PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Universidade Federal do Rio de Janeiro

JOÃO LUIZ REBELO MOREIRA

ONTOWAREHOUSING MULTIDIMENSIONAL DESIGN FOR HETEROGENEOUS DATA SUPPORTED BY FOUNDATIONAL ONTOLOGY: a temporal perspective

MASTER DISSERTATION

Rio de Janeiro
2014

Instituto de Matemática

NCE UFRJ Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE MATEMÁTICA
INSTITUTO TÉRCIO PACITTI DE APLICAÇÕES E PESQUISAS COMPUTACIONAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

JOÃO LUIZ REBELO MOREIRA

ONTOWAREHOUSING
MULTIDIMENSIONAL DESIGN FOR HETEROGENEOUS DATA SUPPORTED BY FOUNDATIONAL ONTOLOGY:
a temporal perspective

Master's thesis submitted to the Programa de Pós-Graduação em Informática, Instituto de Matemática, Instituto Tércio Pacitti de Aplicações d Pesquisas Computacionais, Universidade Federal do Rio de Janeiro as a partial requirement to obtain the title of Master in Informatics.

Advisor: Prof.ª Maria Luiza Machado Campos, Ph. D.

Rio de Janeiro
2014

M841 Moreira, João Luiz Rebelo

Ontowarehousing multidimensional design for heterogeneous data supported by foundational ontology:a temporal perspective. / João Luiz Rebelo Moreira. – 2014.

179 f.: il.

Master's thesis in Informatics -- Universidade Federal do Rio de Janeiro, Instituto de Matemática, Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais, Programa de Pós-Graduação em Informática, 2014.

Advisor: Maria Luiza Machado Campos

1. Multidimensional Design. 2. Heterogeneous Data Supported. I. Campos, Maria Luiza Machado (Adv.).II. Universidade Federal do Rio de Janeiro, Instituto de Matemática, Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais, Programa de Pós-Graduação em Informática. III. Title

CDD

João Luiz Rebelo Moreira

ONTOWAREHOUSING
MULTIDIMENSIONAL DESIGN FOR HETEROGENEOUS DATA SUPPORTED BY FOUNDATIONAL ONTOLOGY: A TEMPORAL PERSPECTIVE

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática, Instituto de Matemática e Instituto Tércio Pacciti de Aplicações e Pesquisas Computacionais, Universidade Federal do Rio de Janeiro, como requisito parcial à obtenção do título de Mestre em Informática.

Aprovada em 22 de agosto de 2014.

____

Prof.ª Maria Luiza Machado Campos, Ph. D, UFRJ

____

Prof. João Paulo Almeida, Ph. D, UFES

____

Prof.ª Jonice de Oliveira Sampaio, D.Sc., UFRJ

____

Prof. Pedro Manoel da Silveira, Ph. D, UFRJ

Acknowledgments

> “Gratitude can transform common days into thanksgivings, turn routine jobs into joy, and change ordinary opportunities into blessings.”
>
> *William Arthur Ward*

I can’t say that this dissertation is only mine because many people were involved in its construction. At first, I would like to thank God for this great life I have.

Thanks to my "academic mother" Maria Luiza Machado Campos – sometimes a stepmother during the reviews – for everything she has being doing for me in the last years, witch includes: helping me to finish the undergraduation course (when it was almost lost), supporting my doubts about the IT professional life (when I was thinking on leaving it), encouraging me to have the master course, always believing on my potential, insisting on teaching me formal ontology and UFO (even when I underestimated the research topic). For being such a good person for me in a lot of aspects, thank you very much.

Thanks to all colleagues from GRECO/PPGI/UFRJ, especially to my "academic sister" and surfer friend Kelli Faria, for being a great colleague in the last years: the partnership, all discussions, ideas exchanged and graphic design services. To Professors Jonice Olivera and Pedro Manoel, for being part of the examination committee. Thanks to Maria Ines Bosca, for helping me with several issues regarding ontologies.

Thanks to NEMO research group, for supporting me on both theoretical and practical questions about formal ontology, UFO and OLED tool. Special thanks to the colleagues Bernardo, Tiago and John, their help was fundamental to this achievement. To Professor João Paulo Almeida for accepting to be in the examination committee, providing essential comments to the final version of this dissertation.

Thanks to ONS organization, for giving me the opportunity to apply our proposal in the Brazilian electric system domain and for providing adequate conditions during the master course. To all friends from ONS who encouraged me during this process.

Thanks to my family, especially my parents José and Lucia, for providing me all the necessary education to achieve this title and the unconditional love even when I was absent. Thanks to the best person I ever met in this life: my great-aunt Maria (“Tia”) for doing everything I asked. To my brother and all my friends who also supported me.

A special thanks to my dear wife Bel, for all friendship and comprehension in those hard last years. Thanks also to her family, especially to her parents who always supported me.

To all who somehow participated and I did not mention above: thanks very much!

“Make it a habit to keep on the lookout for novel and interesting ideas that others have used successfully.”

Thomas Edison

Resumo

Moreira, João Luiz Rebelo. **Ontowarehousing multidimensional design for heterogeneous data supported by foundational ontology: a temporal perspective** 2014. Master's thesis (Mestrado em Informática) – Programa de Pós-Graduação em Informática, Instituto de Matemática, Instituto Tércio Pacciti de Aplicações e Pesquisas Computacionais, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2014.

A escolha de como representar a informação é extremamente importante para alcançar requisitos analíticos, fazendo da modelagem multidimensional (MD) uma tarefa fundamental no ciclo de vida de soluções de Business Intelligence (BI) e Data Warehousing (DW). Para isso, necessita-se de um processo de engenharia capaz de capturar a semântica das entidades do negócio e suas relações e juntamente com as necessidades de BI identificadas, avaliar para as possibilidades oferecidas pelos dados existentes, a melhor forma de organizá-los para o processamento analítico. A expressividade semântica na modelagem MD é um assunto que vem sendo estudado há alguns anos. Porém, a falta de construtos para expressar a conceitualização de fenômenos do mundo real ainda apresenta desafios, refletindo-se também na dificuldade em escolher as representações corretas para expressá-los no modelo MD, de forma a melhor explicitar restrições, dependências e regras de negócio em geral, sendo o problema tratado aqui. Nessa dissertação é apresentada uma nova abordagem ontológica para a derivação de conceitos e esquemas MD, sugeridos ao modelador, a partir de categorias da ontologia de fundamentação Unified Foundational Ontology (UFO), usadas para classificar o domínio dos dados de origem durante a modelagem MD. Propomos uma automação da abordagem híbrida, onde a ontologia de domínio é construída com base em dados heterogêneos (fontes estruturadas e não estruturadas) e posteriormente classificada com conceitos da UFO. Então, os conceitos MD são derivados a partir da ontologia de domínio por regras de mapeamento: (i) Eventos como Fatos; (ii) Participações de objetos como Dimensões e Hierarquias; (iii) Relações Temporais como um esquema Snowflake; (iv) Relação de causalidade como dicotomia Fato / Dimensão; (v) Mudanças de situações como um esquema MD para análises causa-efeito. A abordagem é validada através de argumentação das evidências obtidas na aplicação no cenário do sistema elétrico brasileiro, para exploração conjunta de informações de perturbações elétricas e sua repercussão em notícias. Uma discussão sobre causalidade e mudanças de situações é apresentada usando uma ontologia do processo ITIL como exemplo.

Abstract

Moreira, João Luiz Rebelo. **Ontowarehousing multidimensional design for heterogeneous data supported by foundational ontology: a temporal perspective** 2014. Master's thesis (Mestrado em Informática) – Programa de Pós-Graduação em Informática, Instituto de Matemática, Instituto Tércio Pacciti de Aplicações e Pesquisas Computacionais, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2014.

The choice on information representation is extremely important to fulfil analysis requirements, making the multidimensional (MD) modelling task a fundamental phase in Data Warehousing (DW) lifecycle and Business Intelligence (BI) solutions. For that, an engineering process to capture semantics from business entities and their relations is required. This process must take in account the identified BI needs and evaluate the best ways to organize them for analytical processing, considering the possibilities offered by the existing data. The semantic expressiveness in MD design is an issue that has been studied for some years now. Nevertheless, the lack of conceptualization constructs from real world phenomena in MD design is still a challenge, reflecting the difficulty in choosing the correct representations to express concepts in a MD model, considering identity principles, restrictions, dependencies and business rules, which is the problem treated here. Therefore, in this dissertation, it is introduced a novel ontological approach for the derivation of MD concepts and schemas, suggested for the modeller, using categories from a foundational ontology (FO) to analyse the data source domains as a well-founded ontology, supporting MD design. We propose a systematic automation of the hybrid approach, where the domain ontology is built based on heterogeneous data (structured and unstructured sources), classified with the Unified Foundation Ontology (UFO) conceptualization, increasing its expressiveness. Thus, MD concepts are derived from the domain ontology by a set of mapping rules: (i) Events as Facts; (ii) Object Participations as Dimensions and Hierarchies; (iii) Time Interval Relations as Snowflake Schema; (iv) Causality Relation as Fact/Dimension dichotomy; (v) Situation Changes as MD schema for cause-effect analysis. The approach is validated through arguing the evidences obtained by its application in Brazilian electrical system scenario, supporting joint exploration of electrical disturbances information, as structured data; and their possible repercussion on the news publications, as unstructured data. In addition, a discussion is presented for causality and situation changes, exemplified within ITIL ontology.

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
BI	Business Intelligence
CM	Conceptual Modeling
CMS	Content Management System
DB	Database
DBMS	Database Management System
DE	Domain Engineering
DL	Descriptive Logic
DM	Data Mart
DSS	Decision Support Systems
EDW	Enterprise Data Warehouse
ETL	Extract, Transforming and Loading
DW	Data Warehouse
FO	Foundational Ontology
IR	Information Retrieval
IS	Information System
MD	Multidimensional
MDA	Model-Driven Architecture
MDD	Model-Driven Development
NLP	Natural Language Processing
NOSQL	Not Only SQL
OCL	Object Constraint Language
OLAP	On-Line Analytical Processing
SE	Software Engineering
SW	Semantic Web
UFO	Unified Foundational Ontology
UML	Unified Modeling Language
V&V	Verification and Validation

Contents

Introduction

Business Intelligence (BI) solution based on Data Warehouse (DW) architecture is a well-accepted approach for analytical information systems (KIMBALL e ROSS, 2013). For the last 30 years it has become a major industrial domain and economic driver (TDWI, 2013). From a research (DECISIONPATH, 2010), it is estimated that 90% of all enterprises use this type of solution in their business decisions, with 70% using BI solutions across more than one department and approximately 20% of them use BI solutions widespread across most or all of their departments. Many organizations have been adopting this type of solution to support decision making processes and even for operational concerns. Most often (64%) BI solutions are directly related to traditional reporting used mainly by power users, but in 32% of the cases it can be used by all levels of corporations.

Both academic and industrial efforts have embraced the evolution of techniques and tools for BI/DW solutions. The number of courses and academic schools with BI/DW disciplines has been increasing for the last years, from latu sensu to stricto sensu (TDWI, 2010), such as the IT4BI (Information Technologies for BI) European master and doctoral programmes[1], which counts with experienced researchers and professors of the area. Conferences such as DaWaK (DW and Knowledge Discovery) and DOLAP (International Workshop on DW and OLAP) represent some of the main international events that address the research topics of BI. Moreover, some institutions were created to provide in-depth and high-quality education and training in BI/DW industry, such as TDWI (The DW Institute)[2], which provide recognized best practices reports about the strategies, techniques and tools required to design, build and maintain DWs.

BI/DW initiatives in companies are aware of the challenges that face their projects. Among the main factors that contribute to BI success, the maturity of the development methodology is crucial. The scope of BI environments, centralized and decentralized BI resource organization, budgets, FTE (full-time equivalent) employees and team sizes are also relevant issues that must be addressed for a successful BI/DW project (TDWI, 2013). A

[1] https://it4bi-dc.ulb.ac.be/

[2] http://tdwi.org/

BI/DW solution usually counts with a set of techniques and tools. Examples of these are DBMS (Database Management Systems), ETL (Extract, Transform and Loading), data discovery, data quality evaluation, OLAP (On-Line Analytical Processing), predictive analysis and data mining tools. The most common modelling technique in BI/DW solution is the so called multidimensional (MD) design, which is based in the dimension/fact dichotomy. It is a method to deliver understandable information for users in a simple, concrete and tangible way (KIMBALL e ROSS, 2013).

BI encompasses several scientific and technological fields including information integration (HAAS e SOFFER, 2009), large-scale processing (HOANG, TRAN, *et al.*, 2011), big data analytics (CUZZOCREA, SONG e DAVIS, 2011), collaboration (MARSHALL, WOBBER, *et al.*, 2012), privacy (CUZZOCREA e BERTINO, 2011), modelling and semantics (JOVANOVIC, ROMERO, *et al.*, 2014). Each of these fields presents research topics to be evolved, such as the optimization of user-defined ETL activities (GALHARDAS, LOPES e SANTOS, 2011), streaming data treatment (LIU, LITA, *et al.*, 2008), data integration for semantic data (BERKANI, BELLATRECHE e KHOURI, 2013), flexible and efficient MD data processing (MUSLEH, COLL. OF COMPUT. SCI. & ENG., *et al.*, 2013), data-intensive analytical algorithms (SHAH, JAITLY, *et al.*, 2009), graph analytics (SATISH, SUNDARAM, *et al.*, 2014), query processing for big time-series (BIEM, FENG, *et al.*, 2013), DW in cloud environments (MA, SCHEWE, *et al.*, 2011), measurement of intangibles (LIU, XIE e WU, 2009), among others.

The MD design task is a fundamental core phase in BI/DW lifecycle (KIMBALL e ROSS, 2013). It requires an engineering process to capture semantics from business entities and their relations, dealing with restrictions, existential dependencies among analytical perspectives and business rules. The difficulty in choosing the correct representations to express the conceptualization constructs in a MD model is still an issue in MD design because it can limit the accuracy of business analyses or even compromise the model semantic (PARDILLO e MAZÓN, 2011). Furthermore, considering unstructured data during the MD design activity is a challenge because of the difficulty in representing concepts from large and ambiguous textual sources. This is not addressed by typical dimensional modelling methodologies and therefore, most of the data on a company is not used (NESAVICH e INMON, 2007), worsening the problem. Although some solutions, based on Natural Language Processing (NLP) and Information Retrieval (IR) techniques, have been recently proposed for data representations (FREITAS, CARVALHO, *et al.*, 2012), only few researches are adopting unstructured data in BI/DW solutions (PARK e SONG, 2011). With the explosion of the internet, enhanced with hardware and software computing capabilities, this new paradigm needs to be investigated.

From a BI/DW designer point of view, in the Software Engineering (SE) context, capturing essential aspects of domains during BI/DW lifecycle, from the perspective of the subject matter experts, is the specific research topic treated in this work. It includes the identification and analysis of relevant concepts for designing conceptual MD models, coping with analytical requirements and data sources. In this direction, ontologies have been already applied as a mechanism to enhance the semantic expressiveness of domain representations from data sources (ROMERO e ABELLÓ, 2010). In addition, we consider heterogeneous and complex data, basically classified as structured or unstructured data[3]. This dissertation is concerned with the development of derivation rules from a well founded domain ontology to multidimensional (MD) concepts as suggestions for the MD modeller.

In this chapter, at first, the general concepts to support this work are presented. Secondly, the problem definition is formally stated. Afterwards, the methodology used is described, also defining the expected objectives based in the Goal Question Metric (GQM) template. Thereafter, a minimal scope for this work is set. Then, the structure of this dissertation is presented.

1.1 General concepts

To deal with modelling and semantics particularities, the general concepts to support this work are the research topics from BI/DW and formal ontology. Regarding the first and the second, MD design approaches (analysis-driven, supply-driven and hybrid), development methodology (project lifecycle) and unstructured data treatment (NLP and IR techniques) are the basic topics involved. Concerning the formal ontology research area, foundational ontologies (FO), and, specifically, the Unified Foundational Ontology (UFO) (GUIZZARDI, 2005), with its application in different domains to increase model's semantic expressiveness

[3] We consider only textual data as unstructured and disregard images, sounds and others. There is a discussion if a formal text is considered unstructured or not, because it follows morphological and lexical patterns. However, we do not make this distinction in our approach.

are utilized in this dissertation. The definitions regarding FOs, their relations to domain ontologies and their role in the formalization of a domain representation are explored to support our solution proposal.

1.2 Problem definition

The choice of a proper data representation structure is extremely important to fulfil analysis requirements, making the modelling task fundamental in the BI/DW project. A problem in this context is the difficulty in choosing the correct representations to express the concepts in a MD model, considering identity and part-whole principles, existential dependencies, constraints and business rules. Representing conceptualization constructs from real world phenomena is still an issue in MD design where the lack of semantic expressiveness in conceptual models may compromise the accuracy of business analyses or even limit its scope and comprehensiveness. The semantic power in the process of MD modeling is still a challenge that has been studied for some years now (ABELLÓ, 2002) (MALINOWSKI e ZIMANYI, 2004) (ROMERO, 2010). Even some practitioners, such as Kimball (KIMBALL e ROSS, 2013) and Inmon (INMON, 2005), introduced several design guidelines for choosing MD elements to represent domain concepts; they all were stated in an informal way, not considering theoretical foundations from different fields, like metaphysics, for example. Therefore, the main problem addressed here is the lack of formalization in choosing the appropriate concepts from a domain to use in MD design.

1.3 Objective

The objective of this work is to deal with the problem mentioned above by formulating a semi-automatic derivation process based on mapping rules from concepts of a domain ontology, well founded on the foundational ontology UFO, to elements of a MD schema. This process uses ontological analysis based on UFO categories, taking advantage of their precise characterization of domain concepts that are represented in data sources, enriching semantically the modelling activity.

1.4 Methodology

The research methodology adopted in this work counts with bibliographic revision of the general concepts and related works, proposal approach formulation and validation through experimentation and examples. The experimental study follows the model defined in (WOHLIN, RUNESON, *et al.*, 2012), where the plan is specified in order to facilitate its reuse in a future repetition of the study. The definition can be summarized in the following assumptions:

Table 2.1: Description of the methodology adopted in this work

Object of study	The use of ontological approach based in temporal aspects of a FO, specifically UFO, in a hybrid MD design activity for BI/DW solutions.
Purpose	The objective/goal is to formulate derivation rules for MD modelling from UFO concepts, applied to domain ontologies that represent data sources, increasing the semantic expressiveness during the activity of MD design.
Quality focus	The gain achieved by the use of the proposed technique is measured by discussing its effectiveness in choosing the concepts to represent MD concepts from real scenarios and different domains.
Perspective	The view point of the proposed hybrid approach is from the MD modeller perspective for BI/DW solutions development.
Context	BI solution based on DW architecture with MD design as the default representation structure, supported by ontological analysis.

We also state our work with the GQM template (SOLINGEN e BERGHOUT, 1999), where the goal level is the conceptual one, having an objective defined for an object range, respecting quality models from different perspectives relative to a particular environment. The question level is the operational level, where questions are stated to define the assessment of a goal through a characterization model. The objects of measurement characterization are based in quality aspects from a selected viewpoint. The metric level is the quantitative layer, where objectively or subjectively a set of data is linked to each question to answer it in a solid way. In this dissertation we chose the subjective measure of arguing about the results benefits and limitations. Therefore, the GQM template for this research is defined as follows:

Table 2.2: Methodology described in GQM template

To analyse the use of ontological approach based on temporal aspects of a FO, specifically UFO, in the MD design activity for BI/DW solutions.
For the purpose of formulating derivation rules for MD modelling from UFO concepts, applied in domain ontologies that represent data sources, increasing the semantic expressiveness during the MD design activity.
With respect to benefits and drawbacks of adoption the approach.
From the point of view of MD modellers for BI/DW solutions development.
In the context of BI/DW solutions based in MD design activity, supported by ontological analysis.

1.5 Scope

In this work the scope is defined as:

- Revision of the main literature and related works regarding MD design in BI/DW solutions;

- Revision of the main related works to unstructured data use in BI/DW solutions;

- Revision of a BI/DW development lifecycle methodology;

- Revision of related works addressing ontological approaches for BI/DW solutions;

- Revision of the main literature and related works regarding formal ontology, specifically FO and UFO concepts;

- Exploration of *perdurants* aspects from UFO to increase semantic expressivity in MD design activity, considering mereological relations among events; participations of objects in events; time interval relations and causality relation between events; and situation changes related to events;

- Introduction of an ontological approach based in derivation rules from UFO to MD concepts;

- Introduction of a hybrid method considering the prior ontological approach and unstructured data modelling;

- Validation of the approach through the implementation of an example in real scenarios, demonstrating each derivation rule execution and hybrid method application.

1.6 Structure

This dissertation is organized as follows:

- Chapter 2 presents an in-depth characterization of concepts and related works of BI and DW, the types of MD design approaches (analysis-driven, source-driven and hybrid) and the support for unstructured data in BI/DW development methodology. These concepts help to understand the research base of this work;

- Chapter 3 presents the basic concepts of ontologies, how these were already applied in BI/DW solutions. Furthermore, FO is described, particularly UFO and its parts. OntoUML, a language that considers some of UFO's stereotypes, discussing related works and applications;

- Chapter 4 presents the approach proposed in this work, so called OntoWarehousing. A set of mapping rules is introduced, describing how MD concepts can be derived from a domain ontology based on UFO concepts, such as an event as a fact and participation as perspective of analysis. In addition, a hybrid MD design adaptation regarding these mapping rules and the use of unstructured data sources is depicted;

- Chapter 5 presents the experimentation of the approach introduced in section 4. At first, a prototype for MD elements derivation through rules execution is described. Afterwards, a study case exemplification in the Brazilian electrical grid security domain illustrates the proposed hybrid approach, considering the prototype execution and the use of unstructured data sources. At last, a discussion on causality and situation changes rules is made upon an example of ITIL process domain scenario, exemplifying through a MD schema generation;

- The Conclusion describes the main contributions of this dissertation and future works to address on the continuity of this research line.

2 Business Intelligence and Data Warehousing

This chapter presents the main background concepts of the study and related works. The research major topics are *Business Intelligence* (BI), *Data Warehousing* (DW), *Multidimensional* (MD) design and BI/DW Project Lifecycle. The concept of BI was firstly conceived by Hans Peter Luhn in 1958 as “the ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal.” (LUHN, 1958). The BI term got popularity with Decisions Support Systems (DSS), which research began in 1960s, and tied to DW since 1990s. However, BI and DW are different concepts, a BI system can be built with DW architecture or not. BI is the set of architectures, methodologies, technologies and processes to enable analytical information exploring. BI can be understood as the use of multiple sources of information with the main goal to support the definition of strategies for companies.

Some authors state that BI aims to increase the companies profitable and competitiveness in its market (MOSS, 2003). However, we believe that the concept of BI is broader, because it is bound to assist the decisions within a business domain. Independently of organizations objectives, profitable or not, BI solutions can provide analytical information for decision making.

To build a BI information system it is necessary to follow an adequate software development methodology. In addition, it must be based on inter-organizational initiatives, coping with qualified sponsors and appropriate BI project team.

2.1 Data Warehousing

DW is defined as a technology by some authors (SOARES, 1998) (OUESLATI e AKAICHI, 2010). However, it can be better understood as a software architecture (INMON, 2005) because it refers to a high-level design structure, whilst technology refers to specific platforms from vendors, as sets of software and hardware. To avoid mistaken interpretation, DW is not a product that is simply bought and installed in the company, nor an implementation language, nor an isolated single project and nor a copy of transactional systems.

As DSS natural evolution, the term DW was introduced by Bill Inmon in 1990s (INMON, 1992). It was defined as data integration and consolidation process to centralize the necessary information for analytical decision makings from the information systems sources stored in relational DBMS. Its fast absorption from the companies is related to the domain information needs to guarantee analytical responses and actions to ensure their business decisions. Among other reasons, the technological advances, the changes in business structures and economy globalization contributed to it.

The mission of a DW is to publish the organization's data assets to most efficiently support decision making. The BI/DW system requirements can be summarized as: to make information easily accessible, to present it consistently in a timely way, to be adaptable to changes, to be secure and to be a trustworthy foundation for decision making (KIMBALL e ROSS, 2013). The BI/DW system data-flow, i.e. the Extract, Transform and Loading (ETL) process, begins in data extraction from heterogeneous data sources (internal or external, structured or not), then integrates and transforms data and delivers the data to end-users through different data visualization levels, accessible via On-Line Transactional Processing (OLTP) and/or On-Line Analytical Processing (OLAP) tools. In general, architectures oriented to BI/DW solutions consist of a set of tools that must respond to heavy query processing load. Those tools include ETL capability to prepare and deliver the data, OLAP capability to visualize and explore the data, data profiling capability to evaluate data quality in its origins and data mining capability to check data patterns and rules, enabling predictive analysis.

Numerous academic researches and commercial initiatives in BI/DW have been developed for the last 30 years. From 1990s until now, we can cite as significant authors of BI/DW research area: Bill Inmon, Ralph Kimball, Margy Ross, Larissa Moss, Esteban Zimanyi, Elzbieta Malinowski and Alberto Abelló. Some commercial books stand out, such as Building the DW (INMON, 1992) and the DW toolkit editions from Kimball's works (KIMBALL e ROSS, 1996) (KIMBALL e ROSS, 2002) (KIMBALL e ROSS, 2013). The later proposed the MD design activity, describing fundamental concepts, different techniques and application case studies. Regarding MD design, the Advanced DW book (MALINOWSKI e ZIMÁNYI, 2009), originated from the author PhD thesis, introduces extensions for spatial and temporal concepts in MD modeling. Kimball's books about ETL (KIMBALL, 2004) and BI/DW Lifecycle toolkits (KIMBALL, ROSS, *et al.*, 2008) should also be mentioned as important related work.

The definition adopted in this work for a BI/DW solution is "a system that extracts, cleans, conforms, and delivers source data into a dimensional data store and then supports and implements querying and analysis for the purpose of decision making" (KIMBALL, 2004). Therefore, while some works state that MD design is not strictly necessary for a BI/DW solution (MOSS, 2003); we consider the MD design in our BI/DW approach.

2.2 Multidimensional design

> "The ability to visualize something as abstract as a set of data in a concrete and tangible way is the secret of understandability. (…) Albert Einstein captured the basic philosophy driving dimensional design when he said, 'Make everything as simple as possible, but not simpler'. " (KIMBALL e ROSS, 2013)

Also called dimensional modelling, MD design is the most accepted technique for presenting analytic data because it delivers information that is understandable to business users and provides fast performance when querying. It is intuitive to query and presents the information for the user in a concrete and tangible way. The simplicity of MD models is the main reason why MD design is widely employed, being its most important property because it makes the data understandable for non-expert users. For example, it is not necessary to know SQL to retrieve analysis results from a MD model through OLAP tools. Moreover, it allows software to provide navigation and result delivery capabilities in a quick and efficient way. Indeed, the data loaded in MD models represent the same information as operational normalized models. However, it presents the data in a formatted way, delivering understandable information for the user, coping with query performance and resilience to change (KIMBALL e ROSS, 2013). It can be implemented in relational DBs, usually referred to star or snowflake schemas, being available to be accessed by Relational OLAP (ROLAP) tools. It can also be implemented in MD DBs, known as data cubes, being available to be accessed by MD OLAP (MOLAP) tools.

The MD conceptual view of data is based in the fact/dimension dichotomy, where the data items with *n* attributes are represented by points in an *n*-dimensional space (ROMERO e ABELLÓ, 2010). A MD model structures the information into facts and dimensions, basically. A fact represents a focus of analysis (MALINOWSKI e ZIMÁNYI, 2009) or a business process measurement event (KIMBALL e ROSS, 2013) or a subject of analysis (ABELLÓ, 2002). Examples are `sale`, `payment`, `delivery` and any other business processes, such as `product development process` or a `service provision`. Notice that they are all representations of something that happened in time, composed by events, bringing the reality from one situation to another. In addition, they can only happen, i.e. they have existential dependency, with the participation of other things to contextualize it. The dimensions are those things that are associated to the fact, they describe "who, what, where, when, how and why" associated with the event (KIMBALL e ROSS, 2013). For example, a `common sale` depends on a `vendor`, a `client` and a `product`, occurring during a `time` interval in a certain `location`.

The dimension attributes and hierarchies are perspectives of analysis of a fact, commonly identified as the "by" words in report requests. Dimensions and facts are represented in DB as data tables. The dimension is defined by a single Primary Key (PK) and attributes, which may form hierarchies, such as `location` dimensions (e.g. `country`, `state` and `city`) and `time` dimensions (e.g. `year`, `semester`, `month` and `date`). The concept of hierarchy is fundamental in analytical solutions, because human mind is organized hierarchically, being the base of logic in human cognition (ZHOU, JIN e HAN, 2009). In the last years several works have been proposed for hierarchy visualization techniques. The survey (SCHULZ, HADLAK e SCHUMANN, 2011) introduces a systematic design space of these techniques.

The conceptual classification of OLAP hierarchies was introduced in (MALINOWSKI e ZIMÁNYI, 2004) and different usages of them and their representations in graphs were explored in (VIEIRA, 2013). A hierarchy level is the participation of a dimension in the hierarchy. The items comprising the hierarchies are called members or nodes. The sequence of members through the levels is called hierarchical path, where the number of levels is defined as the path length. The first Level of a hierarchical path is the leaf, which is the most detailed, and the highest Level of aggregation is the root. Hierarchies are usually implemented as a flat table (in a star schema) or a normalized structure (in a snowflake schema). For a full understanding about aggregation in star schemas refer to (ADAMSON e KIMBALL, 2006).

DW hierarchies are fundamental in analytical solutions and its conceptual representation can be complex. It deals with aggregation paths, sequence of levels for roll-up/drill-down actions, kinds of hierarchies, instance levels, cardinalities and parent-child relationships. The parallel hierarchy is an aggregation of individual hierarchies, which can be simple or alternative. The former are the ones that can be represented as trees, i.e. all its parent-child relations are one-to-many. It can be balanced, unbalanced (ragged) or generalized. A full description of all these types is presented in (MALINOWSKI e ZIMÁNYI, 2009, page 80). The bridge-table plays a fundamental role in the implementation of hierarchies. It is a many-to-many table used to relate one row of the fact table to multiple rows of the dimension through a group table. It can be applied in the implementation of ragged hierarchies, as well as recursive pointer (KIMBALL, 2004).

The fact table has a set of Foreign Keys (FK) representing each dimension PK. A fact also contains measures, the attributes of the represented event (MALINOWSKI e ZIMÁNYI, 2009). Usually, they are numeric qualities that allow quantitative evaluation through aggregations, e.g. `product sales value`, `sales taxes`, `profits percentage`, among others. The idea is to represent the measurement event of the physical world as a one-to-one relationship to a single row in the fact table (KIMBALL e ROSS, 2013). The additivity of a measure is an essential property. It defines the behaviour of aggregation through different rows when joining and grouping the related dimensions. Common examples of aggregations functions are sum, maximum, minimum and average. Furthermore, calculated measures can be set up with manifold math functions, such as exponential, hyperbolic, logarithms, polynomial and periodic functions. Semi-addictive measures are the ones defined by the modeller to be aggregatable for a subset of dimensions. The non-additive measures are the ones that should not aggregate when drilling-down/rolloing-up.

A DW designed with MD schemas can also be understood as specialized DB aimed to support the decision-making process, which stores and delivers subject-oriented, integrated, nonvolatile and time-varying data. Therefore its design should be made through a method, similar to an information system design activity. Conventional transactional system supports the business operational processes, storing all data input. Conceptual Modelling (CM) is commonly used in software development process and it is revised in chapter 3. A system is generally designed using the conceptual, logical and physical model levels. The first is a high level (abstract) conceptualization, where the most important domain concepts, their relations and some restrictions (business rules) are described. The main goal of conceptual models is to provide a common understanding of the represented domain among the stakeholders (PARENT, SPACCAPIETRA e ZIMÁNYI, 2006). In addition, they serve as system documentation, providing a reference point for software developers. Generally, they are formalized through Unified Model Language (UML) and even through Entity-Relationship (ER) language, describing normalized relations for the correspondent logical schema. That one is typically produced from the conceptual model, where the implementation paradigm is chosen, such as relational, which is typically generated with ER representations, or object-orientation (OO), typically represented with UML. Afterwards, the physical schema is designed from the logical model to describe the intern data structures, e.g. tables, columns, relationships, PKs and FKs, indexes, constraints, among others. In other words, for common transactional information systems, specific features of the DBMS are used in physical models to increase querying performance, improve data normalization and storage.

Several CM researches have been conducted in the last years to deal with designing issues for transactional and analytical systems. The expressivity needed for better describing the real world phenomena in models is one of them. Also called semantic expressiveness or semantic power, it is the measure of how a model describes the reality (SALTOR, CASTELLANOS e GARCÍA-SOLACO, 1991), i.e. how a model best represent conceptual structures. The semantic enrichment of a model occurs when its semantic expressiveness is increased. Unlike the traditional conceptual models, the MD conceptual schemas must be modelled in a way that ensures a better comprehension of the data for common user analysis, but also to increase performance for complex queries (MALINOWSKI e ZIMÁNYI, 2009). In this direction, some works introduced approaches to semantically enrich MD models. MALINOWSKI e ZIMÁNYI introduced ER representations for conceptual MD models (MALINOWSKI e ZIMANYI, 2004), as illustrated in Figure 2.1. By grouping characteristics into their corresponding levels, it is possible to enrich the expression power of the ER model. A dimension is differentiated from a fact by its shape: the former is rectangular, whilst the latter is rhombus. In addition, measures are directly connected to the fact described in a rounded rectangle. The hierarchy is represented by *n*-ary relations between dimensions.

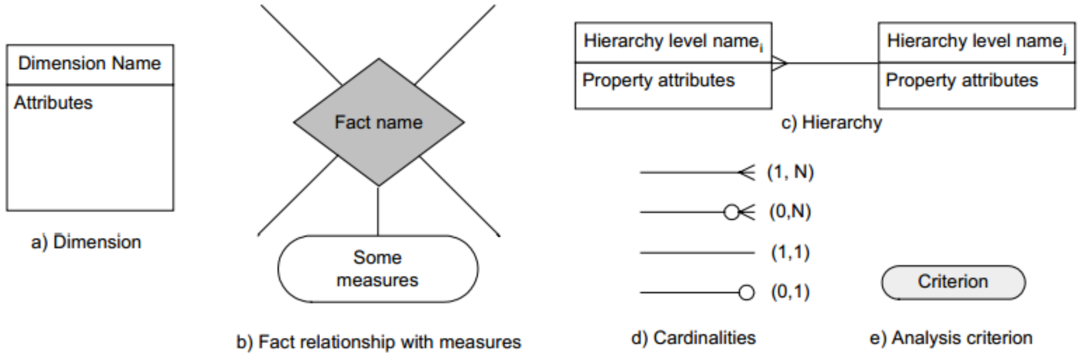


Figure 2.1: Elements to enrich the semantic expressivity of MD models in ER from (MALINOWSKI e ZIMANYI, 2004)

Thereafter, in Advanced DW (MALINOWSKI e ZIMÁNYI, 2009), the ER metamodel was extended to describe MD concepts dealing with temporal and spatial concepts, commonly used in MD models. It is stated that an event correspond to a phenomena at one instant or a set of instants, while a state occurs during an interval or a set of intervals. The temporal data types (Figure 2.2a) consider simple and complex time structures, i.e. a unity or a set, for instants and intervals. Moreover, icons to characterize synchronization relationships between events were introduced based on Allen's temporal predicates (ALLEN, 1983), depict in Figure 2.2b.

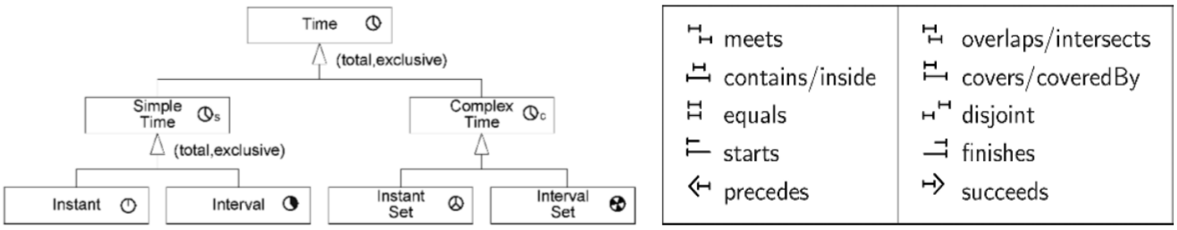


Figure 2.2: Enhancing semantic expressiveness in MD models: (a) temporal data types and (b) syncronization relationships from (MALINOWSKI e ZIMÁNYI, 2009)

Besides temporal data types and temporal relations, temporality types were also explored in this work. The Valid Time (VT) demonstrates a time period in which a fact is true in the modeled reality. The Transaction Time (TT) represents the time period in which a fact is current in the DB, beginning when the row in the data table is inserted or updated and ending when it is deleted or updated, commonly generated by the source system. When both occurs (VT and TT) it can be classified as Bitemporal Time (BT). The Lifespan (LS) is an object existence time in the source application, used to represent the duration of an instance. It is also applied in relationships, demonstrating how long a relation instance can exist. For last, the Loading Time (LT) represents the time since when the data is current in a DW.

The application of these concepts in an example scenario is shown in Figure 2.3. It represents a common MD model with a `sales` fact, which is classified as an event that can overlap, i.e. a `sale` instance overlaps another `sale` instance. Furthermore, it defines that the measure quantity amount is a VT, which means that it keeps track of the changes in its value. The same classification is applied in `product`, `category` and `sales district` attributes. Notice that these types of classification enhance the understanding of a MD model regarding temporality issues. A complete description of this model example can be found in page 192 of (MALINOWSKI e ZIMÁNYI, 2009).

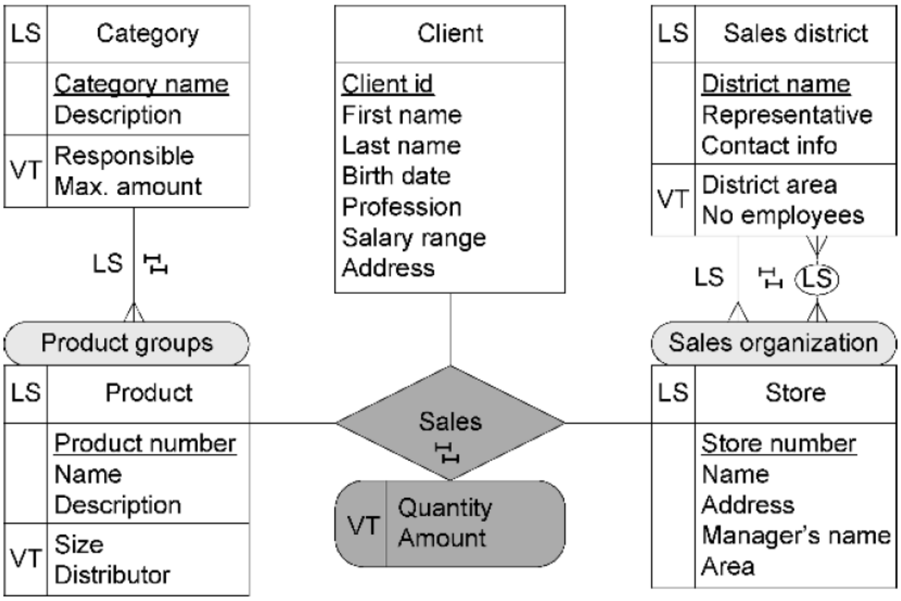


Figure 2.3: A MD model example semantically enriched by temporal concepts from (MALINOWSKI e ZIMÁNYI, 2009)

In (ABELLÓ, 2002) a survey of different metamodels for MD design with UML was made. In addition, it introduces a complete conceptual MD metamodel described with UML (YAM²), coping with semantic OO benefits for stars relations. Among other characteristics, it deals with explicit aggregation and multiple hierarchies, measures at different levels of granularity, generalization and association relationships, many-to-many relationships between two levels and between fact and dimension, inherent integrity constraints and operations (e.g. drill-across, roll-up, projection and dice). Figure 2.4 depicts the main concepts of YAM² MD metamodel and their relations, split in three abstraction levels.

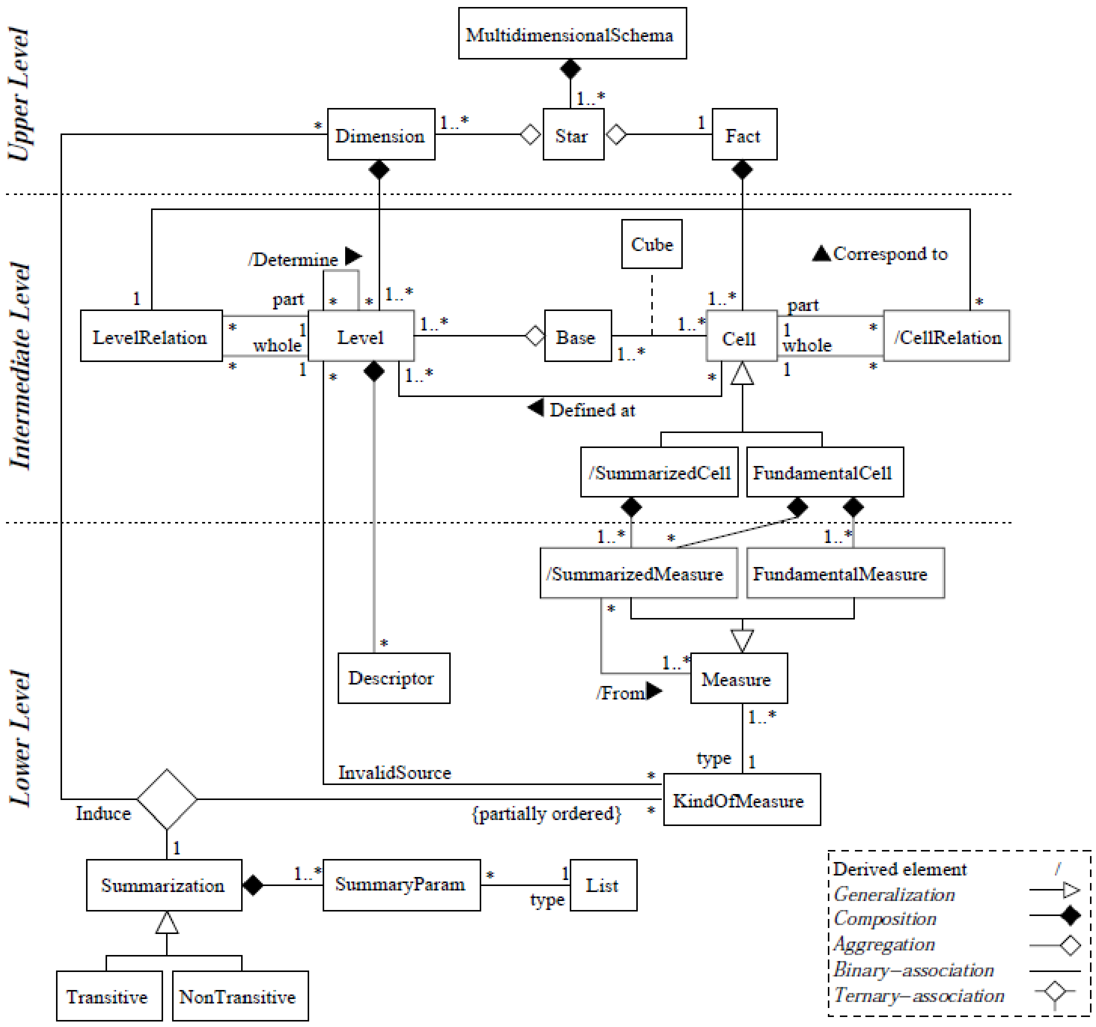


Figure 2.4: YAM² MD metamodel using OO (UML) from (ABELLÓ, 2002)

Regarding MD modelling, the Common Warehouse Metamodel (CWM) is an important research effort to be highlighted (MEDINA e TRUJILLO, 2002). It is an open industry specification of Object Management Group (OMG) and also describes MD concepts as an UML extension, dealing with some of the issues addressed by YAM² – CWM is one of the MD metamodels compared in the survey mentioned. However, the main objective of CWM is to provide a standard metadata definition to ensure interoperability among different DW platforms, such as OLAP, ETL and data mining tools. The CWM architecture is organized in 21 packages, grouped in five layers by means of similar roles. The analysis layer, specifically the OLAP package, can be used for conceptual MD design. Nevertheless, it lacks in characteristics, such as measure sets and additivity semantics, and was not conceived as a conceptual model. More recent works are applying ontologies to represent the domain and the correspondent data sources, dealing with semantic expressiveness issues of conceptual MD models. These works are revised in section 3.1.

