A model-driven approach to industrializing discovery processes in pharmaceutical research

K. Bhattacharya
R. Guttman
K. Lyman
F. F. Heath III
S. Kumaran
P. Nandi
F. Wu
P. Athma
C. Freiberg
L. Johannsen
A. Staudt

Despite year-to-year increases in R & D budgets, the number of successful NCEs (new chemical entities) has continued to decline. Drug companies are looking into new ways to make research processes more efficient, to manage information better, and to improve collaboration among research groups. This shift from an artisan approach to an organized, streamlined discovery process is often termed the "industrialization of discovery processes." This paper presents an approach to industrializing drug discovery that involves the formal modeling of research processes at several layers of abstraction, mappings between adjacent layers, and an implementation of this hierarchy using information technology-level execution elements. This approach is applied to the assay development phase of the drug discovery process. First, a business operations model is built by identifying the business artifacts, developing models for the life cycles of these artifacts, and then creating a comprehensive model that combines these life-cycle models and their interactions. Using the concept of adaptive business objects, a solution composition model that expands on the business operations model is developed. This model is then mapped into an executable platform-specific implementation by using the IBM WebSphere® platform. Our prototype system was built and validated as a joint effort between IBM Research and Bayer HealthCare.

The main business of the pharmaceutical industry is to provide drugs that save and extend lives, cure diseases, and alleviate the burden of sickness or age. High profits are generated mainly with patent-protected drugs during the time of patent protection. The pharmaceutical industry relies heavily on new drugs to generate the revenue needed to drive the expensive process of drug development. Drug development is research-intensive and thus pharmaceutical companies return more of their profits to

research and development than any other industry. The average cost of bringing a drug to market has risen from around US$54 million in 1976 to an estimated US$802 million in 2001. Moreover, this cost is expected to rise to US$900 million in 2003, an almost 17-fold increase since 1976.[1]

Drug discovery researchers strive to produce high-quality new chemical entities (NCEs) that could potentially be turned into effective drugs. Despite year-to-year increases in R & D budgets, the number of successful NCEs has declined significantly in recent years. Facing this challenge, research organizations are looking into new ways of managing processes and information and enhancing collaboration among teams. This shift from an artisan approach to an organized, streamlined discovery process is often termed the "industrialization" of drug discovery.[2,3]

Although increased costs can be attributed mostly to the clinical phase of development, the entire process of R & D has to be streamlined in order to identify the most promising drug candidates earlier in the drug discovery process, which should lead to reduced development times and improved success rates. A streamlined approach to drug discovery should also address several problems identified by industry analysts, such as low productivity in the laboratory, patent expiries, and intense therapeutic competition that have made it increasingly difficult for pharmaceutical companies to produce new drugs.[3,4]

Over the past decade, information technology (IT) has been used in various aspects of the drug discovery process, such as data evaluation, determination of 3D molecular structure, and simulation of biological systems. The successful industrialization of drug discovery, however, needs to go beyond the deployment of specialized applications. Drug discovery research of the future will support "horizontal" integration of various research teams to provide information sharing across "vertical silos of expertise," thus enabling effective collaboration of dispersed research teams and improved decision making.

Working with its partners, IBM is developing a set of IT capabilities to facilitate the industrialization of a wide range of business processes. In particular, IBM Research has developed a model-driven approach to support the collaboration among various teams and the integration of people, information, applications, and systems involved in business processes. Teams from IBM Research and Bayer HealthCare Research have collaborated in the work presented here, which demonstrates the value of the model-driven approach to drug discovery.

Drug discovery process

The goal of drug discovery research is finding and producing chemical compounds that can be used to fight diseases. The viability of chemical compounds as disease-fighting drugs is determined by a set of complex characteristics including potency, selectivity (i.e., effectiveness against the biological target and harmlessness to other tissue), toxicity, solubility, metabolic rate and kinetics within a test organism, ease of chemical synthesis, and patentability. Drug research typically takes between three and five years to determine a suitable compound for a given disease.

This process starts with identifying and isolating the biological target—the biological structure associated with a specific disease. Such targets are mainly enzymes or receptors but may also be regulatory processes within an organism. The target needs to be prepared in such a way that a very large number of compounds can be tested for potency against (or ability to bind with) this target in order to identify a few compounds that best inhibit or neutralize the malignant biological behavior of this target.

Chemical structures selected during a highly automated procedure called high throughput screening (HTS) are further subjected to a battery of molecular tests aimed at determining the value of further investment into these compounds. The parameters studied include chemical characteristics (ease of synthesis, solubility, reactivity, etc.) and biological characteristics (selectivity, activity in living organisms, toxicity, etc.). When a compound that satisfies all the major requirements is found, a chemical synthesis program to produce similar compounds with improved characteristics, such as amenability to clinical testing, might follow.

Consecutive phases in the drug discovery process become increasingly time consuming and costly due to the complexity of the tasks involved, the resources consumed, and the personnel required. Therefore, it is desirable to streamline the process in

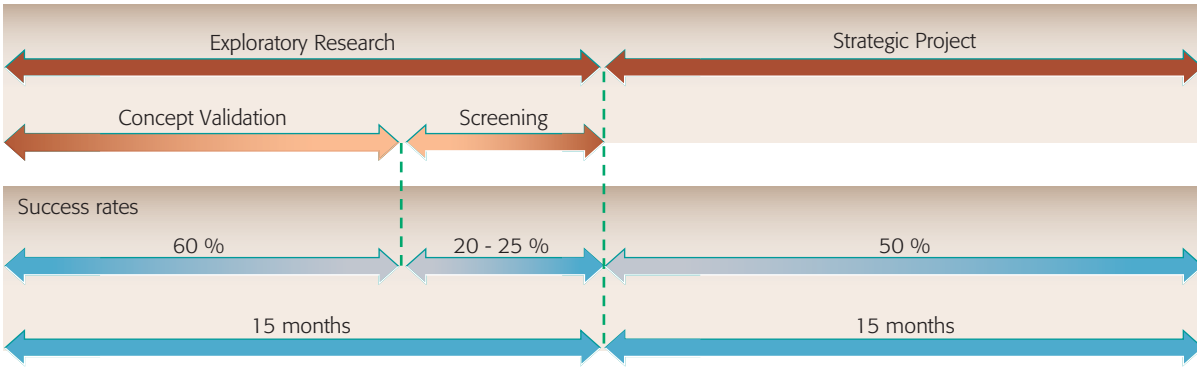

Figure 1
Overview of the drug discovery process

such a way that compounds can be identified early in the process.

In most pharmaceutical organizations the drug discovery process contains the following steps:[3]

1. *Target identification*—A suitable biological target associated with a disease is identified.
2. *Target validation*—The target is validated by conducting several preliminary experiments.
3. *Assay development*—The components to be analyzed are identified; a protocol to sustain the screening of large numbers of compounds against a biological target is developed.
4. *HTS*—Compounds are screened for potency against the biological target.
5. *Hit identification*—A reduced set of chemical compounds, typically of the order of one million, are identified in the HTS phase as inhibitors of the target's biological function.
6. *Lead candidate selection*—The selected set of compounds is further reduced by focusing on the most promising compounds (typical size of this set is 1000 compounds).
7. *Lead optimization*—The chemical structure and biological effects of the selected set are further used to reduce the number of compounds (typical size of the set is 10 compounds).
8. *Preclinical testing*—Selected compounds are pretested before the large scale clinical trials (e.g., testing on animals).

An overview of the drug discovery process is shown in ***Figure 1***. The success rates and time-line values shown are for illustration purposes, and they vary with the environment. The drug discovery process can be divided into two parts: the *exploratory research* part that consists of all the initial phases up to and including lead candidate selection, and the *strategic research* part (we also refer to it as the *strategic project*). Exploratory research starts with an idea for a biological target and ends with lead candidate selection over the course of about 15 months. The strategic project, whose goal is to determine a single candidate compound, is launched only if the exploratory research phase produces a set of promising results. It typically stretches over a 15-month period. Exploratory research can be further divided into the *concept validation* phase, which includes the initial phases of the process up to but not including the HTS phase, and the *screening* phase. Figure 1 shows typical success rates for drug discovery projects. About 60 percent of projects survive the concept validation phase. Of the remaining projects, about 20 to 25 percent survive the screening phase. Then, about 50 percent result in a candidate compound for further processing. In the long run, only 1 percent of undertaken projects lead to a satisfactory return on investment (not shown in Figure 1).[3]

As described earlier, the pharmaceutical industry currently faces the challenge of streamlining the drug discovery process. Although cost reduction is one aspect of this goal, the main goal is to produce promising lead compounds as early in the process as possible. We now describe some of the difficulties that the pharmaceutical industry has to cope with in order to achieve this goal.

