Journal of Biomedical Informatics 42 (2009) 100–112

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Translational integrity and continuity: Personalized biomedical data integration

Xiaoming Wang [a,c,*], Lili Liu [a,c], James Fackenthal [b], Shelly Cummings [b], Oluwatobi I. Olopade [b], Kisha Hope [b], Jonathan C. Silverstein [a,c], Olufunmilayo I. Olopade [b,*]

[a] Biomedical Informatics Core, Computation Institute, University of Chicago, 950 E 61st Street, Room 334, Chicago, IL 60637, USA
[b] Center for Clinical Cancer Genetics, Department of Medicine, University of Chicago, 5841 South Maryland Avenue, Room I-216, Chicago, IL 60637, USA
[c] Computation Institute, University of Chicago, 5801 S Ellis Avenue Suite 400, Chicago, IL 60637, USA

ARTICLE INFO

Article history:
Received 2 February 2008
Available online 12 August 2008

Keywords:
Data integration
Data curation
Data integrity
Data continuity
Translational research

ABSTRACT

Translational research data are generated in multiple research domains from the bedside to experimental laboratories. These data are typically stored in heterogeneous databases, held by segregated research domains, and described with inconsistent terminologies. Such inconsistency and fragmentation of data significantly impedes the efficiency of tracking and analyzing human-centered records. To address this problem, we have developed a data repository and management system named TraM (http://tram.uchicago.edu), based on a domain ontology integrated entity relationship model. The TraM system has the flexibility to recruit dynamically evolving domain concepts and the ability to support data integration for a broad range of translational research. The web-based application interfaces of TraM allow curators to improve data quality and provide robust and user-friendly cross-domain query functions. In its current stage, TraM relies on a semi-automated mechanism to standardize and restructure source data for data integration and thus does not support real-time data application.

1. Introduction

With completion of the human genome project, scientists are systematically studying the molecular basis of human diseases [1–3] to explore effective individualized therapies [4–7]. To achieve this unprecedented goal, investigators are breaking traditional boundaries between research domains from patient bedsides to experimental laboratories to conduct translational research [7–8]. Data generated from research on different topics need to be extensively reviewed and iteratively verified to become reliable clinical or scientific knowledge [9–10]. However, because the majority of clinical and basic research data are currently stored in disparate and separate domain databases, it is often inefficient for a researcher to access these data [11–14]. Furthermore, even where domain data can be aggregated and viewed through a single computational platform, translational researchers still often see incomplete, fragmented, and unverified data in their original forms. These problems greatly impede research efficiency, particularly statistical analysis. Despite overwhelming demands for a modern method to facilitate personalized data tracking, management, and improvement over a translational workflow, few software products that meet these requests are available or widely accepted in the translational research community.

Our goal was to provide a computational system that is able to: (1) integrate data generated from multiple research domains with the flexibility to capture dynamically evolving domain concepts; (2) allow curation for data improvement; (3) support robust and intuitive query functions for biomedical researchers; (4) execute independently from third party products, meaning the system does not have to rely on a direct interaction with source databases (SDBs) or any middleware for its stable performance; and (5) be generic enough that it can be applied to a broad range of translational research. Achieving these goals enables our system to answer important questions that involve data generated in multiple research domains. For example, a translational researcher may ask: (1) How many patients, who were diagnosed with cancer "A" and had pathology records available, share a genetic profile "B"? (2) Which patients who have a special histological cancer type "C" and under a special treatment "D" share a distinct biomarker "E" and a unique family and exposure history? (3) Do these patients have tissue or DNA samples available and where can these samples be obtained for further studies?

Our system, the Translational Data Mart (TraM), was developed upon a domain ontology (DO) [15] integrated entity relationship model (ERM) [16,17] and it has been implemented and in use by several translational researchers. Later in the Section 6 of this paper, we will describe how the TraM system is applied in the real

* Corresponding authors. Address: Biomedical Informatics Core, Computation Institute, University of Chicago, 950 E 61st Street, Room 334, Chicago, IL 60637, USA. Fax: +1 773 834 8647 (X. Wang); +1 773 834 3834 (O.I. Olopade).
E-mail addresses: xiaoming@uchicago.edu (X. Wang), folopade@medicine.bsd.uchicago.edu (O.I. Olopade).

1532-0464/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2008.08.002

world for biomedical data integration and what the TraM data can offer to answer important research questions.

2. Terminology used in this paper

Domain data integrity means the data are "whole" or "complete" according to required information standards set by a particular research domain. For example, microarray data must meet the standards of Minimum Information About a Microarray Experiment (MIAME) as defined by the functional genomics research domain [18].

Translational integrity means that the data completion meets the minimal required standards as defined by a translational research plan, which may include data from multiple research domains. Domain data integrity does not automatically yield translational integrity.

