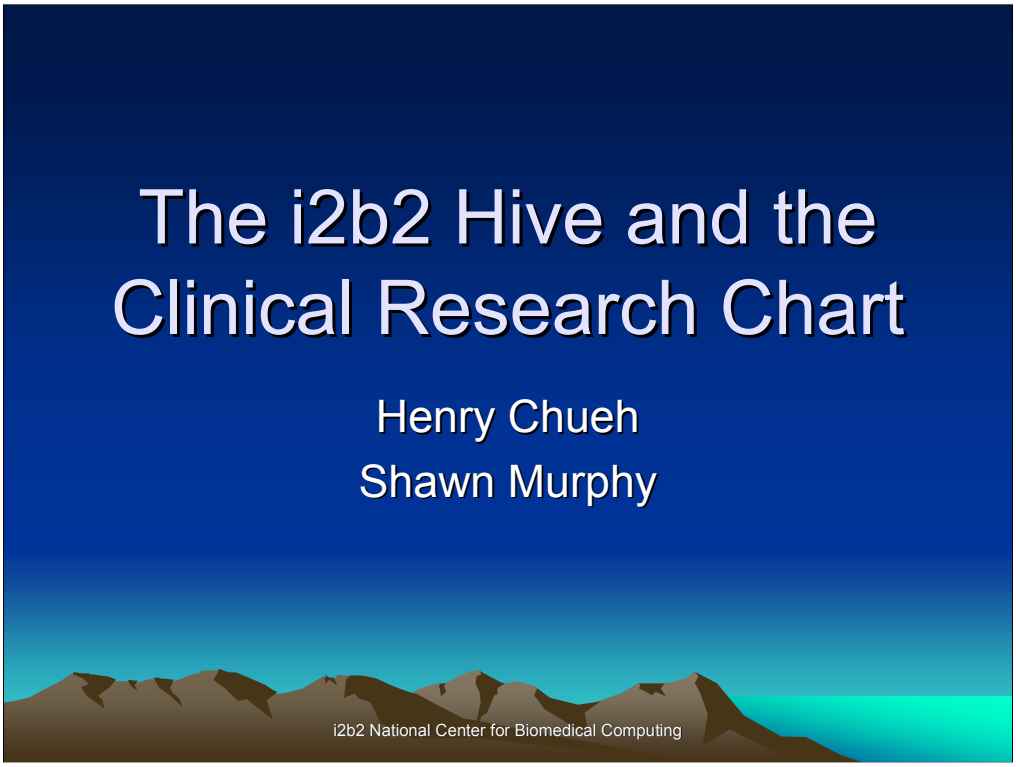
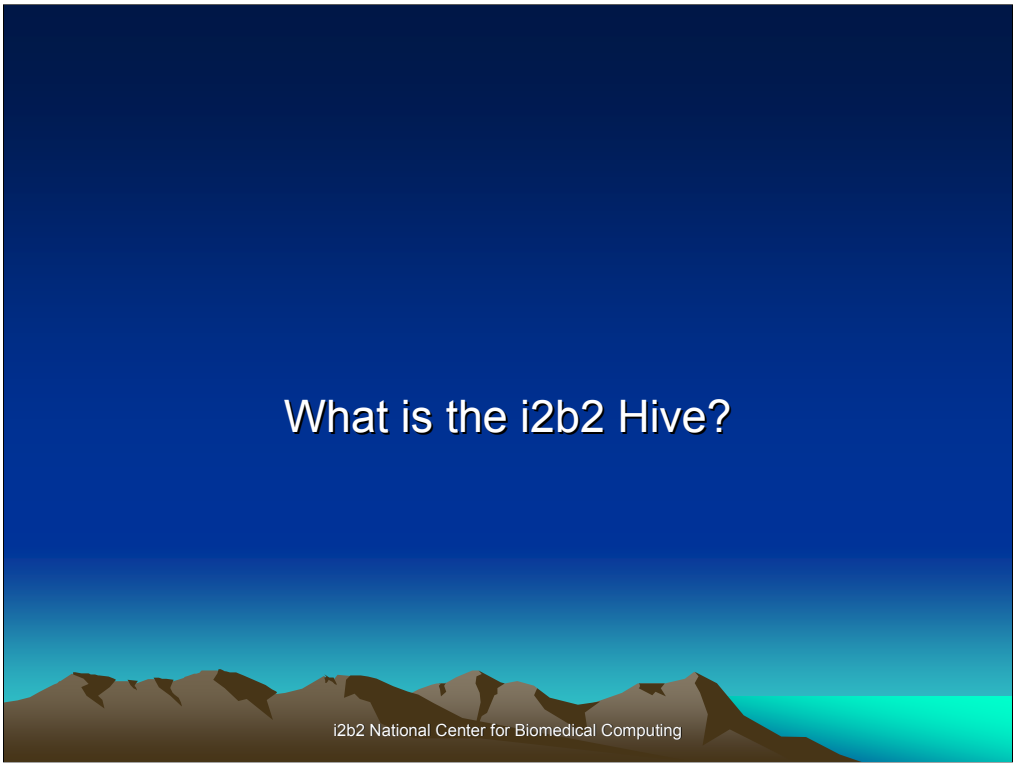
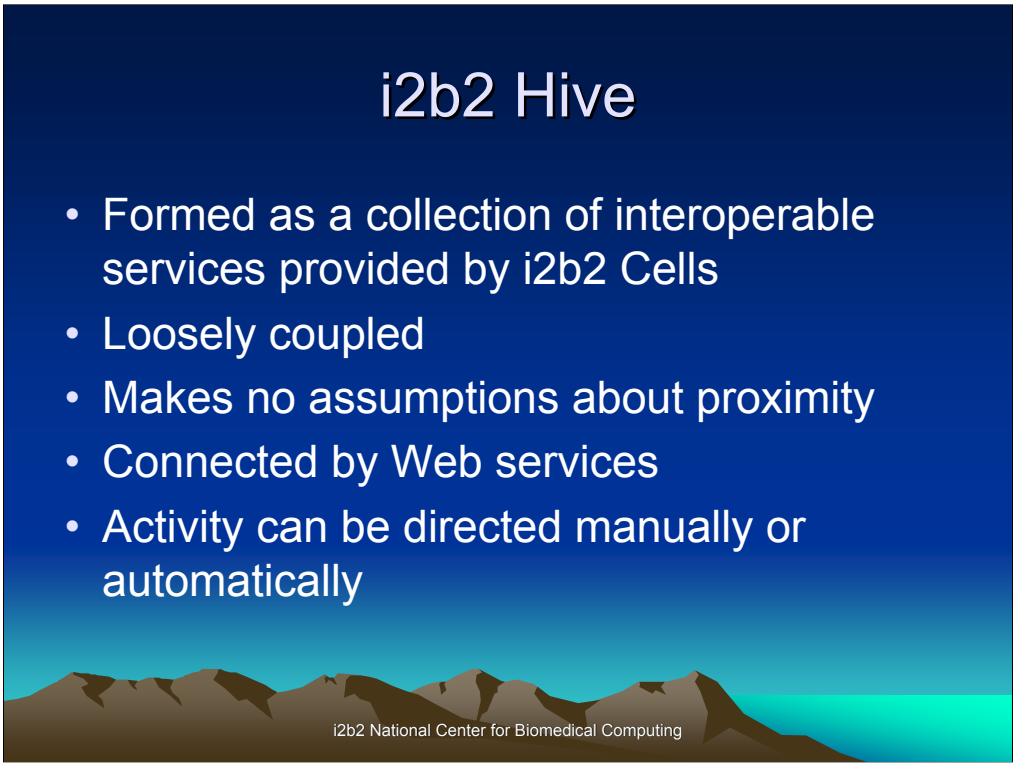


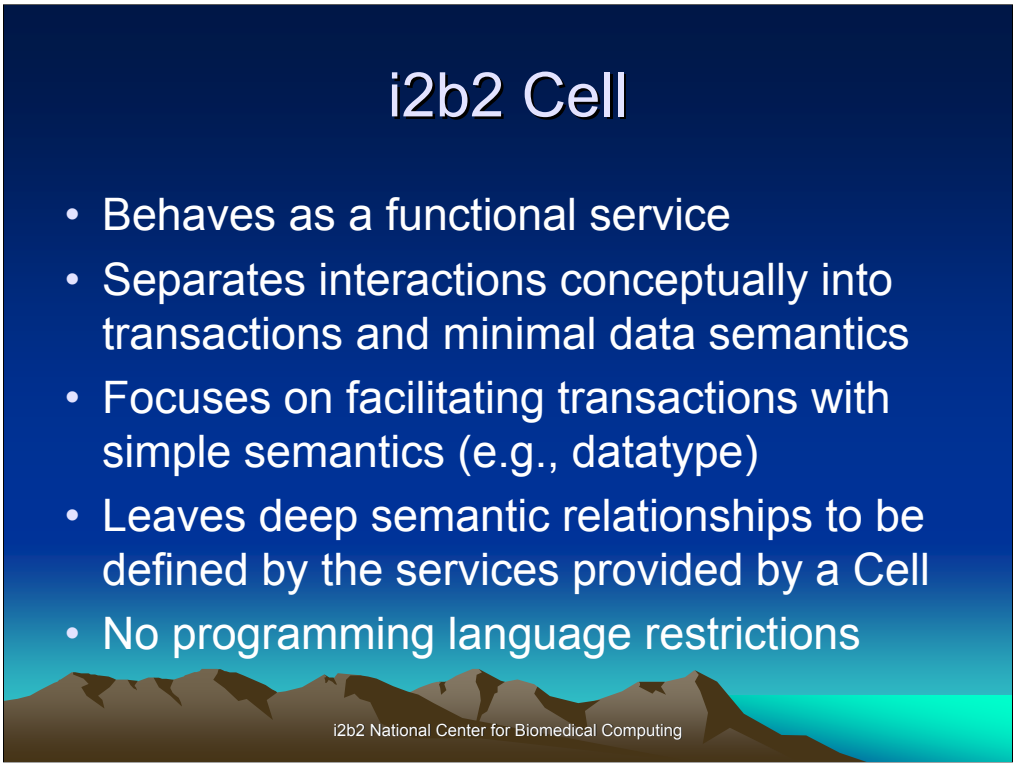
The i2b2 Hive is centered around two concepts. The first concept is the existence of services provided by applications that are “wrapped” into functional units, such that their functionality are exposed as messages that travel to and from the various cells of the hive. The second concept is that of persistent data storage, which is managed by the cell named the “Clinical Research Chart”. This presentation describes the concepts behind the Clinical Research Chart, which serves as the data repository for the Hive.



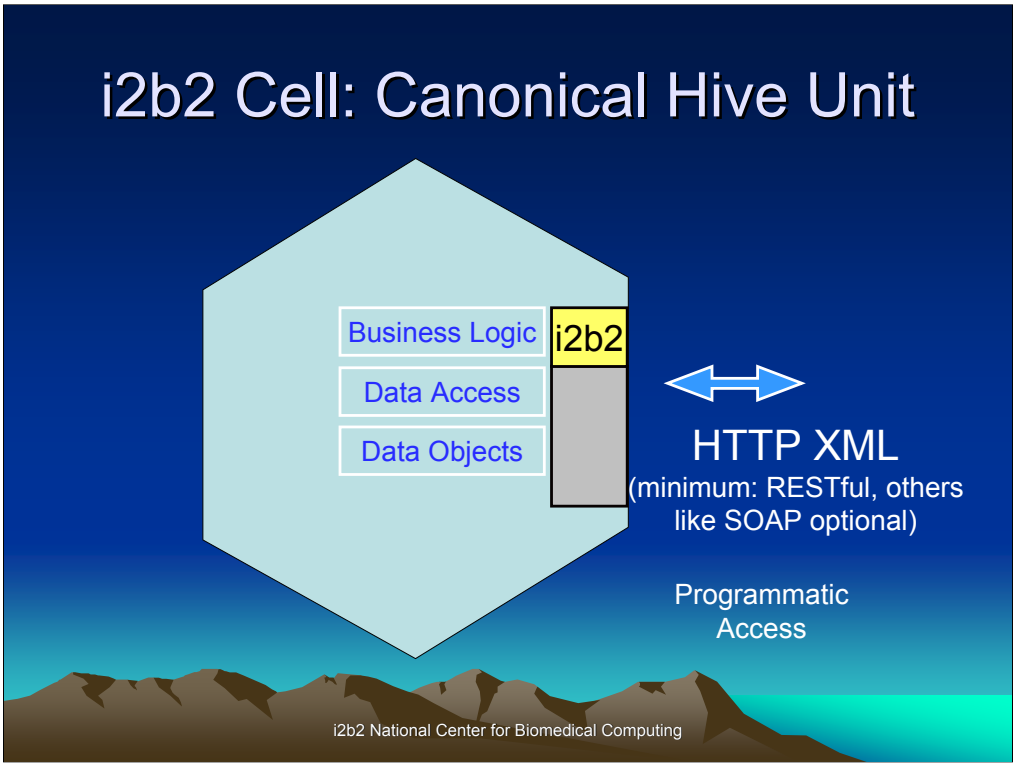
We will begin with discussing the rational behind the creation of an i2b2 Hive.



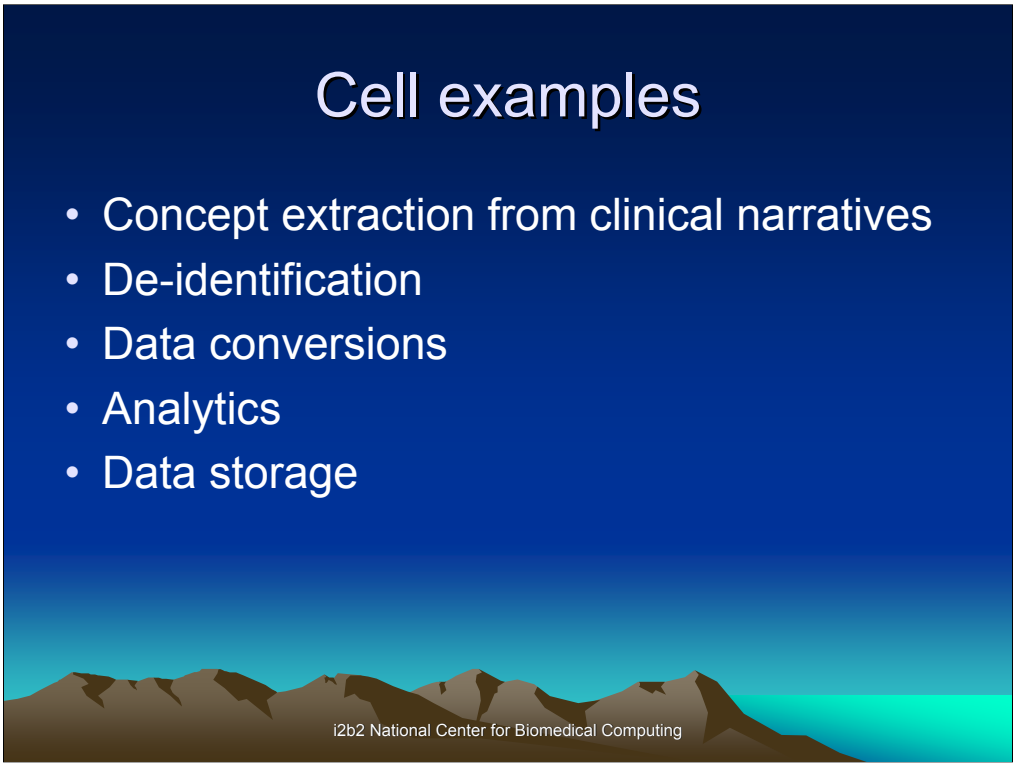
The Hive is a collection of interoperable services provided by i2b2 Cells. As a collection they are loosely coupled and generally do not know about each other, including their relative locality. Instead activity in the Hive is directed through the use of Web services that are invoked manually by the user through a specific user interface, or automatically by workflow engines.



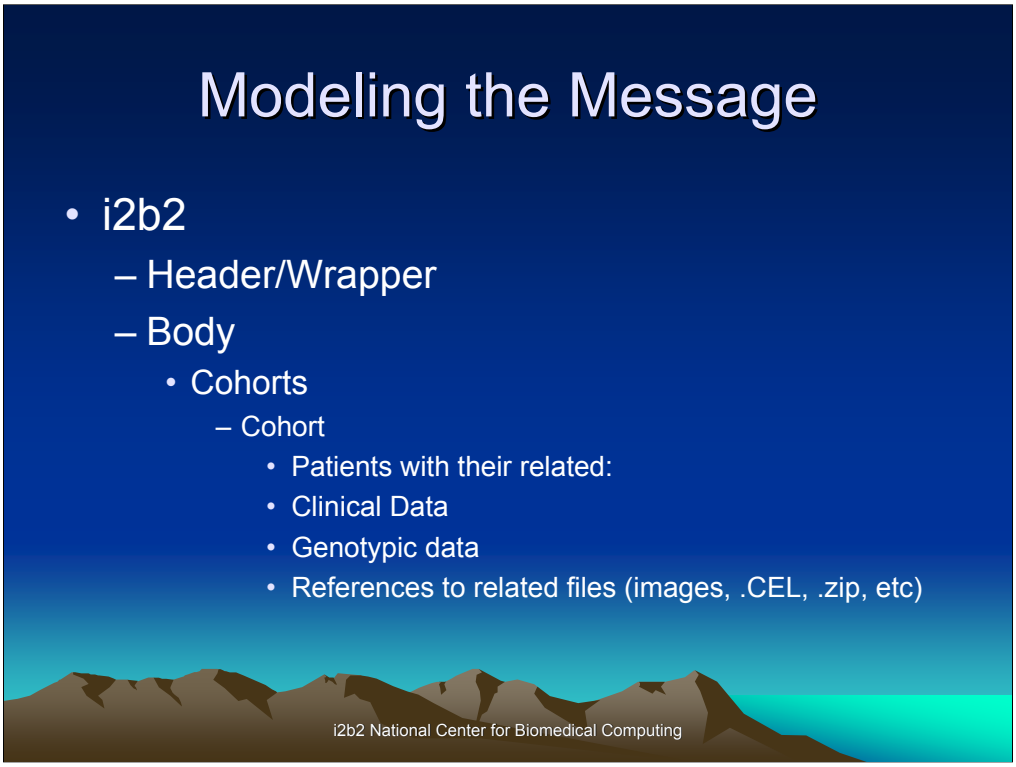
An i2b2 Cell can be considered a functional service with two parts. The transactional component is explicit and defines only minimal data semantics, making this communication straightforward. The deeper semantics that may describe relationships between objects and data are left to be defined by the Cell services and interpreted by the user of those services. Because the interface is a Web service, there are no language restrictions for creating a Cell.



