Health Informatics Meets eHealth
G. Schreier and D. Hayn (Eds.)
© 2018 The authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/978-1-61499-858-7-204

Modeling of ETL-Processes and Processed Information in Clinical Data Warehousing

Erik TUTE[a,1], Jochen STEINER[a]
[a] *Peter L. Reichertz Institute for Medical Informatics of the University of Braunschweig - Institute of Technology and Hannover Medical School, Germany*

Abstract. Background: Literature describes a big potential for reuse of clinical patient data. A clinical data warehouse (CDWH) is a means for that. Objectives: To support management and maintenance of processes *extracting*, *transforming* and *loading* (ETL) data into CDWHs as well as to ease reuse of metadata between regular IT-management, CDWH and secondary data users by providing a modeling approach. Methods: Expert survey and literature review to find requirements and existing modeling techniques. An ETL-modeling-technique was developed extending existing modeling techniques. Evaluation by exemplarily modeling existing ETL-process and a second expert survey. Results: Nine experts participated in the first survey. Literature review yielded 15 included publications. Six existing modeling techniques were identified. A modeling technique extending 3LGM² and combining it with openEHR information models was developed and evaluated. Seven experts participated in the evaluation. Conclusion: The developed approach can help in management and maintenance of ETL-processes and could serve as interface between regular IT-management, CDWH and secondary data users.

Keywords. Data Warehousing, Organizational Models, Health Information Interoperability

1. Introduction

Resulting from the broad adaption of information technology (IT) in hospitals, there is much clinical patient data digitally available. The use of such data for other purposes than the originally indented is commonly referred to as its secondary use or reuse [1]. The Medical Informatics literature describes a big potential for reuse of this data, e.g. [1]. There are still technical and organizational challenges related to secondary use of clinical data [1]. A clinical data warehouse (CDWH) is a facility addressing some of these challenges by integrating data from heterogeneous sources and making it available for reuse, e.g. for analysis. The process of *Extracting* data from a source system (e.g. a patient data management system), *Transforming* it into a target schema (for example a schema facilitating later analyses) and finally *Loading* it into the persistence layer of the CDWH (layer in which data is stored permanently for later reuse) is called ETL-process. These ETL-processes are typically complex and custom-made which makes their management and maintenance challenging. Established tools supporting the graphical user interface (GUI) based development and operation of ETL-processes exist (e.g. Microsoft SQL Server Integration Services, Talend Open Studio). However, to the best of the author's knowledge there is a lack of established standards for the modeling of

[1] Corresponding Author: Erik Tute, Peter L. Reichertz Institute for Medical Informatics of the University of Braunschweig - Institute of Technology and Hannover Medical School, Mühlenpfordtstr. 23, 38106 Braunschweig, E-Mail: erik.tute@plri.de

ETL-processes and the thereby processed information, which are suitable for supporting ETL-process documentation and IT-management.

Modeling techniques and tools aiming to support IT-management in hospitals exist for more than 20 years. Their aim is to provide information or to give the possibility for analysis of business processes, information objects, supporting application systems and underlying hardware [2]. An approach to interoperability between different healthcare information systems through shared implementable definitions of the clinical concepts represented by clinical data are clinical information models (CIM), also referred to as detailed clinical models [3]. CIM-based CDWHs have potential merits ([4], [5]). This work is based on the assumption that there is an overlap between the information needs of regular hospital IT-management, ETL-development/maintenance and users in data reuse scenarios justifying the effort for standardized and machine processable documentation of ETL-related metadata. Figure 1 shows this information overlap without trying to be exhaustive. Metadata needs are derived from Kahn's proposed recommendations on metadata reporting for distributed data networks [6].

