RAMEPS: A REQUIREMENTS ANALYSIS METHOD FOR EXTRACTING-TRANSFORMATION-LOADING (ETL) PROCESSES IN DATA WAREHOUSE SYSTEMS

Azman Taa, Mohd Syazwan Abdullah, Norita Md. Norwawi
Graduate Department of Information Technology
College of Arts and Sciences, Universiti Utara Malaysia,
06010 UUM Sintok, Kedah, Malaysia
{azman, syazwan, norita}@uum.edu.my

Abstract: - The data warehouse (DW) systems design involves several tasks such as defining the DW schemas and the ETL processes specifications, and these have been extensively studied and practiced for many years. The problems in heterogeneous data integration are still far from being resolved due to the complexity of ETL processes and the fundamental problems of data conflicts in information sharing environments. Current approaches that are based on existing software requirement methods still have limitations on translating the business semantics for DW requirements toward the ETL processes specifications. This paper proposes the Requirement Analysis Method for ETL processes (RAMEPs) that utilize ontology with the goal-driven approach in analyzing the requirements of ETL processes. A case study of student affairs domain is used to illustrate how the method can be implemented.

Key-Words: - Requirement Analysis, ETL Processes, Data Warehouse, Ontology, Business Intelligence

1 Introduction

DW is a system for gathering, storing, processing, and providing a huge amount of data with analytical tools to present complex and meaningful information for decision makers. These data are collected, stored, and accessed in centralized databases in order to sustain competitiveness in businesses [1]. However, the DW system is dependent on the ETL to provide the data [2]. In other words, the success of DW system is dependent on the design of ETL processes. There are remaining issues in requirement, modeling, and designing the ETL processes due to the non-standardization of methods imposed by the providers through their own DW tools. Moreover, the design tasks need to tackle the complexity of ETL processes from early phases of DW system development. An early phase is important to ensure the satisfaction of information for the DW systems [3].

The complexity of ETL processes always refers to the problem of generating the transformations for data sources toward the DW structure. These transformations involve the reconciliation semantic of user terms and data source schemas [4]. Generally, an ambiguous definition of user requirements occurs because the users are unable to define their requirements precisely and clearly [1]. Moreover, various meaning of data (i.e. attributes, tables) makes it difficult for integrating the user requirements to the data sources. Thus, reconciliation the appropriate semantic of user terms and data sources are important in generating the transformations accordingly. Generating the transformations are about designing the ETL processes from an early phase of DW system development. This should be based on the systematic method for capturing and analyzing the user requirements toward generating the ETL processes. However, this method is incomplete due to the limitations and linkages in modeling and designing the DW systems. Clearly, these limitations have contributed to the failure of DW projects [3]. Therefore, we propose the RAMEPs, a requirement analysis method based on goal-ontology approaches.

This paper is structured as follows: related work is described in the section 2. Section 3 and 4 explains our approach on RAMEPs, while section 5 discusses a case study on how RAMEPs can be used. Section 6 shows how the case study is implemented on a Jena 2 framework. Finally, section 7 concludes the work and proposes the future research direction.

2 Related Literature

The designing of ETL processes is essential for helping the developer to develop the DW system from the early phases of system development. Due to the heterogeneity problems, the tasks to manage and develop the ETL processes become difficult, tedious and complex. The emergence of ontology as the main artifacts of semantic web technology has been used in resolving the heterogeneity problems in information sharing environments [4]. The ontology has been used to reconcile the semantics within database integration, especially in DW system environments [5]. Moreover, the database schemas can be modeled as an ontology model with respect of the complexity in ontology

construction. Therefore, an effort to simplify these tasks is important through the ETL tools that support the multipurpose data integration platform together with the ontology.

Generally, software design requires unambiguous, complete, verifiable, consistency and usable user requirements that support data analysis and decision-making processes [6]. However, the work of capturing and analyzing the user requirements are not an easy task because it involves various levels of users, departments and organizations. The tasks involve analyzing the goals, resources, realities, and rules that affecting the ETL processes into one place. [3] has applied goal oriented approach in designing the DW structure without extended to the ETL processes. Meanwhile, [5] elaborated the design of ETL processes by using ontology without mentioning how the user requirements are provided. Therefore, this research will present the method which will be applied the goal-ontology approaches to design the ETL processes from early phases of DW system development.

