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 goal of i2b2 is to produce a comprehensive set of software tools to enable clinical investigators to collect, manage and share their project-related clinical and genomic research data, that is, a software suite for the modern clinical research chart.

The model used for i2b2 is the ‘hive’, which is made up of building blocks called ‘cells’. Each cell may be an independently-developed software package, service or set of services with a common function. The cells communicate with each other via document-oriented XML messages. Cells in the i2b2 hive gather and process medical information, and then place this data in the Clinical Research Chart (CRC) via XML messages.

We present an example of a clinical asthma investigation to demonstrate the architecture and tools designed for the i2b2 hive. In this case, only text notes from an asthma clinic and some reports from pulmonary function tests (PFTs) are available. Using this data, the steps to extract, encrypt and collate meaningful concepts and their values using the i2b2 hive are described.

The Workflow Framework (WF) cell directs a series of communications between other cells. The clinical asthma investigation requires a workflow that involves the Identity Management (IM) cell, De-Identification (DID) cell, Data Repository (CRC) cell, Natural Language Processing (NLP) cell and Pulmonary Function Test Processing (PFT) cell.

Before the text notes and PFT reports are added to the CRC cell, the IM cell provides a code to replace patient identifiers and the text is either encrypted or stripped of identifiers by the DID cell. Only users with access rights to see the identified notes will have a key or certificate to view them. This preserves the CRC as a HIPAA-defined limited data set. In order to extract concepts from text notes, such as hospital discharge summaries and electronic medical record notes, the WF cell retrieves notes from the CRC and sends them to the NLP cell. Concepts are extracted, checked for integrity and then placed back into the CRC. In order to extract concepts from PFT reports the WF cell passes PFT reports from the CRC to the PFT Processing cell. PFTs are parsed and the values associated with each test are placed back into the CRC.

Once the data has been processed, it is available to be viewed and manipulated using the i2b2 Workbench. The Workbench is a client application that consists of a collection of plug-ins within a visual framework that is loosely-coupled and allows different groups of developers to contribute their own plug-ins. The open-source Eclipse framework available at www.eclipse.org contains the plug-ins. Plug-ins determine the way users interact with particular cells, so when a cell is developed, a plug-in may also be developed in order to support the intended operations of that cell. A complete package to illustrate the development of a cell and its accompanying plug-ins can be found at www.i2b2.org.

The hive can be thought of a toolkit for clinical discovery, one that provides a number of advantages to the typical approach of creating a unique data set for each new study. The hive allows for the integration of analytical programs built by different groups, and it does so in a way that takes into account unique data ownership and consent issues. Researchers are able to share data and computational methods, which may then be used more than once. Because of its modular design, the hive also makes it possible to combine different types of data; specifically, it allows for genotypic data to be easily integrated into the patient record. Another benefit is scalability; as more and more services are developed by different groups, their cells are added to the hive in a seamless way. When combined with a visualization and analytic tool like the Workbench, the result is an organized and flexible approach to clinical discovery.

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