MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-588-4-193

Methodology of integration of a clinical data warehouse with a clinical information system: the HEGP case

Eric Zapletal[a], Nicolas Rodon[a], Natalia Grabar[ab], Patrice Degoulet[ab]

[a] *Department of Medical Informatics, Georges Pompidou University Hospital, Paris, F-75015 France*
[b]*UMR-S 872, Éq. 20, Centre de Recherche des Cordeliers, INSERM, Université Paris Descarte , Paris, F-75006 France*

Abstract

Clinical Data Warehouses (CDW) can complement current Clinical Information Systems (CIS) with functions that are not easily implemented by traditional operational database systems. Here, we describe the design and deployment strategy used at the Pompidou University Hospital in southwest Paris. Four realms are described: technological realm, data realm, restitution realm, and administration realm. The corresponding UML use cases and the mapping rules from the shared integrated electronic health records to the five axes of the i2b2 CDW star model are presented. Priority is given to the anonymization and security principles used for the 1.2 million patient records currently stored in the CDW. Exploitation of a CDW by clinicians and investigators can facilitate clinical research, quality evaluations and outcome studies. These indirect benefits are among the reasons for the continuous use of an integrated CIS.

Keywords:

Clinical data warehouse, Clinical information system, Heterogeneous data integration, Data security.

Introduction

Clinical Data Warehouses (CDW) were defined in the early 90s as subject oriented, integrated, time-variant, non volatile collections of data used in support of management decisions [1]. They have been extensively used in the industrial field [2] and much more recently in healthcare. CDW now play an important role when addressing issues related to the integration of heterogeneous databases, considered essential for biomedical research [3]. For example, new types of data, arising from tissue bank management systems [4] or biomedical research [5], need to be integrated with legacy data sources to promote networked or translational research.

Although there are examples of studies demonstrating the benefits of using a CDW for clinical and biomedical researches [6-8], the direct reuse of clinical information systems (CIS) for this purpose is still very rare [9].

Here, we present the methodology used at the Georges Pompidou European Hospital (HEGP) to implement a CDW that is closely integrated with the hospital CIS.

Materials and Methods

The HEGP clinical information system

HEGP is an 890-beds university hospital in southwest Paris, France. The total number of employees is 3,200, with 400 full-time equivalent physicians. The mean number of inpatient admissions/month is 4,700 and of outpatient visits/month is 18,000.

The HEGP CIS is built upon on a component-based approach [10]. The healthcare-related components include the patient ADT, healthcare record (EHR), and act management (CPOE) components from MEDASYS®, and an appointment-resource scheduling component ONECALL© from McKesson®. Healthcare components are integrated through an enterprise application integration platform (EAI) developed by THALES® including a reference manager where the different concepts and their relationships are declared. The CIS is accessed from the 3000 PCs, laptops, and thin clients through a common portal associated with a HL7/CCOW manager and a security component where user access rights are described.

A top down design approach

The deployment of this CDW is a four-year project that started in 2008. One of its objectives is to establish a global methodology for integrating a CDW into comparable French institutions. Using a top-down approach (Table 1), we decided to distinguish four domains or realms (technical, data, restitution and administration) in which we successively: 1) model UML use cases related to the CDW design; 2) analyze compliant technological frameworks, and 3) implement functionalities with selected tools.

This approach is described in detail below.

