Data Warehouse Design Methods Review: Trends, Challenges and Future Directions for the Healthcare Domain

Christina Khnaisser [1], Luc Lavoie [1], Hassan Diab [2], and Jean-Francois Ethier [2,3,4]

[1] Département d'informatique, Université de Sherbrooke, Sherbrooke, Canada
{christina.khnaisser, luc.lavoie}@usherbrooke.ca
[2] Centre intégré universitaire de santé et de service sociaux de l'Estrie - Centre hospitalier de Sherbrooke, Sherbrooke, Canada
hdiab.chus@ssss.gouv.qc.ca
[3] Département de médecine, Université de Sherbrooke, Sherbrooke, Canada
[4] INSERM UMR 1138 team 22 Centre de Recherche des Cordeliers, Université Paris Descartes - Sorbonne Paris Cité
ethierj@gmail.com

Abstract. In secondary data use context, traditional data warehouse design methods don't address many of today's challenges; particularly in the healthcare domain were semantics plays an essential role to achieve an effective and implementable heterogeneous data integration while satisfying core requirements. Forty papers were selected based on seven core requirements: data integrity, sound temporal schema design, query expressiveness, heterogeneous data integration, knowledge/source evolution integration, traceability and guided automation. Proposed methods were compared based on twenty-two comparison criteria. Analysis of the results shows important trends and challenges, among them (1) a growing number of methods unify knowledge with source structure to obtain a well-defined data warehouse schema built on semantic integration; (2) none of the published methods cover all the core requirements as a whole and (3) their potential in real world is not demonstrated yet.

Keywords: Data warehouse design, Clinical data warehouse, Secondary data use, Medical informatics, Bioinformatics.

1 Introduction

Large volumes of heavily fragmented healthcare domain (HD) data are generated every day from several healthcare institutions using different knowledge models and terminologies for the same episode of care. Part of this situation can be explained by the fact that patients will see different care providers in various independent organizations for the same problem. Moreover, different care processes mandate

different requirements specific to the specialty and context (e.g. acute care in hospital vs. chronic care by the treating physician in clinic) resulting in heterogeneous data.

Fragmentation must thus be resolved along at least three axes: location, time, and function. The net result is that it is very difficult to have a unified and complete view of a patient's clinical state and history. While each setting may have a very efficient system from a local perspective, not having a complete picture of a patient creates difficulties in providing optimal care, conducting efficient research and managing resources. A data warehouse (DW) is needed to uniformly integrate heterogeneous data from hundreds of independent sources with minimal human resources.

Many DW issues in the HD can be found in other domains. Some of them have been deeply investigated in the DW literature although many proposed solutions are hardly implemented in commercial DW platform. Besides, many issues have been studied independently. Possible incompatibilities or negative interactions between various solutions can be present. Previous surveys [9, 12, 14, 19, 40, 46, 56] do not clearly identify the best methods that suits the HD and some of the comparison criteria used are not well-documented. Furthermore, none of the surveys compare complete methods in the context of a real-world implementation. We took this opportunity to review the scientific literature in order to identify the relevant methods in data warehouse design (DWD). While some end products like I2B2 [36] exist, it is fundamental to first examine the design methods themselves as they will have significant implications in terms of functionality and limitations of the resulting systems. Therefore, in this paper we focus on comprehensive and integrated DWD methods that can be practically implemented in the HD.

The paper is organized as follows: section 2 describes the methodology used to select and compare papers. Section 3 presents interesting points from the evaluation results. Section 4 discusses trends and remaining challenges. Finally, section 5 concludes with open questions and potential research avenues.

2 Study methodology

The aim of this study is to help identify current DWD methods and ongoing challenges as applicable to the HD. Seven requirements have been defined from clinical data characteristics and used to evaluate methods trends and unresolved requirements. Although none of these requirements are unique to HD, they must be fully satisfied together in order to give the intended services to the HD applications.

2.1 Clinical data characteristics

Health care applications range from processing of very low level of data objects (e.g. mass and length) to very higher level of data objects (e.g. patient behavior, organism). Health care data must also be identified in time with multiple degrees of accuracy. Among others, these characteristics raise inevitable special issues and fundamental differences in comparison with many other domains [52].