3 Goal-Ontology for ETL Processes Requirements

Requirement analysis of ETL processes focuses on the transformation of informal statements of user requirements into a formal expression of ETL processes specifications. The informal statements are derived from the requirement of stakeholders and analyzed from the organization and decision-maker perspectives [3]. We argue an analyzing the DW requirements from the abstract of user requirements toward the detail of ETL processes are important in tackling the complexity of DW system design. This widely accepted that the early requirement analysis significantly reduces the possibility misunderstanding of user requirements [7]. The higher understanding among stakeholders possibly increases the agreeable about terms and definitions used during the ETL processes execution. Therefore, our requirement analysis method for ETL processes (RAMEPs) is centered on the organizational and decisional modeling and focuses on the transformation model from the perspective of a developer. By adapting the approach used by [3], the model of our method is presented in Figure 1.

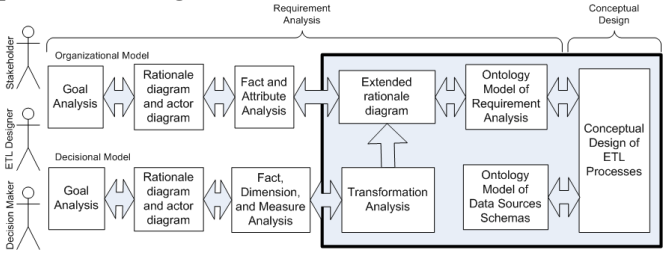

Fig. 1. The RAMEPs

Our extended works in the RAMEPs model are highlighted in the shaded area. The organizational modeling is used to identify the goals that are related to facts, and attributes in DW components. The decisional modeling is directly focused on the information needs by decision makers and related to facts, dimension, and measures. The information needed by the decision maker is provided by the transformational model that related to actions and business rules. These help the developer to generate the appropriate transformations for populating the data sources. In summary, tasks in RAMEPs are presented in Table 1.

Table 1. The RAMEPs Tasks

Steps	Activities	Methodology
1.	Gather and elicit requirements with stakeholders.	Interview, and document analysis
2.	Analyze requirements based on the organization perspective.	Tropos Goal-oriented
3.	Analyze requirements on the decision-maker perspective.	Tropos Goal-oriented
4.	Analyze requirements on the developer perspective.	Tropos Goal-oriented
5.	Ontology construction for requirement analysis.	RDF/OWL Ontology model
6.	Ontology construction for data sources.	RDF/OWL Ontology model
7.	Map and merge the requirements ontology with the data sources ontology.	RDF/OWL Ontology model
8.	Refine the structure of merging ontology and make adjustment to fully satisfy the user requirements.	RDF/OWL Ontology model
9.	Construct the required ETL processes specifications from the merging ontology.	Ontology model, Jena 2 Framework

This paper focused on the steps 5 – 9 and provides a case study to evaluate the proposed method.

4 The RAMEPs Tasks

The RAMEPs is based on the Tropos methodology that was developed from the well-accepted i* conceptual framework of software development [8]. The aim is to define the decisional information from the perspective of organizational, decision-maker, and developer. The goal oriented requirement analysis will determine the components of DW structure through diagrams. The diagrams represented in specific symbols explained their roles and activities (i.e. facts, dimensions, measures, business rules, actions). End of these analyses, the glossaries of facts, dimensions, measures, business rules, and actions will be used to proceed on the conceptual

design of ETL processes. However, these glossaries need to be mapped to the corresponding data sources. The mapping process should be based on a unify model (i.e. ontology) to reduce the uncertainty. Indeed, the heterogeneity problems should be resolved along the process take place.