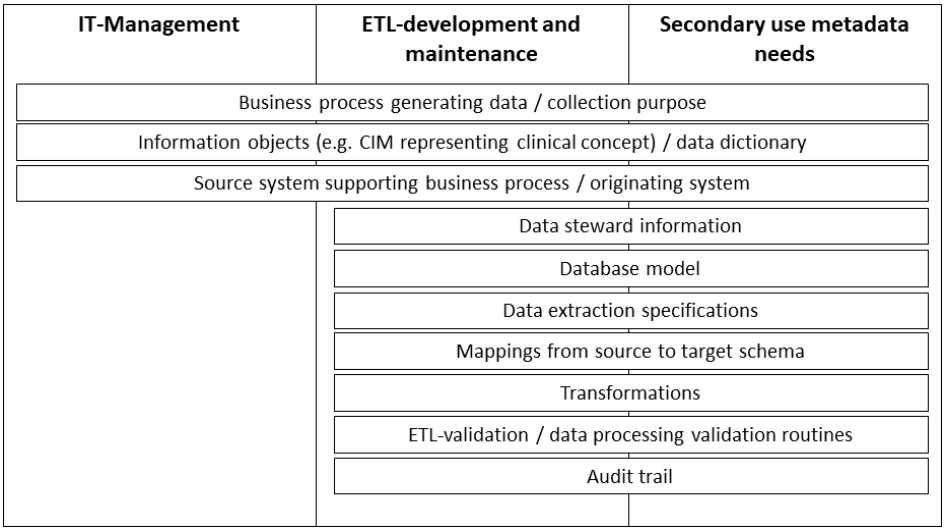

Figure 1. Information overlap between IT-management, ETL-development and secondary use.

Objective of this work is to find an approach for modeling of ETL-processes and thereby processed information, which can be integrated with tools supporting regular IT-management, can help in management and maintenance of ETL-processes and eases the machine processable provision of IT-management and ETL-related metadata for secondary use.

2. Methods

First, an expert survey and a literature review were conducted to find requirements for the modeling approach and to assess existing modeling techniques. Subsequently, an approach for modeling of ETL-processes for clinical data warehousing was developed as extension of an existing modeling technique. Finally, this approach was evaluated through a proof of concept (POC) modeling example and another expert survey.

2.1. Literature review on existing modeling techniques

In a literature review Pubmed, IEEE Explore and Scopus where searched for publications on modeling of ETL-processes. The search terms where developed iteratively for optimizing precision and recall. One reviewer included or excluded results based on previously defined criteria assessing first title, in doubt abstract and if necessary full text. References of included papers were also assessed. The final search terms as well as inclusion and exclusion criteria can be found in [7].

2.2. Expert survey on existing modeling techniques and their shortcomings

A survey employing face-to-face interviews or an online questionnaire depending on the experts' availability was conducted to complement existing ETL-process modeling techniques as well as to find shortcomings of these existing methods and currently practiced ETL-process documentation (for interview structure and online questionnaire see [7]). Experts were selected from the clinical data warehousing staff at Hannover Medical School (MHH) and from members of the HiGHmed consortium. HiGHmed is funded by the Federal Ministry of Education and Research (BMBF) under the Medical Informatics funding scheme, pursuing among others the objective to develop information infrastructures to increase efficiency of clinical research. These infrastructures are facing the problem of ETL-process documentation and metadata provision in a multisite environment [8]. Experts were recruited from all positions (e.g. ETL-programmers, data analysts, management) to gain different views on ETL-modeling and related information needs. Part of the interviews was a self-assessment of the presumed experts on their knowledge about ETL-processes.

2.3. Determination of ETL-modeling technique

Based on literature review and interview results, requirements for modeling of ETL-processes in clinical data warehousing were derived and prioritized as must-, should- or can-criteria. In a further step, evaluation criteria for the found existing modeling techniques where derived under consideration of their relevance for the objectives of the underlying master's thesis [7]. For each found modeling technique and each evaluation criterion an assessment regarding satisfaction of the criterion including documentation of the corresponding reasoning was done.

A definition of the modeling process based on the most promising modeling technique was developed. The intention in this step was to keep the changes or extensions in the modeling technique as little as possible in order to stay compatible with existing tools and not to distort core concepts of the modeling technique.

2.4. Evaluation of modeling technique

As a POC an ETL-process from the MHH clinical data warehouse was modeled using the modeling approach determined in the previous steps. Criteria for the selection of the ETL-process for POC where pragmatic: access to the actual ETL-process and respective documentation, good reachability of the ETL-developer and already available CIMs.

Another expert survey on the resulting ETL-model from the POC, the determined ETL-modeling approach in general and potentials for improvements and further development of the modeling approach complemented the evaluation. The survey

consisted of a short presentation of the modeling approach and the POC results for the experts who participated in the first survey in face-to-face interviews followed immediately by a paper-based questionnaire (see [7]). The presentation was given in groups, the questionnaire was completed by each expert on their own. Criteria for evaluation were shortcomings identified in the first expert survey. Improvements on these shortcomings were assessed based on experts' responses in the second survey.

3. Results

In the following, we briefly cover the results of the literature review and expert survey. We present the developed modeling technique extending 3LGM² in combination with openEHR information models and conclude with a description of the evaluation results. More details can be found in [7].

3.1. Literature Review and expert survey

The search in Scopus listed 299 publications, IEEE listed 420 and PubMed 13. From these 433 items, 76 were removed as duplications. After application of exclusion criteria on title and abstract, 55 publications remained for screening of the full text. Based on the references of not excluded full texts another 25 full texts were screened. Finally, 15 publications were included and found ETL-modeling techniques were documented.

