An improved data workflow for a medical data warehouse

Georg Fette[a] and Maximilian Ertl[a] and Georg Dietrich[b] and Jonathan Krebs[b] and Martin Toepfer[b] and Stefan Störk[a] and Mathias Kaspar[a] and Frank Puppe[b]
[a] *Comprehensive Heart Failure Center, 97078 Würzburg*
[b] *University of Würzburg, Chair VI, 97074 Würzburg*

Abstract. In recent years clinical data warehouses (CDW) have become more and more popular to support scientific work in the medical domain. Despite the tool support for many subtasks it is still a laborious task to establish a CDW in an existing clinical data environment. We present a workflow which can be taken as a blueprint for newly established CDW projects and the implementation of this blueprint at the University Clinic Würzburg.

Keywords. Clinical Data Warehouse, Data Workflow

1. Introduction

In recent years clinical data warehouses (CDW) have become more and more popular to support scientific work in medicine. A CDW can serve for multiple purposes.

The first use case is used to screen patients for study inclusion by modeling the study's inclusion and exclusion criteria as a CDW query respecting data privacy guidelines. Desired patients can either be found retrospectively by searching old clinical data and, if necessary, contacting the patients by mail or by identifying the desired patients online while still hospitalized.

Another way of assisting clinical studies is by supporting data completion. Data from the CDW can be exported and imported into the study database instead of having the data collected anew.

A third application of a CDW is data mining on the stored data. Already existing knowledge can be confirmed by testing it against the clinical routine data or new correlations can be identified. These correlations can either be discovered by automatic hypothesis generation and evaluation or by manual evaluation of hypotheses created by clinician scientists.

We created a data workflow for the creation of a CDW and created an instance of such a CDW at the Universitätsklinikum Würzburg described in the following.

2. System architecture

2.1. Generic data model

Hospitals store their clinical routine data in centralized clinical information systems (CIS). Despite this single point of access these systems consist of many heterogeneous subsystems containing data in various forms (tabular data, parameterized medical forms, office documents, etc.). Nevertheless most of this data can be represented in an abstract information data model. This model contains patients which can appear at the

clinic for various visits ("cases"). Cases can contain documents with heterogeneous information facts. A fact contains a patient-ID, an attribute ID (referencing to an entry in an attribute catalogue), the actual value (e.g. number, text, blob) and the time when the fact was measured. Figure 1 illustrates this scheme.

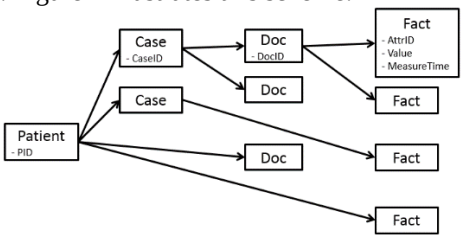

Figure 1. Information model

The paradigm of facts belonging to entities (patients) and referencing attribute catalogue entries is commonly known as entity-attribute-value system [1] (whereas it is here an entity-attribute-measuretime-value system).

2.2. Workflow elements

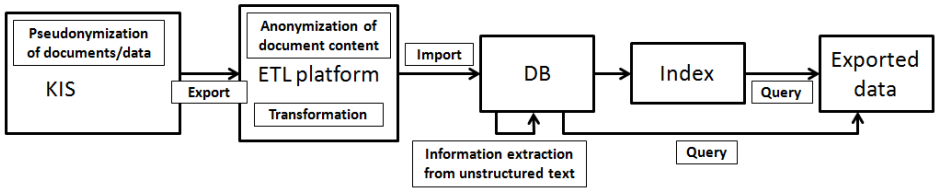

Figure 2. Data process workflow

Assuming the given information scheme in figure 1, we established a data workflow for our CDW. The ETL (extraction transformation load) part of the workflow can be executed on a complete data dump of the CIS or be executed in a daily (or hourly) fashion for subsequent updates. The workflow is depicted in figure 2 and is described as follows:

All data is pseudonymized during export from the CIS by replacing all patient-, case- and document identifier by pseudo numbers. The data is exported into the file system of an ETL (extraction transformation load) staging area. The export uses three file formats: 1) CSV files for tabular data like e.g. laboratory values or diagnosis codes 2) XML files for parametrized medical documents (PMD), e.g. cardiography reports or anamnesis reports 3) doc or pdf files containing discharge letters.