A clinical DW must contend with three important characteristics of clinical data and its use in the context of secondary analysis of operational data. Firstly, clinical data is tightly coupled in nature and highly dependent on contextual information in order to fully derive its semantics. For example, while "diagnosis" may seem like a straightforward concept, many aspects can, and need to be taken into account to fully understand the nature of a diagnostic code present in a database. Is it: a diagnosis given when a patient was first admitted to the hospital (so it might change as more information becomes available), a discharge (final) diagnosis or a diagnosis entered to justify an investigation? Is it a diagnosis made by a medical student, a resident or an attending physician? Is it an active diagnosis (the patient has a pneumonia), a past diagnosis that is now resolved (the patient had pneumonia 2 years ago), or a diagnosis that was first identified in the past but that is chronic (the patient was first diagnosed with diabetes 10 years ago)? Etc. Many other similar of clinical data include the same level of complexity.

Secondly, as illustrated with the pneumonia/diabetes example above, temporality is a significant challenge with medical data. It covers the entire life of an individual. A bacterial infection at the age of three can have an impact on a heart valve disease identified at the age of fifty-five. There is also substantial uncertainty surrounding a significant part of temporal data. It is common to have a patient report that she or he has had diabetes for "more than ten years" (when in reality, the first diagnosis was 12 years ago but the disease has been present for 16 years). Querying and managing such data is challenging. This is compounded by the concept of "episode of care". For example, if a patient suffers from a major depression episode, she or he will likely see a physician multiple times for that episode. Clinical data will then show multiple entries for "major depression" during that time. Nevertheless, it is really only one episode. Now let's consider that the episode is resolved, but two years later, the patient has another episode and seeks medical attention again. Medical data will show again a "major depression" entry. It is very challenging to reconstruct the timeline for this patient and to decipher how many episodes are represented. Did the patient seek care as a follow-up for the previous episode that was never fully gone or is it a completely new episode? This is just one of the simpler situations. When intertwined with medication timing, investigations (process and results) and other care events, handling of temporality becomes quite complex.

Thirdly, the nature of data and its use for clinical care and research bring specific demands. As opposed to some other domains where requirements can be predefined with users and then implemented, clinical DW must support prospective analysis along axes that evolve rapidly as new knowledge arises.

2.2 Method requirements

From these characteristics and existing requirements for management activities, we can derive a list of requirements a clinical DWD method must satisfy:

R1 - Data integrity. The method must preserve (all available) integrity constraints to ensure data quality and correctness [45]. Data in the DW will be used to generate different kinds of reports. Results must be correct and reliable to help different end-

users (e.g. managers, cardiologists or researchers). Data needs to be stored in a neutral way as not to hinder use in one context or another.

R2 - Sound temporal schema design. Information variation over time is crucial for most analysis purposes. Having a well-defined temporal schema ensures correct temporal semantic and temporal constraint management. The final DW schema must be based on a sound, comprehensive and formalized temporal model to improve expressiveness and interoperability (like [8] and [11]).

R3 - Query expressiveness. The final DW schema must simplify the expression of queries, especially temporal ones. This may be reached by automatic generation of views specific to a target problem class expressed in terms of its contributing knowledge elements. It must also be possible to define operators specific to the problem class to facilitate data manipulation (like [50] for OLAP querying).

R4 - Heterogeneous data integration. The method must ensure heterogeneous integration of data extracted from multiple sources in a context of high fragmentation. See [3] and [32] for interesting definitions and propositions.

R5A - Knowledge evolution integration. The method must provide mechanisms to minimize errors and human resources when integrating knowledge changes. Knowledge is in constant evolution and the DW must cope with it, while maintaining earlier knowledge interpretations and preserving coherent data, correctly represented.

R5B - Source evolution integration. The method must cope with new sources integration and structural changes in existing ones with minimal impact on the DW and no impact on the end-user view of the DW (other than the availability of new data and its supporting structure). See [48] for interesting propositions.

R6 – Traceability. The method must keep track of changes in knowledge models, source availability, source structure, schema structure, and designer choices along the DW life cycle. Using mechanisms to coordinate all DWD phases is essential [9]. Traceability helps to assess the impact of structural changes and improve reusability and maintainability [31].