First, drug discovery usually involves multiple teams that are dispersed throughout the world. For

example, an HTS unit takes over the biological target assays from therapeutic area teams and screens a large number of chemical structures supplied by chemistry units. The smooth operation of the process demands efficient collaboration among all parties concerned. This collaboration includes the coordination of resources and the exchange of information.

Second, every large pharmaceutical research organization has a heterogeneous technology infrastructure. In many instances companies have invested heavily in cutting-edge technology without due consideration for its overall impact on the entire discovery process. Research units often are equipped with platforms that are selected based on local considerations and may not be designed to interoperate with other teams' equipment.

Third, pharmaceutical research produces a tremendous amount of data, sometimes too vast to analyze in detail.[3,5] Often data are used by different teams in ways that impede collaboration and the smooth flow of information. For example, the heterogeneous nature of the technology infrastructure leads to information produced in a variety of data formats. Also, data are typically managed locally within research units and thus not easily accessible by other teams.

Although other industries are facing similar situations, the problems are more severe in the pharmaceutical industry because research is at the core of its business. The main asset of pharmaceutical companies is the scientific ingenuity and the creativity of their research staff. The drug discovery process must be streamlined without inhibiting the creativity of researchers. The challenge is to improve the efficiency of existing processes by integrating people, information, IT systems, and various apparatuses to facilitate collaboration among research units. We now discuss our approach to accomplishing this task.

Our approach

In a recent paper one of the authors laid out a conceptual framework for the so-called "model-driven enterprise" that aims at bridging the gap between business goals and the IT systems that support the business processes of the enterprise.[6] The framework consists of a hierarchy of four abstraction layers, each providing a different view (model) of the behavior of the enterprise.

1. *Strategy layer*—The strategy model specifies what the business is intended to achieve. It models the business objectives in terms that executives and business strategists understand. For example, it might specify the objectives in terms of the well-known balanced scorecard perspective.[7]
2. *Operations layer*—The operations model describes what the business is doing to achieve its strategic objectives and how it will measure its progress toward them. It is typically developed by business analysts in conjunction with line-of-business managers. Because the model captures the business operations, commitments, and operational key performance indicators (KPIs) in terms accessible to business users, we refer to it hereafter as the *business operations model*.
3. *Solution composition layer*—The solution composition model describes the processes and information flows that the business uses to implement the operations model. It is platform-independent and allows iterative performance improvement while ensuring consistency with the business objectives. A transformation (mapping) tool is used to create the core elements of the solution composition model from the operations model, which is then manually refined to complete its definition.
4. *Implementation layer*—The implementation model is a platform-specific realization of the solution composition model. Tools are used today to construct portions of the implementation model directly from the solution composition model, much as a compiler translates a high-level language into a machine language. The model links to applications and specifies how to measure the parameters needed to determine the KPIs.

This model-driven approach is now applied to analyzing and capturing pharmaceutical research operations. We model specific research operations at the operations level using the WebSphere* Business Integration (WBI) Modeler.[8] We apply a heuristic transformation algorithm that maps the operations model into an solution composition model. The key component used to create the solution composition model is the *adaptive business object* (ABO)—the solution composition layer is a composition of communicating ABOs.[9]

UML2 is used as the language for defining the models at the various levels of the hierarchy.[10] UML** is the standard modeling language maintained by the Object Management Group** (OMG**). We use the UML2 Profile mechanism to tailor the UML2 metamodel for modeling the enterprise. UML supports metamodels at two levels. Meta Object Facility (MOF**) is a language for defining the models. UML is defined by using MOF as the metamodel. User-defined models may use UML as the metamodel. A profile is used in UML to extend a reference metamodel or another profile. The reference metamodel extended by the profile may be any MOF-based metamodel, including UML. The multilayer framework models consist of sets of such profiles.

User-Centered Design (UCD) methods support and complement the model-driven approach by defining user interactions with the solution, which are otherwise not captured explicitly in the strategy, operations, implementation, or solution composition models.[11] We are using a UCD approach for designing, developing, and implementing the user interfaces.

The paper is organized as follows. In the section "Business operations model," the business operations modeling technique is illustrated by using the assay development process. Then, in the section "Solution Composition Model," the ABO concept is introduced, and the mapping of the operations model into a solution composition model based on communicating ABOs is demonstrated. In the section "System Design," the details of the implementation of our prototype are described. A discussion section compares our approach with related work and is followed by a conclusion.

BUSINESS OPERATIONS MODEL

We omit the formal modeling of the strategic layer and start directly with the business operations layer. The actors (or role players) in these operations are biologists and chemists known as "lab heads," who design and lead the experiments, and lab technicians, who conduct the experiments.

As explained in the previous section, the business operations model describes the way the business plans to achieve its strategic goals in terms amenable to the business user—lab heads and lab technicians in this case. Typically, these business users are not IT people, and they should not be overwhelmed with details that relate to the IT implementation. At the same time, the business operations model is used to generate an IT-centric model, and thus, the business operations model needs to be descriptive enough to allow for its mapping into a formal representation of an IT-centric model, the solution composition model.

Assay development process

The focus of our business operations model is the assay development process, a central aspect of the early phase of drug discovery. The goal of assay development is the preparation of assays and the design of a protocol and instructions for the HTS phase. Assay development is a highly research-driven and collaborative process. Assay development typically takes from two to eight months and is driven by scientific insight into the biological target under consideration.

HTS has frequently been portrayed as the frontline technology within pharmaceutical discovery, and over the past decade the industry has witnessed an apparently astronomical increase in the capabilities of various HTS groups.[4,5,12,13] Running an HTS experiment requires very precise and detailed instructions on the preparation of the biological target (for example, specifying the incubation time or the concentration of solution). The protocol is an exact description of the reagents, parameters, and workflow required to screen the in-house compound library against the biological target in an HTS apparatus. The team of biologists within an assay development project strives to create an optimal HTS protocol. Optimal HTS protocols are designed for maximum signal strength in the HTS apparatus in order to obtain unambiguous results.

Identifying business artifacts

Our approach to the modeling of research processes is based on identifying the relevant business artifacts through interviewing scientists and technicians responsible for these processes. Any business relies on business documents and other "artifacts" that record concrete information pertinent to the business. For example, when bank customers withdraw money from their accounts, they fill out a withdrawal slip and pass it to the teller. The teller uses the slip to check the details of the customer account and, depending on the amount, requests a manager's approval. The manager approves by

Enzyme A Test (EnzA)

A: 1.2 mg/L Enzyme, 250 mg/L BSA, 0.5 mM ATP in C
B: 40 mg/L 23S+16S rRNA in C
C: 50 mM Tris/HCl, 5 mM $MgCl_2$, 1 mM DTT, pH 7.5 in 0.1% DEPC in H2O, autoklaviert
D: 0.04% Malachite green, 5% $(NH_4)2MoO_4$, 3.7% HCl, 10 mM EDTA in H_2O, autoklaviert

Reaction: Test compound in 2 µL DMSO + 25 µL A + 50 µL B > 60 min @ 30°C > + 100 µL D (30 min @ RT) > A @ 630nm

Experiment No.:	990729
ExpDate:	99-07-29
Execution:	Susan Tech
Sample	
MTP Nr.:	Plate name
Description:	
Concentration:	10 mmol
2 µl in 77 µl reaktion volume:	
260 µmol	1:2 serial dilution

Compounds / dilution

	Controls		1:1	1:2	1:4	1:8	1:16	1:32	1:64	1:128	1:256	1:512
	1	2	3	4	5	6	7	8	9	10	11	12
A	0%	100%	BAYA001234									
B	0%	100%	BAYB004321									
C	0%	100%	BAY 123456									
D	0%	100%	BAY 123457									
E	0%	100%	BAY 123458									
F	0%	100%										
G	0%	100%										
H	0%	100%										

Absorption @ 630 nm

	1	2	3	4	5	6	7	8	9	10	11	12
A	0.134	0.572	0.316	0.302	0.240	0.405	0.312	0.323	0.440	0.512	0.514	0.599
B	0.130	0.580	0.254	0.288	0.328	0.388	0.430	0.432	0.562	0.573	0.603	0.612
C	0.131	0.568	0.230	0.176	0.166	0.213	0.253	0.331	0.487	0.578	0.620	0.626
D	0.132	0.579	0.223	0.177	0.160	0.169	0.261	0.477	0.554	0.591	0.583	0.585
E	0.138	0.556	0.233	0.179	0.162	0.221	0.363	0.531	0.620	0.623	0.611	0.603
F												
G												
H												

AVG	0.133	0.57
SD	0.003	0.01

Evaluation

Concentration [µmol]

