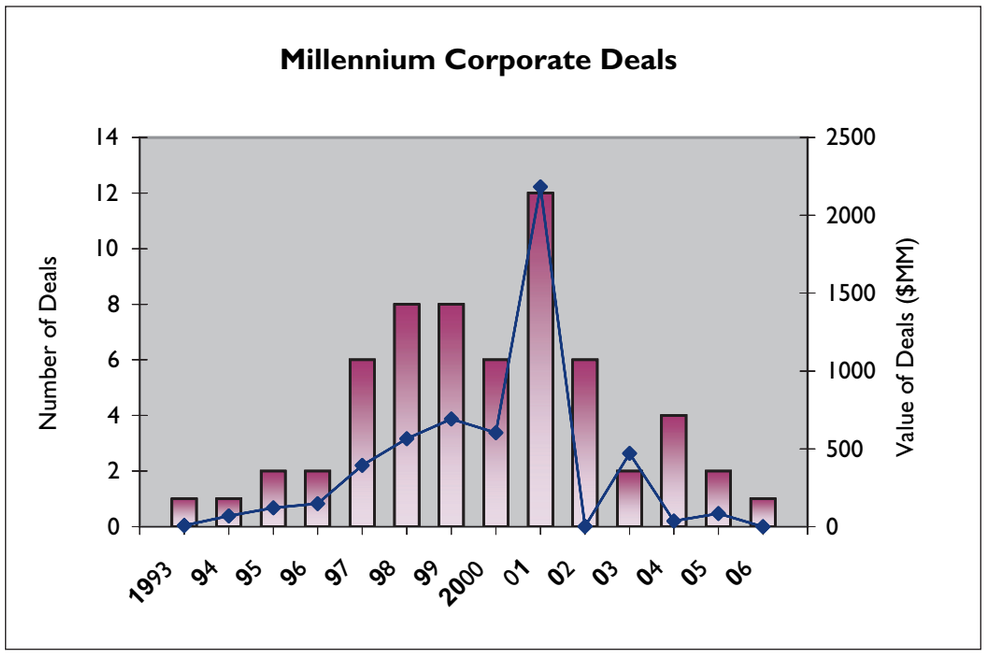
The evolving world of drug discovery data integration in the pharmaceutical industry

The world of drug discovery has been impacted significantly by an ever evolving use of computer systems during the past 25 years with the greatest change occurring during the past 10 years or so as a consequence of the genesis of extremely large datasets from bioinformatics and cheminformatics operations. As new chemical and biological entities make their way through the R&D pipeline, there is a strong need to co-ordinate the use of electronic data and records to manage this process more effectively. We are moving from an era of high throughput data management to a more fully integrated environment where access to full project data is needed for effective and efficient decision management. The evolving environment of Discovery Informatics at Millennium Pharmaceuticals has incorporated many of the changes experienced by the industry during this cycle. The various options and models being explored to create a workable solution for the industry will be described in this article.

Dr David Sedlock

In 1999 the Harvard Business School published two case studies highlighting the meteoric rise of Millennium Pharmaceuticals, Inc as a prominent company in the newly minted world of genomics targeted drug discovery[1,2]. Millennium had reached its notoriety at that time through a combination of timely deal making, a strong commitment to excellence in developing a new technology platform, and demonstrated success with the companies it had developed a business relationship. During its early growth years Millennium built up a strong reputation for being able to 'talk the talk' and 'walk the walk' in the area of new drug target identification and validation. The company sought and signed a significant number of business deals to help it grow as a new biotech company in this rapidly evolving world. Between its genesis in 1993, with its first investment of $8.45 million in seed money, until the launch of Velcade© in 2003, the first drug that it shepherded through clinical testing and the regulatory process, Millennium had completed some 54 deals worth a total of $4.8 billion (**Figure 1**). This included a number of acquisitions and corporate spin-offs as Millennium forged a business model toward the goal of a fully integrated biopharmaceutical company.