Translational continuity refers to a special data completion status that allows one to track a single person's data from one research domain to another over a translational workflow.

A data element (DE) is an atomic element within a database. It is equivalent to an attribute in an ERM [16]. A DE is composed of two function domains: a concept domain that holds the abstract name for a set of data that share the same concept and a value domain that carries the records belonging to this concept. For example, "dosage" is a concept, "15" is a value, "unit of measure" is a concept, and "mg/day" is a value.

Translational element (TE) denotes primary identifiers of related domain databases that are mapped to each other and stored within the databases. For example, when the barcode of a tissue sample (originating from a tissue bank database) is mapped to the medical record number (collected from a clinical database) of a person from whom the sample is derived, we say that the medical record number and barcode are TEs of each other. TEs are the DE to assure translational integrity and continuity. If missing, TEs can be recovered by using other critical DEs stored in both SDBs, such as name, date of birth, race, and gender.

Data aggregation vs. data integration: *data aggregation* is the collective display of data in a unified platform, or physical collection of data within a centralized storage system from separated sources. Aggregated data may or may not relate to each other. *Data integration* is a special type of data aggregation that requires that aggregated data share TEs.

Personalized data are the data that can be identified as being associated with a distinct person, no matter how distant the data origin or derivatives are.

3. Background

3.1. Translational data status and domain database systems

The challenge of integrating source data from various research domains comes from the nature of translational workflows and the conditions of domain databases. In reality, one domain may contain zero, one, or more databases. Different databases designed for the same purpose may have distinct data structures. A database may have multiple versions and each version often results in a set of data that do not share the same data structure with others. The heterogeneity in concept extraction, data modeling, logical interpretation, naming convention, DE configuration, vocabulary used, and format definition all contribute to the challenge of data integration [19,20]. In addition, if SDBs are not designed to store TEs from other domain databases, the connections among these source data will be disrupted, even though domain data integrity within these SDBs might have been achieved. Furthermore, logically consecutive SDBs in a translational workflow often recruit biomedical records in an autonomously administrative manner. If these databases recruit data from unrelated cohorts, personalized data flow can be truncated without being noticed [21]. These problems all lead to one unwanted consequence: data are *inconsistent* in their structure and expressions and *discontinued* in their cross-domain connections. Data in such condition cannot be effectively comprehended and used without thorough cleansing, recovery, reconfiguration, and reorganization.

3.2. Data organization architectures for data integration

Several methods have been proposed to address the problems associated with integrating biomedical data. These methods include semantic mapping [22], ontology and agent methods [23], service-oriented architectures or grids [24–26], distributed search engines [27–29], and federated databases and data warehouse [19,20]. For those interoperable data sharing methods, e.g., service-oriented grids, distributed search engines, and federated databases, the availability of a service-enabled infrastructure is essential. This kind of infrastructure has not yet been established or standardized in most medical institutions. The majority of Health Insurance Portability and Accountability Act (HIPAA)-compliant SDBs are proprietary products and many have neither native web-services nor an accessible application programming interface (API), which makes immediate interoperable data extraction plan not feasible. Even if many SDBs are service-enabled, which undoubtedly will greatly enhance data aggregation ability from disparate sources, translational integrity and continuity will not be automatically achieved simply because of improved interoperability. Thorough data cleansing and verification process is likely required before data can be truly integrated and effectively used [20,23]. Furthermore, it will take a tremendous effort and time to make every required SDB in a translational research plan service-enabled. If one of these SDBs happens to be not interoperable, the data held within this SDB have to find other ways to be extracted and integrated. On the other hand, semantic mapping service has been developed to improve data standardization efficiency [22,26]. However, it alone may not be sufficient to resolve deeper problems caused by the divergence of data modeling methods.

It is generally agreed that no single data integration architecture can satisfy all demands of the entire biomedical research community. For the goals we intend to achieve, in particular to improve translational data integrity and continuity, data warehouse and federated databases are most appealing [19,20] . The two approaches are based upon entirely different design theories and result in distinct system architectures. Each of them has its strengths and limitations. Table 1 (modified from Louie et al. [20]) compares the two systems noting issues specific to translational research. For both architectures, the challenge of achieving broad system adaptability in different SDBs environments is daunting, although the coping methods are different. We believe that the heterogeneity of SDB architectures and segregation of domain database managements in different institutions will have a larger impact on federated databases than on data warehouses. Data warehouse architecture is a stand-alone system, and only access to source data is required for its basic function.

3.3. Data integration methods

Data integration methods are classified into three subtypes [19,23]: (i) information linkage, (ii) query translation, and (iii) data translation. *Information linkage* uses a URL to access data in an HTML form presented by other computation platforms through the Internet [23]. *Query translation* is meant to convert source data on the fly and present data via a virtual data organization structure