Architecture of the Open-source Clinical Research Chart from Informatics for Integrating Biology and the Bedside

Shawn N. Murphy, MD, Ph.D.[1], Michael Mendis[1], Kristel Hackett[1], Rajesh Kuttan[1], Wensong Pan MS[1], Lori C. Phillips MS[1], Vivian Gainer MS[1], David Berkowicz MD[1], John P. Glaser, Ph.D.[2], Isaac Kohane, MD, Ph.D.[3] and Henry C. Chueh, MD, M.S.[1]
[1]Laboratory of Computer Science, Massachusetts General Hospital, Boston, MA, [2]Partners Healthcare Inc., Wellesley, MA, [3]Children's Hospital, Boston, MA

Informatics for Integrating Biology and the Bedside (i2b2) is one of the sponsored initiatives of the NIH Roadmap National Centers for Biomedical Computing (http://www.bisti.nih.gov/ncbc/). One of the goals of i2b2 is to provide clinical investigators broadly with the software tools necessary to collect and manage project-related clinical research data in the genomics age as a cohesive entity, a software suite to construct and manage the modern clinical research chart. The i2b2 "hive" is a set of software modules called "cells" that have a common messaging protocol that allow them to interact using web services and XML messages. Each cell can be developed by independent investigators to achieve specific analytic goals, and then be integrated into the hive to enhance the functionality available in the i2b2 Hive. We have applied this architecture through several ongoing clinical studies and found it to be of high value. The current version of this software has been released into the public domain and is available at the URL - http://www.i2b2.org.

INTRODUCTION

The i2b2 team has developed an interoperable framework of software to implement the vision of a research chart that can be extended for new and unanticipated data types as well as functionality. It is intended to serve the following users:

- Clinical investigators who want to use the software in as "shrink-wrapped" a way as possible,
- Bioinformatics scientists who want the ability to customize the flow of data and interactions, and
- Biocomputational software developers who want to develop new software capabilities that can be integrated easily into the computing environment.

The framework of software modules is called the i2b2 Hive, and is centered upon two integration strategies. The application integration strategy is based upon web services provided by applications that are "wrapped" into functional units called "cells", such that their functionalities are exposed as messages that travel to and from the various cells of the hive. The data integration strategy is modeled after that of the Research Patient Data Registry [1], optimized to serve as a repository for data from the medical record and associated genomic data. The data that are stored in the i2b2 Hive are further organized so that data ownership and data privacy is preserved even when shared across several different groups or entities.

The i2b2 Cell is the basic building block of an i2b2 environment, and encapsulates business logic as well as access to data objects within the cell behind standard Web service interfaces. The communication within the i2b2 Hive occurs through these web services provided by the i2b2 Cells. A Web service describes a standardized way of integrating Web-based applications using the XML, SOAP, REST, and WSDL open standards [2] over an HTTP Internet protocol backbone. XML is used to tag the data, SOAP or REST is used to transfer the data, and WSDL is used for describing the services available. Because all communication is in XML, Web services are not tied to any one operating system or programming language. For example, Java can talk with Perl, and Windows applications can talk with UNIX applications.

As a collection, the i2b2 Cells are loosely coupled and generally known to each other only through the use of the Web services. The web service model was adopted to adapt to the reality that most researchers work in an environment where they are relatively isolated from infrastructure projects. Even when they do work within an infrastructure project, the emphasis is integration into that environment alone. This impairs the reusability of the code produced by much of their work. As a result, many analysis methods developed by outside scientists can not be applied to internal and protected data sets.

The i2b2 Cell can be used to share functionality between Hives. Some examples of an i2b2 Cell would be to perform concept extraction from clinical narratives (natural language processing) and data de-

identification where reports are rendered free of specific names and dates. Remote cells can be hosted so that the code (perhaps proprietary) does not need to be shared. Because the functionality of the cells is exposed through web services, there are no assumptions of proximity (such as on the same server) and a cell could be built at one institution and exposed to another.

The importance of providing an architecture that allows cells to expose their internal workings only through web services is that the developers of the cells do not need to be intimately familiar with other principles used to build the Hive. For example, if a cell wishes to provide the added functionality of natural language processing to the Hive, the developers need to know only how to receive XML messages providing a document from the repository cell, and transmit messages describing the coded concepts that were extracted from the document back to the data repository cell. The work of taking the concepts and recording them in the clinical research chart is taken care of by other Hive services. Thus a scientist is able to focus on the specific functionality which is their expertise, and not be concerned with the other details of managing a clinical research chart.