After exporting documents, their content is anonymized by removing all occurrences of person names and addresses. In a first step, all references to the known patient-ID information are removed by direct search. A second anonymization algorithm based on linguistic patterns is run on all text fragments removing occurrences of further names, e.g. names in the form of patterns like “Dr. X” or “living in Y”, to ensure the deletion of misspelled names or addresses.

Before the import into the database the data is transformed to fit the import specifications of the CDW. The transformations performed by our system are done with standard ETL software processes. Discharge letters that come in the form of doc

files are transformed into HTML and further split into distinct sections (e.g. diagnosis, physical examination, etc.).

The import into the database can take place by either writing data directly into the database, by using a Java API or by using an import tool. The import tool requires the input files to be in given a defined CSV- or XML-schema and a configuration file which describes how the data should be imported. The import incorporates not only the import of the fact data but also the import of the attribute catalogue which contains the meta data about the actual facts. Whenever no exportable meta catalogue for a certain data domain in the CIS exists, the catalogue entries belonging to this data are created on the fly during the fact import and the entries' data types are inferred from the imported data after a full import.

Figure 3 depicts the database table schema of our CDW model: Table DWInfo contains the facts, DWCatalog contains the meta data catalogue and table DWImportDeleted stores the IDs of facts which have been deleted in update import operations, so that subsequent index structures can easily update their state according to the database.

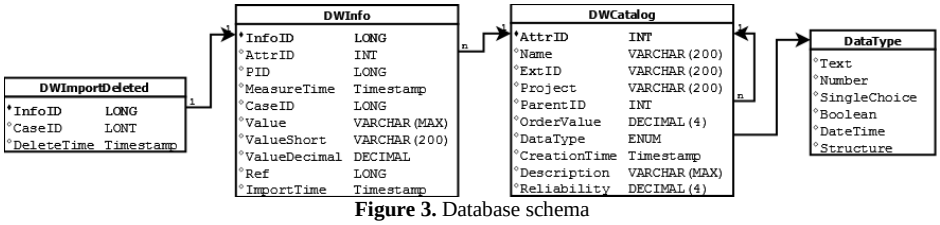

Figure 3. Database schema

As a further improvement of search capabilities on the given data it is possible to perform information extraction [2] on unstructured textual facts in the database. The structured information gained in this extraction process is added to the CDW and linked with its source documents.

When the database content has been completely imported, index structures can be added to greatly improve search performance. We developed a Solr/Lucene-Index with which most necessary query operations can be sped up by about two orders of magnitude [3]. The Solr data model which was implemented is depicted in figure 4.

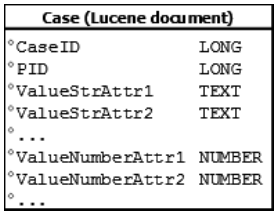

Figure 4. Solr/Lucene data model

Some query operations which cannot be executed with this Solr data model still have to be performed with a query engine directly operating on the database. A query formulated as an XML String in a custom-built medical query language can be interpreted either by the SQL or the Solr query engine and creates a result table. The table contains, depending on the query definition, statistics about the queried data or the concrete data itself and can be exported as CSV or Excel file.

To facilitate querying we developed a query GUI for the CDW [4, 5] with which queries can be easily assembled, managed and executed.

3. The data flow in practice

3.1. Data

Currently the CDW instance of the Universitätsklinikum Würzburg stores data of about 1.2 million patients, having about 4 million cases, containing about 20 million documents and about 250 million facts. The database has a size of 160GB and runs on a MSSQL server 2012. The Solr/Lucene index has a size of 75GB.

3.2. Usage

The access to query statistical data in the CDW is available to all medical staff without special application (simply via clinical standard user login). The access to pseudonymized patient data is possible with a specific access application with the consent of the data privacy official. The access to depseudonymized patient data is possible with an additional access application. The depseudonymization is manually triggered by the project ID gate keeper with the consent of the institutional data protection officer and the ethics committee.