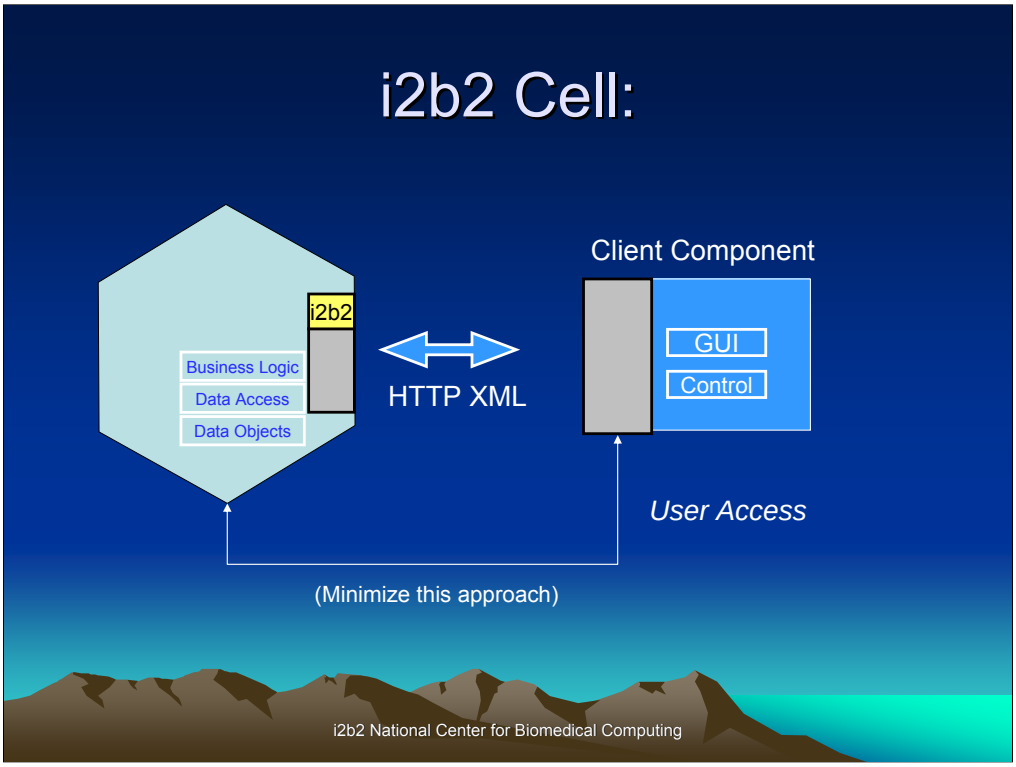
The i2b2 Cell is the basic building block of an i2b2 environment, and encapsulates business logic as well as access to data objects behind standard Web interfaces. These may be as simple as XML or RESTful services, or SOAP.



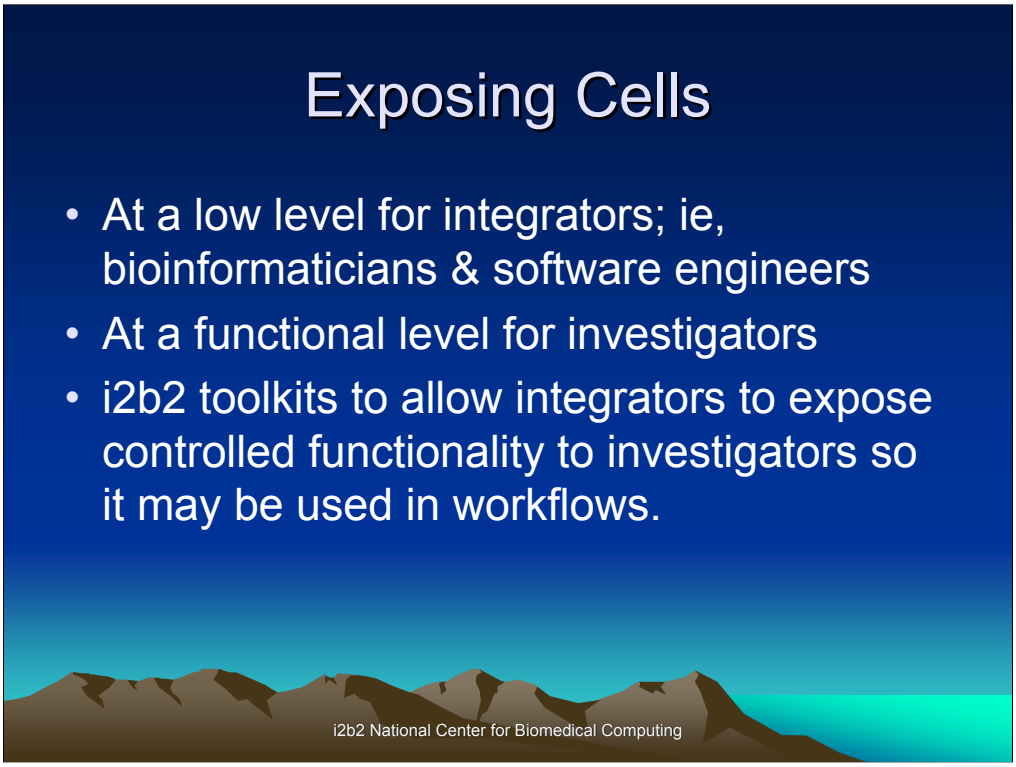
Some examples of cells are noted above, and range from repository services to basic data conversion. There is no restriction on how simple or complex a service a Cell can provide.



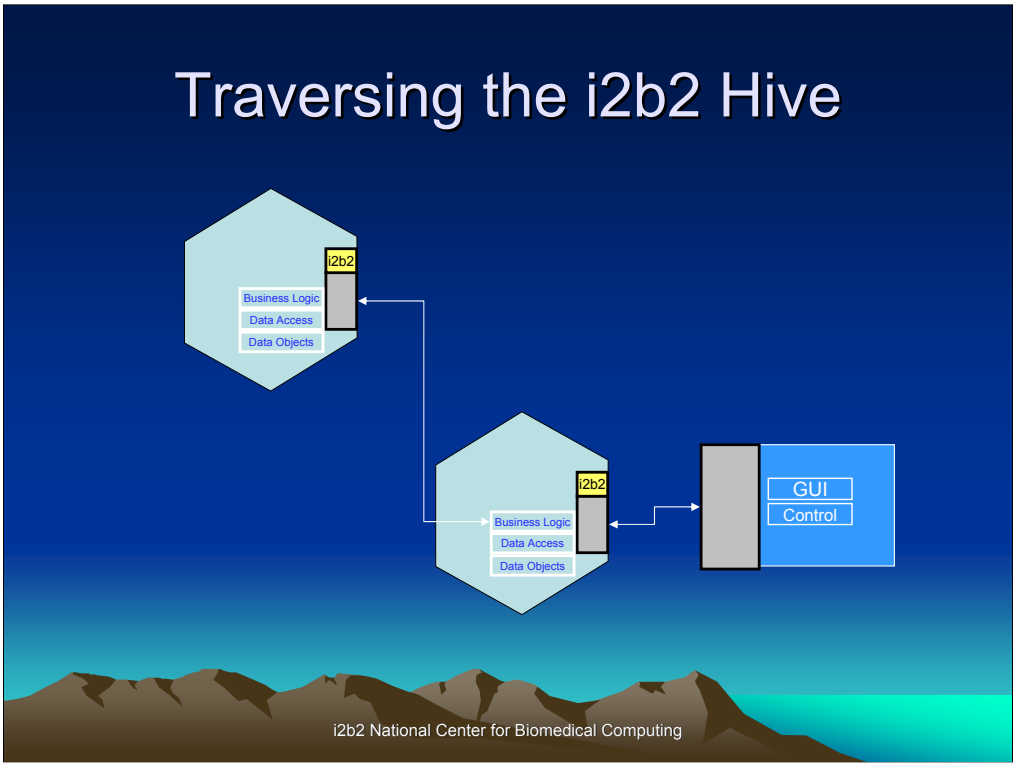
The overall model for an i2b2 compliant message is an XML schema that defines a header or wrapper for management of the basic communication, and then a message body that contains patient sets with their related clinical phenotypic and genotypic data as well as references to other data objects.



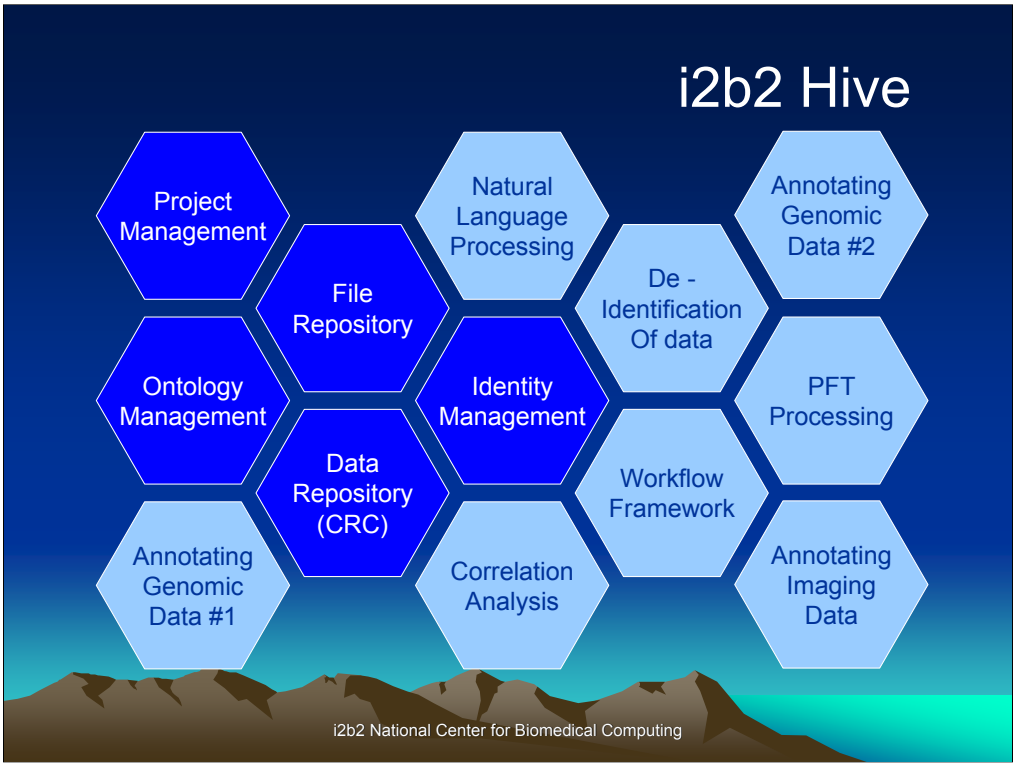
Many i2b2 Cells will have one or more corresponding visible client components that a user can interact with directly. These clients will often be created by the Cell developer, but should utilize the public Web services interface to access the Cell, rather than any private communication mechanisms. There may be some situations where the latter will be unavoidable.



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. A intermediate role may be the most important: those integrators who can construct applications on top of i2b2 Cells to create domain-specific workflows.

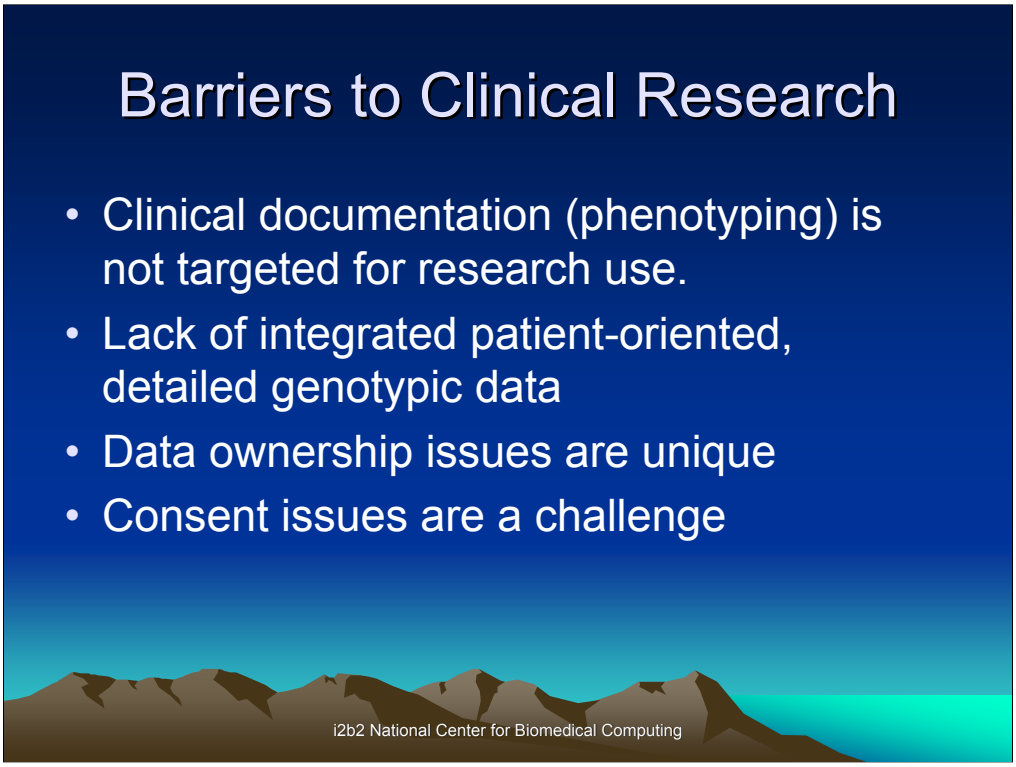


Cells may invoke other Cells. This means that a developer can independently create complex behavior and user interfaces to “wrapper” the functionality of existing Cells.

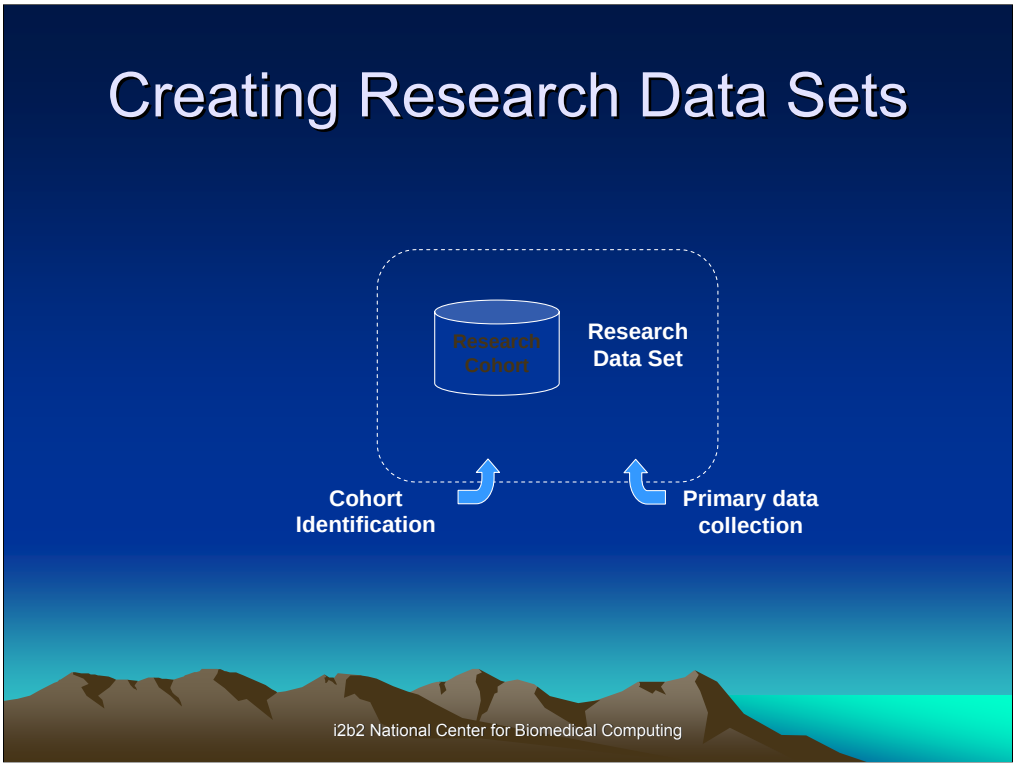


The i2b2 Hive, then, consists of a number of core Cells that establish basic services, as well as any number of additional Cells to provide enhanced services. It is intended to be a scalable approach for managing an increasing number of independently developed software services for clinical research.