As cited before, the design activity of MD modeling is the most important and crucial phase in the development of a BI/DW solution, being a fundamental core phase in the BI/DW project lifecycle. It requires an engineering process to capture semantics from business entities and their relationships. A problem in this context is the difficulty in choosing the correct representations to express the concepts in MD models. Moreover, MD design depends mostly on a prior knowledge from the designer, being error prone. This situation results in the lack of semantic expressiveness in MD models.

A method for MD design was introduced in (MALINOWSKI e ZIMÁNYI, 2009), following the same steps as in transactional systems development. Figure 2.5 illustrates the process, beginning by requirements specification from interviews with stakeholders. Then, the conceptual design phase considers these elicited requirements to describe concepts and their relations to respond the analytic questions. Afterwards, the logical model is designed from the conceptual model and the physical, usually, auto-generated from the logical model. The MD modeling task for BI/DW solutions considers those four phases and may be classified as analysis-driven, supply-driven or hybrid approach.

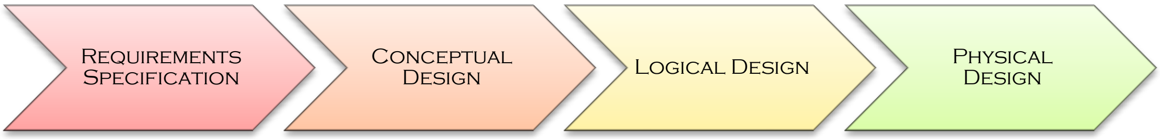


Figure 2.5: Standard design process for transactional and analytical systems

2.2.1 Analysis-driven approach

Also called demand-driven or user-driven, analysis-driven approach is the process where the user is fundamental during the requirements analysis and the design of concepts for facts and dimensions through sessions of interviews and meetings. Kimball's approach

(KIMBALL e ROSS, 2013) can be considered as analysis-driven, illustrated in Figure 2.6. It starts in a preparation activity, where business participants are identified; business requirements are elicited and reviewed. In addition, modeling and data profile tools are chosen and naming conventions are defined. As result, the business case, bus matrix and detailed business requirements are generated, serving as input to the MD design process. Thereafter, business processes to be analyzed are identified and a high-level model is designed, detailing the grain of analysis, the facts and dimensions concepts found. In an interactive and iterative process the MD model is verified and validated with the business representatives. At last, the final MD design documentation is written, with the detailed DB design and an issues log.

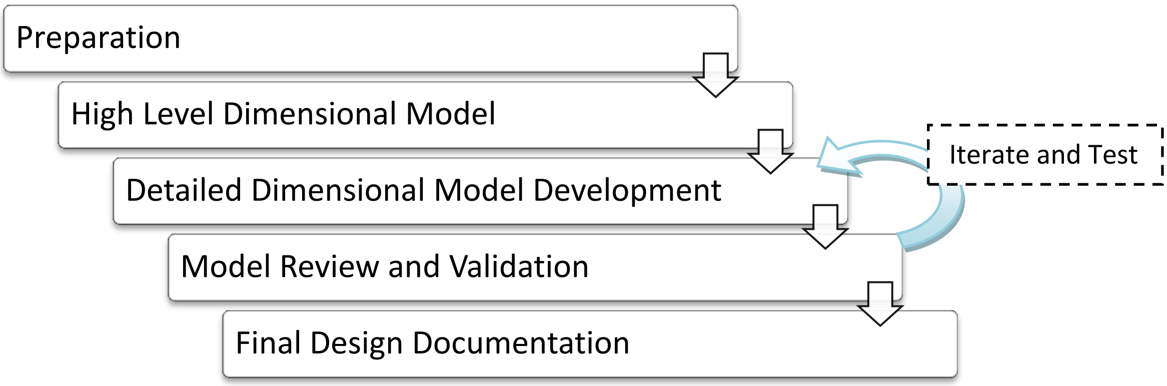

Preparation
High Level Dimensional Model
Detailed Dimensional Model Development
Model Review and Validation
Final Design Documentation
Iterate and Test

Figure 2.6: Kimball’s MD design process

Among the main advantages of Kimball’s analysis-driven approach are: (i) it enables the understanding and formalization of specific business needs; (ii) it provides to users a better understanding about the facts, dimensions, measures and attributes; (iii) it defines AS-IS business process models and increases the acceptance of the BI/DW system. The main disadvantages are: (i) user’s requirements can be different from the business goals; (ii) duration of the project tends to be longer, increasing its cost; (iii) existent information in sources may not be feasible to achieve the requirements. Similar to Kimball’s approach, Malinowski e Zimányi described the analysis-driven method as illustrated in Figure 2.7. Notice that the main differences from Kimball’s are the data availability check, the ETL definition and implementation.

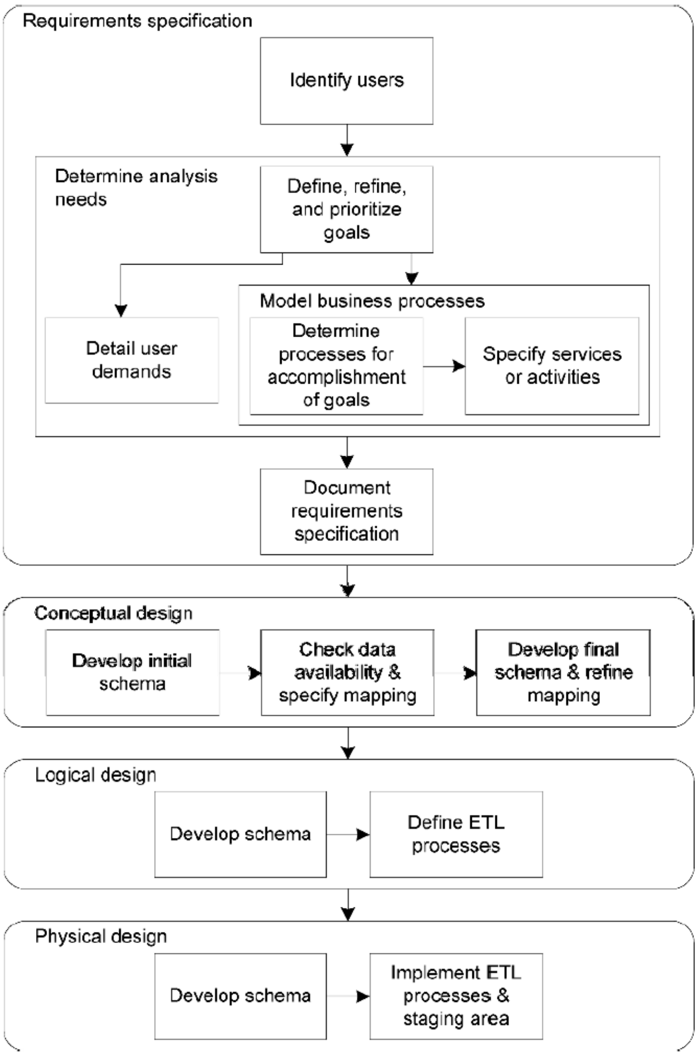

Requirements specification
Identify users
Determine analysis needs
Define, refine, and prioritize goals
Detail user demands
Model business processes
Determine processes for accomplishment of goals
Specify services or activities
Document requirements specification
Conceptual design
Develop initial schema
Check data availability & specify mapping
Develop final schema & refine mapping
Logical design
Develop schema
Define ETL processes
Physical design
Develop schema
Implement ETL processes & staging area

Figure 2.7: Analysis-driven process from (MALINOWSKI e ZIMÁNYI, 2009)

A variation of this approach is the so called business-driven or process-driven or goal-driven or requirements-driven, where the derivation of the concepts of the MD model starts from an analysis of the high-level business requirements or the business processes, existent services and activities specifications (WINTER e STRAUCH, 2003). To a better understanding of detailed differences among these requirements approaches, refer to (BUSSER, 2011).

2.2.2 Source-driven approach

Also called supply-driven or data-driven, the MD model is derived from the source systems analysis, looking for normalized DBs to extract the facts, dimensions, measures and hierarchies concepts. The users are involved only sporadically and the data is typically represented at a low level of detail. Among its main advantages are: (i) it reflects the underlying relationships in the data; (ii) it simplifies the ETL process; (iii) source systems may provide more stable basis then user requirements; (iv) the development process can be faster and if the sources are normalized DBs, then automatic or semiautomatic techniques can be applied, such as reverse engineering. Among its main disadvantages are: (i) business needs gathered are only reflected by the existent data source models; (ii) the DW system may not meet the user's expectations; (iii) the inclusion of hierarchies may be complicated and, in case of large sources data models, it is harder to be understandable. Figure 2.8 bellow demonstrates the activities during the source-driven approach from (MALINOWSKI e ZIMÁNYI, 2009).

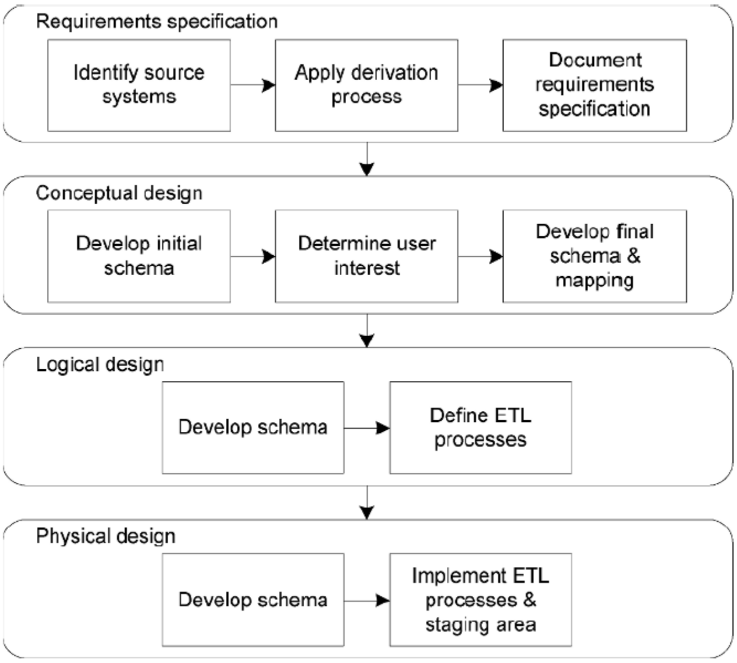


Figure 2.8: Source-driven approach from (MALINOWSKI e ZIMÁNYI, 2009)

A full comparison of the approaches can be found in (MALINOWSKI e ZIMÁNYI, 2009). In the source-driven approach the main step is the derivation process from source systems, which may be performed manually or (semi) automatic. Regardless the automation, it should follow a set of heuristics to find the dimensional concepts. For the last years there are some works in this direction, one of them is presented in (RODRIGUES, 2004). It introduces a proposal to obtain information compatible with the user analytical perception from source DBs, i.e. it classifies and selects the potential MD elements from relational DBs by a set of inference heuristics. Some examples of metadata collected of each element in the sources DB are columns name, data type, length, nullable admission, primary and foreign key relations, index participation, among others. In the end of the derivation process some analysis groups are proposed, composed by elements, tables and columns. They are classified and organized as trees, where roots represent the fact tables and leafs represent the dimensions. For the experimentation of this work it was used the TPC-H benchmark (TPC, 2002), a common DB used for examples regarding DW solutions.

2.2.3 Hybrid approach

Also called analysis/source-driven, it is the combined approach, where a source-driven approach is executed preliminary, providing a sketch of the existent data structures from the source systems. Then, it is executed an analysis-driven approach where the model reflects the user needs. In a third step both models are matched somehow. In many real scenarios of hybrid approach executions, the users usually do not know the potential data for analysis from sources and may not consider them in their requirements. There is a distinction between sequential and interleaved hybrid approaches. The former occurs when demand-driven and source-driven are performed independently and the models conciliated at the end, whilst the later performs both stages simultaneously, using their partial results to support each other, benefiting from their feedbacks and obtaining better result at the end (ROMERO, 2010). The main advantages are: (i) it generates a feasible solution; (ii) it may indicate missing data in operational DBs that is required and the analysis can be expanded to include new issues not considered at first. The main disadvantage in sequential hybrid approach is the need, and therefore major effort, of designing two models to be matched in the end. The greater difficult is in the need of complex techniques for the integration process.

To increase the semantic expressivity of MD models described in ER specifications, (MALINOWSKI e ZIMÁNYI, 2009) introduced a set of concepts to categorize spatial and temporal constructs, as cited before. The schema generated by the hybrid approach is semantically enriched by including the inherent semantic from spatial properties, such as lines, surfaces and topological relationships; and temporal properties, such as temporality types and synchronization relationships, as illustrated in Figure 2.9.

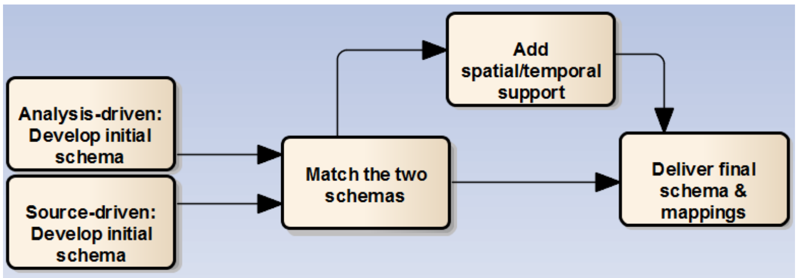

Add spatial/temporal support
Analysis-driven: Develop initial schema
Source-driven: Develop initial schema
Match the two schemas
Deliver final schema & mappings

Figure 2.9: Hybrid approach steps for spatial and temporal DW

To deal with the problem of matching user analysis requirements over the data sources, which is usually done manually in natural language, (ROMERO e ABELLÓ, 2010) proposed the automatic method MDBE, focused on linking end-user requirements with the data sources. It follows a classical approach, considering that the analytical requirements are clear, all gathered by the MD designer and specified as SQL queries to be executed in the data sources. Then, it discovers MD concepts by checking the requirements conciliated with the data sources. Section 3 presents the revision of the second part of this work (called AMDO) which uses ontologies and considers non-clear requirements. The continuation of this work through GEM approach (ROMERO, SIMITSIS e ABELLÓ, 2011) is also revised.

2.3 BI/DW lifecycle and the support for unstructured data

> "We need to look always for the relationships and inter retroactions between every phenomenon and its context, relations of reciprocity whole / parts: as a local modification affects on the whole and as a modification of the whole reflects on the parties" (MORIN, 2003, page 25).

The BI/DW system development methodology is also called the BI/DW project lifecycle. As cited before, the BI/DW solution construction follows similar activities of a transactional (conventional) information system. It should consider the same issues of software engineering, such as the process itself, the project management activities, its metrics, project planning, risk analysis and management, project scheduling and tracking, quality assurance, configuration management, architectural project and test techniques (PRESSMAN, 2002). Many organizations have the necessary infrastructure for the implementation of BI/DW applications. However, it is observed that many companies still lack on maturity in aspects such as understanding the complexity of BI/DW projects and the need of establishing a methodology for developing BI/DW projects. In addition, it is necessary to understand the BI/DW project manager role, business analysts' participation, key activities of standardization, evaluation of the impact of "dirty" data in business and to understand the needs and uses of metadata.

Several factors determine the complexity of a BI/DW project, such as the establishment of a clear difference between a BI/DW project and traditional one, described in Table 2.1. Moreover, understanding the function of each specific infrastructure component in a BI/DW application is important. Recognizing what are the impacting factors on a BI/DW project, determining the amount and types of resources (both technical and human) and defining the architecture of the application (e.g. MD design or ad-hoc queries) are natural concerns. One of the main differences between a BI/DW project to a traditional transactional one is the incremental definition of requirements. For each new iteration in application development, the requirements for strategic information must be reviewed and enhanced. This is due mainly to the fact that a BI/DW application is oriented to business opportunities, making the development process a dynamic and iterative activity. The data and features are available in versions (releases). Each new version starts the process of eliciting new requirements for the next version.

Table 2.3: Differences between BI/DW and transactional systems

	Applications	
	BI/DW	**Transactional**
Orientation / Direction	Business opportunities	Business needs
Implementation	Support organizational strategies for decision making	Support departmental activities
Requirements	Strategic information	Operational functions
Analysis	About business	About system

Kimball (KIMBALL, ROSS, *et al.*, 2008) and Inmon (INMON, 2005) introduced two different approaches to build a BI/DW solution. The first is the bottom-up strategy, where each department vision – also called a Data Mart (DM) – is built and, then, integrated, forming an Enterprise DW (EDW). The second is the top-down, where the whole business of the company is mapped and designed to build the EDW, after the DMs are derived from it.

Moss introduced a BI project lifecycle (MOSS, 2003), defining the process to build a BI solution, illustrated in Figure 2.10. There is an adequacy of this methodology to Kimball's (KIMBALL e ROSS, 2013) and Malinowski's (MALINOWSKI e ZIMÁNYI, 2009) approaches. In addition, it proposes metadata repository construction during the project. It presents a balanced approach, considering complexity and practice. Its acceptance in academic and computer industry solutions is high. Each activity is set to a specific phase.

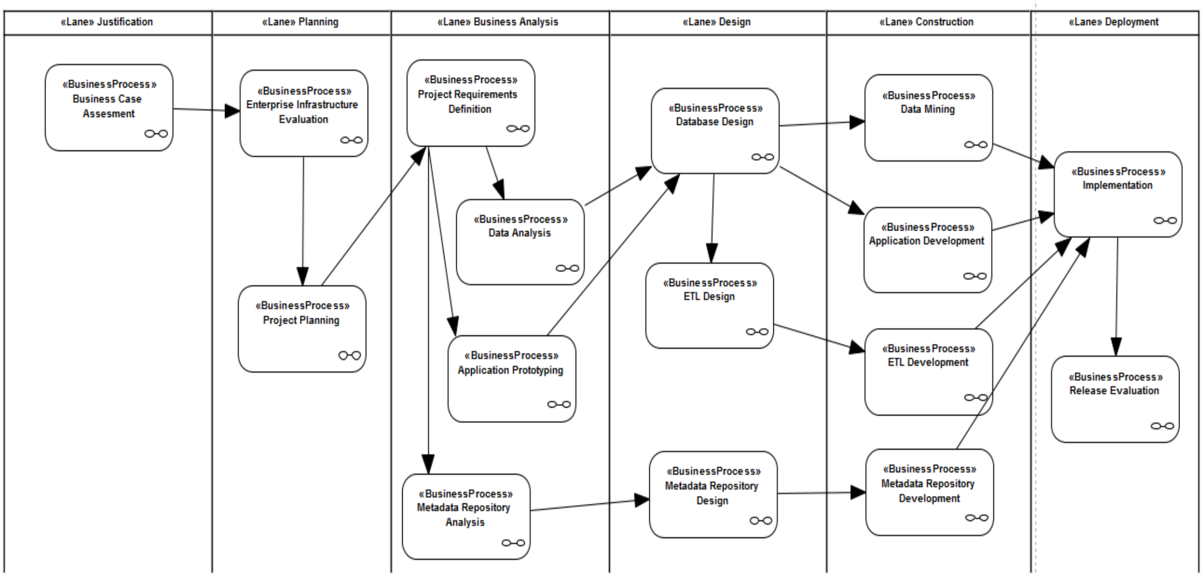

«Lane» Justification
«Lane» Planning
«Lane» Business Analysis
«Lane» Design
«Lane» Construction
«Lane» Deployment
«BusinessProcess» Business Case Assesment
«BusinessProcess» Enterprise Infrastructure Evaluation
«BusinessProcess» Project Planning
«BusinessProcess» Project Requirements Definition
«BusinessProcess» Data Analysis
«BusinessProcess» Application Prototyping
«BusinessProcess» Metadata Repository Analysis
«BusinessProcess» Database Design
«BusinessProcess» ETL Design
«BusinessProcess» Metadata Repository Design
«BusinessProcess» Data Mining
«BusinessProcess» Application Development
«BusinessProcess» ETL Development
«BusinessProcess» Metadata Repository Development
«BusinessProcess» Implementation
«BusinessProcess» Release Evaluation

Figure 2.10: Moss's BI/DW process lifecycle methodology

The necessity of coping with unstructured data in BI/DW solutions is fundamental for business analytics nowadays. According to a TDWI research in 2007 (RUSSOM, 2007), it is estimated that more than 31% of useful information to business is in unstructured format. However, with the advent of big data and cloud computing technologies in the last years, it is believed that this rate is rising exponentially. Even so, almost all BI environments, supported by EDWs or interlinked DMs, are based on structured data coming from relational DBs that store operational data. Analyzing and exploring data from heterogeneous natures, jointly, can enhance the analytical applications potential offered to decision makers of these organizations (INMON, STRAUSS e NEUSHLOSS, 2008).

Many approaches to integrate text through relational DBs for analytical solutions were proposed, such as (GROSSMAN, FRIEDER, *et al.*, 1997) (LEE, GROSSMAN, *et al.*, 2000) (MCCABE, LEE, *et al.*, 2000) (LEE, GROSSMAN e ORLANDIC, 2002) (CHRISMENT, DOUSSET e

ALAUX, 2003) (ROY, MUKESH, *et al.*, 2005) (TSENG e CHOU, 2006) (RAVAT, TESTE e TOURNIER, 2007) (LIN, DING, *et al.*, 2008) (BHIDE, CHAKRAVARTHY, *et al.*, 2008) (MOREIRA, CORDEIRO e CAMPOS, 2009) (ZHANG, ZHAI, *et al.*, 2009) (THOLLOT, BRAUER, *et al.*, 2010) (BARCZYNSKI, BRAUER, *et al.*, 2010) (GARCIA-ALVARADO e ORDONEZ, 2010) (HEUSELER, 2010) (PARK e SONG, 2011) (MOYA, KUDAMA, *et al.*, 2011) (SAIAS, QUARESMA, *et al.*, 2012) (NEVES, 2012) (MOREIRA, CORDEIRO e CAMPOS, 2013). Most of them implement Information Retrieval (IR) and Natural Language Processing (NLP) techniques.

In the following sections we describe each phase issues related to unstructured data needs passing through Moss methodology (MOSS, 2003) phases, discussing possible adaptations, specifically for what affects the MD Design task.

2.3.1 Justification

The "justification for a BI decision-support initiative must always be business-driven and not technology-driven" (MOSS, 2003), so the business drivers and requirements are always the motivator of a BI/DW project. For this reason a BI/DW project cannot be motivated only because of technology challenges. However, the business analysis issues must take into account the textual information sources, once it can provide the data necessary for the high-level requirements, possibly serving as the data sources to attend them. The process must consider the information systems that hold the unstructured data sources, such as Content Management Systems (CMS). In many times, this kind of software can provide important information in text format, through intranets, collaboration (wiki) websites, among others. Other types of unstructured data sources can be verified too, such as e-mails, contracts, corporative procedures, reports, presentations, medical records and other types of documents. In addition, Semantic Web (SW) features has brought, from outside of organizations, important textual data to their processes. Therefore, searching information needed in the web is advisable, even if it has direct costs involved to its acquisition (affecting the cost-benefit analysis). For example, if an organization aims to analyze how energy consumption of a certain region occurs, maybe the data to respond this is available in the internet freely by a government website or costly by specific companies who own the information.

The source data quality analysis of structured data is already a challenge because estimating DBs cleanness and consistency is a hard task. Textual information sources increase its complexity and the cost-benefits analysis have to be improved to consider them. Analyzing how data is unstructured is a necessity, i.e. if they follow a formal natural language structure or if they are totally free. Examples of formal written documents are contracts and corporative procedures, whilst tweets and social media posts are many times written without grammar and other language constructs concerns. The detailed risk assessment matrix should include a line with the degree of how the unstructured data sources quality is measured. An example would be: high formal language used (green), formal and informal language used (yellow) and informal language (red).

Sometimes, to structure free texts generated from business processes can have a better cost-benefit then to use the unstructured data in the BI/DW solution. For example, in electrical security domain, the information system that supports the process of registering electrical disturbances provides a textual field through a form to describe the sequential occurrences of a disturbance event. If the system form changes to have a list of classified (naturally normalized) entities representing the sequential occurrences, instead of free text, then using conventional structured data can be costless to the organization. Moreover, maintenance like that can improve the data assertiveness on representing the steps occurred in the business process, providing better data quality. Therefore, the cost calculation for changing transactional information systems (within processes) or using textual information “as-is” should be made.

As indicated by Moss, in this earlier stage, the BI/DW analyst should start to draft a high-level domain model based on the top-level requirements and main concepts captured. As it is presented in section 4, the application of FO in domain representations can increase the semantic captured from real world. To support this activity, the subject experts should be encouraged to point the structured and unstructured data sources involved to the business requirements, revising the high-level model. Corpora of these documents can be collected and analyzed by the designer. For this task, NLP tools can help for morphological and lexical analysis. An example is VisualText[4] software, which structures the content,

[4] http://www.textanalysis.com/

correlating terms from the text. Moreover, it provides entities recognition from names, locations, dates, and other atomic features of text. Indexing, filtering, mining the text, test grading, summarization, automated coding, natural language query and dissemination are capabilities presented in this type of software. The morphologic analysis can provide a degree of how risky is to use each unstructured data source and how they can be accessed and used by the analytical solution. Indicating how BI/DW project can benefit by using them is fundamental and can be done by demonstrating revenue or profit increases, customer satisfaction improvements or market share gain.

The risk assessment is one of the most important justification components, once it defines the risk conditions that can impact a BI/DW project, such as technology used, which can be incompatible to the unstructured data sources. The complexity of the project aspects, such as textual different formats, should be considered. The integration risk is a variable hard to estimate when it is used unstructured data sources, once it is necessary to check for potential linkages between them and the structured data sources. Having same entities represented in both sources is common and, at this point of the project, making a high level relation between them is necessary. For example, in a BI/DW solution for a medical clinic, the existence of a direct relation between patients account system (as structured data source) and medical records of patients (as unstructured data source) is probable. In this case, the patient entity is represented in both sources and the relation could be done by conforming them.

The business case assessment report is the artefact produced from the justification phase. It contains the strategic business goals of the organization, the BI/DW project goals, the business opportunity, cost-benefit analysis results, risk assessment and improvements considerations for transactional data. The process to build this document is illustrated in Figure 2.11 bellow, which also shows the particular steps to regard unstructured data sources.

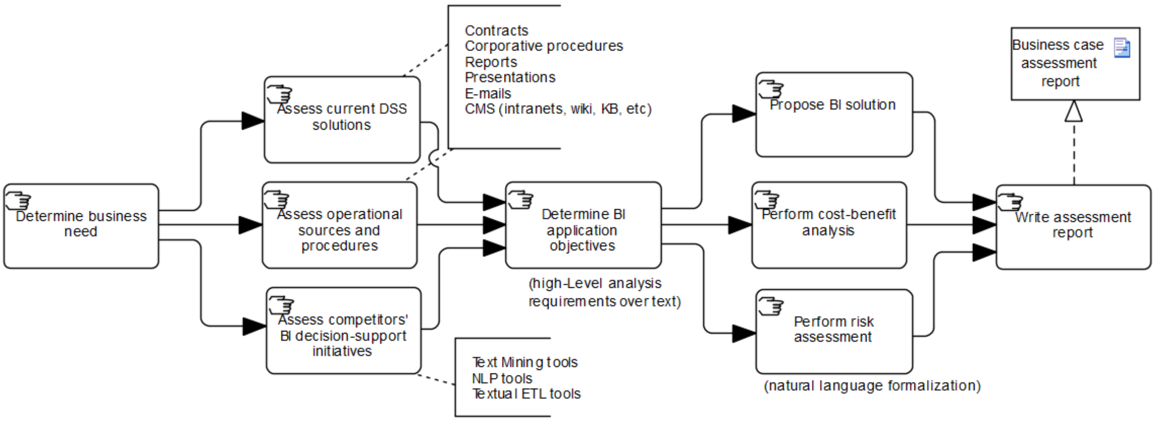

Contracts
Corporative procedures
Reports
Presentations
E-mails
CMS (intranets, wiki, KB, etc)
Business case assessment report
Assess current DSS solutions
Propose BI solution
Determine business need
Assess operational sources and procedures
Determine BI application objectives
Perform cost-benefit analysis
Write assessment report
(high-Level analysis requirements over text)
Assess competitors' BI decision-support initiatives
Perform risk assessment
Text Mining tools
NLP tools
Textual ETL tools
(natural language formalization)

Figure 2.11: Business case assessment tasks considering unstructured data

As shown in Figure 2.11, assessing current DSS solutions, if exists, is important to estimate the technological and complexity risks. The same is applied in assessing operational sources and procedures, as already discussed in this section. Some BI/DW vendors already deal with unstructured data sources through the so called Textual ETL (NESAVICH e INMON, 2007), a similar process of conventional ETL which uses NLP and IR capabilities, such as text mining, terms occurrences indexing, relevant document selection, orthographic correction, case sensitive and stop words removal, tokenizing, stemming, synonymous and homographs resolution, categorization and classification (ALMEIDA e SILVA, 2009). Regarding NLP and IR techniques, there are some APIs mature enough to provide their implementations and should be considered too, such as Lucene[5] and Stanford NLP[6] libraries.

An example of Textual ETL is Forest Rim[7] software. Therefore, checking their particularities, successes and failures in different domains can be helpful. Testing and evaluating POCs can be helpful to justify the use of them. This way it is possible to determine the necessary supplemental code to be written. Furthermore, the analytical solution can consider particular types of analysis, such as sentiment analysis over customer opinions (as free text) in social networks medias (ROY, MUKESH, *et al.*, 2005) (SANTOS, 2013). Therefore, these kinds of techniques may be investigated to be considered in the BI/DW project justification phase and reflected in the business case assessment report.

[5] https://lucene.apache.org/core/

[6] http://www-nlp.stanford.edu/

[7] http://textualetl.com/

2.3.2 Planning

The planning phase is divided in two activities: the enterprise infrastructure evaluation and the project planning. The former is the activity of checking technical and nontechnical issues. Enterprise architecture standards, hardware, network, DBMSs and other specific tools, such as OLAP and metadata repository, are revised for the BI project. The latter is the task in which detailed information from the project is given, considering the goals and objectives, risks, constraints and scope issues. The plan of the project includes the resources, material and human, and the activities, which are divided in tasks and subtasks with their dependencies set.

During the technical enterprise infrastructure assessment the hardware platform requirements have extreme importance, especially if unstructured data sources are considered for the BI/DW project, once their volumes can be huge. Data volume, variety and velocity needs are the three pillars of big data research, providing technologies as NOSQL DBs and Mapreduce algorithm, which treat data storage and processing, respectively. In addition, cloud computing must be considered when planning BI/DW solutions. Platforms for Cloud Warehousing were proposed (MA, SCHEWE, *et al.*, 2011) and vendors sell services for BI/DW infrastructure, such as Amazon Web Services (AWS) Redshift[8], where different types of data sources, such as SW data (XML/RDF/OWL), text, geo-related and sensor data can be explored.

Therefore, estimating the volume of unstructured data which will be used is crucial to the project success. However, at this early stage it is only possible to derive a calculation for an order of magnitude of the data size. The classical approach for storage estimation calculus is checking the average size of each data source, sum them and duplicate the result. For example, if contracts are considered for user analysis and they are stored in network directories, supposing their size sum is 500GB, than the estimated size needed would be 1000GB. In addition, volume increase tax should also be calculated (increase of MB/day expected).

Middleware components are evaluated in this phase and the ability to access unstructured data sources should be checked, once they can provide a bridge to textual data

[8] http://aws.amazon.com/redshift/

extraction. An example is the ODBC provider, which can support the connection to different types of text files, such as *pdf* and *doc* files. Furthermore, web crawlers are powerful tools acting as middleware on extracting texts from web-based interfaces. Regarding the choice of OLAP tools when unstructured data is used, open APIs can represent a better adequacy to for joint exploration and quantitative indicators to reach specific documents.

Nontechnical infrastructure components should also be evaluated for the assessment, such as CASE tools. They are fundamental to support the business analysis and design phases. As it is presented in chapter 4, CASE tools can provide well-founded modelling languages to best represent the semantic from real world. Coupled to them, the verification and validation tools play an important role to increase the maturity of the models, making the BI/DW software construction more assertive to represent the domain observed. In addition, metadata repositories provide navigating capabilities through metadata from different types of data sources, such as DBMSs, information systems, business glossaries, organization charts, among others. The metadata of unstructured data sources, such as official procedures and reports, should also be scanned and available, gathered with the other technical and business information. More expressive representation of metadata concepts can be delivered through ontology representations and SW standards, such as RDF, OWL and SPARQL (NGUYEN THANH BINH e MANGISENGI, 2001) (TORRES, LORENZATTI, *et al.*, 2011) (KUBIK e IWANIAK, 2012), so their exploration ability can evaluate user experience. They can also be improved by reasoning mechanisms presented in particular tools, such as FaCT++[9], Pellet[10] and RacerPro[11] (DENTLER, CORNET, *et al.*, 2011).

Patterns and guidelines from enterprise standards are essential to the success of the development of DBs, ETLs and OLAP cubes. Test, reconciliation, security and SLAs (Service-Level Agreements) are specific subjects that should have their procedures adapted for structured and unstructured data sources. For example, tests should consider a certain volume of documents corpora during the quality assurance phase. Depending on how large are the textual data sources, this can directly impact the analysis tests and if there are not

[9] http://owl.man.ac.uk/factplusplus/

[10] http://clarkparsia.com/pellet/

[11] http://www.racer-systems.com/products/racerpro/

standardized procedures then the results can be depreciated. Those aspects should be analyzed and presented by a full infrastructure assessment report.

The project planning itself should consider unstructured data sources when setting the project scope and deliverables definition. For example, BI/DW solutions scope is traditionally measured by the number of data elements from the sources. Therefore, it should include textual data elements, such as different documents that could be used and their main concepts. Project risks must include the lack of textual tools necessary and the unstructured data limitations, such as incompleteness of information, which will affect directly the cost-benefit analysis task (as cited before). Moreover, hybrid estimating techniques can be taken to measure the effort to integrate textual information, merging projects historical patterns, preview experiences and formula-based over worst, average and best estimates. Critical path method can be applied in all tasks.

Skills management is one of the main parts of the project planning task. Specialized human resources are a key point to the success on the application of unstructured data in BI/DW project lifecycle. The BI analysts need specific capacities to analyze text from linguistic point of view and know-how to manage IR, NLP and Text Mining tools. In addition, human resources should be specialized in domain engineering, being able to work with ontologies, rules formalizations by first order logic for business rules (e.g. OCL) and familiarity with modelling and ontological languages, such as UML and OntoUML.

2.3.3 Business analysis

Business requirements elicitation is an important task which should have the application leader developer time prioritized on it. Functional requirements reflect the types of information the business need, it can origin from both structured and unstructured data sources. In this aspect, business executives are essential during the requirements interviews because of their visionary characteristics. IT managers are important stakeholders too and, commonly, have wide systemic view over the information required and their sources.

Historical requirements can be affected by textual information volume, limiting the DW historic capacity. The security requirements must deal with permissions and roles over documents. For example, the security policy for analyzes of e-mails contents should follow several rules, preventing individual rights. Data requirements elicitation needs to provide the location of the unstructured data sources, how clean they are and if they can be summarized. The definition of the most appropriate sources and their data quality by a tolerance level for dirty data is important. For example, in the health domain, if specific reports regarding heart frequency are required and this information is only available in patients medical records (under textual format), then it is necessary to establish a poor-quality level that can be accepted. In this case it would be necessary to check if the heart frequency property is described in each medical record, if it follows a pattern and their measure ranges.

An application requirements document is written to describe all these requirements, including the main reports over textual information, which includes graphical representation needs that can be achieved by OLAP tools. The SLAs for the query response time are also included in this document. However, the requirement elicitation task is an interactive process and is enriched by the data analysis phase activities. Once the requirements are represented, e.g. listed in this document, the next step is to expand the high-level logical data model built in previews phases. The newly discovered concepts and relations from both structured and unstructured data should be added to this conceptual model.

Data analysis activity considers if there are multiple sources, if they are internal or external of the organization and the models refinement and expansion. In this phase, the conceptual domain model can be supported by FO formalisms and conceptual MD schemas can be derived from it. Moreover, business rules can be described in the conceptual model by the cardinality of the relationships and the possible constraints existent. The MD concepts resulted are the entry to the DB and ETL design activities. This is the main subject of this work and is detailed in chapter 4.

The application prototyping task is part of the business analysis phase and occurs parallel to the data analysis activity. It serves to validate the project requirements and to evaluate risky issues. Cloud computing environment can provide an interesting cost-benefit on prototyping with greater data volume. Nevertheless, the same best practices over prototyping should be considered, such as limitation of the scope, understanding of data sources, usability tests by OLAP tools and involvement with business people. A key issue in prototyping is the prototype charter, which will set the structured and unstructured data to be used, the hardware and software platforms to deal with them and the prototype measures of success. Furthermore, it also points the reports and queries to be designed and the work necessary. For example, through a proof-of-concept (POC) it is possible to explore the implementation risks of extracting a set of documents by a time frame, applying determined NLP techniques, such as clustering, classification and entity recognition. In addition, demonstrating how reports over textual information will be delivered by visual-design prototype over an OLAP tool can be made.

Metadata repository analysis is a task that supports the other business analysis tasks and should have some adjustments to consider unstructured data sources. First of all, metadata repository initiatives already suffer from budget, political and other technical difficulties. Moreover, BI/DW developers in majority do not use a specific tool to manage the metadata, leaving it in spreadsheets and text documents, which is hard to maintain up to date and has several problems for integrated access and analysis. As cited before, ontology based approaches for metadata repositories can be used.

2.3.4 Design

Moss methodology does not obligate MD design, but considers it because it provides data structures to aggregate and summarize large amounts of information in a practical way for the decision maker analysis requirements in a BI/DW solution. Thus, in this phase the star and snowflake schemas for both logical and physical DB designs are built. From the prototype built in business analysis phase, the DB designer should review reports and queries to best define the DB design schema. Data aggregation and summarization requirements are set to a lowest level. Textual measures can be non numeric and non additive (RAVAT, TESTE e TOURNIER, 2007). In addition, keyword aggregation functions provide navigability through terms by implementing proximity function between them. Therefore, the DB designer should determine the most appropriate DB design based on textual drilling down and rolling up requirements. Moreover, thinking in performance is crucial in this activity and distributed computation capability may be considered.

In DB physical design, data storing, partitioning, clustering and indexing for target BI/DW DB is usual. Furthermore, parallel query execution, backup and recovery activities are maintenance procedures which should be addressed. Because of size increasing due to unstructured data, tuning queries and monitoring the BI application are essential for better performance.

Some authors call the ETL design as conceptual ETL (ROMERO, SIMITSIS e ABELLÓ, 2011), a high level description of where data should be extracted and loaded; and how the transformations should occur from an abstract point of view. Traditionally, a source-to-target mapping document in a spreadsheet (as a matrix) is built to show data origins, targets and transformation specifications. ETL design should be adapted to support unstructured data, once it will lead with both data source types when extracting and transforming them. As cited in planning phase, IR and NLP techniques and components should be used as a base for Textual ETL. Thereafter, both conventional and textual ETL data flow can be designed, componentizing the possible parallel tasks. Performance issues are considered, by analyzing the possibilities to turn off referential integrities among tables which store the indexed data. Data should be manipulated in a staging area and quality metrics to analyze integrity errors should be defined. Parallel to DB and ETL design, metadata repository is designed, regarding concepts represented in tools such as CASE, DBMS, ETL, data-cleansing, OLAP and data mining tools. The meta-meta model can be designed as ontologies, as cited in planning phase.

2.3.5 Construction and deployment

The implementation of the ETL processes, related DBs and OLAP are made in this phase. There are numerous issues regarding construction and deployment phases of a BI/DW solution, but they will not be considered in the scope of this work. To reflect the use of unstructured data, Textual ETL should be applied and OLAP tools adapted, as cited before. ETL testing is a subject with great importance and may consider unit, regression, quality assurance, integration, acceptance and performance tests. As a result, all deliverables must be prepared, such as ETL and application packages, the metadata repository, the target BI DBs and all operation and guide documentations.

Notice the equivalence of the MD design approaches revised in section 2.2 with business analysis, design and construction phases of Moss methodology, specifically in requirements definition, data analysis, DB design and ETL design and development activities. Therefore, in this section, it was described the high level issues regarding the use of unstructured data sources (as text) in a BI/DW solution, serving as a basic theory (and related works) for the experimentation described in chapter 5. However, the specific gap in BI/DW project lifecycle treated in this dissertation is the lack of semantic expressiveness when MD modellers choose the appropriate MD concepts to represent the domain during the MD design activity. In next chapter we describe how related works deals with this problem by applying ontological-based approaches.

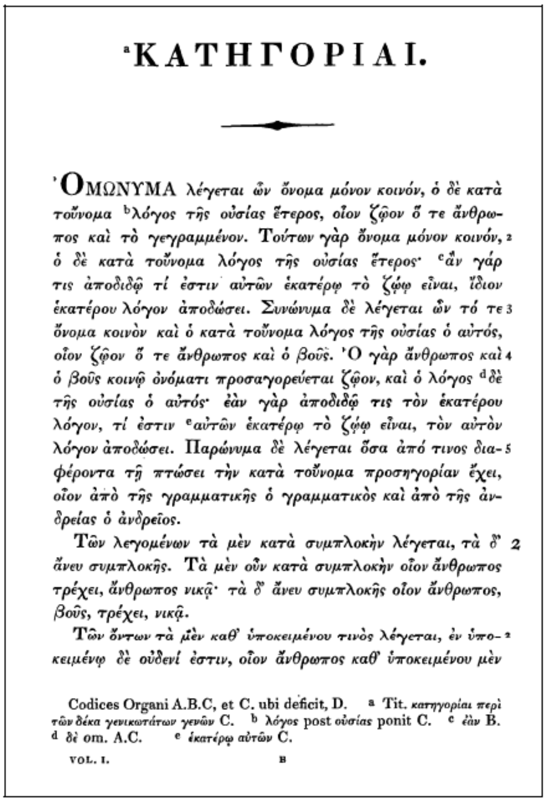
3 Ontologies

Representations of portions of the reality are necessary to understand, communicate and reason about these representations. A model is an abstraction and simplification of reality and, more than that, a description or representation of something from a specific point of view and with a specific purpose. The Universe of Discourse (UofD) is a real world part being modelled, the domain. Models abstract irrelevant details and enable more efficient analysis of past, current and future states of a UofD (HODGES, 1993).

Ontology is a representation schema, i.e. a model that describes the formal conceptualization of a UofD, implemented through schematic diagrams. In Software Engineering (SE) area it is commonly built as UML class diagrams. In Artificial Intelligence (AI) area it is commonly built as semantic networks and in Database (DB) area as ER diagrams. All these models seek to represent entities, relationships, properties, rules and restrictions of areas. It can be considered formal when it is machine-processable, enabling automated reasoning by the semantics described in formal logic.

The term ontology – "onto" (to be) and "ology" (study) – was created in the XVII century, but its concept was systematically built on western ancient philosophy and it is a sub-area of metaphysics that deals with the nature and existence of things. The categories classifications are present in the ideas of Plato. In Categories (Figure 3.1), Aristotle discusses the common and philosophic vocabularies, based in words and propositions. It is a discipline that aims to study the most general features of the reality, differing from other areas of science, such as physics, chemistry and biology, dealing with specific concepts of their domains. It is the philosophical study of the nature of being, of change, of existence, of reality and their relationships.

Aristotle invented the ontology the way we can still recognize today, the study of how the things are organized in the universe as types or kinds or universals or categories. It deals with questions such as what entities exist or can exist and how they can be hierarchically grouped and classified according to their similarities and differences. Basically, the categories are the universals in the highest level in a hierarchical structure that contains these other concepts, defined as top level categories. For Aristotle these categories are intelligible, i.e. we can formulate definitions about them in a way that we can understand them.



[a]ΚΑΤΗΓΟΡΙΑΙ.

