interpretive knowledge or actionable data are generally retained in silos. The challenges in managing large-scale research are related to culture, operational model, and tools. They include creation of a culture where knowledge is shared; incentives to collaborate; consolidation of data and tools that have originated from various mergers and acquisitions; tools that facilitate collaboration across geographical and therapeutic areas; interpretation of data and conversion of raw data to knowledge; curation tools that facilitate deposition of interpreted knowledge; facilitation of utilization of 'silo-ed' experts across the organization by creating Community of Practitioners (CoPs); making knowledge available in proper context and collaboratively learning from successes and failures (learning organization). A data centric view of these challenges includes the acquisition of data from multiple sources; automated or manual curation of data to avoid storing 'junk'; integration of proprietary, in-house data and commercially or publicly available data into data warehouses; federated or consolidated querying tools; and automated tools for data mining, visualization, and utilization in further research.

As described above, the tasks of the modern day drug discoverer have been greatly altered from those of their predecessors because of the numerous scientific and technological advances that have led to an explosion in the amount of data generated in pursuit of the next blockbuster drug. The availability of data from varied and heterogeneous sources, coupled with the desire to build knowledge bases and collaborative networks, has driven the need for improved integration techniques at the fundamental data level. All downstream activities, such as information exploitation and knowledge management, are crucially dependent on the effective integration of data and tools.

Data integration

The simplest forms of integrated data systems use the network connecting the various data sources as the integration layer. Under this paradigm, typical standard query language (SQL) practises implemented in standard database technologies are used to link, join, and merge data sources using key (common) fields. Views of the underlying data can then be implemented to create the appearance of a single merged data source. Extensions of this basic database technology exist in the recently advanced physical and virtual data warehousing technologies.

In the late 1990s, the large repositories of data being created, managed, and mined by data intense financial and related industries drove the need for better data management techniques. The corporate information factory was developed as a response to this need. In short, an information factory manages the full data life cycle by progressing it from the capture to the presentation layers through a series of dedicated storage areas (transactional databases, operational data stores, data warehouses, and operational data marts) [15]. As data move from one level in the factory to the next, there is an opportunity to manage, manipulate, and harmonize it by utilizing extract, transform, and load (ETL) process logic (Figure 2).

Alternatively, virtual data warehousing is undergoing a revival in the industry. One might speculate that the sheer size of a drug discovery company's data warehouse would be cost prohibitive to create and maintain. A virtual data warehouse created as a 'view' into the authoritative data sources, without the physical space requirements, is a financially attractive alternative. An early example of this technology is present in the current IBM Information Server/Federated Server software, which is based on IBM's 'garlic' [16] technologies that IBM introduced in the late 1990s. More recent implementations include MetaMatrix [17] and Composite [18]. In all of these technologies, the importance of metadata, or 'data about the data' that define the rules that are used to determine which data fields are equivalent across the various data sources and can thus be linked are tantamount to a successful implementation.

In February 2004, The World Wide Web Consortium released the Resource Description Framework (RDF) and the Web Ontology Language (OWL) as W3C recommendations. RDF is used to represent information and to exchange knowledge in the Web, while OWL is used to publish and share sets of terms, ontologies, supporting advanced Web search, software agents, and knowledge management. Taken together, these semantic technologies are poised to drive the next generation Internet, or the semantic web [19], as well as provide a flexible and agile data integration methodology with the goal to make context sensitive information available from across the network and facilitate the knowledge sharing. Currently, a number of pilot studies [20] are being performed to demonstrate the value of semantic technologies as data integration tools.

Application integration

In order to support the various individual scientific tasks in a drug discovery workflow, numerous specialty software packages are required; however, the desktop capacity is as limited as the scientists' tolerance for training on disparate applications. Our goal, as informatics professionals, is to support and advance workflow by providing scientists with appropriate software services at the appropriate time.

