MASSACHUSETTS GENERAL HOSPITAL

THE i2b2 HIVE: A SOFTWARE ARCHITECTURE FOR THE CLINICAL RESEARCH CHART

HARVARD MEDICAL SCHOOL

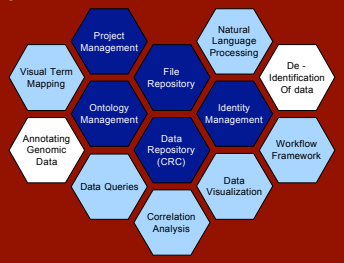
Chueh, Henry C., MD, MS[1]; Berkowicz, David A.[1], MD; Murphy, Shawn N., MD, Ph.D.[1]; and Kohane, Isaac MD, Ph.D[2].

[1] Massachusetts General Hospital, Boston, MA; [2] Children's Hospital, Boston, MA

ABSTRACT

A key goal of i2b2 is to develop methods for systematically collecting and managing clinical research data over time into a modern-day clinical research chart. As with clinical systems, software tools and applications will play an increasing large role in this activity. Unlike clinical systems, the diversity of research practice and its rapidly changing landscape suggests that any particular software suite will likely face early obsolescence. The i2b2 Hive is a software architecture intended to promote a scalable model for continuous development of software components that can service the clinical research chart.

The i2b2 Hive can be described as a collection of Web services, or "Cells", that are aggregated to perform a variety of tasks with data from the clinical research chart. Each Cell can define a functional component that can be developed independently from other Cells, and describes a public software interface minimally as XML over HTTP, and optionally supports a standard for a managed user interface. By leveraging the network for functional messaging, this approach allows free reuse of existing services to be combined into new functional services and Cells -- not unlike the "mash up" programming that is emerging autonomously on the Web. The i2b2 team is developing reference implementations of cells to handle core functions such as repository, workflow, identity, project, and ontology management.



The implementation of an i2b2 Cell is intentionally simple since the i2b2 Hive is broadly inclusive of diverse software development methodologies, languages, and most importantly, researchers who are not expert software engineers but who have developed valuable software services. The i2b2 Cell specification defines basic mechanisms to support authentication, session information, simple routing, and callback notifications, all as part of the i2b2 XML schema for messaging. Support for managing transmission of relationships between de-identified patients and their related clinical research data is also included.

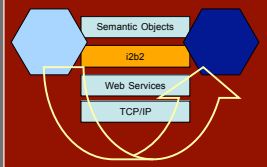
This work was funded by the NIH Roadmap for Medical Research, Grant U54LM008748.

Overview and Goals

The overarching goal of i2b2 informatics is to design and implement the i2b2 Clinical Research Chart (CRC). The CRC has many aspects, some specifically practical and functional, and others intended to establish an open architectural software framework that can support unanticipated future functionality. The i2b2 "Hive" is the software engineering architecture being designed to implement the CRC. The Hive is a collection of software services placed under the control of clinical investigators, bioinformaticians, and software developers. From an architectural perspective, i2b2 takes an approach based on using standardized, lightweight XML-HTTP services to deliver biocomputing resources. This approach supports the vision of expanding the landscape for "usual" biocomputing beyond the desktop and server to the Internet. A low barrier for entry into integrating biocomputing resources in the i2b2 model is intentional, to allow for unanticipated and novel combinations of pre-existing software. From a functional perspective, identifying and developing these services critical to managing clinical research data in a sustainable fashion can ultimately provide clinical researchers a head start on projects, because the CRC will provide an organized repository and platform to extract and analyze their data.

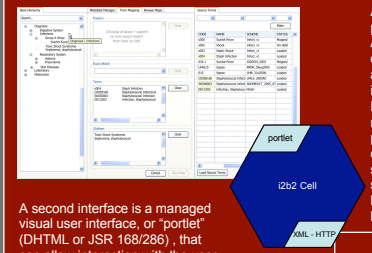
One Hive, Many Cells

An i2b2 Hive is formed as a logical collection of interoperable services provided by i2b2 Cells. This creates a loosely coupled system that makes no assumptions about physical proximity because interactions between Cells are performed via Web services.



