Data Warehouse Design Methods Review for the Healthcare Domain

Christina Khnaisser [1], Luc Lavoie [1], Hassan Diab [2], and Jean-François Éthier [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. This paper presents current research trends and ongoing challenges in data warehouse design methods. 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

With the development of multiple systems and tools over the last decades, the healthcare domain now relies heavily on data consumption for the provision of care to patients in multiple contexts. Electronic health records (EHR) play an essential role and the importance of this field has been recognized by initiatives like Horizon 2020 [17] in Europe and Meaningful use [54] in the United States where

substantial resources have been invested to foster the uptake of existing methods and to stimulate innovation focusing on quality of care, patient safety and cost reduction.

Similar challenges face the region of Estrie, in the province of Quebec, Canada. It provides healthcare to a population of over 500 000 citizens. It includes a tertiary university care center, four local hospitals as well as multiple primary care and specialist clinics, each with a different, independent local EHR system. In order to address data fragmentation on its territory, the region launched a project for the creation of a platform to unify data and to facilitate linkage of its data with external sources like provincial registries.

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 (e.g. primary clinics, specialist clinics, hospitals) 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. More specifically, we are searching for a comprehensive, largely automated and tested design method that meets the requirements needed for a clinical DW. This DWD method must cover the initial schema creation process (which includes knowledge representation, source representation, conceptual design, logical design and physical design), the Knowledge-Requirements-Source (KRS) mapping process, the Extract-Transform-Load (ETL) generation process, and the DW evolution processes. Given the large number of sources, a significant part of the processes must be automated with or without parameterization and guidance. We will then conduct case studies at different scales, from pedagogical examples to realistic tests before choosing the most suitable DWD method for our project.

Data warehouses have been previously used to assemble and present operational data. While the term is used to represent different structures in the literature, the seminal definition published by Inmon [27] is: “a subject-oriented, integrated, time variant and non-volatile collection of data in support of management decisions”. It is usually built and maintained away from operational systems. While some aspects of this definition are still debated, it gives a good overview. It is worth noting here the emphasis on “management decisions”. DW has been used successfully in many fields to facilitate decision-making based on data originally collected during and for business operations. Nonetheless, a fully integrated clinical DW will need to satisfy not only management requirements, but also clinical care and research needs. Three major aspects underpin such needs: clinical data is tightly coupled and requires

contextual information along multiples axes to fully define its semantics; temporal relations span a long period from birth to death with significant uncertainty around temporal data, resulting in complex temporal operations; paradigms for data analysis and query design will evolve rapidly through the life of the DW as new knowledge emerges and so cannot be fully pre-defined. Clinical research also comes with special needs: some request must be formulated from different knowledge models and the resulting information then compared. All these characteristics give the "subject-oriented" qualification a rather large scope. Finally, the complexity inherent to these three aspects is compounded by data fragmentation as discussed above and so the need to host data from multiple sources in the DW keeping track of the original sources used to produce any synthesized information.

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 then be present. Previous surveys [11, 16, 22, 29, 62, 69, 81] do not clearly identify the best methods that suits 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 [57] 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.

As a first step in our research, we reviewed the published methods covering DWD. We then created a list of requirements needed by the creation process of a DW that would address the needs of clinical care providers, researchers and decision makers (end-users). We then reviewed the relevant literature in regards to these requirements and report our findings here.

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 (data integrity, sound temporal schema design, query expressiveness, heterogeneous data integration, knowledge and source evolution integration, traceability and guided automation) and used to evaluate methods by mapping criteria with requirements that allows identifying 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). Finally, health care data must 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 domain data [76].

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 a diagnosis for the patient at hand or a diagnosis of one of his family members? 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 aspects 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, using only EHR data, 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 most of the requirements can be predefined with users and then implemented, clinical DW must be flexible and support prospective analysis along axes that evolve rapidly as new knowledge arises. Knowledge is in constant evolution and data generated a few years ago will need to be re-analyzed based on new paradigms.

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 [68]. 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 [10] and [13]).

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 [74] for OLAP querying granular temporal trends).

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 [49] 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 [72] 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 [11]. Traceability helps to assess the impact of structural changes and improve reusability and maintainability [48].

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 [83] 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 [34].