	% Activity	IC50	260	130	65	32.5	16.3	8.1	4.1	2.0	1.0	0.5
BAYA001234		7.1	42%	39%	24%	62%	41%	43%	70%	87%	87%	106%
BAYB004321		52.0	28%	35%	45%	58%	68%	68%	98%	100%	107%	109%
BAY 123456		7.6	22%	10%	8%	18%	27%	45%	81%	102%	111%	113%
BAY 123457		12.8	21%	10%	6%	8%	29%	79%	96%	105%	103%	103%
BAY 123458		17.5	23%	11%	7%	20%	53%	91%	111%	112%	109%	107%

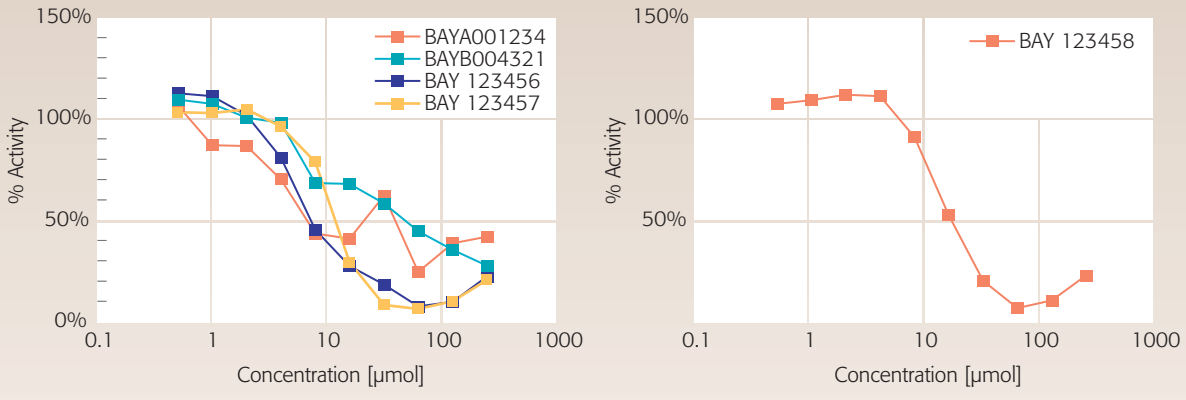

Figure 2
An experiment record

signing the slip and returning it to the teller, who hands out the requested money to the customer and then files the slip for the bank records. The key business artifact in this example is the withdrawal slip, which captures all important data pertinent to the withdrawal process.

A business artifact can be characterized as follows:

1. A business artifact has a unique identity and therefore cannot be broken up and reassembled.
2. A business artifact is self-describing and contains all information pertinent to the business context.
3. A business artifact is a record meaningful to the business user.

We conducted several workshops with the lab personnel in which we identified the business artifacts for the assay development process and how the business artifacts are used by asking the participants two questions: (1) what do you produce? and (2) how do you produce it?

As a result we identified three distinct business artifacts for the assay development process:

- *Candidate HTS protocol record*—This artifact represents the assay development team's main product. The candidate HTS protocol record consists of a set of instructions that describes the steps involved in running an HTS experiment.
- *Experiment record*—An experiment record consists of a protocol section for capturing instructions for experiments, a results section for capturing results from experiments, and a notes section for capturing notes, comments, and observations during the experiment (see ***Figure 2***).
- *HTS protocol record*—The HTS protocol is used by the HTS labs to conduct HTS pre-runs.

The experiment record illustrated in Figure 2 is a spreadsheet designed by the lab head. It is divided into horizontal sections and consists of three distinct parts. The first section at the top is the *protocol section*, which contains instructions for conducting the experiment (steps A, B, C, and D are shown). The second and third sections make up the result part; the data contained are entered and updated by lab technicians who conduct the experiment. The bottom two sections make up the evaluation part; the data are typically entered by the lab head and a lab technician.

Modeling centered on artifacts

The business approach centered on artifacts that is used in this project is based on the *operational specification* as described by Nigam and Caswell (our terminology is somewhat different).[14] In the operational specification approach, the modeling of business processes is done by tracking the life cycle of key business artifacts. There are two modeling primitives used: the task and the repository.

In a business environment, business artifacts are passed from agent to agent (which may be a person, an application, or some other processing system or apparatus). An atomic business transaction that modifies a single artifact or a collection of artifacts is known as a *task*.

The *repository* represents a place to store a specific artifact type. Repositories are modeled after physical, electronic, or logical objects, such as a filing cabinet, a hard drive, or a database.

Whereas the task is an active entity, the repository is a passive one. After receiving an artifact and processing it, the task either releases the artifact and makes it available to another task or stores it in a repository. The repository does not actively initiate the release of a stored artifact, but only responds to requests to retrieve the artifact.

A task owns input and output ports. Each input and output port is associated with an interface that supports one specific type of business artifact; that is, two business artifacts of different types cannot enter or leave a task through the same port. Conditions on the ports control whether an artifact passes through the port.

Developing the business operations model

Figure 3 shows the life cycle of a candidate HTS protocol record. The life cycle includes the following tasks and processes:

- *Design experiment*—The lab head initiates the process illustrated in Figure 3 by creating the initial draft for the candidate protocol. The artifact is stored in a repository.
- *Evaluate protocol*—The protocol is evaluated through a series of experiments performed by lab technicians. The lab head determines whether the candidate HTS protocol is viable.

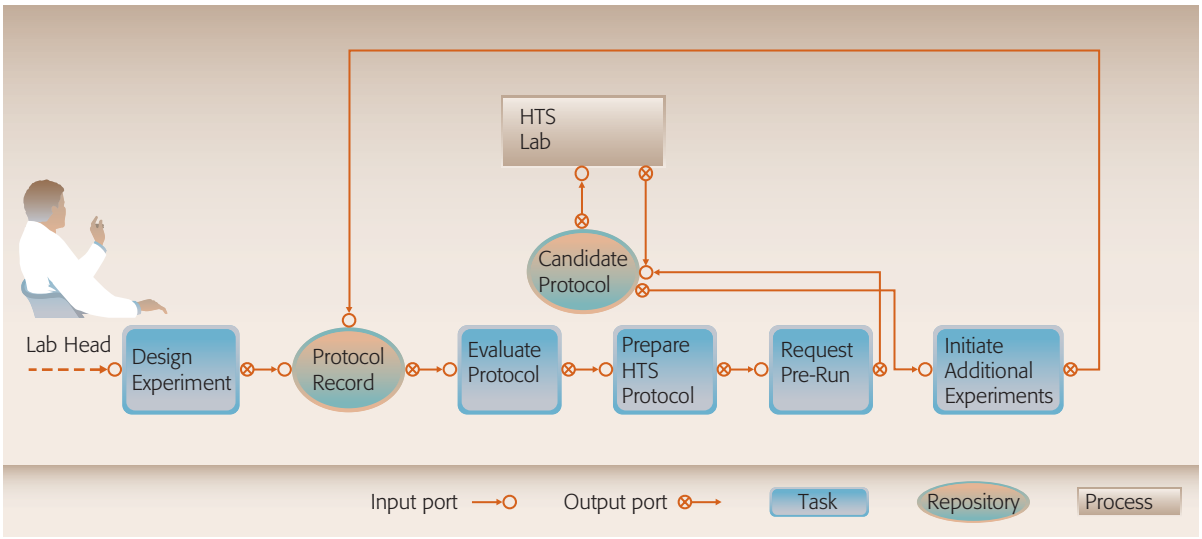

Figure 3
Life cycle of a candidate HTS protocol record

- *Prepare HTS protocol*—The lab head summarizes the experimental results and finalizes the candidate HTS protocol.
- *Request pre-run*—The lab head is now ready to share the candidate protocol with the HTS lab and makes it available to the HTS lab through the repository.
- *HTS lab*—The HTS lab retrieves the candidate HTS protocol for review and may return the protocol with comments and suggestions if more validation is required. Note that this step is modeled as a process instead of a task.
- *Initiate additional experiments*—If the HTS lab asks for further validation, the lab head updates the candidate HTS protocol and queues it up for additional experiments by lab technicians.

Figure 4 illustrates the life cycle of an experiment record.