Figure 1
The history and value of Millennium business deals



Creation of Millennium Discovery Informatics platform

Millennium, along with several other companies of its genre (eg Celera Genomics and Human Genome Sciences) were at the forefront of what many saw as a new revolution in the world of drug discovery to help fill what was commonly seen as a dearth of new chemical entities moving through the regulatory processes for approval as marketed drugs[3]. To accomplish this plan Millennium made a sizeable investment in the creation of a technology platform to help manage the terabytes of data that were being generated as part of the growth of the company (**Table 1**). Managing the massive growth of data became an immense undertaking for the company and Millennium responded through the creation of a significant infrastructure devoted to the storage, analysis, processing, access and viewing of these various data types. The goal has been the same throughout the industry: to develop tools to enable the conversion of data to information, and the assimilation of this information to knowledge, with the intended use of this knowledge to aid in the discovery and development of new treatments for known diseases.

This information flow from data to knowledge creation forms the basic process that comprises our Informatics infrastructure. For the first 10 years of the company's existence, this Informatics platform was geared toward the established need to find more validated drug targets and to apply these targets to the drug discovery process. The various programme technologies cited in **Table 1** that were created and developed to accomplish this task generated significant technology hurdles that were attempted to be solved through a variety of mechanisms, including a substantial investment in software developed by Millennium engineers. The company's foray into the world of software development allowed it to create major new tools in the area of genomics data mining, differential expression analyses, tissue banking, image processing and data viewing that enabled the scientists at Millennium and at its corporate partners to understand quickly the value of the plethora of data fields that were being mined.

Changes to the environment

The evolution of the Informatics environment mirrored the company's growth and business priorities. Almost from its inception the company focused on creating the data mining tools needed to support the explosion of genomics data with a vast infrastructure following to handle the pipelining of sequence data and the concomitant application of these data to the identification of discrete proteins that could be associated with disease targets. A major effort was devoted to the analysis of differential expression profiles and the creation of numerous cDNA libraries for cell expression experiments to help interpret these data and to create a database for target validation. This was

linked with an extensive tissue banking project to handle samples that were being received from a variety of research collaborations, all of this coming together to form an extensive bioinformatics platform that was used by hundreds of scientists at Millennium and its business partners (**Figure 2**).

The creation of this platform provided a mechanism for Millennium to analyse, annotate and identify hundreds of new potential targets that formed the basis for discovery programmes; however Millennium, along with the rest of the industry, soon recognised that it wasn't going to create a new drug discovery paradigm based on the bioinformatics data alone. When you look back on the evolution of the biotech industry, especially when you examine the various technology innovations during the past 20 years, it has become clear that the new technologies that enabled the use of parallel chemical synthesis, the use of high-throughput screening, the analysis of thousands of mRNA expression profiles, etc were not the panacea that many had predicted. The vast swell of new drugs resulting from the use of these tools has not materialised. A more pragmatic assessment of the environment revealed that, in the term 'drug discovery', the key was the use of the word 'drug', that is the industry had lost site of the fact that it was looking for biological and chemical substances that could elicit a beneficial and predictable pharmacological effect in the treatment of a defined disease. The vast investment in technology in the late 1980s into the early 2000s helped fuel the notion that the new drug discovery paradigm was all about moving more data, more compounds and more targets through the pipeline when, in fact, it was and still remains a science of pharmacology, ie the understanding of how a substance, when introduced into the body, elicits and maintains a biological effect. During this technology era we became vastly improved at our ability to discovery, characterise and document bioactive substances tested *in vitro*; however, the ability to translate this information into sustainable increases in our true drug discovery ability did not materialise as predicted.