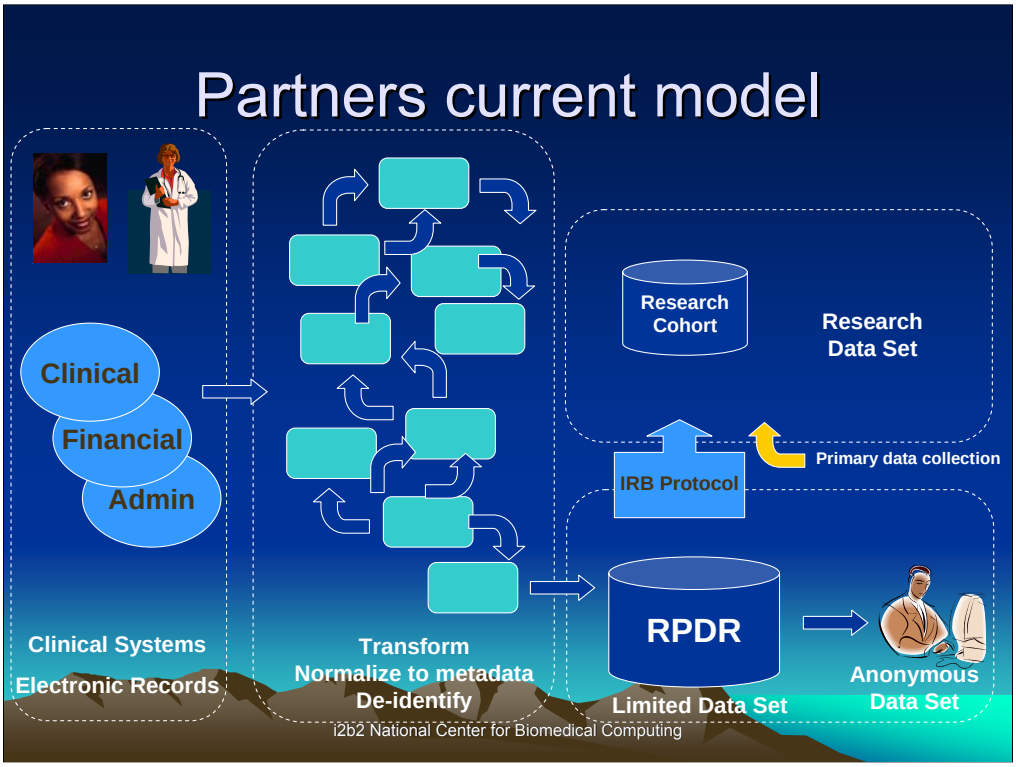
We will continue by discussing the use case for creating a special cell to hold clinical and research data



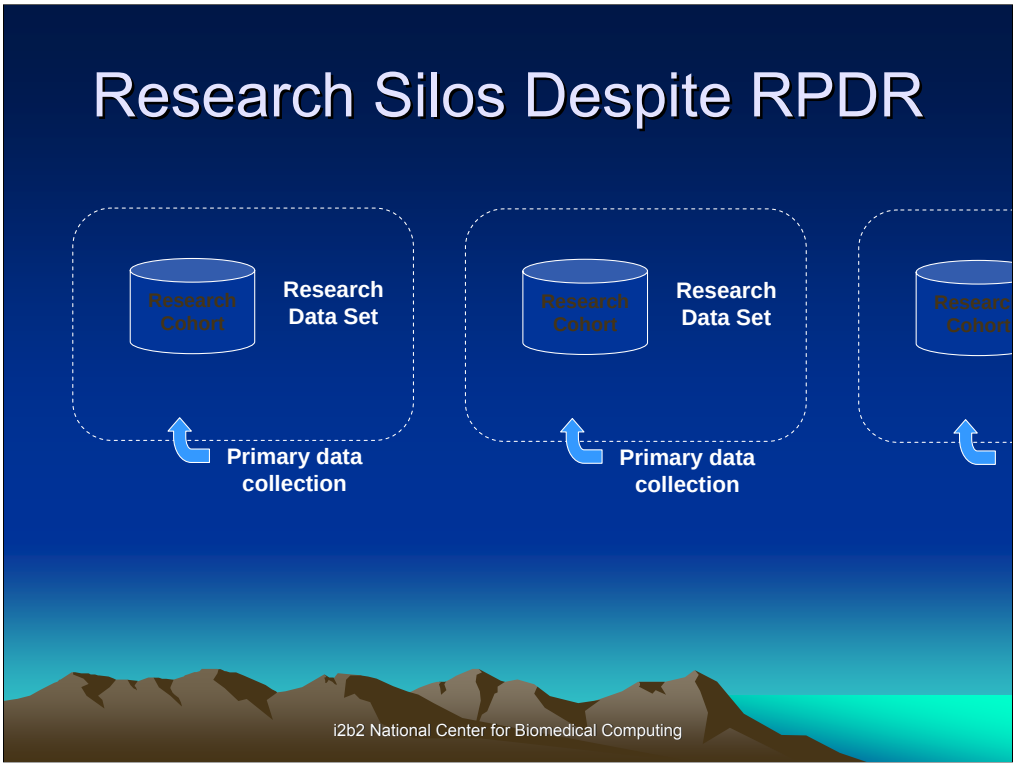
The complexity of raw clinical data is too high for it to be used easily in research. Most genetic data is available by experiment, not by orientation around a patient. Data goes through a cycle of ownership for exclusive use, during which it is not considered for sharing. Managing the consents associated with various data is challenging.



Data for research is created in small sets to achieve the goals of the specific research study.

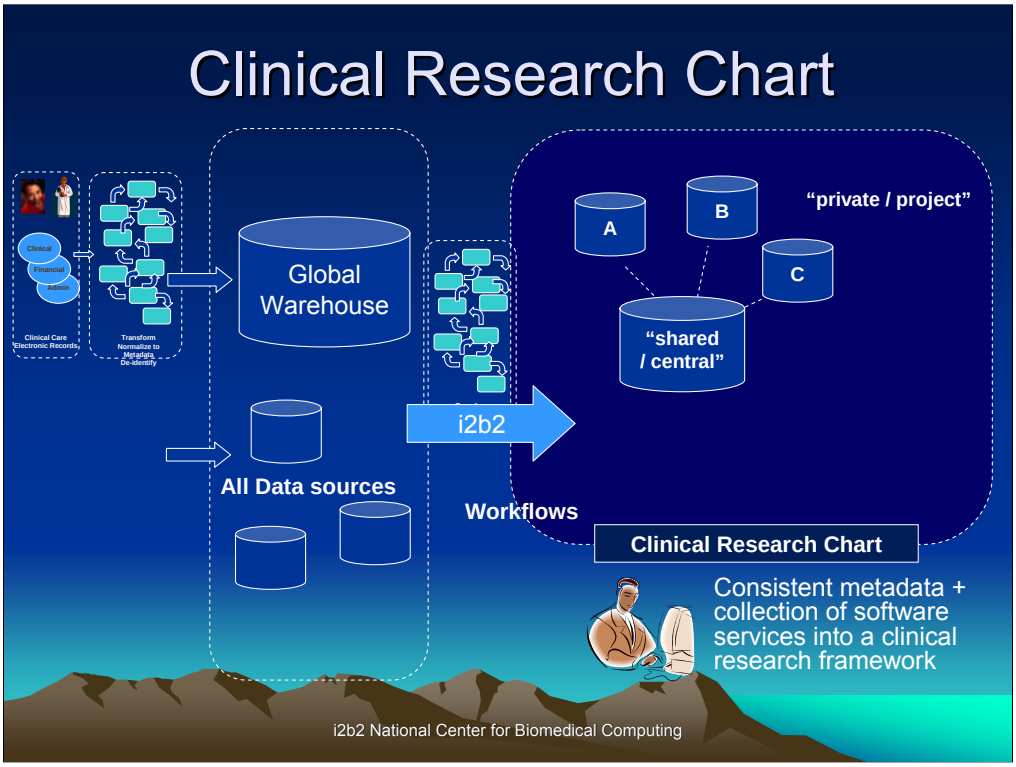


At institutions such as Partners Healthcare, the data is assembled from large amounts of clinical data.

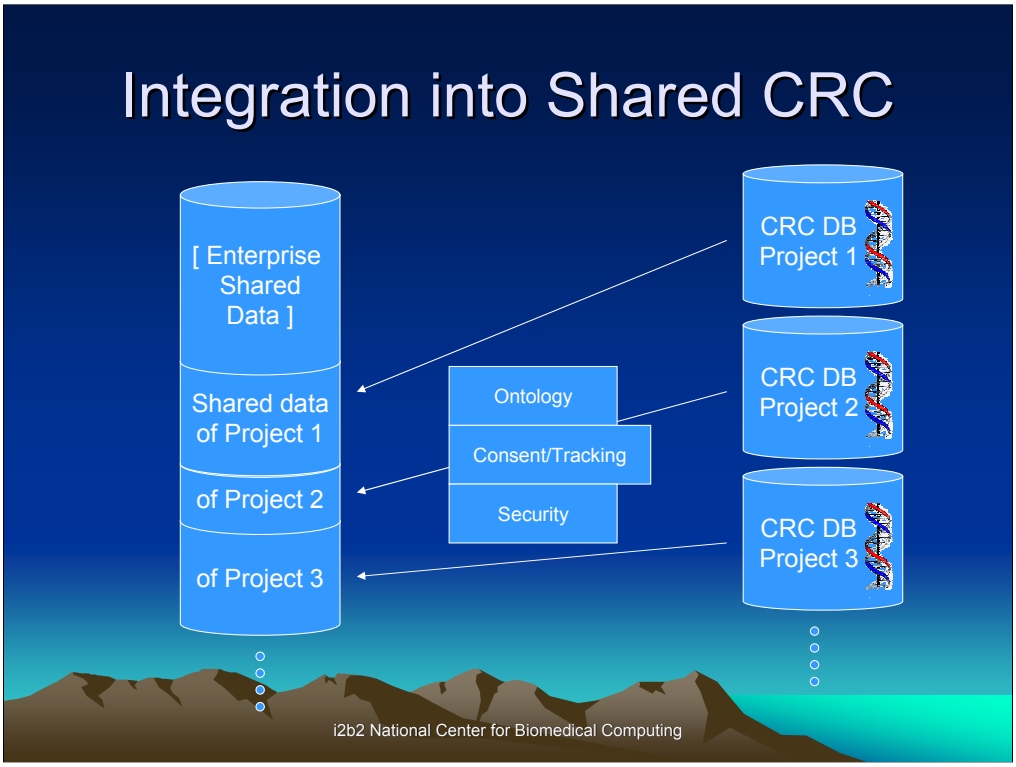


Once assembled, a considerable amount of data cleaning and integrity checking occurs, but this curated data remains in silos because the data formats are now different in each silo.

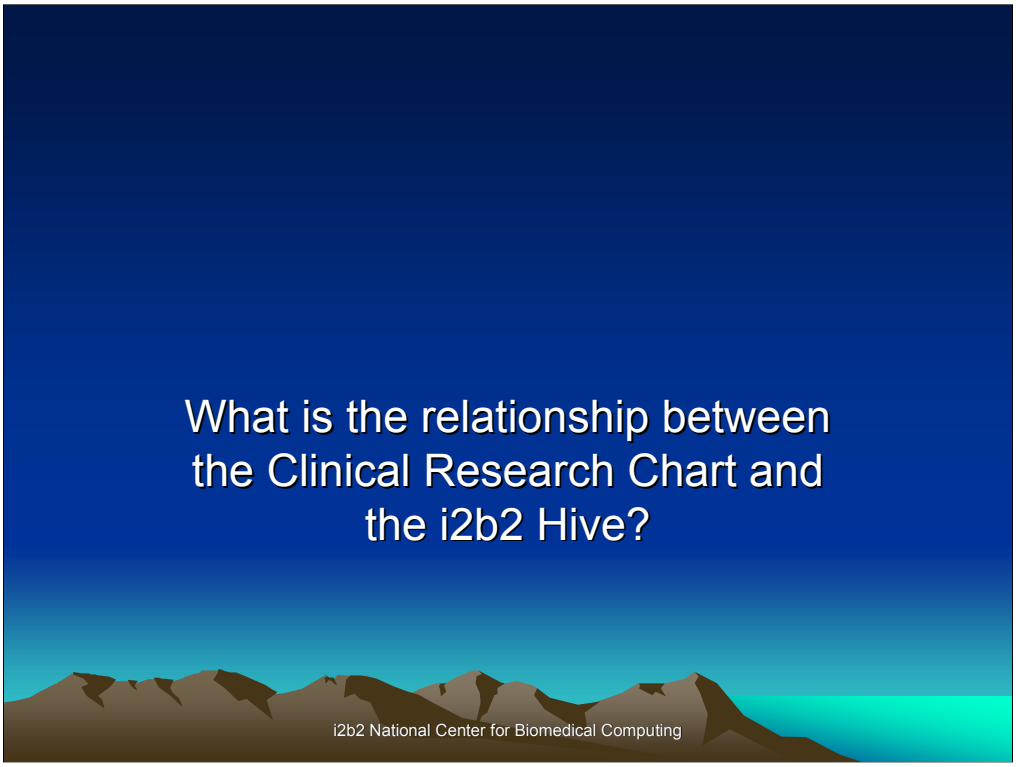
We will now discuss some of the high-level concepts of the Clinical Research Chart



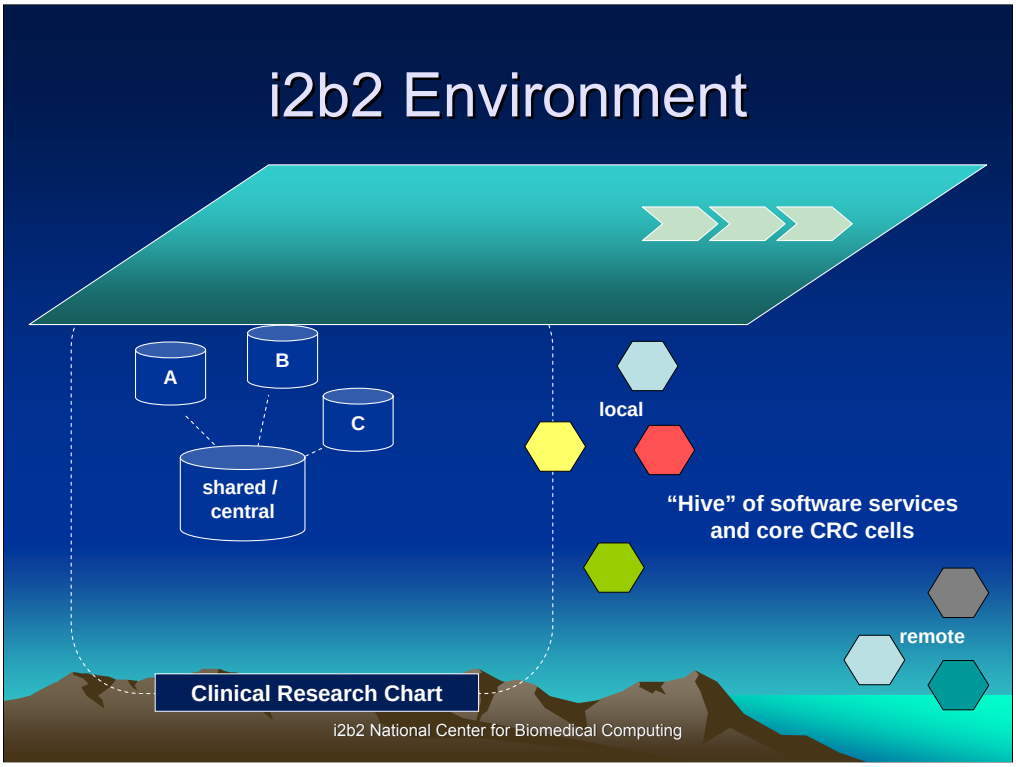
Fundamentally, the clinical research chart (CRC) is built to hold medical data. The cells of the i2b2 Hive contribute to placing the data into the CRC, which ultimately occurs by sending messages to the CRC. Even at the stage of simply being multiple stand-alone CRC's (A, B, and C) they share metadata and structure with a consistent data model.



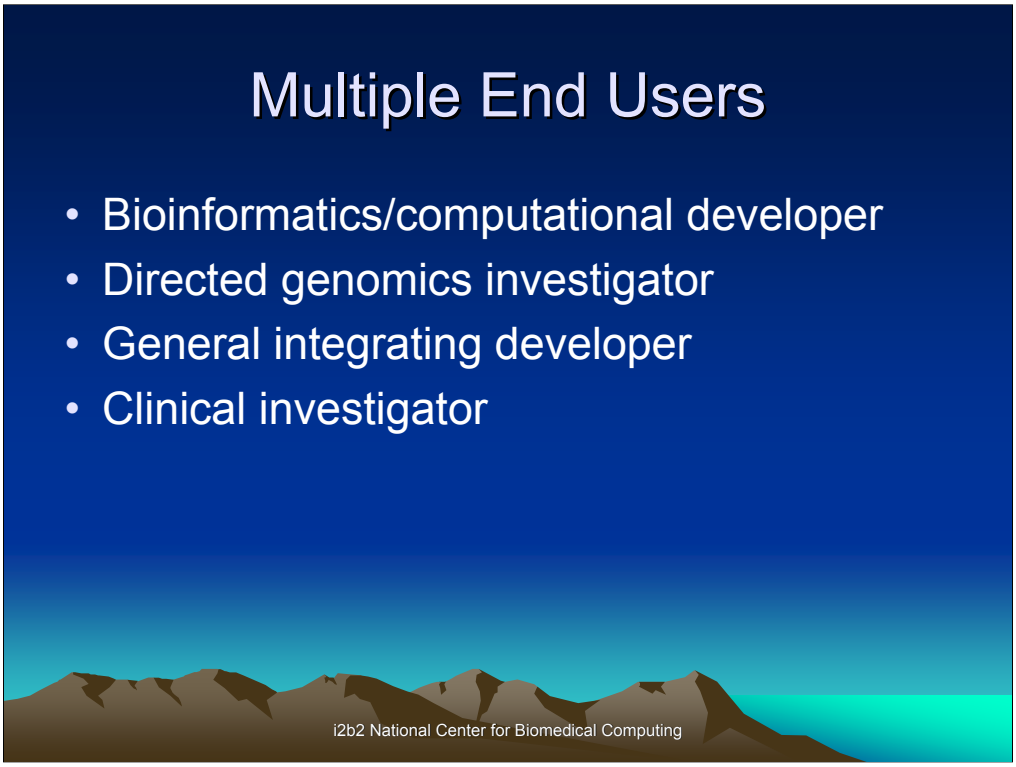
Since during their creation and management they shared a common architecture, they can be joined together at any point to created a large shared CRC.



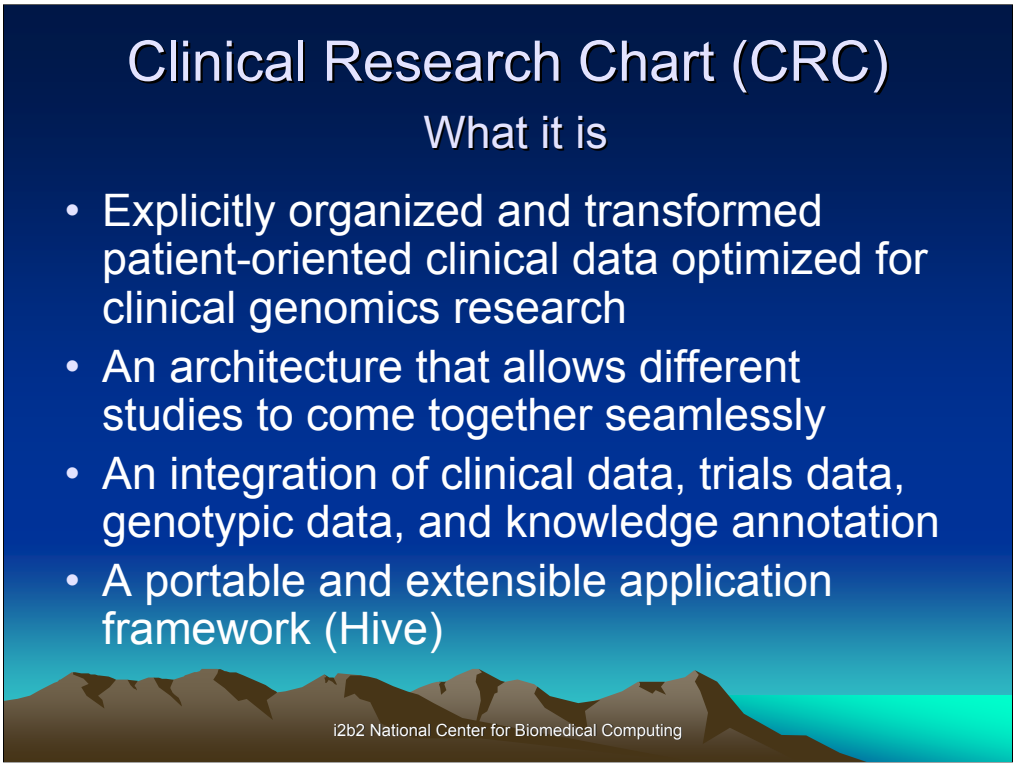
The CRC is a cell in the i2b2 Hive. It needs all of the core cells to gain complete functionality.



