POSITION PAPER

Prepared for the W3C Workshop on Semantic Web for Life Sciences
27-28 October 2004, Cambridge, Massachusetts USA

Knowledge Integrated Modeling (KIM), an application for the Semantic Layered Research Platform (SLRP)

A project led by the IBM Advanced Technology Group in collaboration with MGH/MIT/HMS Martinos Center for Biomedical Imaging.

By Sean Martin[1] and Anne Jackson[2], IBM Corporation.

The research process and information management technology

Information Management technology has come a long way, but the practice of Life Sciences research does not always reflect this. Despite the fact that the entire research process revolves around the production, organization and dissemination of information, much of Life Science research lacks an information management infrastructure that spans and integrates the entire research cycle. With the explosion of data, an accelerated movement towards Digital Biology[3] and collaborations across geographies, disciplines, and institutions, finding a cure for cancer, or any disease, will require a new approach to Life Sciences research and information management. Research practices have yet to take advantage of nascent knowledge integration technologies such as those offered by the Semantic Web. In short, information management and integration technologies are immensely underutilized.

Today, much of the research process remains manual, despite the trends in increased use of information technologies. Researchers may search the web, but they habitually write down web search results in their lab notebooks. They use a pencil to markup hard copies of documents and research papers which are then placed in piles on their desks. It is not uncommon to see ten open web browser windows and a great deal of cutting and pasting of text between them. Experimental simulations are run on computers, but the input parameters and results frequently come from and go back to paper or are sometimes stored somewhere in a file system, in a digital format - usually one that requires some kind of manipulation for use down the line.

A manual process is often inefficient and error prone. Records are mislaid or mixed up, work needs to be repeated and software programs and data versions get jumbled. People

1 Sean Martin sjmm@us.ibm.com

2 Anne Jackson ajackson@us.ibm.com

3 The NIH BISTI initiative http://www.bisti.nih.gov/2003meeting/

make mistakes in transcription. Much potentially interesting information is lost simply because too much effort is required on the part of humans to notice it, let alone capture it along with the necessary context[4] that would make it useful later. Investigators may leave taking a key piece of information about a project history away in their heads or it is lost as a scribble in lab notebook that nobody can decipher.

Collaborations, which are critical to the research process and progress, take place either in person or by laboriously gathering and emailing textual context along with associated spreadsheets, images and documents as attachments. Information, data and ideas, each in its context or described in relation to all the other information in a research endeavor is not easily shared with collaborators across the lab, let alone across the world, reducing the probability of generating the human and machine generated insights we need.

In addition to being manual, each step in the research process is all too often viewed by the researcher as logically and physically discrete. Not unnaturally many scientists tend to look at their problems in terms of the individual tools they will need to complete each individual investigative sub-task. However, it is not just general lab apparatus that is used in this way. Information technology tools are similarly utilized mainly for discrete, separate processes, without researchers realizing the opportunity costs of this single dimensional view of computing infrastructure.

What is lacking is a holistic information gathering and management system that integrates, tracks and aids the end-to-end research process. Examples of features such a system might provide include the following: the recording of provenance and context, or the systematic archiving of raw and processed data; the automated repetition of investigational tasks with variations caused by new information; automated discovery with alerts for newly arrived relevant information; the smoothing of information flows and impedance mismatches between experimental stages – especially in user interfaces; the wider query of and possibly automated inference against better integrated information stores; evolved abstractions to the scientist's view of their computing resources which are bettered suited to achieving the scientist's research goals without requiring advanced computer science skills.

Moreover, scientist access to available High Performance Computing (HPC) resources can be severely gated both by the skills required to employ these somewhat arcane systems and the length of time it takes to prepare for each use. To access HPC resources today, the scientist often needs to be moderately computer programming literate in order to manually string together and debug a script or a database query or to program the pipelines that form their digital experiment. Additional layers of software will help ameliorate this problem and will sharply increase usage; however, scientists don't often have the know-how or resources to produce them.

[4] Context refers to related information that provides important background about the data which helps a user to understand where an item of data fits in and what it is related to..

The formal output of research is often text and diagrams published in journals for peer review, rendered using methods that are still relatively impenetrable to natural language parsing technologies. The recording of research provenance is another usually laborious but still very necessary manual exercise as is the archiving of experimental data and results. How much easier would collaboration be if it were straightforward to accurately reproduce and then easily produce variations of another scientist's methods because there could be no ambiguity about the methods, software and data used in the original experiment?

Many research projects require the relatively manual organization and coordination of multiple disparate, but interrelated, logistical odds and ends of data such as grant proposal documents, collaborator emails, intellectual property agreements, research papers with annotations to read, fragments of research papers being written, refrigerator content lists, protocols, monographs and sketches of ideas along with the growing experimental data collection – the list is almost endless. Where are the information systems that can be used to support the organization of a growing collection of much unstructured data?

In summary, the task-oriented approach to information technology results not only in a manual, inefficient process but ignores the true potential of information tools. Effective information management tools can provide the context, the provenance, the insights and fluid connections to all data in what is, after all, a process intended to create and protect information.