The Discovery Informatics platform at Millennium has evolved to encompass this revised paradigm. With the evolution of the company into a more fully integrated biopharmaceutical entity,

Data Storage Allocation	
INFORMATICS ENVIRONMENT	**ALLOCATED STORAGE (TB)**
Genomics platform	5.5
Expression profiling	2.1
Biobanking	0.1
Image management	2.5
Compound/Bioassay data	0.24
Compound inventory	0.03
Electronic notebook	0.07

Table 1
Compilation of Millennium Discovery Informatics data storage environment allocated across the biology and chemistry platforms

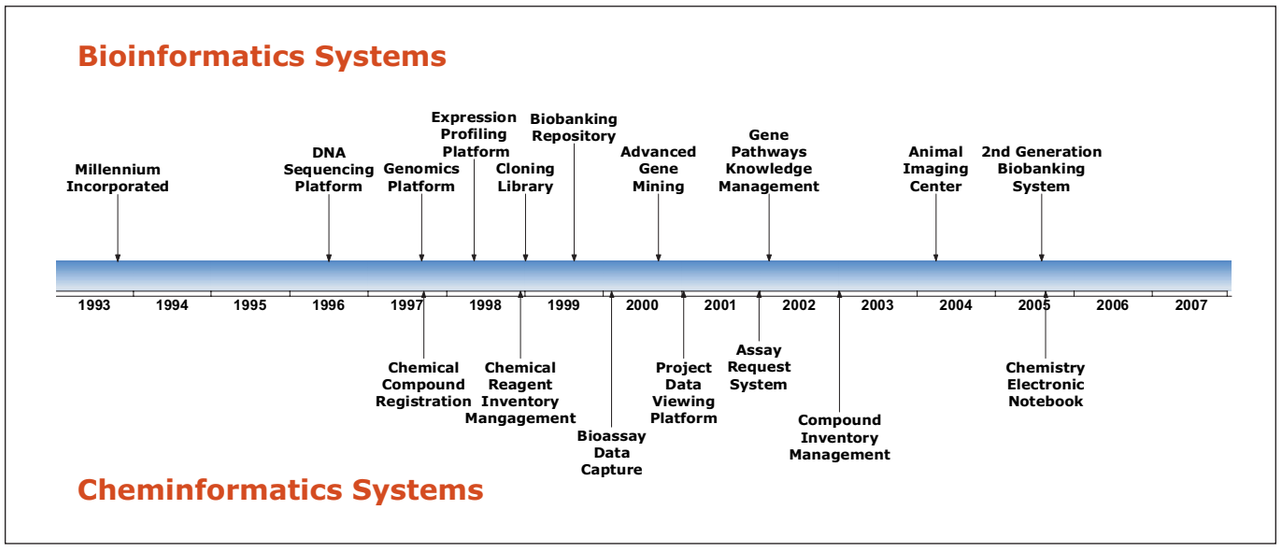


Figure 2
Timeline of the major Discovery Informatics systems deployed at Millennium

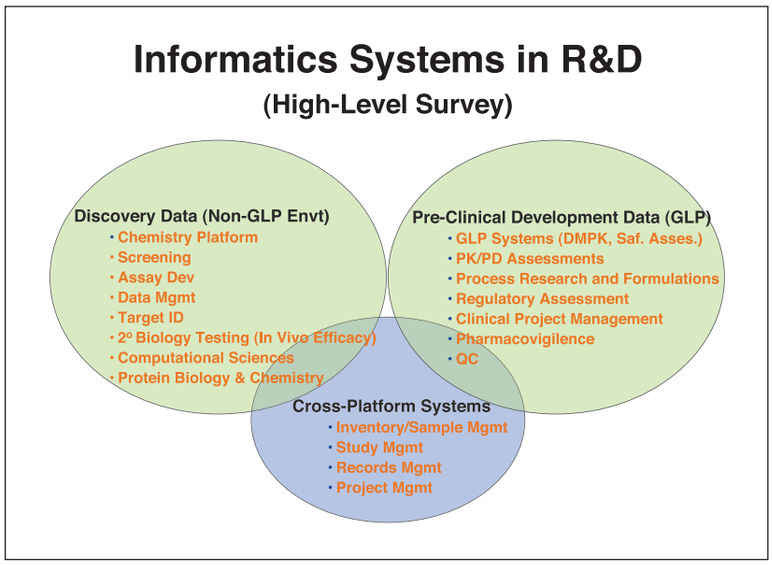


Figure 3
Overview of drug discovery and preclinical drug Development Informatics systems environment