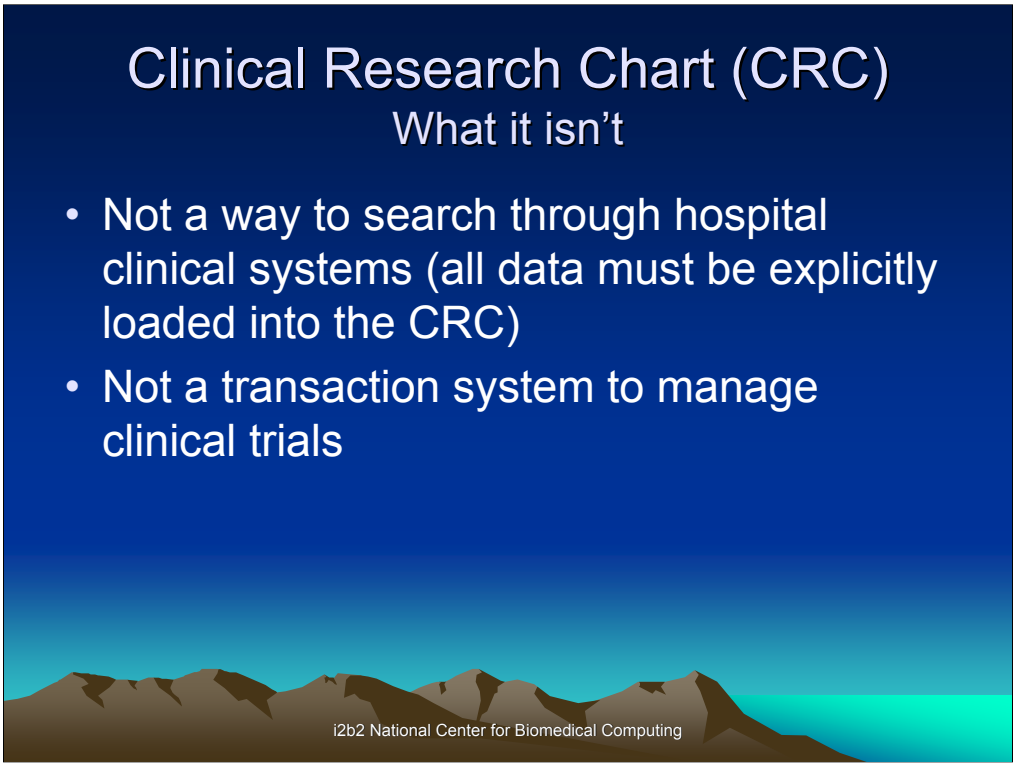
However, it is relatively decoupled from other cells, such that various cells that it depends upon may be replaced by locally built cells. For example, loading occurs through the Identity management cell, and the Ontology management cell is necessary to check codes as they put into the repository.



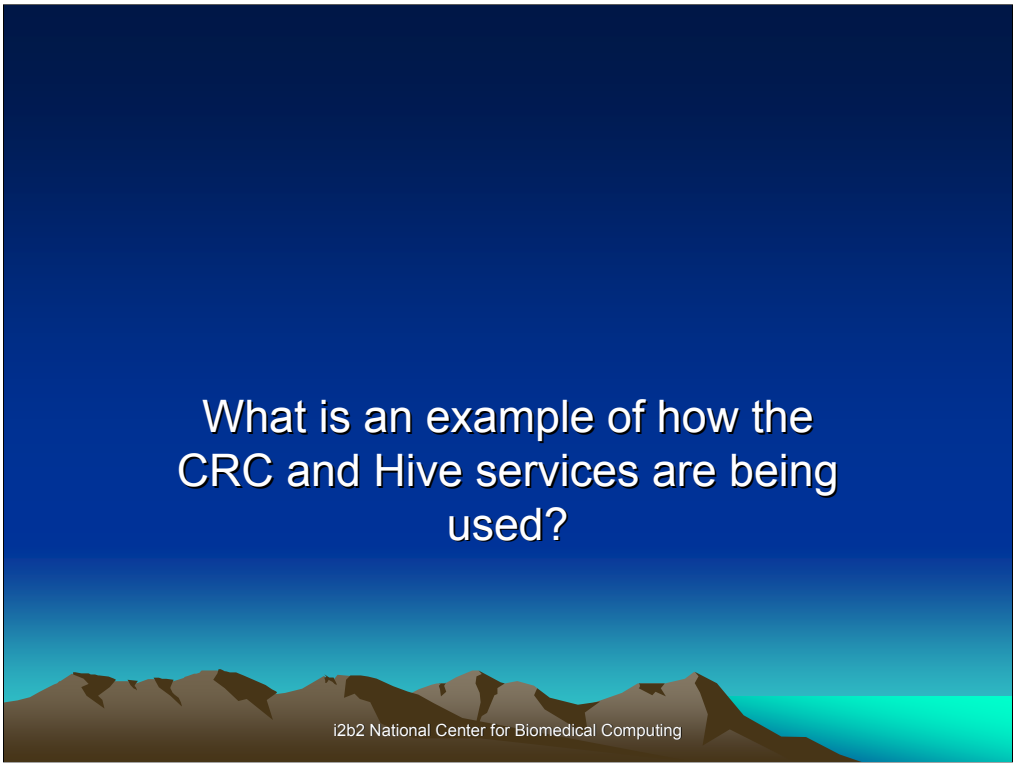
The CRC targets several kinds of end users.



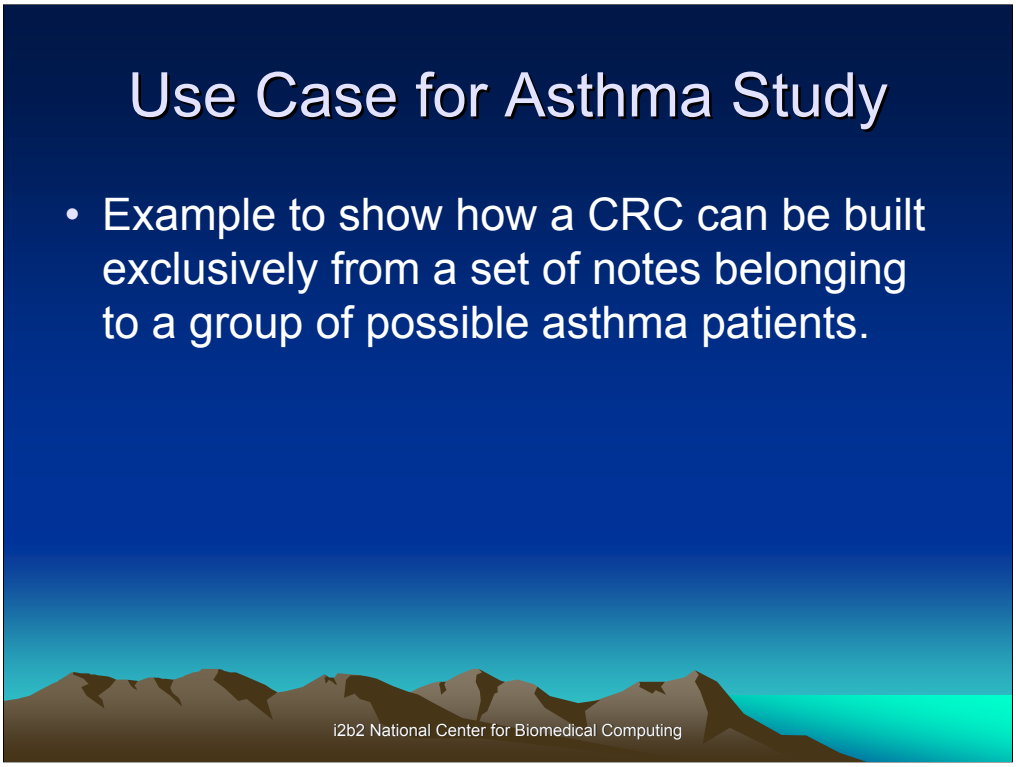
The CRC functions as the integrated data repository for the i2b2 Hive



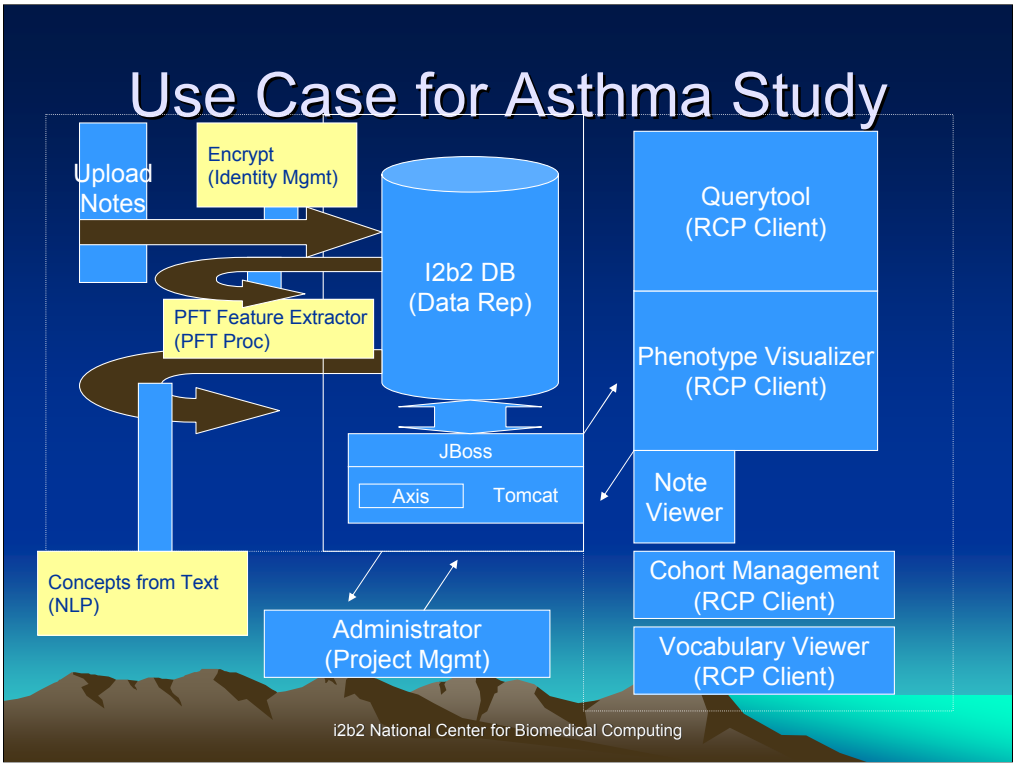
This is functionality that the core cells of the i2b2 Hive do not support.



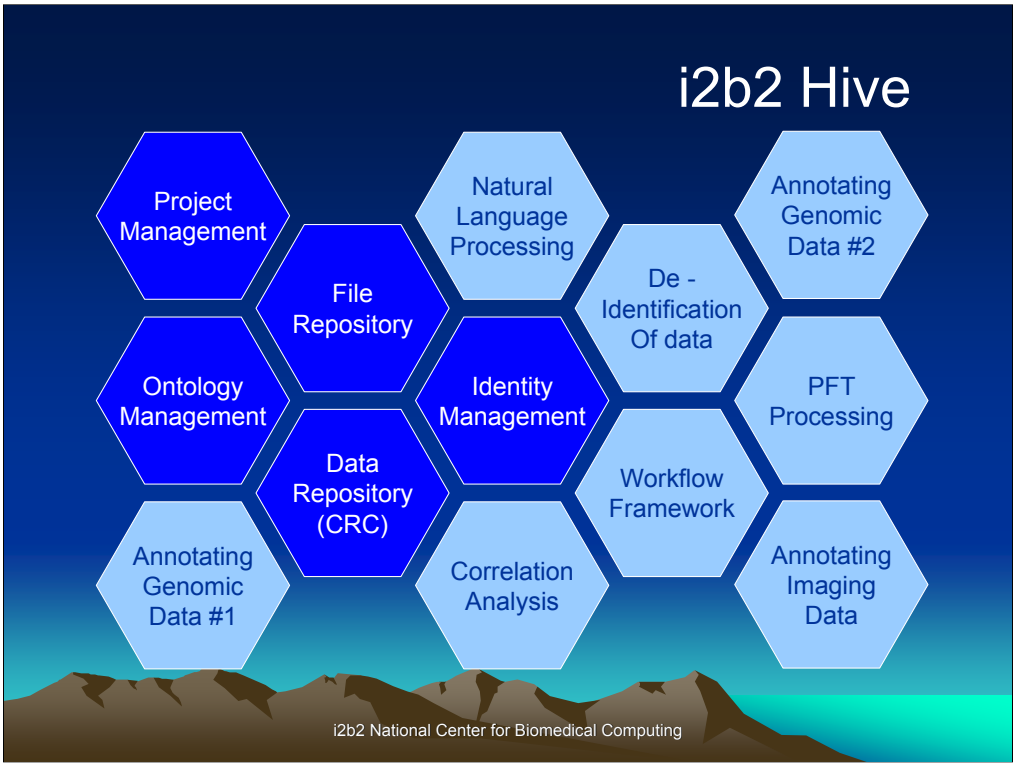
Some of the functionality of the CRC is best explained within an example clinical investigation.



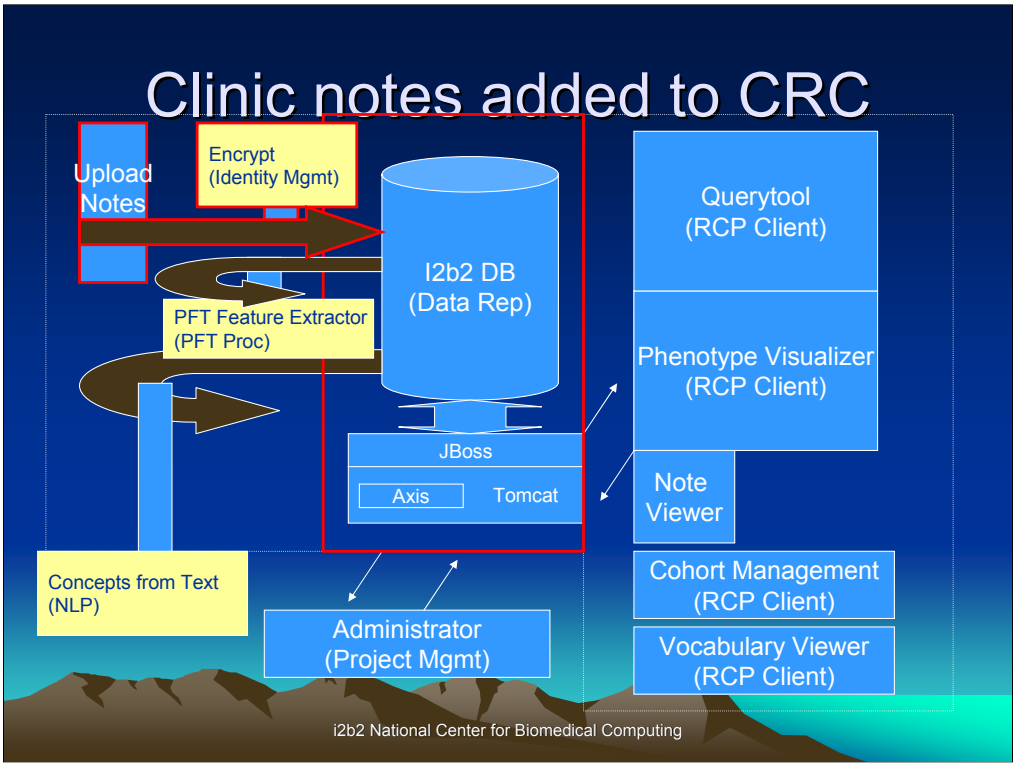
In this investigation of Asthma patients we have only notes from a clinic available. They are going to be processed through the Hive into specific concepts associated with patients, and the concepts will be placed in the CRC.



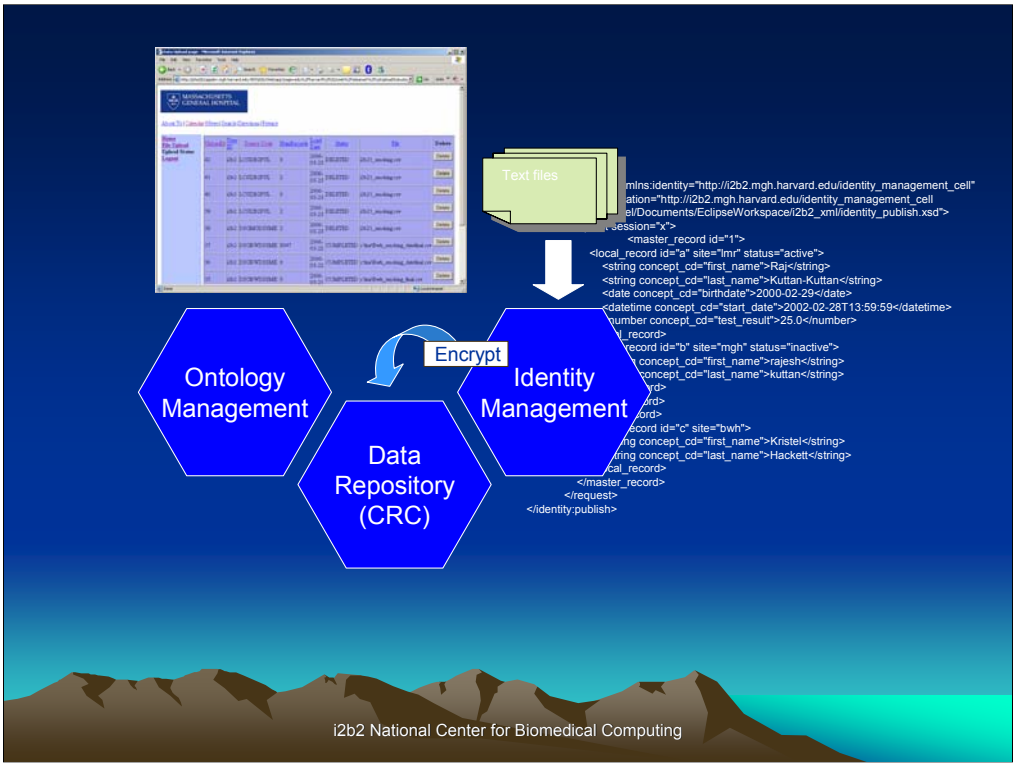
The data flow for the asthma study is diagramed above.