It is common to find software applications that have been designed to support use cases, or singular or small series of related tasks (e.g. enter a list of compounds and compute Rule of Five

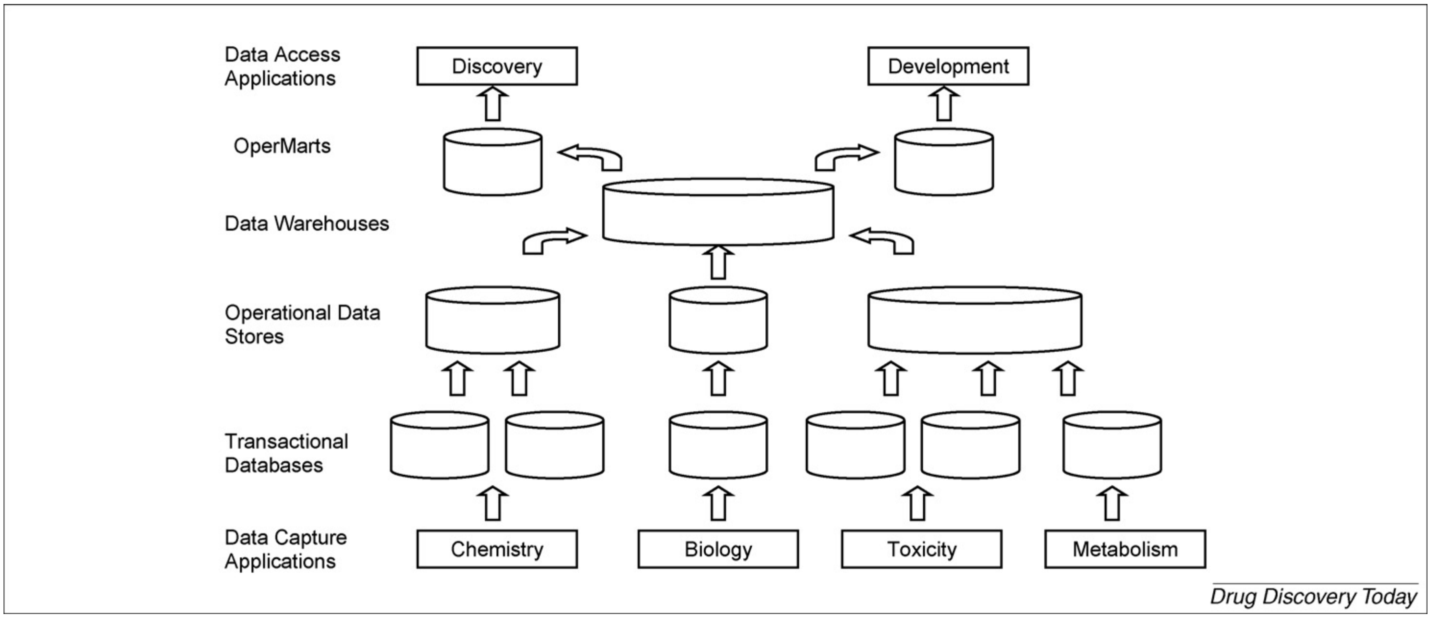


FIGURE 2

An illustrative schematic representation depicting data flow represented by arrows, from data capture mechanisms through an information factor framework to data access mechanisms.

parameters [21]). Less typical, but emerging, are applications that are utilizing workflow engines [22], which directly mimic individual workflow steps (tasks, or capabilities, e.g. design and synthesize a plate compound on the basis of a hit identified in an HTS) in the most flexible manner to enable rapid adaptation to the ever-evolving discovery workflow.

Perhaps the earliest attempts to provide an integrated application suite were seen with the rise of static web portals, or home pages, which incorporated best practise guides and maps to navigate through the use of hyperlinks to relevant applications and data sources. As an evolutionary advance over static portals, dynamic portal solutions [23] provide basically the same functionality in a more user customizable and dynamic environment through the use of widgets, portlets, and page parts.

Much can be learned by examining the techniques employed by Microsoft to facilitate application integration. For example, a workflow such as ‘I want to send an e-mail to a co-worker listed in my contacts to ask them to attend a meeting’ catalyzed the amalgamation of three legacy products—Contacts, Calendar, and Exchange/Mail into one: Outlook. A single user interface with workflow-specific needs is addressed by tabs or card decks and task supported by steps or wizards. The Microsoft Office trio of applications, Word, Excel, and Powerpoint illustrate a different style of application integration. In this example, there was no overarching, or unifying, workflow that necessitated the integration and merging of entire code bases, and these can therefore be regarded as siloed applications. Rather, through the sharing of objects, using object linking and embedding (OLE) technology, it is possible to integrate functions, data, or views from any one of these applications into another in an ‘on demand’ fashion, thereby promoting the reuse of objects.