the informatics environment has also shifted. As the bioinformatics platform was maturing, a new emphasis began to emerge to improve the abilities to manage compound data and a major cheminformatics platform was developed. Admittedly, early on this moved down the path of familiarity to create a substantial data store associated with the large amount of compound screening data that was being generated, and we have deployed an array of tools that integrate our compound and bioassay data (**Figure 2**). This includes the ability to register compounds, manage the storage of our compound inventory, monitor assay requests and publish resulting bioassay test data for viewing by the project teams. While our bioinformatics platform was created to help manage internal as well as partnered projects, our cheminformatics platform has evolved primarily to handle only internal discovery programmes. And, while neither Informatics platform has existed for more than 10 years, we have already begun to experience the pains of the large pharmaceutical company in terms of vast amounts of legacy data described in **Table 1** that need to be considered as part of the full lifecycle of any software system.

Incorporating the development environment

What we have seen at Millennium has been also true for the industry, in general. As the full Discovery Informatics platform has evolved, especially during the past several years, it has become more obvious that a more holistic approach is needed to our use of software systems for discovery data management. As reviewed above, the changes to our informatics environment previously have been driven to a large extent by changes to the technology platform. Science and engineering efforts have been geared toward problem solving those technical hurdles formed by the creation of massive amounts of data. We have developed no lack of applications, data stores, wrappers, daemons, services, etc to manage the migration of data from instrument to server and back to the desktop computer. Where we are currently lacking solutions, however, is in the integration and management of very disparate datasets that are associated with the full drug discovery and development pipeline.

Let me illustrate. If we look at our discovery and early stage development environments at a high level, we see an informatics landscape that can be divided based on the discovery pipeline phase (**Figure 3**) of our projects. To some degree this is in line with the validation state of our software systems. Integrating data across these systems presents numerous hurdles, not only because of the different requirements set for the Gxp and non GxP environments but because of the different data models set up for the discovery and development areas: for discovery we're dealing with recording, managing and displaying large data sets for terabytes of data. In the development world we are dealing with fewer but far more complex sets of assay and data types. The database management world can be reduced to a simple metaphor of a very long and skinny table structure in discovery and a short and very, very fat schema for the development platform. Marrying these two environments is a challenge.

If we zoom in at a lower level, we are looking at the merger of two crucial data generation and management areas: the biology platform focusing on target ID, validation and assay development and the chemistry platform focusing on compound management and testing. The 'touch-point' is the identification of a development candidate to move forward with the focus becoming more around project management and decision making (**Figure 4**). It is really at this stage of managing project data where we have entered a truly new informatics model. There are numerous software packages and systems available to assist the project manager with task management, schedules, resources, budget, etc; however, there is a dearth of tools available to manage the various assay results, efficacy data, and associated pharmacokinetic and pharmacodynamic profiles that are reviewed at project team meetings for crucial decision making.

The most common approach is a mixture of Excel files, PowerPoint slides, and e-mail to compile and communicate these data. During the past few years there has been an attempt to address the integration of these data through electronic notebook technology; however, the deployment of this solution, especially in the biology world, has been slow to materialise. The bottom line is that the lifeblood of any biotech or pharmaceutical company is the ability to make rapid decisions that correctly predict the outcome of a new drug in the next test. If you predict correctly, you succeed; if you don't, the project, the programme, even the company can fail.

A contemporary approach

So, how do you construct an informatics environment that contributes to this process – that positively impacts the decision-making process? Millennium and the rest of the industry are grappling with this issue. We have explored different solutions and, admittedly, do not propose to have developed the ideal approach – we are still learning. Over the years, working through several different options and working solutions, I can comment on a few select jewels that have helped us navigate this minefield. These can be divided into three areas:

1. Software system environment – One recommendation is to implement system lifecycles into the platform so that you are able to plan for technology changes, software system use changes and wholesale environment evolution. Software should be deployed for a purpose and that purpose should translate to use. Once that use changes or even disappears, the software system needs to change as well. All systems require the implementation of a maintenance schedule but too often the maintenance becomes a major cost burden. Every system deployed should have decommissioning built in as part of its lifecycle. A corollary to this is the need to maintain an accurate network architecture map. Our informatics systems environment is under constant pressure from environmental change (application upgrades, OS patching, database maintenance, hardware changes, etc). We place a huge burden on our support staff to maintain this environment amidst these constant intrusions and an accurate map of our network architecture including an updated application inventory, configuration plans, and dependency grid is critical to enable implementing a solid test plan for these changes.