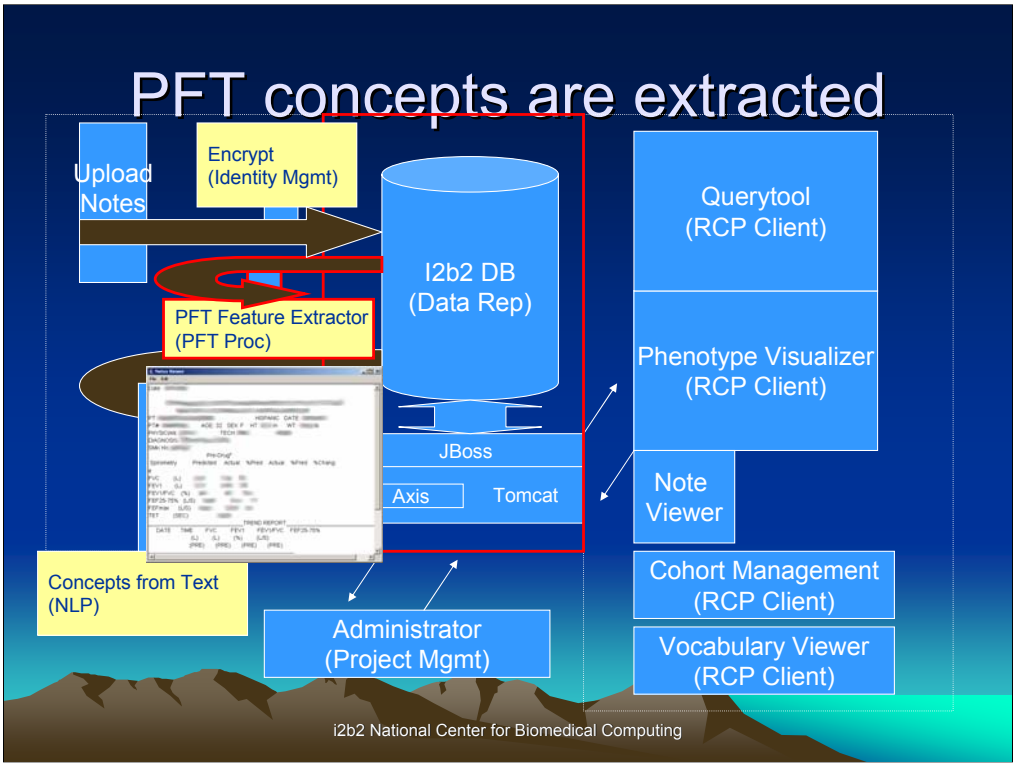
The data will flow through the cells of the hive in order to be placed into the CRC.



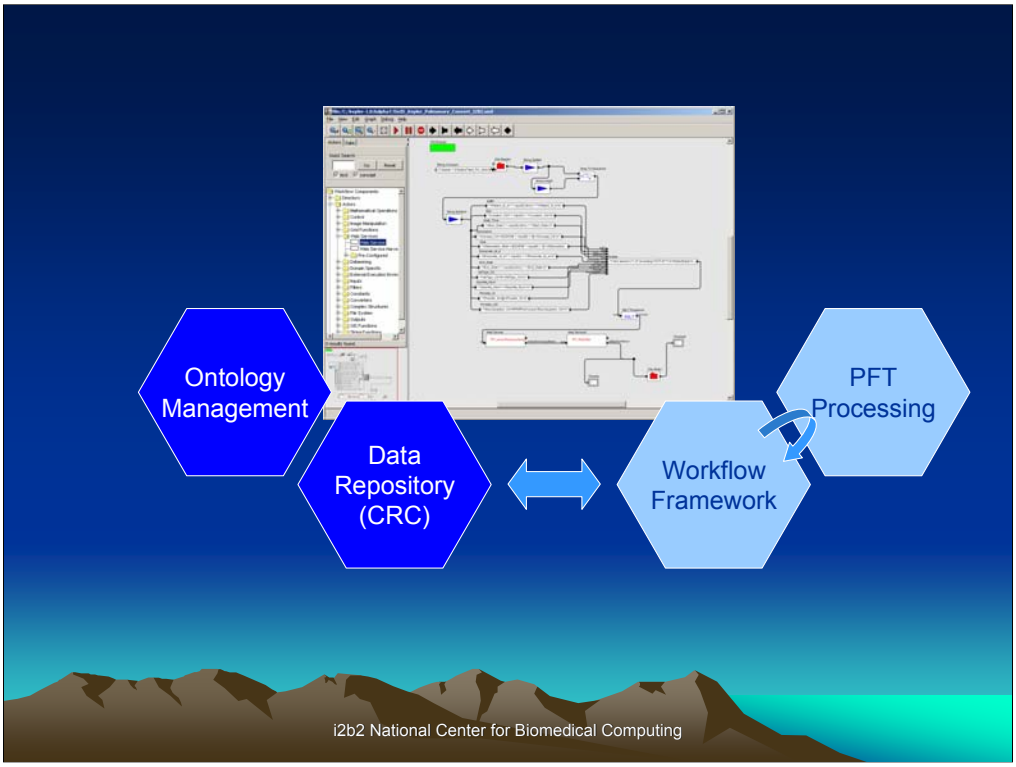
The clinic notes are to be added to the CRC.



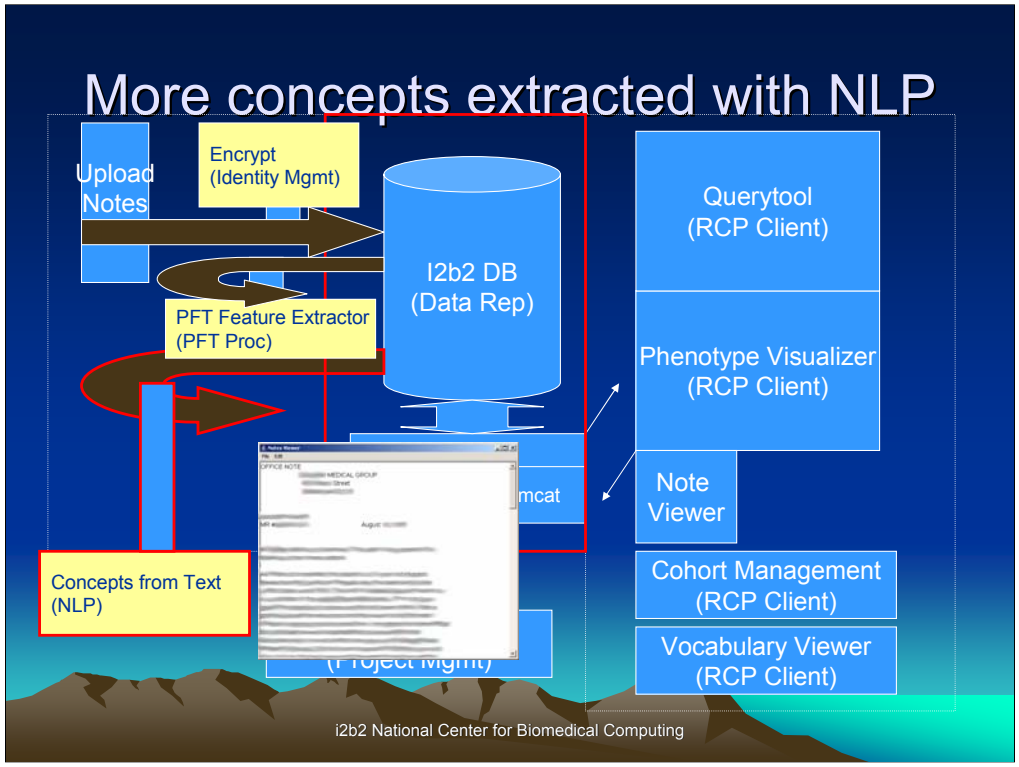
The clinic notes are added through the uploader that is part of the Identity management cell. The names and medical record numbers are resolved and retained in the Identity management cell, from where coded information is fed to the CRC. Notes are added to the CRC in an encrypted format. This preserves the CRC as a HIPAA defined limited data set.



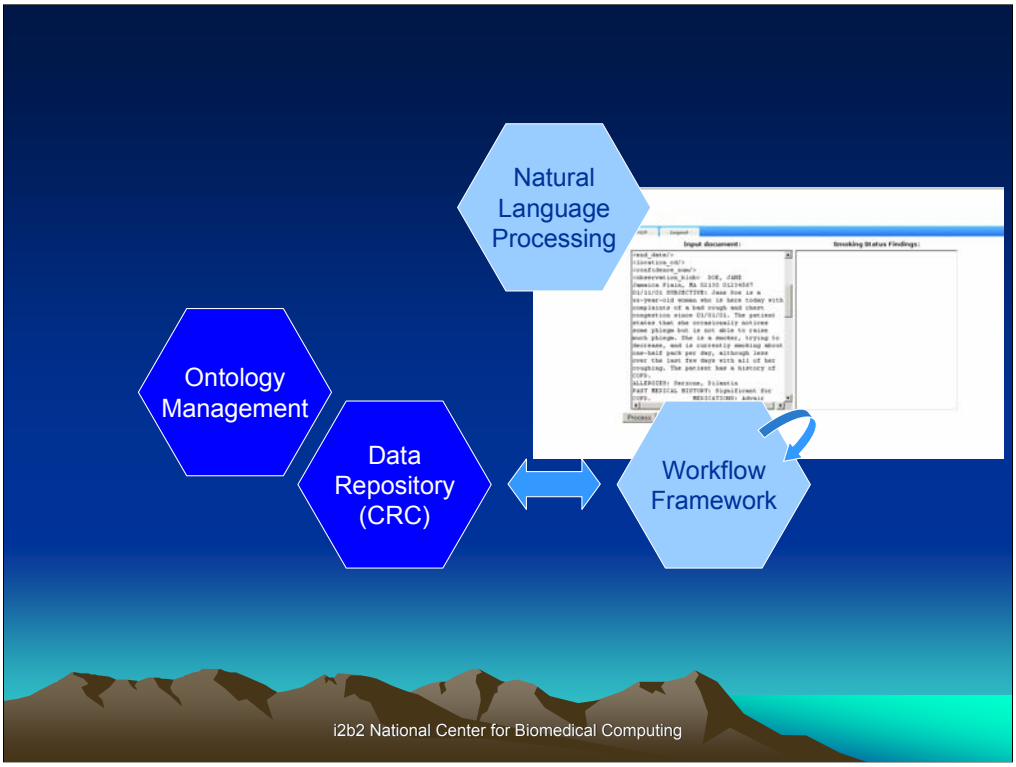
Values are extracted from the notes associated with pulmonary function tests (PFTs) and placed back into the CRC as individual concepts.



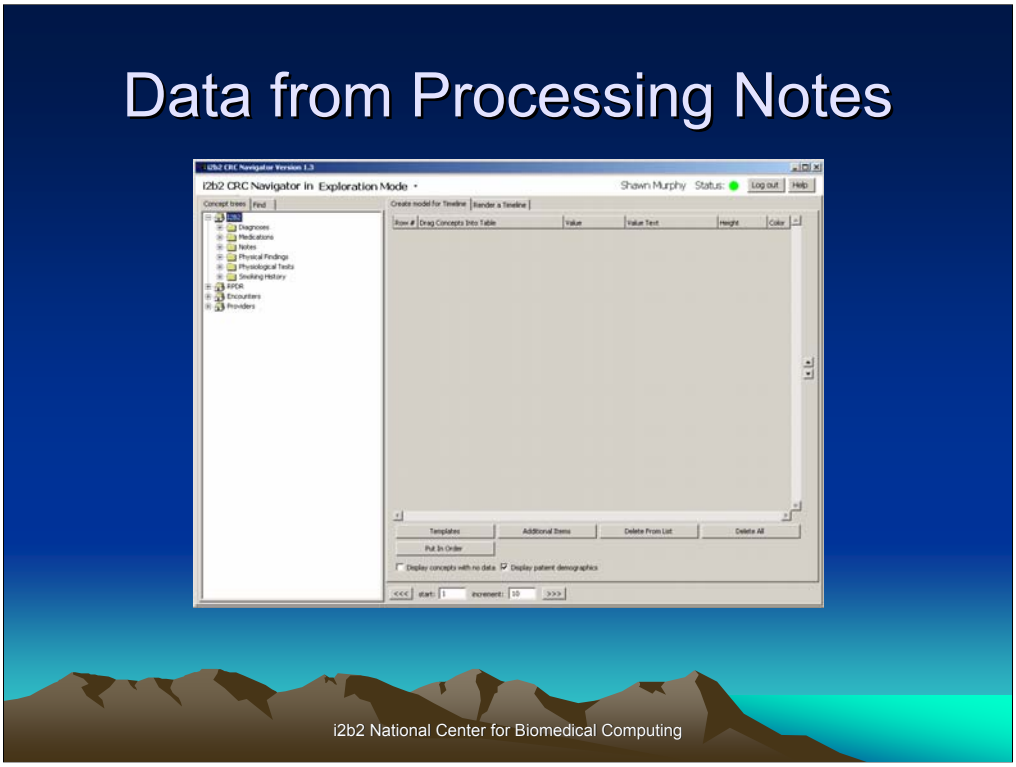
Four cells are involved in the processing on the PFT notes. The data is removed from the CRC using the Workflow Framework cell. It is sent to the PFT Processing cell one by one where the PFTs are parsed, and the concepts are checked for integrity prior to being placed back into the CRC.



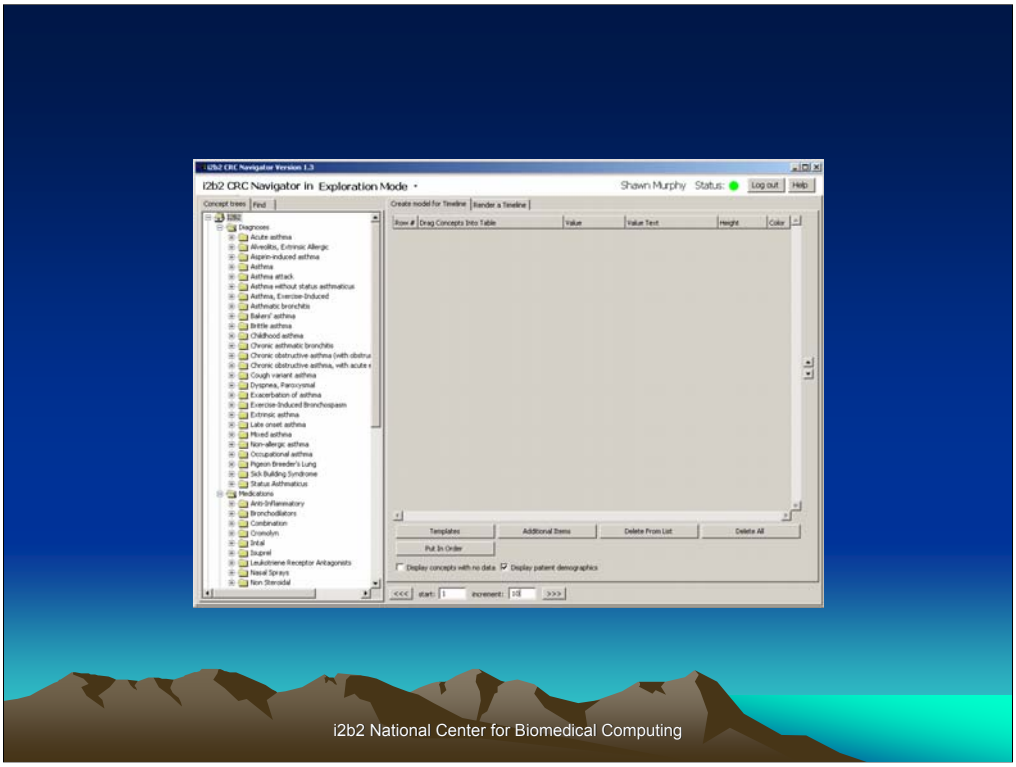
Values are extracted from the hospital discharge summaries and electronic medical record notes and placed back into the CRC as individual concepts.



