

CONNECTING THE DOTS

The Schering-Plough Research Institute has built a discovery data library as part of its knowledge management efforts.

BY TOM LAZ, KEVIN FISHER, MITCH KOSTICH, AND MARK ATKINSON

In just one decade, the World Wide Web has fundamentally changed the way people interact with computers. Users now expect near-instantaneous access to information from an interface tailored to their interests and usage habits. Developers, too, appreciate that Web-based applications can be designed faster and cheaper than client-server platforms.

But simply choosing to develop a Web-based system doesn't guarantee—in the case of scientific software development—happier, more productive scientists. Consider the Web-based systems available from leading scientific software vendors. Some systems are simply shells for delivering access to a vendor's own products. Others omit access to key information sources because of privacy or competitive concerns. The error comes with putting the technology first; vision, ultimately, dictates how well technology performs.

The Schering-Plough Research Institute (SPRI), the R&D arm of pharmaceutical company Schering-Plough, began learning the value of vision over technology in 1997 as we experimented with using HTML to help scientists ferret out the G-protein coupled receptors (GPCRs) among the sequences being characterized by the Human Genome Project. Seven years later, that system underpins SPRI's knowledge management system, the Discovery Data Library (DDL).

Clearly, technology is not the differentiator: The DDL stores information in a variety of flat files and SQL databases, such as

Sybase and Oracle; leverages established vendor standards, such as ISIS from MDL Information Systems and ActivityBase from IDBS; and connects data using standard common gateway



Searching for clues. Schering-Plough Research Institute scientists access the Discovery Data Library for managing drug R&D information.

The Schering-Plough Research Institute's Discovery Data Library helps scientists quickly navigate and view information. On the left is a form used to retrieve screening data from ActivityBase, SPRI's chosen system for managing biological data. Retrieval parameters are easy to specify: In this case, scientists are retrieving only those results with EC₅₀ values less than 5 nM, and have opted to retrieve the results to a Web page in the current browser window (a spreadsheet view is also available). The query results are shown on the right.

interface (CGI) scripts written primarily in Perl. What sets the DDL apart is its vision: that you can, in fact, get there from here.

Every piece of information in the postgenomic world is part of a simple triad: Diseases are mitigated by mechanisms of action in genes and proteins, which can be augmented with drugs. By cross-referencing information from the three parts of the triad, we ensure that our system is the first place scientists go to find out about a disease, sequence, or drug. And because it's easy to add a new source to the DDL—by just creating a new hyperlink—we've built a system that truly reflects the promise of the Web: a repository for SPRI's (and, through patent and competitive report databases, our competitors') discovery research data, accessible to scientists through interfaces that help them work smarter and faster.

INSIDE THE DDL

A common complaint among scientific software developers is that the data is unstructured. Chemical structures, chromosome

maps, 3D gel images, in vivo and high-throughput assay results, and other common discovery data are complex and uniquely individual data types. Just figuring out how best to database these results has been sufficient to tax many pharmaceutical information technology (IT) staffs and has provided myriad niche markets for discovery software vendors. Harder still is the challenge of tying and integrating the data to key contextual metadata to decide what it means and determine what to do next.

A key insight early in the DDL's development was the recognition that even unstructured data has structure. Every piece of information generated during drug discovery has a structure, a context, and a connection. More crucially, this information is all part of an interconnected investigative process, providing vital knowledge that can inform research at every stage. The results of a screening run tell us something about the efficacy of those compounds, which in turn tells us something about the mechanism of action at the target site, which in turn tells us something about the disease mitigated or caused by changes at this target site.

The right knowledge at the right time is critical to "failing early," the oxymoronic key to discovery success. By providing a way for all SPRI researchers—from genomics specialists to medicinal chemists—to traverse accumulated knowledge rapidly, the DDL helps SPRI prioritize its lines of investigation and eliminate time-consuming guesswork about what has been and needs to be done.

Although the DDL was initially envisioned as a portal to bioinformatics information, its flexible data integration platform has been easily extended to accommodate the breadth of data sources required for modern discovery. We built the DDL on an integration layer that parcels out data housed in various workflow applications to users working from a common, but customizable, interface.