Table 1- HEGP top-down approach to integrate clinical data warehouse into the clinical information system

	Technical realm	Data realm	Restitution realm	Administration realm
UML use cases	To ensure security and data protection To optimize usability of IT tools To tune data flows and data volumetry	To make source data available To make target data available	To design and broadcast CDW objects	To manage the life cycle of CDW objects
Technological frameworks	French National body for security recommendations; strong reversible encryption algorithm Open source software; collaborative developments Database mirroring	Navigation through database models; business layers analysis Mapping of data sources to target CDW schema; datamart creation	Business intelligence solutions; enterprise content management systems	Help desk system
Implementation tools	Bouncy Castle® java library I2B2® framework ORACLE® utilities	SchemaSpy; SAP®/BO Universes ORACLE® (materialized) views; Talend®/OS ETL jobs; DocuWiki	SAP®/BO reports; KaliTech®/Kalidoc; I2B2 client	Peregrine®/GPS

UML use cases

The technical realm

The technical realm concerns the deployment of CDW hardware and software infrastructures. We have identified three corresponding top-level use cases:

- To ensure security and data protection: we assume that users with high level permissions (physicians for example) are allowed access to all data concerning their patients, but that users with low level permissions (statisticians for example) should only have access to anonymized data. We refined these use case by analyzing each of the recommendations of the French coordinating body for privacy and security (CNIL) and their impact on every CDW activity [11].
- To tune data flows and database volumetry.
- To provide easy-to-use IT tools: one of the main objectives of the CDW project is to facilitate user access to the clinical data. Therefore, the ergonomics and the usability of the IT tools provided are very important. In particular, they must solve the "not enough time" issue [12].

The data realm

The data realm covers the management of the data sources that feed the CDW, including how data are modeled, accessed, and integrated into the CDW. We identified two corresponding top-level use cases:

- To make the source data available: this use case concerns physical access to data, model documentations, potential reverse engineering techniques and cooperation with applications vendors.
- To make the target data available: in this use case, the focus is on the capacity of the CDW model to support data analysis and design of new objects (reports, indicators, etc.), i.e. to move from a storage/access-oriented data model of operational medical applications to the analysis-oriented model of a CDW.

The restitution realm

The restitution is related to the manner in which users have access to the CDW objects and how new information is generated in the CDW. We identified one top-level use case in this realm:

- To design and to broadcast reports.

The administration realm

The administration realm deals with the exploitation (supervision, parameterization, evolution and maintenance) of the CDW objects. From a quality of service point of view, with the objective of managing the activity of the CDW as a service provided by the IT team to the end users, the main use case is:

- To manage the life cycle of the CDW objects

Results

CDW technological frameworks and implementation tools

The technical realm

To fulfill the constraints related to source code accessibility and collaborative development possibilities, we decided to focus on Open Source software components, whenever possible. A CentOS/Linux operating system and the I2B2 [13] framework upon an ORACLE® database for the storage were selected and installed as the core CDW infrastructure.

The first security principle we applied was the anonymization of all data identifying patients, both for structured data or for non structured data (free text reports). In some cases, it is necessary to restitute patient identifiers, so anonymization has to be reversible. All non identifying data (biological results, physician order entries, ICD-10 codes, etc.) are available without restriction. Patient-identifying data are encrypted with a strong reversible cryptographic procedure based on a AES algorithm implemented using the Bouncy Castle java libraries [14]. For experimentation, we are using a 128 bit-long AES key. This encryption function has been integrated into the jobs used for data migration that also use the java language.

For structured data, HIS patient identifiers and hospital stay identifiers are also eligible to anonymization. These two identifiers are used in the CDW as primary keys in the database tables and the encrypted identifiers are bigger (in terms of internal machine representation) than the non-encrypted identifiers. We therefore modified the storage format of some I2B2 tables in the ORACLE® database. We modified the I2B2 client accordingly to manage the new format of patient identifiers.

For non structured data sources (e.g., text reports from in/outpatient stays or surgical interventions), anonymization involves five successive steps:

1. Extraction of Microsoft® WORD-based text reports from the EHR database: a first PHP script using MS ActiveX objects extracts and uncompresses the native WORD-based text reports stored in the DxCare®/Medasys EHR database.
2. Conversion from WORD format into DocBook XML format with the ANTIWORD program.
3. Tagging of identifying data: a NLP tool is used to mark identifying data inside DocBook XML-based reports with new tags.
4. Encrypting of tagged data: the AES algorithm is used to encrypt reversibly tagged data inside XML-based reports with the 128-bit AES key.
5. Integration of encrypted XML-based reports in the CDW database with a second PHP script.