Design Approach (D. App.). Methods are classified within three design approaches [69]: a source-driven approach (also named data-driven, supply-driven, bottom-up) starts from data sources in order to derive the DW schema, a requirement-driven approach (also named demand-driven, goal-driven, top-down) starts from user requirements, and a hybrid approach (also named mixed approach) combines both approaches. We can distinguish a fourth approach that we named knowledge-driven approach that focus on domain knowledge to identify relevant concepts to structure and design the DW schema. We can thereby extend the hybrid approach definition as combining requirements and sources (R-S), knowledge and sources (K-S) or requirement, knowledge and source (R-K-S).

Process (P-CRE, P-MAP, P-ETL, P-KEV, P-KEV, P-SEV). DWD life cycle combines difficult, complex activities. Kimball [37] presents the whole DW life cycle starting from business requirements definition to the implementation phase including maintenance, evolution and project management. Designers tackle several challenges in all phases. Even so, there is no method addressing the cycle as a whole [53]. In the present study, we focused only on the design phase, including: the initial schema creation process (P-CRE which includes requirement representation, knowledge representation, source representation, conceptual design, logical design and physical design), the Knowledge-Requirements-Source mapping process (P-MAP), the Extract-Transform-Load generation (P-ETL) process, and the DW evolution processes (P-xEVs). The P-xEVs include the knowledge DW evolution (P-KEV), the requirement evolution (P-REV) and the source DW evolution (P-SEV). A detailed definition of each process is out of the scope of this paper.

We evaluate each process (if it is taken into account by the method) according to the level of automation: fully automated, mostly automated, partially automated or not significantly automated. Some activities need user's suggestions to generate a complete output like ontology annotations [68]. In some other cases, the designer's input is required to approve algorithm propositions. For each level, we distinguished the type of automation: with or without parameterization and guidance.

Knowledge representation (K. Rep.). This criterion identifies the model used to represent domain knowledge. Also called model of meaning, it is a sound representation of domain entities and the relations between them [66].

Requirement representation (R. Rep.). This criterion identifies the model used to represent end user's requirements.

Source representation (S. Rep.). This criterion identifies the model used to represent a source (not to confuse with the source type).

Source analysis (S. Ana.). Source analysis can be done on the structure (aka meta-data, S), the data (D) or both (D-S).

Multiple sources (Multi. S.). DW schema can be derived from multiple sources. This assumes that the method takes the integration of the data and the structure of the sources into account.

Algorithms definition (Algo.). Authors published all required algorithms in a manner such that they can be implemented independently.

Conceptual data model (CDM). A CDM aims at identifying and describing the concepts as they are understood by end-users; for further details, see DIV-1 models in [15]. The criterion is used to document the preferred model used by the method, if any. Typical values are Entity-Relationship model (ERM), Ontology model (OM), Dimensional-Fact model (DFM) [20], etc.

Logical data model (LDM). A LDM aims at defining data requirements (as types, constraints, etc.) relative to a deductive framework (usually based on the first order logic) of a corresponding CDM; for further details, see DIV-2 models in [15]. The criterion is used to document the preferred models used by the method, if any. Typical values are the Relational model (RDT), the Star model (Star), the Tagged Graph Model (TGM), etc.

Physical data model (PDM). A PDM aims at defining the representation (as data structures and access methods) of a corresponding LDM; for further details, see DIV-3 models in [15]. The criterion is used to document the preferred mechanisms by which the method transforms its LDM in PDM, if any. Typical values are SQL (more precisely a RDBMS implementing the SQL language, such as Oracle or PostgreSQL), MOLAP, ROLAP, OBDW (like Onto DB [14]) etc.

Temporal data model (TDM). This criterion specifies the temporal data model (BCDM [79], TRM [13], AV [30], etc.) used to design the DW schema (if any).

DW type. This criterion specifies the DW type (Relational [9], Dimensional [37], Anchor [72], Data Vault [25], etc.) produced by the DWD process.

Case study (Case). This criterion specifies if the papers present (refer to) well-documented case studies. We have defined four classes (to provide an approximation of the 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`

Standard data sets. The case study benchmarks are based on publicly available data sets.

Used tools and techniques. A descriptive criterion: the list of tools and techniques used to design the DW schema as reported by the authors.

Complementary papers. A descriptive criterion: the list of other published materials used to extend the method.

2.4 Literature selection process

The literature review process is summarized in figure 1. 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 using: "Data warehouse" AND ("design methodology" OR "design method"). We then targeted more clinical specific method with PubMed using: ("Data warehouse"[Title/Abstract]) AND (Design[Title/Abstract]); "Clinical data warehouse"; "Medical Data warehouse".