Using a Semantic Data Web for an integrated approach

The *Knowledge-Integrate-Modeling* (KIM) system is a custom application designed to support Systems Biology Research. KIM is constructed using IBM's more general *Semantic Layered Research Platform* (SLRP) prototype.

The Semantic Layered Research Platform is principally based on GRID[5], High Performance Computing (HPC) and Semantic Web[6] technologies and concepts. At its heart, the SLRP system has a central data store named CART, in which all stored information is represented in a "Resource Description Framework" (RDF)[7] model. RDF is a flexible method of holding data conforming to practically any schema. CART supports queries against the store using RDQL[8].

Views of this central data store along with its event model will be automatically synchronized or replicated as necessary to and from multiple end points including local RDF cache stores for end user applications; connections to scientific instruments; and even out to individual computer job processes executing on a GRID. An online/offline model is supported for distributed clients. CART also provides complete access control to information at the RDF triple and document levels. Users and administrators of the system, registered in a Light Weight Directory Access Protocol[9] (LDAP) directory system can establish Access Control Lists (ACL's) for any information in the store. CART is designed to treat data in the store as individual RDF triples, multiple RDF documents or one large RDF document for the purposes of query, update and deletion. CART maintains a full revision history for audit and provenance purposes.

All SLRP data and system objects are named using the Object Management Group's Life Science Identifiers (LSID)[10] a new industry standard for uniquely naming and thereby identifying any digital artifact. Life Science Identifiers provide a handle by which that object might be later retrieved and a unique key on which meta-data about that object or its relationship to other objects might be attached, retrieved and merged from multiple sources.

The LSID concepts and protocols are also fully supported by the SLRP system's Distributed Data Repository (DDR), a write once distributed storage system for binary objects including documents, images and experimental data sets. DDR provides managed

5 Foster I., Kesselman, C., Tuecke, S. ,The anatomy of the GRID (http://www.globus.org/research/papers/anatomy.pdf)

6 Berners-Lee,T., Hendler, J., Lassila, O. (2001) The Semantic Web, Scientific American, May 2001 (http://www.scientificamerican.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&catID=2)

7 Miller E, Swick R, Brickley D (2003) 'W3C Resource Description Framework (RDF)' (http://www.w3c.org/RDF/)

8 Seabourne, A HP Labs Bristol, RDQL - A Query Language for RDF (http://www.w3.org/Submission/RDQL/)

9 Benett G., LDAP: A Next Generation Directory Protocol (http://www.intranetjournal.com/foundation/ldap.shtml)

10 EBI, I3C, IBM (2004) 'Life Sciences Identifiers final adopted specification, Object Management Group, Document dtc/04-05-01 ((http://www.omg.org/cgi-bin/doc?dtc/04-05-01)

storage and immediate access to data objects from distributed clients, scientific instruments and for programs running in GRID clusters. It also manages the caching and synchronization of these objects for an online/offline remote client model. Again access to all data objects in the DDR is restricted under user or administrator controlled ACL's.

Slingshot, a further sub-system of SLRP, provides a distributed workflow execution and coordination engine and is implemented as an OWL-S[11] interpreter. Slingshot relies on the central RDF store with reliable distributed eventing provided by CART. Slingshot instances are capable of coordinating the sophisticated "ad-hoc" flows of digital experimental processes that may execute a single flow over the following heterogeneous environments: HPC/GRID, Web Services[12], client-side user interactive programs. Flow processes are described in RDF conforming to the OWL-S 1.1 ontology with extended groundings to support HPC/GRID and local processes.

SLRP will include a client environment development framework that is based on the cross-platform open source Eclipse Framework[13]. While Eclipse started as an advanced Integrated Development Environment, it may now also be used as the runtime for delivering rich cross-platform desktop client applications written in Java. In order to support the RDF data model in Eclipse, SLRP includes Telar, a Java toolkit for Eclipse that enables the mapping of RDF data using the CART application programming interfaces to SWT[14], JFace[15], and Eclipse Forms. Developers can use Telar to define operations that will map to one or more RDF data triples. Operations are mapped to UI widgets that may themselves be clustered into logical units to make complete UI forms or form fragments. In this way, information in each form may be drawn from a variety of disparate data sources. Telar helps the developer to manage the life cycle of the form and its underlying data.

Jastor is another tool that will help SLRP application developers. It is a code generation tool for creating custom libraries to provide an object-oriented view, including events, of RDF that conforms to a chosen OWL ontology. Currently Jastor emits Java classes but is extensible to other object oriented programming languages. Jastor is fully compliant with OWL-Lite, but supports an increasing number of features of OWL-DL and OWL-Full.

GRID technologies provide SLRP with the systems management framework it can use to provide and integrate support for the automatic deployment and tending of software systems, applications and process flows for compute and data intensive tasks in large clusters of computers sometimes deployed over a wide area.