The next generation object reuse and software development will be supplanted by a virtual paradigm where web services are orchestrated dynamically to power composite applications in a service-oriented development paradigm. Web services, the cornerstones to service-oriented architecture are the building blocks, or logical small entities, that are needed to facilitate a need or task in the underlying computing environment and, therefore, can inherently support the recent more and more automation-driven discovery workflow processes. These logical entities – capabilities – provide their business value not in themselves but in their dynamic implementation into ever-changing workflow implementations. The capabilities, which are constant, contrary to their implementations, therefore need to be carefully identified and designed.

Definitions and standards are crucial to make a service-orientated architecture (SOA) software development infrastructure work efficiently and support a drive toward vendor neutral implementations through abstracted access to the vendor proprietary libraries or programs. A clear separation between high performance authoritative sources (many create, read, update, delete (CRUD) databases) and query databases (few) in tandem with their respective services and GUI components is essential to provide optimal data quality and update as well as access performances (Figure 3). ETL processes transform disparate data entities to structured information in support of knowledge discovery through the added value of association (as described in Figure 3) just as workflow applications transform capabilities to business value.

There are a number of technologies on the horizon that could, potentially, dramatically change the way in which informatics organizations design, develop, deliver, and support applications and data infrastructures to deliver maximum value to drug discovery organizations. Previously mentioned was the concept of the dynamic portal, which was noted as being well suited for passive applications where data are pushed or pulled to client. As an alternative, AJAX, a combination of two older technologies in the form of Asynchronous JavaScript and XML, has been