R7 – Guided automation. To account for the characteristics of clinical data and its fragmentation, DWD must support some degree of automation. The resulting DW scale inevitably calls for automated tools to minimize the resources needed. However, human involvement also remains necessary to handle ambiguous situations. Guided automation is a trade-off, balancing automation and human judgment while facilitating traceability efforts and minimizing errors.

2.3 Comparison criteria

Twenty-two criteria are defined to compare DWD methods and evaluate the requirements. Some criteria introduced by [57] were extended, including: automation, design approach, requirement and source representation, source analysis, algorithm, conceptual data model, logical data model, physical data model and used tools. Other criteria were added to support requirements assessment [23].

We have defined four classes (to provide case study implementation's scale):

Table 1. Case study categories

Classes	Sources	Relations	Attributes	Tuples
Pedagogical example (PE)	1	3	12	1E+02
Proof of concept (PC)	3	20	100	1E+04
Scale test (ST)	8	1 000	10 000	1E+08
Realistic test (RT)	50	10 000	100 000	1E+11

The intended use is the following: PE for illustration purpose, PC for coverage demonstration, ST for evaluating practical performance at early stages, RT for benchmarking and road test before ongoing a real deployment effort.

The complete list of criteria and their definitions can be found online at `http://info.usherbrooke.ca/llavoie/projets/epiiramide/DWDMR`

2.4 Literature selection process

Throughout the entire process, we retain only papers from year 2000 and up. At first, we targeted general methods (634 papers) with Google scholar, Summon 2.0 and Engineering Village. The final group was chosen based on the inclusion of some automation (i.e. including some kind of potential automation for the creation process). A total of 40 papers were then evaluated: [1, 4–6, 10, 13, 15–18, 21, 22, 24–31, 33–35, 37–39, 41, 43, 45, 47–49, 51, 53, 55, 57–61].

3 Compilation and results

General observations are presented, based on our results compilation available on the public share [23]. The requirements defined earlier are then reviewed and assessed.

3.1 General observations

Many methods use a hybrid approach (19/40), 6 among them including a knowledge approach. Since 2010, most methods representing requirements and/or knowledge use ontologies (12/18). Extraction for the source representation and data integration is still mostly manual. The relational model is the most common model to represent sources (8/40), although complete information on sources structural representation is rarely available. Only three methods report significant results based on multiple sources test cases. Dimensional modeling is widely used in DWD (26/40), but relational modeling is also quite present (8/40). If we restrict to temporal DWD, (5/8) are relational, (2/8) are dimensional and (1/8) is entity-attribute-value (EAV). We also notice that most authors don't distinguish between conceptual and logical model and, when they do, they may be using different definitions form one to another. Ontology-based DW are an emerging solution to address data heterogeneity [3]. Few methods used standard data sets (6/40).

3.2 Requirements

R1 - Data integrity. Data integrity constraints may come from knowledge models (KM) or, occasionally, from the sources themselves (see R4). Moreover, integrity constraints are often encapsulated in applications (not in the database), thereby increasing the complexity of extraction and validation (even in source-driven approach). Only 5 methods propose a hybrid approach including knowledge and source but no method proposes a dual source analysis (structure and data) with explicit integrity verification and validation. Most methods give very few indications on constraints preservation and propagation by their algorithms. As it stands, R1 is partially satisfied.

R2 - Sound temporal schema design. Only 8 methods address the temporal modeling explicitly. One method provides temporal DW schema based on TRM model [11] while others use *ad hoc* models (5/40). None of these methods offer a significant automation level based on knowledge temporal constraints, source temporal structure and source temporal data. Interesting representations are given in [39] and [48]. As it stands, R2 is partially satisfied.

R3 - Query expressiveness. No method addresses explicitly the issue of query expressiveness. Many of them seem to consider that views directly produced by DM design are adequate. In our experience, they may fulfill some of the managers' needs, but are not adequate when end-users (e.g. care provider or researcher) must be able to query the DW themselves, using multiple and complex knowledge models. As it stands, R3 is not satisfied at the design step.

R4 - Heterogeneous data integration. Data integration has received a large attention by the DW community over the last 30 years. Our hypothesis is that data integration must be guided by knowledge and part of the DWD design method. Only 5 methods explicitly cope with multiple sources and only 3 of them have a knowledge representation that can be used to arbitrate the heterogeneity. Only one of them addresses explicitly the ETL process, but more experiments based on a ST class case study are needed to conclude. As it stands, R4 is partially satisfied.