The SLRP platform provides applications with services to:

11 DAML Services Home page http://www.daml.org/services/owl-s/

12 W3C Web Services Activity page http://www.w3.org/2002/ws/

13 The Eclipse Project home page http://www.eclipse.org/

14 The Standard Widget Toolkit http://www.eclipse.org/articles/Article-SWT-Design-1/SWT-Design-1.html

15 Jface: UI Framework for plugins http://help.eclipse.org/help30/index.jsp?topic=/org.eclipse.platform.doc.isv/guide/jface.htm

- Connect a researcher's desktop directly to the powerful databases in which all the research information collected by a collaborative research effort is held along with its associated semantic linkages to other networked data sources (public and private) as well as the meta-data that provides each item of stored information with a context.

- Provide the information management backbone into which research tools and processes may be integrated in order to achieve an overall information system for all aspects of an ongoing research project. These may include new tools that can help scientists to capture their ideas and concepts and the relationships of these to data stored in the system or as LSID pointers to information remotely accessible.

- Provide a system that supports queries against all aspects of the ongoing research process. Later the system will be able to support inference against the same information and an infrastructure for the automatic generation of alerts or the triggering workflow based on the introduction of new but potentially related information.

- Enable wide area[16] collaboration between researchers by allowing the sharing of any data object or concept in the system along with its semantic context, linkages to information and meta-data, thus allowing an interconnected web of information to easily be communicated between collaborators.

- Allow the structured annotation of any object in the system (e.g. a phrase in a research paper or spreadsheet, a comment in a code module, an interesting area of an image, a parameter in an experimental result, a node in a brainstorming "concept" semantic web) whether it be to add additional meta-data or simply to note a semantic relationship between stored objects or concepts.

- Provide interfaces to the GRID compute platform where models or digital experiments can be executed by the assembly and execution of workflows representing digital experiments by end-user researchers, without the need for technical understanding of the sophisticated underlying computer systems involved.

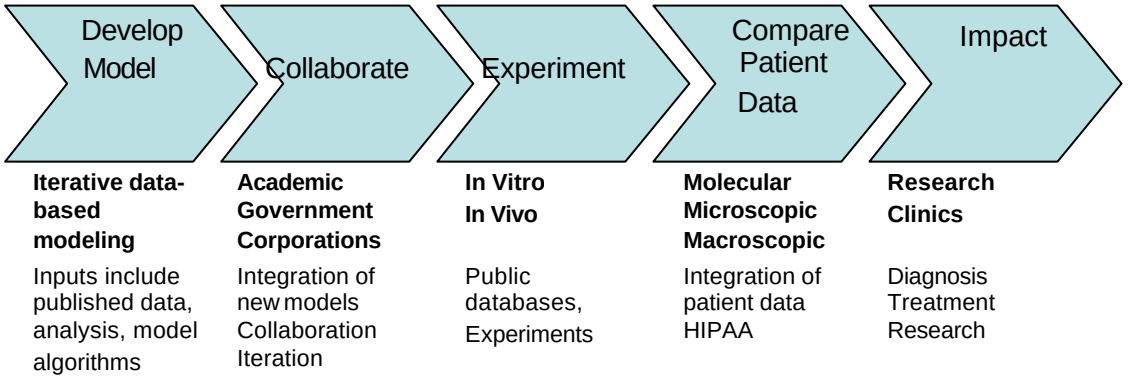
- Provide visual interfaces for the integrated viewing and annotating of all information associated with and created by a research effort.

- Provide a collaborative environment designed for the creation, debugging & maintenance and reference archive of the source code modules associated with a research effort that requires significant computer aided modeling, simulation and data intensive computing.

[16] Wide area collaboration refers to collaborations that span multiple disciplines and geographical locations, as well as ad-hoc collaborations where researchers aren't working together for a long period of time.

A closer look at the KIM application

The Knowledge Integrated Modeling application goals will alter over time. In the first phase the application must support the iterative initial development of a computer model capable of simulating the growth of brain tumors through multiple scales. The second phase includes the introduction and integration of additional models from collaborators in different disciplines working at different institutions. In this phase the system must support new wide area interactions as well as data and compute sharing. During the third phase the application will guide and manage the acquisition of experimental data which will be used for further refining the accuracy of the combined model. In the forth stage, human data will be introduced requiring strict adherence to the HIPAA[17] adding further refinements to the accuracy of the model. In the final phase, the KIM application alters to become a clinical diagnostic and treatment research tool.



End user researchers using the KIM application will be provided with a "Biologists Workbench" desktop application program that can be customized to aid in the tasks that person carries out in their day to day work and collaborations with colleagues. This application will be based on the Eclipse Framework. The Eclipse Framework already provides a great deal of support for code developers including source code editors, source code repository integration, viewers and debuggers.

Through the addition of SLRP components for annotation and connectivity to the central SLRP data stores and custom *Eclipse Perspectives* (a perspective is a mechanism for extending the Eclipse visual framework) for tasks like wide-area collaboration, views for clustering of related data objects or annotating information in PDF (Appendix *Figure 1*) or from the web (*Figure 2),* image manipulation and comparison (*Figure 3*), support for spreadsheets, word processing, and for creating, executing and later reviewing digital experimental work flows (*Figures 4 &5*), this environment will provide an excellent single desktop work space for the collaborating systems biology team.