The data realm

As the first milestones of security and data protection were established, we assessed integrating real patient data into the CDW. Data realm implementation consists of three steps.

Step 1 – Identification of data sources

This step requires knowledge and expertise from legacy data sources necessitating close collaboration with the various application vendors.

Step 2 – Mapping of data sources into the target schema

The target schema is the five-axis star schema of I2B2 (PATIENT, PROVIDER, VISIT, CONCEPT and OBSERVATION) [13]. PATIENT, PROVIDER and VISIT related data are easily extracted from the data sources and mapped to the target schema because our CIS components already structure their data according to these axes and the corresponding items can be almost directly integrated into the I2B2 schema.

CONCEPT mapping

The attributes of the data source related to the CONCEPT axis address the issue of the classifications and terminologies used; these are very intimately linked to the business logic with which the data are created. In the I2B2 framework, each classification or terminology must be transformed into a tree structure (i.e. a hierarchy with single inheritance links). Therefore, the initial phase of CONCEPT mapping consists in providing:

1. A name for the new classification used by the data source;
2. A SQL request (r1) that returns a raw dataset including 1) the names of concepts in the new classification; 2) the identifiers for the concepts; 3) the identifiers of the parents for each concept.

Then, each line returned by (r1) must be transformed into I2B2 storage format with the tree structure constraint: the location in the classification is managed by a text field that is calculated as the sequence of the traversed levels from the root of the hierarchy down to the associated CONCEPT. This step can be performed by a second SQL request (r2).

OBSERVATION mapping

The attributes of the source data related to the OBSERVATION axis are the primary data subject to computation and analysis in the CDW. The initial phase of the OBSERVATION mapping is to provide a SQL request (r3) that extracts raw dated facts which comply with the four following constraints:

1. They must be related to a (single) patient.
2. They must be temporally dated: data that are not dated (for example the first name and the last name of a patient) are theoretically not eligible to be stored in the OBSERVATION axis.
3. They must be "generated" within the scope of the healthcare process, with an identified creator or source (the PROVIDER) and with an identified medical event (the VISIT), for example a consultation or a hospitalization.
4. They must be associated with a concept (an anchor in a classification loaded in the CONCEPT mapping step).

Then, as in the case of CONCEPT mapping, each line returned by (r3) must be transformed to meet I2B2 table format constraints (r4).

For each new DATA SOURCE, we have implemented these four different requests with dedicated ORACLE views using copies of the CIS databases (Figure 1). Copies are managed with ORACLE® import/export utilities to prevent direct use of production databases. The views are used in the Open Source Talend® ETL jobs when data are migrated. A master job encapsulates all ETL jobs to provide a single access point for integrating all CDW data at once.

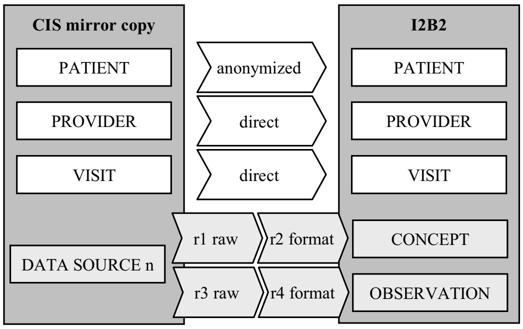

Figure 1 – Overall synopsis of data mapping

Step 3 - Building of dedicated datamarts

An important issue is the availability of data at the relevant level of granularity [15]. In the CDW, data are stored at the lowest possible level of granularity (as observation facts) and this may not be appropriate for users who want highly aggregated data. Hence, when necessary, dedicated datamarts are created for storing aggregated data at a required level with ORACLE materialized views. These materialized views allow specific design and fast access to pre-computed variables.