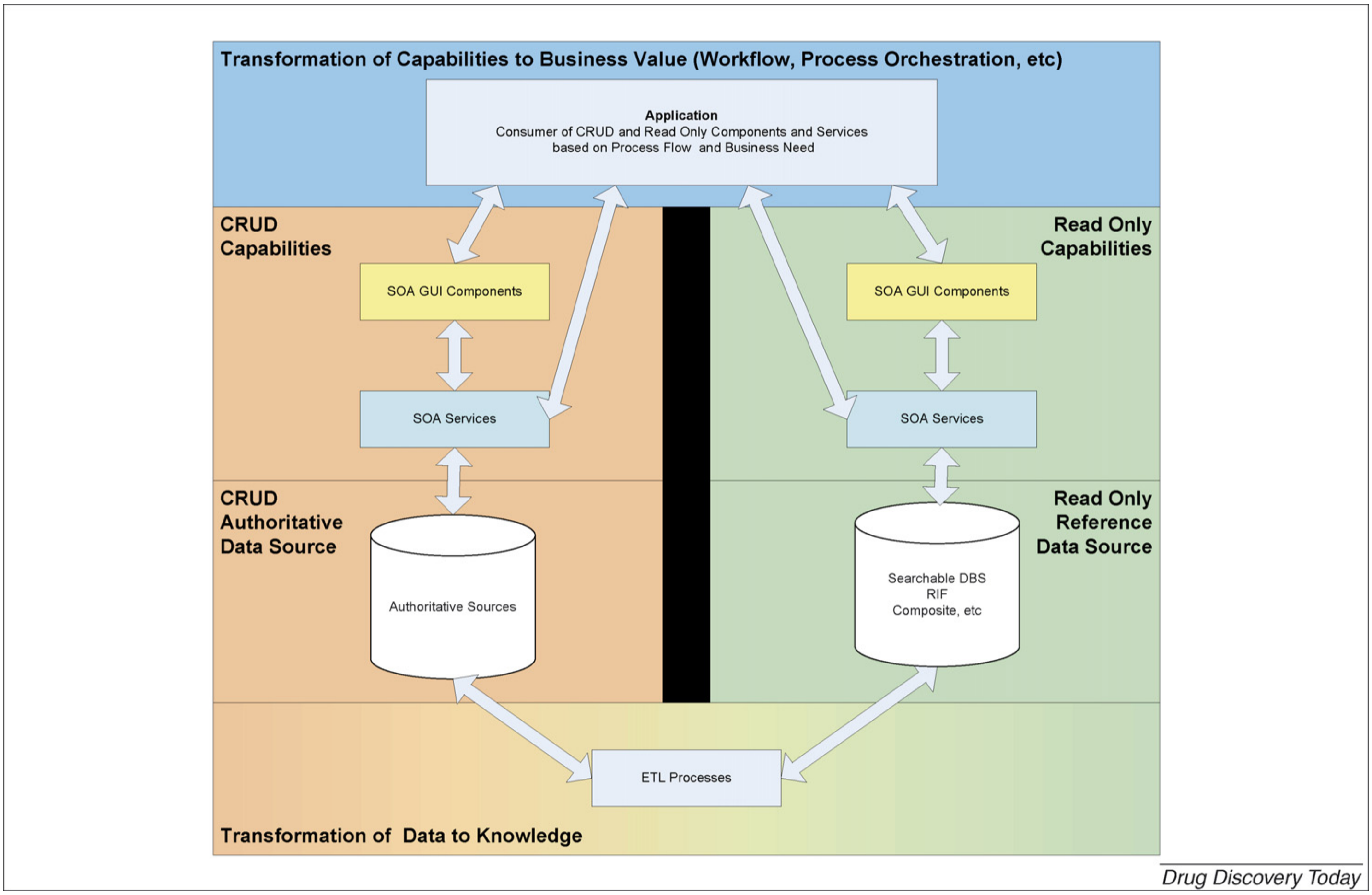

Transformation of Capabilities to Business Value (Workflow, Process Orchestration, etc)
Application
Consumer of CRUD and Read Only Components and Services based on Process Flow and Business Need
CRUD Capabilities
Read Only Capabilities
SOA GUI Components
SOA GUI Components
SOA Services
SOA Services
CRUD Authoritative Data Source
Read Only Reference Data Source
Authoritative Sources
Searchable DBS
RIF
Composite, etc
ETL Processes
Transformation of Data to Knowledge
Drug Discovery Today

FIGURE 3

An illustrative schematic respresentation of a logical architecture of a SOA compliant informatics system.

developed. It is envisioned that this technology might be appropriate for more active portals where data manipulation is required.

As an extension to user customizable interfaces, self-driven or wizard-driven assembly and presentation of GUI elements to support dynamic and variable workflow scenarios could take us into a world where the application becomes a transient entity, existing only as long as required to support the immediate business need. Informatics portfolios under this scenario would comprise underlying services, user interface components, and application frameworks.

Finally, the notion of self-assembling code has been quite compelling for some time, but it has yet to make a real mainstream impact. Attribute-oriented software development technologies and techniques such as xDoclet may ultimately reduce our basic unit of currency in the informatics portfolio to the 'class' level with applications being developed and delivered in an informatics framework that is self-aware and immediately responsive to the ever-changing business needs.

Additional considerations

Informatics organizations are using a number of tactics to reduce the cost base and enhance the productivity of the drug discovery scientist. Primarily among these is the use of lower cost providers for commodity services such as application support, programming, and project management. In April, 2006, Fortune reported that 8 of the top 10 outsourcing firms were engaged in some sort of informatics/information technology support activity (Table 1).

In addition to pursuing lower cost staffing options, informatics organizations are working to evolve the typical vendor relationship model into a more strategic partnership model. However, the current landscape precludes full utilization of such partnerships. Informatics organizations would like to provide best-of-breed applications to discovery scientists, but because of the incompatibility of different vendor-provided applications, a seamless exchange of data and workflow between multiple applications is not possible. While other industries have been successful in evolving Open Standards, attempts at creating standards in computational chemistry and Chemistry Informatics have been less than successful. A primary goal of these partnerships is the development of innovation-enabled applications. Innovation, and applications, may come in an organization from multiple directions—informatics, discovery scientists, vendors, and contractors. How this will be facilitated is open for discussion, but here are some suggested criteria: (1) All parties should adhere to Open Standards for User Interfaces; (2) all parties should adhere to SOA for flexible, efficient, and reusable back-end services; and (3) all applications should provide a framework for plug-ins. On a more functional level, a sandbox environment for code creation and a curation process for promoting code to enterprise level must be developed and adhered to.

TABLE 1

Top 10 outsourcing firms as reported by Fortune magazine in April 2006

Rank	Company	Services
1	IBM	CRM; HR Mgmt.; Information and Communication Technology Mgmt.
2	Sodexho Alliance	Real Estate and Capital Asset Mgmt.; Facility Services
3	Accenture	HR Mgmt.; Information and Communication Technology Mgmt.; Financial Management
4	Hewlett-Packard	Information and Communication Technology Mgmt.; Financial Mgmt.; Imaging and Printing
5	Capgemini	CRM; Information and Communication Technology Mgmt.; Financial Mgmt.
6	ARAMARK	Facility Services; Uniform and Career Apparel
7	Wipro Technologies	CRM; Information and Communication Technology Mgmt.; Transaction Processing
8	CGI Group	HR Mgmt.; Information and Communication Technology Mgmt.; Transaction Processing
9	Unisys	Information and Communication Technology Mgmt.; Corporate Services; Transaction Processing
10	Cognizant	Information and Communication Technology Mgmt.