The goal of the i2b2 Cell is to expose functionality at many levels to many roles in the genomics research domain. Bioinformaticians and software engineers will want to develop and wire together Cells, where investigators will want to use visual tools. An intermediate role may be the most important: those integrators who can construct applications on top of i2b2 Cells to create domain-specific workflows.

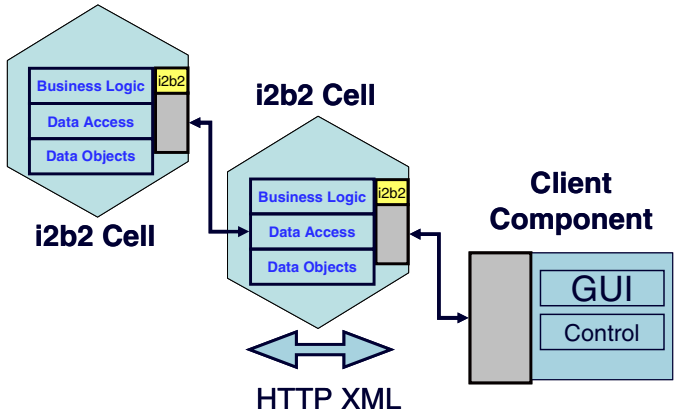

Figure 1 - Unlike traditional client/server models, such as a Web server/Web page system, Web services do not provide a visual user interface. Web services instead share business logic, data and processes through a programmatic interface across a network. These can then be connected together into workflows, administered by a traditional GUI client. Cells may invoke other Cells. This means that a developer can independently create complex behavior and user interfaces that reuse the functionality of existing Cells

METHODS

The i2b2 Hive consists of a number of core Cells that establish basic services to support the activities of the Clinical Research Chart (CRC), as well as any number of additional Cells to provide enhanced services. It is intended to be a scalable approach for managing an increasing number of independently developed software services to be contributed to the CRC.

Fundamentally, the CRC is built to hold medical and medically-oriented genomic data. Any of the various cells of the i2b2 Hive may contribute to placing the data into the CRC, which ultimately occurs by sending XML messages to the CRC.

The overall model for an i2b2 message regarding patient data to or from the CRC is an XML schema that after defining a header for management of the basic communication then defines a message body that contains patient sets with their related phenotypic and genotypic data as well as possible references to other data objects in the Hive. This type of message is referred to as a "document-oriented" message in the web service model. [2] It is different than a remote procedure call (RPC) type message, which is a more traditional function call with parameters that returns a data object. The document-oriented message allows data to be passed to a cell, have new data added, and then moved to the next cell for additional analysis. In this way fairly complicated workflow patterns can be accommodated when working on particular sets of medical records.

The complexity of raw clinical data is very high making it difficult to use for research. Use of data derived from the medical record is governed by HIPAA regulations [3] and must comply with specific security requirements [4]. Specifically, for identified data from the medical record to be used for research, the patient must have been given a notification of such before their data is entered into the medical record. Even with notification, identified data may only be released with approval from the Institutional Review Board (IRB), and in general may not travel outside of the entity (usually the originating

Hospital). However, HIPAA allows identifiers to be removed and have a "Limited Data Set" created [4]. One of the core cells of the Hive (core cells are shown in dark gray) is the Identity Management cell. This cell contains the "code book" that maps real patient identifiers to arbitrary patient numbers in the CRC. The cell needs the identified data to determine if new patients in the Hive map into old patients in the Hive, and associates their data with the correct coded numbers. It is only accessible by those with IRB approval within the entity, but the rest of the Hive is accessible to all those on the project who have signed a HIPAA "Data Use Agreement". The consent process is also managed within the Identity Management cell.