17 Standards for privacy of individually identifiable health information.
http://www.hhs.gov/ocr/hipaa/guidelines/guidanceallsections.pdf

Researchers using the KIM application to track their experimental modeling will automatically create and store semantic links to the inputs and output of any particular model run, along with LSID references to the actual source code modules that were executed for any particular simulation execution run and the data produced. The application will automatically document and make directly accessible the knowledge required to exactly reproduce a digital experiment without fear that the an incorrect version of the source code or input parameter data is being used. The Workbench includes Integrated Development Environment perspectives (*Figure 6)* supporting collaborative work on the source code for the actual simulation models and features shared annotations for source code.

Summary

Information management technology shows great promise for Life Sciences research. As technologies such as Semantic Web are developed and deployed, they make possible applications, such as KIM, which may totally transform the research process. With integration of sources of information, experiment algorithms, results and the scientist's conclusions, the research process is virtualized, collaboration is easier and data with provenance is preserved in context..

Appendix

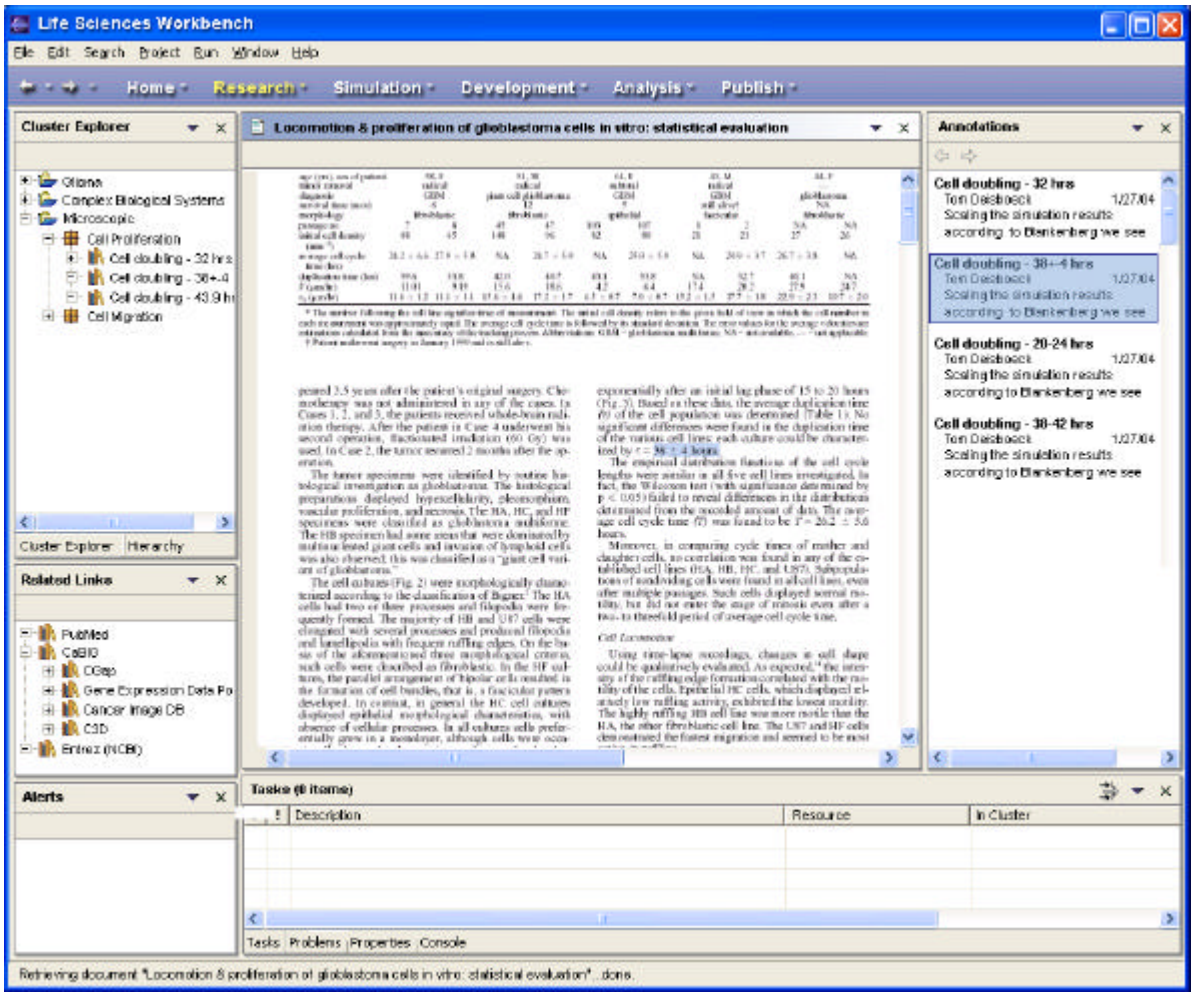


Figure 1 Biologists Workbench - Perspective showing in place viewing and structured annotation of a PDF formatted research paper.