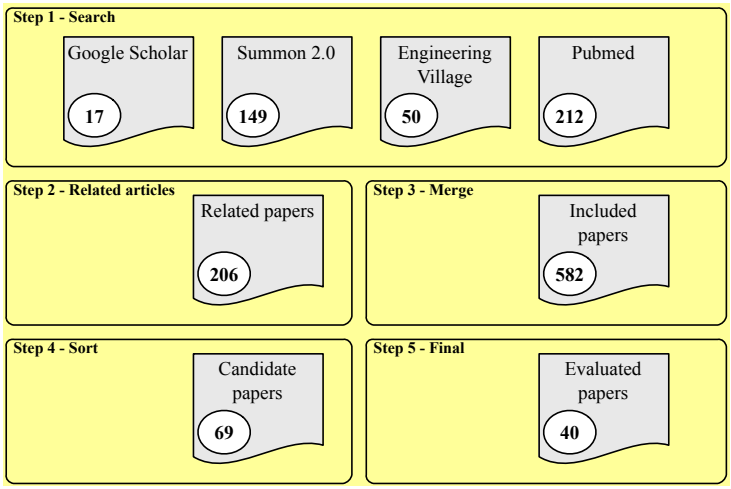

Fig. 1. Selection process (February 2015)

We then looked for citations and related papers of those found in the initial search with addition mostly related to [83] and [20]. Duplicates were then eliminated and we selected only journal and conference papers explicitly addressing data warehouse design methods. Papers related to XML DW, object-oriented DW, single data marts and standalone OLAP cubes were excluded, leaving sixty-nine candidate papers. The final group was chosen based on the inclusion of some automation (i.e. including some kind of potential automation for P-CRE, the creation process). When multiple papers referred to the same method, only one was retained. A total of 40 papers were then evaluated: [1, 6–8, 12, 18, 23, 24, 26, 28, 32, 33, 35, 36, 38, 42–44, 46, 48, 52, 55, 56, 58, 59, 61, 63, 65, 68, 71–73, 75, 77, 80, 83–87].

3 Results compilation

First, we present general observations based on our results compilation available on the public share [34]. The requirements defined earlier are then reviewed and assessed. A summary concludes the section.

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). When they come from the KM, source analysis is required to cope with non-compliant data or non-compliant structure. 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 addressed by some methods, but no rigorous evaluation has been presented to prove the algorithm efficiency in real scale problems, so R1 is only 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 [13] 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 [61] and [72]. 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 [32] and [48] 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 (which is required to obtain full query expressiveness, R3). 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. True guidance requires an easy walk-through between CDM and LDM that suppose a strong compatibility model (or identity) between

them, restricting again the ability to merge methods. Methods using model-driven architecture (MDA) approach can be largely automated (from requirements to physical schema generation) 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.

Requirements	***R1***	***R2***	***R3***	***R4***	***R5A***	***R5B***	***R6***	***R7***
Not satisfied	X		X		X			
Partially satisfied		X		X		X		X
Mostly satisfied							X	

Table 2. Requirements evaluation summary

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. Clinicians and researchers need to explore data to scope their study. At the beginning, requirements are not known or rarely exhaustive. 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 [31] 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. Thus, following Inmon [27] architecture, we suggest building the DW using domain knowledge (K-S

approach) and then building specialized data marts using user requirements (R-S approach, S being the DW). 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, including stars and snowflakes variant), although no consensus on its formalism has been established yet [22]. 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 [9]. DDM is based on fact/dimension dichotomy which is not universal from a problem to another [55]. 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. 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. With vertical representation [39] and in-memory databases [41] at hand, performance may even be better with a RDT DW than with a (denormalized) DDM DW.

4.3 Temporal model

Temporal clinical data warehouses are acquiring increasing importance in the health field [2]. Time-based decision support in healthcare in needed to improve health quality. Reasoning with temporal data provides more accurate representations of the patients' states and the events causing state changes. Temporal data is important, especially for specifying and detecting clinical phenotypes [64]. Accurate data are needed to ensure seemly results. In addition, a temporal sound schema plays an essential role in minimizing data incertitude, data indeterminacy and query expressiveness. Current temporal data models [13] and [79] 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. [67] 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. While we identified three characteristics essential for functional clinical DWD methods, others might emerge and would need to be added to the list.