Ὁμώνυμα λέγεται ὧν ὄνομα μόνον κοινόν, ὁ δὲ κατὰ
τοὔνομα [b]λόγος τῆς οὐσίας ἕτερος, οἷον ζῷον ὅ τε ἄνθρω-
πος καὶ τὸ γεγραμμένον. Τούτων γὰρ ὄνομα μόνον κοινόν, 2
ὁ δὲ κατὰ τοὔνομα λόγος τῆς οὐσίας ἕτερος· [c]ἂν γάρ
τις ἀποδιδῷ τί ἐστιν αὐτῶν ἑκατέρῳ τὸ ζῴῳ εἶναι, ἴδιον
ἑκατέρου λόγον ἀποδώσει. Συνώνυμα δὲ λέγεται ὧν τό τε 3
ὄνομα κοινὸν καὶ ὁ κατὰ τοὔνομα λόγος τῆς οὐσίας ὁ αὐτός,
οἷον ζῷον ὅ τε ἄνθρωπος καὶ ὁ βοῦς. Ὁ γὰρ ἄνθρωπος καὶ 4
ὁ βοῦς κοινῷ ὀνόματι προσαγορεύεται ζῷον, καὶ ὁ λόγος [d]δὲ
τῆς οὐσίας ὁ αὐτός· ἐὰν γὰρ ἀποδιδῷ τις τὸν ἑκατέρου
λόγον, τί ἐστιν [e]αὐτῶν ἑκατέρῳ τὸ ζῴῳ εἶναι, τὸν αὐτὸν
λόγον ἀποδώσει. Παρώνυμα δὲ λέγεται ὅσα ἀπό τινος δια- 5
φέροντα τῇ πτώσει τὴν κατὰ τοὔνομα προσηγορίαν ἔχει,
οἷον ἀπὸ τῆς γραμματικῆς ὁ γραμματικὸς καὶ ἀπὸ τῆς ἀν-
δρείας ὁ ἀνδρεῖος.

Τῶν λεγομένων τὰ μὲν κατὰ συμπλοκὴν λέγεται, τὰ δ' 2
ἄνευ συμπλοκῆς. Τὰ μὲν οὖν κατὰ συμπλοκὴν οἷον ἄνθρωπος
τρέχει, ἄνθρωπος νικᾷ· τὰ δ' ἄνευ συμπλοκῆς οἷον ἄνθρωπος,
βοῦς, τρέχει, νικᾷ.

Τῶν ὄντων τὰ μὲν καθ' ὑποκειμένου τινὸς λέγεται, ἐν ὑπο- 2
κειμένῳ δὲ οὐδενί ἐστιν, οἷον ἄνθρωπος καθ' ὑποκειμένου μὲν

Codices Organi A.B.C, et C. ubi deficit, D. [a] Tit. κατηγορίαι περὶ τῶν δέκα γενικωτάτων γενῶν C. [b] λόγος post οὐσίας ponit C. [c] ἐὰν B. [d] δὲ om. A.C. [e] ἑκατέρῳ αὐτῶν C.

VOL. I. B

Figure 3.1: First page of Categories, Aristotle, 3th century BC

Metaphysics seeks to answer questions like “what is the existence of a being?”, “does essence precede existence?”, “if any, what are the fundamental entities?”, “all entities are objects?”, “what is a physical object?”, “what differentiates objects and events as they relate?”, “what is the identity of an object?”, “an object is equal to the sum of others?”, “how objects relate themselves with their properties?”, among others. To try to answer them, concepts such as universal, individual, substance, accident, abstract, concrete existence, persistence, change and classification were used in formulating hypotheses, e.g. “everything changes”.

The first time that ontologies were mentioned in the context of computer science, in 1967, was in the context of data processing: “This is an issue of ontology or the question of what exists" (MEALY, 1967). This work was about the data modelling formalisms and it discussed the distinction among the definitions of the real world as it really is, i.e. the ideas inherent in the human mind, and the symbols used in some type of storage, such as paper or computer. In the 70s the three-tier architecture – with conceptual, interfaces and implementation schemas – began to popularize. Logical models, such as relational and network, were most important in the SE area, with a large increase in proposals offering better semantic expressiveness of these models through conceptualization languages, such as ER, introduced by Peter Chen in 1976.

The Conceptual Modelling (CM) is the activity of describing formally some aspects of the physical and social world around us, for purposes of understanding and communication between humans and not machines, supported by structures and inferences grounded in psychology (MYLOPOULOS, BORGIDA, *et al.*, 1990). Its goals in computer science are domain learning, database schemas integration, problem solving for domain engineering, services interoperability, information correction and intelligent search in the SW. In CM, concepts are described oriented to specific domains and represented by entities, relationships, goals, actors, among others, being structured according to cognitive principles, such as generalization, aggregation and classification. Onotlogies are commonly used as a tool to support CM of domains, mainly by SW approaches. Analogous to traditional SE development, the process of Ontology Engineering is necessary, dealing with the identification of purpose, requirements specification, the ontology capture, formalization, reuse, integration, evaluation and documentation. The use of CM techniques based on ontological analysis for BI/DW solutions is common and is described as follows.

3.1 Ontologies and their role in BI/DW solutions

The application of ontologies in BI/DW has been increasing in the recent years. In (BERLANGA, ROMERO, *et al.*, 2011) it is introduced how SW technologies can be useful in nonconventional BI scenarios, discussing their advantages, disadvantages and application cases. SW languages such as RDF and OWL, founded in Description Logic (DL), can be used to define rules to be processed by reasoning services, assisting data aggregation. Semantic annotations can be applied in the construction of ETL conceptual data flows to deal with its structural and semantic heterogeneity. In MD modelling, the ontologies and reasoning services are applied to discover dimension hierarchies from sub-domains, integration concepts from different ontologies and MD schema production.

SW technology pattern RDF was also used for to support the definition of relations among the measures and the dimensions, i.e. the measure-dimension consistencies (NIEMI e NIINIMÄKI, 2010). The MD consistency definition is formalized to guarantee that the aggregation results are correct. The inputs are the aggregation function (e.g. SUM and MAX), type of dimension (temporal or non-temporal) and the measure. As a start point it uses an OLAP ontology, illustrated in Figure 3.2. By the execution of an RDF query (RDQL) the summarizability is checked.

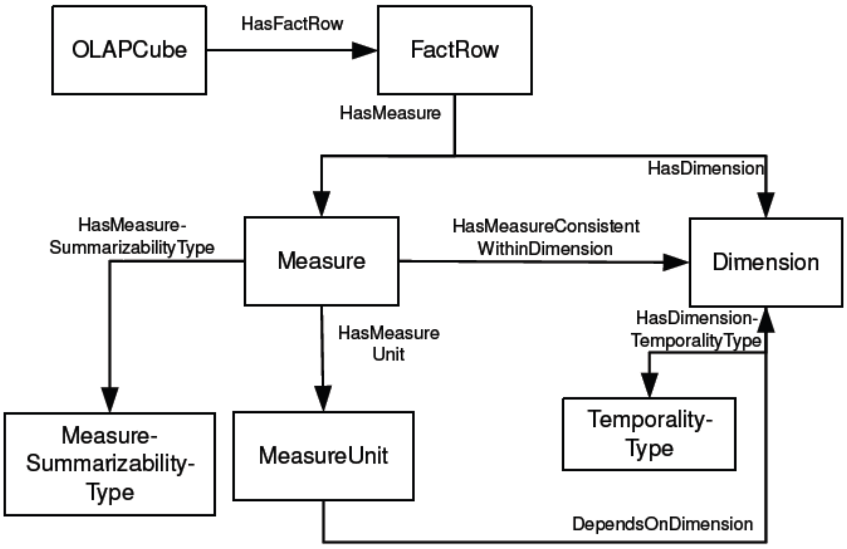


Figure 3.2: OLAP ontology describing OLAP concepts from (NIEMI e NIINIMÄKI, 2010)

Other approaches represent OLAP concepts through MD ontology, describing facts, dimensions, measures, attributes, hierarchies, their relations and specificities. Business entities are described in the domain ontology and considering the business measures and their meaning on the domain, such as a process indicator. A hybrid ontological structure was proposed in (CAO, ZHANG e LIU, 2006) to integrate BI components (including conceptual, analytical and physical views) over DW and Enterprise Information Systems (EIS). They were divided into four layers: (i) business ontologies, such as user profile and interface agents; (ii) DW ontology, containing OLAP and data mining concepts, metadata management and global vocabulary; (iii) EIS ontology, for billing and accounting systems; (iv) mapping and query parsing ontology as mediation layer.

As cited in section 2.2, to enforce the semantics in MD design, Romero et al. (ROMERO, 2010) proposed an approach to enforce the semantic expressiveness of conceptual modelling on BI/DW development, based on end-user requirements elicitation and hybrid MD design. The first step is to represent data sources as a domain ontology. In parallel, the end-user requirements are described as SQL queries over data sources, where the involved concepts are linked to the domain ontology. Then, it uses a supply-driven mechanism (AMDO) to derive MD concepts (facts, measures, dimensions, hierarchies and attributes) from existent functional dependencies in data sources, formalized through first order logic. It proposes a set of derivation rules to discover MD concepts. This is a trend in CM because "if we take into account the expressiveness of the algebras, they might be as semantically rich as conceptual models are" (ABELLÓ, 2002). The MD formalization through Descriptive Logic (DL) is not a new idea, it is an initiative introduced in DWQ project (JARKE, LENZERINI, *et al.*, 2003).

The GEM approach (ROMERO, SIMITSIS e ABELLÓ, 2011) operationalizes the whole process, automating the identification of potential MD concepts by analyzing the domain ontology and the semantic annotations, represented in OWL-DL. It filters the concepts by requirements (like SLAs) described in XML files. It also provides the design of the ETL conceptual representation as a data flow, implementation independent.

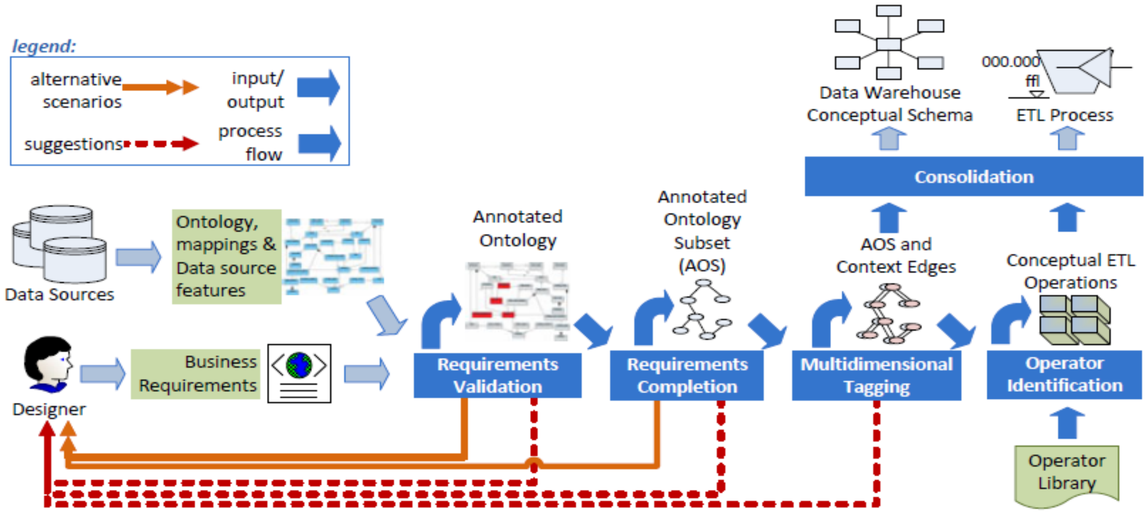


Figure 3.3: GEM - Generation of Conceptual MD and ETL from (ROMERO, SIMITSIS e ABELLÓ, 2011)

Figure 3.3 above illustrates the process that is divided in five phases, beginning in the requirements validation, where the ontology from the data sources are tagged by the functional requirements (f-reqs) concepts and the non-functional requirements (nf-reqs) are analyzed. The second phase is the requirements completion, where the ontology is simplified by ignoring the untagged structures and the unwelcome many-to-many associations. Afterwards, it produces an annotated ontology subset (AOS) to answer the f-reqs. The MD validation is the third phase, which begins by validating untagged concepts and tagging those that could play dimensional or factual roles; then, the rest of the concepts are analyzed and schemas are proposed following MD normal forms. The fourth phase is the operation identification, where the ETL operations are identified by the annotations done in previews phases. Additional information about surrogate keys and slowly changing dimensions are considered. The last phase is the conciliation, where the ETL data flow and the unique conceptual MD schema are merged with the AOS and the ETL operations.

GEM's architecture has been evolving for the last years and a tool was built to support the whole process. In addition, supporting evolution of DW design is under construction as add-ins for both conceptual MD models and ETL processes. ORE module (JOVANOVIC, ROMERO, *et al.*, 2014) was proposed to consider the inevitable complexity of frequent changes in MD design. It integrates each new analytical requirement with the existing MD schemas in an iterative and incremental process. The CoAl module is responsible for integrating the existing ETL flows to these new requirements, producing an unified ETL process, coping with tuning and optimizing both MD and ETL designs.

The representation of MD models as OWL-DL ontologies through a set of transformation rules (MD to OWL concepts) were proposed in (PRAT, AKOKA e WATTIAU, 2012). Examples of those mappings are: *Dimension* and *Fact* as Class, *Hierarchy* as Class linked by Object Property, *Attribute* and *Measure* as a Data Type Property. It also defines the OWL object properties for measures: summable, averageable, maxable, minable and countable. An implementation using Protegé[12] was made considering a case of spatiotemporal DW, as the presented in Advanced DW (MALINOWSKI e ZIMÁNYI, 2009), revised in section 2.2, and a set of transformation rules specializing this case as axioms and

[12] http://protege.stanford.edu/

facts in description logic. The reasoning capabilities of OWL-DL were explored to infer at the classes, properties and instances levels, such as consistency checking, class subsumption and equivalences (explicit and implicit).

Another work deals with the problem of integrating heterogeneous data sources for OLAP operations: the schema inconsistency that may happen (SHAH, TSAI, *et al.*, 2009). It presents the semantic conflicts in heterogeneous data cubes and a survey of the use of SW technologies in DW, also discussing the application of distributed warehouse data model through column-store architecture for large volume of data. It treats the gap between conceptual model and MD implementations of data cubes by proposing OLAP features supported by semantic annotations. The approach starts by mapping the MD schema using the OLAP ontology illustrated in Figure 3.4. Afterwards, a set of semantic services, implemented in Jess[13] and FuzzyJ[14] (Java APIs for rules execution and fuzzy logic, respectively), provide the pointers to different DBs and unique query interface. In energy sensor data application it uses reasoning capabilities to define rules to calculate specific measures, such as average energy consumption. The framework could enhance the reusability of existing data cubes and composed to others to provide an integrated OLAP tool.

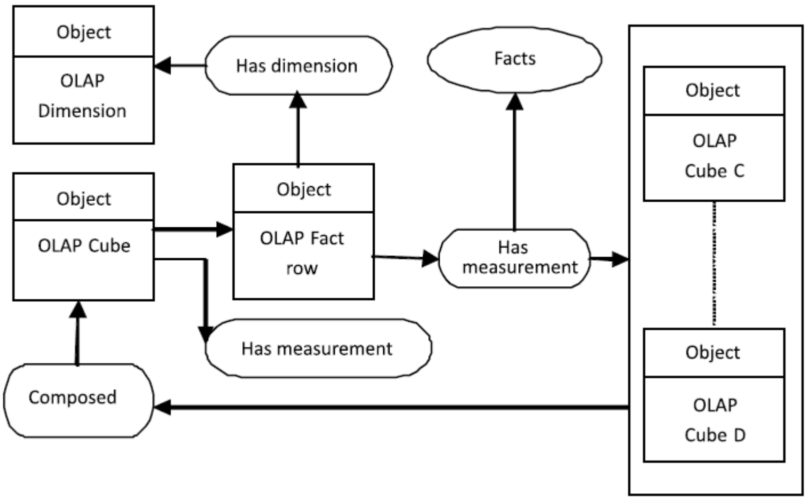


Figure 3.4: Composite OLAP cube ontology in OWL from (SHAH, TSAI, *et al.*, 2009)

[13] http://www.jessrules.com/

[14] http://rochard.github.io/FuzzyJ/

The similarity in using ontologies on the revised works is the mapping from MD schemas to SW standards (RDF or OWL). They use a MD metamodel ontology as a basis to link existing MD models. However, all those works do not consider the cognitive process in MD modelling, assuming that the MD schema designed by humans is always semantically correct or an automatic process can substitute them. This works that automate the MD design are mostly related to the analysis of functional dependencies of the data sources and business requirements representations. Pardillo e Mazón (PARDILLO e MAZÓN, 2011) discusses the use of ontologies for the design approaches of BI/DW solutions based in MD schemas. It brings a set of shortcomings pointing the MD design needs and how each one can be addressed by ontology applications. The adoption of Foundational Ontology (FO) is mentioned for semantic-aware summarizability of measure aggregations to treat the lack of ontological foundation in the measures and constraints representations, which miss formal semantics. Structural (*endurant*) and temporal (*perdurant*) concepts from FO can be used to classify the domain concepts and MD models can be logically derived from the ontologies. The semantic annotation of MD models with FO concepts is called "interpretation mapping", where the process of assigning ontological meaning is made. Then, ontological reasoners can be used to infer concepts such as facts and formalize summarizability constraints. The motivation is set as follows:

> "The problem is that, without an ontological foundation (...) there is no possibility of being sure that these frameworks are complete or formally stated. Simply, there is only tacit knowledge, not formal semantics (two readers may interpret different things). In this way, ontologies enable us to gain better understanding about the very nature of summarizability constraints, reasoning about why, when, and how to hold it. One first issue to deal with is the role that time dimension plays in the summarizability constraints. (...) time is a (probably, the) key dimension in a data warehouse. (...) Therefore, the first dimension to check additivity is usually this one. Here, these facts that are described temporally, may be understood as events, should be described by a foundational ontology such as DOLCE as *endurant* or *perdurant* particulars. The distinction between *endurants* and *perdurants* plays a prominent role in top-level ontologies and it can provide very useful knowledge about what is really managed by data warehouses and how OLAP analyses should be performed over them. " (PARDILLO e MAZÓN, 2011)

3.2 Foundational ontologies

The formal ontology discipline is an analogous reference to the formal logic, which contemplates logic formal structures, such as truth, validity and consistence (NIRENBURG e RASKIN, 2001). The term was firstly cited by the philosopher Edmund Husserl as: "Formal

Logics deals with the interconnections of truths (...) formal ontology deals with the interconnections of things, objects and properties, parts and wholes, relations and collectives". It is founded on the mathematical disciplines of mereology, i.e. the study of part-wholes, also the theory of dependence and topology. In Information Systems (IS) area ontology is used as a system of categories that is independent of language and state of things, in contrast to other areas, where the term is used to refer to a set of artefacts with different characteristics, such as catalogs, glossaries, thesauri and taxonomies. However, they have low complexity and do not have the potential for automatic reasoning and inference. In this work we use the definitions of the concepts of conceptualization, language, logic models and intentional models, as in (GUIZZARDI, 2005) whose relationships are shown in Figure 3.5.

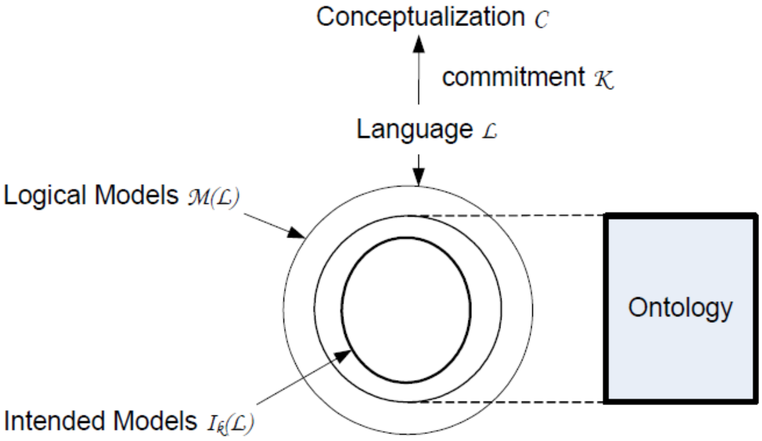

Conceptualization $\mathcal{C}$
commitment $\mathcal{K}$
Language $\mathcal{L}$
Logical Models $\mathcal{M}(\mathcal{L})$
Ontology
Intended Models $\mathcal{I}_{\mathcal{K}}(\mathcal{L})$

Figure 3.5: Ontology relations to conceptualization, language, logic and intended models from (GUIZZARDI, 2005)

The conceptualization *C* is an intentional structure <*W, D, R*>, where *W* is a nonempty set of possible worlds, *D* is the domain observed and *R* is the set of relationships among the concepts considered in domain *C*. There are intentional relations ϸ ε *R* where each *n*-ary relation is a function defined in it for possible worlds *W* for *n* tuples of individuals in *D*. The intended world structure *S* is the state of affairs characterization of world *w* ε *W* considered admissible for *C*. A logic model is described by a language *L* and defined in *S* with the interpretation function *I* of the elements of *D*. The intentional interpretation is a structure of *C* and the intentional function *J* to sign the elements of *D*. It can be seen as the Ullman triangle, depict in Figure 3.6, where a concept is represented by a symbol which refers to an item of reality, i.e. an abstracted concept. The ontological commitment $<C,J>$ of $<S,I>$ model with language L and vocabulary V is defined when (i) all s ε S, (ii) for any constant c then $I(c) = J(c)$ and (iii) there is world w which all predicate ƥ is mapped in J. The logical rendering is the theory T of a specification X in language L which is the description of X in first order logic.

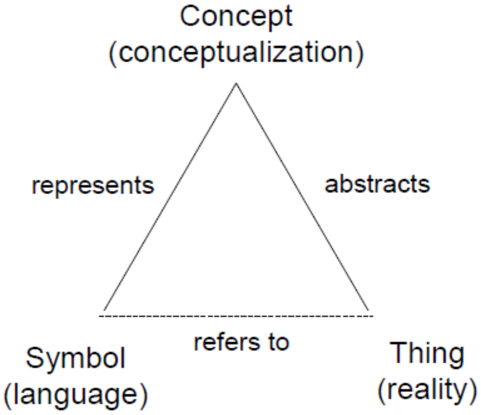


Figure 3.6: The intentional function described as Ullman triangle from (GUIZZARDI, 2005)

Domain Engineering (DE) is a subfield of Software Engineering (SE) and was motivated by the need to reduce the costs of software maintenance through the concept of reuse. In the early 2000s some proposals for the use of ontologies in DE were made, such as ODE approach (FALBO, NATALI, *et al.*, 2003). It proposes an ontology for software development process, showing how to reuse components from activities. The modeling language LINGO (GUIZZARDI, FALBO e PEREIRA FILHO, 2001) was introduced to support the domain models construction based on ontology. A systematic approach to derive object-oriented structures from the ontology was proposed, its methodology comprises a number of techniques such as policies mapping, design patterns and rules formal translation.

A Foundational Ontology (FO) is a high-level category system based on philosophical concepts, a top-level ontology (GUARINO, 1995). The classification of these types of ontologies and levels were introduced in (GUARINO, 1998) and is illustrated in Figure 3.7. There are four types in three levels: the top-level ontologies, the domain and activity ontologies, and the application ontology. Top-level ontologies are the FOs. They describe the most general concepts, such space, time, matter, object, event and action, concepts independent of a domain or a particular problem. The domain ontologies describe the vocabulary related to a generic domain such as astronomy, medicine or genealogy, through specialization of the concepts of a top-level ontology. Activities ontologies describe tasks or activities from business processes, such as sales or operation of the electrical system, also specializing the concepts of the top-level ontologies. Application ontologies describe concepts that depend both on domain and activities ontologies, representing the role of domain entities while performing certain activities.

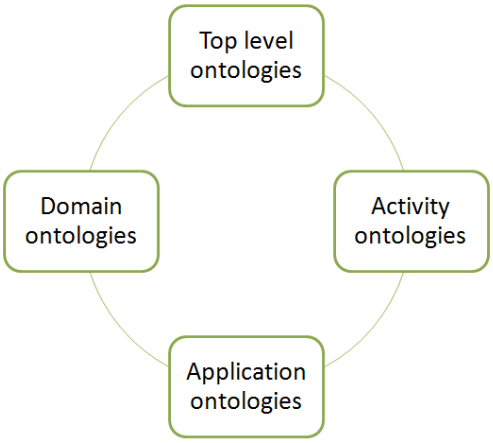


Figure 3.7: The intentional function described as Ullman triangle

The FO is the mechanism of formal ontology to define, through first order logic, the metaphysic questions, such as the general notions of types and instances; objects and their properties; the relation between identity and classification; the distinctions among sorts of types and their admissible relations; characteristics of part-whole relations; dependences; unity theories; events (*perdurants*) and other temporal and social aspects. As described in section 2.2, the semantic expressiveness is an essential property of models, and it is increased in FO due the philosophical constructs. The domain ontology is a representation of business concepts, whilst a FO is a well-founded and domain independent concepts set (GUIZZARDI, 2005). Among the subjects treated by FOs, some principles should be highlighted (ZAMBORLINI, 2011):

- ***Endurants***

Structural concepts, defined as types and instances, object properties, property-value spaces end different types of relations.

- ***Perdurants***

Temporal concepts of changes, defined as events, considering objects participations on them and their relations, changing the reality from a situation to another.

- **Identity**

It is the principle to distinguish and count individuals. E.g: `apple`, `car`, `person`.

- **Rigidity**

A concept is **Rigid** when the individual pertains to it while exists, e.g. `person`, `bird`, `stone`. A concept is **Anti-Rigid** when the individual pertains to it in a contingent way, i.e. can stop being it and continue existing, e.g. `student`, `athlete`. A concept is **Semi-Rigid** when it is optional for an individual and obligatory for other, e.g. `readable` is optional for `road sign` but obligatory to `academic paper`.

- **Dependency**

A concept has **Generic Dependence** to another if the dependency relation can change, but it always instantiates the **part-of** relation to an individual, e.g. an `equipment` (in energy grid domain) always instantiates the **part-of** relation to an `equipment container`, but can change one to another. A concept has **Specific Dependency** to another if the dependency relation can not change, i.e. is defined in terms of mutual disjointness. A concept has **Existential Dependency** to another if the dependency relation is strictly necessary for the individual exists, e.g. a `water reservoir` is always **part-of** a `drainage basin` (`watershed`), being fundamental to its existence.

- **Formal relations**

A relation is **Formal** when it only needs the existence of the individuals which it connects, e.g. `grater-then`, `company-owner`, `father-of`.

- **Material relations**

A relation is **Material** when it only exists depending in a connection entity, e.g. the relation `registered company` can only exists depending in the registration mediating the `company` and the `government`.

- **Meronymic relations**

A relation is **Meronymic** when it refers to a part-whole (**part-of**) property. It can be essential, inseparable or sharable.

3.2.1 Unified Foundational Ontology (UFO)

The application of formalisms through ontologies is a way to better understand the concepts and relations from a domain. The domain ontology built with those concepts categories is classified as reference ontology. The challenge in creating such a common language is to balance expressiveness with computational efficiency, once FO requires the use of high expressive formal language to its characterization. The Unified Foundational Ontology (UFO) was proposed to be a framework for modelling reference ontologies and it is evolving for the last years, being applied in several domains. The lightweight ontologies are versions of the reference ontologies that are created once there is a common sense among the users. They are focused on the required computational properties, and therefore do not prioritize adaptation to reality (GUIZZARDI, 2005). UML and OWL can be considered suitable languages for lightweight ontologies, while UFO purports to be a conceptual modelling language for the development of reference ontologies. Bellow it is described some general characteristics to better understand UFO:

- **Language levels**

The main levels of languages were classified (GUARINO, 1995) as: (i) **Logic** level has predicates as primitive constructs, the formalization as main characteristic and arbitrary interpretation. (ii) **Epistemological** level has concepts and roles as primitive constructs, the structure as main characteristic and its interpretation is also arbitrary. (iii) **Ontological** level has structural relations based in semantic, having the meaning as main characteristic and restricted interpretation. It concerns the nature of phenomena of interest, independently of particular information needs.

The **logic** level is also called **information** level, focused on structure of information about specific phenomena of interest, mainly driven for the development of information systems and DBs. Therefore, **ontological** level languages are the more appropriated to represent real world concepts in a clear and unambiguous way, but, they are not computationally efficient.

- **Ontology Engineering (OE)**

Analogous to SE and IS methodologies, the process distinguishes three levels of models: conceptual, project and implementation. In OE the first has the objective of building the domain reference ontology, to better represent the domain in an ontological level. The second has the objective of choosing the appropriate lightweight ontology for the implementation phase to maximize non-functional requirements. This last one has the objective of representing the domain reference ontology in epistemological or logic level to guarantee the computational characteristics needed.

- **Ontological languages**

The ontological languages are those set of patterns to support modelling, being divided in three levels mentioned above.

The UFO (GUIZZARDI, 2005) is a system of categories developed to support reference ontologies by using formalisms from the formal ontology, philosophical logic, linguistics and cognitive psychology. It was derived from other FO languages: GFO/GOL and OntoClean/DOLCE (GUARINO, 1998), both used on natural science and cognitive engineering. The UFO purpose is to unify those FOs, overcoming the limitations in the ability of capturing basic concepts from them and from CM languages, such as UML. Its formal characterization is based on axioms commonly found in modal logic, describing all admissible relations from a philosophical point of view. Its architecture is divided in three parts: UFO-A, the structural aspects; UFO-B, the temporal aspects; and UFO-C, the social aspects. Figure 3.8 depicts the main ontological subjects treat in each part of UFO, as revised in the following sections.

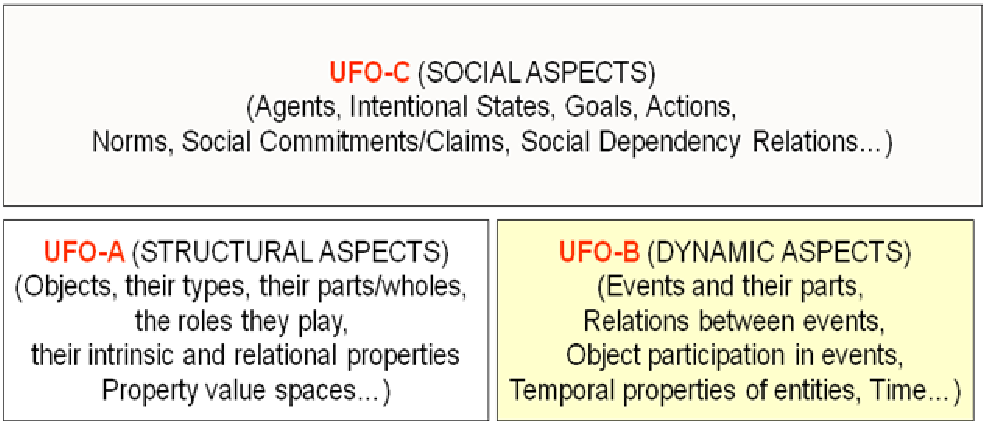

UFO-C (SOCIAL ASPECTS)
(Agents, Intentional States, Goals, Actions,
Norms, Social Commitments/Claims, Social Dependency Relations...)
UFO-A (STRUCTURAL ASPECTS)
(Objects, their types, their parts/wholes,
the roles they play,
their intrinsic and relational properties
Property value spaces...)
UFO-B (DYNAMIC ASPECTS)
(Events and their parts,
Relations between events,
Object participation in events,
Temporal properties of entities, Time...)

Figure 3.8: UFO divisions and their main subjects

UFO was initially built upon UFO-A conceptualization, also called ***Endurant***. Then, it was extended to the ***Perdurant*** (**Event**) formalizations, which has been evolving as UFO-B, supporting UFO-C social aspects (GUIZZARDI, WAGNER, *et al.*, 2013). UFO-B layer provides a set of temporal representations and the relations to UFO-A concepts, such as mereology of **Events**, the **Participation** of ***Endurants*** in **Events**, how an **Event** transforms one **Situation** to another, among others. For last, UFO-C presents the set of concepts that represent social aspects, as **Actions** and **Social Agents**, which are supported by UFO-A and UFO-B concepts.

The UFO top-level concept is the **Thing** which can be seen as the most abstract concept, defined to be the root of the theory about categories of **Universals** and **Individuals**. The former is the set of standard properties and relations (e.g. `dog`, `person`) that can be used to classify the **Individuals**, which represent the instances (e.g. `the dog Kito`, `the person João`). **Universal** is represented in the *conceptual level* as **Monadic**, i.e. an entity, or a **Relation**, i.e. a relationship between **Monadics**. The Individuals are represented in a *conceptual level* and the instances of them in the *instance level*. Figure 3.9 demonstrates those concepts and how they are organized in those abstractions levels.

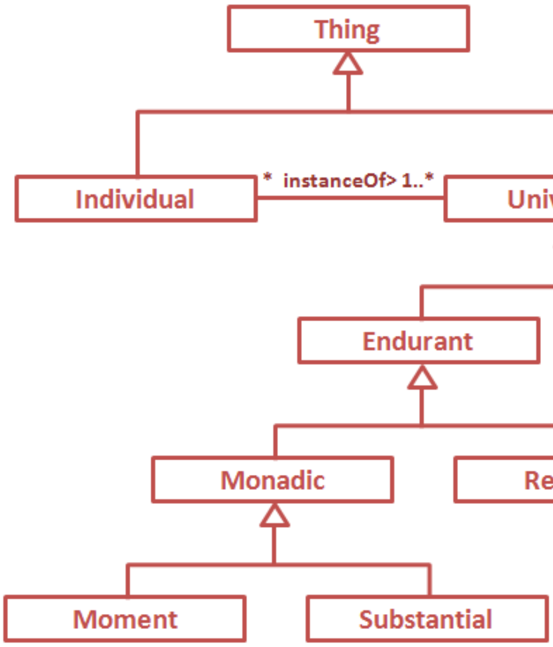


Figure 3.9: The *Endurant* and *Perdurant* (Event) categories from UFO in conceptual levels

Regarding **Individual**, it can be **Concrete Particular** or an **Abstract Particular**. The former is applicable for particular **Objects**, whilst the later is applicable to properties (quality or characteristic). The **Concrete Particular** can be an ***Endurant*** or ***Perdurant***, as previously cited. The **Abstract Particular** can be a **Quality Structure** or a **Quale**. The former is a measure structure, a values space, where a quality is associated to a **Quality Structure** (e.g. age). **Qualities** are "the basic entities we can perceive or measure" (GANGEMI, GUARINO, *et al.*, 2002), being distinguished by their qualia. **Quale** corresponds to a point in the **Quality Structure** and means a conceptualization of an intrinsic property. It "describes the position of an individual **Quality** within a certain **Quality Structure**" (CARRARETTO, 2012).

3.2.2 UFO-A: structural concepts

UFO-A layer presents the structural aspects of the reality, defining how the objects and their relations are formed. ***Endurant*** can be classified as **Substantial** (e.g. `tier`, `car`, `road`) or **Moment** (`e.g. moves`, `part-of`, `operates`, `rides`, `is-a`). **Substantial** is an entity that persists in time and which carries the identity principle and rigidity properties, being specialized as **Sortal** or **Mixin**. The former aggregates individuals with the same identity principle, whilst the later classifies entities that represent individuals with different identity principles, such as insurable item (e.g. car insurance, health insurance, life insurance, etc.). A **Sortal** can be **Rigid** (the individual only pertain to it) or **AntiRigid** (the individual can change).The **Rigid Sortal** can be an **Ultimate Sortal** or a **Subkind**. An **Ultimate Sortal**, i.e. **Kind** or **Collective** or **Quantity**, provides the identity principle to its individuals, whilst a **Subkind** inherits this principle. The **Kind** is a functional complex, one of the most common categories in a domain representation. The **Collective** is a collection of entities and the **Quantity** is a cumulative of entities.

An **Anti Rigid Sortal** can be a **Phase**, i.e. its individuals are classified as it because of their intrinsic properties (e.g. Adult by age property); or a **Role**, i.e. its individuals are classified as it because of their relational properties (e.g. Wife by is-married-with property). **Category** is a **Rigid Mixin** which generalizes **Rigid** concepts with different identity principles. **Mixin** and **RoleMixin** are the **Semi Rigid** and **Anti Rigid Mixins**, respectively. Figure 3.10 illustrates this classification and the domain concepts as instances of them.

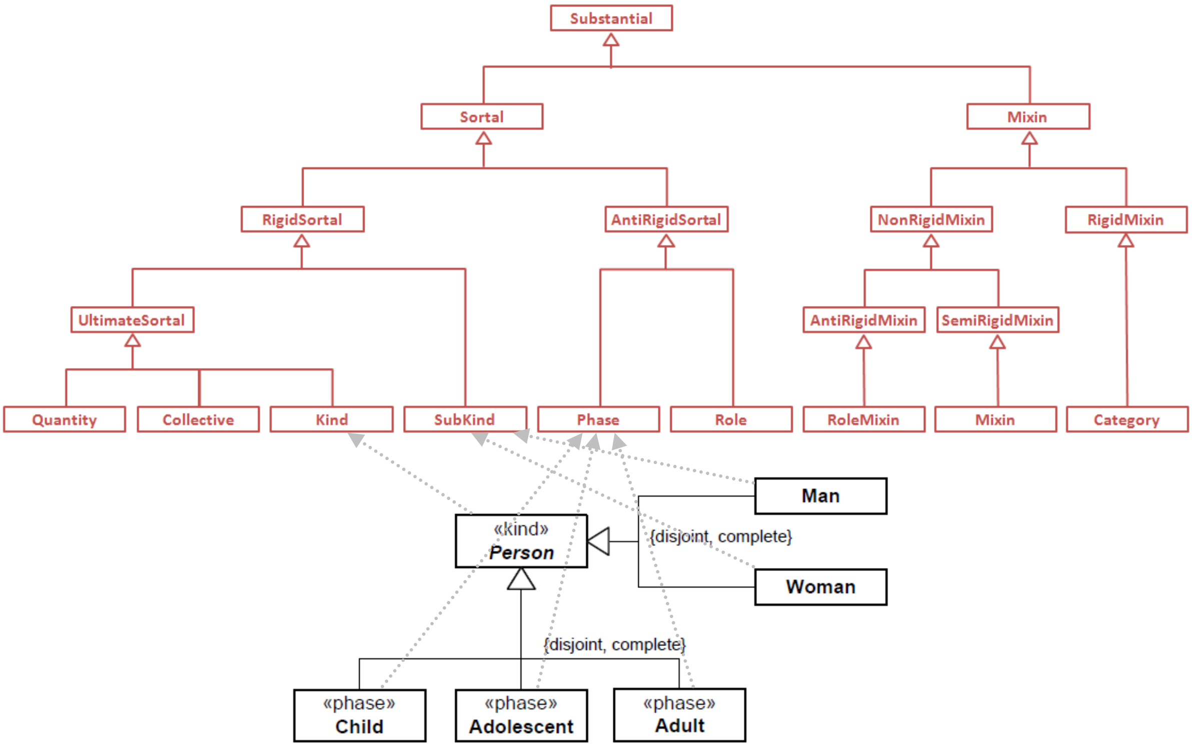


Figure 3.10: A domain ontology classified by **Substantials** concepts

The **Moments** are individuals that are inherent in others, classified as **Relator** or **Intrinsic Moment**. The **Relator** represents the mediation between individuals, i.e. it is a concept that defines a relation, such as a `subscription` of a `student` in the `university`, mediating the `study-in` relation. The **Intrinsic Moment** can be a **Quality** or a **Mode** and it denotes the intrinsic properties of the **Individual** that carries it. The former is a measurable property (e.g. `weight` and `height`), whilst the later cannot be represented in a measure system, having its value in a multidimensional structure (e.g. `colour`, `flavour`, `level of pain`). Those concepts are demonstrated in figure 3.11 bellow.

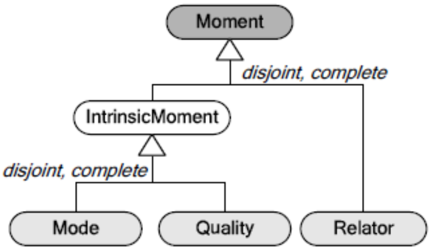


Figure 3.11: Types of **Moments** from (ZAMBORLINI, 2011)

All those concepts presented are **Monadic Universals**, as previews described. The **Relation Universal** can be classified as **Formal Relation**, **Material Relation** or **Meronymic**. The philosophic literature typically considers the first two types of relations. A **Material Relation** depends on an intermediate individual, a **Relator**, whilst a **Formal Relation** is valid only by the existence of the concepts connected. It can be a **Basic Internal Relation**, i.e. represents an existential dependency relation; or a **Domain Formal Relation**, i.e. formal relations specific of the domains. The **Basic Internal Relation** can be **Characterization** or **Mediation**. The former is the **inherits-in** relation to relate **Moment** individuals. The later is the **mediate** relation to relate **Substantials** through **Relators**. The **Meronymic Relations** conceptualize part-whole relations, being classified as **ComponentOf** (among functional complexes), **MemberOf** (among **Collections** and functional complexes), **SubCollectionOf** (among **Collections**) or **SubQuantityOf** (among **Quantities**). They are all illustrated in Figure 3.12 below.

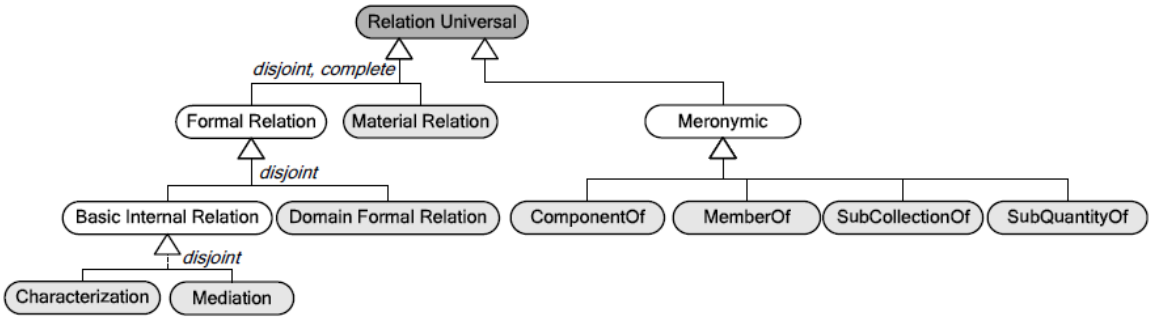


Figure 3.12: Types of **Universal Relations** from (ZAMBORLINI, 2011)

3.2.3 UFO-B: temporal concepts

Perdurant aspect from UFO denotes the **Events** which happen on the timeline, composed from temporal parts that extend in time. They are represented by possible transformations from one portion of the reality to another. It can affect the reality by changing its characteristics from a pre **Situation** to a post **Situation**. UFO-B was firstly introduced in (GUIZZARDI e WAGNER, 2005) and it is advancing, full described with axiomatization in (GUIZZARDI, WAGNER, *et al.*, 2013). It proposes a definition for the **Events Mereology**, where **Event** can be classified as **Atomic** or **Complex Events**, where the first denotes an **Event** that begins (**begin-point**) and ends (**end-point**) in the same **Time Point**, being part-of a **Complex Event**. An **Event** is a concept which is existentially dependent on the **Participation** of **Substantial** classes and it is composed by them. This is an important part of UFO-B because it presents a connection from ***Perdurants*** to ***Endurants***, a formal relation called **participationOf**, as depict in Figure 3.13 below.

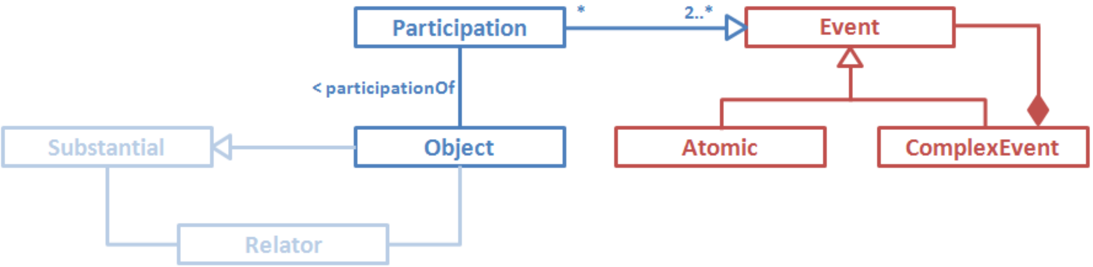


Figure 3.13: **Event** mereology and the **Object's Participation**

It illustrates the representation of **Complex Events** as sums of **Object's Participations**, where each **Atomic Event dependsOn** an **Object** to happen. One of the main parts of UFO-B is the **Time Interval Relation** between two **Events** (**source** and **target**), a **Formal Relation** that is represented by each *Allen's operator* (ALLEN, 1983), such as **during**, **before**, **meets**, **overlaps**, **starts**, **finishes** and **equals**. An example of the use of a **Time Interval Relation** is to relate `promotion` entity (first **Event**) to a `sale` entity (second **Event**) by the **before** relation. So, if we need to consider how this temporal relation occurs, we could model that a `promotion` can happen **before** a `sale` **Event**. Figure 3.14 demonstrates those concepts and their relations.

