Towards to an Oncology Database (ONCOD) using a data warehousing approach

Xiaoming Wang, Ph.D [1*], Lili Liu, MS[1], James Fackenthal, Ph.D.[2], Paul Chang, MD [3], Gilliam Newstead, MD[3], Steven Chmura, MD[4], Ian Foster, Ph.D.[1], Olufunmilayo I Olopade, MD [2*]

[1]Computation Institute, University of Chicago and Argonne National Laboratory; [2]Department of Medicine, [3]Department of Radiology, [4]Department of Radiation Oncology, University of Chicago

Abstract

While data warehousing approaches have been increasingly adopted in the biomedical informatics community for individualized data integration, effectively dealing with data integration, access, and application remains a challenging issue. In this report, focusing on ontology data, we describe how to use an established data warehouse system, named TRAM, to provide a data mart layer to address this issue. Our effort has resulted in a twofold achievement: 1) a model data mart tailored to facilitate oncology data integration and application (ONCOD), and 2) a flexible system architecture that has potential to be customized to support other data marts for various major medical fields.

1. INTRODUCTION

Data integration is a constant challenge in translational science[1, 2]. In the past decade, several data integration regimes, including federated database strategies[3], workflow approaches[4], semantic web[5-7], and warehousing methods [8-11], have been tested in the biomedical informatics community. The strengths and limitations of these approaches have been carefully reviewed [12-14], and a data warehousing approach is considered most suitable because of its desired data integrity and its standalone architecture that is less affected by inadequate infrastructure environments. To date, the approach has been widely adopted in the translational informatics community: in a recent clinical translational science award (CTSA) annual meeting, 23 out of 67 abstracts were related to warehousing strategies [15]. However, there is no consensus on whether, and how, heterogeneous source data need to be processed for integration; how these data should be accessed; and how the data can be shared beyond the local setting [8-11, 15]. In the database layer, the Entity-attribute-value (EAV) scheme is commonly used to manage evolving domain concepts, in conjunction with various modeling concepts [8-10] ; this may further complicate the problem of data semantic inconsistency within, and between, warehouses. At the application level, many locally developed data warehouses do not have an end-user application interface. Users have to rely on programmers or informaticians to retrieve data on a case-by-case basis. Another kind of warehousing system, e.g., Informatics for Integrating Biology and the Bedside (I2B2) [9], provides a two-step data retrieval strategy; users are given a cohort number through some query criteria, and then are required to shape and clean the selected dataset to create their own "mini-marts" for their special needs. Whether through programmers or assisted by computation tools, obtaining data of interest on a case-by-case basis is not a cost-effective solution. In addition, if end-users cannot directly access data values in a database, the quality of this data source could be compromised due to lack of user feedback [18]. Therefore, issues concerning data integration and application of a warehouse system warrant further investigation. In this report, we introduce an alternative approach to address these issues.

Strictly speaking, a data warehouse (DW) is not a simple data repository filled with aggregated source data. Rather, it is a database that integrates data from disparate sources while delivering data with uniformity, semantic consistency, and minimized redundancy[12-14, 19]. Without meeting these criteria, the aggregated data will be of little use. Here data usability is defined as "data + meaning," which can be achieved when data are unified, standardized, connected, and validated [20-22]. To fully benefit from these data, the data mart (DM) concept is introduced to facilitate users to consume data stored in a DW, [23, 24]. oftentimes by providing a user interface for a user group with shared specific interest [25, 26]. The Star schema and Entity-Relationship (ER) schema are the major data organization schemes used to organize DW data [27, 28]. Conceptual data modeling is considered a necessary process to build a flexible warehouse schema that can satisfy various requirements [26, 29, 30]. The ER approach is frequently used in the conceptual design, due to its mathematical foundation [30-32], ability to clarify and annotate data semantics [27, 33], and its adequate support from established SQL functions and industrial grade data management tools [29, 34]. Although a group of pioneer researchers had used such modeling method to successfully manage a centralized clinical data source in 1991 [35], conceptual modeling is often overlooked in translational and clinical informatics practice.

Motivated by various translational research projects, which involve both cancer and non-cancer medical research fields, we have initiated a data warehousing project called Translational data Marts (TRAM)[11]. During the four years that TRAM has been in active use, our local cancer translational research community has further specified their informatics demands, which can be categorized as follows:

1. Researchers want to be able to search and retrieve semantically and descriptively consistent data across domains and longitudinally, and use these data for quantifiable analysis with little or no additional effort for data manipulation and cleansing.
2. Bio-specimen data need to be annotated with available clinical and translational research data.
3. Molecular research (e.g., genotyping) and phenotypic (clinical) records should be interlinked at the level of individuals if they are derived from the same persons.
4. Researchers demand to protect their data privacy for ongoing research but also want to be able to share these data with collaborators.
5. Researchers hope they can curate and annotate integrated data, and eventually develop an evidence-based knowledgebase for all cancers.

When analyzing these requests, one can realize that these specifications are, in fact, not unique to cancer researchers. However, there has not been a warehousing system available publicly to satisfy these application demands. Our TRAM system, on the other hand, has the architecture framework that can be built upon to meet these requirements. Therefore, our objective was to develop an oncology data mart (ONCOD) as a module within the TRAM system to satisfy the need of cancer researchers. Through this effort, we should be able to establish a DW/DM system that can be easily customized to support other marts for major medical fields. In this report, we first introduce our system design and the methods used to develop ONCOD. We then assess the end-results by measuring data quality and performance of ONCOD against the specifications proposed by cancer researchers. We also outline the system architecture that supports ONCOD and its potential. Additionally, we discuss lessons learned in this study, highlight unsolved problems and possible solutions of our current approach, and describe the potential application of the ONCOD/TRAM mechanism beyond our local environment.

2. METHODS

2.1 System design

We use the TRAM DW for individualized biomedical data organization and storage [11]. In conjunction, we use a semi-automated workflow to unify, standardize, and curate text data for integration. The binary datasets, however, will be integrated in an interoperable manner. We define and develop a data access layer upon the TRAM schema that can be customized to support various DMs, with each DM intending to satisfy the needs of translational researchers in a specific medical field. Therefore, the difference between marts in the TRAM system is their focused data contents, not their software architectures. Within a DM, multiple project accounts can be created to assure project-specific data privacy, while all accounts share identical data structure, domain taxonomy, and one version of data. Under each project account, multiple user roles are defined to grant unique data access privileges. The system abstraction is diagramed in Figure 1. ONCOD is the first data mart to be fully developed with this system.

