Requirements Specification and Conceptual Modeling for Spatial Data Warehouses*

E. Malinowski** and E. Zimányi

Department of Informatics and Networks
Université Libre de Bruxelles
emalinow@ulb.ac.be, ezimanyi@ulb.ac.be

Abstract. Development of a spatial data warehouse (SDW) is a complex task, which if assisted with the methodological framework could facilitate its realization. In particular, the requirements specification phase, being one of the earliest steps of system development, should attract attention since it may entail significant problems if faulty or incomplete. However, a lack of methodology for the SDW design and the presence of two actors in specifying data requirements, i.e., users and source systems, complicates more the development process. In this paper, we propose three different approaches for requirements specifications that lead to the creation of conceptual schemas for SDW applications.

1 Introduction

The conventional DWs are designed based on the multidimensional view of data. It consists of fact and dimension tables. A fact table contains numeric data called measures while dimension tables include attributes that allow to explore measures from different perspectives. These attributes can form a hierarchy allowing to see measures at different levels of detail.

Since it is estimated that about 80% of data stored in databases (DBs) has a spatial or location component, the location dimension has been widely integrated in DW systems. Nevertheless, this dimension is usually represented in an alphanumeric, non-cartographic manner (i.e., using solely the place name) since these systems are neither able to store nor to manipulate spatial data.

On the other hand, spatial databases (SDBs) have been used for several decades for storing and managing spatial data. Therefore, bringing together DWs and SDBs, leading to spatial DWs (SDWs), allows to keep the intrinsic concepts of a DW and additionally provide support for managing spatial data.

However, SDWs as a new research field raise several issues [11, 12, 13]. For example, even though, in conventional DWs the advantages of using a multidimensional model for expressing users' requirements are well known, in SDWs this model is seldom used. Therefore, in order to exploit these advantages for

* The work of E. Malinowski was funded by a scholarship of the Cooperation Department of the Université Libre de Bruxelles.

** Currently on leave from the Universidad de Costa Rica.

R. Meersman, Z. Tari, P. Herrero et al. (Eds.): OTM Workshops 2006, LNCS 4278, pp. 1616–1625, 2006.

SDWs, in [12, 13], we propose a conceptual multidimensional model that allows to include spatial support in different elements of a DW.

On the other hand, since there is still a lack of methodology for the DW design, SDW implementers incur to problems not only related to the DW design but also to the inclusion of spatial data in DWs. In particular, requirements specification is a difficult task since it must consider not only users' requirements but also data in source systems that are used to feed SDWs.

In [11] we propose a methodology for the DW design that is in line with the traditional DB methodology, i.e., it includes the requirements specification, conceptual, logical, and physical design phases. Considering users, source systems, and both, we extend this methodology by three different approaches for requirements specifications. In this paper, based on the methodology described in [11], we propose different approaches for the requirements specification and conceptual modeling phases that allow to include spatial support for different elements of multidimensional models.

This paper is organized as follows. Section 2 refers to works related to requirements specifications in conventional DWs. Section 3 describes spatial support that may be included in multidimensional models. Section 4 presents our proposal for different approaches for the requirements specification phase and the creation of conceptual schemas for SDWs. Section 5 concludes this paper.

2 Related Work

To our knowledge, there are not works related to the methodology for the SDW design. For the conventional DWs, several approaches exist for requirements specification and conceptual modeling. In [11] we classify them in order to make them easier to understand. Next, we briefly refer to this classification.

User-driven approach. This approach considers that users play a fundamental role during the requirements analysis and must be actively involved in the elucidation of relevant facts and dimensions [5, 10]. Users from different levels of organization are selected. Then, different techniques, such as interviews or facilitated sessions are used to specify the information requirements [5].

Business-driven approach[1]**.** This approach considers that the derivation of DW structures should start from analysis of either business requirements or business processes [2, 6, 9]. Business requirements specification provides a description of users' needs considering business goals, thus starting from the highest level of the organization. Then, users from lower organization levels may participate; their requirements are aligned with the previously-established business goals. The process of refining business goals is conduced until identifying the necessary multidimensional elements.