2. Data archiving – A second component is the development of an active data archiving model to manage the constant growth of data and the fact that only a very small percentage of stored data is

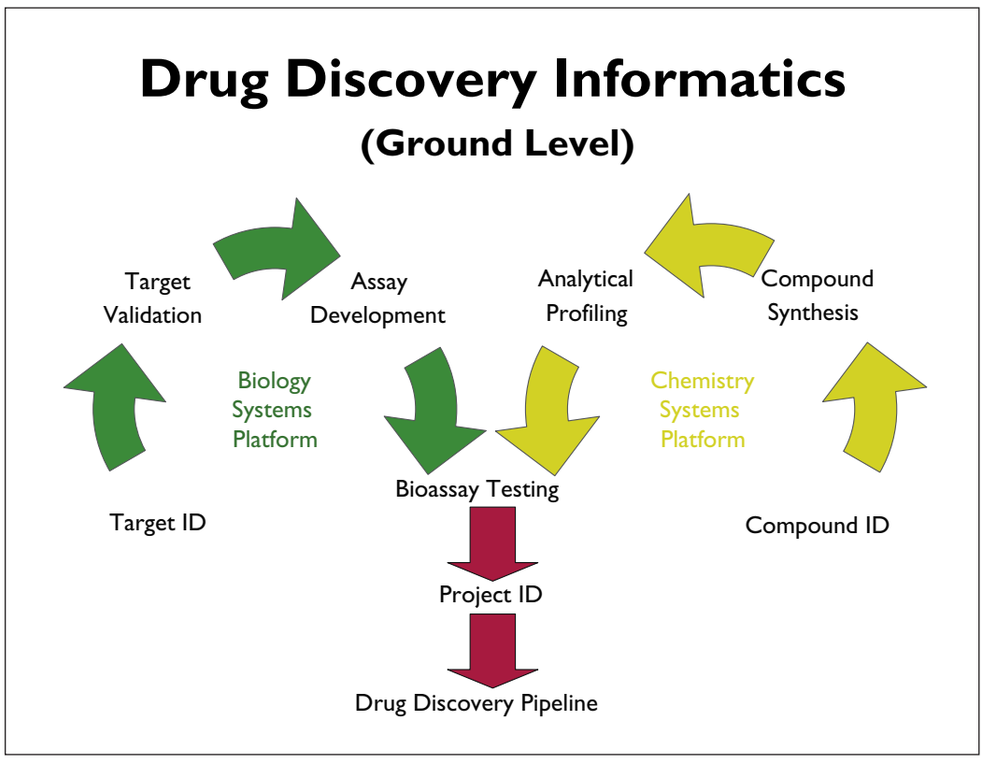


Figure 4
Component analysis of drug discovery activities requiring Informatics support

ever accessed or otherwise used. Over time the value of the data decreases. With ever lower costs for storage, we have been lulled into a sense of data laziness where it is easier to keep adding to the data store than to clean it up. Ask anyone who has worked on merging large datasets from a corporate acquisition (even from the same application, let alone from different systems) about the value of keeping active data to a minimum size.

3. Data integration – A third area for consideration is the most difficult: this is the creation of a data integration strategy that can be used to present needed and relevant data to those who need access to the data for decision-making purposes. When constructing an informatics platform to navigate this minefield, we have taken the approach of looking to the user to capture the processes that control the decisions. When we have constructed these systems, we have tried to incorporate the different elements associated with our varied user community: (a) recognising that it is a diverse group with different needs and different work processes, (b) understanding that there is a key issue dealing with the question of 'data ownership' and 'data sharing', (c) considering roles for the very useful but sometimes demanding power user – they contribute significantly to our needs for problem solving but, at times, can command attention overshadowing the needs of the other less experienced users, (d) and trying to accommodate the fact that the decision-making environment changes as the projects move downstream.