- *Design experiment*—The lab head creates an experiment record and assigns a technician to conduct experiments. The experiment record is made available by storing it in a repository.
- *Perform experiment*—The lab technician performs an experiment and retrieves and updates the experiment record. Note that this step is modeled as a process instead of a task. Perform Experiment could be just another long-running business operation to be conducted in the context of assay development (e.g., a biochemical experiment or a screening experiment).
- *Analyze results*—The lab technician and the lab head analyze results to determine the next steps. There are three possible outcomes. First, the experiment protocol needs to be updated. Second, the experiment needs to be rerun (e.g., control experiments). Third, the experiment needs to be rerun in a different way.
- *Update protocol*—There are two potential outcomes of the Update Protocol task. First, the experiment is completed and the record placed in a repository. Second, the protocol needs to be updated and the experiment repeated.
- *Clone experiment*—The same experiment needs to be rerun (e.g., for control purposes), and the experiment record is updated. The updated record is stored in the repository.
- *Modify experiment*—The experiment record is updated to capture suggested changes in the experiment.

Finally, as shown in ***Figure 5***, when the HTS lab is satisfied with the candidate HTS protocol supplied by the assay development team, it creates its own business artifact, the HTS protocol record.

Figure 5 shows the business operations model for the assay development process. The model combines all artifact life cycles and includes some additional interactions between the life-cycle enti-

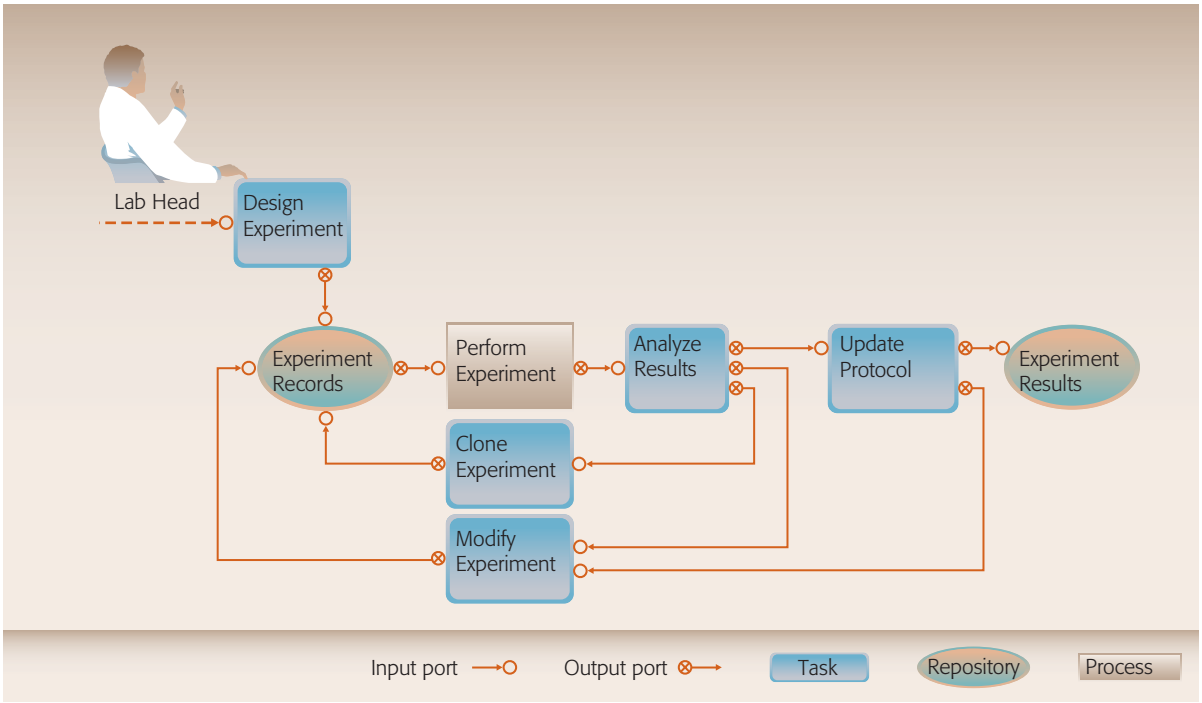

Figure 4
Life cycle of an experiment record

ties, as follows. First, the Design Experiment task has the lab head create both the candidate HTS protocol and the experiment record. Second, the Update Protocol task accesses the candidate HTS protocol to update it based on the experimental results and places it back in the repository. Third, the steps Evaluate Protocol, Prepare HTS Protocol, and Request Pre-Run retrieve the experiment records from the appropriate repository.

In business operations models, a task can encompass a number of activities whose details are not visible at this level. These activities can include automated processing, manual workflow, and unstructured collaboration, such as instant messaging, phone calls, or face-to-face meetings. There can also be explicit communication such as e-mails, information exchange in the form of "team rooms" and common databases, and shared access to documents. Significant results from the activities that take place within a task are recorded on the business artifacts before the task is completed.

The modeling approach allows us to formally describe the work researchers do. Business operations modeling is often used as a way to understand how to best streamline work processes. Although the business operations model shows how the assay development process can be streamlined by modeling the information exchange among scientists and technicians, it does not model the creative side of the research, which takes place within a task.

The business operations modeling methodology is different from standard process modeling or workflow modeling. First, we aim at a representation that is understood by business users. This requires the use of concepts that resonate with business people. Second, on a more formal note, the semantics underlying the artifact-centric approach is driven by the uniqueness property of the artifact. The business artifact can only be at one place at a time. This allows for modeling the life cycle of individual business artifacts independent of each other. The candidate-HTS-protocol artifact can be modeled independent of the experiment-record artifact. The model for the overall process is a composition of the individual life cycles. The compositional semantics is depicted by the interaction points of both artifacts in a task. For example, two artifacts are created by the Design Experiment task, or the candidate-HTS-

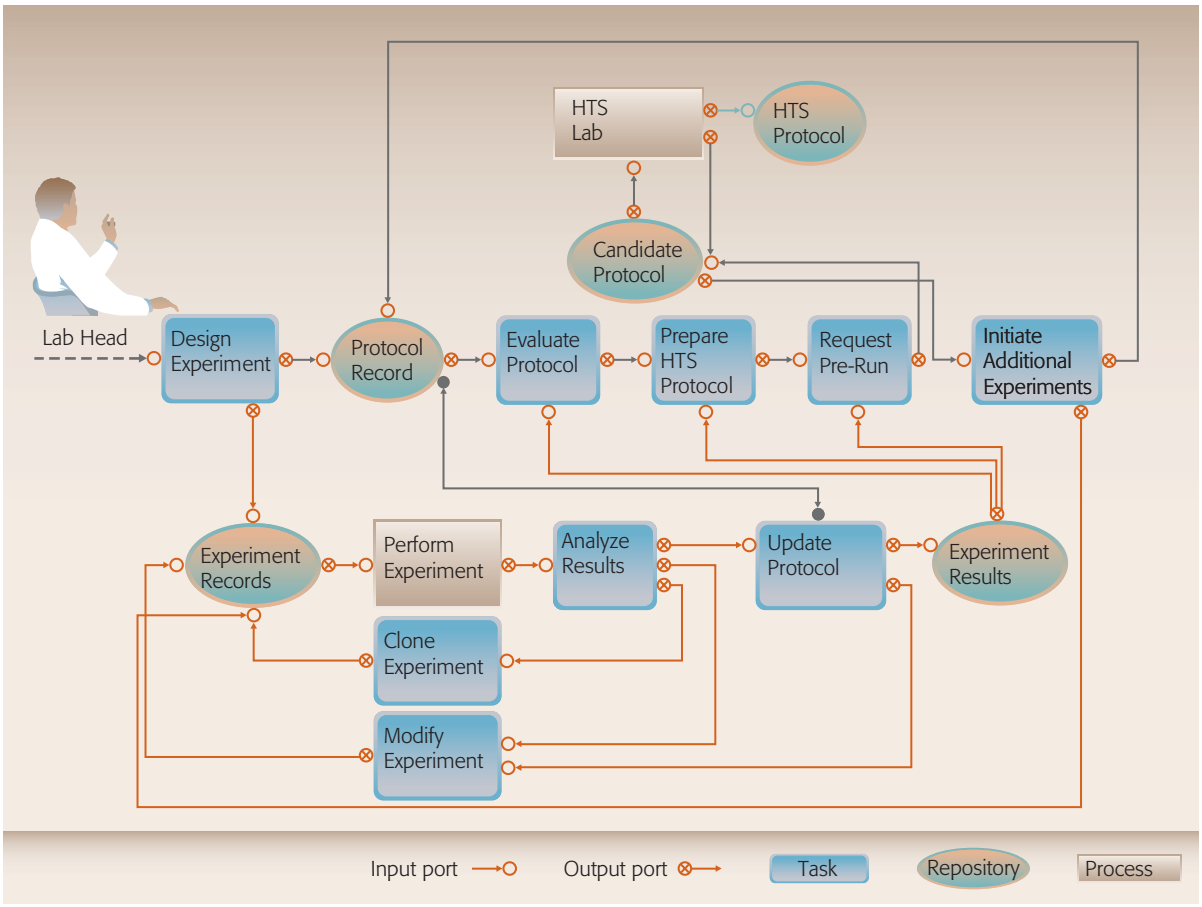

Figure 5
Business operations model for the assay development process

protocol artifact gets updated in the Update Protocol task.