On the other hand, the analysis of business processes requires to specify different business services or activities that ensure to produce a particular output. Since different elements participate in these activities, they may be considered

[1] It is also called process-, goal-, or requirements-driven.

as dimensions. Further, decision makers need metrics to evaluate business activities, which may be considered as measures in the DW schema.

Data-driven approach. In order to obtain the DW schema, the underlying source systems are analyzed [1, 6, 7, 14]. These source schemas should exhibit a good degree of normalization [6] to facilitate the extraction of facts, measures, dimensions, and hierarchies. In general, the participation of users is not explicitly required [8]; however, in some techniques users should either analyze the obtained schema to confirm the correctness of the derived structures [1] or identify facts and measures as a starting point for the design of multidimensional schemas [7, 14]. After schema creation, users can specify their information requirements by selecting items of interest.

Demand/supply-driven approach[2]. This approach is the combination of business- or user-driven and data-driven approaches [3]. Demand indicates business or user data requirements while supply refers to the availability of data in source systems. In the ideal situation these two parts should be equal, i.e., all information that users (business) require for analysis purposes should be supplied by the data included in source systems.

3 Spatial Support for Elements of Multidimensional Models

Requirements specification determines, among others, what data should be available and how it is organized. This specification for DWs should lead to discover the essential elements of the multidimensional model, i.e., facts with associated measures, dimensions, and hierarchies [2, 4, 14], which are required to facilitate future data manipulations.

Similar approach should be applied for SDWs. However, it is necessary to know whether multidimensional models can be used for representing spatial data. In [12, 13] we proposed a spatial extension for the conceptual multidimensional model called MultiDimER. To describe our model, we use an example for the analysis of highway maintenance costs as shown in Figure 1[3]. To better understand the constructs of the MutiDimER model, we first ignore spatial support, i.e., the symbols of different geometries and topological relationships.

The schema in Figure 1 contains dimensions, hierarchies, a fact relationship, and measures. A dimension is an abstract concept for grouping data that shares a common semantic meaning within the domain being modeled. It represents either a level or one or more hierarchies. Levels correspond to entity types in the ER model. Hierarchies contain several related levels. They can express different structures according to an analysis criterion, e.g., geographical location.

Figure 1 includes Road Coating and Time as the one-level dimensions. The County[4] and Highway segment dimensions contain hierarchies. The Geo location

[2] It is also called top-down/bottom-up analysis.

[3] A formal definition of the model can be found in [11].

[4] We call a dimension using the name of the level that is attached to the fact relationship.

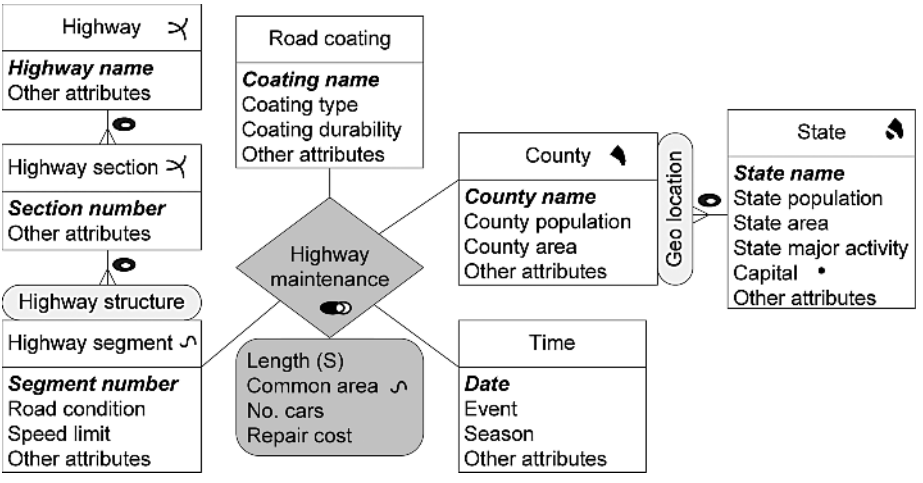

Fig. 1. An example of a multidimensional schema with spatial elements

hierarchy comprises the County and State levels while the Highway structure hierarchy includes the Highway segment, Highway section, and Highway levels.