The issue of data integration and the technologies available to manage this effort have been discussed at some length in previous reviews[4-6] with the two major approaches in the past focusing on the pros and cons of centralised data warehousing vs a distributed approach using a federated architecture[7]. Both approaches have their merits depending on the size and complexity of the impacted data stores, the development and maintenance costs of the respective systems, and the desired outcome (ie how will these 'integrated' data be used). At Millennium we have tended to move down the distributed model, though our current cheminformatics platform is highly dependent on a central data store for all chemistry and bioassay information. In our development of these solutions we have approached the problem using a combination of internally built systems, customised vendor products, customised commercial frameworks, and commercial out-of-the-box systems – a pretty common mixture of software systems in the creation of a large Discovery Informatics architecture.

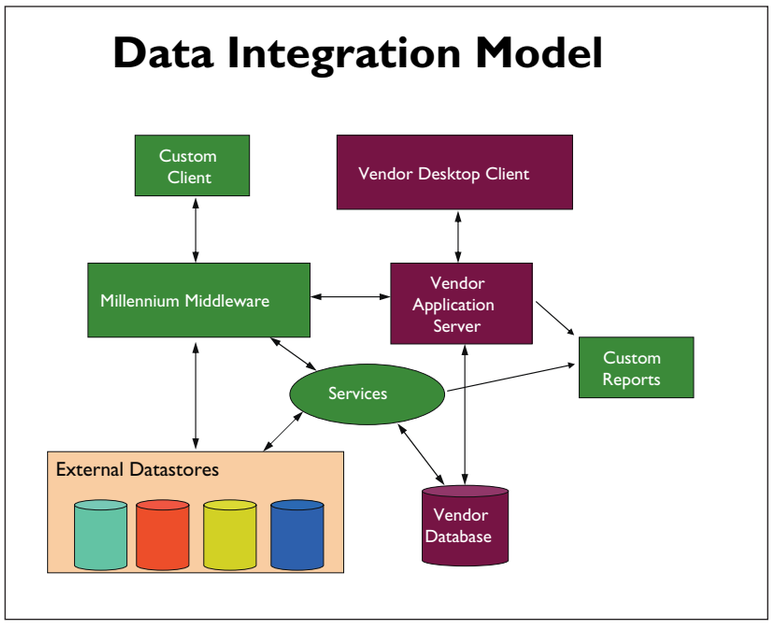


Figure 5
Example of data integration architecture to enable user access to drug discovery and development data

We have considered but have not implemented any reliance on a semantic web infrastructure. This is a new paradigm for data integration that has its roots for the biotech community in the bioinformatics world with special attention to the world of systems biology[8,9]. The major drawback to its use in a full drug discovery environment is the inherent difficulty developing and maintaining a controlled ontology. This presents significant challenges that, at least at Millennium, we have not yet solved.

Millennium Discovery Informatics roadmap

As mentioned above, the development of a data integration strategy has required a rethink as to how we should manage this environment. In the past we focused on solving the data management needs of self-contained user groups (eg combichem, high throughput screening, DNA sequencing, etc) to provide solutions specifically for the 'creator' of the data in their respective data management worlds. As the need for data integration has migrated from department functional unit to discovery project, and, eventually, to our development environment, there has been a concomitant increase in the complexity of the data management architecture needed to handle these disparate data types. We are no longer simply trying to integrate protein structure information with the structure-activity profile of a compound series. Now the problem is how to merge our animal efficacy data with our preclinical safety assessment and

References

1 Matthews, Sarah. Strategic deal-making at Millennium Pharmaceuticals. 1999. Case Study 9-800-032. Harvard Business School Publishing. Cambridge, MA.