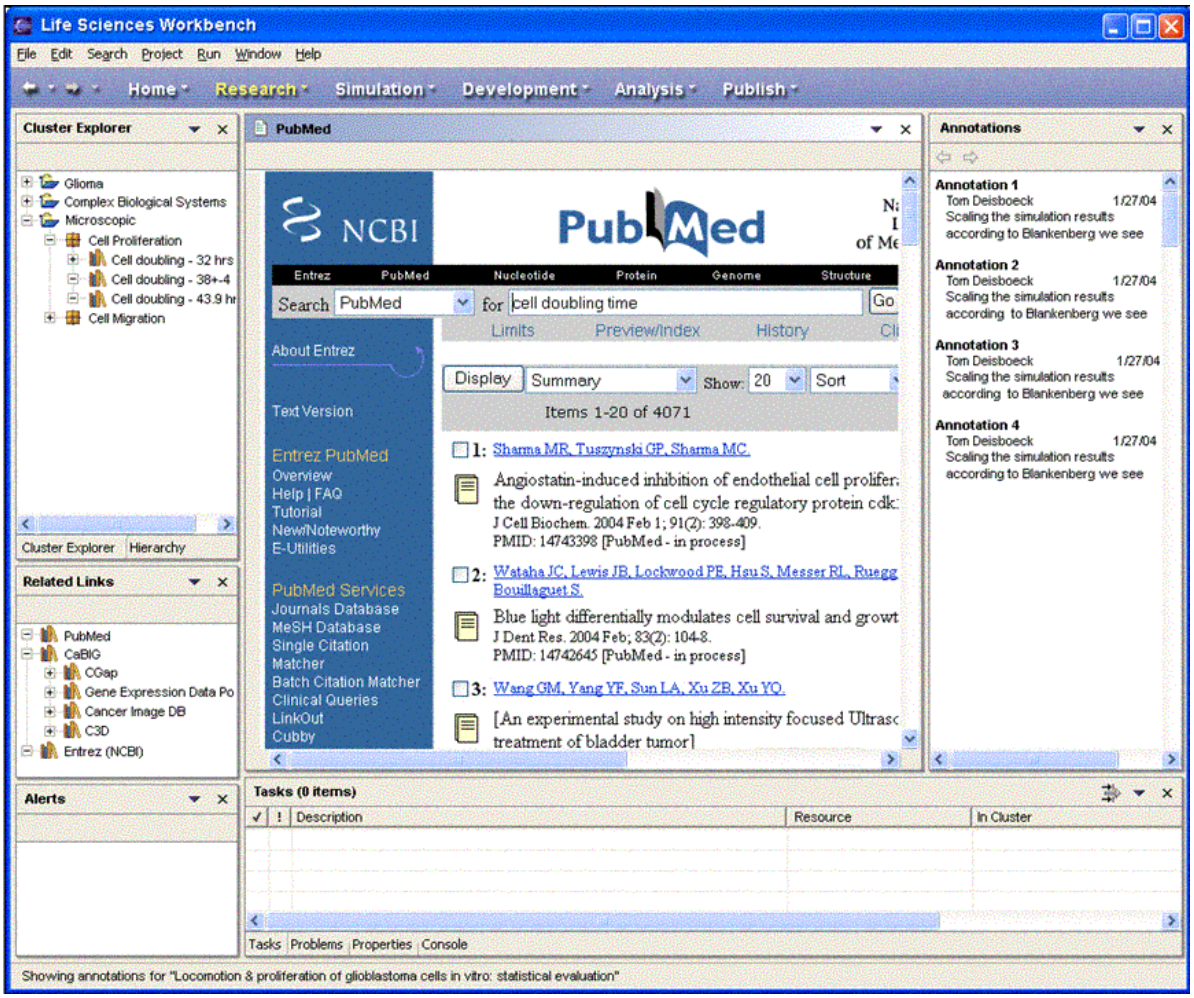


Figure 2. Biologists Workbench – Perspective for searching web based medical research information with annotation.

