A New Methodology for Elicitation of Data Warehouse Requirements based on the Pivot Table Formalism

Sandro Bimonte*, Amir Sakka*,**, Lucile Sautot***,

* IRSTEA, UR TSCF, 9 Av. B. Pascal, 63178, Aubière, France
amir.sakka@irstea.fr, sandro.bimonte@irstea.fr
** Université Paul Sabatier, IRIT, Toulouse, France
guy.camilleri@irit.fr, pascale.zarate@irit.fr
*** AgroParistech, UMR TETIS, 500 rue Breton, Montpellier, France
lucile.sautot@agroparistech.fr

Abstract. Data Warehouses (DWs) are conceived according to data sources and users requirements. Therefore, the more the DW model reflects stakeholders' needs, the more the stakeholders will make use of their data. Therefore, in literature particular attention has been provided to DW requirement elicitation, specification and validation processes. However, most of these approaches are based on the interviews and complex formalisms that cannot be used with unskilled OLAP decision-makers. Therefore, we propose a new elicitation methodology based on the pivot table formalism, since it is well-known and used by decision-makers. We validate our methodology using a real case study.

1 Introduction

Decision Support Systems (DSSs) are flexible and interactive information systems that help decision-makers to extract useful information for identifying and solving problems and make decisions. Among DSSs, Data Warehouse (DW) and OLAP systems are probably ones of the most used in academic and industry communities. A DW is a subject-oriented, integrated, time-variant and non-volatile collection of data to support the decision-making process [Kimball et al. (2015)]. Warehoused data are analyzed using OLAP systems enabling online exploration of data stored according to the multidimensional model. Warehoused data are represented according to analysis different axes (dimensions) and facts. Dimensions are organized in hierarchies composed of levels. Facts represent the analysis subjects, and they are described by numerical measures. Measures are aggregated along dimensions hierarchies using aggregation functions (e.g. sum, min, max, etc.). Since DWs are conceived according to data sources and users requirements, the more the DW model reflects decision-makers' needs, the more the decision-makers will make use of their data. Therefore, in literature particular attention has been provided to DW requirement elicitation, specification and validation processes [Prakash and Prakash (2018)]. Requirements elicitation is the practice of collecting the requirements of a system from users, customers and other stakeholders. Requirements elicitation is non-trivial. Requirements elicitation practices include interviews, questionnaires, user observation, workshops,brainstorming,use cases, role playing and prototyping [Pohl (2010)].

Before requirements can be analyzed, modeled, or specified they must be gathered through an elicitation process. Requirements elicitation is a part of the requirements engineering process, usually followed by analysis and specification of the requirements. In the DW context, on one hand, several works investigate the specification, and their validation, of DW requirements using natural languages, formal models etc. [Romero and Abelló (2009)]. These models are commonly too much complex to be expressed directly by decision-makers without DW skills during the elicitation process. On the other hand, only few works investigate DW requirements elicitation [Prakash and Prakash (2018)]. Most of them uses classical requirements engineering methods (i.e. interview, questionnaire, etc.), and no work provide an ad-hoc methodology for DW requirements elicitation. Therefore, motivated by the lack of a well-defined approach for DW requirement elicitation destined to DW unskilled decision-makers, we propose in this work a new methodology based on the pivot table formalism and prototyping. We also present some experiments issued from a real DW project that validate our proposal. The paper is structured in this way: Section 2 presents related work, the case study is described in Section 3, our elicitation methodology is described in Section 4, and experiments are detailed in Section 5.

2 Related Work

Different approaches have been defined for the elicitation of DW requirements. A review is presented in [Nasiri et al. (2015)]. Among relevant works we cite the following ones.