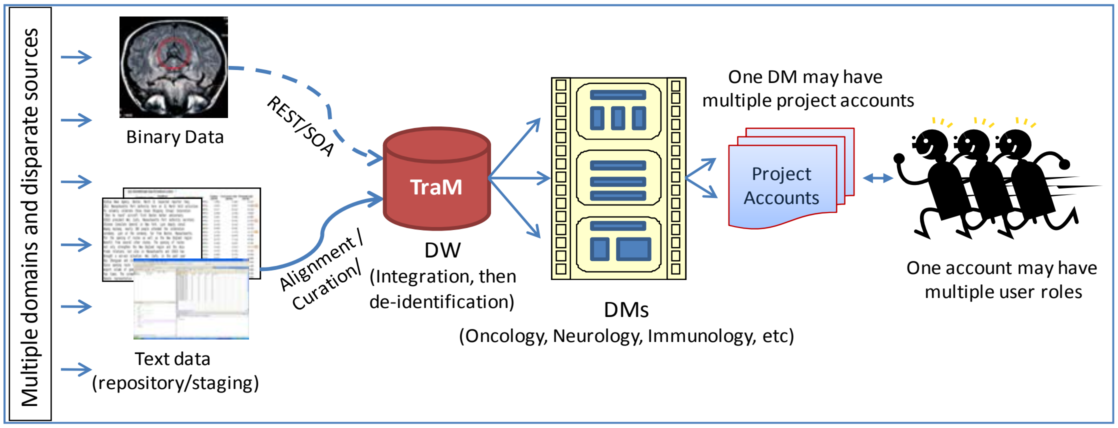


Fig. 1. The components and features of the TRAM system

2.2 System implementation

The ONCOD/TRAM system is implemented using Oracle, Sun LDAP, SSH, Java, and Apache Tomcat technologies. The system also benefits from collaboration with image data researchers, using their established Representational State Transfer (REST) API for interoperable data integration [36]. The entire system is located within an intranet behind a firewall to meet Health Insurance Portability and Accountability Act (HIPAA) compliance, as source data contains an individual's identifiable information, which is required to connect various source data derived from the same person (Fig 1). However, this identifiable information is unlinked when the system creates materialized views, upon which data marts and query programs are developed. The essential technical components to support ONCOD are a data warehouse and its web-based data application interface, and a data supply workflow.

2.3 Conceptual data model

The conceptual model of the TRAM schema is an ER abstraction (Fig 2), which lays the foundation for the entire TRAM system [11]. The conceptual modeling process involves the following:

1. Normalize domain concept entities and research object entities to allow maximal data development flexibility for both;
2. Use a many-to-many relationship data structure to associate these two types of entities for fact data collection, and utilize temporal and spatial stamps to annotate fact data;
3. Ensure individualized data integrity and continuity by enforcing an "is a part of" (one-to-many) data relationship between a person and biomaterials derived from the person.
4. Separate data values from data structures to assure that data can be queried across domains within the warehouse and retrieved from the warehouse effectively.

For the domains where expert-managed ontology is available, we adopt these concept terms and leaf class data structures of the ontologies to form domain concept entities [37, 38]. In the domains where no standard ontology is available, we provide a tree data structure to facilitate community-based and expert-driven domain taxonomy development [11]. Controlled vocabularies created in this way will be used to describe fact data. Through a relationship data structure, these data are simultaneously connected with other domain data through the ER framework (Fig 2).

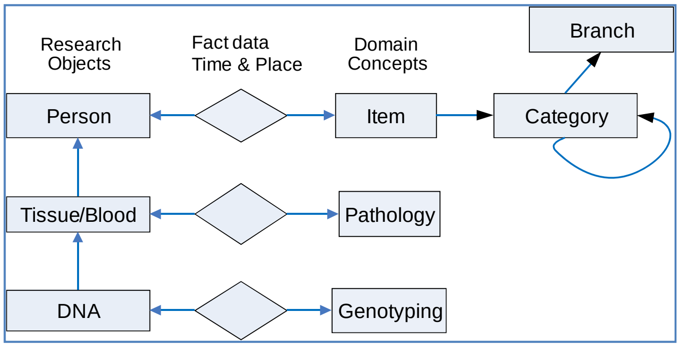


Fig. 2. Conceptual model of the TRAM schema: The rectangles indicate entities and diamond shapes indicate many-to-may relationships. The arrows point to the parent entities from which data are inherited or associated from. The domain concept alignment data structure, indicated in a branch-category-leaf hierarchy, is provided in the domains where no reputable ontology or taxonomy is available.

It is important to point out that the flexibility of this model does not rely on creating attributes in rows as the EAV approach does. Instead, it relies on normalizing domain concept and research object entities and taking advantage of relationship data structures [34, 39]. In addition, domain concept terms and their annotations in our design are treated as master data, not metadata. Therefore, concept definitions and descriptions are always available to end-users together with fact data, which further improves data clarity.

We did not use the popular Star schema to structure our data warehouse or data mart for a variety of reasons [25, 26, 30]. Using advanced relational database management technology, we can assemble various materialized views as the data layer for specific marts, and index them according to query statistics for excellent query performance.

To tailor the TRAM model for ONCOD, we paid close attention to attribute polymorphism – attribute variations existing in various medical disciplines [40]. Examples include cancer staging attributes for solid tumor and early onset mental disorder signs for neurodegenerative diseases.

2.4 Data supply workflow

The TRAM data supply workflow supports two kinds of data integration: binary and ASCII data (Fig 1). Binary data is composed of digital image files, such as MRI datasets. Such data do not require further manipulation except de-identification of heading records within the image. Therefore, we do not physically store these data in the TRAM schema. Instead, we only store the metadata about the images –such as image annotations and URI required to locate the precise image set stored in a picture archiving and communication system (PACS) in the hospital.