Seven experts participated in face-to-face interviews. One interview was excluded afterwards, because the interviewed expert's self-assessment on ETL-knowledge indicated a lack of knowledge on ETL-processes. Two experts participated via the online questionnaire. One of those questionnaires was not completed and thus excluded.

3.2. Existing ETL-modeling techniques

Literature review and expert survey identified six existing ETL-modeling techniques (origin from review or survey specified in brackets):

- UML-based modeling of ETL-processes described in [9] and [10] (review)
- UML-based modeling of ETL-processes employing activity diagrams described in [11]. This method builds on reasoning from [9] (review)
- Arktos – Graph-based modeling of ETL-processes described in [12] (review)
- KoMo – conceptual modeling of ETL-processes described in [13] (review)
- Business Process Model and Notation (BPMN 2.0) based modeling of ETL-processes described in [14] (review)
- A combination of an extension of the 3LGM² technique for modeling of hospital information systems described in [2] with standardized CIMs (survey)

The only modeling technique resulting from the expert survey employs 3LGM² because it was already in use in medical informatics research and teaching contexts at MHH. The idea to combine 3LGM² with a standard able to express implementable definitions of clinical concepts resulted from the information overlap depicted in Figure 1 and aims at increasing reuse of content definitions in different contexts.

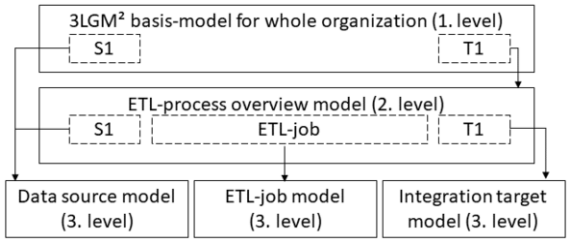

Figure 2. General structure of ETL-process models. S = Source, T = Target, Dashed-boxes: elements linking to sub-models, solid boxes: sub-models, arrows: links to sub-models. This is a modified version of a figure from [7].

3.3. ETL-modeling technique

The assessment regarding satisfaction of the criteria derived from the expert survey lead to the selection of the approach employing an extension of 3LGM². 3LGM² was combined with CIMs complying with openEHR [15], because the Peter L. Reichertz Institute for Medical Informatics at the MHH already has research activities regarding openEHR in clinical data warehousing (for example CIMs see [16]). However, other standards for CIMs could also be employed.

First development result was a mapping of required ETL-process components (e.g. data source, data input description, transformations, error-output description, data output description, integration target) into the 3LGM²-modeling technique. Second result was a reference model for ETL-process modeling with 3LGM² and openEHR defining required components and information to be stored in an ETL-process model.

Figure 2 depicts which kind of information is located in which levels of the model. The first level is a regular 3LGM² model for an organization supporting IT-management in general (for an example view on 3LGM² see [17]).

The second level sub-model gives an overview of an ETL-job by providing information on an ETL-processes' data sources and integration targets. An example view is shown in Figure 3.

On the third level, sub-models describing the data sources (including a description of the data input from the source like tables, fields, datatypes, constraints or a CIM), ETL-job (transformations, error-output) and integration target (including a description of the data output by linking a CIM) are located. Figure 4 shows an example view for an ETL-job. Sub-models can be linked in models of higher level, supporting reuse of information on the different levels and thus, improving maintainability.

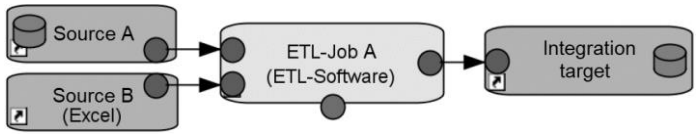

Figure 3. Example view for the ETL-process overview sub-model (level 2). Round edged boxes: application systems, circles: interfaces, arrows: indicate directed data flow, small boxes with arrows: indicate links, e.g. to sub-model or CIM, cylinder: database.