Mazón et al. (2007) used i* framework, which is based on the study of distributed intentionality of stakeholders, answering the 'who' and 'why' questions to model the business goals. Authors structured the business goals (i.e. the goals that data warehouses aid to achieve) into strategic, decision and information goals. They design, using a UML profile, the conceptual model retrieved from the i* diagrams. Prakash and Gosain (2008) used the informational scenario. By focusing on the decision, the informational scenario as an elicitation mechanism has been deployed with a goal-decision-information model composed by a set of interconnected tuples <Q, R>, where Q represents the set of SQL decisional input queries and R the response of the decisional system for each. Salinesi and Gam (2006) introduced the idea of the CADWA method that proposes to anticipate decision-makers' requirements initially from the organization's business plan depending on the interests and activities of each decision-maker. Then, using a goal-based approach of requirements modeling (i.e. Map formalism) they specify the different existing intentions to be satisfied and strategies to do so. For each decision-maker, a macro business plan is defined using distribution matrix, then all together validated for consistency with the global BP. Using the Map formalism, each of the decision-makers is also supposed to define his micro BP for his specific needs. Finally they adopt a linguistic approach to express action plans (i.e. the final level of functional requirements) to allow decision-makers to evaluate and validate their requirements. Romero and Abelló (2010b) described their automatic multidimensional design from ontologies (i.e. AMDO) that, without considering decision-makers requirements, start by a full analysis of the data sources to discover the ontology concepts. Next, using defined filtering functions, users chose their required functionalities from the retrieved knowledge making the elicited requirements already conciliated with the data-sources. Finally, after finishing the elicitation step, AMDO framework create automatically the data warehouse conceptual schema.

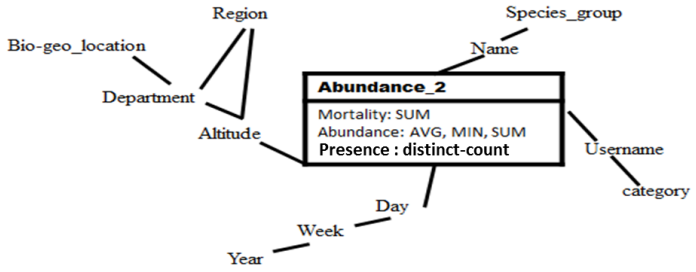


FIG. 1 – *DW model designed by a VGI4Bio volunteer.*

Most recent works [Bhardwaj and Prakash (2016); Nasiri et al. (2015)] model requirements as functions and identify parameters and outputs types. Finally, as described in [Prakash and Prakash (2018)], DW requirements works can be classified in goal oriented, scenario techniques and coupling them.

However, these approaches are all based on interviews and other complex formalisms that are not well-adapted to decision-makers that have no DW skills.

3 Case study

In the context of project VGI4Bio [1], we mobilize two volunteers databases (Visionature and Observatoire Agricole de la Biodiversité - OAB) to build DW applications to analyze farmland biodiversity indicators. Visionature and OAB have 7682 and 1500 volunteers that produce data, respectively. Among possible users interested in analyzing these data, we have identified a huge number of users belonging to diverse categories such as: (i) the volunteers themselves that are interested in analyzing data to improve their data production quality, their related daily practices, etc.; (ii) public and private organisms (DREAL, Chambre d'Agriculture, etc.). At this phase of the project, we have identified some volunteers. Figure 1 shows a multidimensional model defined by a volunteer at the end of the DW design phase. They concern the analysis of the abundance of animals. It allows answering to queries like: "What is the total abundance of birds per altitude, species and week?" (Fig. 1a).

4 DW requirements elicitation methodology

The methodology (Figure 2) is composed of different steps:

STEP 1. OLAP tutorial. This preliminary step consists of presenting to the decision-makers some existing DW applications, and explain them the main concepts of DW and OLAP. A web-based OLAP client is used to allow the decision-makers to "play"

1. https://www.vgi4bio.fr/

with the existing OLAP applications. This step is 2-3 hours long. Let us note that this tutorial only allows decision-makers to have a first idea of analysis possibilities of OLAP systems. At the end of this step, decision-makers have not a sufficient understanding of main OLAP concepts. They will incrementally acquire them in the next steps.

STEP 2. Interview. This step represents a classical interview with the decision-makers to understand their main analysis needs.

Steps 3 and 4 are executed in an iterative way until the decision-makers validate the requirements.

STEP 3. XSL & semi-structured interviews. This step allows the decision-makers to express the requirements by their own by using a pivot table defined in an Excel file. Then, DW experts using a semi-structured interview ask some questions to decision-makers about their pivot tables in order to better understand their needs, and to incite decision-makers to think about their errors and/or modifications.