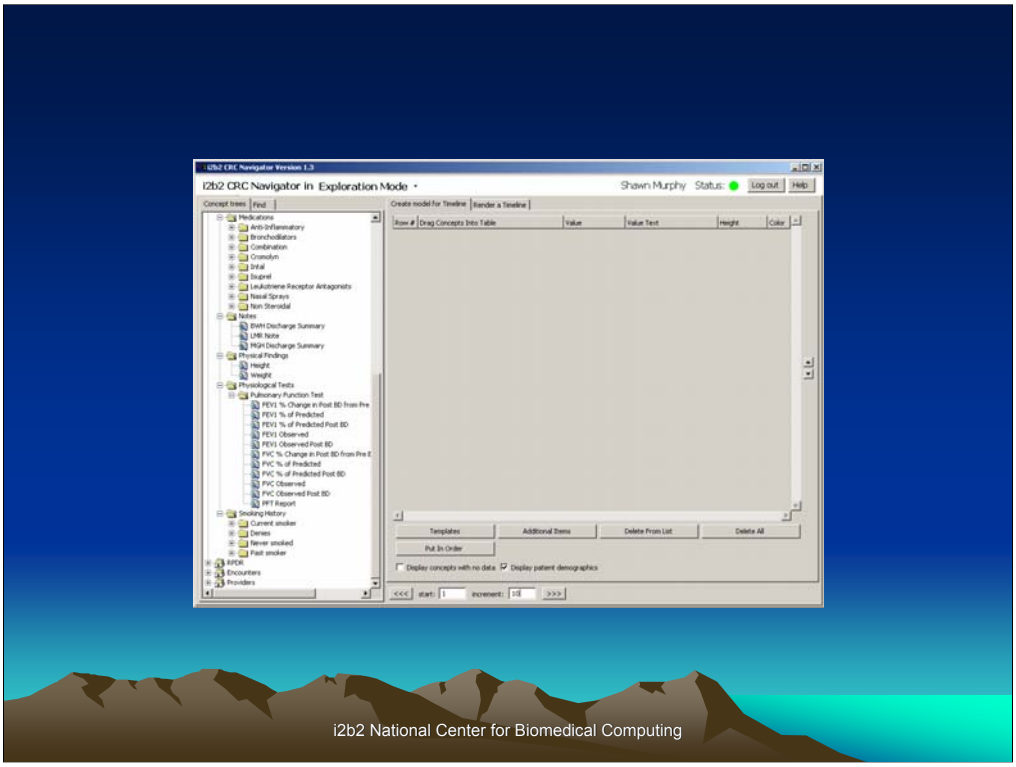
Four cells are involved in the processing on the text notes. The data is removed from the CRC using the Workflow Framework cell. It is sent to the Natural Language Processing cell one by one where the concepts are extracted, and the concepts are checked for integrity prior to being placed back into the CRC. The process is illustrated in the PowerPoint version of these slide by clicking on the “Input document”.



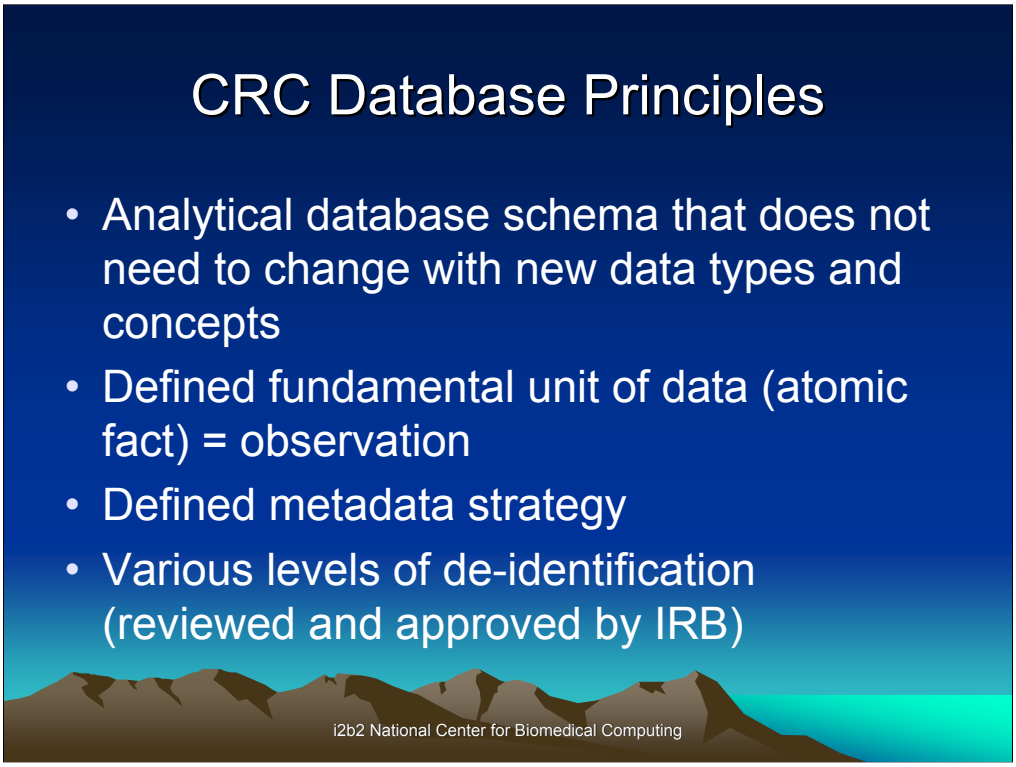
As a result of the import process, phenotypic data is available that may be queried and viewed using the CRC Navigator application.



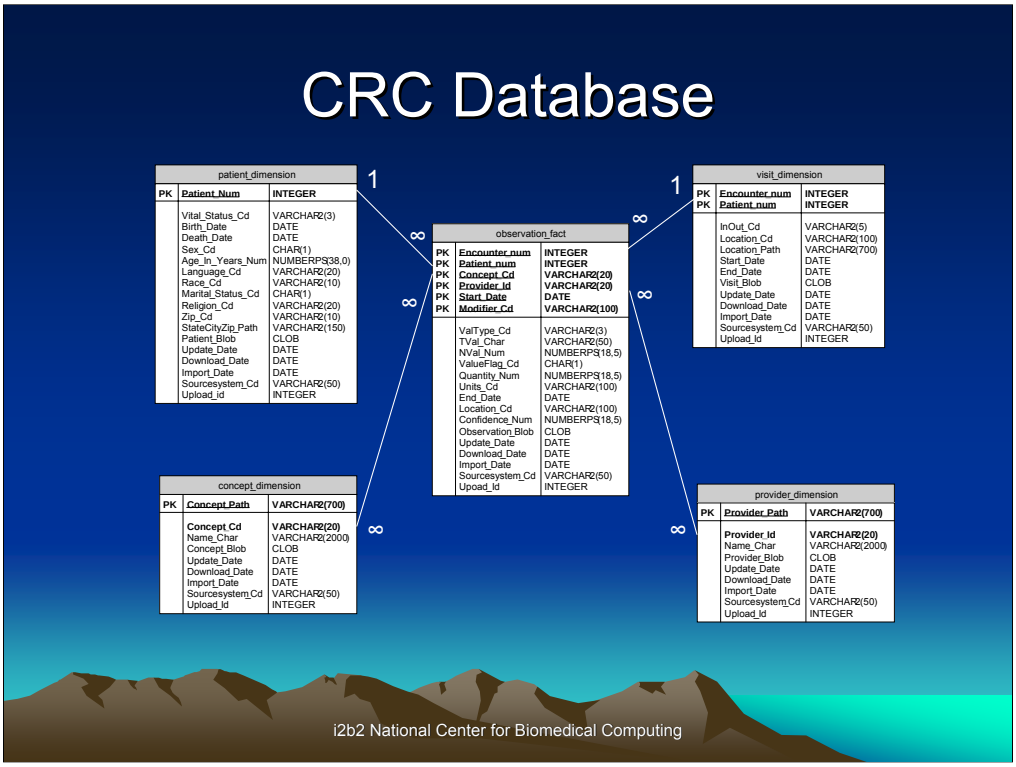
NLP extracted medication and diagnosis concepts related to asthma form the notes.



Physical findings and physiological test results were extracted from the PFTs.

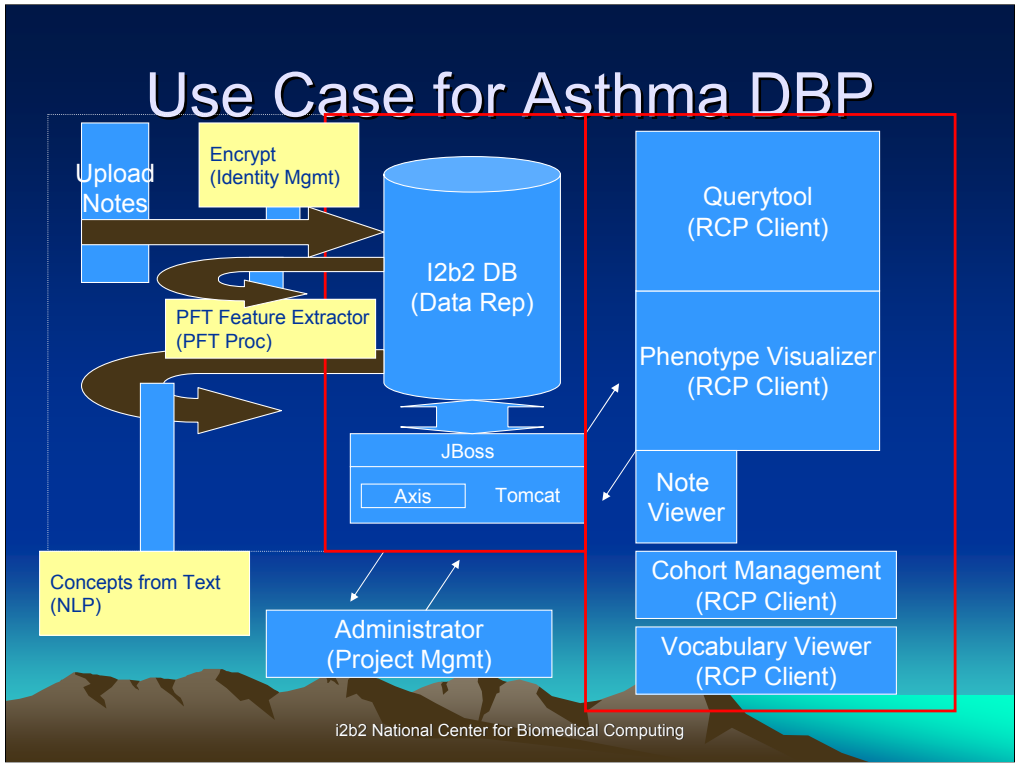


The data now exists in the CRC database, which follows the above principles.

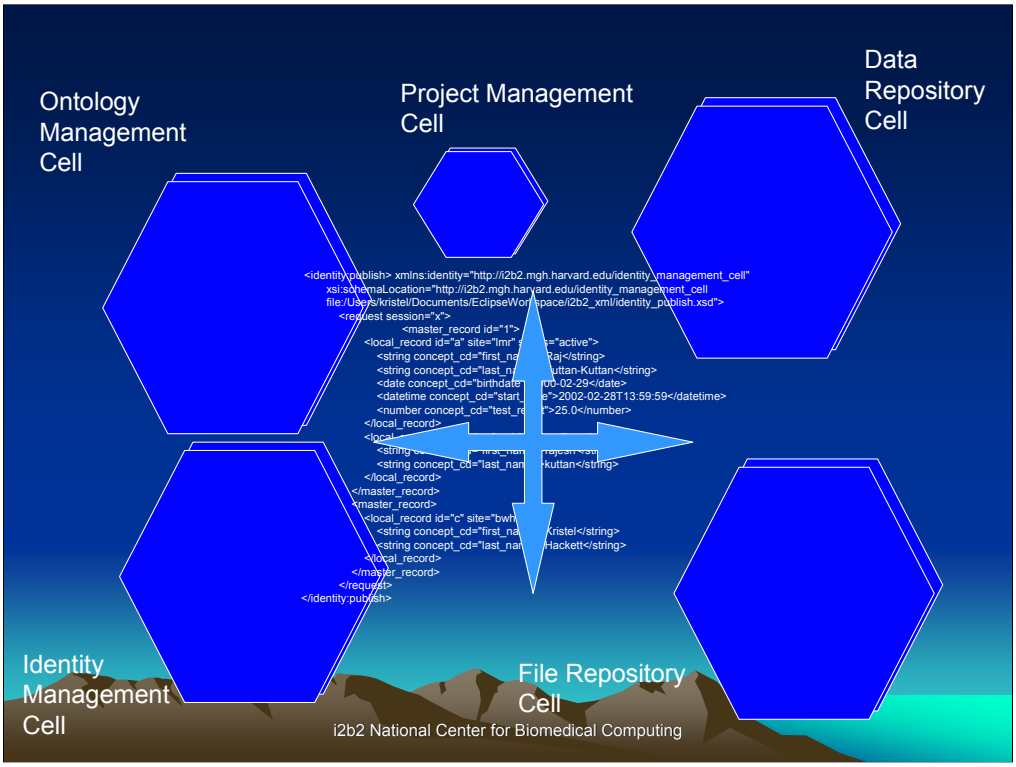


The data model for the CRC database is shown above.

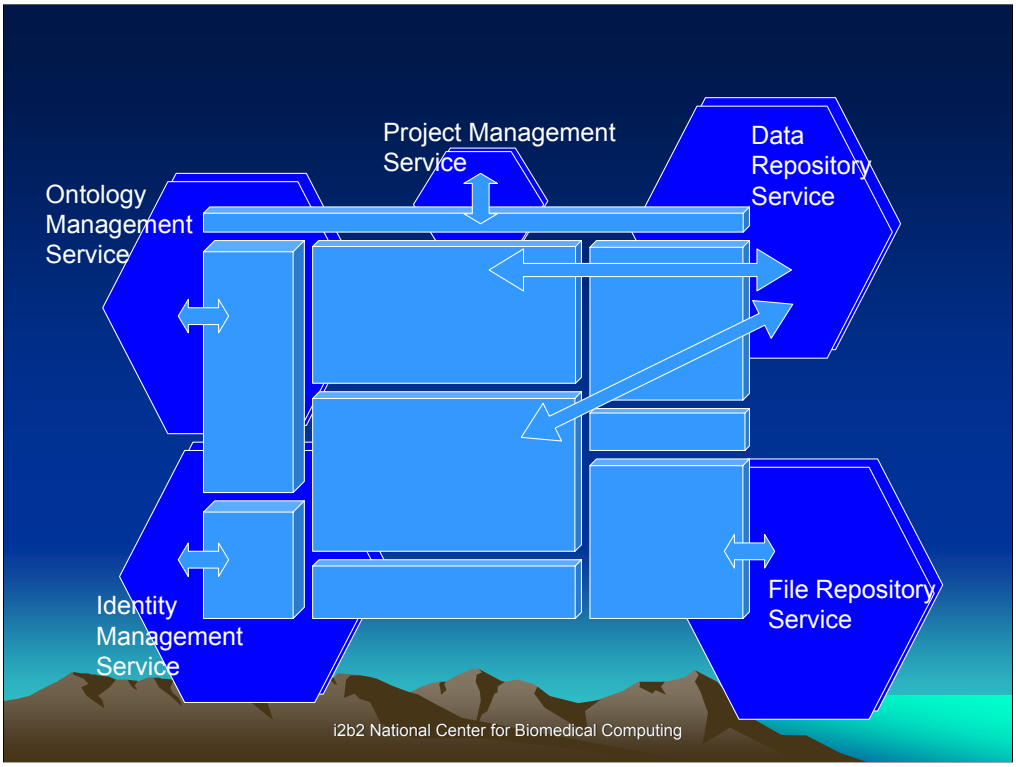
The CRC data model is used as a Reference Information Model to craft the Patient Data Object shown above, that is the fundamental message structure for transferring patient data.



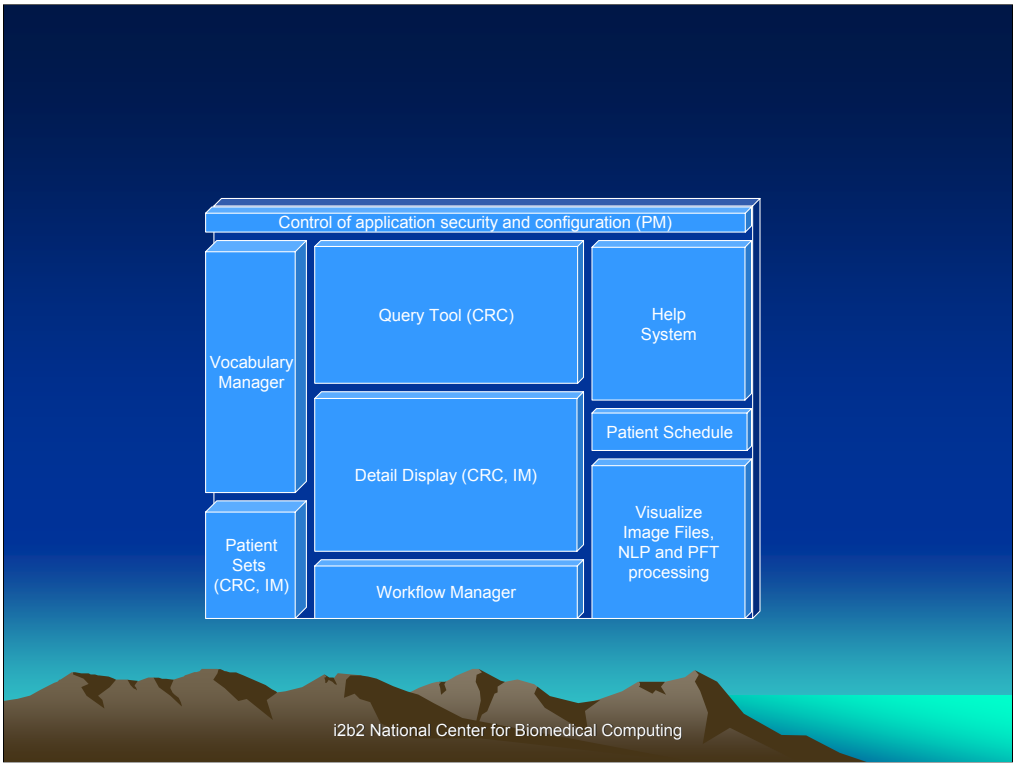
A human interaction with the CRC and the rest of the i2b2 Hive is managed through a set of client applications.