We point out, however, that the semantics of the interaction is not formally defined in the business operations model. For example, the sequence of artifact creation in the Design Experiment task is not modeled explicitly. The synchronization step between both artifacts in the Update Protocol task is also not specified (synchronization is needed if, for example, the candidate-HTS-protocol record is not available for updating).

This level of detail, however, is omitted for a good reason. In our case the detailed information about the sequence of artifact creation or the synchronization mechanism is not important to the business user. This level of abstraction entails ignoring some of the lower-level details, which may not impact the business itself but only the implementation of the system. The main point is that the model is still expressive enough to allow for mapping into well-defined IT solution artifacts. We elaborate on this transformation and the solution composition model in the next section.

SOLUTION COMPOSITION MODEL

The business operations model as described in the previous section captures the business artifacts and models their processing. The solution composition model is a platform-independent IT-level description of the business artifacts and their interactions. Our approach to representing the business artifacts at the solution composition level is based on the ABO component model.[9]

The ABO is a *basic unit of composition* in one component model that consists of three major aspects: (1) the life cycle of the artifact, (2) the information associated with the artifact, and (3) the

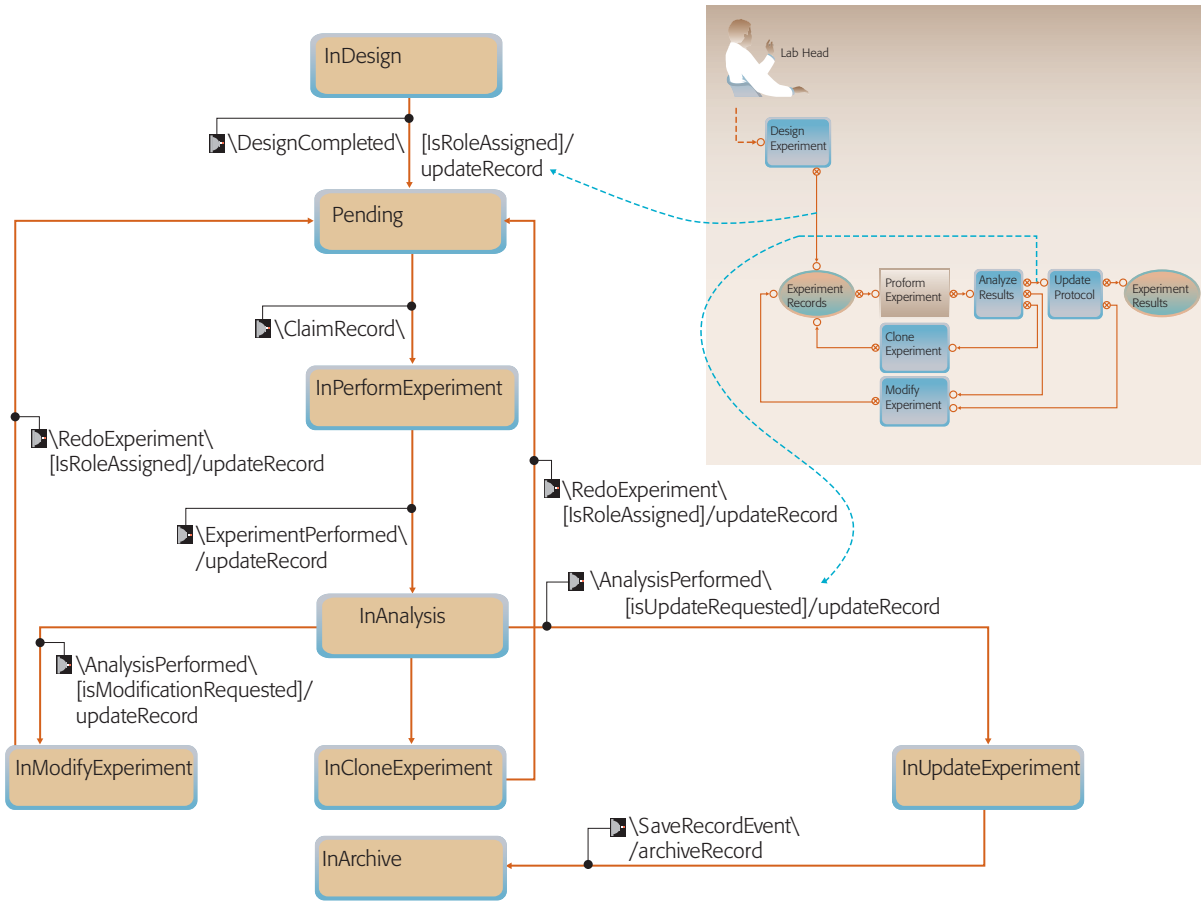

Figure 6
Mapping of the experiment record life cycle to the finite state machine

interaction of the artifact with external agents (such as applications or other artifacts). A solution composition model is thereby a composite of interacting ABOs. The exchange of information between the ABOs can be modeled as a communication protocol with managed state similar to Conversation Policies.[15]

The mapping of the model for the assay development process illustrated in Figure 5 to a solution composition model results in the creation of three ABOs. The three business artifacts (the candidate HTS protocol, the HTS protocol, and the experiment record) are mapped into three corresponding ABOs.

The life-cycle component of an ABO uses finite state machines (FSM) as a computational model. The life cycle of the business artifact at the business operations layer is thereby mapped into a FSM representation at the solution composition level.

Figure 6 illustrates the mapping of the experiment-record life cycle to the FSM of the corresponding experiment-record ABO. State transitions within the FSM (on the left in Figure 6) are labeled as follows: events are enclosed in backward slashes (\Event\); conditions are enclosed in brackets ([Condition]); and actions are preceded by a forward slash (/Action). For example, the completion of the Design Experiment task triggers the event `\DesignCompleted\` associated with the transition between states `InDesign` and `Pending` (note the dotted line that marks this mapping). The state `Pending` corresponds to the repository that holds the experiment record ready to be accessed by the Perform Experiment task. The transition from the Analyze Results task to the Update Protocol task is also mapped to an FSM transition—from the state `InAnalysis` to the state `InUpdateExperiment` (note the second dotted line that marks this mapping). The transition between the two states is

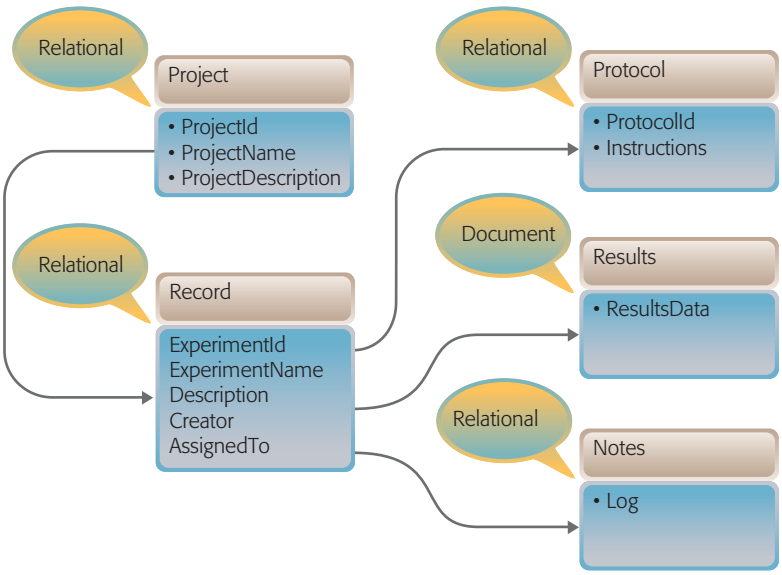

Figure 7
Data graph for experiment record ABO

labeled with the event `\AnalysisPerformed\`, the condition `[IsUpdateRequested]`, and the action `/updateRecord`.

The information managed by the experiment record ABO is distributed among several data stores. In the data graph of the ABO, shown in ***Figure 7***, the nodes of the graph, which correspond to the various data components, are annotated with the type of data store (the yellow balloons next to the nodes). As shown in the figure, the Results are stored in a document store, whereas all the other data components (Project, Record, Protocol, and Notes) are stored in relational stores.

To summarize, the solution composition model represents a different viewpoint of the business operations model. The solution composition model is described by a set of communicating ABOs. Each ABO can be generated by a transformation of the business artifact from the business operations model. The life cycle of the business artifact is represented as an FSM; the data contained in the business artifact is represented as a data graph. Notice that not all the information needed in a solution composition model can be retrieved from the business operations model. For example, the business operations model addresses the concerns of business users, who are typically not interested in the exact location or type of information. The technical aspects of the data graph need to be added by IT personnel, who are the stakeholders responsible for the design of the solution.