The other data supply mechanism is routed through an exhaustive data manipulation pipeline, focusing on text data integration from disparate sources. This workflow is designed to carry out three essential tasks: 1) unify data structure through data semantic alignment; 2) standardize data descriptors through data semantic mapping; 3) validate data on semantically aligned and mapped records. The algorithms for data semantic alignment in the workflow are domain and source data independent, so that they can be reused for data processing of the other marts. During this process, Minimum Information Required for Personalized Data Integration (MIR-PDI), carried by source-specific identifiers (e.g., medical record number, donor ID) or common data elements (birth date, gender, race, etc), needs to be recovered to assure data derived from the same person but stored in different sources can be connected. The roadmap and essential data manipulating components of the workflow is generic to all source data, while taxonomy references for the workflow may be mart specific. For example, ICDO [41, 42], Cancer Collaborative Stage Data Collection System [43], AJCC [44], and UMLS (which contains oncology terms) [45] are used throughout the workflow to support the ONCOD data processing.

2.5 Data application layer

The application layer of ONCOD is decoupled from the data layer. The display of text data is organized in a format that is easy to understand for end-users and can be conveniently applied to various statistics tools. Java Session control techniques, in particular, are implemented to facilitate versatile data accessibility by creating project accounts and assigning user roles. Ajax techniques are used to enable interactive web application, which allows users to query effectively even with minimal knowledge about query terms. Theoretically, there is no limit to the number of project accounts within a data mart.

3. RESULTS

We have implemented the entire system design sketched in Figure 1. ONCOD has been in daily use in our cancer translational research community for four years, with dynamic software improvements and regular data updates.

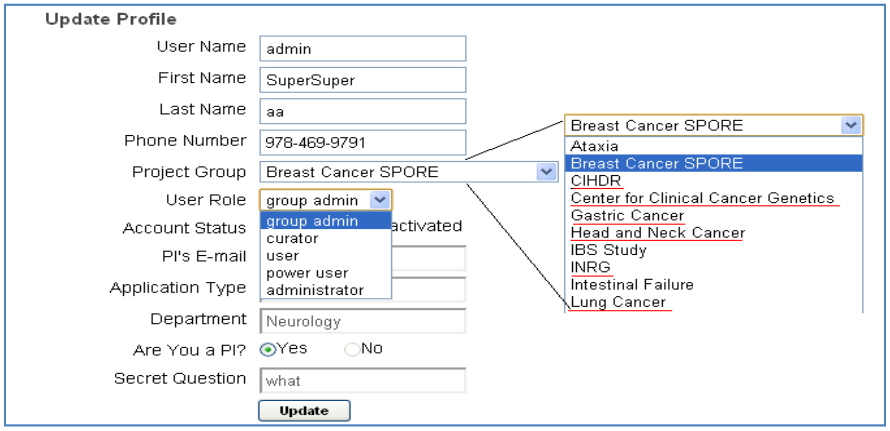
3.1 Data content of ONCOD and performance of the TRAM system

ONCOD currently recruits data only from consented patients and Institutional Research Board (IRB) approved subjects. ONCOD data are granularly formatted and contain records from 12 translational practice and research domains. These domains include epidemiology (medical questionnaire and surveys), clinical genetics (family histories and pedigree records), pathology, clinics, radiology (images and text annotations), radiation, surgery, medication, clinical laboratory testing, specimen banking, tumor signature studies, and clinical and population genotyping research. Data in these domains were collected from more than 120 sources over the past four years. To date, more than 250 million data entries from 70,000 individuals have been processed to feed ONCOD. Note that we do not physically copy DNA micro-array, genotyping, and exome data (usually 3~5 gigabytes per individual) in the TRAM schema. Instead, we only integrate annotations about these data in the system. Examples include research object identifiers (e.g., sample barcodes), experiment identifiers, experimental names and descriptions, summaries of the experimental results (e.g., confirmed genotypes or P-values), and other characteristics of high-throughput molecular biology data. Therefore, analytical results of molecular data are linked to the phenotypic data through our system. Additionally, we do not save digital image data (2~4 megabytes per dataset) for ONCOD. These data are also stored at the source, de-identified and maintained by image data experts, accessible through REST technology.

Currently, 393 query parameters, all Ajax enabled, are available through the ONCOD interface. Cancer researchers can assemble their own query commands dynamically through "query by example" mechanisms [46]. Since query commands are executed upon indexed materialized views, the performance of cross-domain or longitudinal query is excellent. For example, reporting 12,820 individuals' data from 12 domains in a query took less than 2 seconds (not all 70,000 individuals have data from multiple domains or sources). This query result not only provides a cohort number but also displays de-identified granular data in all domains on the screen. Retrieving this dataset (25.2 MB) and loading it into MS Excel took less than 5 seconds (This dataset does not include the genotyping dataset, as genotyping datasets are often too large (over 50,000 rows) to retrieve into Excel).

3.2 Allow data privacy but enforce data integrity

Access to ONCOD data is regulated by IRB protocols. The relationship between users and project accounts is shown in Figure 3, in which a "project group" refers to a project account. The account and its underlying IRB protocols control project-specific data access, while all account users share identical domain taxonomies and use one copy of data, albeit they may only access to certain individuals' data of this copy. Under each account, four kinds of user roles can be assigned to have different data access privileges: 1) assigning roles to other users (account administrator), 2) modifying data and seeing private health information (PHI) (curator), 3) viewing PHI (power user, e.g., authorized clinical researchers), and 4) viewing de-identified data (regular user).