A fact relationship, e.g., Highway maintenance in the figure, represents an n-ary relationship between dimensions. It may contain numeric measures that are used for aggregations, e.g., No.cars or Repair cost.

The spatially-extended MultiDimER model allows to include spatial support for levels, levels' attributes, fact relationships, and measures. Spatial levels are levels, for which the application needs to keep their spatial characteristics. This is captured by its geometry, which is represented using pictograms indicating spatial data types such as point, line, surface, or a collection of these data types. A level may have spatial attributes independently of the fact that it is spatial or not. In Figure 1 the spatial level State contains a spatial attribute Capital.

Two consecutive spatial levels forming a hierarchy are related through topological relationships. The latter is required for determining the complexity of procedures for measure aggregation [13]. Figure 1 shows two spatial levels, County and State, related through the intersect topological relationship ( ).

A spatial fact relationship relates two or more spatial dimensions. It requires the inclusion of spatial predicate for the spatial join operations, e.g., in the figure an intersection topological relationship; it indicates that users require to focus their analysis on those highway segments that intersect counties.

The (spatial) fact relationship may include spatial measures. They can be represented by geometry or calculated using spatial operators, such as distance, area, etc. To indicate that measure is calculated using spatial operators, we use the symbol (S). The schema in Figure 1 contains two spatial measures: Length and Common area. Length is a number representing the length of the part of a highway segment that belongs to a county. Common area is a spatial data representing the geometry of the common part.

4 Requirements Specification Leading to Conceptual Design

The design methodology for conventional DWs proposed in [11] provides three different approaches for the requirements specification that lead to the creation of conceptual schemas. These approaches rely on users' (or business) analysis needs, source data, or both. Since SDWs can be considered as DWs with spatial support in different elements composing a multidimensional schema [12, 13], the proposed methodology in [11] can be applied for the SDW design. However, two aspects should be taken into account: whether the users are able to express their requirements regarding spatial support and whether spatial data is included in the source systems.

Next, we present our proposal for SDW design referring to the requirements specification and conceptual modeling phases. We include three different approaches. For each of them, we present the sequence of phases without indicating different iterations that may occur between them.

4.1 Demand-Driven Approach

In this approach business or user requirements are the driving force for developing a conceptual schema. For the SDW design, two different possibilities exist as shown in Figure 2. The upper line is used when users either are not familiar with spatial data management or have knowledge about spatial data, but they (or designers) prefer first to express their needs related to non-spatial elements and afterwards, include spatial support.

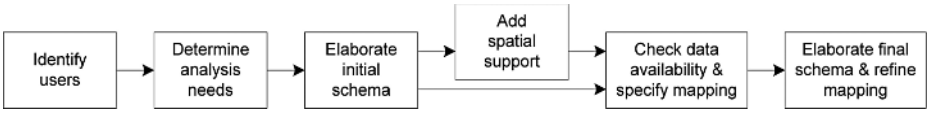

Fig. 2. General phases in the demand-driven approach for the SDW design

The first three phases are developed as for the conventional DWs. In the first phase, in order to ensure that requirements will express the organization-wide goals that a DW is expected to address, users on different management levels are identified. Executives, managers, professionals, and also an enterprise plan will help a developer team to understand the purpose of having a DW and to determine analysis needs (the second phase in the figure). For example, they may express the interest of analyzing highway maintenance cost in different counties and states considering different periods of time and road coating. The gathered information serves as basis in the elaboration of the initial DW schema (the third phase in the figure), e.g.,. of the schema as present in Figure 1 without spatial elements. During the next phase, i.e., add spatial support in Figure 2, this initial DW schema is analyzed to include spatial support.