R5A - Knowledge evolution integration. No method reports support of knowledge evolution integration. As it stands, R5A is not currently satisfied.

R5B - Source evolution integration. Only 3 methods explicitly report support to source evolution integration. No clear indication of the ability to query retrospectively the sources based on a sound temporal model were found. As it stands, R5B is partially satisfied.

R6 – Traceability. Methods [21] and [31] report a convincing traceability approach, at different granularity level, although they don't address explicitly the knowledge representations' changes. Unfortunately, none had linked their framework with a sound temporal model. Finally, more experiments based on a ST class case study are needed. As it stands, R6 is quite fully satisfied.

R7 – Guided automation. As expected, no methods are fully automatized, neither automatized at a level that will make our project feasible. Some methods perform

quite well on discovering dimensional concepts in sources, guided by user suggestions, others, in generating ETL. Mixing best automation results (regardless of the compatibility of their methods) won't even be sufficient for source/knowledge evolution processes at least. Methods using model-driven architecture (MDA) approach can be largely automated but they lack knowledge modeling. As it stands, R7 is partially satisfied.

3.3 Compilation summary.

Within the 40 evaluated methods, no method covers all the design life cycle. When a method shows a good level of compliance on one requirement: (1) supporting algorithms need further documentation to be independently implemented; (2) no evidence, based on an ST class case study, is given that the proposed methods may tackle large problems (only 2 methods report results on a PC class case study).

We conclude that current papers do not satisfy significantly ***R1***, ***R3*** and ***R5A***; partially satisfy ***R2***, ***R4***, ***R5B*** and ***R7***; quite fully satisfy ***R6*** in an integrated method.

4 Discussion

Building a DW, taking into account clinical data characteristics and satisfying the ensuing requirements, is a challenging issue. We will now discuss three fundamental elements, mainly related to requirements R1 to R5.

4.1 Knowledge vs. Requirements

Secondary use of data for analysis is essential to improve the quality of care and conduct optimal research activities. DW will serve many studies for different health fields and medical staff. Moreover, with the opportunity to easily access data, new needs will emerge and existing needs may change. Consequently, DW must contain all available data regardless the requirements that prevailed at initial DWD. Knowledge seems more useful than requirements to decipher source structure and isolate interesting data elements to extract. A recent paper [20] presents a semi-automatic guided method following hybrid requirement/source approach that covers all DWD life cycle. Using requirements for the DWD in health domain is unfeasible regarding the complexity and the diversity of end-users, as well as evolving needs. Moreover, knowledge encapsulated in applications (not in the database) is hardly addressed. To maximize reusability and extensibility, the "ideal" method should (1) take knowledge as the basis of the initial design, (2) "easily" integrate knowledge evolution and (3) be as "requirement neutral" as possible.

4.2 Relational vs. dimensional

By convention, most DW schema are based on dimensional design model (DDM), although no consensus on its formalism has been established yet [14]. Also, DDM

design relies partly on non-consensual “best-known practices”, some of them hardly automatable. Contrariwise, relational design theory (RDT) is algorithmically well defined [7]. DDM is based on fact/dimension dichotomy which is not universal from a problem to another [34]. Furthermore, it relies on processes identification and on requirements that are unknown at DWD time. Even if the processes were all known at design time, DW schema will depend on them, thus any change in the processes may force a change on it. RDT is based on relations and integrity constraints (functional dependencies, referential constraints, temporal constraints, etc.) relying on domain knowledge and sound axioms. DDM schema evolution will be costly and may have a large impact on the whole DW schema. DDM can be used to define known, stable problems using a requirement-driven method to address particular end user’s needs. Finally, RDT can be used to define large domains using knowledge-driven approach to ensure maximum consistency and integrity of data.

Data integrity is critical when integrating a large number of data sources. Heterogeneous data integration is complicated by redundancy. Sound integration cannot be done without minimizing redundancy or adding (costly) constraints. RDT minimizes redundancy and guarantees data integrity on a sound and automatable basis. In light of the recent technology evolution, performance issues related to RDT play a much lesser role, if any.