SYSTEM DESIGN

In this section we describe the system design. We start with the solution composition model described in the previous section and transform it into an executable platform-specific implementation.

UCD

In order to ensure a match between user needs and the functions provided by our prototype, we used UCD methods throughout the design.[11] Specifically, we conducted a user-needs analysis, created a "storyboard" illustrated with user interface designs, and incorporated user feedback throughout the design process.

Initially, we conducted several workshops with pharmaceutical researchers and technicians in order to observe and learn about their work processes. These user studies allowed us to produce profiles for the different role players (lab head, technician, and HTS lab head) involved in the assay development process. We conducted a half-day work session to gather solution requirements from the users; these were used as the basis for our functional requirements. The functional requirements together with the process model were used to create a textual

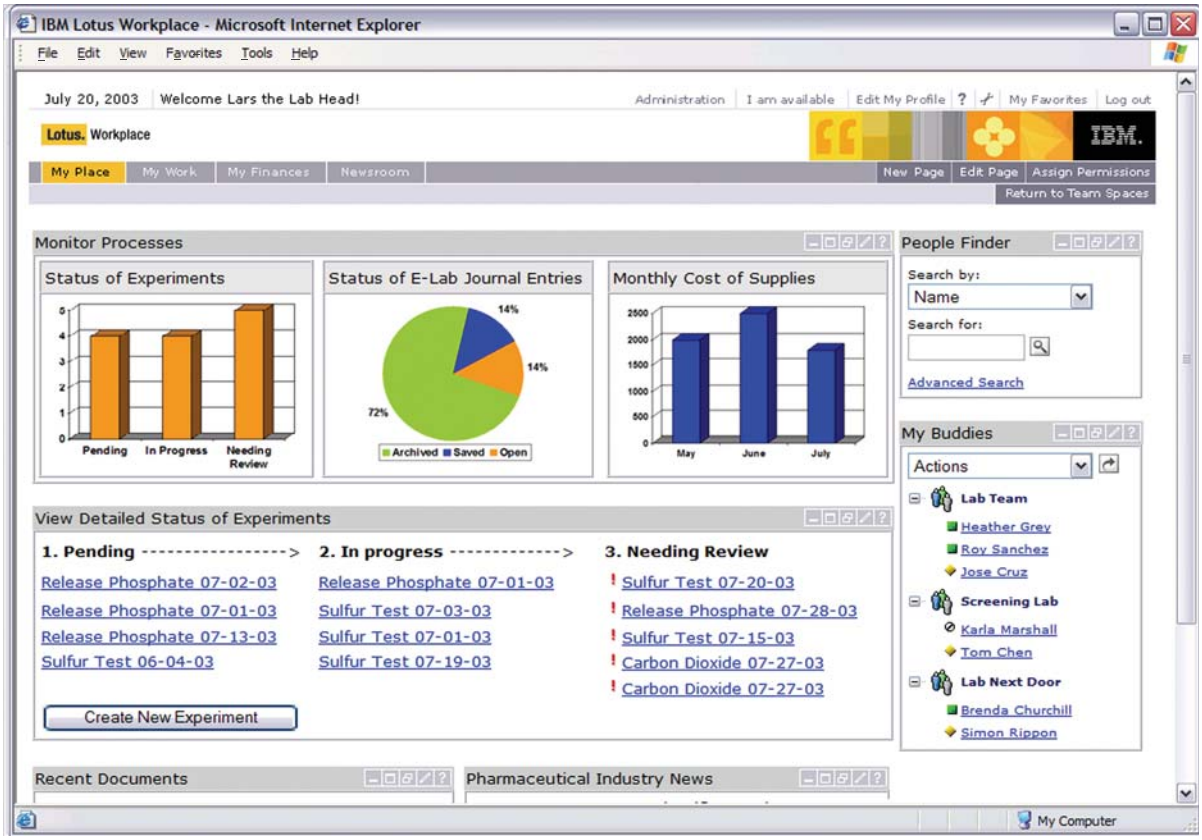

Figure 8
Workplace entry page for the lab head

narrative that describes the way researchers would use the anticipated solution. This narrative was illustrated with rough user interface designs, and then refined to include high-quality user interface designs. The end product of the UCD process was a storyboard, complete with user interface designs that explained the use of the solution by different role players. The user-validated interface designs were then implemented. All these designs were validated during joint meetings between the IBM Research and the Bayer HealthCare Research teams.

The workplace entry page for a lab head is shown in *Figure 8*. The monitoring section at the top of the page shows several views important to the lab head, such as status of experiments and cost of supplies. Below, the detailed status of experiments can be seen. The lab head can create new experiments, assign experiments to lab technicians, monitor the progress of experiments, review the detailed results of experiments assigned to technicians, and approve experimental results. On the right, the workplace contains instant-messaging capabilities to allow for ad hoc collaboration between researchers and technicians. When appropriate, the instant messaging function is context sensitive and enables the user to locate selected colleagues as a function of state.

Platform-specific implementation

Figure 9 shows the solution architecture for processing the experiment record as implemented on the WebSphere* platform. As shown in the figure, access to the data is via the Experiment Record Data Session EJB** (Enterprise JavaBeans). There are various ways to map the data graph to a set of access classes. In our example, we have used a simple mapping to Java classes and have used resource adapters to access the data sources.[16,17] The experiment-record state machine is a specialized implementation for state machine support using container-manager EJBs. The actions and events of the ABO model are represented as remote methods; whereas, data and remote actions are mapped to private methods of the entity EJB. A

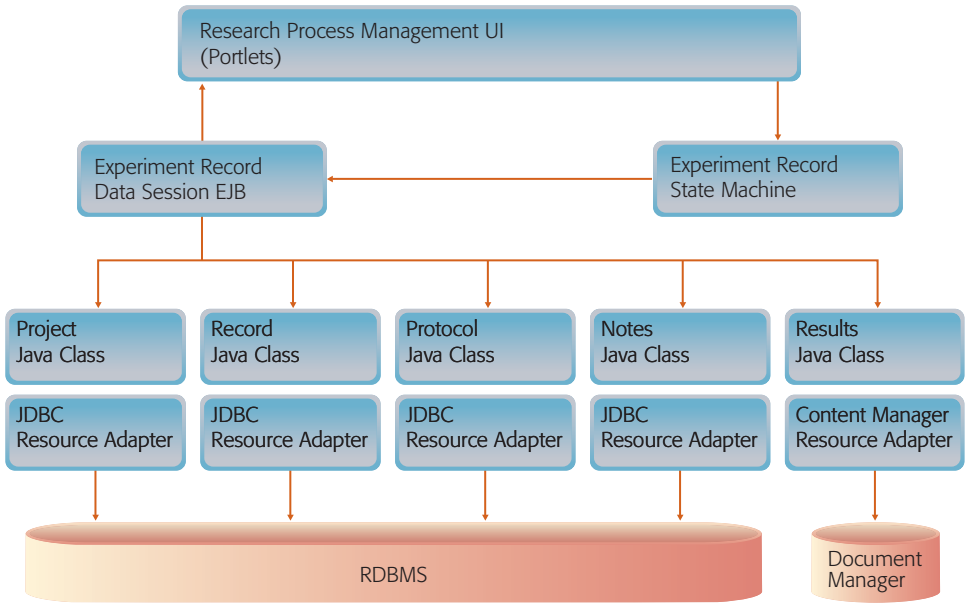

Figure 9
Solution architecture for processing the experiment record

specialized state-machine controller maps remote method invocations to private methods based on the current state of the EJB and the (externalized) state-machine definitions. FSM conditions are also mapped to private methods. In our current example we have handcrafted the portlets running in WebSphere Portal Server Version 5.0 to support the state adaptive access specification (we are in the process of automating the code generation of this function).

An important aspect of the workplace was the integration of typical collaboration tools, such as instant messaging and e-mail. These tools were integrated using IBM Lotus* Workplace technology.[18] We have integrated the collaboration tools so that collaborators can be called on as required, based on the current state of the process. For example, a chat list in the Sametime* panel configures itself to include the appropriate collaborators depending on where the user is in the process.[19]

Tooling

The success of any model-driven approach relies heavily on the tooling available to model business operations and support the implementation phase. In this section we briefly discuss the tools used for implementing our prototype.

The recently released version of WBI Modeler Version 5.1[8] is based on the Business Processes Definition (BPD) metamodel, a recent submission to OMG.[20] The BPD metamodel supports process modeling based on the UML2 token flow semantics. The business artifact modeling approach[14] can be represented using BPD and thereby modeled in WBI Modeler Version 5.