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.

In the end, this review enables us to turn our attention to the next step: evaluation of end products. Classifying them by design methods will allow us to focus on expected gaps and strengths as identified in this study. While further discussions regarding our framework will need to take place in the community to build consensus, we believe it can inform future development in the field. The challenge is to find a way to combine best existing compatible solutions to form an integrated design method with a high automation potential.

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. Bellatreche, L., Khouri, S., Berkani, N.: Semantic Data Warehouse Design: From ETL to Deployment à la Carte. In: Meng, W., Feng, L., Bressan, S., Winiwarter, W., and Song, W. (eds.) Database Systems for Advanced Applications. pp. 64–83. Springer Berlin Heidelberg (2013)
5. Berkani, N., Khouri, S., Bellatreche, L.: Generic Methodology for Semantic Data Warehouse Design: From Schema Definition to ETL. 4th International Conference on Intelligent Networking and Collaborative Systems (INCoS). pp. 404–411. IEEE Computer Society (2012)
6. 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)
7. 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)
8. 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)
9. Codd, E.F.: The Relational Model for Database Management: Version 2. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (1990)
10. 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)
11. Cravero, A., Sepúlveda, S.: Multidimensional design paradigms for data warehouses: a systematic mapping study. J. Softw. Eng. Appl. 2014, 7, 53–61 (2013)
12. Cravero Leal, A., Mazón, J.N., Trujillo, J.: A business-oriented approach to data warehouse development. Ing. E Investig. 33, 1, 59–65 (2013)
13. 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)
14. Dehainsala, H., Pierra, G., Bellatreche, L.: OntoDB: An Ontology-Based Database for Data Intensive Applications. In: Kotagiri, R., Krishna, P.R., Mohania, M., and Nantajeewarawat, E. (eds.) Advances in Databases: Concepts, Systems and Applications. pp. 497–508. Springer Berlin Heidelberg (2007)
15. Deputy Chief Information Officer: DODAF - DOD Architecture Framework Version 2.02, http://dodcio.defense.gov/TodayinCIO/DoDArchitectureFramework.aspx

16. Elamin, E., Feki, J.: Toward An Ontology Based Approach Fro Data Warehousing. (2014)
17. European Commission: Horizon 2020 - The EU Framework Programme for Research and Innovation, http://ec.europa.eu/programmes/horizon2020/
18. Giorgini, P., Rizzi, S., Garzetti, M.: GRAnD: A goal-oriented approach to requirement analysis in data warehouses. Decis. Support Syst. 4–21 (2008)
19. Golfarelli, M., Maio, D., Rizzi, S.: Conceptual design of data warehouses from E/R schemes. Proceedings of the Thirty-First Hawaii International Conference on System Sciences, 1998. pp. 334–343 vol.7. (1998)
20. Golfarelli, M., Maio, D., Rizzi, S.: The Dimensional Fact Model: A Conceptual Model For Data Warehouses. Int. J. Coop. Inf. Syst. 7, 215–247 (1998)
21. Golfarelli, M., Rizzi, S., Saltarelli, E.: WAND: A CASE Tool for Workload-Based Design of a Data Mart. SEBD. pp. 422–426. Citeseer (2002)
22. 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)
23. 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)
24. 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)
25. Hultgren, H.: Modeling the Agile Data Warehouse with Data Vault. Brighton Hamilton, Denver, Colo.; Stockholm (2012)
26. 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)
27. Inmon, W.H.: Building the data warehouse. J. Wiley, Indianapolis, Ind (2005)
28. 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)
29. Jindal, R., Taneja, S., others: Comparative study of data warehouse design approaches: a survey. Int. J. Database Manag. Syst. 4, 1, 33–45 (2012)
30. Johnston, T., Weis, R.: Managing time in relational databases: how to design, update and query temporal data. Morgan Kaufmann/Elsevier, Amsterdam ; Boston (2010)
31. 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)
32. 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)
33. 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)
34. 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