STEP 4. Prototype. At this step, the DW experts implement a prototype of the DW and show it to the decision-makers. Finally, some representative pivot tables obtained with the prototype are saved as Excel files, and sent to the decision-makers.

STEP 5. Validation on data source. Validation on data source. Once a pivot table is validated by the decision-makers, the DW experts validate it on the data source using existing methodologies such as Romero and Abelló (2010a).

STEP 6. Fusion. Finally, since a pivot table represents only a view on the DW, they are merged to obtain the final requirements using the approach proposed in Nabli et al. (2005).

In the rest of the section, we detail the innovative steps of our methodology: step 3 "XSL & semi-structured interviews", and step 4 "Prototype".

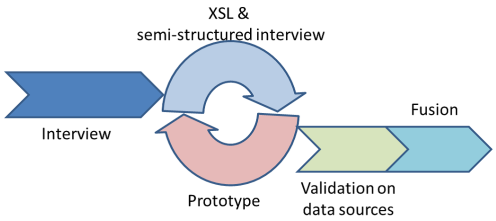


FIG. 2 – *Our pivot table based DW requirements elicitation methodology.*

4.1 XSL and semi-structured interview

As stated in the previous sections, most of DW projects' decision-makers are Information Technology and DW unskilled users. Therefore, they usually present difficulties to identify their analysis needs in terms of DW concepts (i.e. facts and dimensions), but also in terms of

species	location			date		presence
Paridae	ARA			1978	40-1978	1,00
	ARA	puy	montagne	1978	40-1978	1,00
Sturnidae	ARA			1978	3-1978	1,00
					40-1978	0,00
	ARA	puy	mer	1978	3-1978	1,00
					40-1978	0,00

FIG. 3 – *STEP 3: Example of Excel file.*

main concepts of classical requirements engineering methodologies (goals, KPI, etc.). However, as already proved in several works, they can understand the results of DW prototype implementations, which essentially consist of pivot tables issued from OLAP clients. Therefore, we adopt the pivot table as the formalism allowing the decision-makers to express their analysis needs. Then, we provide to decision-makers a simple example of pivot table defined using an Excel file that contain measures and dimension members organized into hierarchies. Thus we suggest to decision-makers to design their pivot tables by their own, and using some sample data that they know.

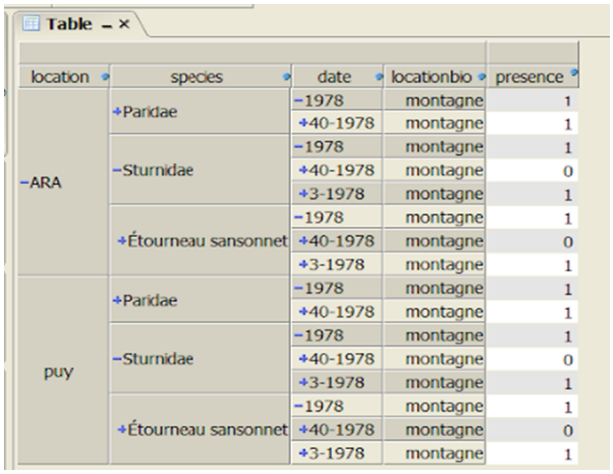
An example of pivot table defined using Excel used to the design of the multidimensional model of Figure 1 is shown in Figure 3. In this example, the decision-makers have identified one measure, aggregated with the sum, and three dimensions (species, location, and date). Figure 3 shows an intermediate interaction of the methodology, therefore it does not represent the final DW model of Figure 1. To help and guide the decision-makers to propose well-defined pivot tables, we associate a semi-structured interview to each pivot table.

This semi-structured interview is composed of different phases: (i) For each dimension, the DW experts ask to the decision-makers to validate or modify their hierarchy, (ii) Then, measures and their aggregation functions are questioned. Examples of queries used in the semi-structured interviews about the pivot table of Figure 3 are:

— "*For the column location, do you think that there is a need to have a coarser level grouping regions into biological locations?*".
This kind queries are used to help the decision-makers to identify hierarchies.
— "*Do you think that a region can have several biological locations?*".
This kind of queries are used to identify complex DW structures [Pedersen et al. (2001)] such as non-strict hierarchies.