As concerns the PATIENT, PROVIDER and VISIT axes, the I2B2 database is loaded with 1,214,000 patients, 1,058,000 stays and 975 hospital units stays (as of September 2009); for the CONCEPT and OBSERVATION axes, we have integrated seven main different data sources into the CDW (Table 2).

Table 2 - Data sources currently integrated into the HEGP CDW (September 2009)

Data sources	Number of concepts	Number of observations
Laboratory results	7,291	62,819,471
Drug prescriptions	31,363	1,002,940
Clinical observations from structured forms ICD-10 codes	5,224 21,356	20,842,494 1,874,639
Medical act codes	10,050	1,775,283
Consultation and hospitalization text reports from the cardiovascular department	39	165,873
DRG codes	3,771	467,281
TOTAL	79,094	88,947,981

The restitution realm

The restitution real analysis involves user interaction and functional integration with other CIS components. The main objective of the CDW is to facilitate access to clinical data for users and these users may have different profiles [12]. We therefore used dedicated tools for each user profile:

- Power users: SchemaSpy (to navigate in database models) and DocuWiki (to maintain technical documentation)
- Standard and occasional users: SAP/Business Objects (to design reports) and KaliTech®/Kalidoc (to integrate reports into the enterprise content management system)
- Searchers: I2B2 dedicated client

The restitution of CDW objects was evaluated through various pilot studies including:

1. EHCR quality management (clinical report exhaustivity for the year 2008)
2. Evaluation of medical prescription practices by studying simultaneous biological prescriptions of ESR and CRP in the hospital (since the year 2003)
3. Evaluation of a rule-based engine dedicated to pharmaceutical validation of drug prescriptions [16].

The administration realm

For the Administration realm, we decided to enlarge the functional perimeter of the Help Desk used in the hospital (Peregrine software®/GPS) to cover CDW functionalities such as "Request for new dashboards". For the two pilot domains, we used the Peregrine®/GPS software to register dashboard creation, and thereby monitor the evolution of these two CDW objects in the future.

Discussion and conclusion

CDW can be used in the health sector with several objectives, including: 1) quality management (for auditing or for outcome studies); 2) identification of best practices; 3) population follow-up (for predictive medicine or disease registries); 4) clinical investigations or case studies; and 5) intervention studies (as in before/after studies or controlled trials). The success of the CDW depends on the preexistence of the CIS producing and storing raw clinical data. The two components evolve in synergy: the CIS is the main source of data for the CDW and the CDW produces data (indicators, reports) that enhance the overall healthcare activity, which is in turn processed by the CIS.

In this paper, we have presented the methodology used for integrating a CDW based on the I2B2 framework into the information system of the Georges Pompidou University Hospital. We were able to initiate a global methodology for integrating heterogeneous databases into an open, community-based framework that allows integration with other HIS components for data restitution.

The open source feature of this framework helps with the adaptation of the core database model to implement our security strategy based on a strong reversible encryption algorithm applied to all patient-identifying data.

The simplicity of the I2B2 star schema model facilitates several tasks:

- Creating generic procedures (either with ORACLE® views or with Talend®/Open Studio jobs) for data integration into the CDW. In the data mapping step (figure 1), View 2 and View 4 are generic, and this accelerates new data source integration.
- Creating dedicated datamarts with ORACLE® materialized views to allow both the design of oriented analysis of new aggregated variables and to accelerate their computation and their integration into users' reports.
- Designing a dedicated SAP®/Business Objects universe layer for building reports and dashboards with raw or aggregated, CDW data.

We have also initiated the use of the Peregrine®/GPS help-desk system but only in the scope of pilot domains. A specific task of user training is required to encourage the use of this help-desk software in the context of the hospital CDW.