35. 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)
36. 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)
37. Kimball, R. ed: The Data Warehouse Lifecycle Toolkit. Wiley Pub, Indianapolis, IN (2008)
38. Krneta, D., Jovanovic, V., Marjanovic, Z.: A direct approach to physical Data Vault design. Comput. Sci. Inf. Syst. 11, 2, 569–599 (2014)
39. Lamb, A., Fuller, M., Varadarajan, R., Tran, N., Vandiver, B., Doshi, L., Bear, C.: The Vertica Analytic Database: C-store 7 Years Later. Proc VLDB Endow. 5, 12, 1790–1801 (2012)
40. Lechtenbörger, J., Vossen, G.: Multidimensional normal forms for data warehouse design. Inf. Syst. 28, 5, 415 – 434 (2003)
41. Lee, J., Kwon, Y.S., Farber, F., Muehle, M., Lee, C., Bensberg, C., Lee, J.Y., Lee, A.H., Lehner, W.: SAP HANA distributed in-memory database system: Transaction, session, and metadata management. 2013 IEEE 29th International Conference on Data Engineering (ICDE). pp. 1165–1173. (2013)
42. 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)
43. 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)
44. 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)
45. Malinowski, E., Zimányi, E.: A conceptual model for temporal data warehouses and its transformation to the ER and the object-relational models. Data Knowl. Eng. 64, 1, 101–133 (2008)
46. 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)
47. Martin, C., Abelló, A.: A Temporal Study of Data Sources to Load a Corporate Data Warehouse. In: Kambayashi, Y., Mohania, M., and Wöß, W. (eds.) Data Warehousing and Knowledge Discovery. pp. 109–118. Springer Berlin Heidelberg (2003)
48. 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)
49. 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)
50. Mazón, J.-N., Pardillo, J., Trujillo, J.: A Model-Driven Goal-Oriented Requirement Engineering Approach for Data Warehouses. In: Hainaut, J.-L., Rundensteiner, E.A., Kirchberg, M., Bertolotto, M., Brochhausen, M., Chen, Y.-P.P., Cherfi, S.S.-S., Doerr, M., Han, H., Hartmann, S., Parsons, J., Poels, G., Rolland, C., Trujillo, J., Yu, E., and

Zimányie, E. (eds.) Advances in Conceptual Modeling – Foundations and Applications. pp. 255–264. Springer Berlin Heidelberg (2007)
51. Mazón, J.-N., Trujillo, J.: A Hybrid Model Driven Development Framework for the Multidimensional Modeling of Data Warehouses. SIGMOD Rec. 38, 2, 12–17 (2009)
52. 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)
53. Mazon, J.-N., Trujillo, J., Serrano, M., Piattini, M.: Applying MDA to the Development of Data Warehouses. Proceedings of the 8th ACM International Workshop on Data Warehousing and OLAP. pp. 57–66. ACM, New York, NY, USA (2005)
54. Medicare, C. for, Baltimore, M.S. 7500 S.B., Usa, M.: Meaningful_Use, http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Meaningful_Use.html
55. 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)
56. 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)
57. 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)
58. 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)
59. Nebot, V., Berlanga, R.: Building data warehouses with semantic web data. Decis. Support Syst. 52, 4, 853–868 (2012)
60. Nebot, V., Berlanga, R., Pérez, J.M., Aramburu, M.J., Pedersen, T.B.: Multidimensional Integrated Ontologies: A Framework for Designing Semantic Data Warehouses. In: Spaccapietra, S., Zimányi, E., and Song, I.-Y. (eds.) Journal on Data Semantics XIII. pp. 1–36. Springer Berlin Heidelberg (2009)
61. 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)
62. Pardillo, J., Mazón, J.-N.: Using ontologies for the design of data warehouses. Int. J. Database Manag. Syst. 3, 2, (2011)
63. Phipps, C., Davis, K.C.: Automating Data Warehouse Conceptual Schema Design and Evaluation. Design and Management of Data Warehouses. pp. 23–32. Citeseer (2002)
64. 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)
65. Prat, N., Akoka, J., Comyn-Wattiau, I.: A UML-based data warehouse design method. Decis. Support Syst. 42, 3, 1449–1473 (2006)
66. Rector, A.L., Qamar, R., Marley, T.: Binding ontologies and coding systems to electronic health records and messages. Appl. Ontol. 4, 1, 51–69 (2009)
67. 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)
68. Romero, O., Abelló, A.: A framework for multidimensional design of data warehouses from ontologies. Data Knowl. Eng. 69, 11, 1138–1157 (2010)
69. Romero, O., Abelló, A.: A Survey of Multidimensional Modeling Methodologies. Int. J. Data Warehous. Min. IJDWM. 5, 2, 1 – 23 (2009)
70. Romero, O., Abelló, A.: Automatic validation of requirements to support multidimensional design. Data Knowl. Eng. 69, 9, 917–942 (2010)
71. 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)
72. 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)
73. Rubin, D.L., Desser, T.S.: A Data Warehouse for Integrating Radiologic and Pathologic Data. J. Am. Coll. Radiol. 5, 3, 210–217 (2008)
74. 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)
75. 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)
76. Shortliffe, E.H., Cimino, J.C. eds: Biomedical informatics: computer applications in health care and biomedicine. Springer, London (2014)
77. 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)
78. Skoutas, D., Simitsis, A.: Ontology-Based Conceptual Design of ETL Processes for Both Structured and Semi-Structured Data: Int. J. Semantic Web Inf. Syst. 3, 4, 1–24 (2007)
79. Snodgrass, R.T.: Developing time-oriented database applications in SQL. Morgan Kaufmann Publishers, San Francisco, Calif (2000)
80. 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)
81. 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)
82. Thenmozhi, M., Vivekanandan, K.: An ontology based hybrid approach to derive multidimensional schema for data warehouse. Int. J. Comput. Appl. 54, 8, 36–42 (2012)
83. 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)
84. 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)
85. 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)
86. 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)