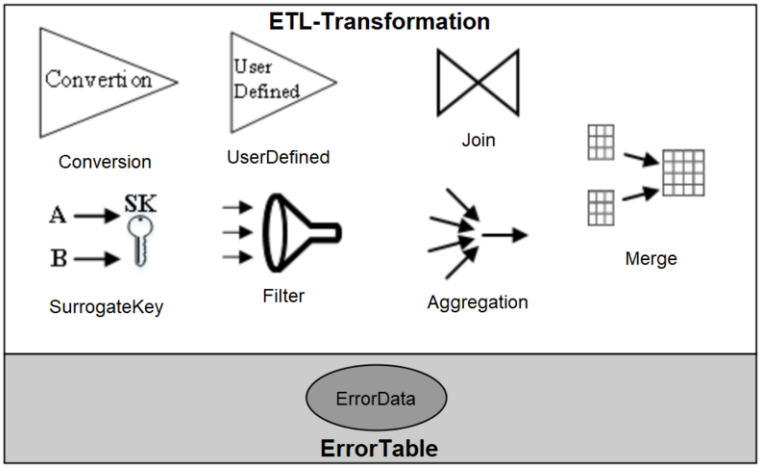

Figure 4. Example view for the ETL-job sub-model (level 3).

3.4. Evaluation results

A POC using the example of an ETL-job integrating assisted ventilation data from two different patient data management systems (PDMS) of a pediatric intensive care unit at the MHH into a common data model of the local clinical data warehouse revealed no issues.

All seven experts who participated in the face-to-face interviews participated also in the evaluation. Experts' answers showed improvements regarding the shortcomings identified in the first expert survey. These shortcomings were (a) missing or incomplete documentation of ETL-processes, (b) too much time necessary to understand ETL-processes implemented by others and (c) data quality issues. On (a): Three of seven experts answered "yes" to the question if they thought that the modeling technique could resolve the problem of no or incomplete ETL-process-documentation. Three experts answered "no" on that question and one expert did answer neither "yes" nor "no". Two of the three experts answering "no" motivated their answer stating that the approach first has to prove its applicability in practice. On (b): Four of seven experts answered "yes" to the question if the proposed modeling technique could resolve the problem of too much time necessary to understand unknown ETL-processes. Two of the three answering "no" to that question stated that the solution could not resolve the problem, but would reduce the burden. On (c): Five of seven experts answered "yes" to the question if they thought that this modeling technique could contribute to an increased data quality.

The experts' answers also indicated potential for further improvements of the modeling technique. The experts mentioned:

- More elements for describing transformations (e.g. union, pivot, sort)
- More details on transformations and better integration with actual ETL-job, e.g. by implementing automatic information extraction from ETL-job files from common ETL-tools
- Functional improvements of supporting tools, e.g. regarding model analysis or merges of models

4. Discussion

Main result of this work is an approach for modeling of ETL-processes and thereby processed information based on 3LGM² and openEHR. The modeling technique can be applied utilizing existing tools. Evaluation with experts from the intended user groups indicate positive expectations regarding its application.

The approach utilizes existing standards originally intended to support regular IT-management and interoperability of routine IT-systems. Thus, there is potential for reuse of common information items from routine IT-management and clinical data warehousing. Such reuse could reduce the efforts necessary to curate models on both sides regular IT-management and clinical data warehousing. Furthermore, there is a potential to reduce a bottleneck experienced in clinical data warehousing which results from the dependency on knowledge and support from experts not explicitly participating in data warehousing, e.g. an operator of a source system [18]. Another potential advantage of the resulting models is that these could be a suitable basis for the automated extraction of metadata in secondary use scenarios as described in the introduction.

As some experts stated in the evaluation an obvious limitation of this work is that we cannot provide sound results on real world application of the presented approach yet. As a result, the potential advantages mentioned above are uncertain and practical issues hampering the adoption of long-known modeling techniques like 3LGM² [2] until today may remain predominant.

Despite of the good intention not to distort core concepts of the modeling technique the third detail level of the ETL-process modeling deviates from the original 3LGM² concept because it merges technical and concept layer. This is not inherent in the approach and could also be properly abstracted, but the proposed reference model merged it simply due to practical reasons.

As mentioned in the introductory section established tools supporting the GUI-based development and operation of ETL-processes exist. These were not found in the literature review although one could argue that such tools implement ETL-modeling techniques. However, ETL-process modeling techniques with a technical detail oriented perspective where not the focus of this work. The objective was to find a technique, which can help in management and maintenance of ETL-processes, as well as to serve as a kind of interface between regular IT-management, clinical data warehousing and secondary data users. Our findings indicate that the presented approach for modeling of ETL-processes and thereby processed information based on 3LGM² and openEHR can support that.

5. References

[1] Martin-Sanchez FJ, Aguiar-Pulido V, Lopez-Campos GH, Peek N, Sacchi L, Secondary Use and Analysis of Big Data Collected for Patient Care, Yearb Med Inform. 2017 Aug;26(1):28-37, doi: 10.15265/IY-2017-008. Epub 2017 Sep 11. PubMed PMID: 28480474.