There is currently no commercial product available for the editing of a solution composition model as we have described. As all our metamodels are defined in UML semantics, we use the Eclipse Modeling Framework (EMF) model editor in Eclipse.[21] This tool allows for editing of metamodel instances, which are handled and stored as XMI** (Extensible Markup Language Metadata Interchange) documents. These XMI documents can be used to generate code for platform-specific implementations.

DISCUSSION

In recent years, most approaches to industrializing the drug discovery process have had a data-centered focus. Indeed, a major challenge in the daily life of a pharmaceutical researcher is finding clues from a vast amount of data. This pivotal fact has driven various research efforts in the areas of data mining and federated databases.[22] From a more business-

related perspective, Davis and Peakman have introduced the concept of *data-driven drug discovery* (4D).[23] In their work they illustrate the problems and the associated costs related to handling the many different types of data produced in pharmaceutical research. Their approach starts out by using a methodology similar to ours. The authors propose an analysis in several business data-related dimensions, such as data lag, data quality, data use, data leverage, data productivity, and data costs. Based on this conceptual framework, the authors created maturity profiles that estimate the individual strengths and weaknesses of companies. This type of approach is intriguing as it tries to go directly to the source of the problem. The authors recognize that most pharmaceutical companies typically suffer from problems in four areas: (1) loading of important data onto corporate systems, (2) sharing of data across organizations, (3) accessing, visualizing, and manipulating of data by scientists, and (4) gathering and presentation of data for decision making. A drawback of their approach, when compared with ours, is the lack of formal modeling techniques. Although their methodology leads to a maturity model that estimates the extent to which the company uses state-of-the-art techniques for data handling, their approach does not offer methods to solve any of the four problem areas.

By comparison, we analyze the current state of the business by modeling business operations and use this model to evaluate the efficiency of the business process through simulations or other techniques. This approach not only leads to similar results in terms of maturity assessment, but it enables the design of solutions to support the processes under consideration. This approach, centered in business operations, acts as an orchestrating principle and helps define goals for associated research activities and, in the process, develop methods for handling the data created. The formal models underlying the layers of the hierarchy and the mappings among them enable us to analyze and refine business requirements and eventually map them into IT solutions, thereby bridging the business-IT gap.

Another very interesting approach described by Peakman et al.[5] is to examine the drug discovery process as a traditional supply-chain model with volatile lead times. For example, maximizing the use of HTS apparatuses requires increased transparency of the various assay development efforts (i.e., the HTS labs can better plan their resource if they can anticipate the number of incoming requests for screening). When such transparency exists, the HTS groups may serve the next HTS request more efficiently, thereby reducing the cycle times of the assay development and HTS processes. In the same way, chemists delivering the compound libraries for the HTS process would benefit from more visibility downstream. Because of volatile lead times, the time to develop an assay and deliver it to the HTS team can significantly vary from project to project. Even for one development project, the time to delivery can vary in the course of the assay development process because a biological system may turn out to be much harder to prepare than initially assumed. Peakman et al. suggest that despite the problem of volatile lead times, collaborative planning concepts from automotive industries may apply to drug discovery as well.

In our approach, monitoring of KPIs, such as cycle times for processes, resource control, and other aspects, across various research processes would allow for gathering the information needed to perform collaborative planning calculations. Monitoring of business operations also allows visibility up and down the discovery supply chain.

We observed a trend towards "e-R&D", that is, the use of information technology to facilitate the overall industrialization of drug discovery R & D.[3] This concept has gained some traction in the areas of e-documentation, clinical trial simulation, and process optimization in clinical trials.

CONCLUSION

The business artifact-centric view we take here brings out a more subtle aspect of the drug discovery process. Although the analysis of data is clearly crucial for understanding the processes analyzed, a purely data-centric approach ignores the fact that the expertise for the interpretation of data belongs to researchers distributed across various organizations doing this work. Facilitating collaboration among researchers amounts to creating a distributed knowledge network. Experience shows that a complex problem is often solved by examining it from different perspectives. Sharing information across vertical "silos" of knowledge has a higher probability of solving such complex issues. Although this is, of course, only a hypothesis, the notion of horizontal integration across functional organiza-

tions is a popular theme in business restructuring and, in particular, in IBM's on demand strategic vision for the IT industry.

How applicable is our approach in the current pharmaceutical environment? During this research project we found out that modeling concepts are well-aligned to current thinking in pharmaceutical research. An implementation of a production system based on our concepts throughout the entire enterprise is a future challenge. One difficulty is the integration of a wide variety of apparatuses, some of which are not equipped with proper interfaces. Intelligent data analysis, which is not addressed by our methodology, is undoubtedly a very important aspect in pharmaceutical research.

The main goal of our work is to enable a methodology that will allow pharmaceutical companies to tackle three major challenges. First, in order to stay competitive in a constantly changing market environment, enterprises require IT solutions that adapt to changing market conditions. A business process management solution should provision capabilities to allow for efficient change management. Our approach accommodates change management by taking a model-driven approach to be applied hierarchically from the strategic layer down to the implementation layer. Changes in business strategy and operations are reflected in changes in the models and can be easily mapped to the IT environment without the need for extensive and costly re-engineering initiatives.

Second, the business value of IT solutions grows substantially when these solutions seamlessly integrate people, processes, systems, and information. This proposition is at the core of our approach as we introduce a novel approach to compose a model-based IT solution, which can be seen as an IT-level blueprint of the business itself.

Third, the ubiquitous nature of IT, its constantly growing capabilities, and its growing cost require a good understanding of how to make best use of both innovations and legacy systems. It is current practice to decide these questions at an IT level. As can be seen in large IT consolidation efforts in the pharmaceutical industry and many others, the question of how to leverage IT will be more and more a business decision. The business strategy and operations will set the context in which IT will be leveraged. Our methodology is an important step towards transforming the business intent and requirements directly into viable IT solutions.

In summary, in this joint project we have taken a model-driven approach to analyzing and capturing pharmaceutical research processes. By applying a heuristic transformation, we have created a solution composition model, which is an abstraction of the actual IT solution. The key technique for developing the solution composition model is the ABO component model. A prototype has been built based on WebSphere technology. We validated each step in the design process in the lab with the participation of the business users. Although we tested and demonstrated the solution in various scenarios, we have not yet deployed our solution into the Bayer production environment.

ACKNOWLEDGMENTS

Several people have contributed to discussions and support over the course of this project, including Anil Nigam, Nathan Caswell, Nitin Nayak, and David Cohn from IBM Research, Anne Aldous and Mark Pease from IBM Life Sciences, Tim Peakman from Biobank UK, and Joerg Weiser from Schroedinger Corporation. Furthermore, we were fortunate to have development help from our colleagues at the IBM China Research Lab.

CITED REFERENCES AND NOTES

1. J. A. DiMasi, R. W. Hansen, H. G. Grabowski, and L. Lasagna, "Cost of Innovation in the Pharmaceutical Industry," *Journal of Health Economics* **10**, No. 2, 107–142 (1991).
2. S. Arlington, *Pharma 2005: An industrial revolution in R & D*, http://www-1.ibm.com/services/us/imc/pdf/gw510-9220-pharma-2005-industrial-revolution.pdf.
3. S. J. Arlington, S. Barnett, S. Hughes, and J. Palo, "Pharma 2010—The Threshold of Innovation," *The Future of the Pharmaceutical Industry*, IBM Corporation (2004), http://www-1.ibm.com/services/us/index.wss/xs/imc/a1001099?cntxtld = a1000060.
4. J. R. Archer, "Faculty or Factory? Why Industrialized Drug Discovery Is Inevitable," *Journal of Biomolecular Screening* **4**, No. 5, 235–237 (1999).
5. T. Peakman, S. Franks, C. White, and M. Beggs, "Delivering the Power of Discovery in Large Pharmaceutical Organizations," *Drug Discovery Today* **8**, No. 5, 203–211 (March 2003).

