User Centred Networked Health Care
A. Moen et al. (Eds.)
IOS Press, 2011

doi:10.3233/978-1-60750-806-9-887

The ONCO-I2b2 Project: Integrating Biobank Information and Clinical Data to Support Translational Research in Oncology

Daniele SEGAGNI[a], Valentina TIBOLLO[a], Arianna DAGLIATI[c], Leonardo PERINATI[a], Alberto ZAMBELLI[a], Silvia PRIORI[a], Riccardo BELLAZZI[b,a]
[a] *IRCCS Fondazione S. Maugeri, Pavia, Italy*
[b] *Dipartimento di Informatica e Sistemistica, Università di Pavia, Italy*
[c] *Institute for Advanced Studies, Pavia, Italy*

Abstract. The University of Pavia and the IRCCS Fondazione Salvatore Maugeri of Pavia (FSM), has recently started an IT initiative to support clinical research in oncology, called ONCO-i2b2. ONCO-i2b2, funded by the Lombardia region, grounds on the software developed by the Informatics for Integrating Biology and the Bedside (i2b2) NIH project. Using i2b2 and new software modules purposely designed, data coming from multiple sources are integrated and jointly queried. The core of the integration process stands in retrieving and merging data from the biobank management software and from the FSM hospital information system. The integration process is based on a ontology of the problem domain and on open-source software integration modules. A Natural Language Processing module has been implemented, too. This module automatically extracts clinical information of oncology patients from unstructured medical records. The system currently manages more than two thousands patients and will be further implemented and improved in the next two years.

Keywords. I2B2, oncology research, biobanks, natural language processing, translational research, hospital information system integration

1. Introduction

ONCO-i2b2 is a project funded by the Lombardia region, in Italy, which aims at supporting translational research in oncology. The project exploits the software solutions implemented by the Informatics for Integrating Biology and the Bedside (i2b2) research center, an initiative funded by the NIH Roadmap National Centers for Biomedical Computing and headed by Partners HealthCare Center in Boston [1]. The i2b2 project developed a data warehouse and a set of software solutions that are based on an architecture called “hive”. The “hive” has different software cells devoted to data extraction, data manipulation or data analysis tasks [2].

Within ONCO-i2b2, the University of Pavia and the hospital IRCCS Fondazione S. Maugeri (FSM) have integrated the i2b2 infrastructure with the FSM hospital information system (HIS) and with a cancer biobank that manages both plasma and cancer tissues. The integration with the HIS provides the access to all the electronic

medical records of cancer patients. The majority of the data collected in the FSM HIS is represented by textual reports. It was therefore necessary to develop and integrate inside the ICT architecture a Natural Language Processing (NLP) module in order to extract important information and clinical tests results, such as patients' histological reports [3]. The oncology biobank provides bio-specimens prepared from a collection of blood and tissue samples, taken with the informed consent of healthy individuals and oncologic patients.

The aim of this paper is to describe the basic steps of the integration process and to present the current status of the ONCO-i2b2 project.

2. Method

The ONCO-i2b2 software implemented at the FSM hospital is designed to integrate data from many different sources and collected for different purposes, in order to allow researchers querying and analyzing the vast amount of information coming from the clinical practice. The main data sources that we have integrated into the i2b2 data warehouse are the hospital pathology unit, the biobank and the HIS. In the following we will describe the detail of the integration process.

2.1. FSM Pathology Operative Unit and Biobank

Data associated to the biospecimens stored inside the biobank are almost automatically uploaded from the hospital pathology unit. A semi-automatic procedure has been implemented to populate the biobank database in order to decrease the time of insertion and reduce the possibility of human error. One of the major efforts made during this implementation was to anonymize each cancer biospecimen, by creating a two-dimensional DataMatrix barcode that does not include any direct reference to the donor patient. Cancer tissues or plasma samples are selected by researchers and placed in new tubes labeled with the new barcode. Granted users may retrieve the information related to the donors through a specialized software application that also show the patient's informed consent.

The biobank database is periodically synchronized (several times during the day) in order to keep biobank samples data constantly updated.

The information on the biological samples contained in a biobank are loaded into the i2b2 data warehouse through a complex series of Extract, Transform, Load (ETL) operations that involve data extraction, processing and mapping in the data warehouse [4]. The ETL activity was performed relying on the KETTLE [5] developed within the Pentaho project [6].

Table 1 shows the number of patients and biological samples currently available in the biobank divided by hospital medical unit of origin and type.

Figure 1 shows the different steps of the integration process. Step 1 is the semi automated data extraction from the pathology unit, step 2 describes the anonymization process involving biosamples before they are stored in biobank. Step 3, instead, represents the i2b2 data warehouse where data from different sources are collected through ETL transformations. Step 4 shows how the information coming from the HIS is integrated, too.