Moreover, this exchange about the Excel file between DW experts and decision-makers is also important in the case the decision-makers have difficulties to define well-formed pivot tables. In this case DW experts help the decision-makers to correct their pivot tables.

The usage of the pivot table formalism has also another important property. As described by Nabli et al. (2005), pivot tables can be formally translated into DW models. Therefore, they represent a simple way for decision-makers to express their analysis needs that can be directly mapped into DW models, which can be easily prototyped by DW experts as described in the next section.



Table

location	species	date	locationbio	presence
−ARA	+Paridae	−1978	montagne	1
		+40-1978	montagne	1
	−Sturnidae	−1978	montagne	1
		+40-1978	montagne	0
		+3-1978	montagne	1
	+Étourneau sansonnet	−1978	montagne	1
		+40-1978	montagne	0
		+3-1978	montagne	1
puy	+Paridae	−1978	montagne	1
		+40-1978	montagne	1
	−Sturnidae	−1978	montagne	1
		+40-1978	montagne	0
		+3-1978	montagne	1
	+Étourneau sansonnet	−1978	montagne	1
		+40-1978	montagne	0
		+3-1978	montagne	1

FIG. 4 – *STEP 4: Example of prototype.*

4.2 Prototype

Prototype is one the most used requirement elicitation and validation methods in software and also in DW development. Therefore, in our methodology starting from the pivot tables defined as Excel files at the previous step, and adopting the approach of Nabli et al. (2005), we obtain DW models corresponding to the pivot tables. Then, DW experts implement these models using the ProtOLAP tool [Bimonte et al. (2013)]. ProtOLAP takes as input an UML model defined using the ICSOLAP UML profile for SOLAP, which is implemented in the CASE tool MagicDraw. It automatically creates the SQL scripts for Postgres (tables' creation and data insertion) and XML configuration Mondrian (the OLAP server) files. In other terms, pivot tables are translated into UML multidimensional models (therefore, *UML is the specification formalism of requirements*), and then a DW prototype is created. An example is shown in Figure 4. It is important note that this step has two important properties.

Firstly, it allows generating a prototype that is shown to the decision-makers.

Secondly, it allows the DW experts to validate the implementation feasibility into existing OLAP servers and DBMSs of complex DW models [Pedersen et al. (2001)], for example using complex hierarchies, facts-dimensions relationships, etc. Indeed, sometimes due to implementation issues, the pivot tables can be translated into quite different implementations. For example, to avoid non-strict problems related to the biological level of the location column of the pivot table of Figure 3, the DW experts can decide to create a new dimension for this level as shown on Figure 4.

Finally, once the prototype is implemented, it is shown to the decision-makers. If they validate them, then the step 4 is applied. Otherwise, according with the decision-makers some pivot tables issued from the OLAP client (Figure 4) are saved as Excel files, they are sent to decision-makers, and the step 2 is applied again.

5 Validation

For the validation of our methodology, we use the VGI4Bio case study.

For the validation of the step 3 (*XSL & semi-structured interviews*), we ask to decision-makers if the XSL files help them to explain and represent their analysis needs. All the 3 decision-makers involved in the experiments evaluate it as "useful". We do not compare our methodology to other approaches since in a real case study it is not possible, due to time and economic costs, provide several different implementations for the same DW.

For the step 4 (*Prototype*), we measure the duration of the implementation of the prototypes with and without our prototyping tool. The manual implementation takes in average 1 hour, since there are systematically SQL, Mondrian and MDX errors. Using ProtOLAP these errors are eliminated, and so the implementation is quasi instantaneous. Per contra, ProtOLAP needs to define a UML model that takes in average half hour. However, these conceptual models are mandatory for the management of the DW project, and thus we do not consider this time in the evaluation.