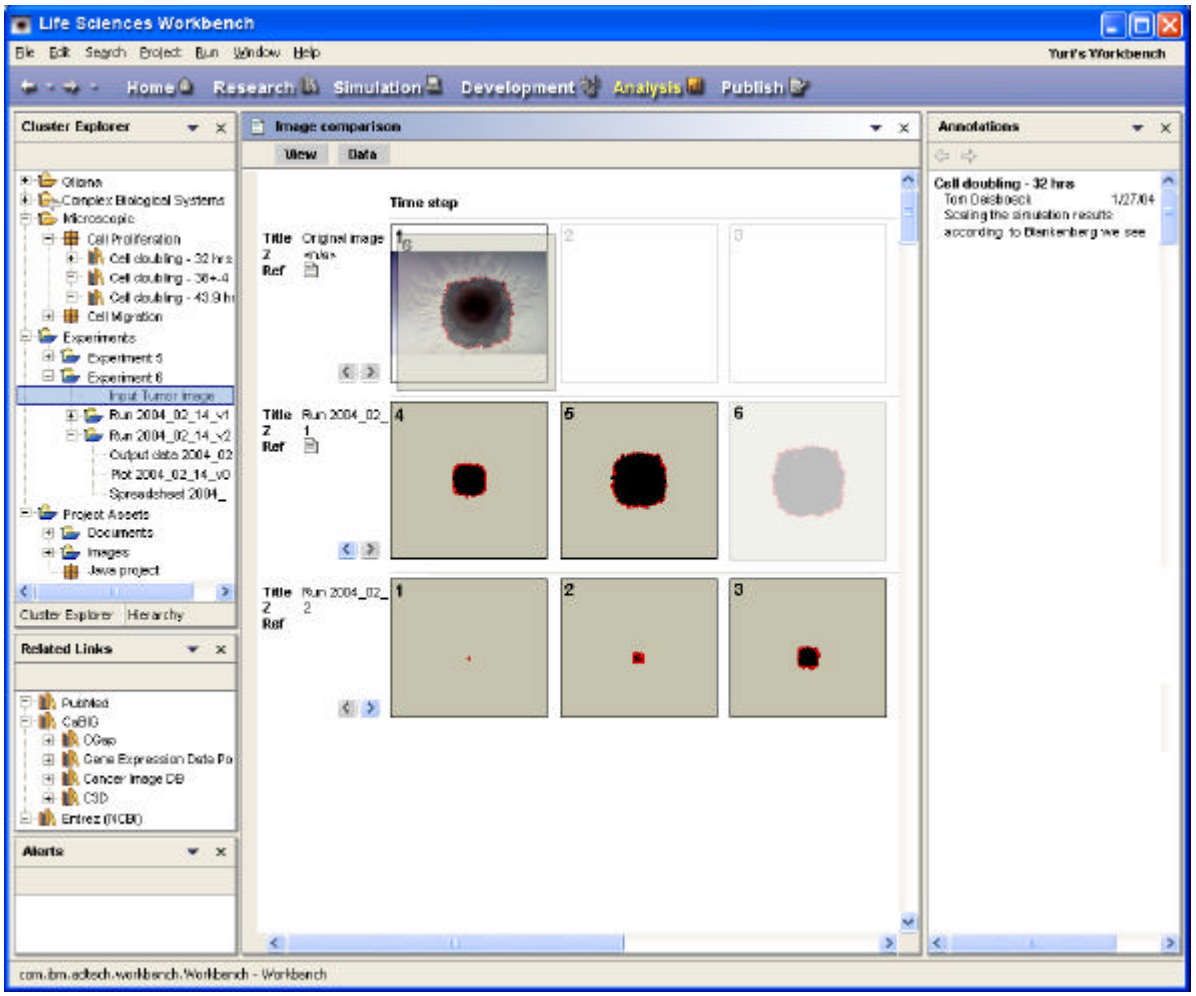


Figure 3 Biologists Workbench – Perspective for i mage comparison & manipulation with annotation.