4.3 Temporal model

Temporal clinical data warehouses are acquiring increasing importance in the health field [2]. Temporal data is important, especially for specifying and detecting clinical phenotypes [42]. In addition, a sound temporal schema plays an essential role in minimizing data incertitude, data indeterminacy and query expressiveness. Current temporal data models [11] and [54] relies on RDT to define design guidelines and constraints regarding temporal representations and constraints. Some methods rely on *ad hoc* models that might work with requirement driven methods but carry limitations. In fact, when applied to a context where prospective operations are not pre-defined, it becomes essential to have a temporal model which stands on its own, provides intrinsic computability soundness, and gives (automatable) provable transformation rules.

5 Conclusion

In 2006, Rizzi et al. [44] wrote: “Though a lot has been written about how data warehouse should be designed, there is no consensus on a design method yet”. This is still valid as none of the evaluated methods cover all the essential requirements, nor was tested in a large-scale implementation.

We presented here a new set of requirements and criteria that can be used to evaluate such methods in the context of clinical DW. This set may be useful in other application domain as well. We also identified certain limitations. Without public standard data sets, it is difficult to measure method efficiency and progress regarding

HD. The specification and creation of such a data set are essential to allow efficient development and evaluation of HD DWD methods.

Another key conclusion of our study is that using domain knowledge is essential to improve relevant data selection and interpretation. It also fosters users' autonomy as they can use data directly through the relevant knowledge representation instead of a requirement driven perspective. As a corollary, methods must tend to unify of source knowledge and domain knowledge, but the optimal knowledge representation method remains elusive at this point in time. In addition, the relational model and a sound temporal model are essential to simplify data queries and management (integrity and evolution).

In conclusion, this review identifies existing gaps between requirements for a fully functional HD DW and existing methods to create one. A large number of independent solutions exist for several requirements, but none of the papers propose a comprehensive and integrated method for the DWD process compliant to the requirements of HD.

References

1. Abelló, A., Martín, C.: A Bitemporal Storage Structure for a Corporate Data Warehouse. Proceedings of the 5th International Conference on Enterprise Information Systems. pp. 177–183. (2003)
2. Adlassnig, K.-P., Combi, C., Das, A.K., Keravnou, E.T., Pozzi, G.: Temporal representation and reasoning in medicine: Research directions and challenges. Artif. Intell. Med. 38, 2, 101–113 (2006)
3. Bakhtouchi, A., Bellatreche, L., Jean, S., Yamine, A.-A.: MIRSOFT: mediator for integrating and reconciling sources using ontological functional dependencies. Int. J. Web Grid Serv. 8, 1, 72–110 (2012)
4. Branson, A., Hauer, T., McClatchey, R., Rogulin, D., Shamdasani, J.: A data model for integrating heterogeneous medical data in the Health-e-Child project. Stud. Health Technol. Inform. 138, 13–23 (2008)
5. Burney, A., Mahmood, N., Ahsan, K.: TempR-PDM: A Conceptual Temporal Relational Model for Managing Patient Data. Proceedings of the 9th International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases. pp. 237–243. World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA (2010)
6. Chute, C.G., Beck, S.A., Fisk, T.B., Mohr, D.N.: The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. J. Am. Med. Inform. Assoc. JAMIA. 17, 2, 131–135 (2010)
7. Codd, E.F.: The Relational Model for Database Management: Version 2. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (1990)
8. Combi, C., Pozzi, G.: HMAP A Temporal Data Model Managing Intervals with Different Granularities and Indeterminacy from Natural Language Sentences. VLDB J. 9, 4, 294–311 (2001)
9. Cravero, A., Sepúlveda, S.: Multidimensional design paradigms for data warehouses: a systematic mapping study. J. Softw. Eng. Appl. 2014, 7, 53–61 (2013)
10. Cravero Leal, A., Mazón, J.N., Trujillo, J.: A business-oriented approach to data warehouse development. Ing. E Investig. 33, 1, 59–65 (2013)