By understanding the behavior of our driving biology projects (DBP's), it is clear that as data are collected and analyzed they typically go through a cycle of ownership for exclusive use, during which they are not considered for sharing. This data for research are created in small sets to achieve the goals of the specific research study. A smaller set of data are picked from larger amounts of clinical data. Once assembled, a considerable amount of data cleaning and integrity checking occurs. Once the original research is prepared and published, The data may be considered for sharing outside of the original project team. This, now curated, data are of much higher value than the original raw clinical data, but this curated data often remains in silos because the data formats are typically different in each silo.

The i2b2 Hive is built on the assumption that sharing is the eventual intention of all research. It accommodates this cycle through the creation of new Hives or "projects" within a single Hive. Each project has its own CRC, and the passwords and certificates are maintained by the Project Management cell. However, the data schema of the CRC and the data curation through the Ontology cell allows the data from several projects to come together after it is deemed to be sharable by a project team. Essentially they share common metadata and are structured with a consistent data model, such that they can be joined together at any point to create a larger, shared CRC.

RESULTS

As part of the i2b2 project, we have been working with our DBP's while they engage in active clinical research, allowing us to test our tools and architectures. Some of the functionality of the CRC is best explained within the context of a sample clinical investigation that begins with a minimal amount of supporting hospital clinical infrastructure.

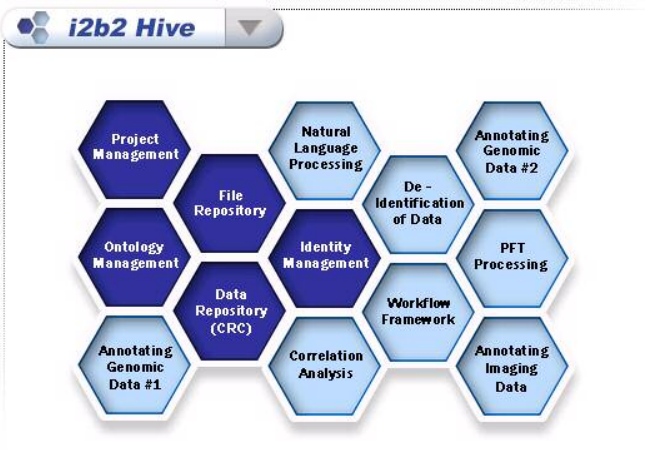

Figure 2 - The i2b2 Hive consists of a set of modules that interact using Web Services. Core modules (dark) are essential for the operation of the Hive. Options modules (light) add functionality to the hive but are not essential.

The example investigation is of Asthma patients with only text notes available from an asthma clinic and some text reports from a pulmonary function test laboratory. Most medical centers will have additional information available through the medical record and other data sources, but here we will focus on the limited case where only text are available to show how medium sized clinics can take advantage of the i2b2 Hive. The text reports will be processed through the Hive into specific concepts associated with patients, and the concepts are placed into the CRC.

The clinic notes are added through the Identity Management cell. The names and medical record numbers are resolved and retained in the Identity Management cell. When information is added to the CRC from identified medical records, it always must pass through this cell. Besides stripping the identifying information from the record and exchanging the identifying numbers for a random code that will be used to represent the patient, the Identity management cell also matches the patient using this information to any patients in the CRC which are computed to be the same person. This corresponds to a master patient indexing service.

Notes are added to the CRC in an encrypted format, and only those with access rights to see the identified notes will have a key/certificate to view them. This