As explained in Section 3, different multidimensional elements are examined. For one-level dimensions users may choose whether a level, its attributes, or both should be represented spatially. Then, every hierarchy level is analyzed in a similar way, e.g., users may require the spatial representation for states and capitals (Figure 1). If a hierarchy includes two or more consecutive spatial levels, e.g., County and State in Figure 1, the topological relationship between them is specified (the intersection relationship in Figure 1). If there are more than two spatial dimensions, e.g., Highway segment and County in Figure 1, designers can help the users to determine whether some topological relationships between dimensions may be of users' interest. In the affirmative case, a specific topological relationship should be included in the fact relationship to indicate a predicate for the spatial join operations (the intersection relationship in the figure). Finally, the inclusion of spatial measures is considered, e.g., Common area in Figure 1.

The next phase of checking data availability and specifying mapping determines whether data required by users is available in source systems. The mapping includes a general description of the correspondences between all elements of a multidimensional schema that match with data in source systems. This description refers also to the required transformations, if necessary. For example, for the attribute Road condition in the Highway segment level (Figure 1) this description could indicate the name of corresponding attribute in operational DBs and the transformation of numeric values to character representation, e.g., that 5 indicates very good, 4 is good, etc. This is necessary before the development of logical and physical schemas. Since spatial data may not be present in the source systems, users may require the access to external sources. Notice that adding spatial support may require additional iterations to include new users, to precise requirements needs, etc.

In the case of lacking some data items in operational DBs or external sources, a modification of schema should be made (the last phase in Figure 2). Modifications to the schema may lead to changes in the mappings.

The lower line in Figure 2 refers to the situation when the users are familiar with concepts related to spatial data. All phases, except adding spatial support, are the same as the ones described above; however, since from the beginning of the requirements gathering process the users are able to express the analysis needs referring them to spatial data, the elaborated initial schema already includes spatial elements.

4.2 Supply-Driven Approach

This approach relies on the data in source systems. It aims at identifying all candidate multidimensional schemas that can be realistically implemented on the top of the available operational DBs.

Similar to the previous demand-driven approach, we refer to two different situations. Since operational DBs are the driving force for this approach, we consider whether these DBs are spatial.

If spatial data is not included in the source systems, the first four phases of the supply-driven approach are the same as specified for the conventional DWs

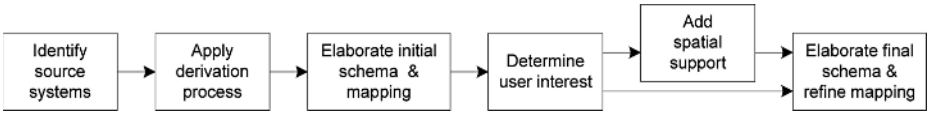

Fig. 3. General phases in the supply-driven approach for the SDW design

[11]. The identification of source systems (the first phase in the figure) aims in determining existing operational systems that can serve as a data provider for a DW. The external sources are not considered in this stage. These DBs are analyzed in an exhaustive manner to discover the elements of multidimensional schemas (the second phase in the figure). For conventional DW design, different techniques can be used [1, 3, 4, 7, 14]. All these techniques require that operational DBs are represented using the ER model or relational tables.

In general, in the first step the facts and measures are determined. This can be done analyzing the existing documentation [1, 4, 14] or the DB structures [7]. Facts and measures are elements that correspond to events occurring dynamically in the organization, i.e., that are frequently updated, e.g., an attribute indicating a highway repair cost in different periods of time. An alternative option may be the inclusion of users that understand the operational systems and can help to determine which data can be considered as measures.

Different procedures can be applied for deriving dimensions and hierarchies. They can be automatic [3, 7], semi-automatic [1], or manual [4, 14]. The process of discovering a one-level dimension or a leaf level of a hierarchy usually starts from identifying in operational DBs the static (not frequently updatable) elements (e.g., an element that corresponds to the County level in Figure 1) that are related to the facts. Then, starting with this element every one-to-many relationship is revised to find other hierarchy levels (e.g., the State level in Figure 1).