[2] Winter A, Brigl B, Wendt T, Modeling hospital information systems. Part 1: The revised three-layer graph-based meta model 3LGM2, Methods Inf Med. 2003;42(5):544-51, PubMed PMID: 14654889.

[3] Goossen W, Goossen-Baremans A, van der Zel M, Detailed clinical models: a review, Healthcare informatics research. 2010 Dec;16:201–214.

[4] Haarbrandt B, Gerbel S, Marschollek M, Einbindung von openEHR Archetypen in den ETL-Prozess eines klinischen Data Warehouse, GMDS 2014. 59. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (GMDS). Göttingen, 07.-10.09.2014. Düsseldorf: German Medical Science GMS Publishing House; 2014. DocAbstr. 230; 2014.

[5] Marco-Ruiz L, Moner D, Maldonado JA, Kolstrup N, Bellika JG, Archetype-based data warehouse environment to enable the reuse of electronic health record data. International journal of medical informatics, 2015 Sep;84:702–714.
[6] Kahn MG, Brown JS, Chun AT, Davidson BN, Meeker D, Ryan PB, et al., Transparent reporting of data quality in distributed data networks, EGEMS (Wash DC). 2015 Mar 23;3(1):1052. doi:10.13063/2327-9214.1052. eCollection 2015, PubMed PMID:25992385; PubMed Central PMCID: PMC4434997.
[7] Steiner J, Evaluation von Methoden zur Modellierung von ETL-Prozessen und der dabei verarbeiteten Daten am Beispiel eines großen Universitätsklinikums [master's thesis], TU Braunschweig, Braunschweig, 2017.
[8] HiGHmed | HiGHmed, http://www.highmed.org, last access: 29.01.2018.
[9] Trujillo J, Luján-Mora S, A UML based approach for modeling ETL processes in data warehouses, In: Song I-Y, Liddle SW, Ling T-W, Scheuermann P, editors. Conceptual Modeling, - ER 2003: Proceedings of the 22nd International Conference on Conceptual Modeling; 2003 Oct 13-16; Chicago, IL, USA. Berlin, Heidelberg:Springer; 2003. p. 307–20.
[10] Luján-Mora S, Trujillo J, A Comprehensive Method for Data Warehouse Design, In: Lenz H-J, Vassiliadis P, Jeusfeld MA, Staudt M, editors. Design and Management of Data Warehouses. DMDW 2003: Workshop Proceedings of the 5th International Workshop on Design and Management of Data Warehouses; 2003 Sep 8; Berlin, Germany. Aachen: CEUR-WS.org [contents]. p. 1.1-1.14.
[11] Muñoz L, Mazón J-N, Pardillo J, Trujillo J, Modelling ETL Processes of Data Warehouses with UML Activity Diagrams, In: Meersman R, Tari Z, Herrero P, editors. On the Move to Meaningful Internet Systems: OTM 2008 Workshops. Berlin: Springer; 2008. 44–53 p.
[12] Simitsis A, Modeling and managing ETL processes, In: VLDB PhD Workshop. Berlin; 2003.
[13] Dupor S, Jovanović V, An approach to conceptual modelling of ETL processes, In: Biljanovic , editor. Information and Communication Technology, Electronics and Microelectronics. MIPRO 2014: Proceedings of the 37th International Convention on Inf Commun Technol Electron Microelectron; 2014 May 26-30; Opatija, Croatia. New Jersey: IEEE; 2014. p. 1485–90.
[14] El Akkaoui Z, Zimanyi E, Defining ETL workflows using BPMN and BPEL, In: DOLAP '09 Proceedings of the ACM twelfth international workshop on Data warehousing and OLAP. Hong Kong, China; 2009. 41–8 p.
[15] openEHR Foundation, openEHR – Clinical Models Program, http://www.openehr.org/programs/clinicalmodels/, last access: 13.12.2017.
[16] openEHR Foundation, Clinical Knowledge Manager, http://www.openehr.org/ckm/, last access: 07.03.2018.
[17] 3LGM²-Team, [3LGM²] – Startseite, http://www.3lgm2.de/, last access: 13.12.2017.
[18] Turley CB, Obeid J, Larsen R, Fryar KM, Lenert L, Bjorn A, Lyons G, Moskowitz J, Sanderson I, Leveraging a Statewide Clinical Data Warehouse to Expand Boundaries of the Learning Health System. EGEMS (Wash DC). 2016 Dec 6;4(1):1245. doi: 10.13063/2327-9214.1245. eCollection 2016. PubMed PMID: 28154834; PubMed Central PMCID: PMC5226381.