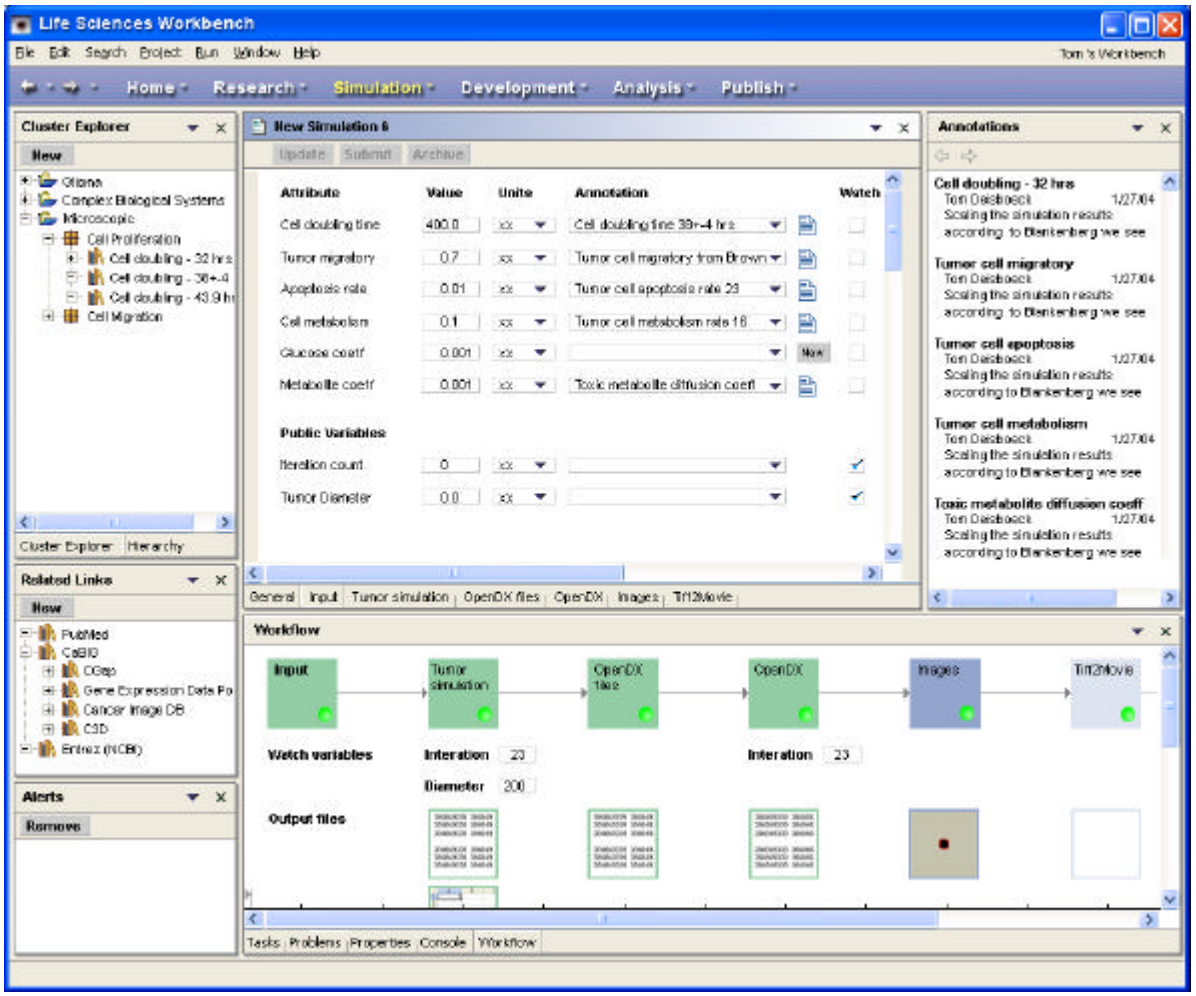


Figure 4 Biologists Workbench – Perspective for viewing experimental workflow and intermediate results in context.