6 Conclusion

Data Warehouses are conceived according to data sources and users requirements. Therefore, the more the DW model reflects stakeholders' needs, the more the stakeholders will make use of their data. Therefore, in literature particular attention has been provided to DW requirement elicitation, specification and validation processes. However, all these approaches are based on the interviews and complex formalisms that cannot be used with unskilled DW and OLAP decision-makers. Therefore, we propose a new elicitation methodology based on the pivot table formalism, since it is well-known and used by decision-makers. We validate our methodology using a real case study. Our current work concerns the definition of some quantitative and qualitative metrics to evaluate our approach and their evaluation in the context of the VGI4Bio project. Future work consists in the implementation of the web-based user-friendly tool to replace the XSL files. This tool will allow decision-makers to sketch well-defined pivot tables without the intervention of the DW experts and so to improve the overall elicitation process.

Acknowledgment

This work is supported by the ANR-17-CE04-0012 project VGI4Bio (www.VGI4Bio.fr). We also thank the volunteers of the project for their participation to the experiments.

References

Bhardwaj, H. and N. Prakash (2016). Eliciting and structuring business indicators in data warehouse requirements engineering. *Expert Systems 33*(4), 405–413.

Bimonte, S., E. Edoh-alove, H. Nazih, M.-A. Kang, and S. Rizzi (2013). Protolap: Rapid olap prototyping with on-demand data supply. In *Proceedings of the Sixteenth International*

Workshop on Data Warehousing and OLAP, DOLAP '13, New York, NY, USA, pp. 61–66. ACM.

Kimball, R., M. Ross, J. Mundy, and W. Thornthwaite (2015). *The kimball group reader: Relentlessly practical tools for data warehousing and business intelligence remastered collection*. John Wiley & Sons.

Mazón, J.-N., J. Pardillo, and J. Trujillo (2007). A model-driven goal-oriented requirement engineering approach for data warehouses. In *International Conference on Conceptual Modeling*, pp. 255–264. Springer.

Nabli, A., J. Feki, and F. Gargouri (2005). Automatic construction of multidimensional schema from olap requirements. In *The 3rd ACS/IEEE International Conference onComputer Systems and Applications, 2005.*, pp. 28–.

Nasiri, A., E. Zimányi, and R. Wrembel (2015). Requirements engineering for data warehouses. In *EDA*, pp. 49–64.

Pedersen, T. B., C. S. Jensen, and C. E. Dyreson (2001). A foundation for capturing and querying complex multidimensional data. *Information Systems 26*(5), 383–423.

Pohl, K. (2010). *Requirements engineering: fundamentals, principles, and techniques*. Springer Publishing Company, Incorporated.

Prakash, N. and A. Gosain (2008). An approach to engineering the requirements of data warehouses. *Requirements Engineering 13*(1), 49–72.

Prakash, N. and D. Prakash (2018). *Data Warehouse Requirements Engineering: A Decision Based Approach*. Springer.

Romero, O. and A. Abelló (2009). A survey of multidimensional modeling methodologies. *International Journal of Data Warehousing and Mining 5*(2), 1.

Romero, O. and A. Abelló (2010a). Automatic validation of requirements to support multidimensional design. *Data & Knowledge Engineering 69*(9), 917–942.

Romero, O. and A. Abelló (2010b). A framework for multidimensional design of data warehouses from ontologies. *Data & Knowledge Engineering 69*(11), 1138–1157.

Salinesi, C. and I. Gam (2006). A requirement-driven approach for designing data warehouses. In *Requirements Engineering: Foundations for Software Quality (REFSQ'06)*, pp. 1.

Résumé

Les entrepôts de données sont conÃ§us en fonction des sources de données et des besoins des utilisateurs. Par conséquent, plus le modèle multidimensionnel reflète les besoins des parties prenantes, plus les parties prenantes utiliseront leurs données. Dans la littérature, une attention particulière a été accordée aux processus d'élicitation, de spécification et de validation des besoins pour la conception d'entrepôt de données. Cependant, toutes ces approches sont basées sur des entretiens et des formalismes complexes qui ne peuvent pas être utilisés avec des décideurs sans compétences spécifiques en informatique. Nous proposons donc une nouvelle méthodologie d'élicitation basée sur le formalisme de la table pivot, puisqu'elle est facilement apréhendée par les décideurs. Nous validons notre méthodologie en utilisant une véritable étude de cas.