11. Date, C.J., Darwen, H., Lorentzos, N.A.: Time and relational theory: temporal databases in the relational model and SQL. Morgan Kaufmann, Waltham, MA (2014)
12. Elamin, E., Feki, J.: Toward An Ontology Based Approach Fro Data Warehousing. (2014)
13. Giorgini, P., Rizzi, S., Garzetti, M.: GRAnD: A goal-oriented approach to requirement analysis in data warehouses. Decis. Support Syst. 4–21 (2008)
14. Gosain, A., Singh, J.: Conceptual Multidimensional Modeling for Data Warehouses: A Survey. Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014. pp. 305–316. Springer (2015)
15. Hachaichi, Y., Feki, J.: An Automatic Method for the Design of Multidimensional Schemas From Object Oriented Databases. Int. J. Inf. Technol. Decis. Mak. 12, 6, 1223–1259 (2013)
16. Hu, H., Correll, M., Kvecher, L., Osmond, M., Clark, J., Bekhash, A., Schwab, G., Gao, D., Gao, J., Kubatin, V., Shriver, C.D., Hooke, J.A., Maxwell, L.G., Kovatich, A.J., Sheldon, J.G., Liebman, M.N., Mural, R.J.: DW4TR: A Data Warehouse for Translational Research. J. Biomed. Inform. 44, 6, 1004–1019 (2011)
17. Husemann, B., Lechtenbörger, J., Vossen, G.: Conceptual Data Warehouse Design. Proceedings of the International Workshop on Design and Management of Data Warehouses, DMDW 2000. pp. 3–9. (2000)
18. Jensen, M.R., Holmgren, T., Pedersen, T.B.: Discovering Multidimensional Structure in Relational Data. In: Kambayashi, Y., Mohania, M., and Wöß, W. (eds.) Data Warehousing and Knowledge Discovery. pp. 138–148. Springer Berlin Heidelberg (2004)
19. Jindal, R., Taneja, S., others: Comparative study of data warehouse design approaches: a survey. Int. J. Database Manag. Syst. 4, 1, 33–45 (2012)
20. Jovanovic, P., Romero, O., Simitsis, A., Abelló, A., Candón, H., Nadal, S.: Quarry: Digging Up the Gems of Your Data Treasury. In: Alonso, G., Geerts, F., Popa, L., Barceló, P., Teubner, J., Ugarte, M., Bussche, J.V. den, and Paredaens, J. (eds.) Proceedings of the 18th International Conference on Extending Database Technology, EDBT 2015, Brussels, Belgium, March 23-27, 2015. pp. 549–552. OpenProceedings.org (2015)
21. Jovanovic, P., Romero, O., Simitsis, A., Abelló, A., Mayorova, D.: A requirement-driven approach to the design and evolution of data warehouses. Inf. Syst. 44, 94–119 (2014)
22. Kerkri, E.M., Quantin, C., Allaert, F.A., Cottin, Y., Charve, P., Jouanot, F., Yétongnon, K.: An Approach for Integrating Heterogeneous Information Sources in a Medical Data Warehouse. J. Med. Syst. 25, 3, 167–176 (2001)
23. Khnaisser, C., Lavoie, L., Diab, H., Éthier, J.-F.: Data Warehouse Design Methods Review for the Healthcare Domain, http://info.usherbrooke.ca/llavoie/projets/epiiramide
24. Khouri, S., Bellatreche, L., Jean, S., Ait-Ameur, Y.: Requirements driven data warehouse design: We can go further. 6th International Symposium on Leveraging Applications of Formal Methods, Verification and Validation, ISoLA 2014, October 8, 2014 - October 11, 2014. pp. 588–603. Springer Verlag (2014)
25. Khouri, S., Boukhari, I., Bellatreche, L., Sardet, E., Jean, S., Baron, M.: Ontology-based structured web data warehouses for sustainable interoperability: requirement modeling, design methodology and tool. Comput. Ind. 63, 8, 799–812 (2012)
26. Krneta, D., Jovanovic, V., Marjanovic, Z.: A direct approach to physical Data Vault design. Comput. Sci. Inf. Syst. 11, 2, 569–599 (2014)
27. Lin, S.-H., Lee, Y.-C.G., Hsu, C.-Y.: Data Warehouse Approach to Build a Decision-Support Platform for Orthopedics Based on Clinical and Academic Requirements. In: Ślęzak, D., Arslan, T., Fang, W.-C., Song, X., and Kim, T. (eds.) Bio-Science and Bio-Technology. pp. 89–96. Springer Berlin Heidelberg (2009)