Table 1. Biobank biospecimens count divided by hospital operative unit of origin and type. Table data are related to the period 1-12-2009 - 15-01-2010.

FSM Unit	Patient	Tissue	Plasma
Senology	237	729	729
Surgery	75	567	243
TOTAL	312	1296	972

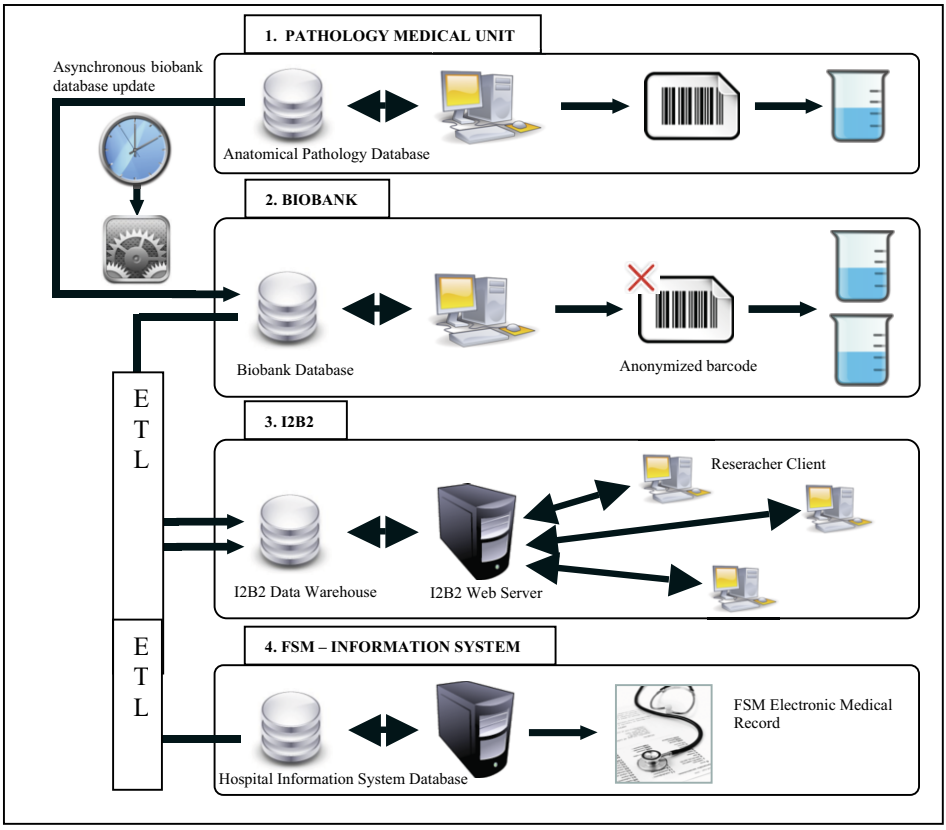

Figure 1. ICT architecture designed to integrate information from the FSM medical units and the hospital information system.

2.2. FSM Hospital Information System

The information collected in the FSM HIS is made available to the I2B2 service through an ETL process that transforms the medical information of interest in concepts that will be queried in the research phase. Some of these oncological concepts refer to various key facts collected in the pathological anatomy electronic report that are only available in textual format.

A NLP software module has thus been developed to extract this information from the FSM HIS for each cancer patient that have at least one biological sample stored in the biobank. To address the problem of extracting structured information from pathology reports for research purposes, we developed an NLP module based on the

GATE system [7] to automatically identify and map anatomic and diagnostic noun phrases found in full-text pathology reports to SNOMED concept descriptors. The pathology unit uses unstructured or semi-structured text documents to represent this information. Therefore, we identified a set of regular expressions that matched clinical phrases commonly found in pathology reports; such expressions are then properly processed by the NLP parser. In particular, the retrieved data relate to a set of oncological SNOMED codes, to the values derived from clinical tests, like scoring breast carcinomas stained with HercepTest or to the scores of the expression of Ki-67, a nuclear antigen protein used to determine the growth fraction of tumors [8]. The system has been internally validated by a manual verification by the medical experts involved in the study on a subset of 100 cases with 100% accuracy. This module is now a part of the overall data warehouse management strategy.