Since facts with measures, dimensions, and hierarchies are already specified, the elaboration of an initial DW schema is straightforward (the third phase in Figure 3). In this phase the specification of mappings between source systems and the proposed schema should be also elaborated.

Until now the participation of the users was minimal responding only to the specific designer's inquiries. In the next phase, i.e., determine user interest in Figure 3, a user input is required in identifying which facts are important since the initial schema may contain more elements than those required by the users.

After determining users' interest related to the conventional DW schema, a new phase of adding the spatial support is realized. Notice that this support is only considered for the previously-chosen elements of the multidimensional schema. The analysis which elements should be spatially represented can be conduced in a similar way as explained for the demand-driven approach above.

Users' recommendations about changes will be reflected in the final schema (the last phase in the figure). Since spatial support does not form a part of the underlying operational systems, external sources should be considered to deliver required spatial data. The modifications in the schema and new data sources may require the changes in mappings.

In another situation when source systems include spatial data, the phases as for the conventional DW design can be used (the lower line in Figure 3). However, a special derivation process should be applied to create an initial schema with spatial elements. Currently this derivation process should be conduced manually, since to our knowledge semi-automatic or automatic procedures for SDWs as the ones developed for conventional DWs do not yet exist.

The phases indicated by the upper line in Figure 3 can also be used when the source systems include spatial data but the derivation process is complex.

4.3 Demand/Supply-Driven Approach

This approach combines both previously-described approaches that may be used in parallel. Therefore, two chains of activities can be distinguished. The first one corresponds to the demand-driven approach and creates a multidimensional schema as it emerges from business requirements. Another chain corresponds to the supply-driven approach and delivers a multidimensional schema that can be extracted from the existing operational DBs.

Similar to the previous approaches, we propose two different solutions considering whether source systems include spatial data and whether users are familiar with the concepts related to spatial data.

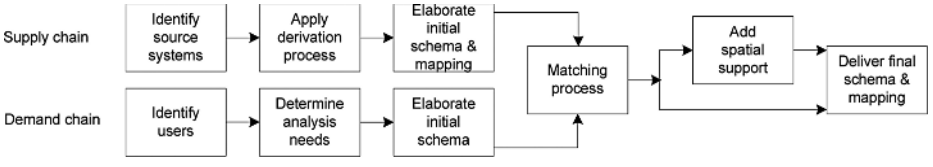

Fig. 4. General phases in the demand/supply-driven approach for the SDW design

If source systems do not include spatial data or users are not familiar with the concepts related to them, all steps until the matching process of schemas from demand and supply chains are the same as for conventional DWs already explained for the demand-driven and the supply-driven approaches above.

After the initial schemas are elaborated using both approaches, the comparison between them is realized. The comparison or integration process is not an easy task. Different aspects should be considered, such as used terminology, degree of similarity between the two solutions for each multidimensional element, e.g., between dimensions, between dimension attributes, or between hierarchy levels. Several solutions already exist for conventional DWs, e.g., [3, 6, 5].

During the matching phase user demands may be covered by data in operational systems and there may be no other data to expand the analysis spectrum, i.e., both schemas cover the same aspects of analysis. This is the ideal situation. Nevertheless, in real-world applications it is difficult to find that both schemas will cover the same analysis aspects. Indeed, the matching process can reveal that either business demands exceed the data availability or operational DBs

provide more analysis scenarios that users did not consider before. In both situations, some actions must be taken to determine the direction of changes in one of the schemas. For example, another iteration in the demand and the supply chains might be required. In this iteration either new users could be involved who are interested in the new solutions provided by source systems or a new initial schema could be elaborated eliminating from the analysis some fact relationships and associated dimensions.

In the next phase, the resulting multidimensional schema is analyzed for inclusion of spatial support (the upper line in the figure). Notice that similar to the previous approaches, external sources may be considered in this stage for obtaining spatial data. Modification to the initial schema leads to elaboration of final schema and the changes in mappings, if required.