The Würzburg CDW has supported several studies in the recent years including all use cases discussed in 1. Cohort acquisition support for clinical studies was given e.g. in the Euroaspire IV study [6], completion of data records with routine data was performed, e.g. in the studies STAAB [7] and AHF [8], data mining and hypothesis confirmation was supported in [9].

4. Discussion

Although the concept of CDWs is not new [10], we developed a full workflow, including existing and, where necessary, own software modules. A popular freely available CDW system is I2B2 [11]. Although the I2B2 system would fit into the presented workflow as replacement of database and query engine, the query modeling possibilities and query speed of I2B2 does not fit the requirements of several of our use case studies. We therefore created the new lightweight database approach and the corresponding index structures discussed above. Further index structures based on ElasticSearch[1] and Neo4J[2], which may be queried by the same query language are in development and could further improve query possibilities and performance.

The process workflow framework and its corresponding software modules are planned to be also used at other hospitals. Thus, they are developed to be as generic as possible.

[1] https://www.elastic.co/products/elasticsearch
[2] http://neo4j.com/

This work was funded by the Bundesministeriums für Bildung und Forschung (BMBF01 EO1004).

References

[1] Dinu V, Nadkarni P, *Guidelines for the effective use of entity-attribute-value modeling for biomedical databases*, Int Journal of Medical Informatics 76 (2007) 769-779

[2] Toepfer M, Corovic H, Fette G, Klueg P, Störk S, Puppe P, *Fine-grained information extraction from German transthoracic echocardiography reports*, BMC Medical Informatics and DecisionMaking, (2015)

[3] Dietrich G, Fette G, Puppe F, *A Comparison of Search Engine Technologies for a Clinical DataWarehouse*, Proceedings of the LWA Vol. 1226 (2014), 235-242

[4] Fette G, Dietrich G, Ertl M, Toepfer M, Kaspar M, Störk S, Puppe F, *Die grafische Benutzeroberfläche PaDaWaN für das klinische Data Warehouse für Routinedaten an der Universitätsklinik Würzburg*, Proceedings of the GMDS (abstract) (2015)

[5] Dietrich G, Fell F, Fette G, Krebs J, Ertl M, Kaspar M, Störk S, Puppe F, *Web-PaDaWaN: Eine Web-basierte Benutzeroberfläche für ein klinisches Data Warehouse*, submitted at the GMDS (abstract) (2016) https://go.uniwue.de/ls6-pub-2016-web-padawan

[6] De Smedt D, Clays E, Höfer S, et al, *Validity and reliability of the HeartQoL questionnaire in a large sample of stable coronary patients: The EUROASPIRE IV Study*, Eur J Prev Cardiol, published online before print September 10 (2015)

[7] Tiffe T, Wagner M, Morbach C, Reuter M, Kircher J, Gelbrich G, Störk S, Heuschmann PU, *On behalf of the STAAB–Consortium: "Characteristics and Course of Heart Failure STAges A-B and Determinants of Progression – The STAAB cohort study: study design and preliminary results"*, 10th annual meeting of the German Society of Epidemiology, Potsdam, (2015) (abstract)

[8] Kaspar M, Fette G, Ertl M, Dietrich G, Nagler N, Störk S, Angermann C, Puppe F, *Extraktion und Transfer patientenbezogener Daten aus klinischen Informationssystemen in Studiendatenbanken – effektive Unterstützung klinisch-epidemiologischer Forschung durch ein Data Warehouse*, Proceedings of the GMDS (2015) (abstract)

[9] Wallenborn J, Störk S, Herrmann S, Kukuy O, Fette G, Puppe F, Gorski A, Hu K, Voelker W, Ertl G, Weidemann F, *Prevalence of severe mitral regurgitation eligible for edge-to-edge mitral valve repair (MitraClip)*, Clinical Research in Cardiology Feb. 2016 (2016), 1-11

[10] Prokosch HU, Ganslandt T, *Perspectives for medical informatics - reusing the electronic medical record for clinical research*, Methods of Information in Medicine 48/1 (2009), 38-44

[11] Murphy SN, Weber G, Mendis M, Chueh HC, Churchill S, Glaser JP, Kohane IS, *Serving the Enterprise and beyond with Informatics for Integrating Biology and the Bedside (i2b2)*, J Am Med Inform Assoc. 17(2) (2010) 124-130.