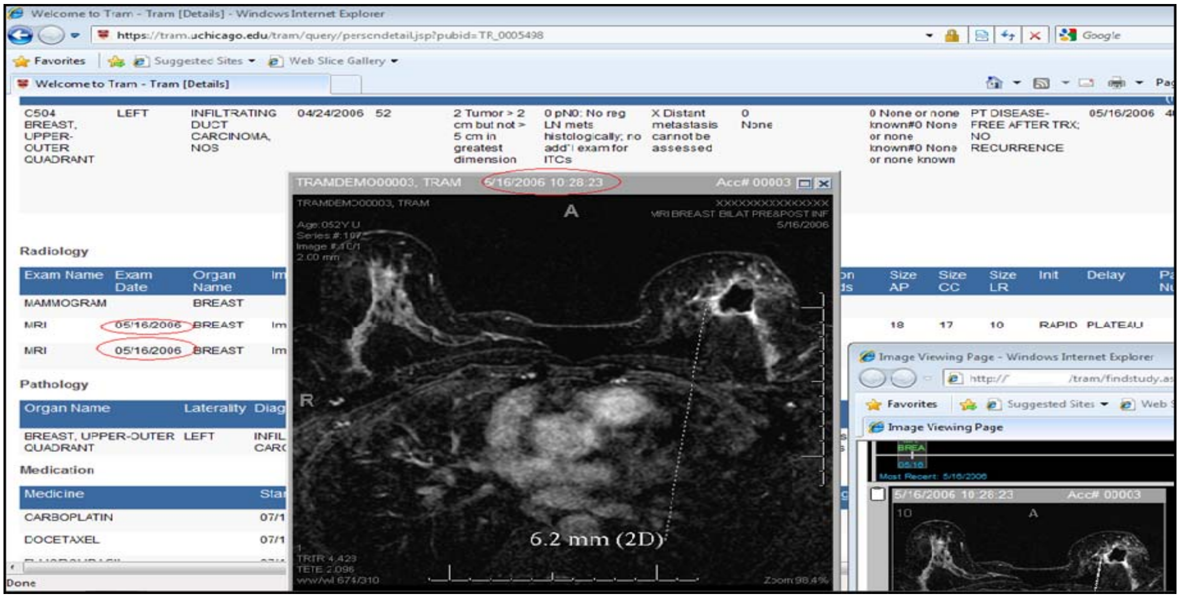
Update Profile

User Name: admin
First Name: SuperSuper
Last Name: aa
Phone Number: 978-469-9791
Project Group: Breast Cancer SPORE
User Role: group admin
Account Status: activated
PI's E-mail:
Application Type:
Department: Neurology
Are You a PI? ● Yes ○ No
Secret Question: what

Fig. 3. Screenshot of the TRAM system administrator web-interface: in the project group list, breast cancer SPORE (Special Program of Research Excellence), CIHDR (Center for Interdisciplinary Health Disparity Research), Center for Clinical Cancer Genetics, Gastric Cancer, Head and Neck Cancer, INRG (International Neuroblastoma Research Group), and Lung Cancer are the project accounts within the ONCOD mart.

3.3 Access of Binary data and ASCII data

After integration, binary and text data can be accessed through a single platform. Digital images can be located on-demand by querying their text annotation which has been previously stored within the system. Users can analyze the image using software viewer in PACS. Because of the ER framework, each image dataset is instantly integrated with the other domain text data (e.g., pathology or clinical diagnosis data) once shown on the screen (Fig 4).



Welcome to Tram - Tram [Details] - Windows Internet Explorer

https://tram.uchicago.edu/tram/query/persondetail.jsp?pubid=TR_0005498

C504 BREAST, UPPER-OUTER QUADRANT | LEFT | INFILTRATING DUCT CARCINOMA, NOS | 04/24/2006 | 52 | 2 Tumor > 2 cm but not > 5 cm in greatest dimension | 0 pN0: No reg LN mets histologically; no add'l exam for ITCs | X Distant metastasis cannot be assessed | 0 None | 0 None or none known#0 None or none known#0 None or none known | PT DISEASE-FREE AFTER TRX; NO RECURRENCE | 05/16/2006

Radiology

Exam Name	Exam Date	Organ Name	Im
MAMMOGRAM		BREAST	
MRI	05/16/2006	BREAST	Im
MRI	05/16/2006	BREAST	Im