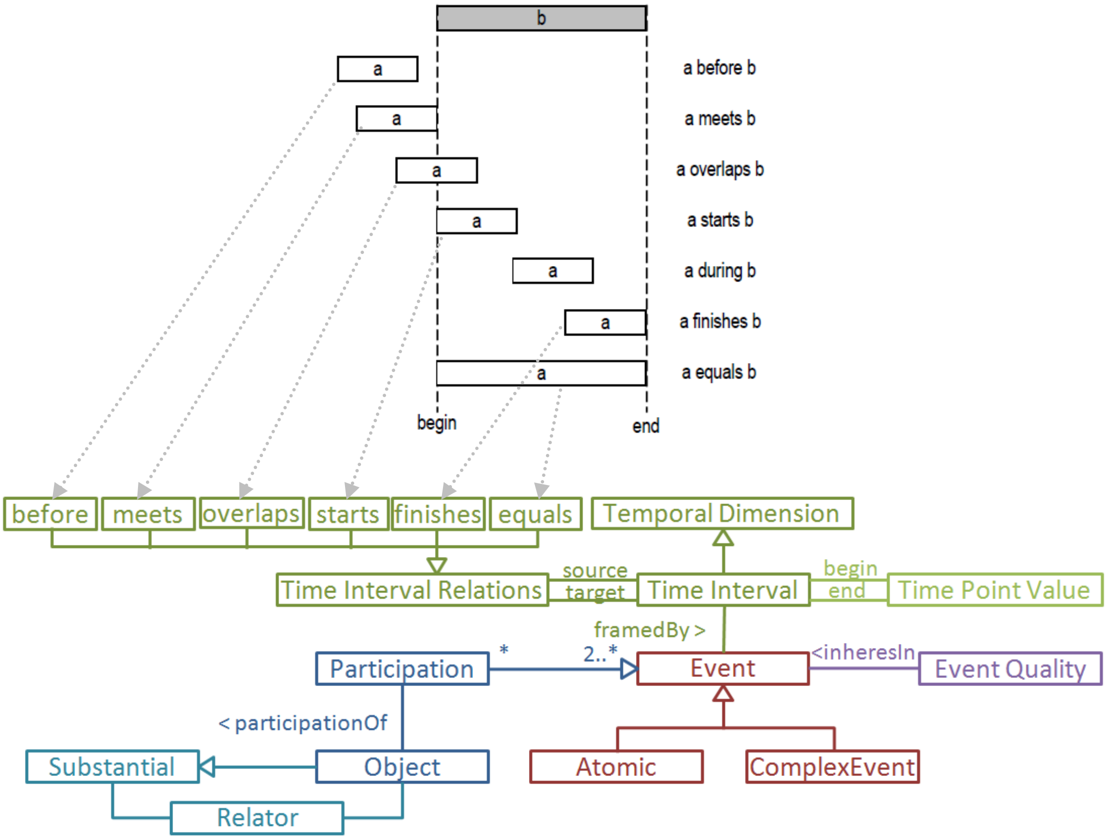
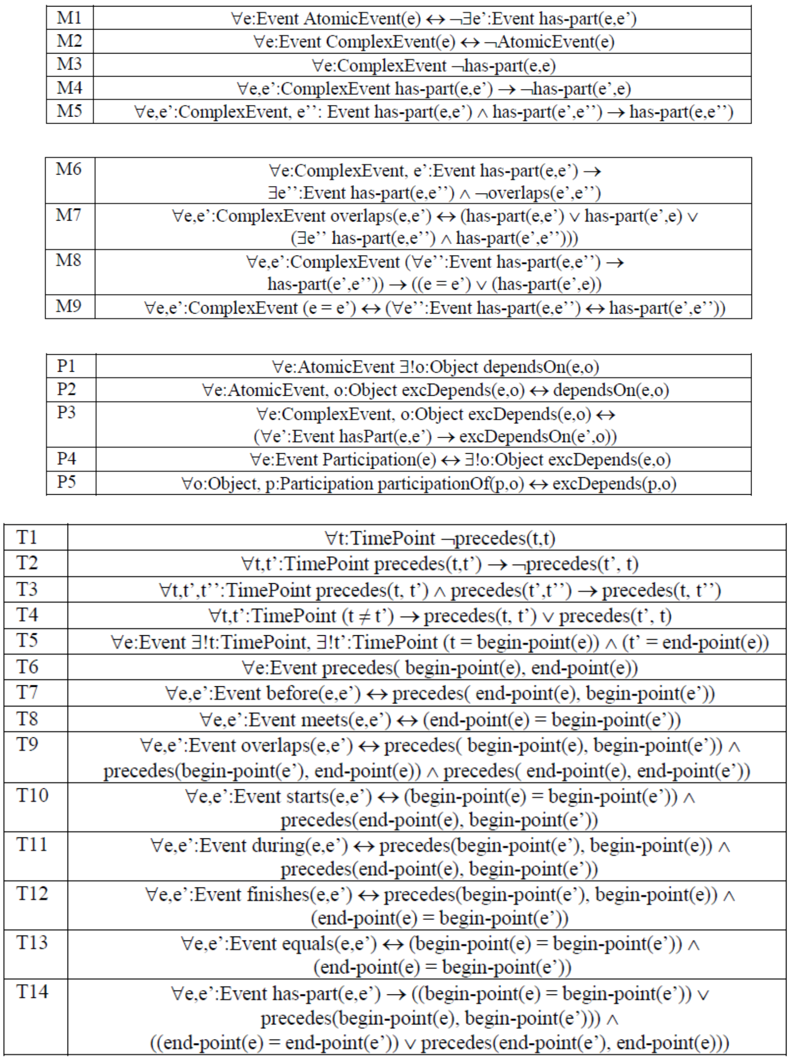


Figure 3.14: **Events Relations** and their **Time Points**

Note that, in Advanced DW (MALINOWSKI e ZIMÁNYI, 2009), page 186, the **Event** concept definition is different from UFO. There, the *state* definition is the same as the **Event** definition in UFO-B, which is framed by **begin** and **end Time Points**. Moreover, the *event* definition in Advanced DW can be seen in UFO as an **Atomic Event**. The conceptualization of UFO-B was formalized in groups of axioms (GUIZZARDI, WAGNER, *et al.*, 2013): M1 to M9 for the **Events** mereology, P1 to P5 for **Participations** and T1 to T14 for **Temporal Relations**. Table 3.1 shows these rules.

Table 2.4: Axiomatization of **Events** mereology, **Participations** and **Temporal Relations** from (GUIZZARDI, WAGNER, *et al.*, 2013)



M1	∀e:Event AtomicEvent(e) ↔ ¬∃e':Event has-part(e,e')
M2	∀e:Event ComplexEvent(e) ↔ ¬AtomicEvent(e)
M3	∀e:ComplexEvent ¬has-part(e,e)
M4	∀e,e':ComplexEvent has-part(e,e') → ¬has-part(e',e)
M5	∀e,e':ComplexEvent, e'': Event has-part(e,e') ∧ has-part(e',e'') → has-part(e,e'')

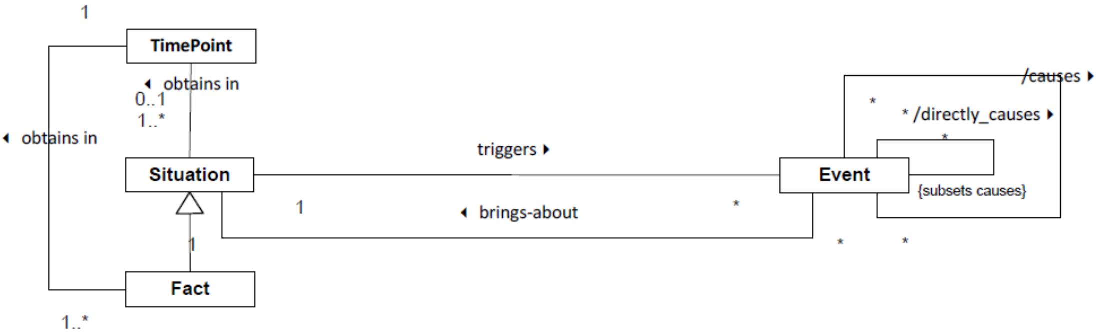
M6	∀e:ComplexEvent, e':Event has-part(e,e') → ∃e'':Event has-part(e,e'') ∧ ¬overlaps(e',e'')
M7	∀e,e':ComplexEvent overlaps(e,e') ↔ (has-part(e,e') ∨ has-part(e',e) ∨ (∃e'' has-part(e,e'') ∧ has-part(e',e'')))
M8	∀e,e':ComplexEvent (∀e'':Event has-part(e,e'') → has-part(e',e'')) → ((e = e') ∨ (has-part(e',e))
M9	∀e,e':ComplexEvent (e = e') ↔ (∀e'':Event has-part(e,e'') ↔ has-part(e',e''))

P1	∀e:AtomicEvent ∃!o:Object dependsOn(e,o)
P2	∀e:AtomicEvent, o:Object excDepends(e,o) ↔ dependsOn(e,o)
P3	∀e:ComplexEvent, o:Object excDepends(e,o) ↔ (∀e':Event hasPart(e,e') → excDependsOn(e',o))
P4	∀e:Event Participation(e) ↔ ∃!o:Object excDepends(e,o)
P5	∀o:Object, p:Participation participationOf(p,o) ↔ excDepends(p,o)

T1	∀t:TimePoint ¬precedes(t,t)
T2	∀t,t':TimePoint precedes(t,t') → ¬precedes(t', t)
T3	∀t,t',t'':TimePoint precedes(t, t') ∧ precedes(t',t'') → precedes(t, t'')
T4	∀t,t':TimePoint (t ≠ t') → precedes(t, t') ∨ precedes(t', t)
T5	∀e:Event ∃!t:TimePoint, ∃!t':TimePoint (t = begin-point(e)) ∧ (t' = end-point(e))
T6	∀e:Event precedes(begin-point(e), end-point(e))
T7	∀e,e':Event before(e,e') ↔ precedes(end-point(e), begin-point(e'))
T8	∀e,e':Event meets(e,e') ↔ (end-point(e) = begin-point(e'))
T9	∀e,e':Event overlaps(e,e') ↔ precedes(begin-point(e), begin-point(e')) ∧ precedes(begin-point(e'), end-point(e)) ∧ precedes(end-point(e), end-point(e'))
T10	∀e,e':Event starts(e,e') ↔ (begin-point(e) = begin-point(e')) ∧ precedes(end-point(e), begin-point(e'))
T11	∀e,e':Event during(e,e') ↔ precedes(begin-point(e'), begin-point(e)) ∧ precedes(end-point(e), begin-point(e'))
T12	∀e,e':Event finishes(e,e') ↔ precedes(begin-point(e'), begin-point(e)) ∧ (end-point(e) = begin-point(e'))
T13	∀e,e':Event equals(e,e') ↔ (begin-point(e) = begin-point(e')) ∧ (end-point(e) = begin-point(e'))
T14	∀e,e':Event has-part(e,e') → ((begin-point(e) = begin-point(e')) ∨ precedes(begin-point(e), begin-point(e'))) ∧ ((end-point(e) = end-point(e')) ∨ precedes(end-point(e'), end-point(e)))

The **Situation** concept from UFO-B has the same meaning of the state of affairs defined in philosophical literature. The **Situations** are snapshots from a part of the world in a specific **Time Point**, involving intrinsic aspects, such as **Endurants** participations and even other **Situations**. Once they are representations of the reality changes, they have two relations to **Events**: (i) A **Situation** can **trigger** a set of **Events**. (ii) A **Situation** can be **brought about** a set of **Events**. The relations **triggers** and **brings-about** are illustrate in Figure 3.15 and its axioms formalized. Therefore, it is possible to model how a change in the world (an

Event) makes the state of affairs changes from one **pre-situation** (the **Situation** which **triggers** the **Event**) to a **post-situation** (a **Situation brought about** the **Event**).



S1	∀s:Situation, e:Event triggers(s,e) → obtainsIn(s, begin-point(e))
S2	∀s:Situation, e:Event brings-about(e,s) → obtainsIn(s, end-point(e))
S3	∀e:Event ∃!s:Situation triggers(s,e)
S4	∀e:Event ∃!s:Situation brings-about(e,s)
S5	∀s:Situation fact(s) ↔ ∃t:TimePoint obtainsIn(s,t)

currence of e. In other words, we can state that e *directly-causes* e' iff:

S6	∀e,e':Event directly-causes(e,e') ↔ ∃s:Situation brings-about(e,s) ∧ triggers(s,e')

Finally, we define a *causes* (S7) relation between events as follows:

S7	∀e,e'':Event causes(e,e'') ↔ directly-causes (e,e'') ∨ (∃e':Event causes(e,e') ∧ causes(e',e''))

Figure 3.15: **Situations** metamodel and axiomatization from (GUIZZARDI, WAGNER, *et al.*, 2013)

The causality relation among **Events** is also explored in UFO-B, stating that **Events** can **cause** other **Events** to happen. An **Enabling Condition** is a **Proposition** which is the description of the necessary conditions for the occurrence of an **Event**. Once a **Situation** satisfies a **Proposition**, the **enable** relation between a **Situation** and the **Event** can be derived.

There are works that discuss the **Situation** conceptualization for simulation and Context-Aware Systems (COSTA, 2007). The concept of **Situation** was included in the conceptual level and its contextual specialization (**Context Situation**) was formalized with a relation to the **Context**, which is an individual inherent to other entities participations. The **Context Situation** can be **Relational** (e.g. `João being married to Bel`), **Intrinsic** (e.g. `João being with fever`), **Formal Relation** (e.g. `João being more then 10cm taller than Pedro`), **Situation of Situation** (e.g. `João being with recurrent fever for 2 days`) or **Combined Situation** (e.g. `João being older` than 30 years while Dilma is president of Brazil). A network domain ontology was constrained by OCL rules. **Situation of Situations** are composition of situations of any types, i.e. it is possible to group other **Situations**, having a relation (called **sitSituation**) to **Context Situation** with cardinality 1..*.

As cited before, Context-Aware Systems can be developed based in MDD approaches, considering Situations modelling. In this line of research, a language for modeling situations, the SML (Situation Modelling Language) was proposed in (COSTA, MIELKE, *et al.*, 2012) and evolved in (MIELKE, 2013). Using rules-based platform for the management of situations is a subject that was applied in (PEREIRA, COSTA e ALMEIDA, 2013). In addition, UFO-B defines an **Event** as **Disposition** of objects, that is, the properties that are activated by some **Situations** and manifested-by **Atomic Events**. Several issues are open about the ***Perdurant*** part of UFO and they are under discussion.

3.2.4 UFO-C: social concepts

UFO-C is the part of UFO that deals with Social entities, classified by UFO-A and UFOB concepts. The **Substantial** is divided as an **Object** or an **Agent**. The **Object** is an inanimate **Substantial**, unable to act, unlike an **Agent**. They can be classified as **Social**, such as a society (**Agent**) and a language (**Object**), or **Physical**, such as a person (**Agent**) and a book (**Object**). An **Agent** can be **Human**, **Collective Social** or **Organization** and creates **Actions**, which is an **Intentional Event**. **Normative Description** is a specialization of a **Social Object** and a **Plan Description** is a **Normative**, which describes **Complex Actions**. A set of concepts of UFO-C was described and formalized in first order logic and some works are applying and discussing it (ROCHA, 2012). Among others concepts, stand out: **Social Role**, **Goal**, **Intentional Moment**, **Mental Moment**, **Commitment**, **Claim**, **Belief**, **Desire**, **Social Relator**, **Delegation**, **Resource Participation** and **Action Contribution**.

3.2.5 OntoUML

OntoUML, an extension of UML, was built to represent UFO concepts (GUIZZARDI, 2005). It contains the elements that represent ontological distinctions and the constraints that govern possible relations which these elements can have, from UFO-A perspective.

These constraints are derived from the UFO axiomatizations for **Substantials** and **Moments**, such as **Kind**, **SubKind**, **Role**, **Relator**, **formal** relation, **material** relation and the **meronymic** relations.

The admissibility of some sates of affair in domains depends on factual knowledge, which only human cognitive can be used to validate. Therefore, as the model-driven approaches, a verification and validation process is required. It is done by the visual simulation of possible words, i.e. by automatically generate examples and counter examples of the concepts and their relations presented into the domain ontology. To support this, it was proposed a set of patterns and anti-patterns as best practices in (SALES, BARCELOS e GUIZZARDI, 2012). This kind of approach is available through the OLED software[15], a tool which provides OntoUML domain ontology verification, a set of anti-patterns detection (common design errors), OCL parser and Alloy Analyzer (JACKSON, 2012) for visual validation.

A plug-in of OntoUML for Enterprise Architect (EA) was made available by NEMO research group[16] and there are several applications considering UFO as a top-level ontology for different goals. Also, diverse domains have been modelled well-founded in UFO, such as ITIL process (CALVI, 2007), healthcare for context aware system (COSTA, 2007), tuberculoses disease (PEREIRA, COSTA e ALMEIDA, 2013), banking (COSTA, MIELKE, *et al.*, 2012), news publications (ROCHA, 2012), soccer game narratives (PENA, 2012), marriage and genealogy (GUIZZARDI e ZAMBORLINI, 2012), biodiversity domain, normative acts, army equipments, federal universities organizational structure, transport optical network architecture, heart electrophysiology, on-line mentoring activities (SALES, BARCELOS e GUIZZARDI, 2012), among others.

A remarkable work regarding OntoUML was introduced in (CARRARETTO, 2012). It proposed a model-driven approach to derive information models from well-founded domain ontologies classified with UFO concepts. A process based in transformation rules was introduced, stating a set of mapping patterns from ontological concepts formalized through

[15] https://code.google.com/p/ontouml-lightweight-editor/

[16] http://nemo.inf.ufes.br/

OntoUML to object OO concepts formalized through UML. Informational concerns about history and time tracking were discussed. For example, if the informational interest regarding a **Kind** `Person` is only the past, then the *Class* `Past Person` is mapped. If the informational interest is both past and present, a *Class* `Person` containing `current` *property* (a Boolean) is mapped. When analyzing **Relator** concept, a similar strategy for history decision with `current` *attribute* is defined, but it also considers the **Relator** cardinalities in the mapping rule. For time tracking decisions of **Relators**, the *Class* derived counts with three attributes: `startTime` and `endTime` as *TimeInstant* data type; and `duration` as *TimeInterval* data type. In this approach, it is also discussed reference decisions via attributes and data types by identifiers, where it can be mapped to primitive or user-defined data type. In addition, measurement representation by attributes is demonstrated from **Quality** in OntoUML. Figure 3.16 illustrates an example of the execution of the approach considering the pattern for Relator types and Role associations.

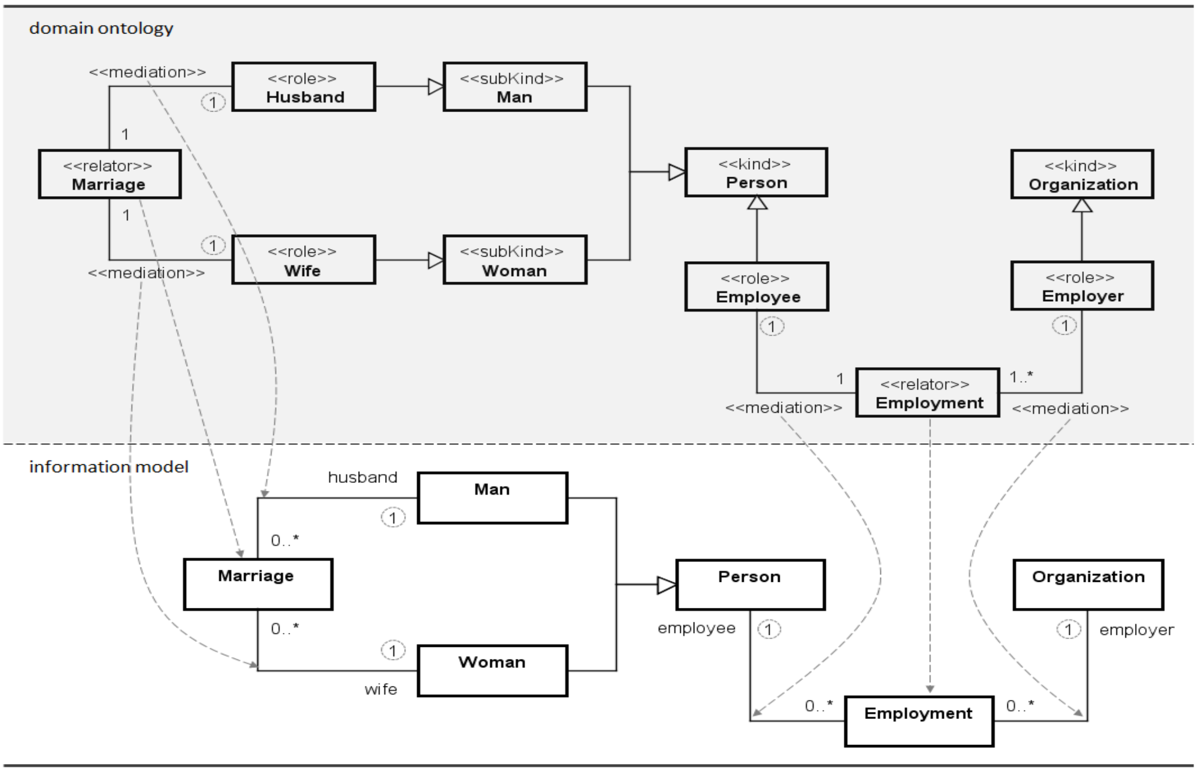


Figure 3.16: An example of domain ontology described in OntoUML mapped to UML from (CARRARETTO, 2012)

There are substantial research efforts[17] in the direction of evolve UFO and apply it in distinct ways.

[17] http://nemo.inf.ufes.br/en/publications

4 Proposal

This chapter introduces a novel approach to increase multidimensional (MD) design activity automation in BI/DW solutions from the point of view of a MD modeler. We propose a systematic automation of the hybrid approach, where the domain ontology is built based on UFO conceptualization, increasing its expressiveness. The semantic enrichment of MD schemas is made by exploring the temporal concepts defined in UFO-B profile. A derivation process is suggested for the design of MD concepts from well-founded domain ontology by a set of rules. The approach includes support for unstructured data in BI/DW lifecycle process, once this kind of data source is fundamental for business analytics.

The following sections describe the proposal details and its overview is illustrated in Figure 4.1. At first, in section 4.1, a set of derivation rules is proposed to increase the semantic expressiveness of the MD design process, orienting the MD modeler how to choose the MD concepts based on UFO concepts, such as events, participations, temporal relations and situations. These rules execution occurs in “MD concepts derivation through rules (UFO -> MD)” activity, where MD concepts can be inferred through them. In a prior step, the domain ontology is built well founded on UFO categories. In section 4.2 an ontological hybrid MD design method is introduced (based on Figure 4.1), considering both structured and unstructured data sources during the source-driven phase. Finally, we conclude the approach proposal in section 4.3.

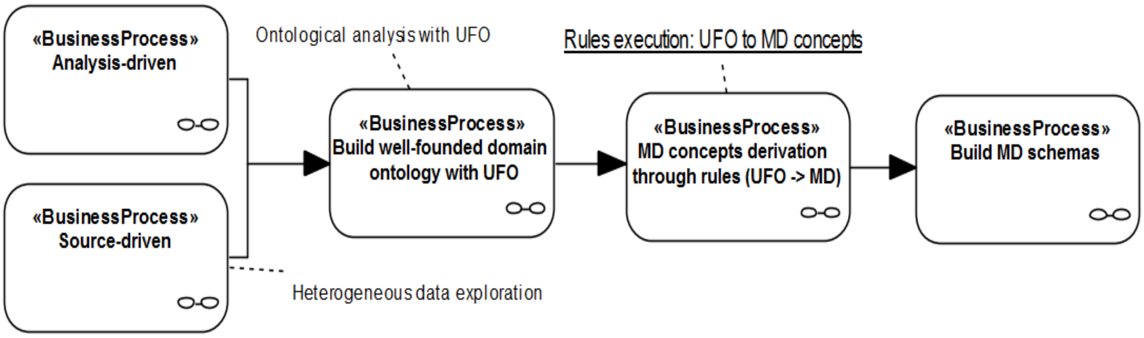

«BusinessProcess»
Analysis-driven
«BusinessProcess»
Source-driven
Ontological analysis with UFO
Heterogeneous data exploration
«BusinessProcess»
Build well-founded domain ontology with UFO
Rules execution: UFO to MD concepts
«BusinessProcess»
MD concepts derivation through rules (UFO -> MD)
«BusinessProcess»
Build MD schemas

Figure 4.1: Proposal overview as MD design process

4.1 OntoWarehousing

UFO deals with temporal aspects (UFO-B), which is an important characteristic of MD design, having intuitive relation to MD concepts such as the fact (in MD terminology) as a business event. In addition, it provides a top-level ontology well-founded in metaphysics, axiomatized in descriptive logic (DL). Therefore, in our approach, we adopted UFO and ontological analysis to support the development of a domain ontology used for MD schemas derivation, based on a set of rules. We concentrated in analysing the *perdurants* (events), their temporal aspects, participations and situations.

It is proposed here the semantic enrichment of MD schemas by using the temporal concepts defined in UFO-B profile (presented in section 3). First, the domain representation is semantically enriched using UFO. Then, a set of mapping rules is applied and alternative elements for the MD schema are identified. Finally, based on analytical requirements, the appropriate elements are chosen. Therefore, the approach defines how to map top-level concepts from a domain (UFO categories) to MD concepts, represented as *Facts*, *Measures*, *Dimensions*, *Hierarchies* and *Attributes*. In this direction, our proposal resembles the mapping from ontological level to the information level introduced in (CARRARETTO, 2012), revised in chapter 3.

As cited in section 3, how to classify domain ontologies based on UFO has been extensively covered in many efforts and particularly on the original work (GUIZZARDI, 2005). One must assume that a well-founded domain ontology is available, marked with UFO’s stereotypes and complemented by the corresponding restrictions. MD elements are mapped from the domain ontology elements: (i) Events and their mereology (complex and atomic events) are mapped to facts and measures. (ii) Participants in events determine possible perspectives of analysis (dimensions and hierarchies). (iii) Time interval relations between events could derive a fact pattern for the analysis of those temporal relations. (iv) Causality relation between events can be analysed through fact/dimension dichotomy. (v) Situation changes and their relations to events as a MD schema for cause-effect analysis.

This derivation process is independent of data sources type: structured or unstructured. It is assumed that the domain ontology here already represents the concepts from both universes. However, there are some issues regarding modelling textual information, such as using reverse engineering from text to the domain ontology, which is treated in the hybrid approach proposal in section 4.2.

4.1.1 Events as *Facts*

The facts are representations of business events with measurements to be analyzed by theirs dimensions, attributes and hierarchies (perspectives of analysis). By identifying the domain events and their mereology, i.e. how complex and atomic events are related in a taxonomy, we can derive possible facts and measures to the MD modeller. To do so, all complex events and its parts should be listed, as other complex sub events (parts of the whole event). The numeric values represented by qualities in events can support useful aggregations, so they can be set as measures of a fact. The set of axioms that defines the mapping rules to propose facts and measures from complex and atomic events mereology can be described as (Figure 4.2):

R1.a. If *e* is a **Complex Event**, then *e* can be represented as a *Fact f*;

R1.b. If *f* is a *Fact* defined by **Event** *e*, then the numeric **quality** of *e* can be derived as *Measures* of *Fact f*;

R1.c. If *f* is a *Fact* from **Event** *e* and *e’* is an **Event** which has mereologic relation (**partOf**) to *e* and there exist **quality** values *qs* of *e’*, then *e’* can be derived as a *Fact* or *qs* as *Measures* in the *Fact f*.

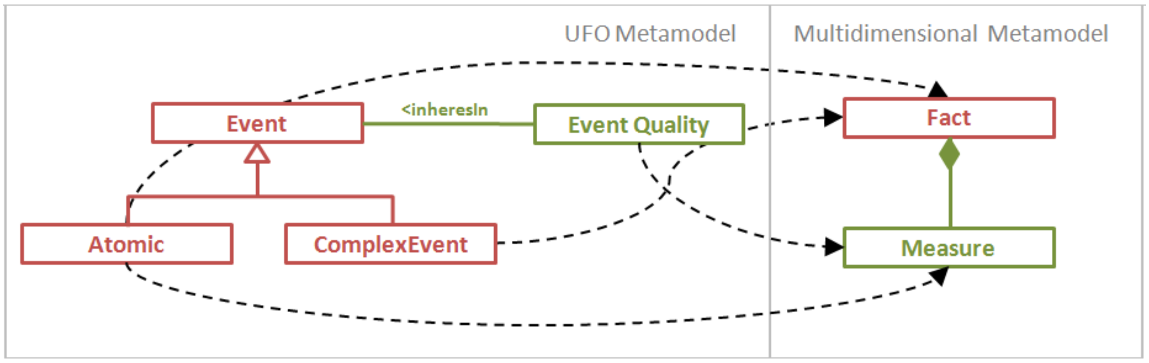


Figure 4.2: Mappings **Events** mereology (UFO) as *Facts* and *Measures* (MD)

Notice that, as events can bear qualities (GUIZZARDI, WAGNER, *et al.*, 2013), the second axiom the *Measure* derived from an event can be fully evaluated by its particular qualities (their attributes). In other words, checking the numeric qualities of the event (or its parts), they can be described as *Measures*. For example, the sale event can be composed by other events, such as product request and payment. The payment tax is a quality from payment event and can be designed as a *Measure* for both cases: if sale event (Figure 4.3a) or the payment event (Figure 4.3b) is chosen as a *Fact*, having different aggregation constraints.

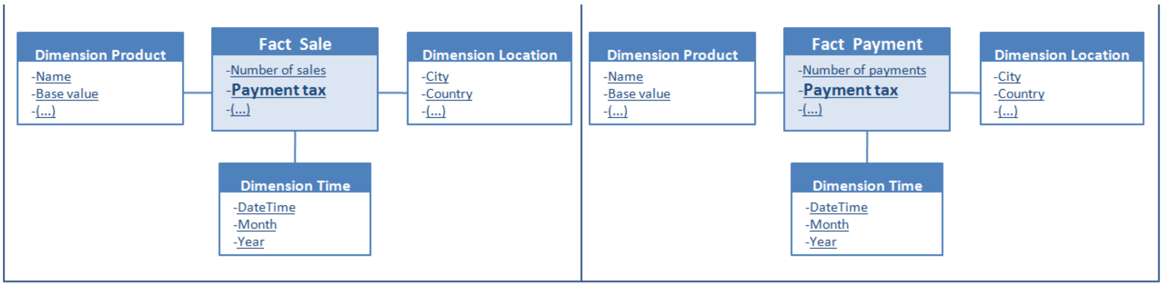


Figure 4.3: (a) sale as *Fact*; (b) payment as *Fact*; both with payment tax *Measure*

An electric disturbance in Brazilian integrated system is a complex event which is composed by other sub events and it is suggested as a *MD Fact* by the rule definition R1a. The starting event of a disturbance occurs in one particular equipment, this participation defines the disturbance cause and can be a natural phenomenon, such as an atmospheric electrical discharge or a storm. It can also be a human failure, or an accident in some transmission line, such as animals hitting it or a flood, among others types of failures in the network. According to UFO they are classified as participations of objects in a disturbance. Transmission line (an equipment container) or a power transformer (an equipment), classified as substantials in UFO, are affected by those types of disturbance causing events (disturbance cause), then the disturbance event begins to occur. Depending on a set of parameters (classified as quality structures in UFO) the disturbance can have other sequential, even parallel, sub events. The disturbance accumulates all its temporal parts, such as a forced shutdown complex event, which is an event of automatically or manually turning off components of the electrical network. It can also be suggested as a *MD Fact* by rule R1a, pointing to a real necessity for the business, once the analysis over the forced shutdown event is an important issue to the electric high-voltage post-operation process.

During a disturbance occurrence, the load cutting event can happen. It is measured by an energy interrupted charge (in MW), which is a quality structure of the electrical network affected. When this charge is turned off, blackouts can happen depending on its value and the resources involved, affecting several regions. By applying rule R2b, load cut can be derived as a possible *Measure* of the *Fact* disturbance. This example is thoroughly explained in chapter 5.

4.1.2 Objects Participations as *Dimensions* and *Hierarchies*

Dimensions and their hierarchies are perspectives of analysis according to different level of details over a business event, derived as a fact by R1. Substantial objects can be participants in an event (through participation). An event is existentially (ontologically) dependent on other objects and a complex event is represented as a sum of object's participations. Therefore, object's participations in events represented in the domain ontology can be viewed as possible perspective of analysis for the fact, i.e. dimensions, attributes and hierarchies. Typical examples of object participants in a sale event are: the sold product with its associated category, which forms a product hierarchy; the spatial region where it was sold classified by country, state, city; the product supplier (companies); the sale vendor; among others. Notice that we can derive possible dimensions and their hierarchies by analyzing the relations between its concepts, such as the entities classified as role mixin participant of the participation, described as following (Figure 4.4):

R2.a. If *f* is a *Fact* representing an **Event** *e*, *o* is an **Object** (**Substance**) which is mediated by **formal** relation (**participationOf**) to a **Participation** *p*, which has a mereologic relation (**partOf** *e*), then *o* can be derived as a *Dimension* of *Fact f*;

<u>R2.b.</u> If *f* is a *Fact* representing an **Event** *e*, *d* is a *Dimension* defined by **Object** *o* which is **Participant** of *e*, and *o* has relationships to other **Substances** *S*, then each **Substance** *s* of *S* can be suggested as *Hierarchies* of the *Fact f* through *Dimension d.*

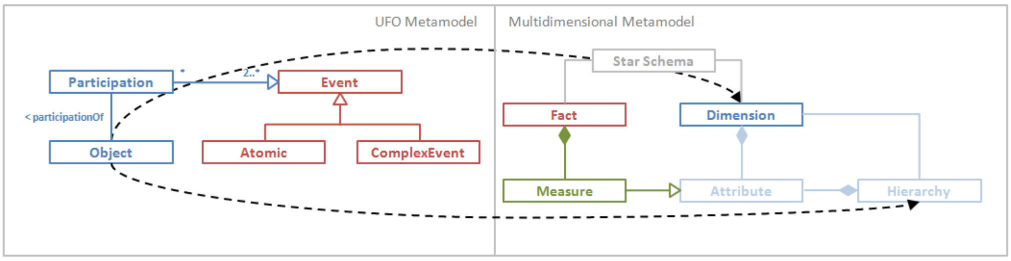


Figure 4.4: Mappings **Participations** (**UFO**) as *Dimensions* and *Hierarchies* (MD)

For example, the `sale` event is the sum of `client` and `product` participations. Therefore, `client` and `product` participations in a `sale` are relations between the `client` and `product` participants (role mixins), respectively, with `sale` event. Therefore, we can represent `client` and `product` as dimensions in a MD schema, as illustrated in Figure 4.5.

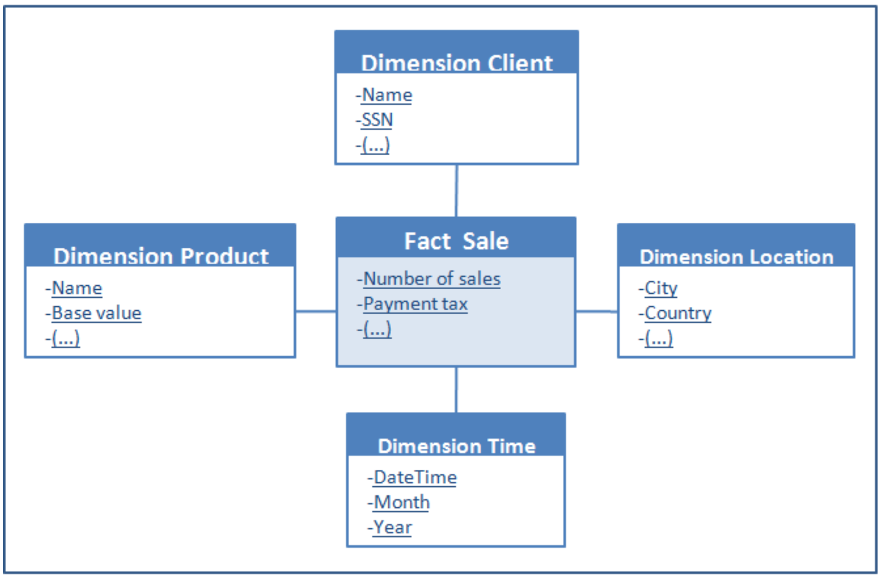


Figure 4.5: Example of `sale` fact with `product` and `client` participants as dimensions

In another example, the disturbance is existentially dependent from equipment participation. The participation is an event and can be complex or atomic as well. So, the transmission line participation in a disturbance is a mediation relation between the transmission line participant (a role mixin) to a disturbance event. Therefore, we can represent the transmission line as a perspective of analysis (a dimension) at the MD schema, as depict in chapter 5.

4.1.3 Time Interval Relations between Events as a *Snowflake Schema*

The temporal properties of events are represented by a quality structure composed of time intervals. In our work we adopt linear ordered time points as a chronic due to its formalization in (GUIZZARDI, WAGNER, *et al.*, 2013). Each event must be framed (associated) by a time interval that is defined by its begin and end time points. The temporal relations between events in the domain ontology design are represented by the Allen's operators, i.e. before, meets, overlaps, starts, during, finishes and equals formal relations. Those operators are represented as time interval relations and join the time intervals of two events by their time points. The use of these representations by the modeller indicates that the time interval relations are important issues to the business needs. Therefore, analysis over them can be derived through a MD schema, which relates two events by their temporal parts. The mapping rule checks for the related events in the domain ontology and is formalized as:

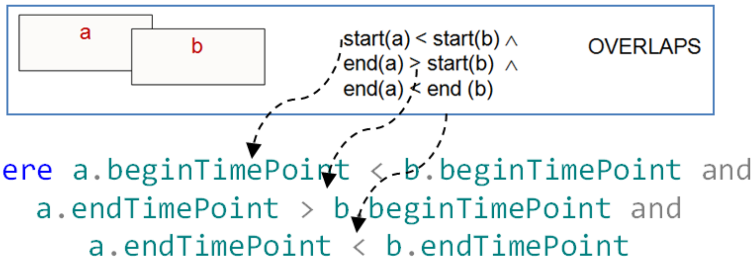
<u>R3.a.</u> If *e1*, *e2* are **Events** and they are modelled with one or more **Time Interval Relations** between them [18], then a *Fact f* can be represented by the **Time Interval Relations** between the **Events** *e1* and *e2*, where *e1* is represented by *Dimension d1* and *e2* represented as *Dimension d2*. Each *Dimension* (*d1* and *d2*) has a *Time Hierarchy* for the **Begin** and **End Time Points**;

<u>R3.b.</u> If a *Fact f* is represented by the **Time Interval Relations** from the **Events** *e1* and *e2*, having *e1* represented as *Dimension d1* and *e2* represented as *Dimension d2*, then

[18] In our approach we only consider the temporal relations between two events which was specified in the domain ontology, even knowing that any event can be formal related with another by their time points intervals.

the fundamental **Participants** of the **Events** can be derived as *Attributes* or *Hierarchies* for *d1* and *d2*.

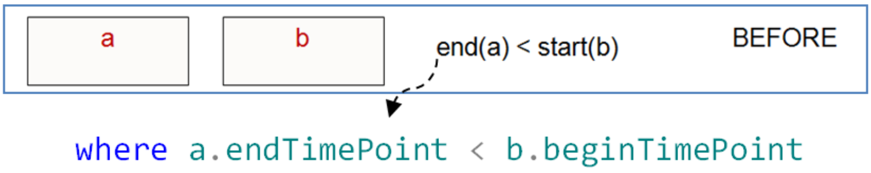
A temporal fact pattern is established as a snowflake schema, supporting analysis of all possible temporal relations between the two events being observed. The events interval times are represented by time dimensions for each event with typical time hierarchies, such as year, semester, month and day for the begin and end time points. Each event (dimension) also has its participants designed as (shared or not shared) dimensions, attributes or hierarchies. Complementary, exploring UFO time interval relations and their intrinsic constraints, it is possible to derive data loading rules for the ETL process through begin and end time points relations between the events. The Allen´s operators play this role because they provide the specific constraints for loading data into the fact table and in the evaluation of the temporal measures, considering summarizability issues. For example, the measure can represent how long the event *e1* overlaps the event *e2*, stating the relations among the events time points (begin and end). Those constraints can be mapped to SQL queries as WHERE clauses to join the events structures and load the fact table (Figure 4.6).



```
where a.beginTimePoint < b.beginTimePoint and
    a.endTimePoint > b.beginTimePoint and
       a.endTimePoint < b.endTimePoint
```

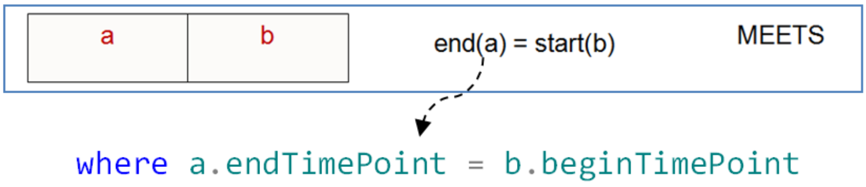
Figure 4.6: Example of overlapping Events and the resulted *WHERE* clause

The axiomatization of the **Allen’s operators** is available in UFO-B formalization (GUIZZARDI, WAGNER, *et al.*, 2013) and each one is mapped to its respective WHERE clause as follows (Figures 4.7 to 4.12):



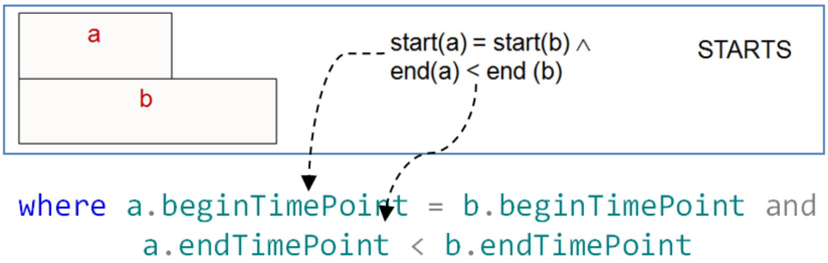
```
where a.endTimePoint < b.beginTimePoint
```

Figure 4.7: Example of **before** relation as *WHERE* clause



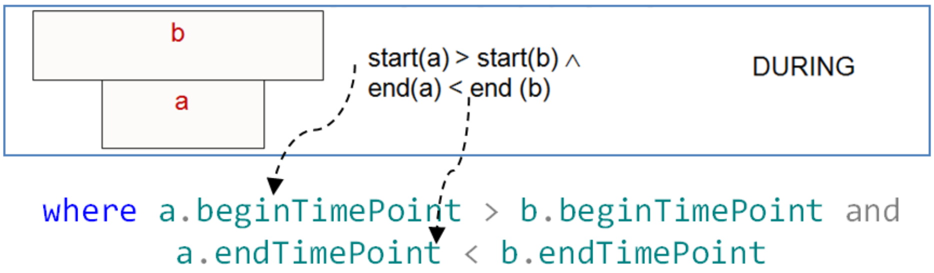
```
where a.endTimePoint = b.beginTimePoint
```

Figure 4.8: Example of **meets** relation as *WHERE* clause



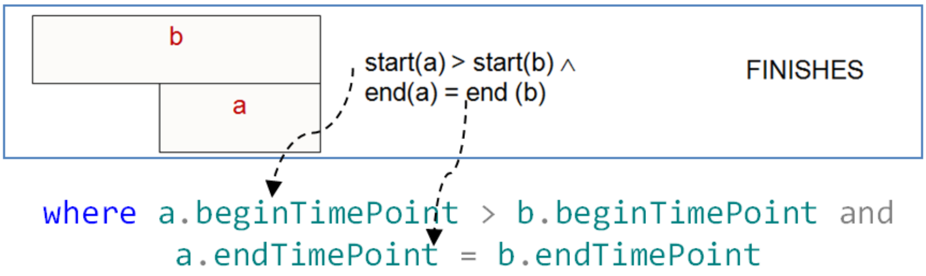
```
where a.beginTimePoint = b.beginTimePoint and
      a.endTimePoint < b.endTimePoint
```

Figure 4.9: Example of **starts** relation as *WHERE* clause



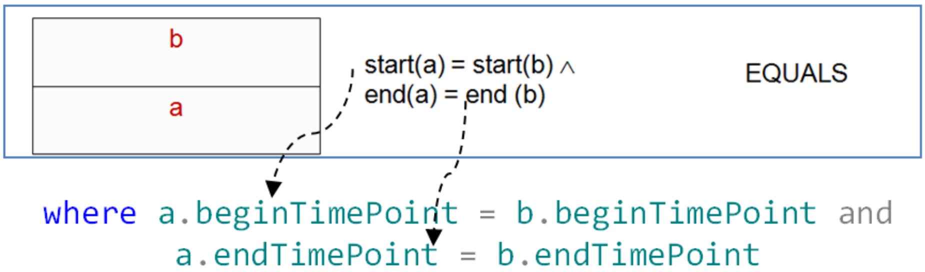
```
where a.beginTimePoint > b.beginTimePoint and
      a.endTimePoint < b.endTimePoint
```

Figure 4.10: Example of **during** relation as *WHERE* clause



```
where a.beginTimePoint > b.beginTimePoint and
      a.endTimePoint = b.endTimePoint
```

Figure 4.11: Example of **finishes** relation as WHERE clause



```
where a.beginTimePoint = b.beginTimePoint and
      a.endTimePoint = b.endTimePoint
```

Figure 4.12: Example of **equals** relation as WHERE clause

Adding possible perspectives over the event by their participants, which define the structural objects, is important. For example, to analyse the possible impacts of a `disturbance` event to the society by the `news publications` event, it is possible to model that a `disturbance` occurs before the `news publication` which reflects the `disturbance` impact in the community. Particularly, the before relation must be accompanied by a maximum number of time points quantity when implementing the ETL process, as we shall demonstrate in the case study, meaning "how much in advance" an event may occur before another.