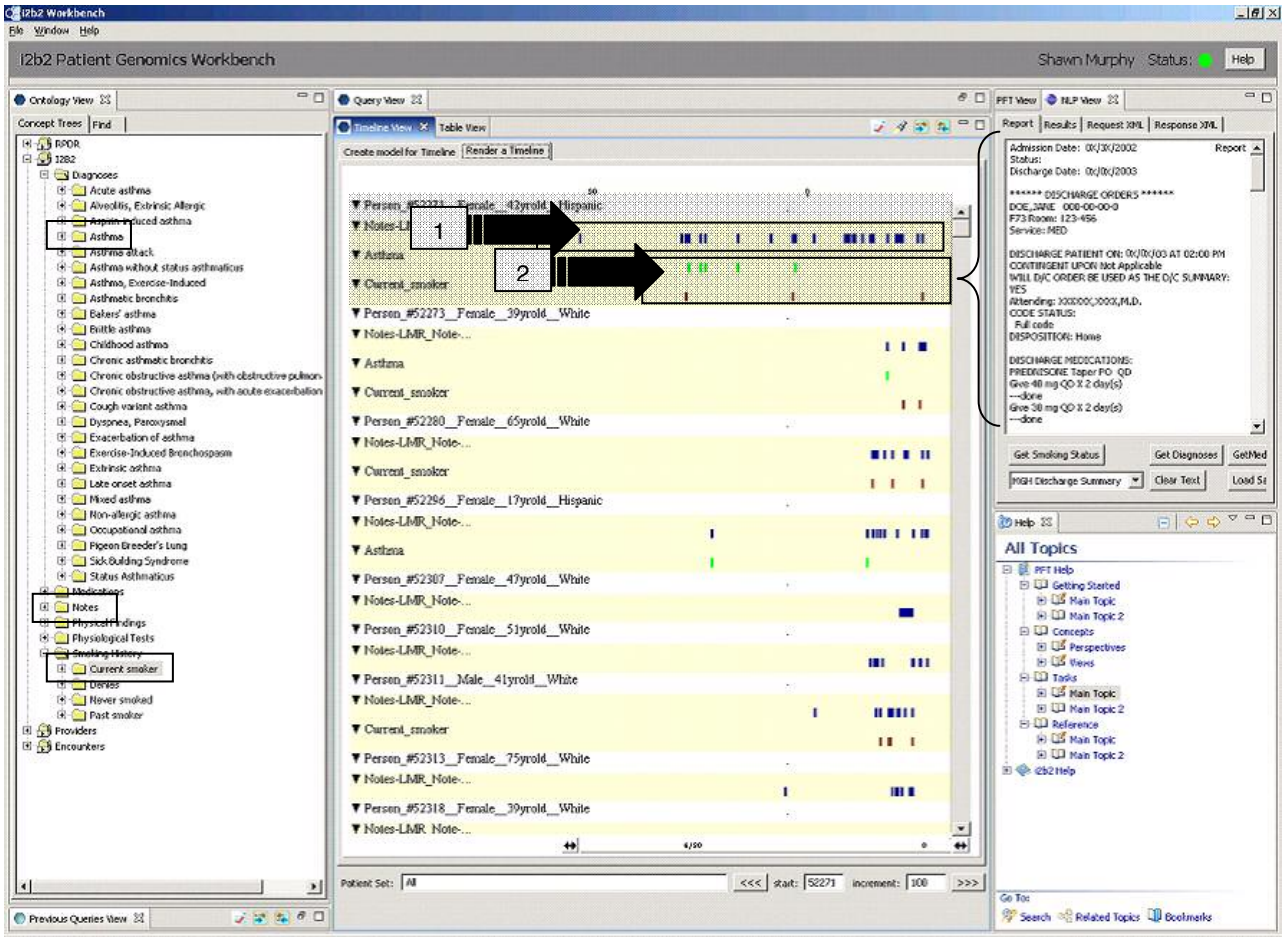

Figure 3 – The i2b2 Workbench shown above is used to display the data placed into the Data Repository Cell (CRC). The terms from the ontology cell are shown to the left, and represent concepts that have been entered into the CRC. The data representations of these concepts are shown along a time line that extends across 10 years. The Notes written on the patients are shown as #1, and the concepts of "asthma" and "current smoker" that have been extracted from the notes are shown in box #2. This is repeated for each of the patients in the CRC.

preserves the CRC as a HIPAA defined limited data set. Notes that are not encrypted are passed through the de-identification cell that will obfuscate names, places, and dates prior to being placed in the CRC.

Concepts are extracted from the text notes and placed back into the CRC as individual codes. Four cells are involved in the processing of the text notes. The notes are pulled from the CRC using the Workflow Framework cell. They are sent to the Natural Language Processing cell one by one where the concepts are extracted, and the concept codes are checked for integrity and then placed back into the CRC. As a result of the import process, phenotypic data are available that may be queried and viewed using the i2b2 Workbench application. In a similar way, the PFT cell extracts numeric values from the reports on pulmonary function tests (PFTs) and places values into the CRC associated with these individual concepts. For example, a concept of "Percent of predicted Vital Lung Capacity" comes back from the PFT cell associated with the concept "i2b2:pulfvcpred" and the value "65" with units "%".