Size AP	Size CC	Size LR	Init	Delay
18	17	10	RAPID	PLATEAU

Pathology

Organ Name	Laterality	Diag
BREAST, UPPER-OUTER QUADRANT	LEFT	INFIL CARC

Medication

Medicine	Star
CARBOPLATIN	07/1
DOCETAXEL	07/1

TRAMDEMO00003, TRAM 5/16/2006 10:28:23 Acc# 00003

TRAMDEMO00003, TRAM
Age:052Y U
Series #:1972
Image #:10/1
2.00 mm

XXXXXXXXXXXXXX
MRI BREAST BILAT PRE&POST INF
5/16/2006

TRIR 4.423
TETE 2.096
ww/wl 674/310

Zoom 98.4%

Image Viewing Page - Windows Internet Explorer

http:// /tram/findstudy.as

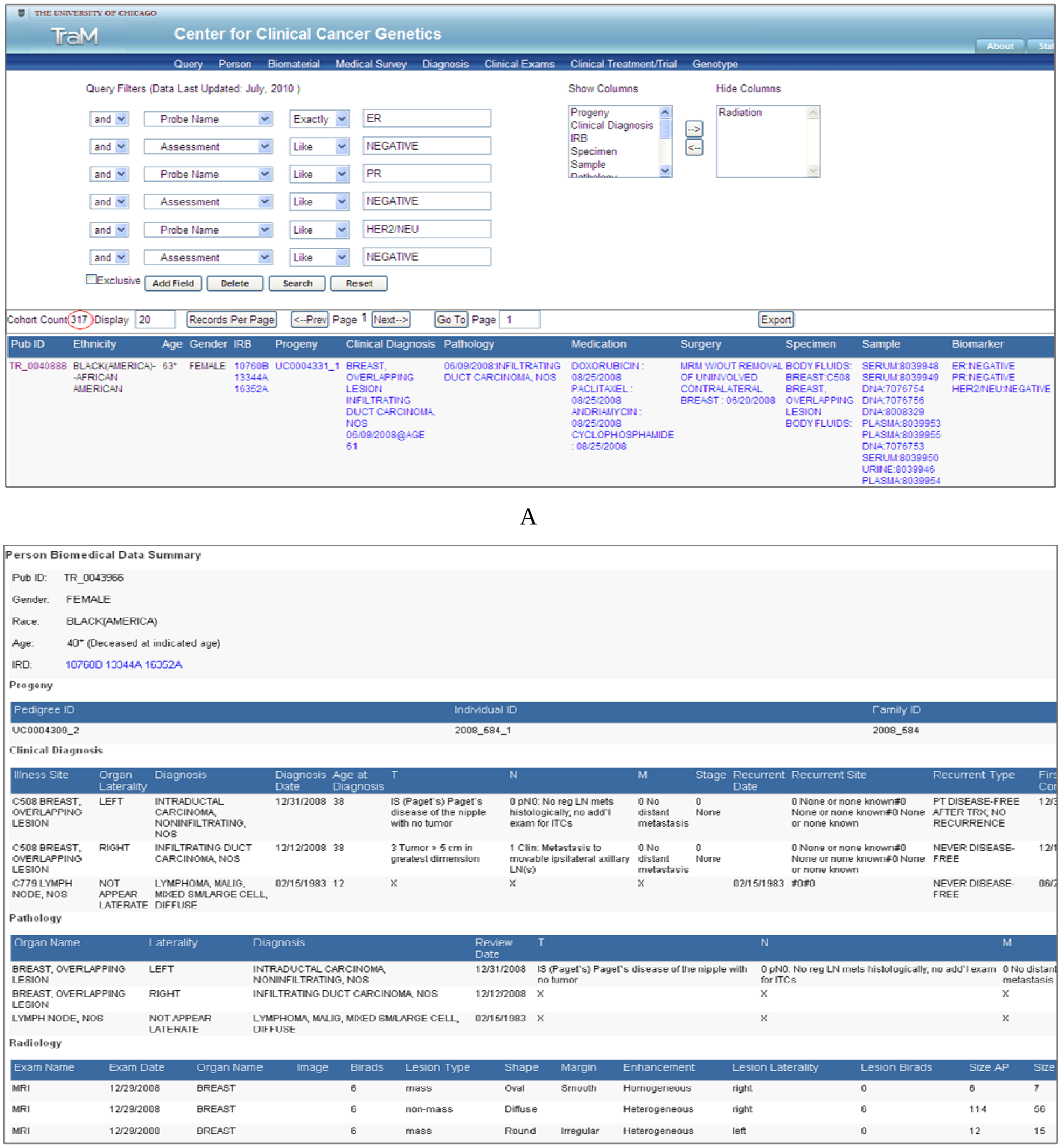
Most Recent: 5/16/2006

5/16/2006 10:28:23 Acc# 00003

Fig. 4. On-demand image data integration through REST protocol: the image was taken in the axial place, the patient identifiable information has been removed from the digital picture, and the lesion at the left breast is indicated on the image, which can be compared with the clinical and pathology records on the screen.

3.4 Longitudinal and cross-domain data continuity

Individualized ONCOD/TRAM data are measured through the integrity and continuity of cross-domain and longitudinal data. Quality of data in these two dimensions is shown in Figure 5. The face value of data scope, consistency, and granularity in various domains is shown in Figure 5A. The longitudinal continuity of individuals' data is presented in Figure 5B.



Center for Clinical Cancer Genetics

Query Filters (Data Last Updated: July, 2010)

and	Probe Name	Exactly	ER
and	Assessment	Like	NEGATIVE
and	Probe Name	Like	PR
and	Assessment	Like	NEGATIVE
and	Probe Name	Like	HER2/NEU
and	Assessment	Like	NEGATIVE

□ Exclusive

Show Columns: Progeny, Clinical Diagnosis, IRB, Specimen, Sample, Patholog

Hide Columns: Radiation

Cohort Count 317 Display 20 Page 1 Page 1

Pub ID	Ethnicity	Age	Gender	IRB	Progeny	Clinical Diagnosis	Pathology	Medication	Surgery	Specimen	Sample	Biomarker
TR_0040888	BLACK(AMERICA)--AFRICAN AMERICAN	63*	FEMALE	10760B 13344A 16352A	UC0004331_1	BREAST, OVERLAPPING LESION INFILTRATING DUCT CARCINOMA, NOS 06/09/2008@AGE 61	06/09/2008:INFILTRATING DUCT CARCINOMA, NOS	DOXORUBICIN : 08/25/2008 PACLITAXEL : 08/25/2008 ANDRIAMYCIN : 08/25/2008 CYCLOPHOSPHAMIDE : 08/25/2008	MRM W/OUT REMOVAL OF UNINVOLVED CONTRALATERAL BREAST : 06/20/2008	BODY FLUIDS: BREAST:C508 BREAST, OVERLAPPING LESION BODY FLUIDS:	SERUM:8039948 SERUM:8039949 DNA:7076754 DNA:7076756 DNA:8008329 PLASMA:8039953 PLASMA:8039955 DNA:7076753 SERUM:8039950 URINE:8039946 PLASMA:8039954	ER:NEGATIVE PR:NEGATIVE HER2/NEU:NEGATIVE

A

Person Biomedical Data Summary

Pub ID: TR_0043966
Gender: FEMALE
Race: BLACK(AMERICA)
Age: 40* (Deceased at indicated age)
IRB: 10760B 13344A 16352A

Progeny

Pedigree ID	Individual ID	Family ID
UC0004309_2	2008_584_1	2008_584

Clinical Diagnosis

Illness Site	Organ Laterality	Diagnosis	Diagnosis Date	Age at Diagnosis	T	N	M	Stage	Recurrent Date	Recurrent Site	Recurrent Type	Firs Cor
C508 BREAST, OVERLAPPING LESION	LEFT	INTRADUCTAL CARCINOMA, NONINFILTRATING, NOS	12/31/2008	38	IS (Paget`s) Paget`s disease of the nipple with no tumor	0 pN0: No reg LN mets histologically; no add'l exam for ITCs	0 No distant metastasis	0 None		0 None or none known#0 None or none known#0 None or none known	PT DISEASE-FREE AFTER TRX; NO RECURRENCE	12/3
C508 BREAST, OVERLAPPING LESION	RIGHT	INFILTRATING DUCT CARCINOMA, NOS	12/12/2008	38	3 Tumor > 5 cm in greatest dimension	1 Clin: Metastasis to movable ipsilateral axillary LN(s)	0 No distant metastasis	0 None		0 None or none known#0 None or none known#0 None or none known	NEVER DISEASE-FREE	12/1
C779 LYMPH NODE, NOS	NOT APPEAR LATERATE	LYMPHOMA, MALIG, MIXED SM/LARGE CELL, DIFFUSE	02/15/1983	12	X	X	X		02/15/1983	#0#0	NEVER DISEASE-FREE	06/2

Pathology

Organ Name	Laterality	Diagnosis	Review Date	T	N	M
BREAST, OVERLAPPING LESION	LEFT	INTRADUCTAL CARCINOMA, NONINFILTRATING, NOS	12/31/2008	IS (Paget`s) Paget`s disease of the nipple with no tumor	0 pN0: No reg LN mets histologically; no add'l exam for ITCs	0 No distant metastasis
BREAST, OVERLAPPING LESION	RIGHT	INFILTRATING DUCT CARCINOMA, NOS	12/12/2008	X	X	X
LYMPH NODE, NOS	NOT APPEAR LATERATE	LYMPHOMA, MALIG, MIXED SM/LARGE CELL, DIFFUSE	02/15/1983	X	X	X