4.1 Ontology for Requirements Glossaries

The DW requirements contain facts (F), dimensions (D), measures (M), business rules (Br), and Actions (Ac). This explains that the DW requirements contain Facts with the set of dimensions, set of measures, set of business rules, and set of actions. In ontology, facts (F), dimensions (D), measures (M), and actions are defined as set of classes, whereas business rules (Br) and relationships among them are defined as set of properties. The relationships referred the link between class to class, class to property, and property to property. As described in ontology definition, set of axioms used to assert subsumptions between classes are defined from the business rules and actions. The business rules specify the domain and range properties, cardinality constraints, disjointness class, and others. The actions defined a new class for aggregation functions used for each fact. Formally, the DW requirements ontology (DWRO) can be defined:

DWRO = (F, D, M, Br, Ac)
Where: F = Facts
D = Set of Dimensions (D_1, D_2, …. D_n)
M = Set of Measures (M_1, M_2, …. M_n)
Br = Set of Business Rules (Br_1, Br_2, .. Br_n)
Ac = Set of Actions (Ac_1, Ac_2, …. Ac_n)

The type of class values are not defined in DWRO because the values were not available yet at this level.

4.2 Ontology for Data Sources (DSO)

The method of semantic mapping from a relational model to RDF/OWL is adapted to facilitate the transformation of data sources into RDF/OWL based structure [9]. The tasks to transform the database to the ontology structure are known semantic reengineering of the legacy information system. These tasks are as follows:

i. Apply the reverse-engineering approach to define the conceptual model of existing data sources system through any modeling tools (e.g. PowerDesigner).
ii. Construct the ontology structure by using semantic mapping rules. The ontology tuple will consist of concepts, relations, function, axioms, and instances: O = (C, R, func, A, I).
iii. The ontology structure will be constructed by using Protégé-2000. The Protégé-2000 is used because of its ability to produce OWL/RDF automatically.

Since the data sources are heterogeneous, the basic mapping principles applied as follows:

i. One or more similar relations R_i is mapped to one related concept C_i.
ii. Primary-foreign relationship R_i is mapped to property OP_i.
iii. Tuple of a relation R_i is mapped to an instance I_i

4.3 Mapping the Requirements to the Data Sources

The need of mapping and matching the DW requirements toward the associated data sources are important in order to construct the single view of ontology. The different view of the ontology model (i.e. DWRO and DSO) in the same domain is known as heterogeneity in the ontologies [10]. Since the heterogeneity problems in data sources have been tackled via ontology representation of data sources, then the same approach has been applied in mapping and matching mechanism. However, the matching ontologies are supported from the domain knowledge of user requirements and application knowledge of existing application system.

The DWRO should be able to describe the semantics of the user requirements toward the semantics of data sources in order to establish mapping between classes and properties. Furthermore, the process of mapping is possibly implemented by the appropriate software and tools with the reasoning functionality. The DWRO was modeled the information according to the following elements:

i. The concepts of the domain
ii. The relationships between the concepts
iii. The attributes characterizing the concepts
iv. The different representation format or value for each of the attributes
v. The restriction impose by attributes or relationships

These elements can be represented in the ontology structure such as {concept ↔ classes}, {relationship ↔ properties}, {type of format or value ↔ classes in the hierarchy}, {specific element in ETL setting ↔ new classes}, and {restrictions ↔ axioms}. Based on these representations, the characteristics of DWRO and DSO can be mapped as shown in Table 2.

Table 2. DWRO and DSO Mapping

DWRO elements	DSO elements	Ontology mapping elements
Fact	Concept	Concept ↔ Fact
Dimension = (dim_1, dim_2, dim_3, … dim_n)	Table: ConceptName (tbl_1, tbl_2, … tbl_n)	Class: ConceptName ↔ dim_1, dim_2, dim_3, … dim_n
Measure = (m_1, m_2, m_3, … m_n)	Attribute: m_1 = $Action_1(attr_1, attr_2, … attr_n)$, m_2 = $Action_2(attr_1, attr_2, … attr_n)$ M_n = $Action_n(attr_1, attr_2, … attr_n)$	Property: ConceptName ↔ [m_1 = $Action_1$ ($attr_1$, $attr_2$, … $attr_n$)], [m_2 = $Action_2$ ($attr_1$, $attr_2$, … $attr_n$)], [m_n = $Action_n$ ($attr_1$, $attr_2$, … $attr_n$)
Business Rule = (br_1, br_2, br_3, … br_n)	Attribute/Relationship	Property: m_1 ↔ [$attr_1$ (br_1), $attr_2(br_2)$, … $attr_n(br_n)$], m_2 ↔ [$attr_1$ (br_1), $attr_2(br_2)$, … $attr_n(br_n)$], …
Action = (ac_1, ac_2, ac_3, … ca_n)	Behavior/Constraint	Axiom: $ac_1..ac_n$ ↔ [ConceptName ↔ m_1.. m_n]
-	Data	Instance/Individual