87. 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)

Appendix

Cross-References

#	Authors	Base paper	Complementary paper
1	Jovanovic et al.	[32]	[31]
2	Khouri et al	[35]	[4]
3	Krneta et al.	[38]	-
4	Maté and Trujillo	[48]	[50]
5	Moreira et al.	[55]	-
6	Neil et al.	[61]	-
7	Cravero Leal et al	[12]	-
8	Hachaichi and Feki	[23]	-
9	Thenmozhi and Vivekanandan	[83]	[82]
10	de Mul et al.	[56]	-
11	Di Tria et al.	[84]	[51]
12	Khouri et al.	[36]	[5]
13	Nebot and Berlanga	[59]	[60]
14	Hu et al.	[24]	-
15	Romero et al.	[71]	[78]
16	Zekri et al.	[86]	-
17	Burney et al.	[7]	-
18	Chute et al.	[8]	-
19	Nazri et al.	[58]	-
20	Romero and Abelló	[68]	[70]
21	Rönnbäck et al.	[72]	-
22	Zepeda et al.	[87]	[53]
23	Lin et al.	[42]	-
24	Lowe et al.	[43]	-
25	Branson et al.	[6]	-
26	Giorgini et al.	[18]	[19]
27	Rubin and Desser	[73]	-
28	Mazon et al.	[52]	[40]
29	Sahama and Croll	[75]	-
30	Song et al.	[80]	-
31	Lujan-Mora and Trujillo	[44]	-
32	Malinowski and Zimányi	[46]	[45]
33	Prat et al.	[65]	-
34	Sitompul and Noah	[77]	[19]
35	Jensen et al.	[28]	-
36	Abelló and Martín	[1]	[47]
37	Wisniewski et al.	[85]	-
38	Phipps and Davis	[63]	[21]
39	Kerkri et al.	[33]	-
40	Husemann et al.	[26]	-