Radiology

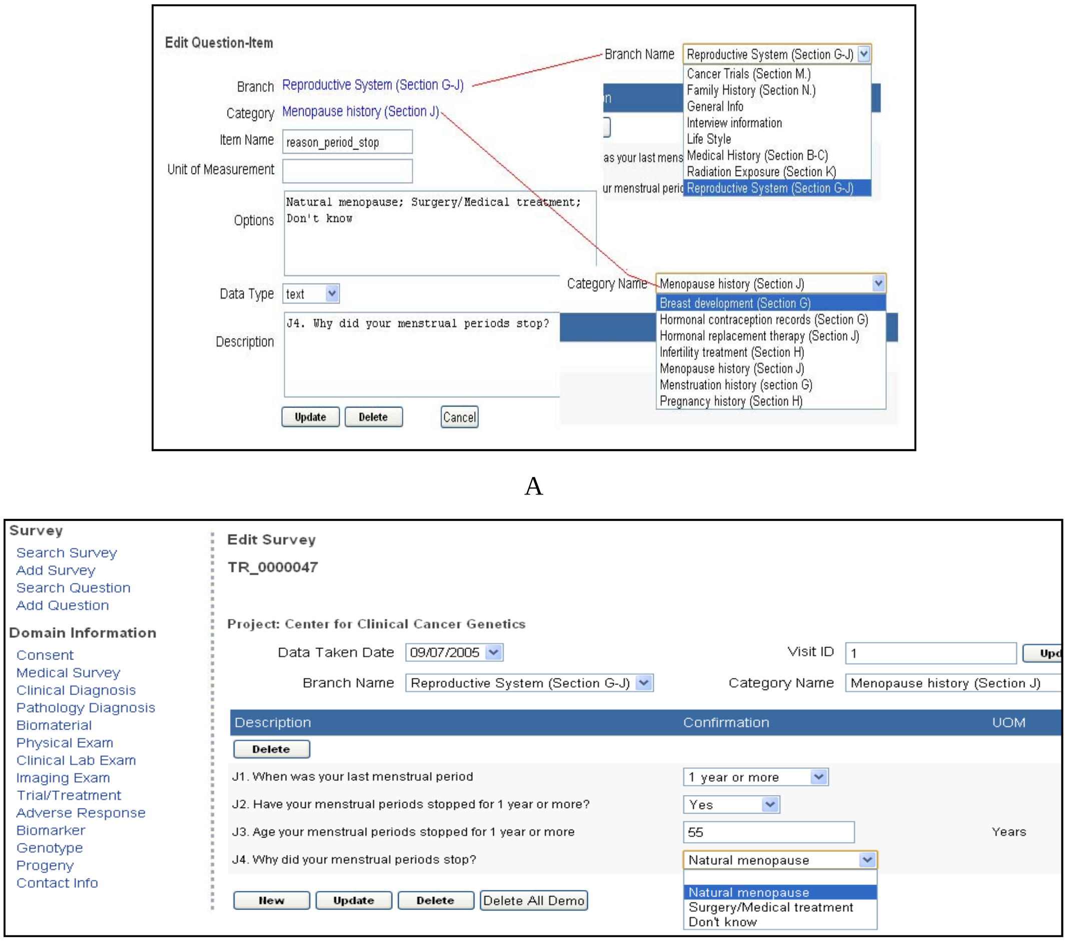
Exam Name	Exam Date	Organ Name	Image	Birads	Lesion Type	Shape	Margin	Enhancement	Lesion Laterality	Lesion Birads	Size AP	Size
MRI	12/29/2008	BREAST		6	mass	Oval	Smooth	Homogeneous	right	0	6	7
MRI	12/29/2008	BREAST		6	non-mass	Diffuse		Heterogeneous	right	6	114	56
MRI	12/29/2008	BREAST		6	mass	Round	Irregular	Heterogeneous	left	0	12	15

B

Fig. 5. Screenshots of the ONCOD/TRAM user interface: Panel A indicates cross-domain data continuity, displaying the project account name in the header: "Center for Clinical Cancer Genetics," showing a cohort number in this query return (circled in red), and exhibiting data contents on the screen; Panel B shows longitudinal data integrity of a person's data.

3.5 Iterative data curation

One unique feature of ONCOD is that it allows users to curate data, which is necessary to building a knowledgebase. Two kinds of curation activities are supported through the ONCOD user interface. One is driven by domain experts (e.g., MD or PhD researchers), who are allowed to edit domain taxonomies by working with an informatician who understands data property abstraction and semantic alignment. Another is operated by a data manager, who can only edit fact data by using taxonomies either adopted or predefined by local experts. This feature greatly boosts flexibility of adding new domain concepts in a domain where no standard ontology is available, and at the same time assures that only one set of taxonomy will be created in this particular domain and the taxonomy is maintained with controlled vocabulary. The curator interface for creating a new domain concept and using the concept to recruit new data or change data is shown in Figure 6. By doing so, we have successfully aligned seven heterogeneous medical survey datasets collected over the past 19 years, and integrated data from disparate sources for tens of thousands of individuals.



Edit Question-Item

Branch: Reproductive System (Section G-J)

Category: Menopause history (Section J)

Item Name: reason_period_stop

Unit of Measurement:

Options: Natural menopause; Surgery/Medical treatment; Don't know

Data Type: text

Description: J4. Why did your menstrual periods stop?