2 Thomke, S, Nimgade, A. Millennium Pharmaceuticals, Inc. 1999. Case Study N9-600038. Harvard Business School Publishing. Cambridge, MA.

3 Challenge and opportunity on the critical path to new medical products. 2004. Food and Drug Administration Report. March.

4 Augen, Jeffrey. The evolving role of information technology in the drug discovery process. 2002. Drug Discovery Today, Vol. 7, No. 5: 315-323.

5 Davies, N, Peakman, T. 2004. Making the most of your discovery data. Drug Discovery World. Spring Issue: 17-23.

6 Koontz, Bryan. Redefining drug discovery informatics. 2005. Drug Discovery World. Spring Issue: 85-90.

7 Claus, BL, Underwood, DJ. 2002. Discovery informatics: its evolving role in drug discovery. Drug Discovery Today. Vol 7, No. 18: 957-966.

8 Wang, X, Gorlitsky, R, Almeida, JS. 2005. From XML to RDF: how semantic web technologies will change the design of 'omic' standards. Nature Biotechnology. Vol 23, Issue 9: 1099-1103.

9 Mukherjea, S. 2005. Information retrieval and knowledge discovery utilising a biomedical semantic web. Briefings in Bioinformatics. Vol 6, Issue 3: 252-262.

drug disposition parameters, or integrating in-life study parameters with bioanalytical profiles of new drug candidates.

Our focus has progressed during the past few in three stages: from a data storage and access paradigm centred on protein targets to the development of a compound-centric view integrating early discovery data to the state we have entered, today, where we are focusing on the complete project and the recognition that the most valuable data for this project tends to be the late-stage study data that deal with the compilation of numerous animal tests. Additionally we need to be able to apply traceability to the data, themselves, as they are analysed, processed, translated and transcribed into study reports. Integrating across this diverse platform has definitely presented numerous challenges. One basic strategy that we are taking is to spend considerable time analysing work processes and use cases to before we incorporate any additional changes into our Informatics environment. We have found it imperative that the user be a key participant in the design of any solution and a significant investment in time is taken to work through an iterative process of system deployment. Since we have already built out a number of substantial platforms, the most cost-effective approach has been to leverage against the present state as we decommission old, rebuild active applications where new features are truly needed, and deploy selected new systems (from completely home-grown to completely commercial 'out-of-the-box' solutions) where completely new solutions are required.

The general model that we are adopting as we migrate toward a full integration strategy relies on a service-oriented framework to connect our various disparate data stores through dedicated client interfaces that have been deployed for specific user groups. These client tools might be standard vendor products (thick client or web application) or our own internally built product for a particular process. Using our electronic chemistry notebook as a model we have developed a very facile architecture that consists of a number of layers. The key is in the development of a middleware layer that allows us to customise the client interface without impacting our vendor products (**Figure 5**). This way we can maintain a vendor neutral approach to system maintenance and focus only on those components that we have created when we undertake a system change. This also allows us to control data flow, access and viewing in a manner that is customised for the particular user or user group. This approach does not come without consequence. There is predictable overhead associated with the deployment of this architecture – the most noteworthy being the enhanced complexity of the environment impacting the testing that is needed when making changes to any of the system components, but we have found the maintenance of this architecture to be manageable with a small support staff.

The migration to this architecture has allowed us to begin planning a roadmap toward a fully integrated discovery (including preclinical development) data environment. Our approach, to date, has relied primarily on lightweight point solutions governed by the aforementioned architectural model and driven by use cases and user need associated with specific data integration requirements. This model is working for us now, will certainly evolve as the technology evolves, and gives us a framework that allows evolution either within our proprietary environment or with vendor applications without the need to completely rebuild as we migrate this forward. DDW

Dr David Sedlock *is currently the Director of Research Systems at Millennium Pharmaceuticals, Inc and is responsible for the development and management of the Informatics environment that supports drug discovery and early-stage preclinical activities. David received his PhD in Bacteriology/Biochemistry from the University of Wisconsin, Madison, USA and has been directing drug discovery functions in the pharmaceutical and biotechnology industries for more than 25 years.*