Within the Hive, Cells behave as a functional services that can be called on demand. The approach of i2b2 separates the interactions between these functional services into two categories conceptually: 1) transaction semantics, and 2) functional semantics. i2b2 protocols and specifications are focused on facilitating transaction semantics -- resolving issues such as common data types. The deeper functional semantics, such as the behavior of data/objects, are left independent of the i2b2 architecture, to be defined freely by the developers of Cells. Users of an i2b2 Hive will need to make decisions about which Cells fulfill desired functions with the appropriate semantics, but connectivity within the Hive can be taken for granted.

Multiple Faces of a Cell



An i2b2 Cell must expose one or more public interfaces for access to its services. A lightweight messaging interface (XML over HTTP) is the minimum standard that allows other i2b2 Cells to interoperate with it in an automated way. For example, the Metadata Cell will allow the addition of a new concept to the ontology through a public Web service. These messages consist of a combination of i2b2 data schemas as well as Cell-specific data schemas. Use of XML HTTP messaging helps to avoid restricting the programming language used to implement a Cell.

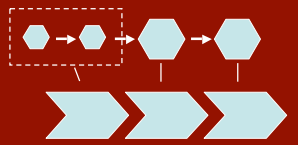
A second interface is a managed visual user interface, or "portlet" (DHTML or JSR 168/286) , that can allow interaction with the user, but within the control of a managing container or "portal". Other optional interface specifications can be added in the future – making for a flexible, evolving architecture.

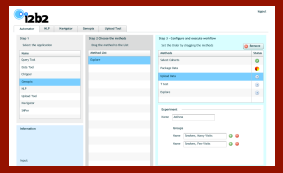
addBaseConcept ( )

```
<i2b2_message>
 <version>1.09</version>
 ...
 <body>
  <new_base>
   <base_cd>S0123</base_cd>
   <scheme_cd>BASE-V1</scheme_cd>
   <name_char>Group A Strep</name_char>
   <path_char>/diagnosis/infections/Group A Strep</path_char>
   <comment>Streptococcus pyogenes</comment>
  <new_base>
   ...
```

Workflow

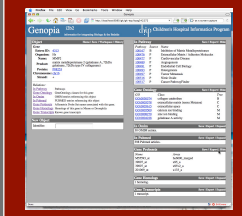
An i2b2 Hive is automated by a Workflow Cell that directs the choreography among Cells for complex interactions. Steps in a workflow may be as basic as the invocation of a single service of one Cell, or can be a complex interaction involving many Cells. Tools for workflow creation will address both informatics integrators as well as investigators.





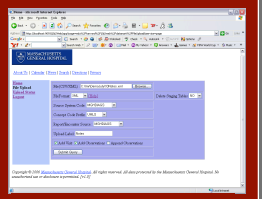
A prototype of the Automator interface to the workflow engine is shown above. The goal of the Automator is to allow complex workflow and computation to be scripted and saved through simple, incremental steps. As long as a Cell adheres to the lightweight i2b2 protocols, its public interfaces can be exposed through the Automator interface.

i2b2 Cell Sampler



Genopia

Genopia starts with a gene, probe set, or collection of either, and allows the user to explore public databases through data relationships. Items of interest can be retrieved to be saved independently from Genopia.



Upload Tool

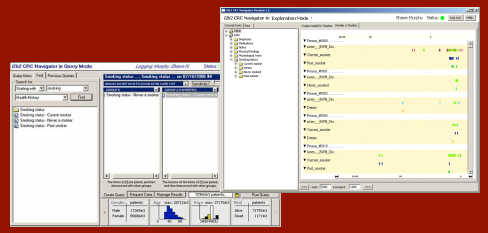
The upload tool provides a mechanism to perform bulk data loading of the Clinical Research Chart. This can be done through its user interface and also using its Web-based XML services.

Chipper

Services allow the definition of a microarray experiment, upload of experimental data, and statistical analysis of differences in gene expression.

NLP

Services support concept extraction from clinical narratives to help annotate phenotypes in the Clinical Research Chart.



Navigator

The Navigator allows a user to query the Clinical Research Chart for patient populations and then select specific phenotypic observations to plot in a timeline visualization.