The first three sets of mappings are represented in Figure 4.13 by colours equivalences. Figure 4.13a presents a subset of UFO concepts and the derived MD schema is represented in Figure 4.13c. Just as a reference, a MD metamodel is depicted in Figure 4.13b, also using colours for the corresponding concepts.

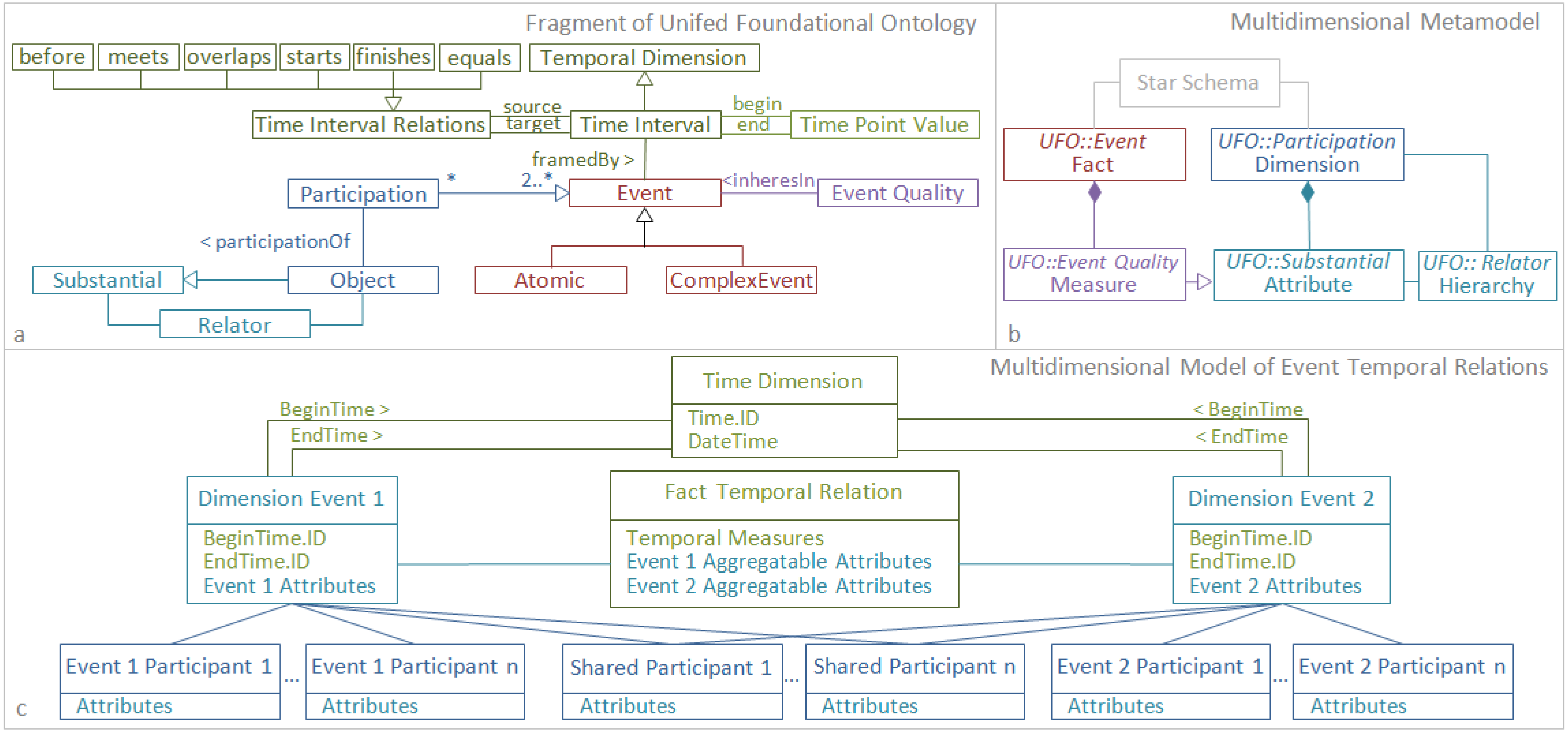


Figure 4.13: Mapping rules represented from UFO to MD concepts by colors coding

4.1.4 Causality relation between Events as *Fact/Dimension* dichotomy

An event can cause another event to happen, as described in section 3.2.3. This causality relation is a type of dependency relation and can provide a perspective of analysis about the cause of a business event. Therefore, in this rule we state that if a cause relation is found in the domain ontology, then the causing event is set as a dimension and the caused event is set as a fact. As defined in rule R1b, any numeric quality of the caused event can represented as the fact's measures. The rule is described as (Figure 4.14):

R4.a. If *e1, e2* are **Events** and *e1* **causes** *e2*, then *e1* can be derived as a *Dimension* and *e2* can be derived as a *Fact*;

R4.b. If *f* is a *Fact* defined by R4.1 from **Event** *e*, then the numeric attributes *attr* of *e* can be suggested as *Measures* for *Fact f*.

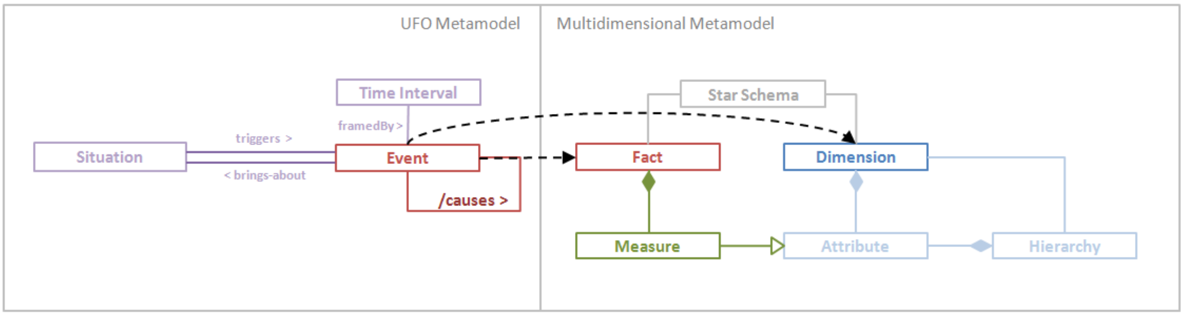


Figure 4.14: Mappings **Events Causality** (**UFO**) as *Dimension* / *Fact* (MD)

In an example domain scenario where the `payment` event is designed with causality relation to a `product delivery`, the execution of this rule can derive the fact `product delivery` and the dimension `payment`. An analysis example can be the `average delivery time` by `payment` attributes, such as date/time. Notice that, by combining these rules with R2, it is possible to set several attributes and hierarchies from the causing event participations. In this context, the `payment` event can have the participations of `client`, `vendor`, `product` and `store`. Therefore, it is possible to generate a set of analysis over the `product delivery` fact by the `payment` dimension and the `store location` hierarchy, for example. Figure 4.15 illustrates this MD schema example.

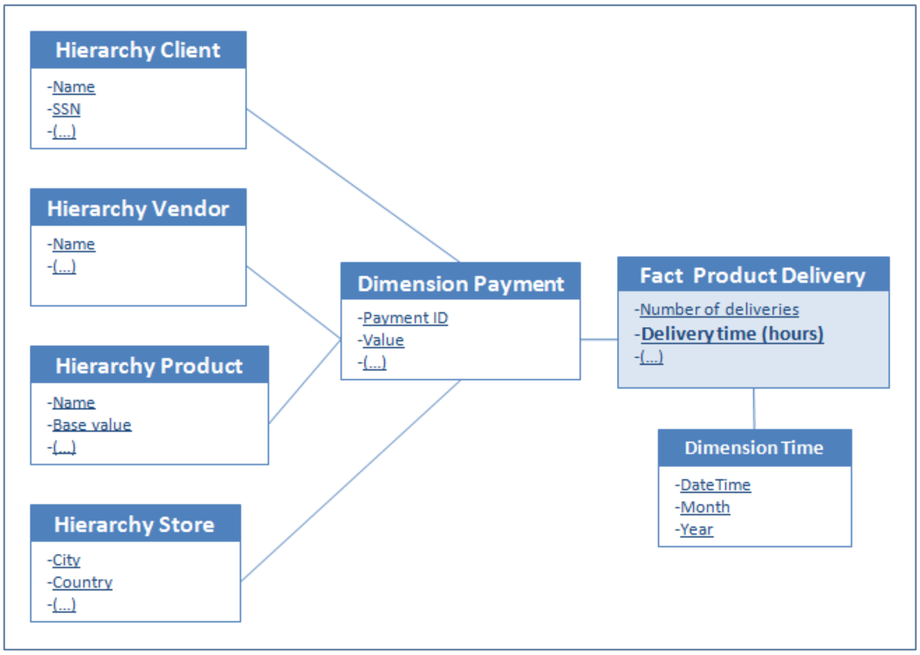


Figure 4.15: `Payment` event causing `product delivery` as MD schema

4.1.5 Situation changes as *MD schema* for cause-effect analysis

Situations represent possible states of affairs of the reality and changes through events (section 3.2.3). An event can be triggered by a situation or it can bring about a situation. In this rule we introduce a MD schema pattern to provide analysis of situation changes by these two relations with events. To do so, a fact is created to describe the situation change by one event and triggered by another event. The situation itself is set as a dimension as well as each event. Similar to rule R3, attributes and hierarchies of those dimensions can be derived from the events participations. The formalization of this rule is set as (Figure 4.16):

R5.a. If *e1, e2* are **Events** and *s* is a **Situation** that has a **triggers** relation to *e1* and a **brought-about** relation to *e2*, then *e1*, *e2* and *s* can be derived as *Dimensions d1*, *d2* and *d3* respectively, and a *Fact f* that relate these three *Dimensions*;

R5.b. If *f* is a *Fact* defined by R5.1 with **Situation** *s,* **Events** *e1* and *e2*, then the numeric attributes of *s, e1* and *e2* can be suggested as *Measures* for *Fact f*;

R5.c. If *ds* is a *Dimension* provided by R5.1 from a **Situation** *s*, *s* is a **Situation of Situation** and *s'* is the **Situation** composition of *s*, then it can be derived as a *Ragged Hierarchy*.

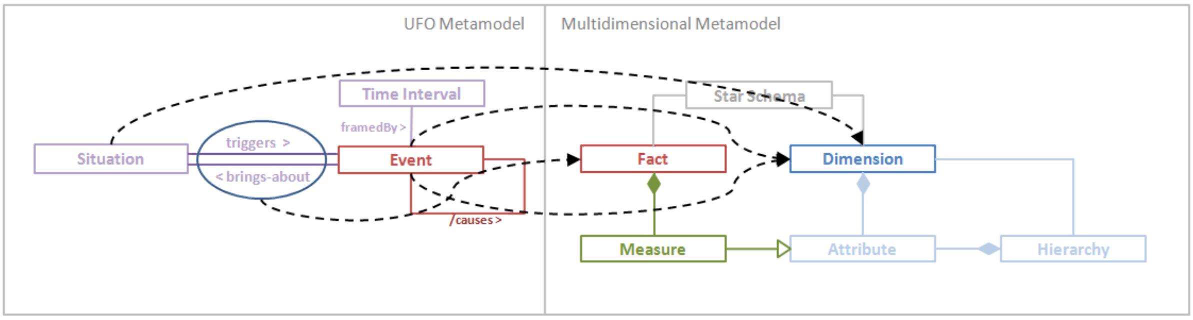


Figure 4.16: MD schema pattern to analyse **Situation** cause-effect

An interesting cause-effect analysis of the events that affects foreseen situations can be performed over this MD schema. Suppose an internet banking domain, inspired in (MIELKE, 2013), where security concerns about connected situations is necessary. A situation of suspicious parallel login can be brought about the login event when two or more connections happen for the same login from different hosts. In addition, as a result action, this situation can trigger the alarm event, for example. By the application of this rule the resulting MD schema can be used to analyze the behaviour of the causing (login) and performed (alarm) events before and after the suspicious parallel login situation. Moreover, the participations in both events can be considered perspectives of analysis, such as the login location and the alarm target, as shown in Figure 4.17. An analysis example would be: how often the situation suspicious parallel login occurs by country/city location, IP address and date (year, month and day). Regarding situation of situation instruction as a ragged hierarchy, this is due to the fact that both concepts are used to group other situations and dimensions, respectively.

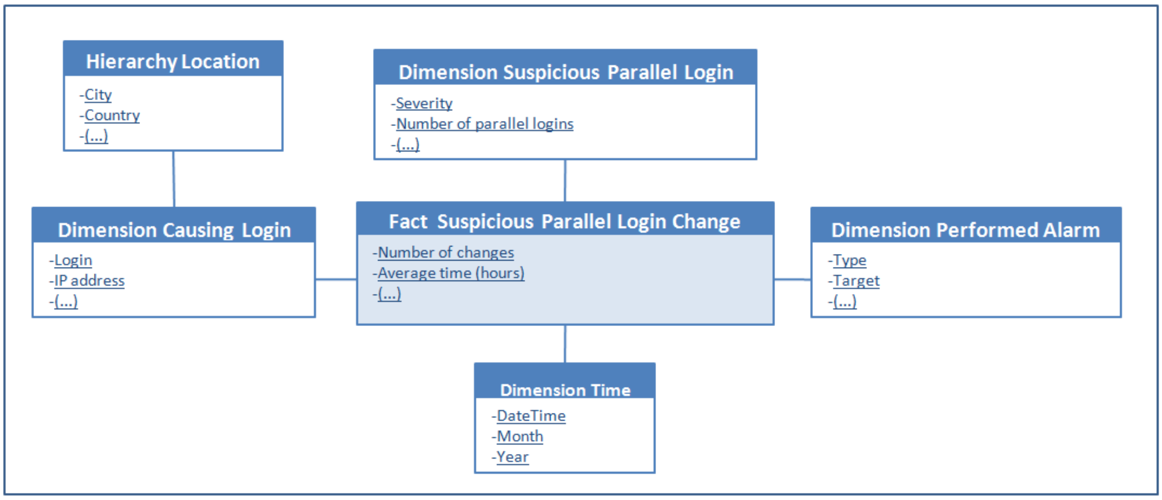


Figure 4.17: MD schema for cause-effect analysis of `suspicious parallel logins`

4.2 Hybrid multidimensional design task for heterogeneous data

In this section we discuss how the derivation rules introduced in the prior section fits in a MD design approach. In addition, we demonstrate how unstructured data sources can be used during the modelling activity. Business analysis and design phases are considered the most important activities of this methodology because it can be a critical divisor between the success and the failure of the project. The increase of the MD modelling efficacy can reduce future costs in maintaining the BI/DW solution by avoiding conceptual errors through some type of formalization of the common understanding. Moreover, the hybrid MD design activity, i.e. source-driven and analysis-driven as parallel processes, is the most common technique. It can produce more complete and assertive MD schemas, trying to combine what already exists in data sources with tacit domain knowledge from end-users analytical requirements. Therefore, this line of thought is maintained in our proposal, being classified as a hybrid MD design task adaptation.

Our adaptation is a similar process to the proposed in Advanced DW (MALINOWSKI e ZIMÁNYI, 2009) and GEM (ROMERO, SIMITSIS e ABELLÓ, 2011), revised in section 2.2. Both begin with parallel source-driven and analysis-driven activities to develop the initial schemas. Thereafter, these schemas are reconciled somehow and, then, enriched with particular semantics. Advanced DW adds spatial and temporal aspects to the ER representation of the, so called, conceptual MD schema, whilst the later explores functional dependencies of the data sources trough an ontological approach, deriving MD concepts and combining them with the analysis requirements (as SQL queries).

In Moss's methodology (MOSS, 2003) the MD design activity fits into business analysis and design phases. Moss lifecycle process has an alignment with Kimball's and Malinowski's methodologies. The main difference is the addition of the metadata repository construction during the project. It presents a balanced approach, considering complexity and practice. Each activity is set to a specific phase, as described at the left of Figure 4.18. As revised in chapter 2, it starts with the justification phase, where the business case assessment is created. Then, the planning phase uses it as input to produce the infrastructure evaluation, project plan, activities, resources and schedules. Next, in the business analysis phase, information requirements are collected and sources of data are analyzed, starting the MD and ETL process design. Then, in the design phase, the DB physical model is designed and the ETL process and OLAP interface are implemented during the construction phase. At last, the BI/DW solution is deployed in the production environment.

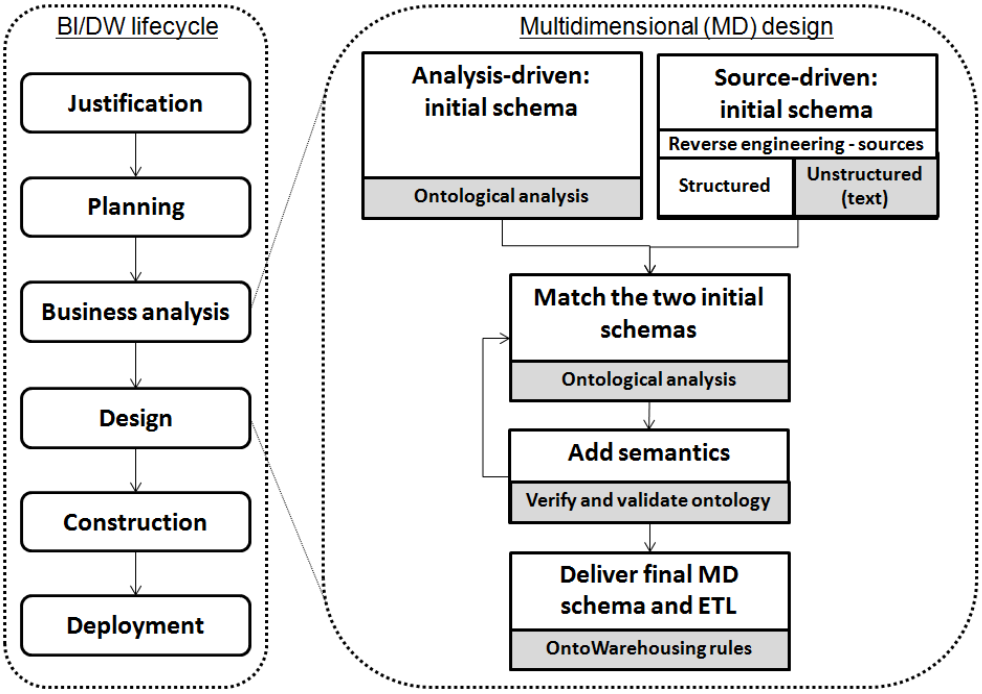

BI/DW lifecycle
Justification
Planning
Business analysis
Design
Construction
Deployment
Multidimensional (MD) design
Analysis-driven: initial schema
Ontological analysis
Source-driven: initial schema
Reverse engineering - sources
Structured
Unstructured (text)
Match the two initial schemas
Ontological analysis
Add semantics
Verify and validate ontology
Deliver final MD schema and ETL
OntoWarehousing rules

Figure 4.18: MD design process adaptation proposal

Our aim in this section is to describe our adaptation of the hybrid MD design process through a methodological process definition, being agnostic of technologies and expansible, so other existent hybrid approaches can be integrated in future work. As seen in last section, once the domain ontology is well founded with UFO through the ontological analysis, the set of rules introduced in section 4.1 can be used to derive MD concepts as suggestions to the MD designer. In addition, from UFO concepts, their relations and intrinsic constraints, it is possible to define high-level extraction and transformation rules, as described in rule R3. Therefore, the main goal of this process is to produce the MD schemas from well-founded domain ontologies.

In the analysis-driven part, the modeller can use the domain knowledge from domain experts, existing procedures, glossaries, taxonomies or other terminological standards. Moreover, if other hybrid approaches are used, then their ontologies (representing data sources) can be matched. Having the domain ontology constructed and well founded with UFO, the modeller can verify and validate it, refining the model in a cyclical way, increasing the domain ontology quality. Afterwards, the rules defined in section 4.1 can be applied to derive possible MD concepts. At last, the MD conceptual schema is built based on the concepts derived from the rules. The final MD schema derived can be adapted to others existent MD schemas. Each of the activities is fully explained as follows:

- **Analysis-driven design: analyse business**

Analysing the business domain is a common activity in a BI/DW project lifecycle. We propose that it occurs as a top-down conceptualization process, having the observed reality represented by the designer point of view through ontological analysis approach. The business experts should be consulted to assert business rules, even if they are not the main stakeholders of the project. Each domain concept should be correctly named, uniquely identified and validated by all business people who will be accessing the data. This task can be supported by interviews, as usual in requirements elicitation, and business official vocabulary definitions, such as glossaries and standards. The well-founded domain ontology should be built upon UFO concepts and independent of processes and technologies, i.e. the domain representations should not be influenced by any type of software (e.g. DB) or hardware. AS-IS business processes should be understood, so the behaviour of the concepts, e.g. their creation or modification, is mapped to the entities in the domain ontology.

- **Source-drive design: reverse engineering from structured data sources**

  Reverse engineering from structured data sources were already addressed by related works as supply-driven approach, such as AMDO, revised in chapter 2. In the majority, they check the existent functional dependencies among tables and their relationships, cardinalities and constraints. Then, MD concepts can be derived automatically based on a set of heuristics. This activity is usual in BI/DW projects, being useful because it can capture some important business rules and policies, such as integrity rules, that could not be captured during the interview sessions, enriching the domain ontology. Nowadays the majority of CASE tools can support this activity, such as Power Designer and Enterprise Architect (EA).

- **Source-drive design: reverse engineering from text**

  The goal of the reverse engineering from text is to generate the representations of the entities and their relations from unstructured data sources. This can be made in two ways: manually or automatically. In both cases a set of text corpora is selected from the sources, with the support of the business experts, and its content is analyzed. In the former, the designer sketches the domain ontology based on the main entities found in the text by reading it. The later is the case where specific Natural Language Processing (NLP) and Information Retrieval (IR) techniques are applied to the corpora, which can automatically generate models suggestions. In this sense, entity and relations recognition techniques play an important role on automatically generating the domain ontology. Tools that implement these techniques are based on lexical methods, such as orthographic correction, stop word elimination, tokening, synonymous resolution, stemming, morphological classification and some type of semantic categorization from business terms. The result artefact from this activity is a sketch of the domain ontology extracted from the text samples, containing the domain represented by the unstructured data-sources. From both reengineering processes (structured and unstructured), it is crucial to make annotations about the origins of the information, i.e. the data source elements for the concepts sketched.

- **Match two initial schemas: enrich the well founded domain ontology**

The possible inputs for this phase are the model sketches from both analysis-driven and supply-driven approaches. The output of this activity is the domain ontology modelled based on UFO concepts, consolidating both structured and unstructured data sources. Common concepts found in those representations should be matched or associated, by annotating their data structures origins.

For example, in an analytical solution of `medical appointments` and their relations to `patient diseases`, registered textually in `medical records`, the representations from the unstructured data source could contain `patient, complaint, disease, heart frequency and blood pressure` concepts. In addition, by analysing the business domain, the modeller could have designed a model with `patient, doctor and appointment` entities. An example of UFO usage, to enrich semantically the conceptual domain design, would be setting the `patient` entity as a **Role** of the **Kind** `person`, having the formal relation **participationOf** to an `appointment` **Event**, which also has the `doctor` participation. Through a source-driven approach, an ontology could be generated from the structured data sources of the `clinical` transactional information system, containing `patient, appointment, payroll, schedule` entities. `Patient` entity is the same (unique) ontological class from all sources, so it should be matched. Therefore, it can be extracted from both unstructured and structured data sources, annotating its origin to be referenced afterwards in the conceptual ETL. In addition, `patient` entity from the textual `medical` record could have its identity set by his `name` in the header of the document. From the structured data source its identity is set by a primary key, but the table also provides the `name` attribute. All those information will be necessary to build the linkages between the structured and the unstructured universes when designing the ETL process. After matching all entities, annotating their origins, a consolidated model should be built.

Therefore, the main advantage of our adaptation in this activity is the semantic power increase with the application of UFO categories in the ontology construction to better represent the domain. Besides all well-known advantages of ontological analysis, we can derive possible MD concepts by applying the rules introduced in this work. Nevertheless, there are critical limitations related to the amount of entities found in the unstructured data sources analysed. An example is in a situation of a distributional semantic method being applied to automatically generate the domain ontology from textual sources results ten thousand entities with several relations among them. It is almost impossible for a human to design a well-founded ontology based on so larger ontologies.

- **Add semantics: verify and validate (V&V) the domain ontology**

In the Verify and Validate (V&V) task, the designer analyzes the foundational constructs and checks if the entities and relations from the model are semantically consistent, also verifying business rules violations. This is made by analysing the foundational constructs and checking if the entities and relations from the model are semantically consistent. This kind of activity is common in CM and it improves the quality of the designed ontologies. It receives as input the domain ontology built upon UFO concepts, as seen in the prior step. The participation of the domain experts during this task is fundamental to improve the quality of the domain ontology, ensuring that the model is semantically correct and covers the main entities involved in the business requirements, avoiding ambiguity among concepts. This is a cyclical process because, once the designer finds an error or an inconsistency, he can fix the model and validate it again, improving the model expressiveness.

The domain ontology can be verified by a formal constraint language, such as OCL , in case of using UML as language to represent the domain, and validated via visual simulation, such as Alloy analyzer. Thus, the designer can try to certify that the modelled world is the intended state of affairs admitted for the domain. Verifying and validating ontologies with many concepts can be unfeasible for humans because of its size and complexity. Thus, a common practice is to choose ontology parts (sub-domains), apart from the rest of the model, validate each part separately and then merge these parts. The resulting artefact in this activity is the well founded and validated domain ontology. The main advantage of performing this task is to refine and test the domain ontology, increasing its quality, so it reflects the domain being modelled by representing the admissible state of affairs. Furthermore, by using UFO temporal concepts, such as events, participations and temporal relations, it is possible to derive MD concepts from the set of the rules suggested in section 4.1. The main limitation is, again, the size of the ontology. If it provides too many concepts, even by separating the sub-domains to validate, it can be a painful task.

- **Deliver final schema: derive multidimensional concepts from UFO concepts**

In this activity the final MD schemas are designed based on the domain ontology built on the prior activity and other existent MD schemas. In common DW methodologies this task depends purely on decisions from the MD designer based on informal guidelines, i.e., it depends on tacit knowledge, and therefore, it is error prone. In our approach we defined a set of mapping rules to derive possible MD structures from the well-founded domain ontology and its temporal aspects, such as events, temporal relations and objects participations. Therefore, the MD designer can use this method to increase its assertiveness in designing the final MD schemas. The mechanism to derive MD concepts begins by reading the domain ontology and looking for the foundational ontology categories. Once they are found, it executes the mapping rules and presents to the designer the possible MD structures inferred.

Thereafter, the MD concepts derived are conciliated with existent ones if MD schemas are already used in the organization, so the MD concepts can be conformed and refined, providing new analytical possibilities. At last, the designer defines the final MD schemas with data sources annotations for the ETL processes, as comments in natural language. In addition, other ontological MD hybrid approaches, such as [25] can be used in parallel, combining the final MD concepts produced.

4.3 Conclusion

In this chapter we have systematically introduced OntoWarehousing approach, proposing a set of mapping rules from UFO concepts to MD concepts, considering temporal classifications in the domain ontology. With these rules, the process of choosing accordingly the representations of real world can enhance the MD design activity. At first, the events were explored to derive facts. Secondly, possible perspectives of analysis could be mapped from events participations. Thereafter, time interval relations could be mapped to a snowflake schema pattern. Then, a fact/dimension dichotomy was suggested to represent the causality relation between two events. At last, situation change was analysed and, based on it, a MD schema was proposed for cause-effect analysis.

The second part of the approach contextualizes how to adopt the derivation rules, above mentioned, in a hybrid MD design activity through a method. In addition, it considers prior works (MOREIRA, CORDEIRO e CAMPOS, 2009) (ALMEIDA e SILVA, 2009) (MOREIRA, CORDEIRO e CAMPOS, 2013) to cope with a valuable source of data: textual. Therefore, our own adaptation of an existing apparatus was explained. Moreover, we could specify the general abstract concepts in a valid way to the BI/DW lifecycle domain, specifically in MD design activity. In addition, the adequate motivation for the proposed approach was presented, ensuring that it follows logically from the research problem and the theoretical framework. We believe that sufficient information is provided for a reader to replicate or evaluate our proposal.

5 Application examples

Our aim in this chapter is to present how the theoretical concepts, revised in chapters 2 and 3, were represented and applied. At last, we state how our goals have been corroborated, also discussing the limits of our findings for each example. We validate the proposed approach effectiveness by discussing its application in two different scenarios, evidencing its benefits in choosing appropriate MD concepts from the rules executions in well-founded domain ontologies. Our focus is the derivation process, but the adapted hybrid MD design method considering heterogeneous data sources is also exemplified in the first example. The method execution is experimented as part of the BI/DW solution lifecycle for ONS corporate image analysis in the Brazilian electrical system domain. Rules R1, R2 and R3 are executed by a prototype tool and a MD schema is designed for the analysis needs. Thereafter, the ETL process and involved DBs were implemented and data loaded in a data cube, being explored through an OLAP tool. The second example discusses rules R4 and R5, by analysing causality and situation changes in the ITIL process ontology from a prior work (CALVI, 2007). Mapping executions are simulated and analyses are exemplified based on the result MD schemas.

In section 5.1 we present a prototype implementation to semi-automatically derive MD concepts from an UFO-based domain ontology by applying the rules proposed in section 4.1. The hybrid MD design activity proposed in section 4.2 is exemplified in section 5.2, considering both prototype execution and unstructured data source use. The example is presented by passing through the main activities in the BI/DW lifecycle (Moss's methodology), considering news clippings as unstructured data source and disturbances DM as structured data source. We present: (i) analysis requirements; (ii) domain ontology construction; (iii) derivation of MD concepts from prototype execution; (iv) MD schema development; (v) ETL process deployment and; (vi) an OLAP application exploring joint analyses of the data cube. In section 5.3 the example regarding R4 and R5 on ITIL ontology is presented, supporting causality and situation changes. For each example, the results are analysed, presenting advantages and limitations of our approach.

5.1 Prototype implementation

In this section we describe the implementation of a prototype to support the derivation process of MD concepts from a domain ontology designed upon UFO concepts. Our aim is to present an automatic mechanism to exemplify the mapping rules proposed in section 4.1. We chose to develop our prototype based on the visual modelling CASE tool Enterprise Architect (EA)[19] with OntoUML plugin because it already presents the language which provides UFO stereotypes, as described in section 3.2.5. The OLED[20] software provided by NEMO research group was also chosen. It provides verification and visual validation by examples/counterexamples and the possibility of using OCL reasoning for ontologies written in EA/OntoUML.

Therefore, the prototype non-functional requirement for interoperability is to accept the exported domain ontology from EA/OntoUML, specifically in XMI 2.1 format with UFO main formalisms. At the time of writing, OntoUML language does not yet implement *perdurants* (UFO-B) stereotypes. Therefore, the technique used to deal with this problem was based on the following choices:

1. Use of the non-official UFO package library in EA/OntoUML provided by NEMO with separated packages that represent concepts such as **Event**, **Participation**, **Temporal Relation**, among others (Figure 5.1):

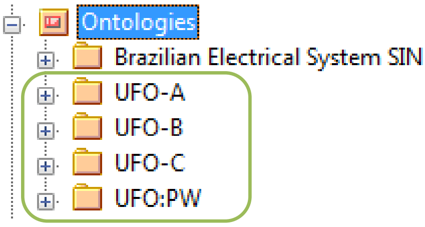


Figure 5.1: UFO packages used in the solution

Observation 1: These packages do not implement exactly OntoUML nor UFO and there are still some related conceptual issues under discussion about UFO parts integration. For example, in those packages, the “hou” stereotype, meaning **High**

[19] http://www.sparxsystems.com.au/

[20] https://code.google.com/p/ontouml-lightweight-editor/

Order Universal, was used. It is not represented in OntoUML yet, but it represents an universal that classifies other classes from second order level;

Observation 2: UFO:PW package was created for integration convenience, from which atomic-complex concepts from UFO can specialize. These packages try to solve some preliminary issues regarding UFO-C formalization and present parts that are also still under discussion. For example, the **Kind** representation in UFO-A package was chosen to harmonize the language, but there are inconsistencies.

2. The **Event** stereotype provided by EA default structural class (UML) was defined to represent an **Event** (as the same meaning from UFO) when it is specialized from **Event Individual** class (represented as **Category** in UFO-B package).

3. **Participation** is represented by a class stereotyped as an **Event**, specialized from the **Participation Individual** class (represented as **Event** in UFO-B package).

4. The **participationOf** (**dependsOn**) relationship is defined as an OntoUML **formal** relationship, which links a class represented as **Participation** to a class defined as **Participant** (represented as **RoleMixin** in UFO-B package).

5. The **temporal relations** (**Allen's operators**) are represented by OntoUML **formal** relationships between two **Events**. To differentiate them (e.g. **overlaps**, **meets**, etc) an EA metadata property was chosen: *Alias*.

5.1.1 Functional requirements

The highest level functional requirement (f-req) of this prototype is to implement rules R1, R2 and R3 proposed in section 4.1, having as input a domain ontology designed in EA tool with OntoUML ad-in profile, available as XMI file in version 2.1 format. Based on this assumption, the functional requirements are described below:

1. Realization of R1: The derived *Facts* from UFO-B **Events** (and their mereological pattern representing **Atomic** and **Complex Events**) should be listed and presented to the user;

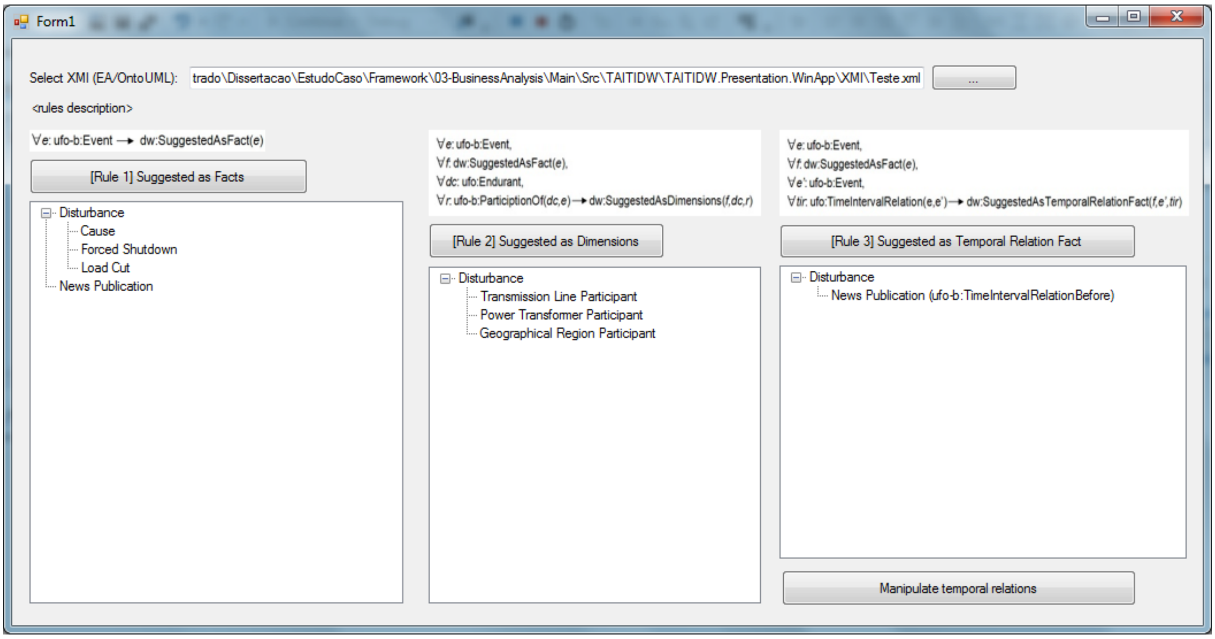
2. Realization of R2: Once an **Event** is chosen to be represented as a *Fact* (in the last step, output of f-req 1), the prototype should list all its **Participants**, which can potentially be derived as *Dimensions*. Also, it should present the associations to the classes defined as structural concepts (from UFO-A) from each **Participant** of the selected **Event**, potentially suggesting them as *Hierarchies* of the *Dimension* for the selected *Fact*;

3. Realization of R3 - first interface: It should provide a mechanism through which all **Temporal Relations** (**Allen's operators**) found in the domain ontology are presented to the user;

4. Realization of R3 - second interface: It should provide an interface where the user can choose two **Events** and then the **Temporal Relations** between them are listed. Also, it should present the MD pattern defined in R3 of both selected **Events** and their **Temporal Relations**. In addition, it should provide the SQL constraints derivations to be used for data transformation and loading (ETL process) of the *Fact* at the *MD schema*, according to the **Temporal Relations**.

5.1.2 Construction

The prototype was built under .NET platform (framework 4.0), C# language and Windows Form interface within Visual Studio 2012 IDE. The main reason for this choice was the author experience with this technology. The architecture chosen for the prototype solution contemplates two tiers: a business layer and an interface layer. The business layer is responsible for the algorithm for mapping the XMI file to the ***Perdurant*** classes in OO classes: **Atomic** and **Complex Events** and **Participations**, **Time Point** and **Interval**, **Relationship** and **Substantial**. Also, it executes the rule mappings to MD concepts. The interface supporting the requirements is illustrated in Figure 5.2.



Form1

Select XMI (EA/OntoUML): trado\Dissertacao\EstudoCaso\Framework\03-BusinessAnalysis\Main\Src\TAITIDW\TAITIDW.Presentation.WinApp\XMI\Teste.xml

<rules description>

∀*e*: ufo-b:Event → dw:SuggestedAsFact(*e*)

∀*e*: ufo-b:Event,
∀*f*: dw:SuggestedAsFact(*e*),
∀*dc*: ufo:Endurant,
∀*r*: ufo-b:ParticiptionOf(*dc*,*e*) → dw:SuggestedAsDimensions(*f*,*dc*,*r*)

∀*e*: ufo-b:Event,
∀*f*: dw:SuggestedAsFact(*e*),
∀*e'*: ufo-b:Event,
∀*tir*: ufo:TimeIntervalRelation(e,e') → dw:SuggestedAsTemporalRelationFact(*f*,*e'*,*tir*)

- Disturbance
  - Cause
  - Forced Shutdown
  - Load Cut
- News Publication

- Disturbance
  - Transmission Line Participant
  - Power Transformer Participant
  - Geographical Region Participant

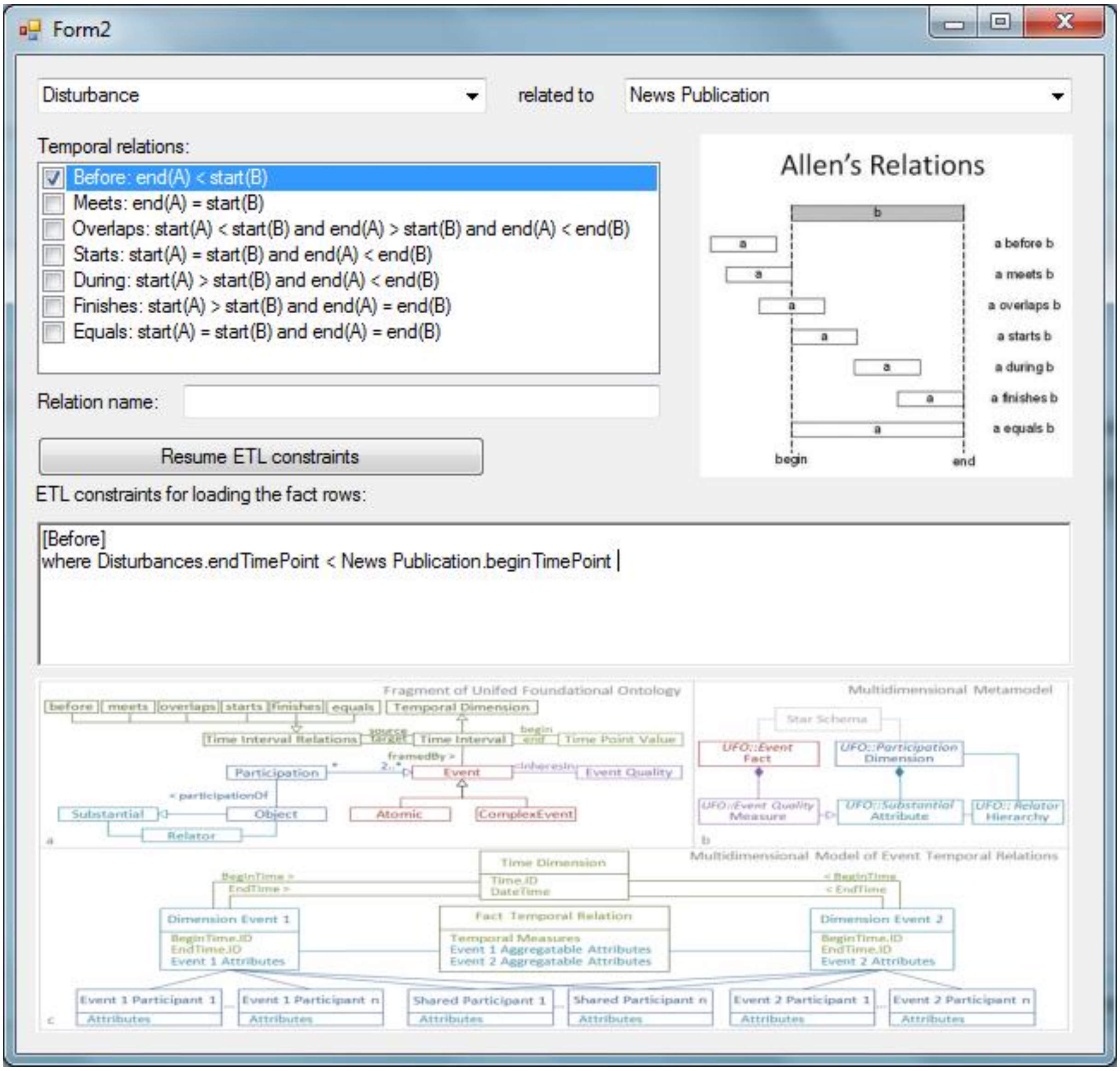
- Disturbance
  - News Publication (ufo-b:TimeIntervalRelationBefore)

...
[Rule 1] Suggested as Facts
[Rule 2] Suggested as Dimensions
[Rule 3] Suggested as Temporal Relation Fact
Manipulate temporal relations

Figure 5.2: Prototype main screen

To use it, at first, the path to the domain ontology as a XMI file should be informed. Thereafter, the user can execute the first requirement (R1) by clicking on the first button (at left), where all **Events** and their parts (**Complex** or **Atomic Events**) are listed as a tree view (hierarchically). The user can then select one of them to be set as a *Fact*, an **Event** to be analysed from different perspectives. By pressing the second button (in the middle), the direct *Dimensions* suggested to be linked with the *Fact* are listed in the first level. In addition, the possible *Hierarchies* for them are presented as a tree view, at lower levels. All the **Events** set with **Temporal Relations** to other **Events** in the domain ontology are listed in the third tree view when pressing the third button (at right). The tree view presents in the first level the **Events**, in the second level the related **Event** and the correspondent temporal relation(s) (e.g. **before**, **overlap**) in parenthesis.