28. Lowe, H.J., Ferris, T.A., Hernandez, P.M., Weber, S.C.: STRIDE – An Integrated Standards-Based Translational Research Informatics Platform. AMIA. Annu. Symp. Proc. 2009, 391–395 (2009)
29. Lujan-Mora, S., Trujillo, J.: Applying the UML and the Unified Process to the design of Data Warehouses. J. Comput. Inf. Syst. 47, 5, 30–58 (2006)
30. Malinowski, E., Zimányi, E.: A Conceptual Solution for Representing Time in Data Warehouse Dimensions. Proceedings of the 3rd Asia-Pacific Conference on Conceptual Modelling - Volume 53. pp. 45–54. Australian Computer Society, Inc., Darlinghurst, Australia, Australia (2006)
31. Maté, A., Trujillo, J.: Tracing conceptual models' evolution in data warehouses by using the model driven architecture. Comput. Stand. Interfaces. 36, 5, 831–843 (2014)
32. Mate, S., Köpcke, F., Toddenroth, D., Martin, M., Prokosch, H.-U., Bürkle, T., Ganslandt, T.: Ontology-Based Data Integration between Clinical and Research Systems. PLoS ONE. 10, 1, (2015)
33. Mazón, J.-N., Trujillo, J., Lechtenbörger, J.: Reconciling requirement-driven data warehouses with data sources via multidimensional normal forms. Data Knowl. Eng. 63, 3, 725–751 (2007)
34. Moreira, J., Cordeiro, K., Campos, M.L., Borges, M.: OntoWarehousing – Multidimensional Design Supported by a Foundational Ontology: A Temporal Perspective. In: Bellatreche, L. and Mohania, M.K. (eds.) Data Warehousing and Knowledge Discovery. pp. 35–44. Springer International Publishing (2014)
35. De Mul, M., Alons, P., van der Velde, P., Konings, I., Bakker, J., Hazelzet, J.: Development of a clinical data warehouse from an intensive care clinical information system. Comput. Methods Programs Biomed. 105, 1, 22–30 (2012)
36. Murphy, S.N., Weber, G., Mendis, M., Gainer, V., Chueh, H.C., Churchill, S., Kohane, I.: Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J. Am. Med. Inform. Assoc. 17, 2, 124–130 (2010)
37. Nazri, M.N.M., Noah, S.A., Hamid, Z.: Using Lexical Ontology for Semi-automatic Logical Data Warehouse Design. In: Yu, J., Greco, S., Lingras, P., Wang, G., and Skowron, A. (eds.) Rough Set and Knowledge Technology. pp. 257–264. Springer Berlin Heidelberg (2010)
38. Nebot, V., Berlanga, R.: Building data warehouses with semantic web data. Decis. Support Syst. 52, 4, 853–868 (2012)
39. Neil, C.G., De Vincenzi, M.E., Pons, C.F.: Design method for a Historical Data Warehouse, explicit valid time in multidimensional models. Ingeniare Rev. Chil. Ing. 22, 2, 218–232 (2014)
40. Pardillo, J., Mazón, J.-N.: Using ontologies for the design of data warehouses. Int. J. Database Manag. Syst. 3, 2, (2011)
41. Phipps, C., Davis, K.C.: Automating Data Warehouse Conceptual Schema Design and Evaluation. Design and Management of Data Warehouses. pp. 23–32. Citeseer (2002)
42. Post, A.R., Kurc, T., Cholleti, S., Gao, J., Lin, X., Bornstein, W., Cantrell, D., Levine, D., Hohmann, S., Saltz, J.H.: The Analytic Information Warehouse (AIW): A platform for analytics using electronic health record data. J. Biomed. Inform. 46, 3, 410–424 (2013)
43. Prat, N., Akoka, J., Comyn-Wattiau, I.: A UML-based data warehouse design method. Decis. Support Syst. 42, 3, 1449–1473 (2006)
44. Rizzi, S., Abello, A., Lechtenborger, J., Trujillo, J.: Research in data warehouse modeling and design: Dead or alive? 9th ACM International Workshop on Data Warehousing and OLAP - DOLAP'06, held in conjunction with the ACM 15th Conference on Information