In the DDL's case, a Web browser is used to query data. CGI scripts between the browser and the server process the queries, retrieve data from the appropriate application source, and format the results as a Web page. Higher-level

The Schering-Plough Research Institute's Discovery Data Library uses hotlinks to direct scientists to detailed information shown on the left, such as that on a protein screened in an assay (through the TARGET column and with hotlink results shown on the right), information on the experiment (through the EXPT ID column), and detailed reports for each compound (through the SCH NO column). The TARGET_ID condition stores an identifier for the actual construct used in the assay. Through this identifier, the DDL can track which versions of the same gene have been used in different assays to provide more precise information on target performance.

data is presented first, such as what information exists and where; scientists decide whether to drill directly to the application source for further details.

Our goal was to make the DDL the first place to go for information on a disease, gene, or drug, and this federated design strategy has enabled this vision. Scientists can get to only and all the discovery information they require using the same tool that they use to surf the Web. The DDL can handle almost any type of data that scientists want to access, without requiring custom applications to be purchased or built and subsequently maintained. This flexibility has freed our developers to focus on designing useful, appealing interfaces that show scientists data that they want and expect to see.

Taken as a whole, the federated informatics approach implemented at SPRI has benefits across the organization. The DDL does not force scientists to abandon their beloved workflow tools for the sake of integration. In fact, because of the DDL's flexibility, SPRI can confidently invest in the best workflow tools serving different areas of discovery—even tools built by different vendors or on different software platforms. And by relying on vendor know-how to keep popular workflow tools up-to-date, our discovery informatics team can focus on building value-added applications, connecting the DDL to additional data sources, and tailoring the DDL interface to make it more supportive of discovery research tasks.

THE BIOLOGICAL "PILLAR"

The benefits of federation are exemplified by our experience strengthening biological data management at SPRI. Key data domains, which we call pillars, reside within each of the components of the discovery triad. SPRI's pillars encompass chemical information, biological data, compound management (inventory information), genomics studies, proteomics studies, and document management.

Concurrent with the effort in 2000 to extend the initial version of the DDL beyond bioinformatics, SPRI conducted an extensive review of the informatics strategies deployed within each pillar. We found several weaknesses in how we were capturing and managing biological data. First, our home-grown application for registering biological data centered on a multistep data-submission

process. Data was housed in Oracle but was submitted for inclusion in the database by individual scientists using Microsoft Excel.

Second, we suffered from a bottleneck. The discovery informatics department used a sophisticated data-loader to upload data into Oracle from the scientists' Excel spreadsheets. Given that the bottle was not just narrow, but not at all big enough, the entire process of registering data was frustrating to biologists and others seeking biological data. In addition, the complexities of our biological data submission process made data difficult to track and monitor, let alone mine. Large silos of data were often sent directly from the tester to a scientist requesting results. This satisfied the immediate need, but prevented the data from being used for future data mining.


Finally, we had no mechanism for efficiently capturing *in vivo* assay data. The results of these assays are extremely valuable because they focus on compounds that have already demonstrated activity in prior screening runs. Our proprietary system lacked the flexibility to manage the multiple, variable parameters associated with *in vivo* assays. This forced each lab scientist to find ways to track data from these assays, usually by creating from scratch an Excel spreadsheet for each test run.

Having already experienced the pains of building and maintaining a proprietary system for biological data, we opted to explore the established systems available from outside vendors. We selected a team of scientists from multiple biology therapy groups at SPRI to evaluate various commercial software options. Ultimately, we chose

ActivityBase from IDBS to serve as our central biological data management system.

The selection of any vendor solution is a personal one—we mention it to illustrate the power of the DDL infrastructure. The DDL's flexibility frees us to select the right tool for the task. In our case, scientists liked the familiar feel of ActivityBase, which is built on Excel, the tool they were already using in conjunction with our proprietary system.

ActivityBase also had the advantage of being able to handle data generated by both *in vivo* and high-throughput screens. The system defines the details of a specific assay as a protocol, a user-definable template that can handle "typical" high-throughput parameters such as dose, compounds, and well location, along with other key con-



THE ERROR COMES
WITH PUTTING THE
TECHNOLOGY FIRST. VISION,
ULTIMATELY, DICTATES HOW
WELL TECHNOLOGY
PERFORMS.