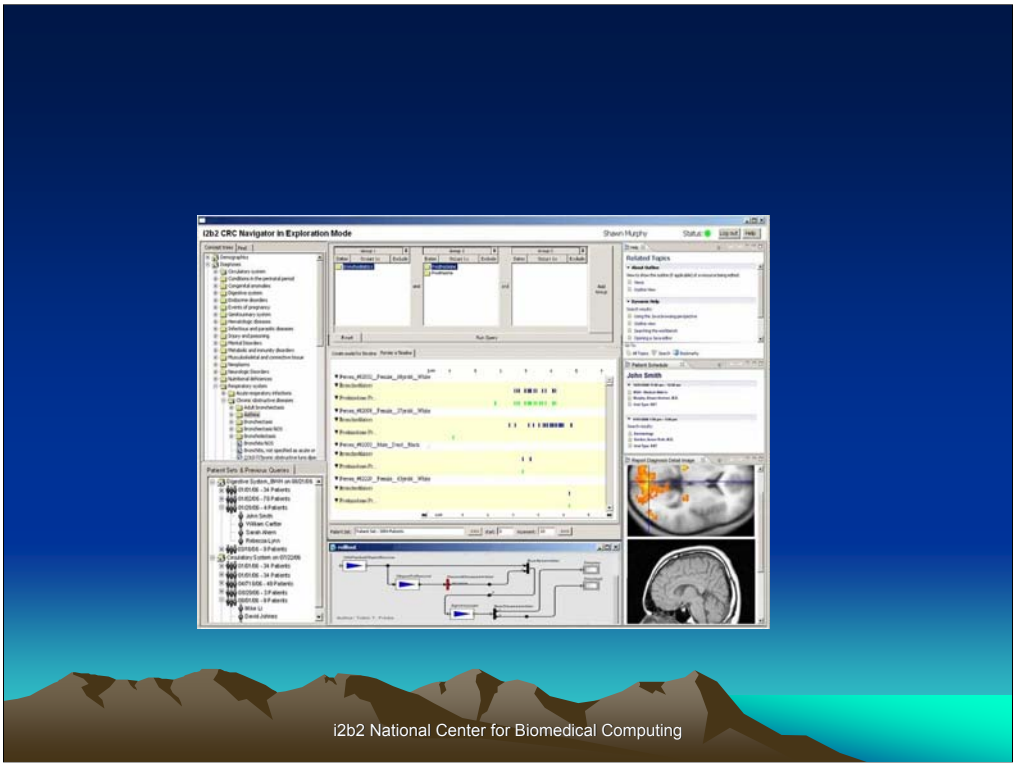
The cells of the i2b2 Hive communicate between each other to maintain their organization. For example, the Identity Management Cell will update the Data Repository Cell when what was formerly recorded as two separate patients is recognized by the Identity Management Cell to be both the same patient.



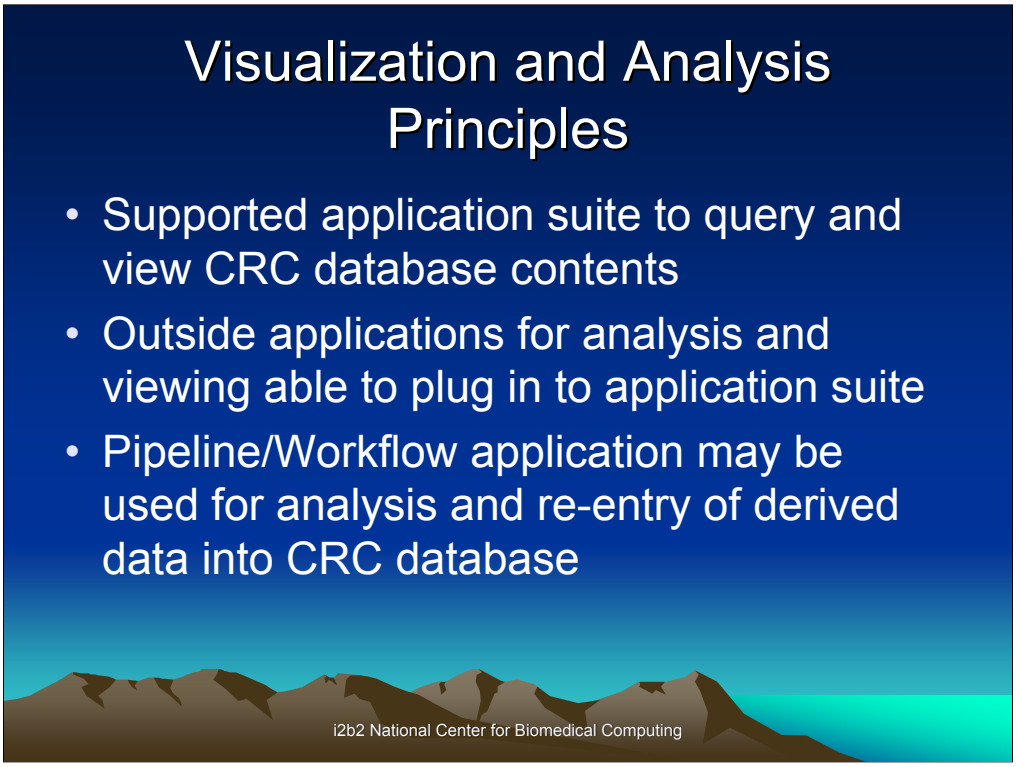
For human communication, cells will usually have a corresponding plug-in that goes onto the i2b2 Navigator.



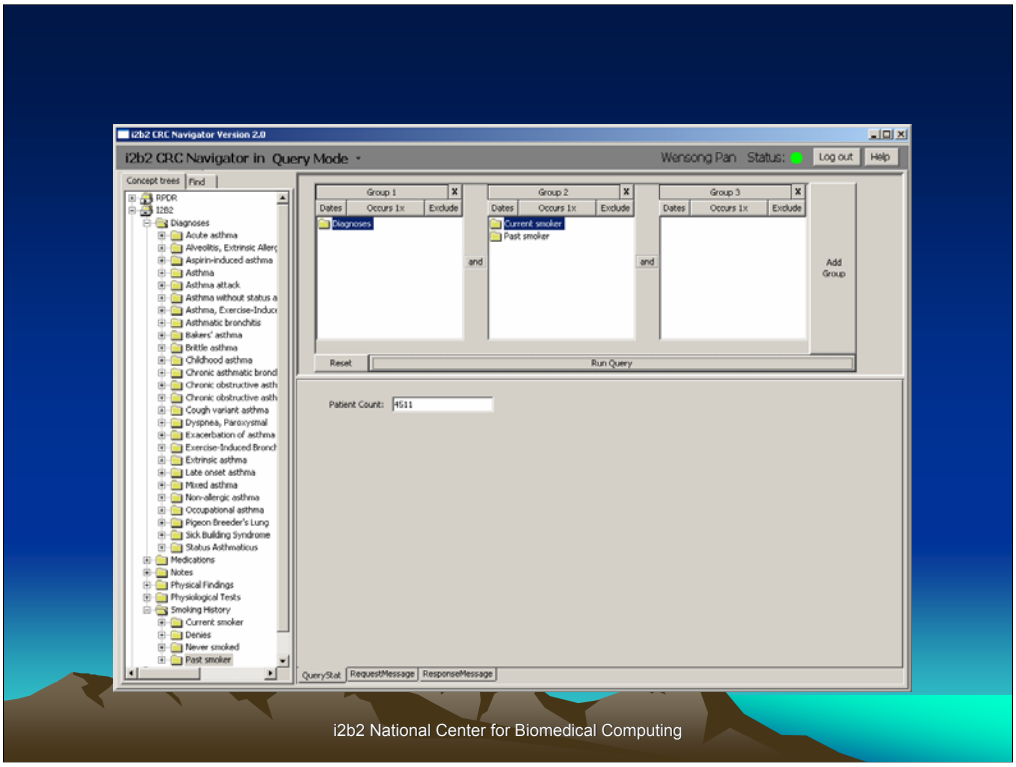
The i2b2 Navigator uses the Eclipse framework available at www.eclipse.org. The client applications are plug-ins, and are the most visible part of the Hive.



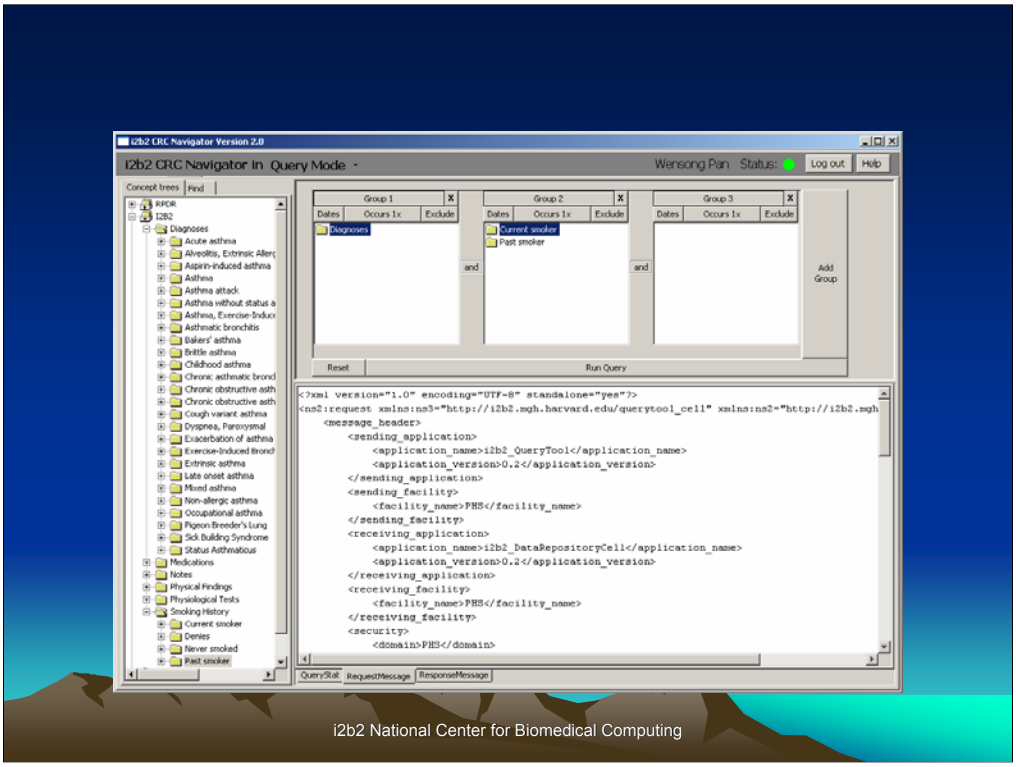
The exposure of these Eclipse plug-ins to the user makes the hive appear as a workbench for managing patient-oriented phenotypic and genotypic data.



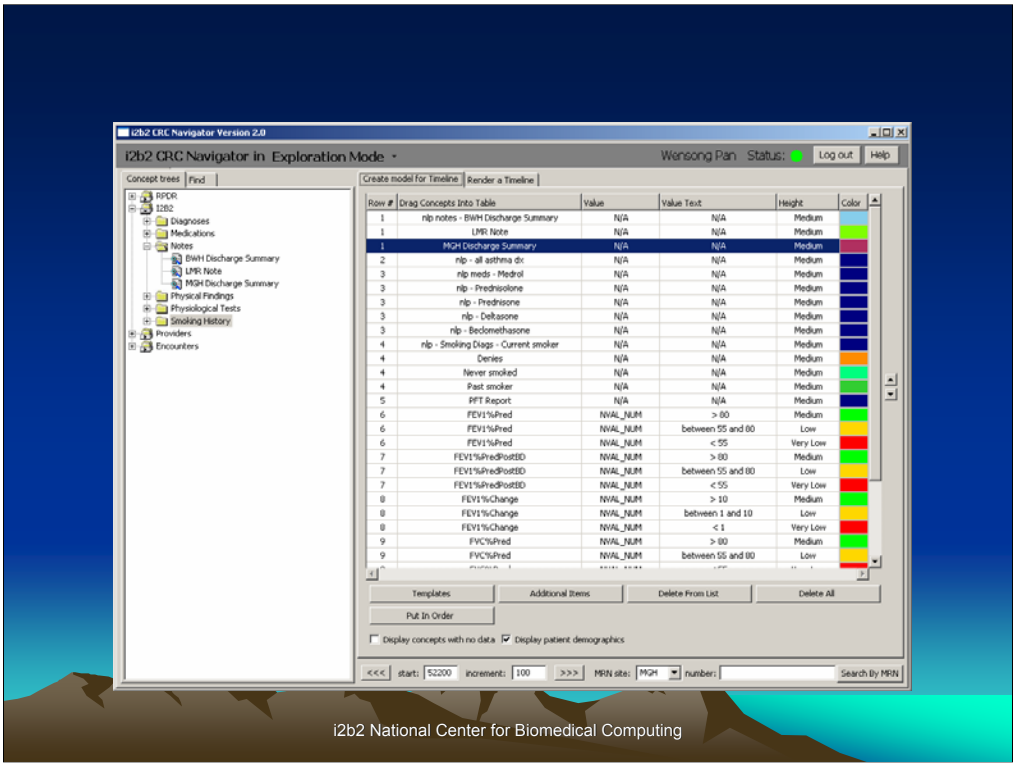
Principles guiding the development of the workbench include a loosely coupled visual framework where independent work from various teams of developers can fit together.



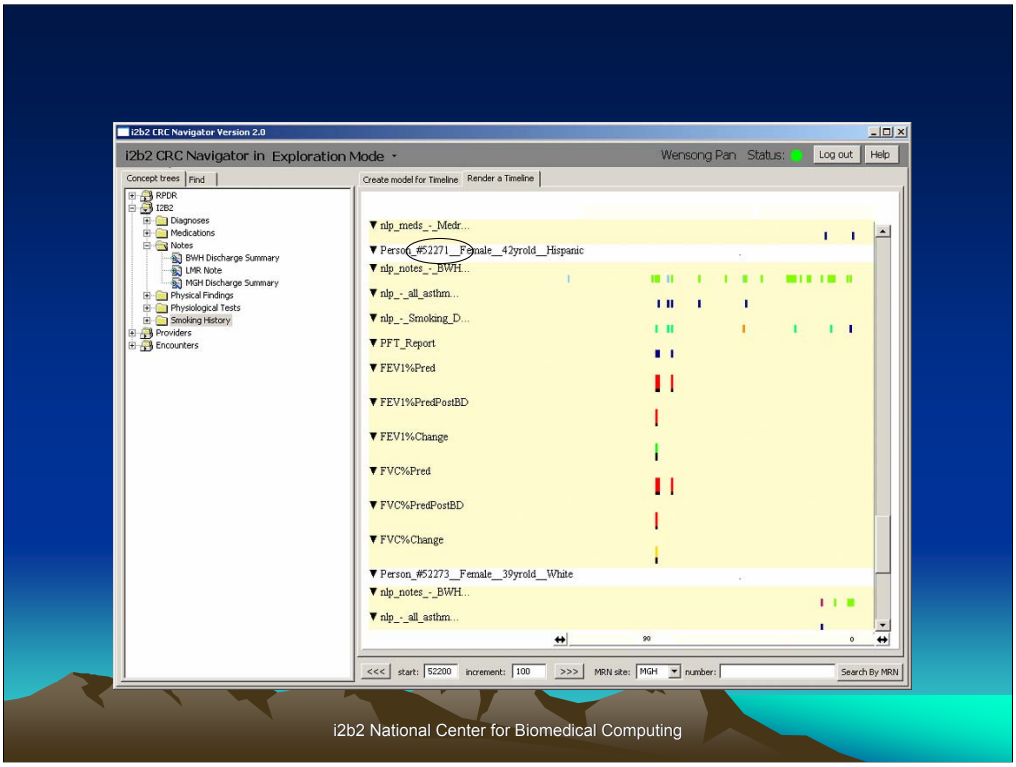
Included in the Navigator is a Query tool for the CRC database.



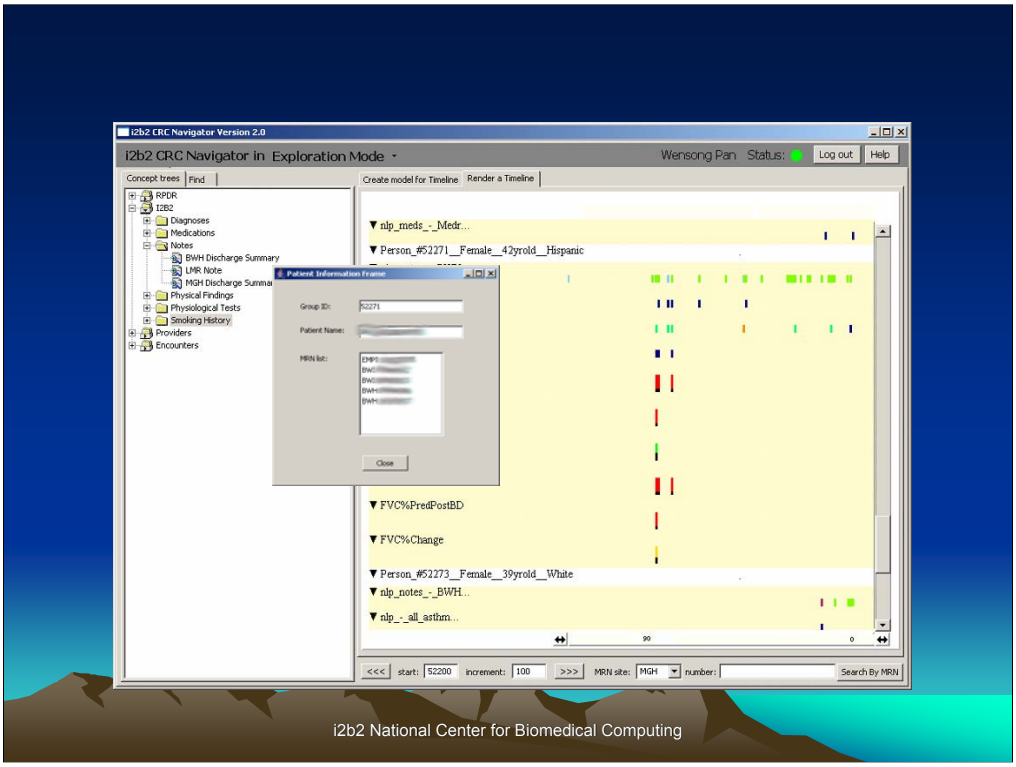
The plug-ins for the Navigator communicate with the Data Repository Cell through the standard i2b2 XML messages that are common in the i2b2 Hive.