#	P#	Year	D. App.	P-CRE	P-MAP	P-ETL	P-KEV	P-REV	P-SEV	K. Rep.	R. Rep.	S. Rep.	S Ana.	Multi. S.	Algo.
1	[32]	14	R	mg	ma			mg	ma	OM	Tagged Graph	none	No	Yes	Yes
2	[35]	14	R	ma	ma					OM	OM	none	??	No	inc.
3	[38]	14	S	pg	??					none	none	RDT	S	Yes	Yes
4	[48]	14	R-S	pg	xx			??	??	none	MD-UML	CWM	S	No	No
5	[55]	14	K-R-S	pg	xx					OM	Text	??	S	No	No
6	[61]	14	S	mg						none	??	ERM	S	No	No
7	[12]	13	R	xx						none	BMM	??	??	??	No
8	[23]	13	S	mg		ma				none	none	ODMG	S	No	Yes
9	[83]	13	R-S	pg	ma					none	OM	OM	S	No	Yes
10	[56]	12	R	xx		??				none	Text	??	??	No	No
11	[84]	12	R-S	pg	mg					OM	MD-UML	exDFM	S	??	Yes
12	[36]	12	R-S	ma	ma	??	??	??	??	OM	Goal Model	none	No	Yes	inc.
13	[59]	12	K-R	pg	pa	pa				OM	DL Exp.	??	No	Yes	Yes
14	[24]	11	K-S	xx		??				OM	none	EAV	No	??	n/a
15	[71]	11	R-S	mg	ma	pg				none	i* model	OM	S	??	Yes
16	[86]	11	S	pa						none	Graph	Graph	S	No	No
17	[7]	10	K	n/a						OM	DFD	??	??	??	No
18	[8]	10	K-R-S	??						OM	??	??	??	??	No
19	[58]	10	S	xx	ma					OM	none	RDT	S	No	inc.
20	[68]	10	S	mg						OM	Math. Exp.	none	??	No	inc.
21	[72]	10	R	pg				pg		none	??	??	No	No	No
22	[87]	10	R-S	pa	??					none	goalModel	RDT	S	No	No
23	[42]	09	R	xx						none	Text	??	??	No	No
24	[43]	09	?	??						OM	??	??	??	No	No
25	[6]	08	K-S	xx						OM	??	??	??	No	No
26	[18]	08	R-S	pg	??					none	i* model	RDT	??	??	No
27	[73]	08	K-S	xx						none	none	Excel .cvs	S	No	No
28	[52]	07	R-S	pg	pg					none	MD-UML	CWM	S	No	No
29	[75]	07	R	xx	xx	xx				none	Text	Report	n/a	?	No
30	[80]	07	R-S	pg						none	??	??	S	No	Yes
31	[44]	06	R-S	xx	pg	xx				none	UML	UML	??	No	No
32	[46]	06	R-S	xx						none	MultiDimERM	??	??	??	No
33	[65]	06	R	pg	pg					none	UML	??	??	No	inc.
34	[77]	06	S	pg						??	none	??	S	No	No
35	[28]	04	S	ma						none	??	RDT	S	No	Yes
36	[1]	03	S	pg		pg				none	??	RDT	S	Yes	No
37	[85]	03	S	xx	xx					none	none	RDT	??	No	No
38	[63]	02	R-S	pg						none	MDX	ERM	S	No	Yes
39	[33]	01	S	??	??					??	??	RDT	Yes	No	No
40	[26]	00	R	xx						none	??	ERM	S	No	No