If source systems include spatial data and users have knowledge about it, the first three phases are realized as explained above for the demand-driven and the supply-driven approaches considering the spatial support from the first phase in both chains. Then, the matching process must also refer to spatial data that is included in both schemas, i.e., obtained from the demand and the supply chains. If the result of this matching process is satisfactory, the final schema is delivered (the lower line in Figure 4). In other case, additional iterations as explained above may be necessary.

5 Conclusions

In this paper we refer to two phases of the design methodology for SDWs: requirements specification and conceptual modeling. First, we presented the MultiDimER model that allows the conceptual representation of multidimensional data with spatial support. Then, we proposed three different approaches for requirements specification that lead to the creation of conceptual schemas. These approaches take into account whether the data requirements for SDW application are based on users' specification, available data in source systems, or both. For each approach we also considered the situation whether users have knowledge about spatial data or whether spatial data is included in source systems.

Proposed approaches provide different options for implementers during the first phases of the SDW development. They can choose an approach that fits better according to users' needs and particularities of the SDW project.

References

1. M. Böehnlein and A. U. vom Ende. Deriving initial data warehouses structures from the conceptual data models of the underlying operational information systems. In *Proc. of the 2nd ACM Int. Workshop on Data Warehousing and OLAP*, pages 15–21, 1999.
2. M. Böehnlein and A. U. vom Ende. Business process oriented development of data warehouse structures. In *Proc. of Data Warehousing 2000, Physica-Verlag*, pages 3–22, 2000.

3. A. Bonifati, F. Cattaneo, S. Ceri, A. Fuggetta, and S. Paraboschi. Designing data marts for data warehouses. *ACM Transactions on Software Engineering and Methodology*, 10(4):452–483, 2001.
4. L. Cabbibo and R. Torlone. The design and development of a logical system for OLAP. In *Proc. the 2nd Int. Conf. on Data Warehousing and Knowledge Discovery*, pages 1–10, 2000.
5. G. Freitas, A. Laender, and M. Campos. MD2 - getting users involved in the development of data warehouse application. In *Proc. of the 4th Int. Workshop on Design and Management of Data Warehouses*, pages 3–12, 2002.
6. P. Giorgini, S. Rizzi, and M. Garzetti. Goal-oriented requirements analysis for data warehouse desing. In *Proc. of the 8th ACM Int. Workshop on Data Warehousing and OLAP*, pages 47–56, 2005.
7. M. Golfarelli, D. Maio, and S. Rizzi. Conceptual design of data warehouses from E/R schemes. In *Proc. of the 31st Hawaii Int. Conf. on System Sciences*, page 334, 1998.
8. B. List, R. Bruckner, K. Machaczek, and J. Shiefer. Comparison of data warehouse development methodologies. case study of the process warehouse. In *Proc. of the 13th Int. Conf. on Database and Expert Systems*, pages 6–1–6–11, 2002.
9. B. List, J. Shiefer, and A. Tjoa. Process-oriented requirement analysis supporting the data warehouse design process - a use case driven approach. In *Proc. of the 11th Int. Conf. on Database and Expert Systems*, pages 593–603, 2000.
10. S. Luján-Mora and J. Trujillo. A comprehensive method for data warehouse design. In *Proc. of the 5th Int. Workshop on Design and Management of Data Warehouses*, 2003.
11. E. Malinowski. *Designing Conventional, Spatial and Temporal Data Warehouses: Concepts and Methodological Framework*. PhD thesis, Université Libre de Bruxelles, 2006.
12. E. Malinowski and E. Zimányi. Representing spatiality in a conceptual multidimensional model. In *Proc. of the 12th ACM Symposium on Advances in Geographic Information Systems*, pages 12–21, 2004.
13. E. Malinowski and E. Zimányi. Spatial hierarchies and topological relationships in the Spatial MultiDimER model. In *Proc. of the 22nd British Nat. Conf. on Databases*, pages 17–28, 2005.
14. D. Moody and M. Kortink. From enterprise models to dimensional models: a methodology for data warehouse and data mart design. In *Proc. of the Int. Workshop on Design and Management of Data Warehouses*, page 5, 2000.