We will continue to deploy a procedure to update the CDW data continuously from the currently integrated data sources, to add new data sources such the hospital bio-bank, and to evaluate the benefits and drawbacks of the CDW in a production environment. An important issue is the usefulness of the CDW for routine selection of candidate patients for clinical research studies. Another is the value of the system for feeding CDISC electronic case report forms (e-CRF), partly or completely, with data stored in the CDW to evaluate research fostering strategies such as those described in [17].

References

[1] Inmon WH. Building the data warehouse (2nd ed.). New York, NY, USA : John Wiley & Sons, Inc. 1996.

[2] Kimball R and Ross M. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. New York, NY, USA : John Wiley & Sons, Inc. 2002.

[3] Bichutskiy VY, Colman R, Brachmann RK, and Lathrop RH. Heterogeneous Biomedical Database Integration Using a Hybrid Strategy: A p53 Cancer Research Database. Cancer Inform 2007; 2: 277-87.

[4] Amin W, Parwani AV, Schmandt L, Mohanty SK, Farhat G, Pople AK, Winters SB, Whelan NB, Schneider AM, Milnes JT, Valdivieso FA, Feldman M, Pass HI, Dhir R, Melamed J, and Becich MJ. National Mesothelioma Virtual Bank: a standard based biospecimen and clinical data resource to enhance translational research. BMC Cancer 2008; 8: 236.

[5] Wehling M. Translational medicine: science or wishful thinking? J Transl Med 2008; 6: 31.

[6] Shah SP, Huang Y, Xu T, Yuen MM, Ling J, and Ouellette BF. Atlas - a data warehouse for integrative bioinformatics. BMC Bioinformatics 2005; 6: 34.

[7] Hu H, Brzeski H, Hutchins J, Ramaraj M, Qu L, Xiong R, Kalathil S, Kato R, Tenkillaya S, Carney J, Redd R, Arkalgudvenkata S, Shahzad K, Scott R, Cheng H, Meadow S, McMichael J, Sheu S, Rosendale D, Kvecher L, Ahern S, Yang S, Zhang Y, Jordan R, Somiari SB, Hooke J, Shriver CD, Somiari RI, and Liebman MN. Biomedical informatics: development of a comprehensive data warehouse for clinical and genomic breast cancer research. Pharmacogenomics 2004; 5: 933-41.

[8] Viangteeravat T, Brooks IM, Smith EJ, Furlotte N, Vuthipadadon S, Reynolds R, and McDonald CS. Slim-prim: a biomedical informatics database to promote translational research. Perspect Health Inf Manag 2009; 6: 6.

[9] Prokosch HU and Ganslandt T. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. Methods Inf Med 2009; 48: 38-44.

[10] Degoulet P, Marin L, Lavril M, Le Bozec C, Delbecke E, Meaux JJ, and Rose L. The HEGP component-based clinical information system. Int J Med Inform 2003; 69: 115-26.

[11] http://www.cnil.fr/la-cnil/ (October 2009)

[12] Puhr C. The clinical data warehouse. PHD thesis (MIAS student). Univ. of Wien. 2002.

[13] Murphy SN, Mendis M, Hackett K, Kuttan R, Pan W, Phillips LC, Gainer V, Berkowicz D, Glaser JP, Kohane I, and Chueh HC. Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside. AMIA Annu Symp Proc 2007; 548-52.

[14] http://www.bouncycastle.org/ (October 2009)

[15] Parmanto B, Scotch M, and Ahmad S. A framework for designing a healthcare outcome data warehouse. Perspect Health Inf Manag 2005; 2: 3-19.

[16] Boussadi A, Bousquet C, Sabatier B, Colombet I, and Degoulet P. Specification of business rules for the development of hospital alarm system: application to the pharmaceutical validation. Stud Health Technol Inform 2008; 136: 145-50.

[17] Ohmann C and Kuchinke W. Future developments of medical informatics from the viewpoint of networked clinical research. Interoperability and integration. Methods Inf Med 2009; 48: 45-54.

Address for correspondence

Eric Zapletal : eric.zapletal@egp.aphp.fr