Branch Name: Reproductive System (Section G-J)

Category Name: Menopause history (Section J)

Cancer Trials (Section M.)
Family History (Section N.)
General Info
Interview information
Life Style
Medical History (Section B-C)
Radiation Exposure (Section K)
Reproductive System (Section G-J)
Breast development (Section G)
Hormonal contraception records (Section G)
Hormonal replacement therapy (Section J)
Infertility treatment (Section H)
Menopause history (Section J)
Menstruation history (section G)
Pregnancy history (Section H)
Update
Delete
Cancel

A

Edit Survey

TR_0000047

Project: Center for Clinical Cancer Genetics

Data Taken Date: 09/07/2005

Visit ID: 1

Branch Name: Reproductive System (Section G-J)

Category Name: Menopause history (Section J)

Description	Confirmation	UOM
J1. When was your last menstrual period	1 year or more	
J2. Have your menstrual periods stopped for 1 year or more?	Yes	
J3. Age your menstrual periods stopped for 1 year or more	55	Years
J4. Why did your menstrual periods stop?	Natural menopause	

B

Fig. 6. Screenshots of the curator interface: Panel A is for domain experts to define questionnaire hierarchy and answers to question; Panel B is for data managers to use pre-defined questions and answer options to collect survey data. Data for both questionnaire and survey data are generally integrated using computational methods. The curators mainly use this interface for ad hoc (small batch) data integration or for error correction. Each data entry person is limited to edit data in the account s/he is granted access to. When a person's data is shared by multiple accounts, a warning message will appear to block the editing attempt until all account managers agree upon the modification on the records.

3.6 Usability of ONCOD data

We have shown data uniformity and consistency, and their individualized cross-domain and longitudinal integrity in Section 3.4. Without these criteria, it is impossible to use even a modest dataset collected from two or more sources. Here we use data from a breast cancer project account shared by 18 principal investigators to further describe the impact of the ONCOD effort on data usability.

In this account, each domain dataset is collected from at least two sources. In the past four years, data of 12 domains (detailed in Section 3.1) from 13379 individuals between 1992 to present have been collected. All bio-specimens (28641 specimens and sample records from 7383 individuals) are fully annotated with available demographic, clinical, and basic research information. Furthermore, genotyping results, clinical records, family history, medical surveys, and pedigree annotations for 1767 persons from more than 16 sources collected over the past 20 years have also been integrated at the level of individuals. Since users of this account are also the data contributors, they themselves participate in data curation. The curated data stay in the TRAM system to benefit all users within the account, and some of these data are also shared with other accounts or external users through researchers' collaborations. Currently, ONCOD data are mainly used for hypothesis generation, grant applications, research planning, bio-specimen searching and sharing, and multidisciplinary and cross-institutional collaborations.

4. DISCUSSION

In order to find a more effective way to manage and use warehouse data, we have developed a model data mart, ONCOD, within our previously developed data warehouse system, TRAM. The capability of ONCOD and the quality of ONCOD data has met all user specified requirements described in Section 1 (Section 1). Although the ONCOD data is still modest in size, the ONCOD/TRAM system has been tested to successfully integrate very complex data contents. We believe that ONCOD will further reveal its strength when more data are consented and large quantities of data become available. Through the ONCOD effort, we have also enhanced the warehouse system architecture with a data access layer that can be customized to support other marts with cost-effective effort. In addition, ONCOD is built upon a set of materialized views created from the TRAM schema, so both share identical attribute semantics and one version of truth about data. Therefore, no additional workload is required to deal with potential data inconsistency between marts and a data warehouse, which allows the system to gain long-term benefits from this costly endeavor.

From the ONCOD project, we have learned two major lessons. 1) Conceptual data modeling is critical for a sustainable warehouse and we have benefitted enormously from this design investment. However, the overhead of delivering a stabilized ER-based conceptual model is larger than that of the EAV approach. Because the ER approach demands predefinition of data elements, one should expect iterative data model improvement, especially in the early stage of system development. 2) To build a translational-scale DW instead of a clinical data dump, it is necessary to analyze a wide spectrum of biomedical domain data in large quantities from multiple sources. The representative data properties can only be identified when the data complexity and quantity is sufficient enough to reveal their common characteristic profiles.

One of the challenges of using TRAM is to feed quality data to the system. Persistent heterogeneity of source data, dynamically evolving data sources, fragility of data access channels, and lack of eligible standards for integrated data presents a daunting challenge to data processing professionals on a daily basis. The ER-based TRAM schema requires column (attribute) semantic homogeneity, which adds another layer of difficulty for data integration, as heterogeneous source data need to be aligned to the predefined semantics of the TRAM model. Although our data transformation workflow has shown consistent throughput capacity and quality of output data, it lacks robustness in error-handling and reasoning ability in semantic alignment. In addition, specific manpower (with domain knowledge and data modeling skills) is required to operate the workflow, limiting the portability of ONCOD/TRAM to those who can provide it. However, if the TRAM system is implemented in a private (secured) Cloud and maintained by professionals, more informaticians and end-users will benefit from this service. Technically, cloud-enabled data warehousing is a better way to improve data semantic interoperability than the current silo setting and autonomous operation in warehousing practice.

From database to dataspace, the TRAM (or TRAM-like) system could function as a data conduit to bridge the wide gap between operational data sources and the semantic web. The ER-based TRAM schema can be easily converted into a resource description framework (RDF) scheme. Together with its controlled vocabulary or ontology-driven data description, this kind of warehouse can be an ideal resource for effective data exchange or web-scale data integration, if all regulatory requirements are met.

Corresponding Authors

Xiaomng Wang: xiaoming@uchicago.edu; Olufunmilayo I Olopade: folopade@medicine.bsd.uchicago.edu

Availability

http://tram.uchicago.edu

Acknowledgement

This work was supported by NIH Grants 1-P50-CA125183 and U54 RR023560-02. Authors thank Peng Liu for providing REST API, Kyle Chard for thoughtful comments, and Cecilia Zora for editing the manuscript.

References

1. Horig H, Marincola E, Marincola FM. Obstacles and opportunities in translational research. Nat Med. 2005 Jul;11(7):705-8.

2. Evans WE, Relling MV. Moving towards individualized medicine with pharmacogenomics. Nature. 2004 May 27;429(6990):464-8.

3. Bug WJ, Astahkov V, Boline J, et al. Data Federation in the biomedical informatics research network: tools for semantic annotation and query of distributed multiscale brain data. AMIA Annu Symp Proc. 2008:1220.

4. Tiwari A, Sekhar AK. Workflow based framework for life science informatics. Comput Biol Chem. 2007 Oct;31(5-6):305-19.

5. Mukherjea S. Information retrieval and knowledge discovery utilising a biomedical Semantic Web. Brief Bioinform. 2005 Sep;6(3):252-62.

6. Deus HF, Stanislaus R, Veiga DF, et al. A Semantic Web management model for integrative biomedical informatics. PLoS ONE. 2008;3(8):e2946.