Based on the mapping results, new classes and properties pertaining to the merging requirement ontology (MRO) will be produced. These new classes and properties captured the knowledge of ETL processes such as aggregated, aggregation, range, table, formation, and others. Through reasoning (e.g. Pellet), the inferred MRO is semantically organized in presenting the knowledge of ETL processes [5]. Therefore, by using semantic web programming (i.e. Jena 2 Framework), the ETL processes specifications can be produced for designing the ETL processes.

5 Case Study

The RAMEPs is validated through DW-Tool for goal-oriented analysis, Protégé-2000 and Pellet reasoned for ontology model. The evaluation process is carried out in the real case study of academic domain. The Academic Affair Director (AADD) is one of the main decision makers that require the information about student registers and performances in each of an academic session. The focus goals associate to AADD (e.g. analysis student register, analysis student performance) is decomposed into sub-goals (e.g. analysis total register, analysis total unregister, analysis student excellence, analysis student examination). In transformation analysis, the relevant plans are connected to the decision goals. The plans are presented as an abstract level of ETL processes, which is implemented in the implementation phase. By using means-end and contribution analysis, the abstract of ETL can be determined. There are no activities to determine the appropriate data source schemas toward DW structure at this level. However, as the transformation analysis is carried out, the facts, dimensions, attributes, measures, and abstract processes of ETL (refer as actions) can be used to design the ETL processes as required by goal to be fulfilled. The final diagram of DW requirements is presented in Fig. 2.

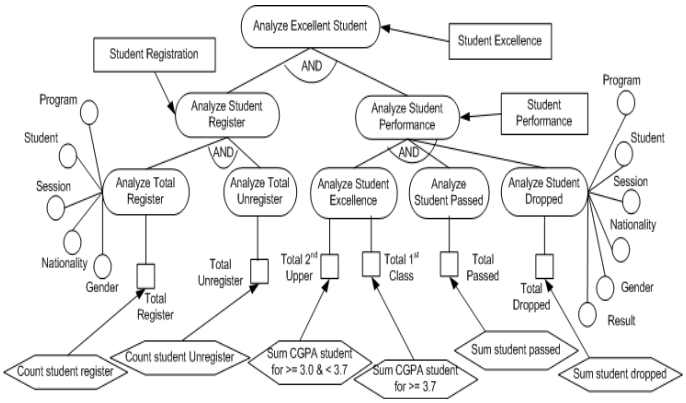

Fig. 2. Final Diagram of DW Requirements

5.1 DW Requirements Ontology

The DW requirements were modeled into ontology structure according to the final diagram of requirement analysis. By using Protégé-2000, the constructed DW requirements ontology is shows in Fig. 3.

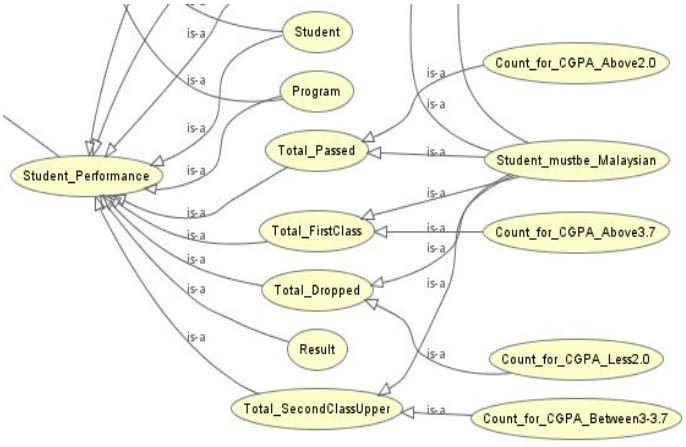

Fig. 3. The DWRO

5.2 Data Sources Ontology

The ontology model defined from two different applications that is Academic Student Information System (ASIS) and Graduate Student Information System (GAIS). The concepts of Student, Gender, Session, Program, Nationality, and Result were

introduced to reconcile the agreeable semantics among the data sources. This can be viewed in Fig. 4.