2.3. The Integrated Architecture: i2b2

The i2b2 data warehouse, called Clinical Research Chart (CRC), is designed to manage data from clinical trials, medical record systems and laboratory systems, along with many other types of clinical data from heterogeneous sources [9]. The CRC stores this data in three tables, the patient, the visit and the observation tables. The three data tables, along with two of the lookup tables (concept and provider), are the main components of the so-called star schema of the data warehouse. The most important aspect of the construction of a star schema is identifying what is a "fact". In healthcare, a logical fact is an observation on a patient. The dimension tables contain further descriptive and analytical information about attributes in the fact table.

The i2b2 infrastructure installed at FSM provides a web-based access to any type of data described in the previous paragraphs. Data information are stored in the i2b2 data warehouse through complex ETL transformations following a cancer-specific ontology that combines atomic information to create a well defined medical observation. The extracted information can be analyzed through the i2b2 web client with appropriate plug-ins specially configured [10, 11].

3. Results

Since December 2010 the entire software system has been installed and is currently running at FSM. The aim of the implementation of the architecture was to allow the FSM researchers to exploit i2b2 query capabilities relying on the user-friendly web interface available. To achieve this goal we focused on the development of data integration processes, on the design of NLP modules and on the management and anonymity of the biological samples contained in biobank.

Integration of these data from heterogeneous sources required several key steps: i) creation of a specific software to upload the information available in the pathology unit; ii) generation of new barcodes when the biosamples are archived in biobank; iii) design and configuration of an NLP software module to extract information from unstructured text documents relevant to the clinical characterization of patients in cancer research; iv) creation of ETL transformations to populate the i2b2 data warehouse with concepts related to cancer research.

Currently, the i2b2 instance installed in FSM consists in 2214 patients (312 of them have at least one biological sample in the cancer biobank), 25826 visits, 163

concepts (divided into demographic data, diagnoses, clinical measurements, histological reports, therapies and biobank samples) and 93680 observations.

4. Discussion

The novel IT architecture created at FSM is a concrete example of how integration between different information from heterogeneous sources can be correctly implemented and make available for scientific research. In order to continuously improve i2b2 easiness of use for hospital researchers, we added at the i2b2 web client application novel plug-ins for data export and for phenotype exploration [12]. One of the major efforts made during the implementation of the i2b2 extensions was to be fully compliant with i2b2 development guidelines, so that our software modules and architecture can be reused by the other researchers of the i2b2 community.

Exploiting the potential of this IT architecture, the next steps of the project will involve the extension of the data set imported by the HIS as well as the management of data from laboratory tests. We also plan to continue extending the capabilities of the FSM i2b2 architecture by implementing new plug-in devoted to data analysis; in particular, we are working on an extension of the i2b2 query engine by adding temporal query capabilities Finally, another important point for the future development of the project will be the integration of patient's genotype data, which will require careful evaluation both in terms of the data representation and storage and of data security and privacy.

Acknowledgements. This paper describes the ONCO-i2b2 project, funded by the Lombardia Region, in Italy. We gratefully acknowledge Prof. Carlo Bernasconi and the Collegio Ghislieri in Pavia for their active support.

References

[1] Murphy SN, Mendis M , Hackett K, et al. Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside, *AMIA Annu Symp Proc.* (2007), 548-52.
[2] Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2), *J Am Med Inform Assoc*. (2010), 124-30.
[3] Jurafsky D, Martin JH. *Speech and Language Processing, An Introduction to Natural Language Processing,Computational Linguistics, and Speech Recognition* Second Edition, Prentice Hall, 2008.
[4] Kimball R, Ross M, Thornthwaite W, Mundy J, Becker B. *The Data Warehouse ETL Toolkit* (2nd edition), 2008
[5] Pentaho Corporation, *Pentaho Data Integration Kettle Documentation* (http://kettle.pentaho.com), 2011
[6] Bouman R, J. van Dongen. *Pentaho Solutions* , Wiley, 2009
[7] The University of Sheffield, *GATE software* (http://gate.ac.uk/sale/tao/split.html), 2011
[8] Broyde A , Boycov O, Strenov Y, Okon E, Shpilberg O , Bairey O. Role and prognostic significance of the Ki-67 index in non-Hodgkin's lymphoma, *Am J Hematol*. 2009 Jun;84(6):338-43.
[9] Partners HealthCare Systems, *I2B2 software (v.1.5) documentation*, 2008.
[10] Mendis M, Wattanasin N, Kuttan R, et al. Integration of Hive and cell software in the i2b2 architecture, *AMIA Annu Symp Proc*. (2007), 1048.
[11] Murphy SN, Churchill S, Bry L, et al. Instrumenting the health care enterprise for discovery research in the genomic era. *Genome Res*. (2009), 1675-81
[12] Bellazzi R, Segagni D et al. R Engine Cell: integrating R into the i2b2 software infrastructure, *J Am Med Inform Assoc* (2011)