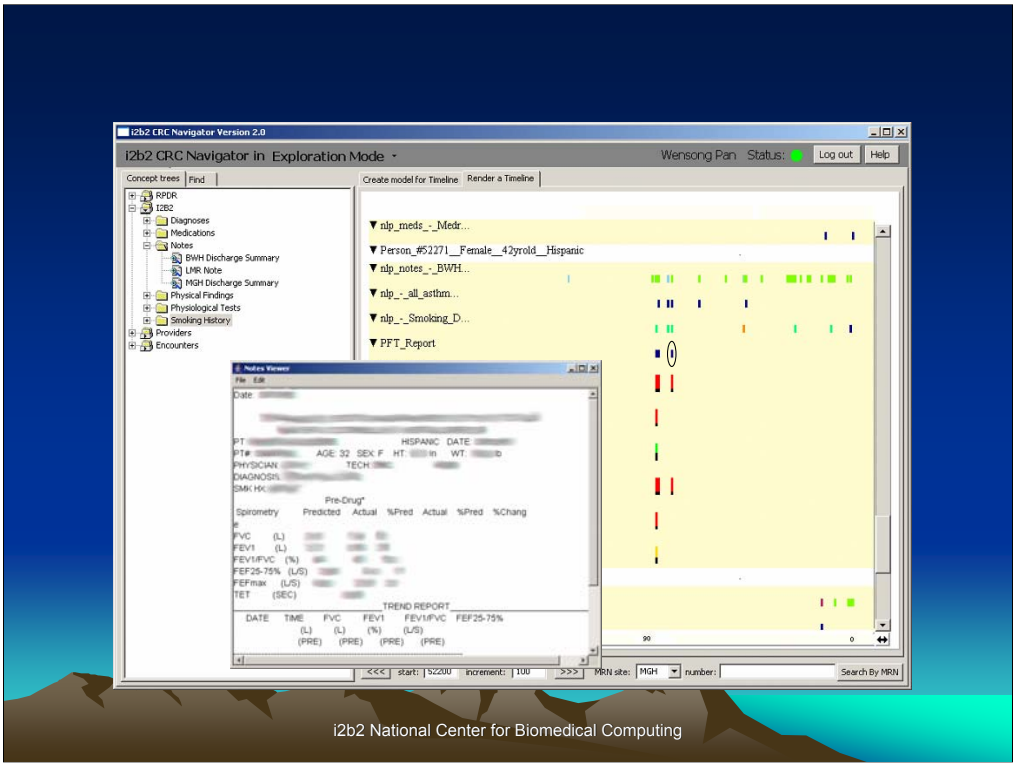
Visualization plug-ins provide a way to understand the data in the CRC.



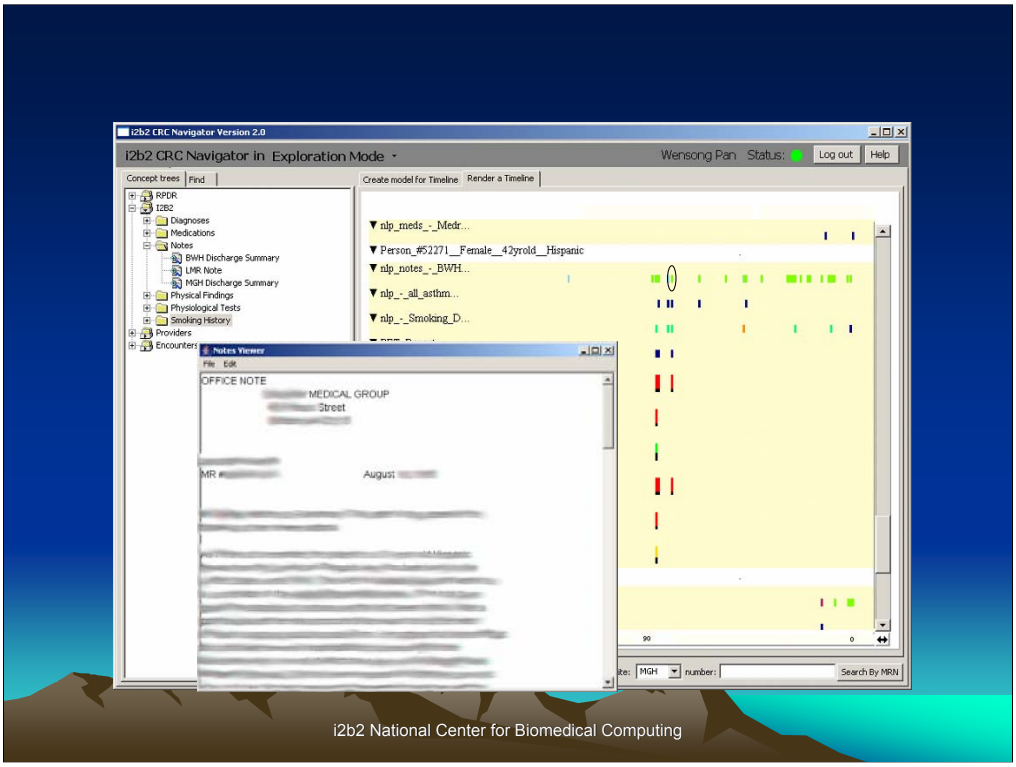
The concepts from the previous slide are shown on a timeline display for each patient. The timeline shows the concept on the left, such as notes and smoking diagnoses derived from the notes, along a time course that runs from left to right. There is one tan band fro each patient that is labeled on the top of the band in white.



If one is properly authorized, clicking on the randomly assigned Patient Number along the top of the band (52271 in this case) will yield identifying information of that patient. This is achieved through a connection to the Identity Management Cell.

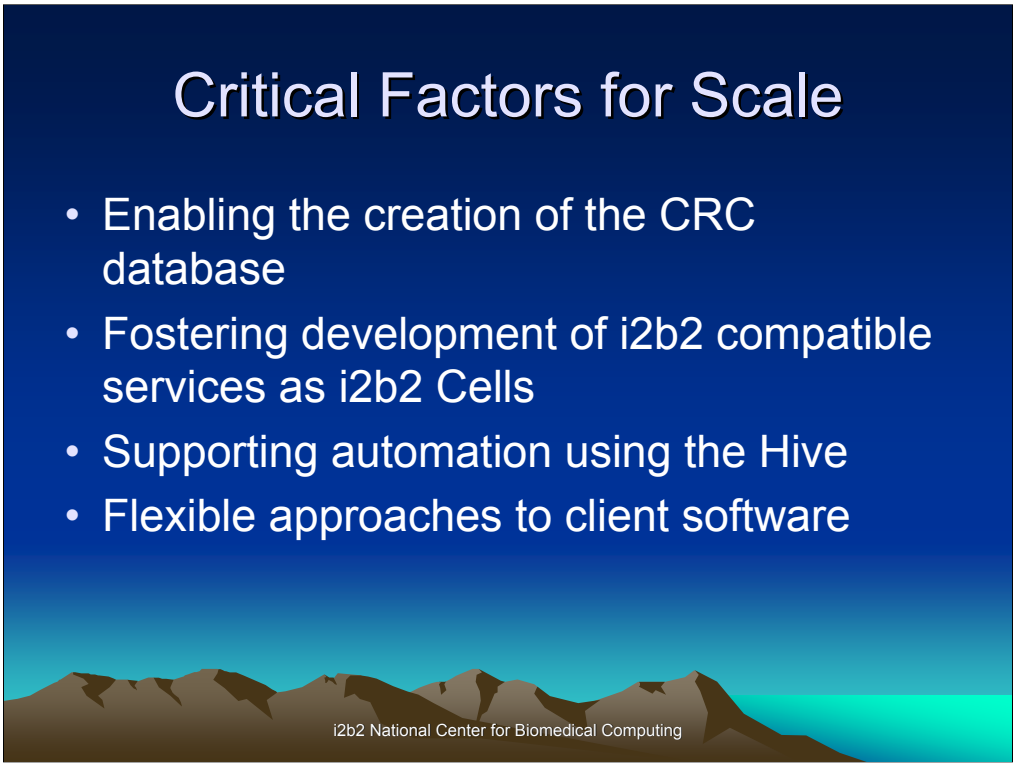


If one clicks on one of the bars in the visual display that represents a report, the report will be decrypted using a decryption key, and the report will be displayed.

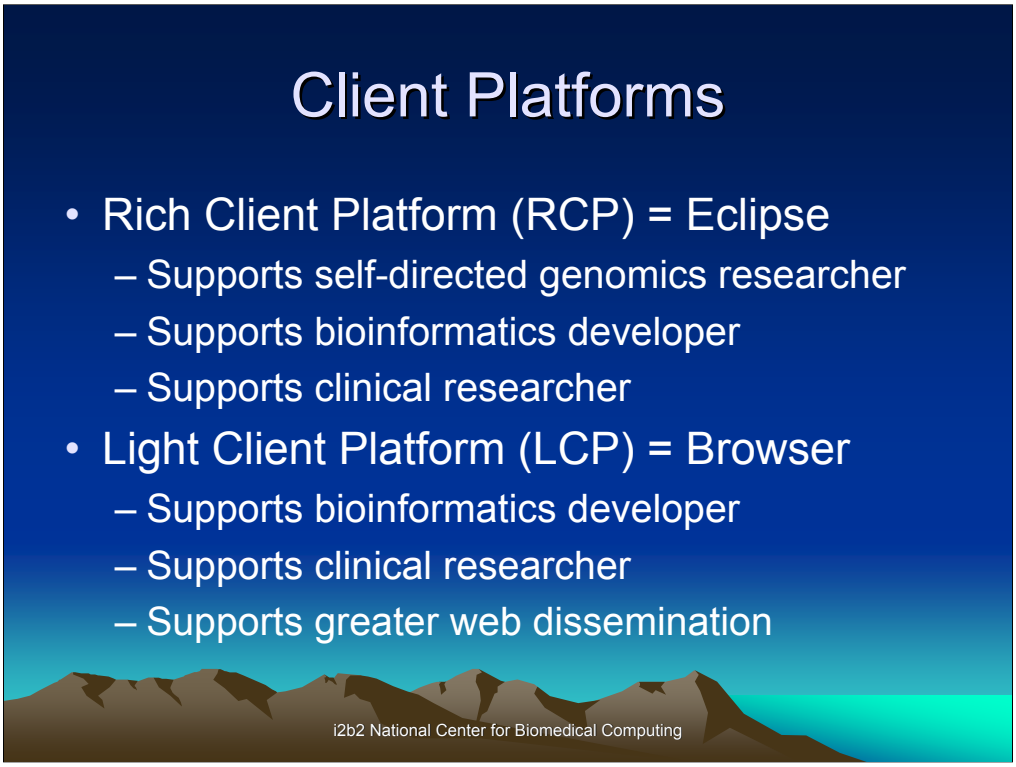


This allows the direct viewing of reports and images in the CRC.

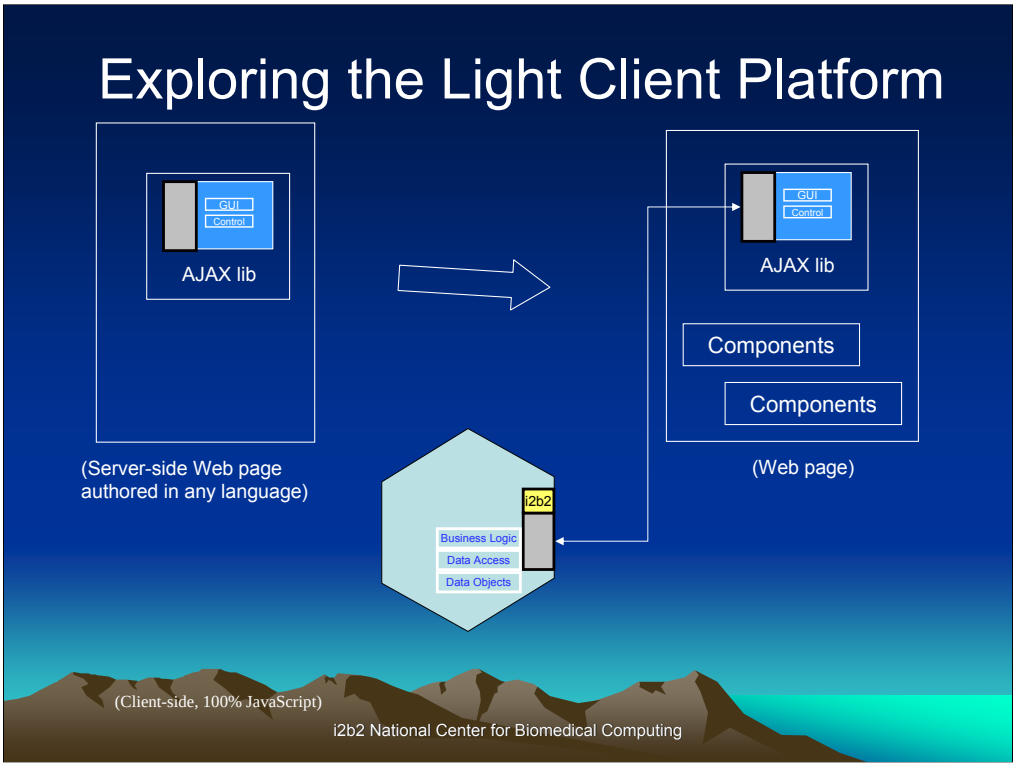
What is important for the i2b2 Hive to scale, and appeal to groups beyond our site?



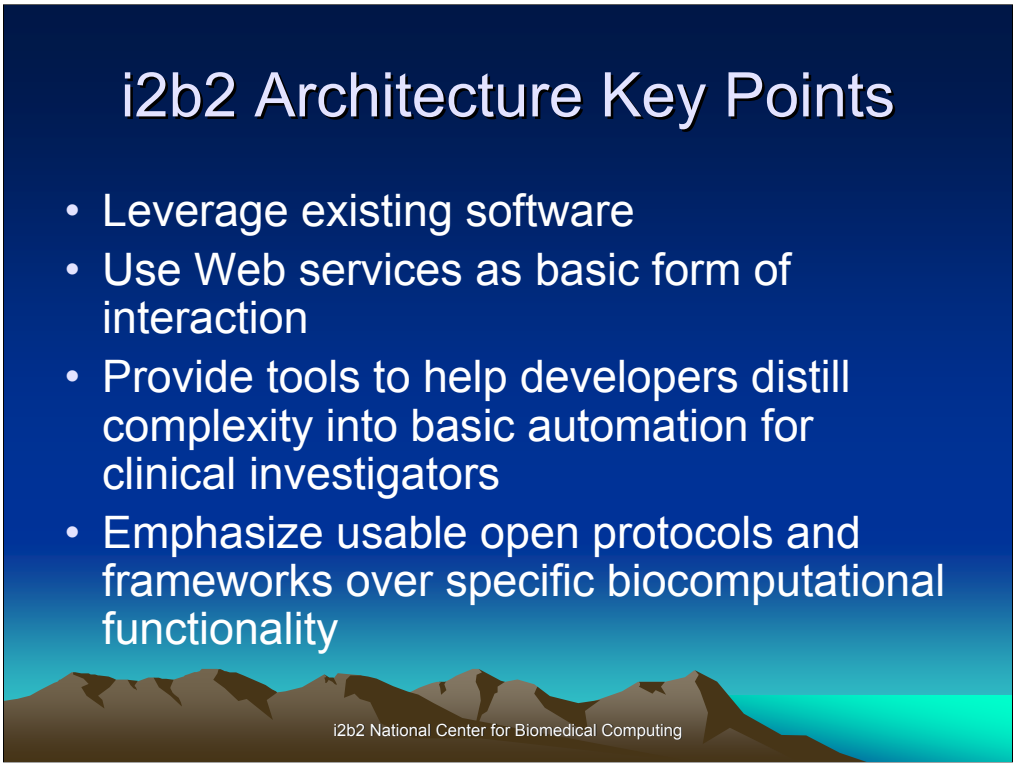
There are several critical factors for getting i2b2 to scale, and these are noted above. Perhaps the most critical are fostering tools to allow sites to populate an instance of the CRC database from their own local systems, and to promote the development of compatible services.



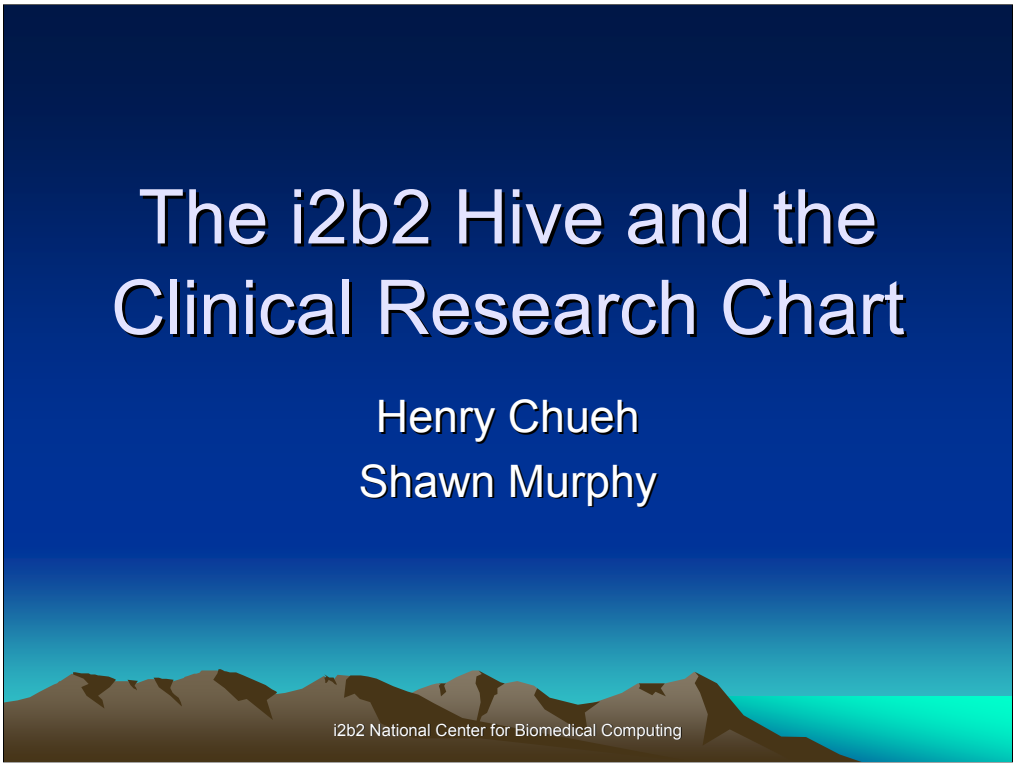
Different client platforms can be established to meet different needs.



Though initial efforts have been to develop i2b2 Cell clients in the Eclipse RCP, we are also exploring the Web 2.0 LCP. Use of AJAX techniques and libraries may allow server-side Web applications to be authored in any language, while still taking advantage of i2b2 services.



One of the main points about the i2b2 architecture is that it emphasizes open protocols over specific software or even functionality. The concept is that software and functionality will need constant renewal, and that the i2b2 platform should enable and facilitate this approach.


The i2b2 Hive and the Clinical Research Chart
Henry Chueh
Shawn Murphy
i2b2 National Center for Biomedical Computing