Architectural Comparison of Three Healthcare Integrated Data Repositories: Quest for Data Representation Best Practices

Vojtech Huser, MD, PhD

Laboratory for Informatics Development, National Institutes of Health Clinical Center

NIH National Institutes of Health *Clinical Center*

Abstract

The importance of Integrated Data Repositories (IDRs) in research is rapidly increasing. We compare the architecture of three IDRs. Based on this analysis, we formulate a list of research data warehouse desiderata which attempt to formulate a set of desired characteristics for an optimal IDR.

Introduction

Current clinical and translational research increasingly relies on existence of robust IDRs with administrative, clinical, and -omics data. Following clear warehouse design principles can lower long-term maintenance costs for organizations which are currently building or significantly re-structuring their data warehouses. Maintenance of those warehouses is very costly and architectural changes are complicated by existing dependencies. Getting the right architecture early during the warehouse creation is crucial.

Methods

We set out to analyze common themes and principles of IDR architecture and IDR maintenance on a comparison of three IDRs: (1) Informatics for Integrating Biology & the Bedside (i2b2), (2) Virtual Data Warehouse (VDW) created by HMO Research Network and (3) Observational Medical Outcomes Partnership (OMOP). We analyze the architectures in two aspects: architecture for storing facts as well as structures for representing the terminology layer of the warehouse. For each warehouse, we look at how several sample clinical events would be stored by each IDR (laboratory result, procedure, clinical document). We also consider how a given warehouse would store a novel data domain (e.g., genomic sequence or data from clinical trial case report forms). Based on this analysis and comparison of the three IDRs, we formulate a list of warehouse desiderata which deal with optimal representation format, metadata representation and management, data lineage, and terminology and maintenance issues. We claim that formulating a set of requirements for a data warehouse may prove similarly beneficial as was formulation of desiderata for controlled terminologies [1]. We build on several prior efforts to formulate healthcare specific set requirements: Huff formalized an event based model [2]; Murphy describes several optimizations for relational database [3]; Nadkarni offers an extensive account on database design [4] and Gilchrist looked at query speed optimizations [5].

Property	i2b2	OMOP	VDW
Generic fact data structure	OBSERVATION_FACT	OBSERVATION	none
Designated data structures	PATIENT_DIMENSION, VISIT_DIMENSION, PROVIDER_DIMENSION	PERSON, VISIT_OCCURENCE, DEATH, COHORT, PROVIDER, CARE_SITE DRUG_ERA, DRUG_EXPOSURE, CONDITION_ERA, CONDITION_OCCURENCE, PROCEDURE_OCCURENCE	DEMOGRAPHICS, ENCOUNTERS, CENSUS, ENROLLMENT, DEATH, PROVIDER LAB_RESULTS, DIAGNOSES, PROCEDURES, PHARMACY, TUMOR, VITAL_SIGNS
Terminology layer	CONCEPT_DIMENSION	CONCEPT, CONCEPT_RELATIONSHIP, CONCEPT_ANCESTOR, SOURCE_TO CONCEPT_MAP	No generic terminology table EVER_NDC table (for drug codes only)
Fact nesting	Generic *modifier_cd* column (coded in native terminology) in the OBSERVATION_FACT table	Generic *obs_value_as_concept_id* column (coded in native terminology) in the OBSERVATION table. Domain-specific columns in designated tables. Additional fact grouping (temporal, functional) via PAYER_PLAN_PERIOD table and several _ERA tables.	No generic fact nesting structure. Numerous domain-specific columns in designated tables (e.g., encounter type in PROCEDURES). Additional fact grouping (temporal) via ENROLLMENT table.
Designated columns in fact table	*valtype_cd, units_cd, encounter_num, provider_id, location, confidence_num, valueflag_cd, observation_blob*	*observation_type_concept_id,associated_provider_id, obs_range_low, obs_range_high, source_obs_code, unit_concept_id*	n/a

Table 1: Comparison table summarizing selected properties for each analyzed warehouse.

- single patient identifier
- consistent naming strategy
- information storage model
- terminology model
- value-sets management within the terminology layer
- data request audit log and table/column usage analysis
- capture data warehouse historical evolution
- shadow ID management
- metadata documentation platform with collaborative functions
- maintenance of value-sets for identifying data (PHI)
- support multiple views of data
- multiple query platforms (self-service/human mediated)

Figure : Partial list of desiderata

Results

We classified the schemes into three basic models for organizing the warehouse: (1) *EAV model* which stores several attributes in a more generic table (e.g., both lab result and procedure event would be a fact instance and stored in one structure). This principle can also be applied at single or at multiple layers. For example, each item in an EAV-based event table (e.g., biopsy event) may have an infinite number of event attributes (e.g., who ordered the biopsy) stored in an associated attribute-EAV-based table (2) *A hybrid model* where some elements are stored in an EAV mode but certain common event attributes have a designated column (e.g., fact_source_system, observation_type, or observation_value_text) [6]. Populating all such hard-coded columns is not required and they may be empty for some facts. Frequently, event_time is one of such attribute and EAV is sometimes extended to entity-attribute-value-time (EAVT) model. And finally ,(3) *traditional* column-based model where each data domain (e.g., encounters or procedures or demographics) is stored in a more specialized table with columns representing necessary fact attributes (e.g., tumor table with tumor stage and tumor type columns).

Table 1 provides an overview of selected comparison aspects. The following relative advantages were found during the comparison: i2b2: full use of entity-attribute-value paradigm; VDW: incorporation of data quality checking scripts; OMOP: excellent documentation, elaborate terminology model with template terminology queries. Figure 1 provides a list of identified desiderata. This list of desiderata is not intended to be complete, and it should serve to facilitate discussion. Additional architecture documentation and analysis is available at http://code.google.com/p/desiderata.

References

[1] Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med. 1998;37:394-403.
[2] Huff SM, Rocha RA, Bray BE, Warner HR, Haug PJ. An event model of medical information representation. J Am Med Inform Assoc. 1995;2:116-34.
[3] Murphy SN, Morgan MM, Barnett GO, Chueh HC. Optimizing healthcare research data warehouse design through past COSTAR query analysis. Proc AMIA Symp. 1999:892-6.
[4] Nadkarni PM. Metadata-driven software systems in biomedicine. New York: Springer; 2011.
[5] Gilchrist J, Frize M, Ennett CM, Bariciak E. Performance Evaluation of Various Storage Formats for Clinical Data Repositories. Instrumentation and Measurement, IEEE Transactions on. 2011;60:3244-52.
[6] Marenco L, Tosches N, Crasto C, Shepherd G, Miller PL, Nadkarni PM. Achieving evolvable Web-database bioscience applications using the EAV/CR framework: recent advances. J Am Med Inform Assoc. 2003;10:444-53.

Acknowledgements: This work is supported by intramural research funds from the National Library of Medicine and the NIH Clinical Center. **Contact:** vojtech.huser@nih.gov