The data from the text reports in the medical record and from the pulmonary lab now exist in the CRC database as specific concepts and associated values. These concepts and values may now be manipulated and displayed in ways that are familiar to clinical researchers. The human interaction with the CRC and the rest of the i2b2 Hive is managed through a client application named the i2b2 Workbench.

The same principle guiding the development of the i2b2 Hive guides the development of the i2b2 Workbench. The Workbench consists of a collection of plug-ins contained within a loosely-coupled visual framework in which independent plug-ins from various teams of developers can fit together. The i2b2 Workbench uses the open-source Eclipse framework [5] to contain the plug-ins. These plug-ins provide the manner in which users interface with the cells of the Hive. When a cell is developed, a plug-in can be developed along side it to support many of its operations. A complete package to illustrate the development of a cell and its accompanying plug-in can be found at www.i2b2.org,.

DISCUSSION

The i2b2 hive is an open source software platform for managing medical record and associated genomic data for research. It has an architecture that is based upon loosely-coupled, document-style web services to allow researchers, who typically do not fit their work into overall infrastructures, expose their work. Without the i2b2 Hive, one may have to wait many years before the engineers are able to perform this integration work. Issues of patient privacy and complex ontologies also hinder connections with the researchers developing new methods. The i2b2 Hive incorporates a secure infrastructure, and data-driven ontology tools. Data is stored in a relational database that is able to fuse with other i2b2-compliant repositories. The data in these fused repositories can of course use the same analytical tools as the isolated repositories or silos. Overall this enables data and tools to be collected and developed on a small scale, and then come together on a larger scale without infrastructure modification. Repeated analysis and re-analysis can be accommodated by building workflows that fit the services together

We are working with 4 separate groups to create cells for the hive outside of our own development efforts, including the Laboratory for Computer Science at the Massachusetts Institute of Technology, the DSG at Brigham and Women's Hospital, the CHIP at Children's Hospital in Boston, and the HCIL at the University of Maryland. The hive is utilized for active clinical research by groups throughout the Partners Healthcare System.

The i2b2 Hive encourages the use of e-Science [6] by promoting a framework that uses web services to support its data transformations. The e-Science program is a UK initiative that enables science through the use or the Grid and web services.

The main drawback that we have encountered with this web service architecture is that services are not always as flexible as raw SQL to query the data repositories, being limited to specific pre-specified queries and result sets. To overcome this, we are developing a second layer of SQL access to the CRC. This layer will also need to comply with the need for security and strict ontology.

The result of this architecture has been an understanding of medical data such that clinicians in our driving biology projects who were unable to use the data in its raw form have now produced several publishable outcomes in which this architecture was utilized [7]. We anticipate a more formal evaluation at a later stage of development to decide upon new architectural directions.

Acknowledgements

This work was funded by the National Institutes of Health through the NIH Roadmap for Medical Research, Grant U54LM008748. Find information on the National Centers for Biomedical Computing at http://nihroadmap.nih.gov/bioinformatics.

References

1. Murphy, S. N., Morgan, M. M., Barnett, G. O., Chueh, H. C. Optimizing Healthcare Research Data Warehouse Design through Past COSTAR Query Analysis. Proc AMIA Fall Symp. 1999; 892-6.
2. W3C Web Services Activity (2007) Retrieved March 3rd, 2007 from W3C Architectural Domain: http://www.w3.org/2002/ws/
3. HIPAA Privacy Rule, **Federal Register** / Vol. 67, No. 157 / Wednesday, August 14, 2002 / Rules and Regulations http://www.hhs.gov/ocr/hipaa/privrulepd.pdf
4. HIPAA Security Rule, **Federal Register** / Vol. 68, No. 34 / Thursday, February 20, 2003 / Rules and Regulations http://aspe.hhs.gov/admnsimp/FINAL/FR03-8334.pdf
5. Eclipse (2007) Retrieved March 3rd, 2007 from Eclipse: http://www.eclipse.org/
6. About the UK e-Science Programme (2007) Retrieved March 5th, 2007 from Research Councils UK: http://www.rcuk.ac.uk/escience/ http://www.rcuk.ac.uk/escience/hea-ex.htm
7. Informatics for Integrating Biology and the Bedside (2007) Retrieved March 15th, 2007 from: http://www.i2b2.org