6. S. Kumaran, "Model-Driven Enterprise," *Proceedings of the Global Enterprise Application Integration (EAI) Summit 2004*, Banf, Canada (2004), pp. 166–180.
7. R. S. Kaplan and D. P. Norton, "The Balanced Scorecard—Measures that Drive Performance," *Harvard Business Review*, 71–79 (January-February 1992).
8. WebSphere Business Integration Modeler, IBM Corporation, http://www.ibm.com/software/integration/wbimodeler/.
9. S. Kumaran and P. Nandi, "Adaptive Business Objects: A New Component Model for Business Applications," in preparation.
10. *Unified Modeling Language*, Object Management Group, http://www.uml.org/.
11. K. Vredenburg, S. Isensee, and C. Righi, *User-Centered Design: An Integrated Approach*, Prentice-Hall, Upper Saddle River, NJ (2002).
12. M. Beggs, "HTS—where next?" *Drug Discovery World* **2**, 25–30 (2000).
13. M. Beggs and A. C. Long, "High Throughput Genomics and Drug Discovery—Parallel Universes or a Continuum?" *Drug Discovery World* **3**, 75–80 (2002).
14. A. Nigam and N. S. Caswell, "Business Artifacts: An Approach to Operational Specification," *IBM Systems Journal* **42**, No. 3, p. 428–445 (2003).
15. J. E. Hanson, P. Nandi, and S. Kumaran, "Conversation Support for Business Process Integration," *IEEE International Enterprise Distributed Object Conference (EDOC)*, 65–74 (2002).
16. *The J2EE Connector Architecture's Resource Adapter*, Sun Microsystems, Inc., http://java.sun.com/developer/technicalArticles/J2EE/connectorclient/resourceadapter.html.
17. In the future we will make use of a more flexible and powerful approach using Service Data Objects (SDO) (see *Service Data Objects*, IBM Corporation and BEA Systems, Inc. (2003), Java Community Process, http://www.jcp.org/en/jsr/detail?id=235).
18. Lotus Workplace Products, IBM Corporation, http://www.lotus.com/products/product5.nsf/wdocs/workplacehome.
19. A demo and screenshots of the solution are available upon request from the first author.
20. *Business Processes Definition Metamodel: Concepts and Overview*, Object Management Group, http://www.omg.org/docs/bei/04-05-03.pdf.
21. *Eclipse Modeling Framework*, Eclipse Foundation, http://www.eclipse.org/emf/.
22. L. M. Haas, E. T. Lin, and M. A. Roth, "Data Integration Through Database Federation," *IBM Systems Journal* **41**, No. 4, 578–596 (2002).
23. N. Davis and T. Peakman, "Making the Most of Your Discovery Data," *Drug Discovery World*, 17–23 (April 2004).

Accepted for publication August 30, 2004.
Internet publication January 31, 2005.

Kamal Bhattacharya
IBM Thomas J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, NY 10598 (kamalb@us.ibm.com).
Dr. Bhattacharya, a member of the Business Informatics Department, focuses on the application of model-driven architecture concepts to real-world business scenarios, adaptive enterprises, and business performance management. Prior to joining IBM Research in 2001, Dr. Bhattacharya served as an e-business IT Architect for IBM Global Services in Germany and worked on several large scale e-business projects in the automotive and travel industry. He received a doctoral degree in theoretical and computational physics from Georg-August University, Goettingen, Germany, in 1999.

Robert Guttman
IBM Thomas J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, NY 10598 (rguttman@us.ibm.com).
Mr. Guttman leads a company-wide initiative that advocates the use of a model-driven approach to map the strategic goals of businesses into IT implementations. He received a B.S.E. degree in computer engineering from the University of Michigan in 1992. Upon receiving his M.S. degree in 1998 from the Massachusetts Institute of Technology (MIT), he founded Frictionless Commerce, a leading enterprise-sourcing software vendor, based on software agent technologies that he invented at MIT's Media Laboratory. These technologies are automating sourcing business processes, providing visibility into sourcing activities and information through role-based dashboards, and helping procurement professionals make optimal award allocation decisions resulting in direct, ongoing bottom-line savings. These innovations earned him a 2002 MIT Technology Review Top 100 Young Innovators Award (TR100).

Kelly Lyman
IBM Thomas J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, NY 10598 (kbowles@us.ibm.com).
Kelly Lyman is a User-Centered Design (UCD) lead in the Business Informatics Department. Her expertise is in researching target market groups and then using her findings to design innovative e-business solutions. Prior to joining IBM, she managed a user experience group at PeopleSoft. She completed her education at Carnegie Mellon—she holds a Master's degree in human-computer interaction, a Bachelor's degree in human-computer interaction, and a Bachelor's degree in information design.

Fenno F. Heath III
IBM Thomas J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, NY 10598 (theath@us.ibm.com).
Terry Heath, a Senior Software Engineer, has been a research engineer for over 14 years and has participated in many prototyping and customer engagement efforts in industries such as manufacturing, automotive, and electronics. He is a member of the Business Informatics Department where he focuses on formal modeling of collaboration and user interaction processes in business process management systems.

Santhosh Kumaran
IBM Thomas J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, NY 10598 (sbk@us.ibm.com).
Santhosh Kumaran leads a team of researchers in the area of model-driven business integration. His research interest is in using formal models to explicitly define the structure and behavior of an enterprise and employing these models to integrate, monitor, analyze, and improve its performance.

Prabir Nandi
IBM Thomas J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, NY 10598 (prabir@us.ibm.com).
Mr. Nandi, a member of the Business Informatics Department at the Thomas J. Watson Research Center, received a B.E. degree in electronics and communications engineering from Birta Institute of Technology, Ranchi, India, in 1990, and an M.S. degree in computer science from the College of William and Mary in 1997. He subsequently joined the Thomas J. Watson

Research Center, where he has worked on business process integration and management. He co-invented the Adaptive Document (ADoc) technology and pioneered the business artifact-centric way of modeling, composing, and implementing business process integration solutions. He also developed the Adaptive Business Object (ABO) concept and the related programming model. Mr. Nandi has authored a number of conference publications, journal articles, and patents.

Frederick Wu
IBM Thomas J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, NY 10598 (fywu@us.ibm.com). Dr. Wu, a research staff member at the IBM Thomas J. Watson Research Center, has worked in the area of electronic commerce and business integration for the past nine years. He holds S.B., S.M., and Ph.D. degrees from the Massachusetts Institute of Technology.

Prasanna Athma
IBM Healthcare and Life Sciences, Route 100, Somers, New York 10589 (pathma@us.ibm.com). Dr. Athma is a Bioinformatics Domain Expert in the IBM Life Sciences Solutions Development. She joined the Computational Biology Center at the Watson Research Center in 2000 as a research associate after working as an Associate Research Professor in New York Medical College, Hawthorne, New York. Her initial work at IBM was focused on structure prediction of proteins. She participated in Critical Assessment of Techniques for Protein Prediction (CASP) experiments, an international competition held every two years, in which protein sequences are released as targets, and the participating teams predict the structures in a blind manner before the actual structures are available. Her recent work at IBM Healthcare and Life Sciences has involved providing domain expertise in a broad range of areas such as genomics, transcriptomics and proteomics. She provides scientific domain knowledge in the development of middleware components for health care and life science solutions and customer engagements. She participates in MAGE and Proteomics standards by driving standards into IBM solutions. She has authored many journal and conference papers and holds two patents.

Christoph Freiberg
Bayer Healthcare, Pharmaceutical Research, D-42096 Wuppertal/Germany (christoph.freiberg@bayerhealthcare.com). Dr. Freiberg studied biology at the University of Göttingen and earned his Ph.D. from the University of Jena. He held research positions at the University of Marburg and the Institute for Molecular Biotechnology in Jena. For the past seven years, he has served as laboratory head at Bayer's Antiinfectives Research unit in Wuppertal. He is responsible for bioinformatics and genomics applications and for screening assays development in the field of antibacterial research. He also coordinates the biology activities of advanced drug discovery and leads structure-optimization projects.

Lars Johannsen
Bayer Healthcare, Pharmaceutical Research, D-42096 Wuppertal/Germany (lars.johannsen.lj@bayerhealthcare.com). Dr. Johannsen earned his Ph.D. in biochemistry from the Free University of Berlin and held research positions at the Bundesgesundheitsamt in Berlin and the University of Tennessee, Memphis. He has 10 years of research experience at Bayer's Anti-infectives unit. For the past three years he has been with the Information Services Department, the Scientific Information and Documentation Group, where he is responsible for the processing of document-based internal information.

Andreas Staudt
Bayer Healthcare, Pharmaceutical Research, D-42096 Wuppertal/Germany (andreas.staudt@bayerhealthcare.com). Dr. Staudt has more than 12 years of experience in information management in the pharmaceutical industry. He has held a number of international management positions, with increasing responsibilities, in the global Bayer HealthCare organization. His current position is Director, Proprietary Information and Research Support at Bayer HealthCare's Pharmaceutical Division. His responsibilities include the coordination of information technology for Global Pharmaceutical Research and the coordination of document management activities across Bayer's worldwide pharmaceutical organization. He earned his doctoral degree in physics from the University of Heidelberg, Germany. He was a postdoctoral fellow at the Max-Planck Institute for Nuclear Physics. He is co-author of a textbook on modern physics for graduate students and professionals. ■