Behind the scenes

The Schering-Plough Research Institute's Discovery Data Library currently links to databases and software tools specific to the three parts of our discovery triad, as well as to general information applicable across the triad, including:

- ▶ gene information: nomenclature, descriptions, taxonomic information, gene ontology (GO) classification, chromosome maps, biochemical pathway information, protein-protein interactions, protein family relations, motif identification, expression data, sequence data, and single nucleotide polymorphism (SNP) data;
- ▶ disease information: descriptions, epidemiology, molecular mechanisms, model organisms, and disease association data;
- ▶ compound/drug information: chemical structures (with the ability to search by structure and intelligently cluster data), high-throughput screening schedules, internal development efforts and projects, assay protocols, specifications, and results; and
- ▶ general information: competitive reports and activity, patent information, license opportunities, literature references, internal documents, and marketing reports.

text crucial to in vivo assays, such as information on test organisms or the drug formulation used. ActivityBase not only captures all of this important “metadata,” but also automatically calculates results and fits curves so that biologists can focus on validating and interpreting the data. Most importantly, this flexible data model applies to all of our biological information, enhancing activities within the pillar and expediting the integration of this data with the rest of our discovery information.

Because ActivityBase is well equipped for tying together compound and screening data, we found it relatively straightforward to connect ActivityBase to the DDL’s triad of discovery data. The Object ID column in ActivityBase, which identifies registered compounds or objects, connects ActivityBase instances to all other DDL information—through it, scientists find out whether there is biological data associated with a DDL object and can, through hyperlinks, drill down to the specific results from ActivityBase.

Linking protein target information to high- or low-throughput screening results was more complicated, particularly because many of our assays require proteins to be modified from their wild-type form. To ensure that all the variants were tracked separately, we opted to preregister the protein and nucleotide sequences used in assays and manually associate them with the wild-type sequences in the DDL. We then aligned these various “protein units” to a single “target ID”

to which all ActivityBase data is tied. Through this mechanism, scientists can retrieve assay results and drill into specifics on the proteins used in those assays, or, conversely, they can start with the protein and drill into the stored parameter information.

IMPACTS OF INTEGRATION

As the DDL has been implemented across discovery activities, it has changed the way scientists work. Workflows have become more streamlined, particularly in biology, where many of the time-consuming steps associated with defining assay parameters, running curve-fitting calculators, and uploading results are handled automatically. Decision-making, as well, is more informed.

The DDL remains a work in progress as we tie in additional data sources and work to implement and integrate the DDL with an electronic laboratory notebook system. Ultimately, though, our experience proves that when it comes to scientific IT, technical innovation follows insight. It’s the vision behind the DDL—supported by the right technology—that has made the system such a valuable tool for scientists looking to connect the dots at SPRI.

Tom Laz, senior principal scientist, **Kevin Fisher**, manager, chemical and biological data, **Mitch Kostich**, associate principal scientist, and **Mark Atkinson**, section leader, are on the staff of the Schering-Plough Research Institute. ■

ChemOvation 

ChemOvation - the perfect partner for successful drug discovery.

INNOVATION ■ DISCOVERY

ChemOvation is a leading provider of drug discovery services to the pharmaceutical and biotechnology sector. We provide a fully integrated service, encompassing drug design, synthesis, screening and early toxicology.

To discuss how we can help you, please contact our Business Development team on +44 (0) 1403 248844.

ChemOvation 

CHEMOVATION PROVIDES DRUG DISCOVERY THROUGH:

- › MEDICINAL CHEMISTRY
- › ORGANIC CHEMISTRY
- › COMPUTATIONAL CHEMISTRY
- › SCREENING
- › ADME & EARLY TOXICOLOGY

INNOVATION ■ DISCOVERY

ChemOvation Limited
Foundry Lane
Horsham
RH13 5PX
UK

T: +44 (0) 1403 248844
F: +44 (0) 1403 248855
E: enquiries@chemovation.com

W: www.chemovation.com

Request more at AdInfoNow.org