It is expected that more strategic partnerships will result in product offerings that are better suited for the needs of drug discovery organizations, which could manifest themselves in the delivery of components, services, and frameworks instead of individual applications or ‘one size fits all’ product suites. As discussed above, these needs are multi-pronged and must strike a balance between maintaining cost while promoting operational efficiencies.

Finally, advances made in the networking arena have provided large pharmaceutical companies some level of confidence and promoted the exploration of opportunities to leverage externally hosted data, services, and applications. VPN security now enables access to disparate sources of data as feeds or as on-demand services. The advantages of this type of data access model are exemplified by such data repositories as those that house inventory on commodity items (e.g. chemical reagents). Real-time access to current inventory information is provided by linking the internal chemical ordering application to the external data host that in turn is linked to the inventories of suppliers. The end result is a clear example of increased efficiency in the order fulfillment process and decreased cost base realized through fewer internal data servers, data management systems, and administrators.

Conclusion

The cost to bring a new drug to the market in 2007 will exceed $1.3B, and it will have taken 16 years to move from inception to market. Pharmaceutical research and development costs will exceed $35B in 2007, up from $3B in 1980 (Figure 1). Productivity, measured in terms of NCE or NME approvals per year, will remain flat or will decrease—a trend that has been consistent since 1996 [1,2]. Pharmaceutical companies must look for ways to decrease R&D costs and increase productivity and profitability. Informatics organizations have risen to the challenge of decreasing operating costs while delivering increased business value through a variety of innovative technologies discussed in this review.

References

1 Gilbert, J. *et al.* (2003) Rebuilding big pharma’s business model. In *In Vivo: The Business and Medicine Report* (vol. 21)

2 Goozner, M. (2004) *The $800 Million Pill: The truth Behind The Cost Of New Drugs.* University of California Press

3 Wang, A.E. (2003) Rx for pharmaceutical companies: internal collaboration is the key to improved innovation, *IBM Institute for Business Value,* IBM Corporation

4 Gao, H. *et al.* (1999) Binary QSAR analysis of estrogen receptor ligands. *J. Chem. Inf. Comput. Sci.* 39, 164–168

5 Venkatachalam, C.M. *et al.* (2003) LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graph Model* 21, 289–307

6 Kitchen, D.B. *et al.* (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* 3, 935–949

7 Balakin, K.V. *et al.* (2006) Rational design approaches to chemical libraries for hit identification. *Curr. Drug Discov. Technol.* 3, 49–65

8 Letovsky, S.I., ed. (2006) *Bioinformatics: Databases and Systems,* Kluwer Academic Publishers

9 Vajda, S. and Guarnieri, F. (2006) Characterization of protein-ligand interaction sites using experimental and computational methods. *Curr. Opin. Drug Discov. Dev.* 9, 354–362

10 Willett, P. (2005) Searching techniques for databases of two- and three-dimensional chemical structures. *J. Med. Chem.* 48, 4183–4199

11 Kearsley, S.K. *et al.* (1994) Flexibases: a way to enhance the use of molecular docking methods. *J. Comp.—Aided Mol. Des.* 8, 565–582

12 Hand, D. *et al.* (2001) *Principles of Data Mining,* MIT Press

13 Drucker, P.F. *et al.* (1998) *Harvard Business Review on Knowledge Management.* HBS Press

14 http://vivisimo.com/html/biometacluster

15 Inmon, W.H. *et al.* (2001) *Corporate Information Factory.* John Wiley and Sons, Inc.

16 Haas, L.M. *et al.* (2000) Integrating life sciences data-with a little garlic. *IEEE International Symposium on Bio-Informatics and Biomedical Engineering*

17 http://www.metamatrix.com

18 http://www.compositesoftware.com

19 http://en.wikipedia.org/wiki/Semantic_web

20 http://www.w3.org/2005/04/swls/BioDash/

21 Lipinski, C.A. (2000) Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* 44, 235–249

22 http://www.teranode.com

23 http://www.plumtree.com