Fig. 4. The DSO

Consequently, the semantics mapping between data sources to the DW requirements can be established during the mapping process. Therefore, the semantic heterogeneity problems have been resolved prior to the generation of ETL processes specifications.

5.3 Merging Requirement Ontology

The construction of MRO is depended on the mapping between DW requirements and data sources. This involved the identification of similarity and dissimilarity of concepts and their associate attributes toward the data sources as follows:

- Concept ↔ Classes (e.g. Student Register, Student Examination)
- Relationship ↔ Properties (e.g. hasDimension, hasMeasure)
- Type of format or value ↔ Classes in the hierarchy (e.g. currency – RM, Dollar)
- Specific element in DW setting ↔ new Classes (e.g. SUM, COUNT, MERGE)
- The restriction ↔ Axioms (e.g. "Student must be Malaysian")

Based on our definition, the mapping between DW Requirement ontology and DS ontology is shown in Table 3.

Table 3. DW Requirements and DS Mapping

DWRO	**DSO**	**The mapping elements**
Fact (Student Register)	-	Concept: Student Register
Dimension (Student, Semester, Course, Sex, Race, Result)	Concept: Student (t210student, t801studmas) Concept: Gender (t012jantina, t801jantina) …	Student ↔ Concept Student Semester ↔ Concept Session Course ↔ Concept Program Sex ↔ Concept Gender …
Measure (Total student register, Total student Unregister)	- Concept: Student (t210student, t801studmas) *- Total student unregister unable to count from Student.	[Total student register] ↔ [Student record]
Business Rule (Student must be Malaysian)	Concept: Student (t210student, t801studmas), Concept: Nationality (t016warga, t016warga)	[Student must be Malaysian] ↔ [Student (t210student, t801studmas), Nationality (t016warga, t016warga)]
Action (COUNT Student Register, SUM Student passed, Student dropped, Student 1st Class, Student 2nd Class, FILTER for Student must be Malaysian)	Concept: Student (t210student, t801studmas), Concept: Nationality (t016warga, t016warga)	[COUNT for Student Register] ↔ [Student (t210student, t801studmas) [SUM for Student passed] ↔ [Result (t312result_exam, t804result) …

The mapping setting in Protégé-2000 is defined such as:

MERGE DS_1, DS_2
Classes Student : asis:t210student ∪ gais:t801studmas
Classes Gender : asis:t012jantina ∪ gais:t801jantina
Classes Session : asis:t005term ∪ gais:t005term
…

Based on mapping mechanism, the MRO is derived as shown in Fig. 5.