and Knowledge Management, CIKM 2006, November 10, 2006 - November 10, 2006. pp. 3–10. Association for Computing Machinery, New York, NY, USA (2006)
45. Romero, O., Abelló, A.: A framework for multidimensional design of data warehouses from ontologies. Data Knowl. Eng. 69, 11, 1138–1157 (2010)
46. Romero, O., Abelló, A.: A Survey of Multidimensional Modeling Methodologies. Int. J. Data Warehous. Min. IJDWM. 5, 2, 1 – 23 (2009)
47. Romero, O., Simitsis, A., Abelló, A.: GEM: Requirement-Driven Generation of ETL and Multidimensional Conceptual Designs. In: Cuzzocrea, A. and Dayal, U. (eds.) Data Warehousing and Knowledge Discovery. pp. 80–95. Springer Berlin Heidelberg (2011)
48. Rönnbäck, L., Regardt, O., Bergholtz, M., Johannesson, P., Wohed, P.: Anchor modeling — Agile information modeling in evolving data environments. Data Knowl. Eng. 69, 12, 1229–1253 (2010)
49. Rubin, D.L., Desser, T.S.: A Data Warehouse for Integrating Radiologic and Pathologic Data. J. Am. Coll. Radiol. 5, 3, 210–217 (2008)
50. Sabaini, A., Zimányi, E., Combi, C.: An OLAP-Based Approach to Modeling and Querying Granular Temporal Trends. In: Bellatreche, L. and Mohania, M.K. (eds.) Data Warehousing and Knowledge Discovery. pp. 69–77. Springer International Publishing (2014)
51. Sahama, T.R., Croll, P.R.: A Data Warehouse Architecture for Clinical Data Warehousing. Proceedings of the 5th Australasian Symposium on ACSW Frontiers. pp. 227–232. Australian Computer Society, Inc., Darlinghurst, Australia, Australia (2007)
52. Shortliffe, E.H., Cimino, J.C. eds: Biomedical informatics: computer applications in health care and biomedicine. Springer, London (2014)
53. Sitompul, O.S., Noah, S.A.: A Transformation-oriented Methodology to Knowledge-based Conceptual Data Warehouse Design. J. Comput. Sci. 2, 5, 460–465 (2006)
54. Snodgrass, R.T.: Developing time-oriented database applications in SQL. Morgan Kaufmann Publishers, San Francisco, Calif (2000)
55. Song, I.Y., Khare, R., Dai, B.: SAMSTAR: a semi-automated lexical method for generating star schemas from an entity-relationship diagram. Proceedings of the ACM tenth international workshop on Data warehousing and OLAP. pp. 9–16. ACM (2007)
56. Tebourski, W., Karâa, W.B.A., Ghezala, H.B.: Semi-automatic Data Warehouse Design methodologies: a survey. Int. J. Comput. Sci. Issues IJCSI. 10, 5, 48 (2013)
57. Thenmozhi, M., Vivekanandan, K.: A Tool for Data Warehouse Multidimensional Schema Design using Ontology. Int. J. Comput. Sci. Issues IJCSI. 10, 2, 161–168 (2013)
58. Di Tria, F., Lefons, E., Tangorra, F.: Hybrid methodology for data warehouse conceptual design by UML schemas. Inf. Softw. Technol. 54, 4, 360–379 (2012)
59. Wisniewski, M.F., Kieszkowski, P., Zagorski, B.M., Trick, W.E., Sommers, M., Weinstein, R.A.: Development of a Clinical Data Warehouse for Hospital Infection Control. J. Am. Med. Inform. Assoc. JAMIA. 10, 5, 454–462 (2003)
60. Zekri, M., Marsit, I., Adellatif, A.: A New Data Warehouse Approach Using Graph. 2011 IEEE 8th International Conference on e-Business Engineering (ICEBE). pp. 65–70. IEEE Computer Society (2011)
61. Zepeda, L., Ceceña, E., Quintero, R., Zatarain, R., Vega, L., Mora, Z., Clemente, G.G.: A MDA Tool for Data Warehouse. 2010 International Conference on Computational Science and Its Applications (ICCSA). pp. 261–265. (2010)