7. DiBernardo M, Pottinger R, Wilkinson M. Semi-automatic web service composition for the life sciences using the BioMoby semantic web framework. J Biomed Inform. 2008 Oct;41(5):837-47.

8. Lowe HJ, Ferris TA, Hernandez PM, et al. STRIDE--An integrated standards-based translational research informatics platform. AMIA Annu Symp Proc. 2009 Sep-Oct
Nov-Dec;2009(5):391-5.

9. 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 Mar-Apr;17(2):124-30.

10. Chute CG, Beck SA, Fisk TB, Mohr DN. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. J Am Med Inform Assoc. 2010 Mar-Apr;17(2):131-5.

11. Wang X, Liu L, Fackenthal J, et al. Translational integrity and continuity: personalized biomedical data integration. J Biomed Inform. 2009 Feb;42(1):100-12.

12. Louie B, Mork P, Martin-Sanchez F, Halevy A, Tarczy-Hornoch P. Data integration and genomic medicine. J Biomed Inform. 2007 Feb;40(1):5-16.

13. Goble C, Stevens R. State of the nation in data integration for bioinformatics. J Biomed Inform. 2008 Oct;41(5):687-93.

14. Stein LD. Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. Nat Rev Genet. 2008 Sep;9(9):678-88.

15. SOURCE. Meeting Abstracts: 23 out of 67 abstracts from different institutions are related to data warehousing 2011 CTSA Annual Informatics Meeting in Natcher Conference Center, October 12-13, 2011 2011:https://www.signup4.net/Upload/BOOZ12A/CTSA42E/CTSA_IKFC_2011_Abstracts_Oct07.pdf.

16. Nadkarni PM, Brandt C. Data extraction and ad hoc query of an entity-attribute-value database. J Am Med Inform Assoc. 1998 Nov-Dec;5(6):511-27.

17. Nadkarni PM, Marenco L, Chen R, Skoufos E, Shepherd G, Miller P. Organization of heterogeneous scientific data using the EAV/CR representation. J Am Med Inform Assoc. 1999 Nov-Dec;6(6):478-93.

18. Quackenbush J. The power of public access: the human genome project and the scientific process. Nat Genet. 2001 Sep;29(1):4-6.

19. Bernstein PA, Haas LM. Information integration in the enterprise. Commun ACM. 2008;51(9):72-9.

20. Bernstam EV, Smith JW, Johnson TR. What is biomedical informatics? J Biomed Inform. 2010 Aug 13;43(1):104-10.

21. Bellinger G, Castro D, Mills A. Data, Information, Knowledge, and Wisdom. www systems-thingking 2004;http://www.systems-thinking.org/dikw/dikw.

22. Ackoff R. From data to wisdom. J Appl Syst Anal 1989;16(1):3-9.

23. BONIFATI A, CATTANEO F, CERI S, FUGGETTA A, PARABOSCHI S. Designing data marts for data warehouses. ACM Trans Softw Eng Methodol. 2001;10(4):452-83.

24. Inmon W. Data Mart Does Not Equal Data Warehouse. http://wwwinformation-managementcom/infodirect/19991120/1675-1html. 1999.

25. Inmon WH, Inmon WH. Building the data warehouse (3nd ed.): John Wiley and Sons, Inc.; 2002.

26. Imhoff C, Geiger JG, Galemmo N. Relational Modeling and Data Warehouse Design: John Wiley \\& Sons, Inc.; 2003.

27. Inmon WH, Rudin K, Buss CK, Sousa R. Data warehouse performance: John Wiley and Sons, Inc.; 1999.

28. Kimball R, Reeves L, Thornthwaite W, Ross M, Thornwaite W. The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing and Deploying Data Warehouses with CD Rom: John Wiley \&amp; Sons, Inc.; 1998.

29. Golfarelli M, Rizzi S. A comprehensive Approach to Data Warehouse Testing DOLAP'09 2009 November 6, 2009;November 6, 2009:17-24.
30. Kamble AS. A conceptual model for multidimensional data. Proceedings of the fifth on Asia-Pacific conference on conceptual modelling - Volume 79. 2008(79):29-38.
31. Codd EF. A relational model of data for large shared data banks. Commun ACM. 1970;13(6):377-87.
32. Chen PP-S. The entity-relationship model: a basis for the enterprise view of data. Proceedings of the June 13-16, 1977, national computer conference. Dallas, Texas: ACM; 1977.
33. Bornberg-Bauer E, Paton NW. Conceptual data modelling for bioinformatics. Brief Bioinform. 2002 Jun;3(2):166-80.
34. Chen PP. Entity-Relationship Modeling: Historical Events, Future Trends, and Lessons Learned Software Pioneers 2002. p. 100-14.
35. Johnson S, Friedman C, Cimino JJ, Clark T, Hripcsak G, Clayton PD. Conceptual data model for a central patient database. Proc Annu Symp Comput Appl Med Care. 1991:381-5.
36. Fielding RT, Taylor RN. Principled Design of the Modern Web Architecture. *ACM Transactions on Internet Technology* 2002;**2** (2):115-50.
37. Sugumaran V, Storey VC. The role of domain ontologies in database design: An ontology management and conceptual modeling environment. ACM Trans Database Syst. 2006;31(3):1064-94.
38. Noy NF. Semantic Integration: A Survey Of Ontology-Based Approaches. SIGMOD Record, December, 04 2004 December, 2004;33(4):65-70.
39. Codd EF. Normalized data base structure: a brief tutorial. Proceedings of the 1971 ACM SIGFIDET (now SIGMOD) Workshop on Data Description, Access and Control. San Diego, California: ACM; 1971.
40. Cardelli L, Wegner P. On understanding types, data abstraction, and polymorphism. ACM Comput Surv. 1985;17(4):471-523.
41. SEER:. ICD-O-3 Coding Materials. http://seercancergov/icd-o-3/.
42. WHO. International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3). http://wwwwhoint/classifications/icd/adaptations/oncology/en/.
43. CS:. http://web2.facs.org/cstage/schemalistabc.html. 2010.
44. Edge SB, Byrd DR, Compton CC, Fritz AG, Greene FL, Trotti A, editors. AJCC Cancer Staging Handbook. 7th ed: Springer; 2010.
45. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267-70.
46. Zloof M. Query by example: the innovation and definition of tables and forms. Proceedings of the 1st international conference on very large databases. 1975:1-24.