When selecting the "Manipulate **Temporal Relations**" button, a form is opened as a pop-up (Figure 5.3). It provides two items lists (drop-down components), which contain the **Events** found in the domain ontology. The previously selected **Events** are marked as default in these drop-down lists. Moreover, the user can select different **Events** to manipulate other possibilities of **Temporal Relations**. Once each drop-down has an **Event** selected, than the existent **Temporal Relations** from the original domain ontology are presented and checked in the checkbox list. By checking the temporal relations and clicking Resume ETL constraints, depending on the **Temporal Relations**, the list of WHERE clauses (in SQL) considering Allen's constraints – included later in the ETL process to load data to the *MD schema* – are presented in the text box beellow. It considers each **Event** (and fundamental **Participants** properties) as a *Dimension*, implemented through one data table. In the Figure 5.3, **before Temporal Relation** between `Disturbance` and `News Publication` **Events** is checked and it's respective *WHERE* clause statement is presented.



Form2

Disturbance related to News Publication

Temporal relations:

☑ Before: end(A) < start(B)
□ Meets: end(A) = start(B)
□ Overlaps: start(A) < start(B) and end(A) > start(B) and end(A) < end(B)
□ Starts: start(A) = start(B) and end(A) < end(B)
□ During: start(A) > start(B) and end(A) < end(B)
□ Finishes: start(A) > start(B) and end(A) = end(B)
□ Equals: start(A) = start(B) and end(A) = end(B)

Relation name:

Resume ETL constraints

ETL constraints for loading the fact rows:

```
[Before]
where Disturbances.endTimePoint < News Publication.beginTimePoint
```

Figure 5.3: Prototype interface to manipulate the temporal relations pattern

5.1.3 Limitations

The main limitation is outside the prototype scope, which is that, until that time, OntoUML language does not provide the stereotypes for representing of ***Perdurants*** (UFO-B) concepts, such as **Events**, **Participations**, **Temporal Relations** and **Situations**. Therefore, there is the disadvantage of losing the expressiveness given by the first order logic formalisms and rules. Moreover, representing **Temporal Relations** through EA metadata property *Alias* do not contemplate temporal constraints defined by the axioms for each relationship.[21]

5.2 Application example 1: impact of disturbances on institutional image

In this section we discuss the application of our approach in a real scenario of the Brazilian electric system. First, we present an overview of the electric grid security domain regarding disturbances in the national electric system and associated news publications. Next, we describe the existing information systems that support both transactional (operational) and analytical processes related to the disturbances registration, triage and analysis. Afterwards, the business needs for integrating the disturbance analysis system with published news about the electrical system (for institutional image evaluation) are listed. Then, we present how the hybrid process proposed in section 4.2 can be used to model the MD schema. To evidence its effectiveness, historic data is loaded through an ETL process and possible analyses are presented with their results. At last, results of the approach application in this scenario are analyzed.

5.2.1 Business scenario

The Electric System National Operator[22] (ONS) is an entity in law private and non-profit organization, performing its duties under the supervision and regulation of the

[21] The prototype was built without some common concerns when developing software, such as: code quality, testing mechanisms and algorithms complexity (e.g. to find the stereotypes in the XMI XPath queries it were not used nor mapping patterns). The decision to build with these problems was made based in the assumption that it will have to be re-built from scratch when OntoUML supports *perdurants*.

[22] http://www.ons.org.br/institucional_linguas/relacionamentos.aspx?lang=en

National Electric Energy Agency (ANEEL). Its mission is to operate the Integrated National System (SIN – "Sistema elétrico Interligado Nacional" in Portuguese) with transparency, fairness and neutrality in order to ensure security, continuity and economy optimization of electricity supply in Brazil.

The SIN is the set of facilities and their equipment – such as power plants, substations, transmission lines, power transformers, among others – responsible for supplying electricity for 97% of the Brazilian territory. It is a large hydrothermal system, with a strong predominance of hydroelectric plants with multiple owners – the companies of the electrical sector (known as "agents"). Hydroelectric power plants are usually located far from the load centres, requiring an extensive network of electricity transmission, which covers companies in geoelectrical regions (south, southeast, mid-west, northeast and parts of the north). Equipment in the system grid are aware to faults and failures of various natures, causing forced shutdowns of one or more devices in the transmission system and can interrupt the power supply to consumers. These occurrences are known as electrical disturbances and may be caused by atmospheric electrical discharges, floods, fires or human failures, among others. In the ONS official technical terms glossary, an electrical disturbance is defined as:

"An occurrence in SIN characterized by forced shutdown of one or more of its components, which cause any of the following consequences: loss of load, shutdown of the system components, equipment damage or violation of operating limits."

The processes to fulfil the coordination and the control of the SIN operation are based on technical procedures, rules and criteria defined in normative documents, called Network Procedures [23]. They are divided in 26 modules which standardize ONS macro processes. The analysis of occurrences and disturbances is the main related process in our experimentation.

Information systems were developed by ONS to support the registration of disturbance occurrences (categorized as abnormalities, undesirable events or unsatisfactory performance), such as the Disturbances Integrated System (SIPER), Disturbance Oscilograms Integration System (SPERT) and the Calculation of Transmission

[23] http://www.ons.org.br/procedimentos/index.aspx

System (SATRA). They are all integrated through ONS master relational DB, called Electrical System Technical DB (BDT).

In SIPER, all disturbances involving equipment belonging to the network are analysed. Operating companies should inform all relative data from disturbances in the equipment and facilities under their responsibilities. A process of consolidation is made between ONS and companies through this system, providing precise and high quality information, which is stored in BDT as structured data. Textual information is also provided in this process, such as the protection team action descriptions and the detailed sequence of events of a disturbance (which are recorded in DB table columns as free text). Later, the information is analysed, though a conventional BI/DW solution, in order to assess the behaviour of the network during the disturbances. Based on it, solutions are provided for the problems encountered, supporting the operation planning and execution.

To address analytical aspects of disturbances and forced shutdown occurrences, supporting disturbances analysis macro-process, ONS has a “tailor made” analytical information solution based on BI/DW architecture, called Disturbances BI. It consolidates the data from transactional systems which support the process of information registration and classification from BDT (SIPER, SPERT and SATRA). The integration is made through a conventional ETL process (implemented in SSIS[24]) over structured data from BDT (DBMS Informix[25]), being available in a Disturbances DM (stored in MS SQL Server DB engine). The users can navigate and generate reports to analyse data through OLAP tools, such as MS Excel and Business Objects (InfoView and Xcelsius) [26].

At first, the ETL process extracts data from BDT to load in on an intermediate repository, called Operational Data Store (ODS) DB. This data represents information about equipment, facilities, companies, geographical regions, disturbances, forced shutdowns, outages, among others. After being available in ODS, a second ETL process extracts data from ODS to load in a MD schema in the DW relational DB. This model is represented by the conceptual star schema in Figure 5.4 below, having as the main fact the disturbances

[24] ETL tool from Microsoft: SQL Server Integration Services (SSIS).

[25] Informix is a DBMS provided by IBM.

[26] Business Objects is the BI platform provided by SAP, having the OLAP tools InfoView (web-based) and Xcelsius (desktop based in MS Excel).

with measures: disturbances quantity, number of load cuts, load cut value (in MWh), interrupted energy value (in MWh), disturbance recovery time, load cut recovery time and the connectivity node recovery time. The disturbance occurrence and its measures can be analysed from different perspectives (dimensions), such as voltage level, equipment, disturbances causes, detailed characteristics, related regional areas, related companies and time.

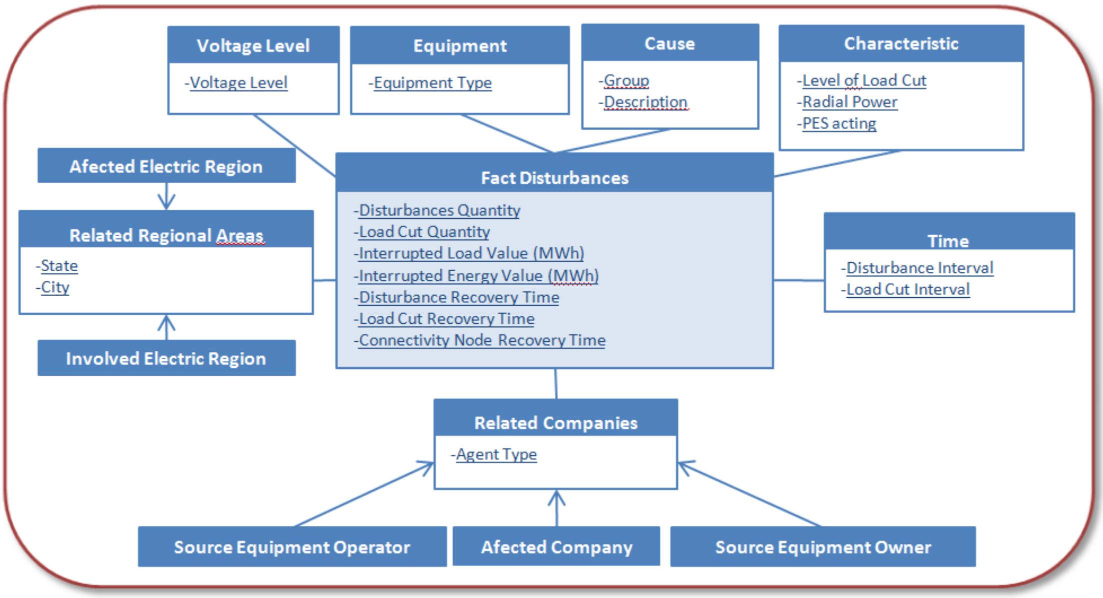


Figure 5.4: Disturbance conceptual MD schema

This MD conceptual schema was physically designed and implemented in a relational DB and is available to be accessed by ROLAP tools. A third ETL process was built to deliver this relational DB to a data cube, available to be explored through MOLAP tools. Important Key Performance Indicators (KPIs) can be extracted from this MD schema, such as the severity indicator for the system performance, based in the calculated measure Minute-System (SM), defined in the Disturbance Analysis process as the following formula:

$$SM = \frac{Interrupted\ Demand\ Load\ (MW) * Average\ Interruption\ Time\ (minutes)}{Maximum\ Peak\ Load\ Demand\ in\ the\ period\ (MW)}$$

The indicator is defined in five levels of severity:

- Normal: $SM < 1$
- Not severe: $1 < SM < 10$

- Severe: $10 < SM < 100$
- Very severe: $100 < SM < 1000$
- Extreme severe: $SM > 1000$

Despite the SM KPI, important analyses can be made when navigating through the data cube, such as the comparison of disturbances with and without load cuts, most common source equipment type, which originate disturbances, disturbances originated in transmission lines and cause-consequence analysis regarding affected areas, among others. Common analyses made in the original data cube are illustrated in appendix.

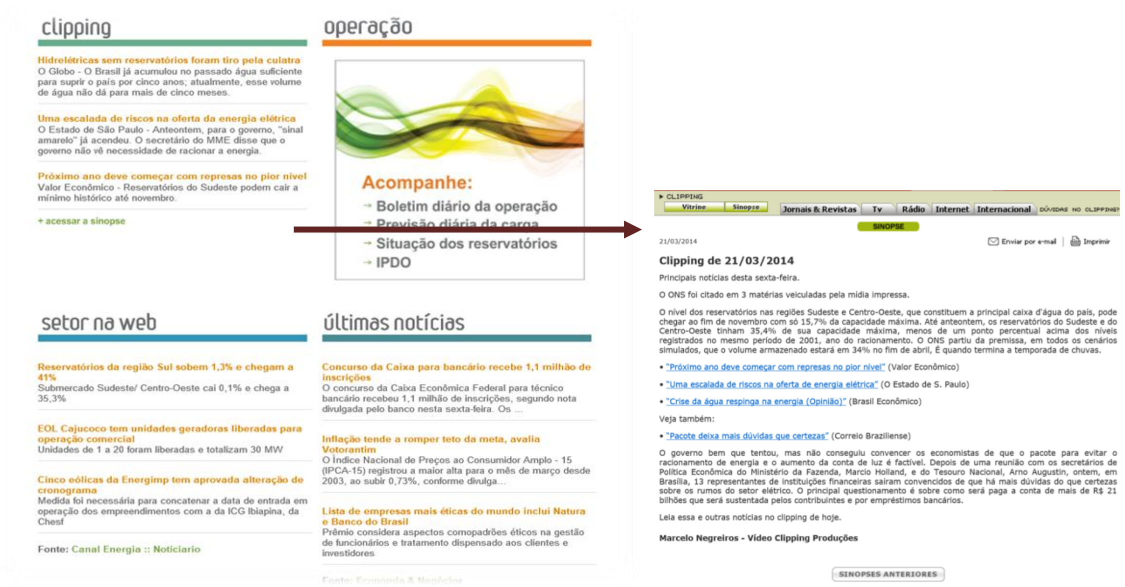
5.2.2 Application of the proposed approach

We follow the essential activities of the BI/DW methodology, particularly the hybrid MD approach proposed in section 4.2, testing the mapping rules defined in section 4.1, supported by the prototype built and the appropriate tools. To contextualize the MD design activity, we described justification, business analysis, design and construction phases of the chosen BI/DW methodology (MOSS, 2003), revised in section 2.3. This was made because the whole process is complex, having direct interconnections with MD modeling activity, being affected by the kind of data used (structured or unstructured).

5.2.2.1 Justification

ONS provides a daily summary of news related to the electricity sector in its Intranet home page, named Clippings. A web-part presents three publications considered most important in the week with the link to access their synopsis, showing where ONS is cited. The main purpose of having such information system is to provide to collaborators of the company what is being published about the electric sector, quoting the organization when it is mentioned by the Brazilian press. Figure 5.5 illustrates the web-part highlighted in the top with a link for each of the three top daily news and a link to access all news publications. Be selecting it, a pop-up is opened presenting a summary of the news of the day, how many times ONS was cited, a brief summary of each one and the links to each news article available in PDF – the copy of the original news article published – and in

HTML. This information system is provided by an outsource firm, mostly with manual efforts (reading daily news from various press companies).



Hidrelétricas sem reservatórios foram tiro pela culatra
O Globo - O Brasil já acumulou no passado água suficiente para suprir o país por cinco anos; atualmente, esse volume de água não dá para mais de cinco meses.

Uma escalada de riscos na oferta da energia elétrica
O Estado de São Paulo - Anteontem, para o governo, "sinal amarelo" já acendeu. O secretário do MME disse que o governo não vê necessidade de racionar a energia.

Próximo ano deve começar com represas no pior nível
Valor Econômico - Reservatórios do Sudeste podem cair a mínimo histórico até novembro.

+ acessar a sinopse
operação
Acompanhe:
→ Boletim diário da operação
→ Previsão diária da carga
→ Situação dos reservatórios
→ IPDO
setor na web

Reservatórios da região Sul sobem 1,3% e chegam a 41%
Submercado Sudeste/ Centro-Oeste cai 0,1% e chega a 35,3%

EOL Cajucoco tem unidades geradoras liberadas para operação comercial
Unidades de 1 a 20 foram liberadas e totalizam 30 MW

Cinco eólicas da Energimp tem aprovada alteração de cronograma
Medida foi necessária para concatenar a data de entrada em operação dos empreendimentos com a da ICG Ibiapina, da Chesf

Fonte: Canal Energia :: Noticiario

Concurso da Caixa para bancário recebe 1,1 milhão de inscrições
O concurso da Caixa Econômica Federal para técnico bancário recebeu 1,1 milhão de inscrições, segundo nota divulgada pelo banco nesta sexta-feira. Os ...

Inflação tende a romper teto da meta, avalia Votorantim
O Índice Nacional de Preços ao Consumidor Amplo - 15 (IPCA-15) registrou a maior alta para o mês de março desde 2003, ao subir 0,73%, conforme divulga...

Lista de empresas mais éticas do mundo inclui Natura e Banco do Brasil
Prêmio considera aspectos comopadrões éticos na gestão de funcionários e tratamento dispensado aos clientes e investidores

Fonte: Economia & Negócios

21/03/2014

Clipping de 21/03/2014

Principais notícias desta sexta-feira.

O ONS foi citado em 3 matérias veiculadas pela mídia impressa.

O nível dos reservatórios nas regiões Sudeste e Centro-Oeste, que constituem a principal caixa d'água do país, pode chegar ao fim de novembro com só 15,7% da capacidade máxima. Até anteontem, os reservatórios do Sudeste e do Centro-Oeste tinham 35,4% de sua capacidade máxima, menos de um ponto percentual acima dos níveis registrados no mesmo período de 2001, ano do racionamento. O ONS partiu da premissa, em todos os cenários simulados, que o volume armazenado estará em 34% no fim de abril, É quando termina a temporada de chuvas.

- "Próximo ano deve começar com represas no pior nível" (Valor Econômico)
- "Uma escalada de riscos na oferta de energia elétrica" (O Estado de S. Paulo)
- "Crise da água respinga na energia (Opinião)" (Brasil Econômico)

Veja também:

- "Pacote deixa mais dúvidas que certezas" (Correio Braziliense)

O governo bem que tentou, mas não conseguiu convencer os economistas de que o pacote para evitar o racionamento de energia e o aumento da conta de luz é factível. Depois de uma reunião com os secretários de Política Econômica do Ministério da Fazenda, Marcio Holland, e do Tesouro Nacional, Arno Augustin, ontem, em Brasília, 13 representantes de instituições financeiras saíram convencidos de que há mais dúvidas do que certezas sobre os rumos do setor elétrico. O principal questionamento é sobre como será paga a conta de mais de R$ 21 bilhões que será sustentada pelos contribuintes e por empréstimos bancários.

Leia essa e outras notícias no clipping de hoje.

Marcelo Negreiros - Vídeo Clipping Produções

Figure 5.5: The Clippings in ONS intranet homepage and its link to the publications

The corporate image is the way the organization is perceived by society, tending to be classified as positive or negative, varying in intensity and depending on variables such as the opportunities, threats and its competence (CARDOSO e POLIDORO, 2011). Different from its identity, which is constructed from its policies and procedures, the corporate image has an external origin: the public mind.

The result of a disturbance in the system can lead to cutting of the power supply of a geographical area, popularly known as blackout, which has a direct relation to the load cut level in the Disturbance BI analytical solution. The negative consequences of a blackout to the population are numerous, generating large financial losses in all sectors of the economy. For this reason, Brazilian press gives great focus to the subject, often citing ONS when such situation occurs, which may influence its corporate image. Because of the nature of ONS work, most often it is noticed by the public when problems occur. Moreover, among ONS main concerns in the electrical security domain, to analyze faults caused by disturbances in the system and its impact on users' lives (reflected in the media) is much relevant. Therefore, the organization needs analytical tools to address this issue.

Current organization information systems to address these analytical needs present the information of disturbances and news about SIN independently and hard manual work is necessary for a joint analysis over large amounts of historic data, often making it impossible to reach the desired results. Therefore, an analytical information system for joint exploration of disturbances and their impact in news publications is necessary. A BI/DW project can be built to acquire this need.

The highest-level requirement for the proposed BI/DW solution was to provide business analysis capabilities regarding ONS corporate image affected by news publications when load cuts happen in SIN. Analyses about the terminology used in news publications from specialized press of electric sector when a disturbance occurs can support this high-level requirement. Therefore, the existing DSS solution that was mapped, at first, was the Disturbance BI solution. It also represents one of the operational data sources that could provide the necessary information, as the intranet Clippings information system. Both can provide transactional records from business processes involved in ONS macro process of analysis of occurrences and disturbances and corporate image analysis. Therefore, they are pointed as data sources in business case assessment. No specific procedures for corporate image analysis were found in network procedures.

5.2.2.2 Business analysis

From the high-level business requirements provided in the justification phase, the analysis of disturbances related to news articles publications, a solution based on the approach proposed in chapter 4 was conceived. Through its application, the prototype built based in derivation rules from UFO temporal aspects was executed, detailed as follows. The data analysis activity could be performed by considering both data sources listed in business case assessment (Disturbances BI and Clippings) and the hybrid MD task method proposed in section 4.2 was tested. It begins with analysis-driven execution parallel to source-driven execution. Afterwards, the domain ontology was designed, using OntoUML, in a cyclic process of verification and validation. Then, MD concepts were derived from rules execution and a MD conceptual schema was proposed. These steps are detailed as follows.

1. **Analysis-driven**

At first, the analysis of the domain to represent it associated concepts was made, supported by ONS official glossary, some domain analysts (power systems engineers) and the CIM IEC 61970[27] international standard. They are described below:

a. ONS official glossary and domain engineers

ONS official glossary has the main terms used in the electrical sector and their natural language (textual) definition. It is the sub-module 20.1 of the Network Procedures, serving as a common understanding about the main business concepts among ONS departments and companies actuating in the sector. Consultations on this document were made several times to assert the initial domain representation. In addition, when a specific term was not encountered in the glossary or there was an ambiguous conceptualization, occasionally, the domain experts, mostly power systems engineers (ONS collaborators) were consulted. They could assert specific rules, such as the membership relation between a disturbance and a `forced shutdown`, where a `forced shutdown` is part of one unique `disturbance`, for example.

b. CIM IEC 61970

CIM IEC 61970 is the international standard developed by the electric power industry and adopted by the International Electro-technical Commission (IEC) for information systems interoperation and common concepts agreement (common understanding). Particularly, the main part of this standard was chosen, the IEC-61970 (for energy management), because it brings core definitions of the electric power transmission and distribution domain, such as `equipment` and their sets (`equipment container`) as `power system resources`. As a practical advantage, this standard is available in EA tool as UML class structural package, organized in `Base` and `Dynamics` sub-packages. It was imported in the EA solution to serve as a conceptual base. Figure 5.6 illustrates these packages in the solution, highlighted with one of the most important entities: `equipment`.

[27] http://www.iec.ch/smartgrid/standards/

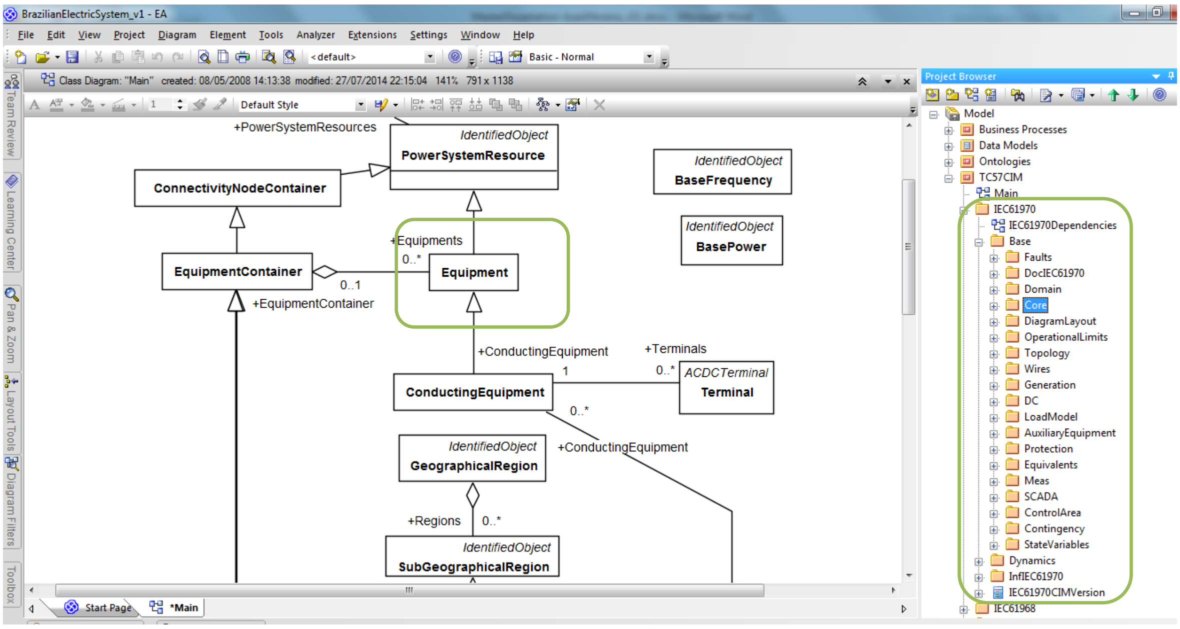


Figure 5.6: CIM structural class package in EA

2. Source-driven approach

As source-driven approach, the involved data sources were checked to assist in domain modelling. BDT data model (part under responsibility of SIPER), BDT-CIM entities mapping, Disturbances BI solution and Clippings information system documents (news publications) were analyzed.

a. BDT

As cited before, BDT is the master DB which provides all representations of the main domain entities, their relations and some restrictions of SIN. The physical data model, designed in Power Designer[28] CASE tool, was used to check tables, attributes and relationship integrities and constraints that implement the domain behaviour. For example, different kinds of `companies` in the `electrical sector`, such as a `transmission company` and a `generation company` are implemented in BDT by an association table that links `company` and `company type` tables. It's conceptualization in DB data model is depict in Figure 5.7.

[28] http://www.sybase.com.br/products/modelingdevelopment/powerdesigner

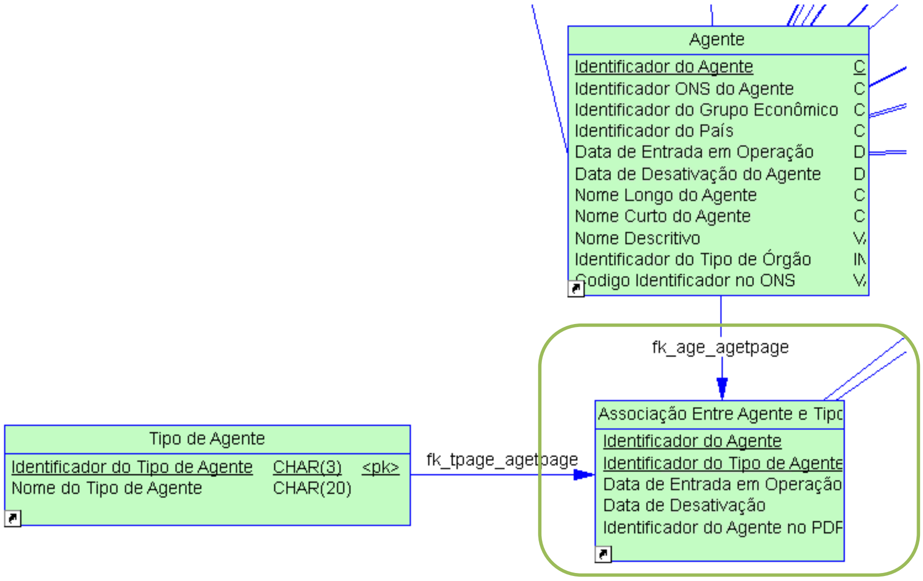

Agente
Identificador do Agente
Identificador ONS do Agente
Identificador do Grupo Econômico
Identificador do País
Data de Entrada em Operação
Data de Desativação do Agente
Nome Longo do Agente
Nome Curto do Agente
Nome Descritivo
Identificador do Tipo de Órgão
Codigo Identificador no ONS
fk_age_agetpage
Associação Entre Agente e Tipo
Identificador do Agente
Identificador do Tipo de Agente
Data de Entrada em Operação
Data de Desativação
Identificador do Agente no PDF
Tipo de Agente
Identificador do Tipo de Agente CHAR(3) <pk>
Nome do Tipo de Agente CHAR(20)
fk_tpage_agetpage

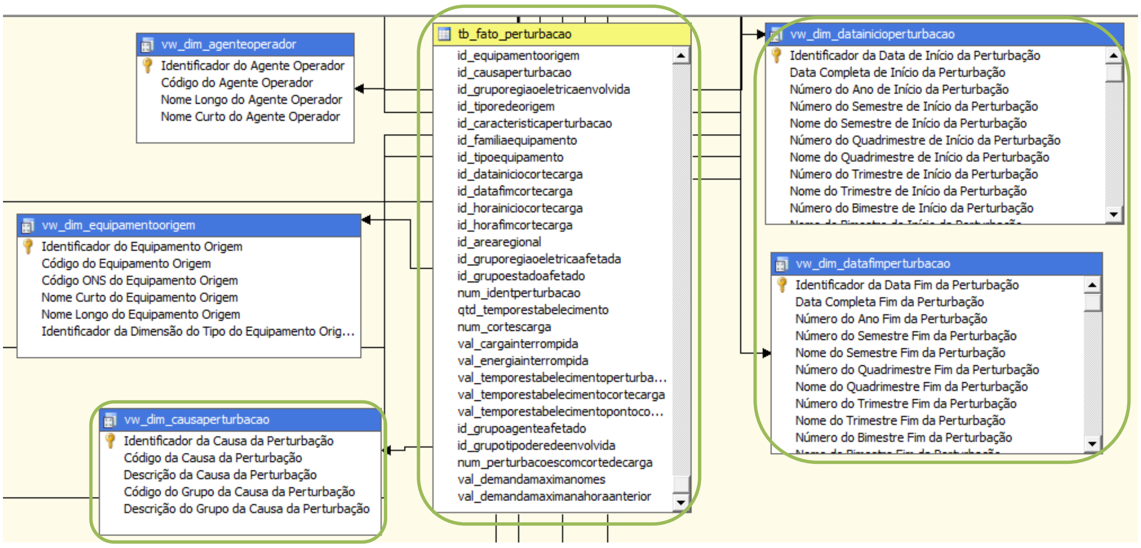
Figure 5.7: Company data table and association table that implement company types

b. BDT- CIM entities mapping

A mapping document specification from entities represented in BDT to entities represented in CIM, in majority with similar concepts, was also used to support the domain representation. This documentation was built under an ONS large project (REGER) and an ETL process was built based on it, where it extracts data from BDT, transforms and load it to a CIM file representation (as RDF). In this document, for each CIM entity (e.g. PowerTransformer), the source definition in BDT and the execution conditions (insert, update and delete) was mapped through SQL queries. That way, the domain could be modelled in English terms, reusing the knowledge existent.

c. Disturbances BI

As cited in the last section, we also used Disturbances BI solution, mainly Disturbance DM. Through this solution, analyzing the available ETL process, we could check how the tables and relations from BDT were extracted and transformed (following transformation rules) to load the Disturbances MD schema. From the DM we could compare some domain concepts and conciliate with the other representations, such as the disturbance source equipment. In Figure 5.8 the fact disturbance has associations to dimensions: owner agent, source equipment, disturbance cause, begin and end time, among others.



vw_dim_agenteoperador
- Identificador do Agente Operador
- Código do Agente Operador
- Nome Longo do Agente Operador
- Nome Curto do Agente Operador

vw_dim_equipamentoorigem
- Identificador do Equipamento Origem
- Código do Equipamento Origem
- Código ONS do Equipamento Origem
- Nome Curto do Equipamento Origem
- Nome Longo do Equipamento Origem
- Identificador da Dimensão do Tipo do Equipamento Orig...

vw_dim_causaperturbacao
- Identificador da Causa da Perturbação
- Código da Causa da Perturbação
- Descrição da Causa da Perturbação
- Código do Grupo da Causa da Perturbação
- Descrição do Grupo da Causa da Perturbação

tb_fato_perturbacao
- id_equipamentoorigem
- id_causaperturbacao
- id_gruporegiaoeletricaenvolvida
- id_tiporedeorigem
- id_caracteristicaperturbacao
- id_familiaequipamento
- id_tipoequipamento
- id_datainiciocortecarga
- id_datafimcortecarga
- id_horainiciocortecarga
- id_horafimcortecarga
- id_arearegional
- id_gruporegiaoeletricaafetada
- id_grupoestadoafetado
- num_identperturbacao
- qtd_temporestabelecimento
- num_cortescarga
- val_cargainterrompida
- val_energiainterrompida
- val_temporestabelecimentoperturba...
- val_temporestabelecimentocortecarga
- val_temporestabelecimentopontoco...
- id_grupoagenteafetado
- id_grupotipoderedeenvolvida
- num_perturbacoescomcortedecarga
- val_demandamaximanomes
- val_demandamaximanahoraanterior

vw_dim_datainicioperturbacao
- Identificador da Data de Início da Perturbação
- Data Completa de Início da Perturbação
- Número do Ano de Início da Perturbação
- Número do Semestre de Início da Perturbação
- Nome do Semestre de Início da Perturbação
- Número do Quadrimestre de Início da Perturbação
- Nome do Quadrimestre de Início da Perturbação
- Número do Trimestre de Início da Perturbação
- Nome do Trimestre de Início da Perturbação
- Número do Bimestre de Início da Perturbação

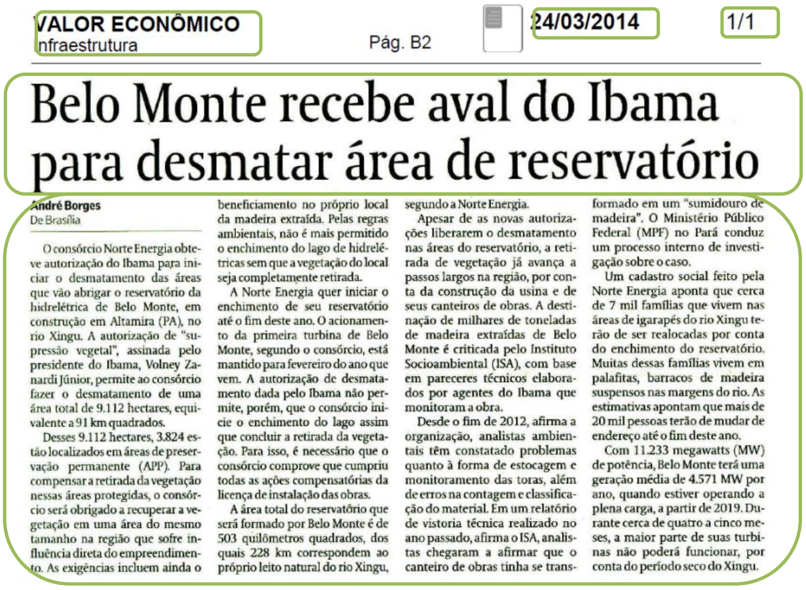
vw_dim_datafimperturbacao
- Identificador da Data Fim da Perturbação
- Data Completa Fim da Perturbação
- Número do Ano Fim da Perturbação
- Número do Semestre Fim da Perturbação
- Nome do Semestre Fim da Perturbação
- Número do Quadrimestre Fim da Perturbação
- Nome do Quadrimestre Fim da Perturbação
- Número do Trimestre Fim da Perturbação
- Nome do Trimestre Fim da Perturbação
- Número do Bimestre Fim da Perturbação

Figure 5.8: Disturbance DM – disturbance fact, its cause, begin and end time

d. Clippings

The Clippings website was used to check and design the `news article papers` sub-domain, as textual information, which metamodel is independent from `electrical sector`. The header structure of the `news documents` was also verified, so patterns could be listed, such as `publication date`, `press company` and `news article`, as illustrated in Figure 5.9. A Web Crawler[29] was built to navigate through the `Clippings` website, visiting each `news article webpage` from 2011 to 2013, saving each one as a text document (.txt).

[29] Web Crawler is a type of program which provides the access and navigation over web pages in an automatic way. In this implementation a program was built to access the Clippings URL, manipulating the query string news identification to access the news text by date range approximation.



VALOR ECONÔMICO
nfraestrutura

Pág. B2

24/03/2014

1/1

Belo Monte recebe aval do Ibama para desmatar área de reservatório

André Borges
De Brasília

O consórcio Norte Energia obteve autorização do Ibama para iniciar o desmatamento das áreas que vão abrigar o reservatório da hidrelétrica de Belo Monte, em construção em Altamira (PA), no rio Xingu. A autorização de "supressão vegetal", assinada pelo presidente do Ibama, Volney Zanardi Júnior, permite ao consórcio fazer o desmatamento de uma área total de 9.112 hectares, equivalente a 91 km quadrados.

Desses 9.112 hectares, 3.824 estão localizados em áreas de preservação permanente (APP). Para compensar a retirada da vegetação nessas áreas protegidas, o consórcio será obrigado a recuperar a vegetação em uma área do mesmo tamanho na região que sofre influência direta do empreendimento. As exigências incluem ainda o beneficiamento no próprio local da madeira extraída. Pelas regras ambientais, não é mais permitido o enchimento do lago de hidrelétricas sem que a vegetação do local seja completamente retirada.

A Norte Energia quer iniciar o enchimento de seu reservatório até o fim deste ano. O acionamento da primeira turbina de Belo Monte, segundo o consórcio, está mantido para fevereiro do ano que vem. A autorização de desmatamento dada pelo Ibama não permite, porém, que o consórcio inicie o enchimento do lago assim que concluir a retirada da vegetação. Para isso, é necessário que o consórcio comprove que cumpriu todas as ações compensatórias da licença de instalação das obras.

A área total do reservatório que será formado por Belo Monte é de 503 quilômetros quadrados, dos quais 228 km correspondem ao próprio leito natural do rio Xingu, segundo a Norte Energia.

Apesar de as novas autorizações liberarem o desmatamento nas áreas do reservatório, a retirada de vegetação já avança a passos largos na região, por conta da construção da usina e de seus canteiros de obras. A destinação de milhares de toneladas de madeira extraídas de Belo Monte é criticada pelo Instituto Socioambiental (ISA), com base em pareceres técnicos elaborados por agentes do Ibama que monitoram a obra.

Desde o fim de 2012, afirma a organização, analistas ambientais têm constatado problemas quanto à forma de estocagem e monitoramento das toras, além de erros na contagem e classificação do material. Em um relatório de vistoria técnica realizado no ano passado, afirma o ISA, analistas chegaram a afirmar que o canteiro de obras tinha se transformado em um "sumidouro de madeira". O Ministério Público Federal (MPF) no Pará conduz um processo interno de investigação sobre o caso.

Um cadastro social feito pela Norte Energia aponta que cerca de 7 mil famílias que vivem nas áreas de igarapés do rio Xingu terão de ser realocadas por conta do enchimento do reservatório. Muitas dessas famílias vivem em palafitas, barracos de madeira suspensos nas margens do rio. As estimativas apontam que mais de 20 mil pessoas terão de mudar de endereço até o fim deste ano.

Com 11.233 megawatts (MW) de potência, Belo Monte terá uma geração média de 4.571 MW por ano, quando estiver operando a plena carga, a partir de 2019. Durante cerca de quatro a cinco meses, a maior parte de suas turbinas não poderá funcionar, por conta do período seco do Xingu.

Figure 5.9: A news article example, published in March/2014 available at the Clippings website

3. Well founded domain ontology

Based on the information provided from analysis-driven and source-driven activities, the domain ontology was built using in EA and OntoUML plug-in tools. In addition, the library of UFO conceptualization provided by NEMO (depict in section 5.1) was used, basically by importing the UFO-A, UFO-B, UFO-C and UFO:PW packages. Then, the `SIN` domain package was created and divided in five sub-domains, the structural behaviour of `companies`, `facilities`, `equipments` and `geographical region`. In addition, the dynamic aspects of `disturbances` and `published news` were added, as illustrated in Figure 5.10.

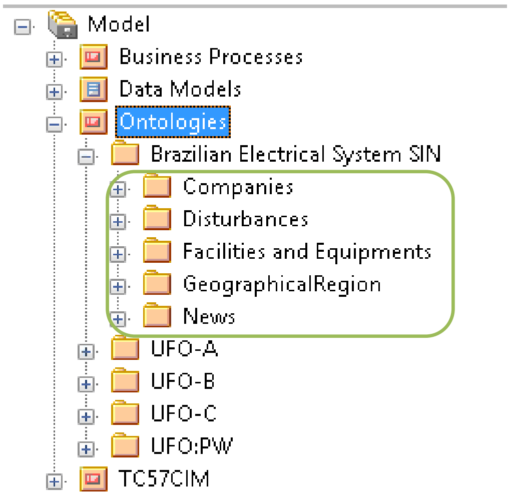


Figure 5.10: Brazilian Electrical System (SIN) domain and its main parts designed in EA tool

In this division, classes and relationships were created based on UFO concepts, having basic UFO structural concepts available through OntoUML language as stereotypes (e.g. **Kind**, **Collective**, **Quantity**, **SubKind**, **Role**, **Phase**, **Category**, **RoleMixin**, among others). The `company` sub-domain is illustrated in Figure 5.11, where the `company` itself was defined as a **Kind**. The `operation company` was defined as **RoleMixin** and was specialized, as disjoint and complete, in `GenerationCompany`, `DistributionCompany` and `TransmissionCompany` (all classified as **Roles**). Also, there are **material** relations to other concepts, such as `ownerCompany` relation from `Equipment` to `Company` and `operatedByCompany` relation from `Transmission Line` to `Transmission Company`. The complete EA solution is available in appendix.

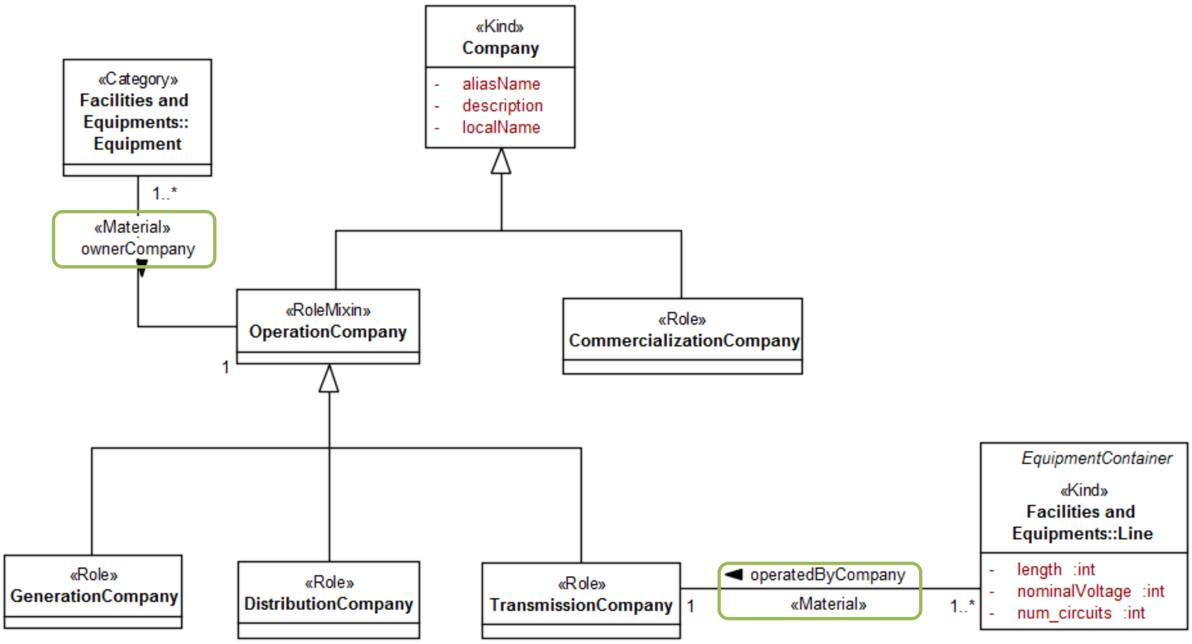


Figure 5.11: Company types ontology cut (well-founded with UFO)

A cyclic verification and validation activity was performed using OLED/Alloy Analyzer software to increase the ontology quality and assertiveness in representing real world concepts. After the first time designing the main concepts in the domain ontology, we made cut offs to verify them. That way we could generate small words to assert the ontology by the generation of examples and counterexamples, simulating instances of classes and their relations. The criteria utilized for the cuts was to bring to V&V each part of the domain ontology. Figure 5.12 illustrates the relations asserted for DisturbanceComplexEvent class and its parts, as participations.