#	#P	Year	CDM	LDM	PDM	TDM	DW type	Case	Data set	Nb. S.	Nb. Rel.	Nb. Att.	Nb. Tup.
1	[32]	14	No	Tagged Graph	No	No	DM	Yes	LEARN-SQL	3	16	?	?
2	[35]	14	OM	RDT	OBDW	No	DM	Yes	LUBM	n/a	n/a	n/a	n/a
3	[38]	14	No	Data Vault	SQL	No	Data Vault	Yes	Ad Hoc	1	8	44	>1M
4	[48]	14	MD-UML	No	No	No	DM	Yes	Ad Hoc	1	>100	??	??
5	[55]	14	??	No	No	Ad Hoc	DM	Yes	Ad Hoc	??	??	??	??
6	[61]	14	TAG	RDT	SQL	Ad Hoc	RDT	Yes	Ad Hoc	1	4	7	??
7	[12]	13	MD-UML	No	No	No	DM	Yes	Ad Hoc	??	??	??	??
8	[23]	13	DFM	No	No	No	DM	Yes	Ad Hoc	4	??	??	??
9	[83]	13	OM	Star	No	No	DM	Yes	EU-Car	n/a	n/a	n/a	n/a
10	[56]	12	ERM	RDT	SQL	No	DM	RC	n/a	??	??	??	??
11	[84]	12	exDFM	RDT	No	No	DM	Yes	Ad Hoc	1	8	28	??
12	[36]	12	OM	Star	OBDW	No	DM	Yes	EU-Car	n/a	n/a	n/a	n/a
13	[59]	12	No	No	RDF	No	DM	RC	n/a	1	??	??	??
14	[24]	11	OM	No	RDF	Ad Hoc	EAV	Yes	Ad Hoc	1	??	??	>5000
15	[71]	11	??	No	No	No	DM	Yes	TPC-DS	??	28	??	??
16	[86]	11	No	Star	No	No	DM	Yes	Ad Hoc	1	16	??	??
17	[7]	10	TempR	RDT	No	TempR	RDT	Yes	Ad Hoc	??	??	??	??
18	[8]	10	??	??	SQL	Yes	RDT	RC	n/a	??	??	??	??
19	[58]	10	ME/R	No	No	No	DM	Yes	Ad Hoc	1	7	41	??
20	[68]	10	No	Constellation	No	No	DM	Yes	EU-Car	n/a	n/a	n/a	n/a
21	[72]	10	Anchor	No	SQL	TRM	Rel-Anchor	Yes	Ad Hoc	1	25	25	1M
22	[87]	10	No	Star	SQL	No	DM	No	n/a	n/a	n/a	n/a	n/a
23	[42]	09	ERM	??	No	No	RDT	RC	n/a	??	??	??	??
24	[43]	09	??	UML-Class	RDF	No	??	inc.	Ad Hoc	??	??	??	??
25	[6]	08	??	No	No	No	??	RC	n/a	??	??	??	??
26	[18]	08	DFM	No	No	No	DM	RC	n/a	??	??	??	??
27	[73]	08	??	??	SQL	No	RDT	inc.	Ad Hoc	??	??	??	??
28	[52]	07	??	RDT	No	No	DM	inc.	Ad Hoc	??	6	22	??
29	[75]	07	??	??	??	No	??	RC	n/a	??	??	??	??
30	[80]	07	No	Star	No	No	DM	inc.	Ad Hoc	??	??	??	??
31	[44]	06	MD-UML	No	UML	No	DM	Yes	Ad Hoc	1	??	??	??
32	[46]	06	ERM	No	No	Ad Hoc	DM	Yes	Ad Hoc	1	3	15	??
33	[65]	06	UML	UMD	MOLAP	No	DM	Yes	Ad Hoc	1	20	28	??
34	[77]	06	DFM	No	No	No	DM	Yes	Ad Hoc	1	7	16	??
35	[28]	04	No	Snowflake	No	No	DM	No	Ad Hoc	n/a	n/a	n/a	n/a
36	[1]	03	No	No	??	Ad Hoc	RDT	No	n/a	n/a	n/a	n/a	n/a
37	[85]	03	??	??	SQL	No	RDT	RC	n/a	13	600	?	>32M
38	[63]	02	ME/R	No	No	No	DM	inc.	TPC-H	??	??	??	??
39	[33]	01	??	??	??	No	??	No	n/a	n/a	n/a	n/a	n/a
40	[26]	00	??	??	??	No	DM	inc.	Ad Hoc	1	6	18	??

Caption

Annotations	
n/a	non applicable
pg	partially automated with guidance required;
pa	partially automated (with or without parameterization);
xx	not significantly automated;
mg	mostly automated with guidance required;
ma	mostly automated (with or without parameterization);
??	information not explicit
inc.	incomplete description
fg	fully automated with guidance required;
fa	fully automated (with or without parameterization; guidance may be available but not required);

Crietria Acronyms	
P#	Paper number
D. App.	Design Approach
P-CRE	Creation process
P-MAP	Mapping process
P-ETL	Extract-Load-Transform process
P-KEV	Knowledge evolution process
P-REV	Requirement evolution process
P-SEV	Source evoluation process
K. Rep.	Knowledge representation
R. Rep	Requirement representation
S. Rep.	Source representation
S. Ana.	Source analysis
Multi. S.	Multiple source
Algo.	Algorithme
CDM	Conceptual design model
LDM	Logical design model
PDM	Physical design model
TDM	Temporal design model
DW type	Data warehouse type
Case	Case study
Nb. S.	Case study Source count
Nb. Rel.	Case study Relation count
Nb. Att.	Case study Attribute count
Nb. Tup.	Case study Tuple count

Value Acronyms	
BMM	Business Motivation Model
CWM	Common Warehouse Metamode
DFD	Data Flow Diagram
DL exp	Description Logic expression
DFM	Dimensional Fact Model
DM	Dimensional Model
EAV	Entity-Attribut-Value
ERM	Entity-Relationship Model
GEM	Generating ETL and Multidimensional designs
ME/R	Multidimensional Entity-Relationship
MDX	Multidimensional Expressions (MDX queries)
OCL	Object Constraint Language (OMG)
ODMG	ODMG object data model
OBDW	OM based date warehouse
OBDB	OM-based Database
OM	Ontology Model
RDT	Relational Design Theory
TAG	Temporal Attribut Graph
TMD	Temporal Multidimentional Model
TRM	Temporal Relational Model
MD-UML	UML Profile for multidimensional modeling [Luján-Mora et al. 2006]
UMD	Unified Multidimensional Model