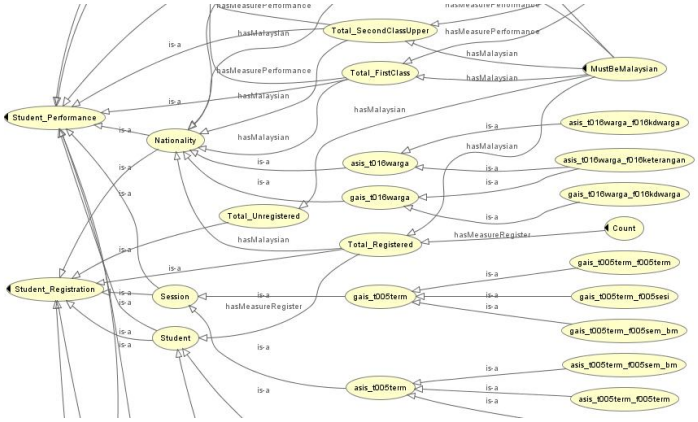

Fig. 5. The MRO

6 Implementation

To generate the ETL processes specifications, a prototype of application has been developed by using Java programming. The ontology structure as shown in Fig. 6 is manipulated through the Jena 2 framework.

```
http://www.semanticweb.org/ontologies/2009/1/GoalReq
uirementOntology.owl#Sum -->
   <owl:Class
   rdf:about="&GoalRequirementOntology;Sum">
    <rdfs:subClassOf>
    <owl:Restriction>
       <owl:onProperty
rdf:resource="&GoalRequirementOntology;hasMeasureP
erformance"/>
            <owl:someValuesFrom
rdf:resource="&GoalRequirementOntology;Total_Droppe
d"/>
       </owl:Restriction>
       </rdfs:subClassOf>
       <rdfs:subClassOf>
…
```

Fig. 6. A snippet of MRO

The ontology is representing by the RDF/OWL language. By using an appropriate algorithm (e.g. as proposed by [5]), the ETL processes specifications can be generated. A part of the results from the prototype application is shown in Fig. 7.

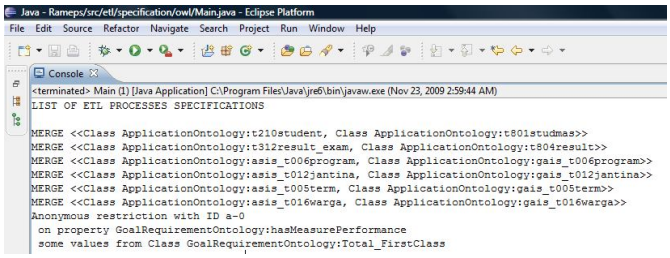

Fig. 7. List of ETL Processes Specifications

7 Conclusion

The RAMEPs has proven the ETL processes specifications can be derived from the early phases of DW system development. The methodology used in analyzing the user requirements has been validated by DW-Tool and Protégé-2000 successfully. Indeed, current work is pressing on the evaluation of the proposed RAMEPs. The evaluation approach is carried out by implement the RAMEPs into various domains of case studies. This will gives the multi views of information in DW systems. Further works will be completing the application prototype and finalize the validation and evaluation process. We believe the adoption of our method can help developers to clearly define the ETL processes prior to the detail design of DW systems. The RDF/OWL language is easy to define and maintain makes the design of ETL processes specifications can be managed easily even the changes in user requirements are frequently occurred.

References:

[1] Inmon, W.H., *Building the Data Warehouse - Third Edition*. 2002: John Wiley & Sons, Inc. 97.

[2] Kimball, R. and J. Caserta, *The Data Warehouse ETL Toolkit. Practical Technique for Extracting, Cleaning, Conforming and Delivering Data*. 2004: Wiley Publishing, Inc., Indianapolis. 491.

[3] Giorgini, P., S. Rizzi, and M. Garzetti, GRAnD: A Goal-Oriented Approach to Requirement Analysis in Data Warehouses. *Decision Support Systems*, 2008. 45: p. 4-21.

[4] Alexiev, V., et al., *Information Integration with Ontologies: Experiences from an Industrial Showcase*. 2005: John Wiley & Son Ltd. 180.

[5] Skoutas, D. and A. Simitsis, Ontology-Based Conceptual Design of ETL Processes for Both Structured and Semi-Structured Data. *Semantic Web & Information Systems*, 2007. 3(4): p. 1-24.

[6] Bruckner, R.M., B. List, and J. Schiefer. *Developing Requirements For Data Warehouse Systems With Use Cases*. in 7th Americas Conference on Information Systems. 2001.

[7] Yu, E., *Modeling Strategic Relationships for Process Reengineering*, in Department of Computer Science. 1995, University of Toronto.

[8] Bresciani, P., et al., Tropos: An Agent-Oriented Software Development Methodology. *Kluwer Academic Publishers*, 2003: p. 1-40.

[9] Shen, G., Huang, Z., Zhu, X., & Zhao, X. (2006). *Research on the Rules of Mapping from Relational Model to OWL*. Paper presented at the OWLED'06, Athens, Georgia (USA).

[10] Aleksovski, Z. (2008). *Using Background Knowledge in Ontology Matching*. Vrije University.