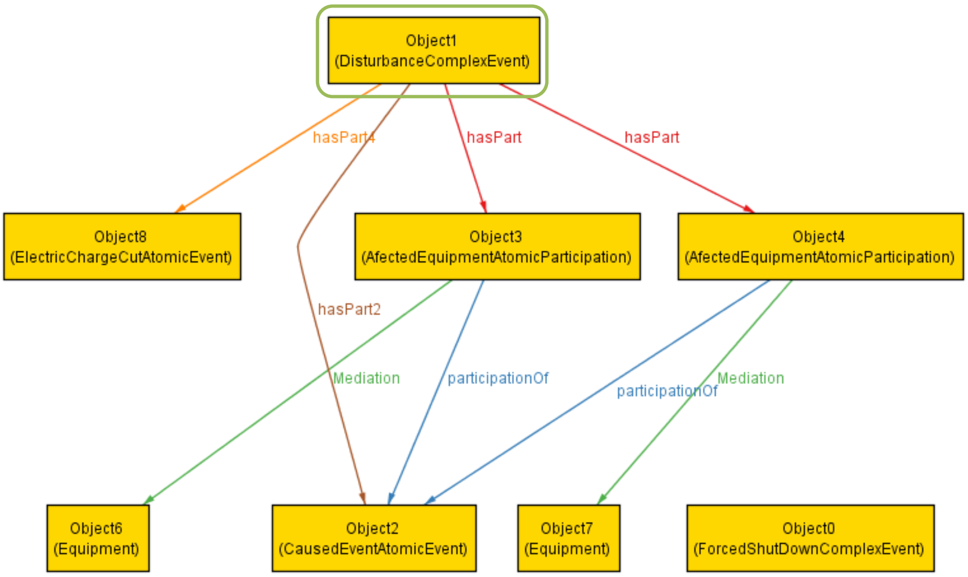


Figure 5.12: Example of visual validation in OLED/Alloy software

During this process some assertions were made and the domain ontology could be evolved and semantically enhanced. As a result from this process, the ontology matured to a better level of common understanding. To achieve our objective in this work, the main parts to be analyzed are the disturbances and news sub-domains. The former is illustrated in Figure 5.13, where the Disturbance is set as **Event** stereotype and specialized from **Complex Event** (from UFO-B package). The participations of equipments as sources of disturbance occurrences are represented by the Source Equipment Participation class, stereotyped as **Event** and specialized from **Participation** (from UFO-B package). This **Participation** has Load Cut Value and Restoration Time attributes and is a part of the Disturbance **Event**, formalized by the **memberOf** mereology relation. Moreover, this **Participation** is different depending of the equipment type, which is represented by the disjoint specialization for Transmission Line **Participation** and Power Transformer **Participation**. As described in section 5.1, the **participationOf** relation was defined as **Formal** relation from an **Event** to a **Substantial** (from UFO-A package). Therefore, the **RoleMixins** Transmission Line **Participant** and Power Transformer **Participant**, which specialize the Transmission Line **Kind**, the Power Transformer **SubKind** and the Equipment, respectively (all from equipment and facilities sub-domains), are linked by the **participationOf** relation to their analogous names. We used the blue colour to represent UFO concepts, green for concepts in BI/DW analytical project context and blank for outside our interest for the experimentation.

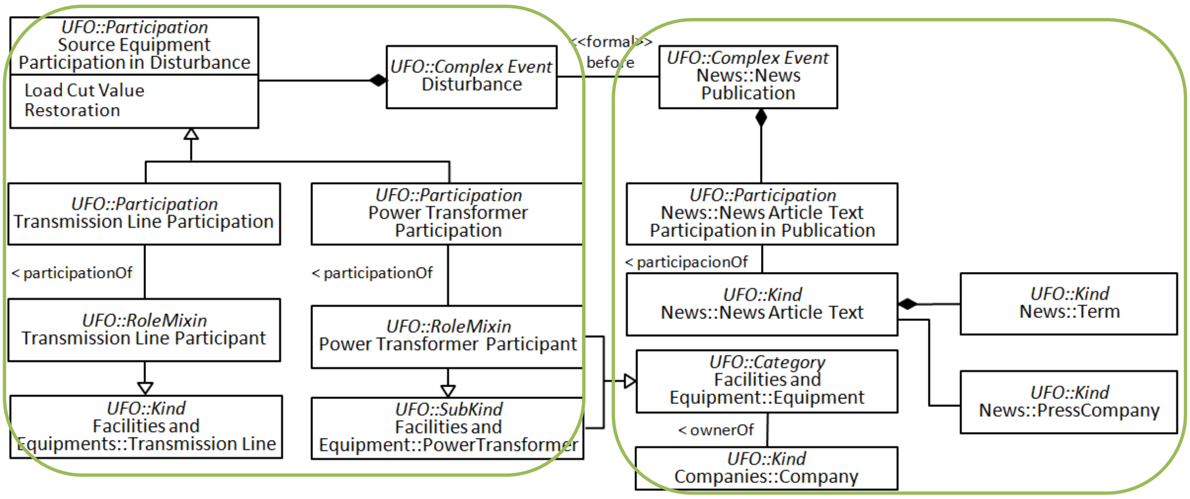


Figure 5.13: Disturbances and News publications domain ontology

The news article publications sub-domain are also illustrated in Figure 5.13, which presents the concept of news publication as a complex event, having the News article Text participation as mereological relation (partOf). The News article text is the document itself, composed by Terms, being owned by a press company and written by an author, all of them represented as kind, with ownerOf and writtenBy formal relations. However, the most significant representation in this ontology for our experimentation is the before temporal relation between a News publication and a disturbance. It means that for each disturbance there is a before relation to a news publication with 0..* cardinality between them. It is a weak relationship, but it is sufficient important to the business needs to formalize it.

4. Derivation of multidimensional concepts from Rules

Once having the domain ontology verified as a valid structure, the prototype execution could be made to derive MD concepts. We exported the full ontology as a XMI file and load it in the prototype. As a result of the first rule execution, the following MD concepts were derived:

- **<u>From R1:</u> Events** as *Facts*

R1.a. If *e* is a **Complex Event**, then *e* can be represented as a *Fact f* (Figure 5.14):

- `Disturbance`
- `News Publication`

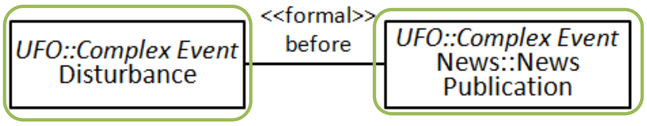


Figure 5.14: **Complex Events** as *Facts* in Domain Ontology

R1.b. If *f* is a *Fact* defined by **Event** *e*, then the numeric **quality** of *e* can be derived as *Measures* of *Fact f*:

- **Complex Event** with attributes was not found.

R1.c. If *f* is a *Fact* defined by **Event** *e* and *e'* is an **Event** which has mereologic relation (**partOf**) to *e* and there exists quality values *qs* of *e'*, then *e'* can be derived as a *Fact* or *qs* as *Measures* in the *Fact f* (Figure 5.15):

- *Fact f* as `Disturbance`, *e'* as `Source Equipment` **Participation**, *Measures* as attributes *attr* `Load Cut Value` and `Restoration Time`:

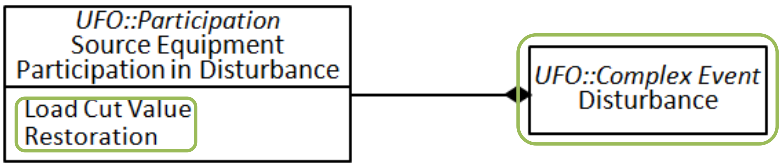


Figure 5.15: *Measures* derived from **Event** attributes

- **<u>From R2:</u> Objects Participations** as *Dimensions* and *Hierarchies*.

R2.a. If *f* is a *Fact* representing an **Event** *e*, *o* is an **Object** (**Substance**) which is mediated by **formal** relation (**participationOf**) to a **Participation** *p*, which has a mereologic relation (**partOf** *e*), then *o* can be derived as a *Dimension* of *Fact f* (Figure 5.16):

- `Transmission Line`
- `Power Transformer`

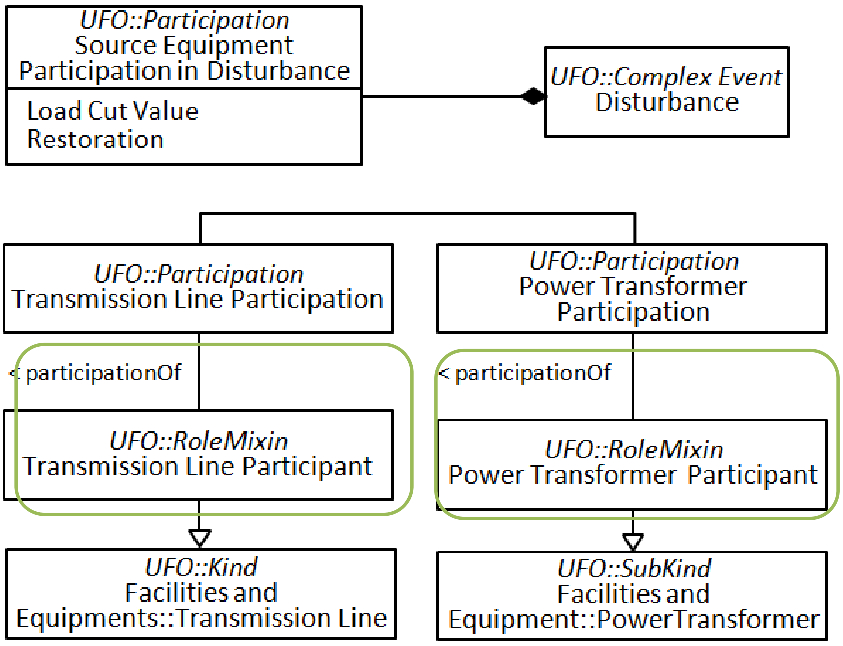


Figure 5.16: **Participants** as *Dimensions*

R2.b. If *f* is a *Fact* representing an **Event** *e*, *d* is a *Dimension* defined by **Object** *o* which is **Participant** of *e*, and *o* has relationships to other **Substances** *S*, then each **Substance** *s* of *S* can be suggested as *Hierarchies* of the *Fact f* through *Dimension d.* (Figure 5.17):

- Owner Company *Hierarchy* of Transmission Line and Power Transformer.

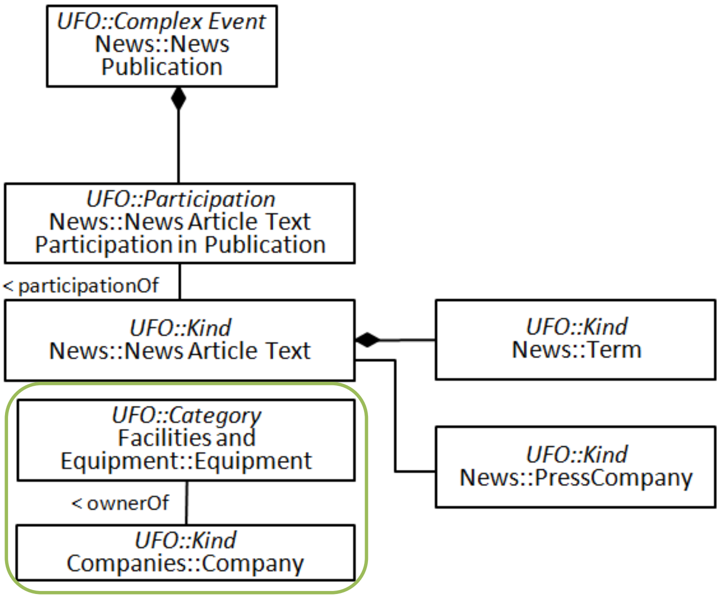


Figure 5.17: *Measures* derived from **Event** attributes

An observation that must be mentioned in this step is that the Transmission Line Owner Company *Hierarchy* could be derived because Transmission Line is an Equipment Container and is related to Equipment **Category** class.

- __**From R3:**__ **Time Interval Relations** between **Events** as *Snowflake Fact*

R3.a. If *e1*, *e2* are **Events** and they are modelled with one or more **Time Interval Relations** between them, then a *Fact f* can be represented by the **Time Interval Relations** between the **Events** *e1* and *e2*, where *e1* is represented by *Dimension d1* and *e2* represented as *Dimension d2*. Each *Dimension* (*d1* and *d2*) has a *Time Hierarchy* for the **Begin** and **End Time Points** (Figure 5.18):

- **Temporal Relation** (**before**) between Disturbance and News Publication.

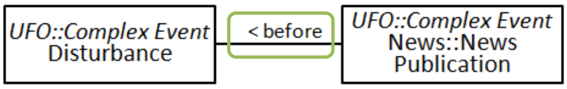


Figure 5.18: **Temporal Relation** between Disturbance and News Publication

Based on this rule we could design a MD schema to analyze the temporal relation **before** defined for Disturbances and Published News. Figure 5.19 depicts this *Snowflake Conceptual Schema* derived from the domain ontology. There are two *Dimensions*, Disturbance and Terms Published. Both present their main **Participants** as **Dimension Attributes** or **Hierarchies**. For example, the Transmission Line *Hierarchy* of the Disturbance *Dimension* was defined with *first level* the Company Owner, *second level* the Transmission Line and *third level* the Code. Analogous situation occurs with Power Transformer *Hierarchy*. The Disturbance Cause is grouped by Categories, than they can also be a perspective of analysis of the temporal relation *Fact*. The Terms Published *Dimension* also contains *Hierarchies* set, for example the Category of Terms, which is a perspective of analysis by grammar categories (verbs, nouns, adjectives, etc). The Press Companies is also a *Hierarchy* set.

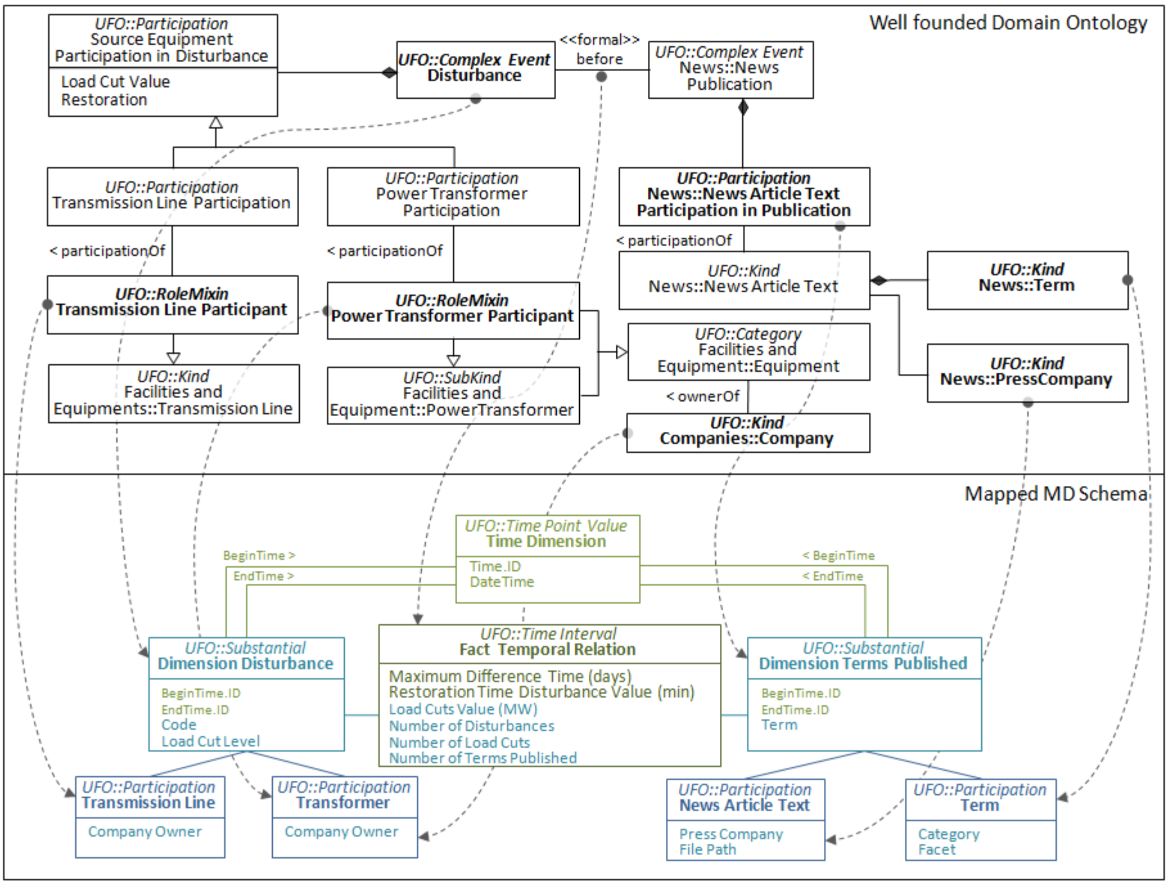
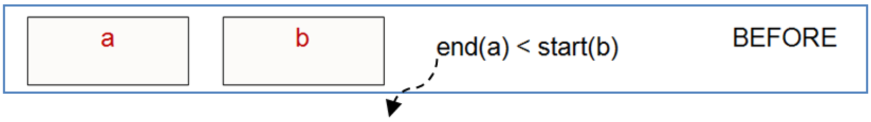

Well founded Domain Ontology
UFO::Participation Source Equipment Participation in Disturbance
Load Cut Value Restoration
UFO::Complex Event Disturbance
<<formal>> before
UFO::Complex Event News::News Publication
UFO::Participation Transmission Line Participation
UFO::Participation Power Transformer Participation
UFO::Participation News::News Article Text Participation in Publication
< participationOf
< participationOf
< participationOf
UFO::RoleMixin Transmission Line Participant
UFO::RoleMixin Power Transformer Participant
UFO::Kind News::News Article Text
UFO::Kind News::Term
UFO::Category Facilities and Equipment::Equipment
UFO::Kind News::PressCompany
UFO::Kind Facilities and Equipments::Transmission Line
UFO::SubKind Facilities and Equipment::PowerTransformer
< ownerOf
UFO::Kind Companies::Company
Mapped MD Schema
UFO::Time Point Value Time Dimension
Time.ID
DateTime
BeginTime >
EndTime >
< BeginTime
< EndTime
UFO::Substantial Dimension Disturbance
BeginTime.ID
EndTime.ID
Code
Load Cut Level
UFO::Time Interval Fact Temporal Relation
Maximum Difference Time (days)
Restoration Time Disturbance Value (min)
Load Cuts Value (MW)
Number of Disturbances
Number of Load Cuts
Number of Terms Published
UFO::Substantial Dimension Terms Published
BeginTime.ID
EndTime.ID
Term
UFO::Participation Transmission Line
Company Owner
UFO::Participation Transformer
Company Owner
UFO::Participation News Article Text
Press Company
File Path
UFO::Participation Term
Category
Facet

Figure 5.19: Mapped MD schema for **temporal relation** analysis of `Disturbances` and `News`

Finally, the *MD schema* could be designed by the MD concepts derived from the proposed rules and also conciliated with the existent *MD schema* from the `Disturbances` DM. Moreover, by asserting the **before** relation we could annotate the constraint for the *Fact* load during the ETL design process as represented in Figure 5.20:



```
where a.endTimePoint < b.beginTimePoint
```

Figure 5.20: Before mapped to ETL constraint: "a" as Disturbance and "b" as News Publication

5.2.2.3 Design

A DB was physically designed and its creation scripts can be found in appendix, reflecting the MD conceptual schema generated from rule R3 to analyze the **temporal relation** between `Disturbances` and `News Publications`. The ETL process was also designed based on JointOLAP architecture (see appendix). The application of this architecture in the `Disturbances` and `Clippings` BI/DW project followed each of the mentioned steps. The conceptual data flow presented in Figure 5.21 describes the ETL steps.

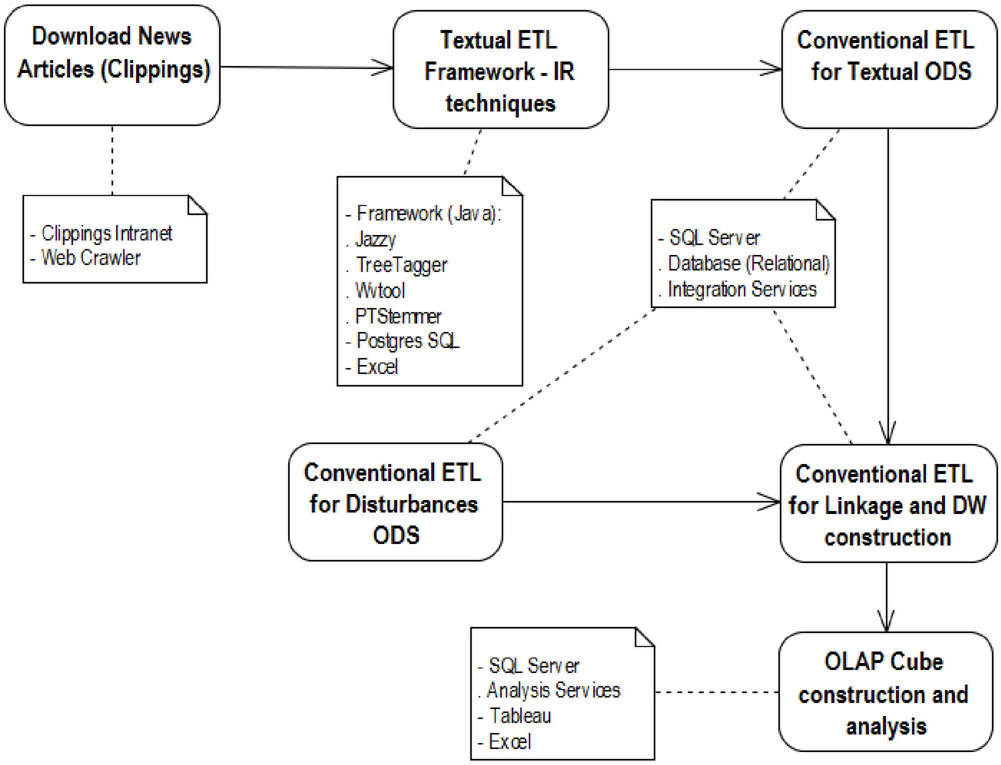

Download News Articles (Clippings)
Textual ETL Framework - IR techniques
Conventional ETL for Textual ODS
- Clippings Intranet
- Web Crawler
- Framework (Java):
. Jazzy
. TreeTagger
. Wvtool
. PTStemmer
- Postgres SQL
- Excel
- SQL Server
. Database (Relational)
. Integration Services
Conventional ETL for Disturbances ODS
Conventional ETL for Linkage and DW construction
- SQL Server
. Analysis Services
- Tableau
- Excel
OLAP Cube construction and analysis

Figure 5.21: ETL conceptual data flow and OLAP cube development

The initial step is to download the `news articles` from the `Clippings` information system by a web crawler. Then, Textual ETL framework is performed in those documents, resulting in the Terminological DB loaded. Textual ETL is one of the main components of JointOLAP, which starts reading each news article as a text file, than, it finds expressions patterns, such as title and press company. The main pattern in the text is the publication date, which is the basis for temporal analysis of clippings used as a temporal qualifier. The textual extraction and indexing is detailed in (ALMEIDA e SILVA, 2009). During this process, in parallel, a conventional ETL process extracts the structured data from Disturbances BI. Thereafter, an ETL process runs the *Linkage* activity, responsible for loading the temporal relation *Fact* between `Disturbances` and `Terms Published in News`. In this stage it is much important to annotate the constraint definition from temporal relation provided by rule R3 execution (to load the *Fact*). At last, the OLAP cube and the user interface for the integrated analysis are built. Each step is supported by a set of tools, as depicted in Figure 5.21.

5.2.2.4 Construction and execution

The full ETL process construction is described in appendix, where the experimental environment and ETL development steps are described, including the acquisition of the news article publications, the involved DBs, ETL data flows and the *Fact* data load, coping with the constraint from rule R3. The result is the MD schema in built in relational DB, which have to be prepared to be explored by an OLAP interface. The process of building the data cube was made in SSAS[30], defining facts, how dimensions are related and how their measures should behave on aggregating, responding to summarizability issues, considering hierarchies, attributes and key attributes relationships.

In appendix each part of the data cube development is represented, such as preparation of dimensions, fact and measures. In addition, each OLAP analysis to support the main business analytical requirements, regarding energy supply security subject and the impact of disturbances in ONS corporate image, is explained in detail. Examples of such analyses are:

- Number of terms published in news articles reflecting the related disturbances severity and the increased proportion of the “blackout” word when the load cut level is higher;
- The Press companies that have more publications related to the national electric system;
- How a severe disturbance can affect the number of publications by time, as the blackout occurred in February 2011, which reflected the raise of terms published in the following months;
- The most common causes, from the disturbance cause dimension, evidence that “natural phenomena” is quite usual. Nevertheless, when checking the associated terms, the “operation error” appears as one of the top causes. This evidences that in several situations the media supposes that an operational error is the origin of a

[30] Data Cube tool from Microsoft: SQL Server Analysis Services (SSAS).

disturbance, but in fact they more often accused natural phenomena. In this case, ONS could invest in marketing actions when blackouts occurred to clarify the causes.

5.2.3 Result analysis

As results from the approach application to impact analysis of disturbances in ONS corporate image we can highlight:

- The mapping rules derived, through the prototype execution, necessary MD concepts to achieve the analysis requirements, evidencing their effectiveness. The main representations for the final MD schema were addressed, such as transmission line dimension and owner company hierarchy;

- The mapping rules derived some existent MD concepts, such as the disturbance fact from R1. This evidence that, if there was not yet a Disturbances BI solution and all its concepts were modelled in the domain ontology, the rules could derive the same MD concepts as Disturbances BI;

- The hybrid approach application in this scenario could cope with unstructured data sources (Clippings information system) for the BI/DW solution. In addition, it considered entities from the unstructured data sources during the MD design activity, such as the `terms published dimension` from the participation of a `news article` in a `news publication`;

- The temporal relation `before` between `disturbance` and `news publication` could be analyzed through the MD snowflake schema derived from R3, bringing new perspective of analysis about this relation.

- A limitation in this example case was that the ETL process and the developed application did not follow common software concerns or best practices (as the prototype) because of the necessary effort. They should only be considered as POCs.

5.3 Application example 2: causality and situation changes in ITIL process

In this section we discuss examples of how rules referring to causality relation between events (R4) and situation changes (R5) can enhance temporal analysis in MD schemas. At first, the ITIL ontology, from a prior work, is shortly depicted. Thereafter, the application of the above mentioned rules is experimented, discussing possible analysis over the resulting MD concepts. Finally, the results and implications of the domain ontology changes to consider an incident as event are analyzed.

5.3.1 Business scenario

A well-founded domain ontology from previews work (CALVI, 2007) was chosen and its representation was designed in EA tool with support of the NEMO package, as in the first example application. It contemplates the scenario of the ITIL incident process. In ITIL terminology an `incident` is defined as an **Event** that is not part of the `standard operation of a service`, which **causes** its `quality reduction` or its `interruption`. Examples are `services` affected by `software and hardware failures`, `service requests`, `infrastructure tasks`, `consultation services` and `questions` from clients and customers. An `incident call` is an **Action** (**Event**) executed from an **Agent** (the `user`) to resolve an `incident` **Situation**, which is the pre-state of (**triggers**) an `incident call`. A `root case` in ITIL can be defined as an **Event**, **bringing about** (has a post-state) the `incident` **Situation**. Moreover, a `root case`, such as an `infrastructure failure` (e.g. `broken internet link`), can **cause** an `incident call`.

Upon the original ontology it was added a data type attribute: the `operatorCost` (float), which indicates the `operator cost` for the company of an `incident call`, illustrated in Figure 5.22 bellow.

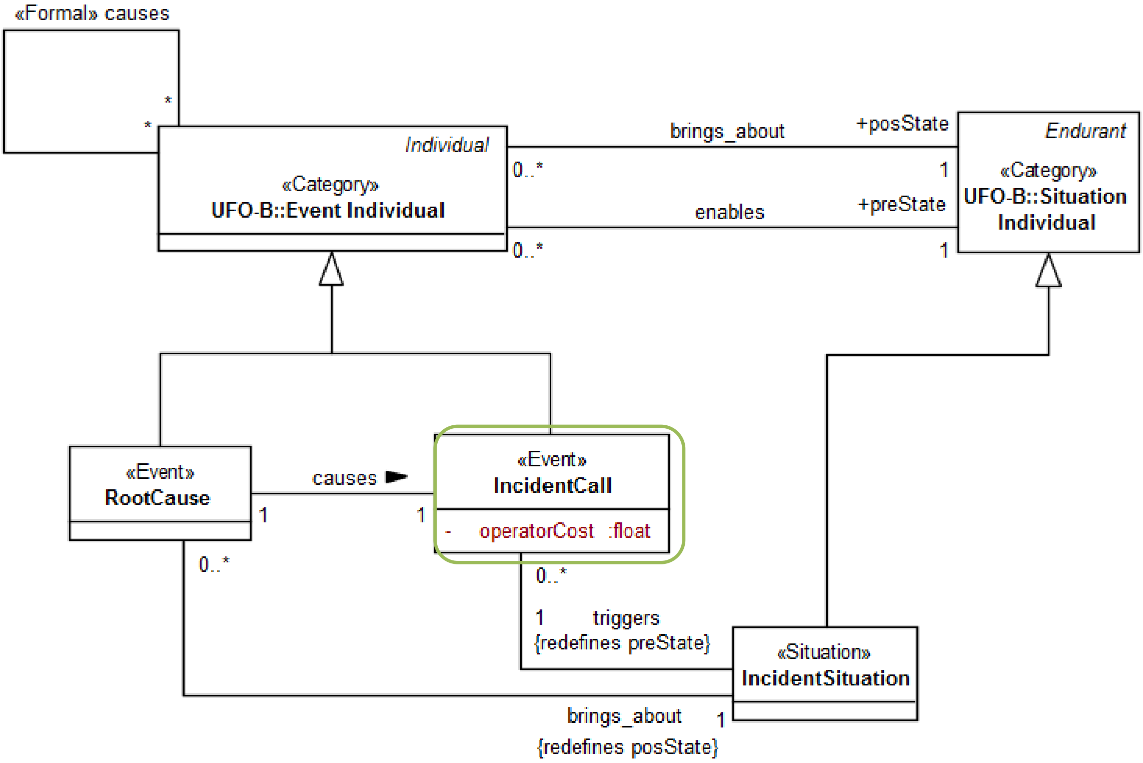


Figure 5.22: ITIL ontology changed with `operatorCost` based on (CALVI, 2007)

5.3.2 Application of the proposed approach

In this section, we discuss the rules R4 and R5 execution on ITIL ontology by simulating them, showing the possible resulting MD concepts. In addition, analyses over the resulted MD schemas are exemplified with their hypothetical data results. The execution of rules R4 and R5 in this ontology snippet resulted in the MD concepts, which were illustrated with *Dimension* and *Fact* stereotypes, as follows:

- **<u>From R4:</u>**

R4.a. If *e1, e2* are **Events** and *e1* **causes** *e2*, then *e1* can be derived as a *Dimension* and *e2* can be derived as a *Fact* (Figure 5.23).

- o `Root Cause` (*Dimension d1*) from `Root Cause` (**Event** *e1*);
- o `Incident Call` (*Fact f*) from `Incident Call` (**Event** *e2*) and **causes**(*e1*,*e2*) relation.

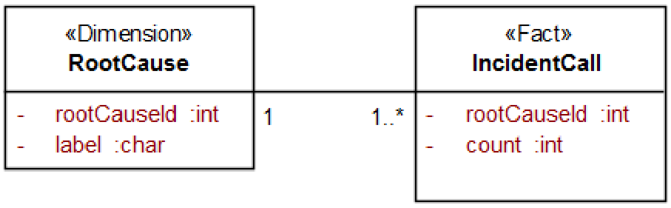


Figure 5.23: Incident Call Fact and Root Cause Dimension derived

- <u>Analysis Example:</u> Most commons `Root Causes` (number of `Incident Calls` by `Root Cause`).

Table 2.5: Analysis example – Incident calls by root causes

Root Cause	# Incident Calls
Internet link broken	50
Active Directory out of service	30
Network synchronization failure	23
Application configuration error	15

R4.b. If *f* is a *Fact* defined by R4.1 from **Event** *e*, then the numeric attributes *attr* of *e* can be suggested as *Measures* for *Fact f* (Figure 5.24).

- `operatorCost` (*Measure m*) from `Incident Call` (*Fact f*), `Incident Call` (**Event** *e*) and `operatorCost` (*Attribute attr*).

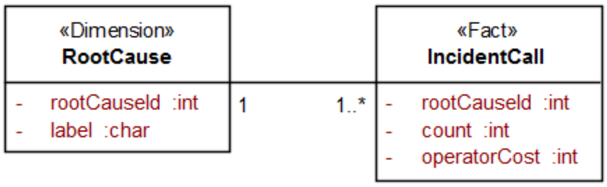


Figure 5.24: `operatorCost` *Measure* in `Incident Call` *Fact*

- <u>Analysis Example:</u> Most expensive `Root Causes` (number of `Incident Calls` by `Root Cause`).

Table 2.6: Analysis example – Operator cost by root causes

Root Cause	# Incident Calls	Operator Cost $
Active Directory out of service	30	2,519.93

Application configuration error	15	1,980.00
Internet link broken	50	1,230.00
Network synchronization failure	23	450.00

- **From R5:**

R5.a. If *e1, e2* are **Events** and *s* is a **Situation** that has a **triggers** relation to *e1* and a **brought-about** relation to *e2*, then *e1*, *e2* and *s* can be derived as *Dimensions d1*, *d2* and *d3* respectively, and a *Fact f* that relate these three *Dimensions* (Figure 5.25).

- Incident Call (*Dimension d1*) from Incident Call (**Event** *e1*);
- Root Cause (*Dimension d2*) from Root Cause (**Event** *e2*);
- Incident Situation (*Dimension d3*) from Incident Situation (**Situation** *s*);
- Incident Situation Change (*Fact f*) from triggers(*s,e1*) and brought-about(*s,e2*) relations.

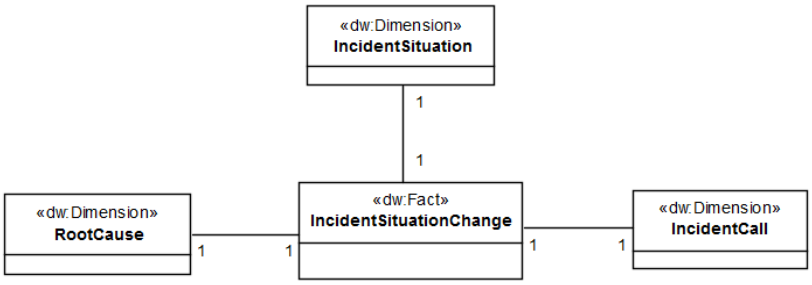


Figure 5.25: MD schema for Situation Change by related **Events** and **Situation**

- Analysis Example: Most common Root Causes of Incident Situations (number of Situation changes by Root Cause and Incident Situation).

Table 2.7: Analysis example – Incident Situations by root causes

Incident Situation	Root Cause	Situation Changes
Network Error	Application configuration error	87
Network Problem	Active Directory out of service	60

Application Error	Error during deployment	34
Application Problem	Missing configuration	21

R5.b. If *f* is a *Fact* defined by R5.1 with **Situation** *s,* **Events** *e1* and *e2*, then the numeric attributes of *s, e1* and *e2* can be suggested as *Measures* for *Fact f* (Figure 5.26).

- operatorCost (*Measure m*) from Incident Situation Change (*Fact f*), Incident Call (**Event** *e1*) and operatorCost (*Attribute attr*).

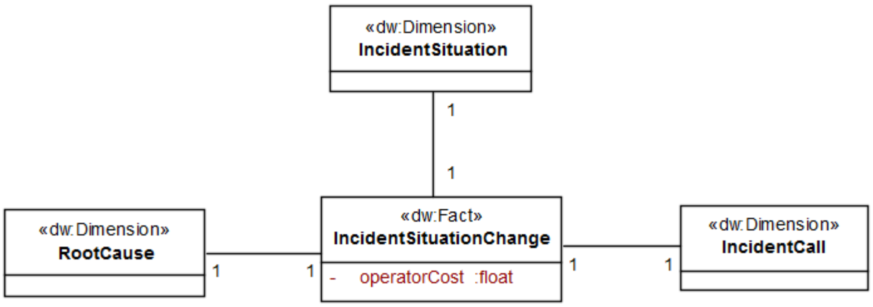


Figure 5.26: Addition of operatorCost *Measure* to **Situation** Change MD schema

- Analysis Example: Most expensive Incident Situations caused by application configuration error (sum of operatorCost of situation changes by Incident Situation where root cause = 'application configuration error').

Table 2.8: Analysis example – Incident Situations by root causes

Incident Situation	# Situation Changes	Operator Cost $
Application Problem XPTO01	5	1,455.09
Application Problem XPTO02	3	1,280.00
Application Problem XPTO03	7	1,125.32
Application Problem XPTO04	2	410.00

5.3.3 Result analysis

The results from rules R4 and R5 manual execution on the ITIL process ontology can be analyzed as:

- The mapping rules derived MD concepts for causality analysis of incident calls (as fact) by root causes perspective (as dimension). This type of analysis could be shown by a simple analysis;

- A numeric attribute of an event could be derived as a measure, evidenced through the operatorCost usage in incident call. In the example, the finance effect of root causes could be explored using this measure;

- A MD schema to analyze incident situation changes could be derived from R5, where two essential events (root cause and incident call) were correlated to the incident situation dimension through the triggers and brought-about relations.

We considered exactly the ITIL ontology from a prior work to serve as domain ontology to apply the rules. Nevertheless, some changes may be discussed in its representation, such as the classification of an incident as an event, instead of a situation. If we change it to be an event that is caused by a root cause and which causes an incident call, then we could derive two dimension/fact dichotomies (from R4) as: root cause (dimension) and incident (fact); incident (dimension) and incident call (fact) with operator cost (measure). They may be also interesting perspectives of analysis when exploring the amount of incidents by root causes and total operator costs in incident calls by incident dimension.

We evaluate our approach by discussing and evidencing benefits of the proposal through the application examples, that a FO can be applied in designing MD constructs in a more comprehensive way than the one described in (PARDILLO e MAZÓN, 2011). We could enrich the semantic expressiveness of MD models, by increasing the level of conceptualization things from real world to MD concepts. The derivation rules that maps foundational temporal concepts from UFO to MD concepts (facts, measures, dimensions, attributes and hierarchies) were prototyped and exemplified through two domains to evidence their effectiveness.

The application of the hybrid MD design approach proposed was made in a real scenario: the impact of disturbances on ONS institutional image. A full set of tools for designing and construction was used to evidence our approach feasibility, following the main phases of a BI/DW lifecycle methodology. The process considers heterogeneous data sources – structured and unstructured (text) – in a step prior than the rules execution. Therefore, we evidence that the use of FO in MD design does not depend on data source types. Moreover, MD modelling of unstructured data is a decoupled activity in the process, being part of supply-driven phase by reverse engineering of text corpora, using NLP and IR techniques to develop an initial ontology – which is enriched by UFO classification in next step. The application of rules R4 and R5 was evidenced through examples of their execution in a domain ontology from a prior work: the ITIL process ontology well-founded with UFO. Sample analyzes were listed to argue about the rules executions results.

With these two examples we could achieve a semi-automatic way to design MD schemas from an ontological aspect, respecting foundational conceptualization of UFO categories. Nevertheless, some possibilities were not covered in the domain ontology classification that could derive interesting MD schemas, such as blackout situations representations in news publications for cause-effect analysis.

6 Conclusion

In this chapter we conclude this dissertation, first by making a general description of the work, second by listing the contributions achieved, then listing limitations and, at last, we discuss future works.

We introduced a MD design process based on UFO ontological analysis, taking advantage of the semantic expressiveness gained by using a well-founded domain ontology. The mereology relation among events and their qualities are explored by the first rule (R1), which derives facts and measures. The participations of substantials in events are explored by the second rule (R2), which derives possible perspectives of analysis of an event (dimensions, hierarchies and attributes). The temporal relations between two events are explored by the third rule (R3), which derives a fact pattern, as a MD snowflake schema, representing each event as dimension and a temporal measure. In addition, Allen's operators are explored to define ETL constraints to load the fact pattern. The causality relation between events are explored by the fourth rule (R4), which derives a fact/dimension dichotomy, where the causing event is mapped to a dimension and the caused event is mapped to a fact having its numeric attribute set as measure of the fact. The situation changes through triggers and brought about relations are explored by the fifth rule (R5), which derives a MD schema for cause-effect analysis, where the events that triggers and brings a situation (as well the situation itself) are mapped to dimensions, having a fact defined for the situation change. It also considers the numeric attributes of the events as measures for the fact.

The feasibility of the use of these conceptual apparatus was experimented through application examples. Besides the FO application, we considered heterogeneous data sources (unstructured data) in the study case through IR and NLP techniques applied to text sources, such as morphological analysis and entity recognition. However, the application of this approach is not restricted to unstructured data sources context, and can be applied in conventional BI/DW solutions over structured data.

6.1 Contributions

The main contribution of this dissertation is the introduction of an unexplored research topic regarding the application of foundational ontologies in MD design (MOREIRA, CORDEIRO, *et al.*, 2014). The rules introduced in this work are part of a research effort in analysing all UFO concepts and map them to possible MD conceptualizations. Until now, as far as we know, there are no other works proposing this kind of mapping rules. Knowing the nature of domain concepts by UFO ontological classification, such as event and temporal relation, we could benefit of correctly choosing MD concepts and schemas representations in a semi-automatic way by following established rules derivations. For example, identifying dimensions is facilitated by checking participants and causality relation in events. Facts can be derived from events, situations changes and temporal relations. Measures can be derived from events quality structures. In addition, by formalizing textually those derivation rules we evidence that they can be fully axiomatized, so it can serve as basis for reasoning approaches that explore inference capabilities.

Therefore, we believe we can gain by increasing the productivity in business analysis and design phases of BI/DW solutions, through the semi-automation of the MD design activity by an ontology-based approach, where schemas can be more correctly built because they follow specific patterns. A prototype was developed, as a POC, allowing the automatic generation of MD concepts from a domain ontology, ensuring greater correction and coverage in the MD design activity. Indeed, some of the proposed rules are intuitive, such as events as facts (R1). However, this work can lead to new initiatives in the modelling and semantics of BI/DW solutions research topic and may drive potential new results. In the limit, having information systems designed in a well-founded manner, i.e. based on a foundational ontology, and linked somehow to their data sources, it may be possible to automatically derive significant part of an analytical information systems based in BI/DW architecture (MD design and ETL) by applying a full set of mapping rules like those introduced here.

A MD hybrid approach for heterogeneous data (that can be expanded) was introduced in this work, also considering the rules execution. Following this method both structured and unstructured data sources can be used when modelling BI/DW solutions. To support it, an ontological-based analysis process, coping with UFO classification, was also described. Another contribution is the bibliographic revision of the main concepts and related works regarding BI/DW lifecycle (specifically MD design) and formal ontology, also describing how ontologies were applied in BI/DW solutions until now.

6.2 Limitations

During the time period when this dissertation as being written, the OntoUML language did not yet implement *perdurants* (UFO-B) as native stereotypes. Their theory and formalization are recent and several issues about consistency are under discussion. Therefore, applying our approach in real scenarios expecting the full axiomatization from UFO, for now, may not be possible. Moreover, the tools involved for this MDD automation approach are extremely dependent on the language definition. In addition, the rules introduced in this work are not yet described in first order logic, so they cannot be processed by AI reasoners tools. Only a few set of UFO stereotypes were analyzed and exemplified in few domains, therefore, more experimentation are necessary to prove the feasibility of our approach.

Even using our approach to semantically enrich the MD schemas during the modelling activity, some semantics from the well founded conceptualization can be missed for the end users. For example, the meaning of a relation that is transformed to a hierarchy may not be explicit in its visualization through the OLAP tool. At last, the choice of MD concepts for the final MD schemas continues to be a tacit activity where the modeller should have some project decisions when modelling, being error prone.

6.3 Future work

There is a full set of concepts in UFO to be analyzed and mapped to MD concepts, such as objects dispositions, actions and agents, i.e. each of UFO concepts revised in sections 3.2.2 (UFO-A), 3.2.3 (UFO-B) and 3.2.4 (UFO-C). We believe that the representation of hierarchies in MD schemas can be enriched by formal and mainly by material relations among substantial concepts (e.g. roles, kinds, collectives, etc), providing more meaning to the relations between levels in a hierarchy. In addition, the conceptualization of hierarchies can be explored by OLAP tools for summarization purposes and semantic enhancement of user interface. Moreover, the definition of measures aggregations through dimensions and their additivity properties may be set by analyzing event qualities and their dependence relations to objects, as cited in (PARDILLO e MAZÓN, 2011). In addition, cardinality and dependencies issues must be considered in the rules. Different time granularities in time point representations of events are important characteristics that may be explored for extracting aggregation rules and for ETL concerns.