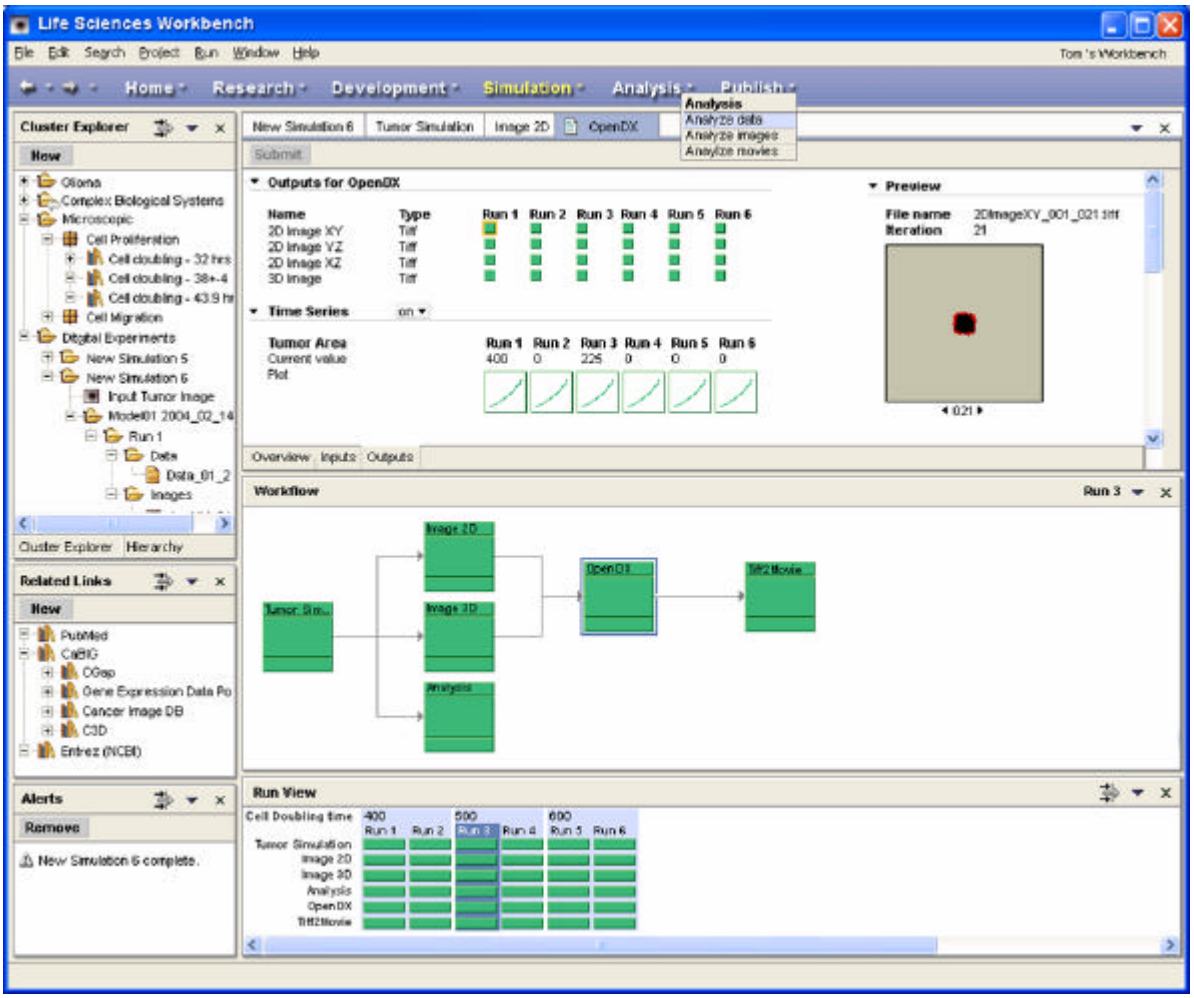


Figure 5. Biologists Workbench - Perspective for exploring and analyzing the results of a digital experiment.

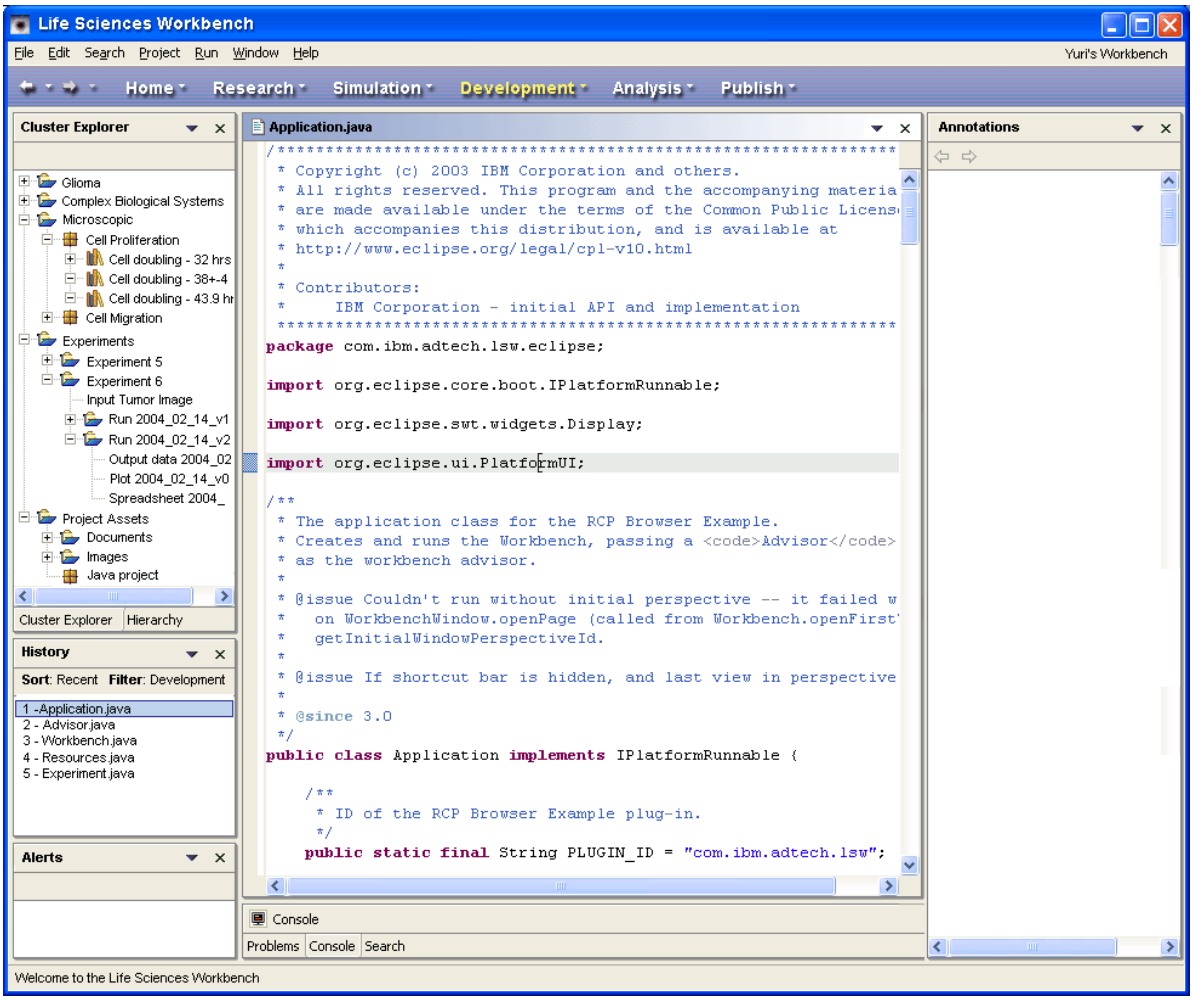


Figure 6 Biologists Workbench - Perspective for editing and debugging experimental simulation logic.