For now, we have formalized the rules only textually, but they can be fully axiomatized in first order logic, serving as a basis for automatic reasoning in the derivation process. This lack of complete formalization means that the full potential has not yet been reached, e.g. no full reasoning can be performed by Artificial Intelligence tools. After axiomatizing the rules, a set of metrics for evaluating MD model quality resulted from rules execution can be used to prove the efficiency of the approach, such as described in (CALERO, PIATTINI, *et al.*, 2001) (SERRANO, TRUJILLO, *et al.*, 2007) (GOSAIN, NAGPAL e SABHARWAL, 2011). It would be relevant to compare each rule with informal design guidelines provided by practitioners such as Kimball (KIMBALL e ROSS, 2013).

Regarding the involved tools, the prototype can be developed from scratch and transformed in an extension of OLED software when OntoUML implement *perdurants* and social aspects, possibly by using rule-engines library. That way it would be possible to deal with the dynamic aspect of changes in axioms during the formulation of the full set of mapping rules. In addition, the "export to OWL" capability of OLED software could use an implementation of Allen's operators constraints through SWRL, as presented in (O'CONNOR e DAS, 2011). Using a mature CASE tool (EA) integrated to OLED was a great strategy from NEMO group and evolving this integration considering our MDD approach for MD design would enrich the tool capabilities.

Once our hybrid approach was designed as a decoupled process, prepared to interoperate with another one, integrating it with another mature hybrid method – from both theory and supporting tool points of view – could enhance the efficacy in automatically generating MD schemas. From the review on other approaches, GEM (and ORE module) looks the best ontological approach for this initiative, having the domain ontologies enriched with UFO and discovery mechanisms enhanced with our mapping rules. In addition, automated code generation based in MDA is a technique that can enrich this integration, such as the proposed in (GARCÍA-DÍAZ, FERNÁNDEZ, *et al.*, 2009) and (SADIQ, FAZZIKI e SADGAL, 2014).

Although our main aim was not focused on unstructured data modelling and querying, we described, in a high-level way, how to cope with textual data in the hybrid approach. There are several issues related to the reverse engineering from text activity by using NLP and IR techniques to be explored. The most remarkable, in our opinion, is the use of distributional models techniques, mainly the recent distributional semantics with best-effort method (FREITAS, CARVALHO, *et al.*, 2012). It can build ontologies from text, presenting balanced efficiency and efficacy results. Furthermore, we believe that the entity recognition and relation extraction activities can also be classified considering UFO concepts, making the step of building well-founded domain ontologies feasible to large ontologies when conciliating the result concepts from analysis-driven and source-driven. For the ETL integration approach there are some issues to be evolved too. For example, a gap in the JointOLAP process is the definition of how incremental load should work in DMs that stores textual information.

We believe that reasoning capabilities from formal ontology research area combined to NLP and IR approaches can improve BI/DW solutions, approximating the meaning of the "I" from BI to the "I" from AI.

References

ABELLÓ, A. **YAM² -** a multidimensional conceptual model. 2002. (PhD Thesis) -- Universitat Politecnica de Catalunya, Barcelona, 2002.

ADAMSON, C. ; KIMBALL, R. **Mastering data warehouse aggregates -** solutions for star schema performance. 1. ed. Indianapolis: John Wiley & Sons, 2006.

ALLEN, J. F. Maintaining knowledge about temporal intervals. **Communications of the ACM**, New York, v. 26, n. 11, p. 832-843, Nov. 1983. ISSN 10.1145/182.358434.

ALMEIDA, D. L. ; SILVA, T. L. **Estratégias e mecanismos para ETL textual**. 2009. Monografia (Projeto Final de Curso) – GRECO, PPGI, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2009.

BARCZYNSKI, W. M. et al. BI-style relation discovery among entities in text. In: INTERNATIONAL CONFERENCE ON DATA ENGINEERING WORKSHOP, 26., 2010, Long Beach. **Proceedings …** Washington, DC: IEEE, 2010.p.. 171-174.

BERKANI, N. ; BELLATRECHE, L. ; KHOURI, S. Towards a conceptualization of ETL and physical storage of semantic data warehouses as a service**. Journal Cluster Computing**, New York, v. 16, n. 4, p. 915-931 , Dec. 2013. ISSN 10.1007/s10586-013-0266-7.

BERLANGA, R. et al. Semantic Web Technologies for Business Intelligence. In: ZORRILLA, M. E. et al. **Business intelligence applications and the web** - models, systems, and technologies. Hershey: IGI Global, 2011. p. 310-339.

BHIDE, M. et al. Business intelligence using EROCS enhanced. In: IEEE INTERNATIONAL CONFERENCE ON DATA ENGINEERING, 24., 2008, Cancun. **Proceedings …** Washington, DC: IEEE, 2008. P. 1616-1619.

BIEM, A. et al. Real-time analysis and management of big time-series data. **IBM Journal of Research and Development**, Armonk, v. 57, n. 3-4, May/Jul. 2013. ISSN 10.1147/JRD.2013.2243551.

BUSSER, D. P. **An Information requirements collection and analysis model for business intelligence**. 2011. (Master Dissertation) -- Delft University of Technology, Delf, 2011.

CALERO, C. et al. Towards data warehouse quality metrics. In: INTERNATIONAL WORKSHOP ON DESIGN AND MANAGEMENT OF DATA WAREHOUSES, 3., 2001, Interlaken. **Proceedings ...** Interlaken: CEUR, 2001.

CALVI, C. Z. **Gerenciamento de serviços de TI e modelagem do processo de configuração ITIL em uma plataforma de serviços sensíveis a contexto**. 2007. Dissertação (Mestrado em Informática) – Programa de Pós-Graduação em Informática, Universidade Federal do Espírito Santo, Vitória, 2007.

CAO, L. ; ZHANG, C. ; LIU, J. Ontology based integration of business intelligence. **Web Intelligence and Agent Systems Journal**, Amsterdam, v. 4, n. 3, p. 313-325, Dec. 2006. ISSN 1570-1263.

CARDOSO, C. ; POLIDORO, M. **Gestão do risco da imagem institucional**. In: CONGRESSO DE COMUNICAÇÃO EMPRESARIAL, 3., Salvador. **Anais ...** Salvador: ABERJE, Associação Brasileira de Comunicação Organizacional , 2011.

CARRARETTO, R. **Separating ontological and informational concerns** – a model-driven approach for conceptual modeling. 2012. Dissertação (Mestrado em Informática) – Programa de Pós-Graduação em Informática, Universidade Federal do Espírito Santo, Vitória, 2012.

CHRISMENT, C. ; DOUSSET, B. ; ALAUX, J. DocCube: multi-dimensional visualisation and exploration of large document sets. **Journal of the American Society for Information Science and Technology**, New York, v. 54, n. 7, p. 650-659, May 2003. ISSN 10.1002/asi.10257.

COSTA, P. D. **Architectural support for context-aware applications:** from context models to services. 2007. (Ph.D Thesis) -- University of Twente, Enschede, 2007.

COSTA, P. D. et al. A Model-driven approach to situations: situation modeling and rule-based situation detection. In: IEEE INTERNATIONAL ENTERPRISE DISTRIBUTED OBJECT COMPUTING CONFERENCE, 16., 2012, Beijing. **Proceedings …** Los Alamitos: IEEE, 2012. p. 154-163. ISSN 978-0-7695-4785-5.

CUZZOCREA, A. ; BERTINO, E. Privacy preserving OLAP over distributed XML data: a theoretically-sound secure-multiparty-computation approach. **Journal of Computer and System Sciences**, New York, v. 77,n. 6, p. 965-987, Nov. 2011.

CUZZOCREA, A. ; SONG, I.-Y. ; DAVIS, K. C. Analytics over large-scale multidimensional data: the big data revolution! In: ACM 14TH INTERNATIONAL WORKSHOP ON DATA WAREHOUSING AND OLAP, 14., 2011. Glasgow. **Proceedings …** New York: ACM, 2011. p. 104, 2011. ISSN 978-1-4503-0963-9.

DECISIONPATH. **How effectively are companies using business analytics**. Gaithersburg: DecisionPath Consulting Research. . 2010.

DENTLER, K. et al. Comparison of reasoners for large ontologies in the OWL 2 EL profile. **Semantic Web Journal**, Amsterdam, v. 2, n. 2, p. 71-87, Apr. 2011. ISSN 10.3233/SW-2011-0034.

FALBO, R. A. et al. ODE: Ontology-based software development environment. In: ARGENTINE CONGRESS ON COMPUTER SCIENCE, 9., 2003, La Plata. **Proceedings …** La Plata: Universidad Nacional de la PlataArgentina, 2003. p. 1124 - 1135. ISSN 0104-6500.

FREITAS, A. et al. A Semantic best-effort approach for extracting structured discourse graphs from wikipedia. In: INTERNATIONAL SEMANTIC WEB CONFERENCE, 11., 2012, Boston. WORKSHOP ON THE WEB OF LINKED ENTITIES , 1., 2012, Boston. **Proceedings …** Boston: [s.n], 2012. p. 70-81, 2012.

GALHARDAS, H. ; LOPES, A. ; SANTOS, E. Support for user involvement in data cleaning. In: INTERNATIONAL CONFERENCE ON DATA WAREHOUSING AND KNOWLEDGE DISCOVERY, 13., 2011, Toulose. **Proceedings …** Toulouse: [s.n.], 2011. p. 136-151. ISSN 978-3-642-23543-6.

GANGEMI, A. et al. Sweetening ontologies with DOLCE. In: GÓMEZ-PÉREZ, A. ; BENJAMINS, V. R. (Eds). **Knowledge and engineering knowledge** management :ontologies and semantic web. 13th: International Conference, EKAW 2002 Sigënza Spain Proceedings. London: Springer-Verlag, 2002. p. 166- 181. (Lecture Notes in Computer Science, v. 2473). ISSN 3-540-44268-5.

GARCIA-ALVARADO, C.; ORDONEZ, C. Keyword search across databases and documents. In: KEYS 10 - INTERNATIONAL ON WORKSHOP KEYWORD SEARCH ON STRUCTURED DATA, 2.,2010, Indianapolis. **Proceedings …** New York: ACM, 2010. p. 1-6. ISSN 978-1-4503-0187-9.

GARCÍA-DÍAZ, V. et al. Automated code support generation for BI with MDA TALISMAN. **International Journal of Artificial Intelligence and Interactive Multimedia**, [S.l.], v. 1, n. 2, p. 87-93, Dec. 2009.

GOSAIN, A. ; NAGPAL, S. ; SABHARWAL, S. Quality metrics for conceptual models for data warehouse focusing on dimension hierarchies. **ACM SIGSOFT Software Engineering Notes**, New York, v. 36, n. 4, p. 1-5 , Jul. 2011. ISSN 10.1145/1988997.1989015.

GROSSMAN, D. A. et al. Data and text: a relational integrating structured approach**. Journal of the American Society of Information Science**, Washington, v. 48, n. 2, p. 122–132, Feb. 1997.

GUARINO, N. Formal ontology and information systems. In: INTERNATIONAL CONFERENCE ON FORMAL ONTOLOGY AND INFORMATION SYSTEMS, 1., 1998, Trento. **Proceedings …** Amsterdam: IOS Press, 1998. v. 1. p. 6-8. ISSN 9051993994.

GUARINO, N. A. G. P. Ontologies and knowledge bases: towards a terminological clarification. In: MARS, N. J. I. (Ed). **Towards very large knowledge bases**. 1. ed. Amsterdam: IOS Press, 1995. p. 25-32. ISSN 968739.

GUIZZARDI, G. **Ontological foundations for structural conceptual models**. 2005. (Ph.D. Thesis) -- University of Twente: Enschede, 2005.

GUIZZARDI, G. et al. Towards ontological foundations for the conceptual modeling of vents. In: NG, W. ; STOREY. V. C. ; TRUJILLO, J. C. (Eds). **Conceptual modeling**. 32nd International Conference ER 2013, Hong-Kong , Proceedings. Berlin: SpringeVerlag, 2013. p. 327-341. (Lecture Notes in Computer Science, v. 8217).

GUIZZARDI, G. ; FALBO, R. A. ; PEREIRA FILHO, J. G. From domain ontologies to object oriented frameworks. In: ONTO WORKSHOP ON ONTOLOGIES, 2001, Vienna**. Proceedings …** Vienna: [s.n.], 2001.

GUIZZARDI, G.; WAGNER, G. Some applications of a Unified Foundational Ontology in Business Modeling. **Business Systems Analysis with Ontologies**, [S.l.],. 345–367, 2005. ISSN 10.4018/978-1-59140-339-5.ch013.

GUIZZARDI, G. ; ZAMBORLINI, V. A Common foundational theory for bidging two levels in ontology-driven conceptual modeling. In: CZARNECKi, K. ; HEDIN, G. (Eds.). **Software Language Engineering**. International Conference of Software Language Engineering SLE 2012. Dresden, Germany. Revised Selected Papers.Berlin: Springer-Verlag, 2012. p. 286-310. (Lecture Notes in Computer Science, v. 7745).

HAAS, L. M. ; SOFFER, A. New challenges in information integration. In: PEDERSEN, T. D. ; MOHANIA, M. K. ; TJOA, A. M. (Eds.). **Data warehousing and knowledge discovery**. 11th International Conference on DaWarK. LinZ Austria 2009 Proceedings. Berlin: Springer-Verlag, 2009. (Lercture Notes in Computer Science, v. ISSN 978-3-642-03729-0.

HEUSELER, F. M. **Uma abordagem multifacetada para exploração integrada de dados estruturados e não-estruturados em ambientes OLAP**. 2010. (Mestrado em Informática) – Programa de Pós Graduação em Informática, Instituto de Matemática, Núcleo de Computação Eletrônica, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2010.

HOANG, D. T. H. et al. Towards the development of large-scale data warehouse application frameworks. In: MOLLER, C. ; CHAUDHRY, S. (Eds). **Re-conceptualizing enterprise information systems.** 5th IFIP WG 8.9 Working Conference COFENIS 2011, Aalborg, Denmark, Revised Selected Papers. Berlin: Springer-Verlag, 2011. (Lecture Notes in Business Information Processing, v. 105).p. 16-18,

HODGES, W. **Model theory**. Cambridge, UK: Cambridge University Press, 1993.v.1

INMON, W. H. **Building the data warehouse**. 4. ed. Indianapolis: John Wiley & Sons, Inc, 2005.

_______.**Building the data warehouse**. 1. ed. New York, NY: John Wiley & Sons, Inc, 1992.

INMON, W. H. ; STRAUSS, D. ; NEUSHLOSS, G. **DW 2.0 - The Architecture for the next generation of data warehousing**. 1. ed. Amsterdam: Morgan Kaufmann, 2008. (The Morgan Kaufmann Series and Data Management Systems).

JACKSON, D. **Software abstractions:** logic, language, and analysis. 1. ed. Cambridge: The MIT Press, 2012.

JARKE, M. et al. **Fundamentals of data warehouses**. 2. ed. Berlin: Springer-Verlag, 2003.

JOVANOVIC, P. et al. A requirement-driven approach to the design and evolution of data warehouses. **Information Systems**, Philadelphia, v. 44, p. 94-119, Aug. 2014.

KIMBALL, R. **The Data warehouse toolkit.** 1. ed. New York: John Wiley & Sons, Inc., 1996.

KIMBALL, R. ; Caseta, J. **The Data warehouse ETL toolkit -** practical techniques for extracting, cleaning, conforming and de ivering data. Indianapolis: John Wiley & Sons, 2004.

KIMBALL, R. ; ROSS, M. **The Data warehouse toolkit:** the complete guide to dimensional modeling. 2.. ed. Indianapolis: John Wiley & Sons, 2002.

_______. **The Data Warehouse Toolkit:** The definitve guide to dimensional modeling. 3. ed. Indianapolis: John Wiley & Sons, 2013.

KIMBALL, R. et al. **The Data warehouse lifecycle toolkit**. 2. ed. Indianapolis: John Wiley & Sons, 2008.

KUBIK, T. ; IWANIAK, A. Building and maintaining metadata repositories with the aid of ontology tools and technologies. In: GLOBAL SPATIAL DATA INFRASTRUCTURE ASSOCIATION WORLD CONFERENCE, 2012, Quebec. **Proceedings ...** New Brunswick, GSDI, 2012. p. 143-150.

LEE, J. et al. Integrating and structured data text: a multi-dimensional approach. In: ITERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY: CODING AND COMPUTINNG , 2000, Las Vegas. **Proceedings ...** Los Alamitos: IEEE, 2000. p. 264-269. ISSN 0-7695-0540-6.

LEE, J. ; GROSSMAN, D. ; ORLANDIC, R. MIRE: a multidimensional information retrieval engine for structured data and text. In: INTERNATIONAL CONFERENCE ON CODING INFORMATION TECHNOLOGY AND COMPUTING, 2000, Las Vegas. **Proceedings …** Los Alamitos: IEEE, 2000. p. 224-229. ISSN 0-7695-1506-1.

LIN, C. X. et al. Text cube - computing IR measures for multidimensional text database analysis. In: IEEE INTERNATIONAL CONFERENCE ON DATA MINING, 8., 2008, Pisa. **Proceedings …** Los Alamitos: IEEE, 2008. p. 905-910. ISSN 1550-4786.

LIU, H. ; XIE, D. ; WU, S. Notice of retraction research approach measuring the strategic readiness of intangible assets. In: INTERNATIONAL CONFERENCE ON FUTURE INFORMATION TECHNOLOGY AND MANAGEMENT ENGINEERING, 2., 2009, Sanya. **Proceedings …** Los Alamitos: IEEE, 2009. p. 99-103. ISSN 978-0-7695-3880-8.

LIU, Y. et al. Real-time data pre-processing technique for efficient feature extraction in large scale datasets. In: ACM CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 17., 2008. Napa Valley. **Proceedings …** New York: ACM, 2008. p. 981-990. ISSN 978-1-59593-991-3.

LUHN, H. P. A Business intelligence system. **IBM Journal of Research and Development**, Armonk, v. 2, n. 4, p. 314-319, 1958. ISSN 10.1147/rd.24.0314.

MA, H. et al. Cloud warehousing. **The Journal of Universal Computer Science** ,[S.l.], v. 17, n. 8, p. 1183-1201, 2011

MALINOWSKI, E.; ZIMÁNYI, E. **Advanced data warehouse design - from conventional to spatial and temporal applications**. Berlin: Springer, 2009.

_______. OLAP hierarchies: A conceptual perspective. In PERSON. A. ; STIRNA, J. (Eds.). **Advanced Information Systems Engineering.** 16th International Conference CAíSE 2004. Riga Latvia. Proceedings. Berlin: Springer-Verlag, 2004. p. 477-491. (Lecture Notes in Computer Science, v. 3084). ISSN 978-3-540-25975-6.

_______. Representing spatiality in a conceptual multidimensional model. In: GIS'04 - ANNUAL ACM INTERNATIONAL WORKSHOP ON GEOGRAPHIC INFORMATION SYSTEMS, 12., 2004, Washington. **Proceedings …** New York: ACM, 2004. p. 12-22. ISSN 1-58113-979-9.

MARSHALL, C. C. et al. Supporting research collaboration through bi-level file synchronization. ACM INTERNATIONAL CONFERENCE ON SUPPORTING GROUP WORK, 17., 2012, Sanibel Island. **Proceedings …** New York: ACM, 2012. p. 165-174. ISSN 978-1-4503-1486-2.

MCCABE, M. C. et al. On the design and evaluation of a multidimensional approach to information retrieval. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 23., 2000, Athens. **Proceedings …** New York: ACM, 2000. p. 363-365. ISSN 1-58113-226-3.

MEALY, G. H. Another look at data. In: AFIPS'67 JOINT FALL COMPUTER CONFERENCE, 1967, Anaheim, CA. **Proceedings …** New York: AFIPS/ACM, 1967**.** p. 525-534. ISSN 10.1145/1465611.1465682.

MEDINA, E. ; TRUJILLO, J. A Standard for representing multidimensional properties: the common warehouse metamodel. In: MONOLOPOULOS, Y. ; NÁVRAT, P. (Eds.). **Advances in Databases and Information Systems**. 6th East European Conference ADBIS 2002. Bratislava, Slovakia. Proceedings. London: Springer-Verlag, 2002. p. 232-247. ISSN 3-540-44138-7.

MIELKE, I. T. **Uma abordagem baseada em modelos para especificação e detecção de situações em sistemas sensíveis ao contexto**. 2013. Dissertação (Mestrado em Informática) – Departamento de Informática, Centro Tecnológico, Universidade Federal do Espírito Santo, Vitória, 2013.

MOREIRA, J. L. R. **HiDoctor OLAP - Abordagem facetada para análise integrada de dados heterogêneos:** um exemplo da área médica. [S.l.]: Undergraduate report, 2008.

MOREIRA, J. L. R. ; CORDEIRO, K. F. ; CAMPOS, M. L. M. **DoctorOLAP:** ambiente para análise multifacetada de prontuários médicos. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, 24., 2009, Fortaleza. **Anais...** Fortaleza, SBC, 2009.

_______. **JointOLAP –** sistema de informação para exploração conjunta de dados estruturados e textuais - um estudo de caso no setor elétrico. SIMPÓSIO BRASILEIRO DE SISTEMA DE INFORMAÇÃO, 9., 2013, João Pessoa. **Anais ...** João Pessoa: SBC, 2013**.**

MOREIRA, J. et al. OntoWarehousing - multidimensional design supported by a foundational ontology – a temporal perspective. In: BELLATRECH, L. ; MOHANIA, M. K. (Eds.). **Data warehousing and knowledge discovery**. 16th International Conference (DaWaK). Munich, Germany Proceedings. [S.l.]: Springer-Verlag, 2014. p. 35-44. ISSN LNCS 8646.

MORIN, E. **A Cabeça bem feita**. 1. ed. Rio de Janeiro: Bertrand Brasil, 2003.

MOSS, L. T. ; ATRE, S. **Business intelligence roadmap -** the complete project lifecycle for decision-support applications. 1. ed. [S.l.]: Addison-Wesley Professional, 2003.(Addison Wesley Information Technology Series).

MOYA, L. G. et al. . Integrating web feed opinions into a corporate data warehouse. In: INTERNATIONAL WORKSHOP ON BUSINESS INTELLIGENCE AND THE WEB, 2.,2011, Uppsala. **Proceedings …** New York: ACM, 2011. p. 20-27, ISSN 978-1-4503-0610-2.

MUSLEH, D. et al. Efficient multidimensional simple path query processing algorithm for XML data. In: CONFERENCE ON INFORMATION SCIENCE: INTERNATIONAL AND APPLICATIONS, 2013, Pattaya. **Proceedings …** Piscataway: IEEE, 2013. p. 1-5. ISSN 978-1-4799-0602-4.

MYLOPOULOS, J. et al. Telos: representing knowledge about information systems. **ACM Transactions on Information Systems,** New York, v. 8, n. 4, p. 325-362, Oct. 1990. ISSN 10.1145/102675.102676.

NESAVICH, A. ; INMON, W. H. **Tapping into unstructured data:** integrating unstructured data and textual analytics into business intelligence. 1. ed. [S.l.]: Prentice Hall, 2007.

NEVES, P. I. **Uma estratégia para apoioar a decisão baseada em mineração de textos livres**. Dissertação. 2012. (Mestrado) – Departamento de Ciência e Tecnologia, Instituto Militar de Engenharia, Rio de Janeiro, 2012.

NGUYEN THANH BINH, A. M. T. ; MANGISENGI, O. MetaCube-X: an XML metadata foundation for interoperability search among web warehouses. In: INTERNATIONAL WORKSHOP ON DESIGN AND MANAGEMENT OF DATA WAREHOUSES, 3., 2001, Interlaken. **Proceedings …** Interlaken: CEUR, 2001.

NIEMI, T. ; NIINIMÄKI, M. Ontologies and summarizability in OLAP. In: ACM SYMPOSIUM ON APPLIED COMPUTING, 2010, Sierre. **Proceedings …** New York: ACM, 2010. p. 1349-1353. ISSN 978-1-60558-639-7.

NIRENBURG, S. ; RASKIN, V. Ontological semantics, formal ontology, and ambiguity. In: INTERNATIONAL CONFERENCE ON FORMAL ONTOLOGY IN INFORMATION SYSTEMS, 2., 2001, Ogunquit, Maine. **Proceedings …** New York: ACM, 2001. p. 151-161. ISSN 1-58113-377-4.

O'CONNOR, M. J. ; DAS, A. K. A Method for representing and querying temporal Information in OWL. In: FRED, A. ; FELIPE, J. ; GAMBOA, H. **Biomedical engineering systems and technologies**, Valencia, Spain,Revised Selected Papers.Berlin: Springer-Verlag, 2011. p. 97-110, (Communications in Computer and Information Science, v. 127). ISSN 978-3-642-18472-7.

OUESLATI, W. ; AKAICHI, J. A Survey on Data Warehousing Evolution. **International Journal of Database Management Systems** [S.l.], v. 2, n. 4, p. 11-24, Nov. 2010. ISSN 10.5121/ijdms.2010.2402.

PARDILLO, J. ; MAZÓN, J.-N. Using ontologies for the design of data warehouses. **International Journal of Database Management Systems** [S.l.], v. 3, n. 2, p. 73–87, May 2011.

PARENT, C. ; SPACCAPIETRA, S. ; ZIMÁNYI, E. **Conceptual modeling for traditional and spatio-temporal applications**. New York: Springer, 2006.

PARK, B.-K. ; SONG, I.-Y. Toward total business intelligence incorporating structured and unstructured data. In: INTERNATIONAL WORKSHOP ON BUSINESS INTELLIGENCE AND THE WEB, 2.,2011, Uppsala. **Proceedings …** New York: ACM, 2011. p. 12-19, ISSN 978-1-4503-0610-2.

PENA, R. A. P. **Suporte semântico à publicação de conteúdo jornalístico na web**. 2012. Dissertação (Mestrado em Informática) – Programa de Pós-Graduação em Informática, Departamento de Informática do Centro Técnico e Científico, PUC, Rio de Janeiro, 2012.

PEREIRA, I. ; COSTA, P. ; ALMEIDA, J. A Rule-based platform for situation management. In: 2013 IEEE INTERNATIONAL MULTI-DISCIPLINARY CONFERENCE ON COGNITIVE METHODS IN SITUATION AWARENESS AND DECISION SUPPORT, 2013, San Diego, CA. **Proceedings …** Los Alamitos: IEEE, 2013. p. 83 - 90. ISSN 978-1-4673-2437-3.

PRAT, N. ; AKOKA, J. ; COMIN-WATTIAU, I. Transforming multidimensional models into OWL-DL ontologies. In: IEEE International Conference on Research Challenges in Information Science, 6. 2012, Valencia**. Proceedings …** Valencia: IEEE, 2012. p. 1-12.

PRESSMAN, R. **Engenharia de software**. 5. ed. New York: Makron Books, 2002.

RAVAT, F .; TESTE, O. ; TOURNIER, R. OLAP aggregation function for textual data warehouse. In: INTERNATIONAL CONFERENCE ON ENTERPRISE INFORMATION SYSTEMS, 9., 2007, Funchal, Madeira – Portugal. **Proceedings ...** New York: ACM, 2007.

ROCHA, L. B. L. **Ontologia de notícias - um modelo para classificação do conteúdo dos jornais on-line brasileiros, segundo a lógica da web semântica**. 2012. Dissertação (Mestrado em Design) – Programa de Pós-Graduação em Design, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2012.

RODRIGUES, A. D. S. **Extração semi-automática de percepção analítica em bases relacionais**. Dissertação (Mestrado em Informática) – Programa de Pós-Graduação em Informática, Instituto de Matemática, Núcleo de Computação Eletrônica, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2004.

ROMERO, O. **Automating the multidimensional design of data warehouses**. 2010. Thesis (Doctor en Informàtica) – Programa en Doctorat en Software, Departmente de Llenguatges I Sistemes Informàtics, Universitat Plitècnica de Catalunya, Barcelona, 2010.

ROMERO, O. ; ABELLÓ, A. A framework for multidimensional design of data warehouses from ontologies. **Data & Knowledge Engineering,** Amsterdam, v. 69, n.11, p. 1138-1157, Nov. 2010. ISSN 10.1016/j.datak.2010.07.007.

ROMERO, O. ; SIMITSIS, A. ; ABELLÓ, A. GEM: requirement-driven generation of ETL and multidimensional conceptual designs. In: CUZZOCREA, A. DAYAL, U. (Eds.). **Data warehousing and knowledge discovery**. 13th International Conference, DaWaK 2011. Toulouse, France, August 2011, Proceedings Berlin: Springer-Verlag, 2011. p. 80-95. (Lecture Notes in Computer Science, v. 6862). ISSN 978-3-642-23544-3.

ROY, P. et al. Towards automatic association of relevant unstructured content with structured query results. In: ACM INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 14., 2005, Bremen, Germany. **Proceedings …** New York: ACM, 2005. p. 405-412. ISSN 1-59593-140-6.

RUSSOM, P. **BI search and text analysis tics -** new additions to the BI technology stack. Reston: TDWI Best Practices Report, 2007.

SADIQ, A. ; FAZZIKI, A. E. ; SADGAL, M. An Agent based etl system: towards an automatic code generation. **World Applied Sciences Journal**, Dubay, v. 31 n. 5, p. 979-989, Jun. 2014. ISSN 95301833.

SAIAS, J. et al. BINLI: An Ontology-based natural language interface for multidimensional data analysis. **Intelligent Information Management,** [S.l.], v. 4 n. 5, p. 225-230, Sept. 2012. ISSN 10.4236/iim.2012.45033.

SALES, T. ; BARCELOS, P. P. ; GUIZZARDI, G. Identification of semantic anti-patterns in ontology-driven conceptual modeling via Visual Simulation. In: INTERNATIONAL CONFERENCE ON FORMAL ONTOLOGY IN INFORMATION SYSTEMS, 7., 2012, Graz. 4th INTERNATIONAL WORKSHOP ON ONTOLOGY-DRIVEN INFORMATION SYSTEMS, 4., 2012, Graz. **Proceedings …** Graz: Medical University of Graz, 2012.

SALTOR, F. ; CASTELLANOS, M. ; GARCÍA-SOLACO, M. Suitability of datamodels as canonical models for federated databases. **ACM SIGMOD Record**, New York, v. 20, n. 4, p. 44-48, Dec. 1991. ISSN 10.1145/141356.141377.

SANTOS, F. B. **Um Arcabouço genérico para análise de opiniões expressas em mídias sociais, ano de obtenção**.2013. Dissertação (Mestrado em) – Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2013.

SATISH, N. et al. Navigating the maze of graph analytics frameworks using massive graph datasets. In: 2014 ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 2014, Snowbird. **Proceedings …** New York: ACM, 2014. p. 979-990. ISSN 978-1-4503-2376-5.

SCHULZ, H.-J. ; HADLAK, S. ; SCHUMANN, H. The Design space of implicit hierarchy visualization: a survey. **IEEE Transactions on Visualization and Computer Graphics**, Washington, v. 17, n. 4 , p. 393-411, Apr. 2011. ISSN 1077-2626.

SERRANO, M. et al. Metrics for Data warehouse conceptual models understandability. **Information and Software technology Journal**, Butterworth-Heinemann Newton, MA, v. 49, n. 8, p. 851-870 , Aug. 2007. ISSN 10.1016/j.infsof.2006.09.008.

SHAH, N. et al. Ontological on-line analytical processing for integrating energy sensor data. **IETE Technical Review,** 26, p. 375, 2009. ISSN 0.4103/0256-4602.55271.

SOARES, V. J. A. **Modelagem incremental no ambiente de data warehousing**. Dissertação (Mestrado em Informática) – Programa de Pós-Graduação em Informática, Instituto de Matemática, Nucleo de Computação Eletrônica, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 1998.

SOLINGEN, R. V. ; BERGHOUT, E. **The Goal/question/metric method -** a practical guide for quality improvement of software development. 1. ed. London: McGraw-Hill, 1999.

TDWI. **2013 TDWI Benchmark Report**. 2013. organizational and performance metrics for business intelligence teams. [S.l.]. TDWI, 2013.

THOLLOT, R. et al. Text-to-query - dynamically building structured analytics to illustrate textual content. In: INTERNATIONAL CONFERENCE ON EXTENDING DATABASE TECHNOLOGY, 13.,2010, Lausanne. **Proceedings …** New York: ACM, 2010. p. 1-8. ISSN 978-1-60558-990-9.

TORRES, G. M. et al. Collaborative construction of visual domain ontologies using metadata based on foundational ontologies. In: JOINT IV SEMINAR ON ONTOLOGY RESEARCH IN BRAZIL, 4., 2011, Gramado. INTERNATIONAL WORKSHOP ON METAMODELS, ONTOLOGIES AND SEMANTIC TECHNOLOGIES, 6., 2011, Gramado, **Proceedings …** Gramado: CEURS , 2011. p. 201-206.

TRANSACTION PROCESSING PERFORMANCE COUNCIL. **TPC benchmark**, 2002. Disponivel em: <http://www.tpc.org/>. Acesso em: jul. 2014.

TSENG, F. S. C. ; CHOU, A. Y. H. The concept of document warehousing for multi-dimensional modeling of textual-based business intelligence. **Decision Support Systems**l, Amsterdam, v. 42 n. 2, p. 727-744, Nov. 2006. ISSN 10.1016/j.dss.2005.02.011.

VIEIRA, T. J. S. **Análise visual para sistemas OLAP -** visualização de hierarquias em modelos multidimensionais. 2013. Dissertação (Mestrado) –Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2013.

WINTER, R. ; STRAUCH, B. A. Method for demand-driven information requirements analysis in data warehousing projects. In: ANNUAL HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES, 36., 2003, Waikoloa. **Proceedings ...** Waskington, DC: IEEE, 2003. Track 8. v. 8. p. 231.1. ISSN 0-7695-1874-5.

WOHLIN, C. et al. **Experimentation in software engineering**. New York: Springer, 2012.

ZAMBORLINI, V. **Estudo de alternativas de mapeamento de ontologias da linguagem ontoUML para OWL abordagens representação de informação temporal**. 2011. Dissertação (Mestrado em Informática) – Programa de Pós-Graduação em Informática, Centro Tecnológico, Universidade Federal do Espírito Santo, Vitória, 2011.

ZHANG, D. et al. Topic modeling for OLAP on multidimensional text databases: topic cube and its applications. **Statistical Analysis and Data Mining Journal - Best of SDM'09**, New York, v. 2 n. 5-6, p. 378-395 , Dec. 2009. ISSN 10.1002/sam.v2:5/6.

ZHOU, J. ; JIN, L. ; HAN, S. Unified hierarchical iterate model of human conceptualization and cognition. In: IEEE INTERNATIONAL CONFERENCE ON COGNITIVE INFORMATICS,8., 2009, Hong Kong. **Proceedings…** Los Alamitos: IEEE, 2009. p. 44-51. ISSN 978-1-4244-4642-1.

Attachments

ATTACHMENT A – DB SCRIPTS

Available at links:

https://drive.google.com/file/d/0B36SsM5rekUCM1YwVEVBV015Wnc/edit?usp=sharing

https://drive.google.com/file/d/0B36SsM5rekUCWWxQVUpFM0NVd3M/edit?usp=sharing

https://drive.google.com/file/d/0B36SsM5rekUCZ3MteEE2M0hLMnM/edit?usp=sharing

ATTACHMENT B – EA SOLUTION

Available at link:

https://drive.google.com/file/d/0B36SsM5rekUCMWdaVXM0UFdreDQ/edit?usp=sharing

ATTACHMENT C – PROTOTYPE SOURCE CODE

Available at link:

https://drive.google.com/folderview?id=0B36SsM5rekUCdWxCbkZCaUExRmc&usp=sharing

Appendices

APPENDIX A – JOINTOLAP FRAMEWORK FOR TEXTUAL ETL

Based in a prior works (MOREIRA, 2008) (MOREIRA, CORDEIRO e CAMPOS, 2009) (ALMEIDA e SILVA, 2009) an architecture to integrate Textual ETL and conventional ETL was designed and experimented in (MOREIRA, CORDEIRO e CAMPOS, 2013). At first, the unstructured data is extracted from the data sources and loaded in a Textual ODS DB (named ODS_Textual). Parallel to it, conventional ETL process extracts structured data, as usual, loading it in a conventional ODS DB (ODS)[31]. A third process, called *Linkage* – based in the concept presented in (NESAVICH e INMON, 2007), is responsible of populating the relational DB of the DW solution. It joins data arisen from textual information and structured data sources. During this phase the linkage is physically built, providing an integrated way for joint exploration over both data source types through OLAP tools. Figure 01 illustrates this architecture.

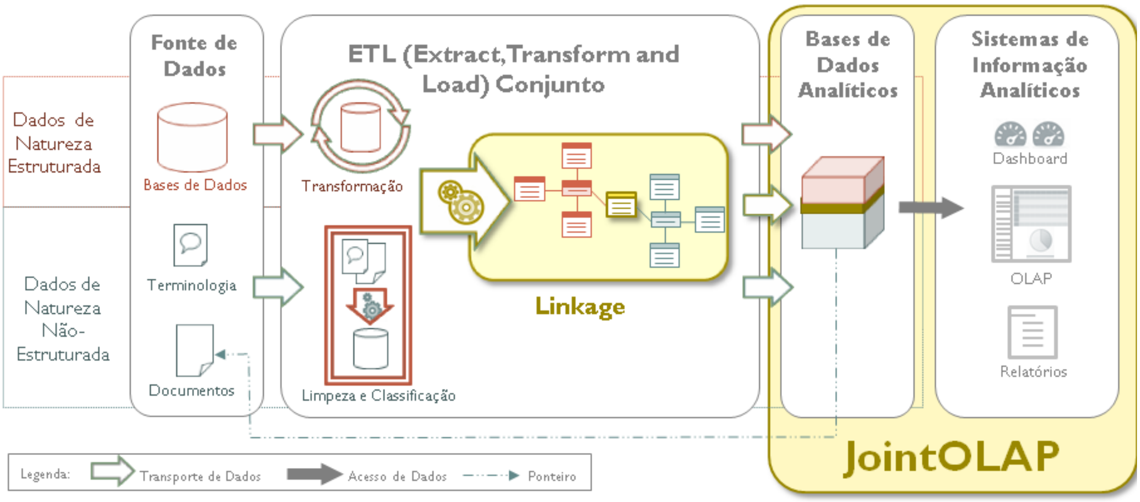


Figura 0.1: JointOLAP architecture from (MOREIRA, CORDEIRO e CAMPOS, 2013)

[31] We decide to set a DB for each ODS (structured and unstructured) most because of size considerations. Other types of approaches can also be applied, as having only one ODS DB or even distributed DBs. Also, having two separated DB is an organized manner such as using schemas (prefixes).

In the Extracting phase a heavy work is made to bring semantic from the text documents. Firstly, the Textual ETL performs a set of IR operations, such as orthographic correction, stop word elimination, tokening, synonymous resolution and stemming. The output of the Textual ETL is a relational DB schema loaded with indexed text, which structures the relation of files with terms and their morphological classifications in the context, called Terminological DB. Moreover, language libraries can provide terms synonymous and radicals, permitting a high quality classification and different perspectives of analysis over text. Those terminology libraries for natural language can be found, for example, in terms glossaries and taxonomies, exercising a fundamental role in processing text documents.

Thereafter, the terms are categorized into facets, which can be enriched with concepts found in terminological tools such as domain glossary of terms mentioned above. The concern with the correct usage and interpretation of data is a constant in this architecture, representing a severe risk for the BI/DW project. This treatment will make navigation easier over the textual information in the presentation layer. Afterwards, conventional ETL is executed to extract data from the Terminological DB and loads it in the MD schema, computing terms occurrences for documents analysis. Textual information can also be the starting point of analysis. The result of the Textual ETL is the analytical database with the unstructured data in a relational DB.

Having both ODS DBs loaded (structured and unstructured), the *Linkage* process constructs the connection between them in the final MD schema. The linkage can be implemented based on the temporal relation fact pattern resulted from rule 3, for example. However, in previews work (HEUSELER, 2010), the linkage could also be implemented through shared dimensions. By mapping dimensions from structured data sources to entities found in text it is possible to use them in joint exploration. The facts are connected by the shared dimensions, called *Linkage Dimensions*, permitting the association and navigation in the facts from both data universes. Most of the time the time dimension is one of those dimensions and through this analysis perspective it is possible to analyze the behaviour of a certain fact over the years, and from that, make decisions to enhance or inhibit that fact. Moreover, with textual information available, these analyzes are enriched allowing its correlation with facts that were not originally described and represented in a structured way. However, the shared dimensions do not allow an easy and usable way to explore the linkage, lacking in semantic expressivity and needing SQL (or MDX) queries for joint analysis most of the times.

The default MD snowflake schema for unstructured data counts with a factless fact pattern for text analysis. It presents a fact of terms occurrences and dimensions for documents, its creation date (with conventional time hierarchy) and a term hierarchy, which counts with the facets in the highest level and categories in a middle level.

APPENDIX B – COMMON ANALYSES MADE IN DISTURBANCES BI

The analyses listed here are common analyses made in Disturbance BI solution by domains specialists.

1. Number of disturbances

The total number of disturbances involving equipment in SIN has been increasing for the last years. This is an expected behaviour because of the SIN continuous expansion, aggregating new transmission lines, power transformers, among other equipment and power plants. Figure 0.2 illustrates in blue the number of disturbances and in orange the number of load cuts. The load cuts indicate wherever an electricity interruption to the consumers occurs. By comparing those measures, the difference between them, we perceive that the tax of load cuts to disturbances is decreasing. In 2007 it was 12.6%, in 2008 it was 11.9%, in 2009 it was 9.5%, in 2010 it was 8.5%, in 2011 it was 7.9%, in 2012 it was 6.9%, in 2013 it was 4.4% (until July). This indicates the quality increase of the protection and emergence actions in the SIN operation for the last years.

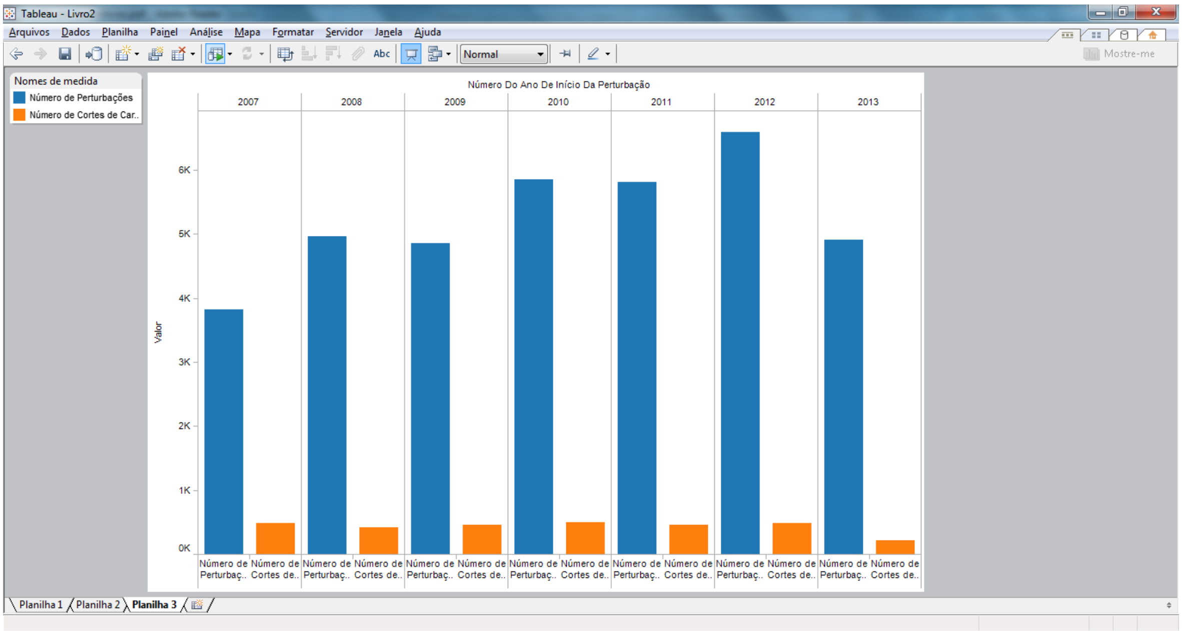


Figure 0.2: Comparison between disturbances number and load cuts number measures

In the Figure 0.3 the analysis over the severe disturbances is generated by applying a filter in the load cut level attribute, selecting the “Level of load cut grater then 99MW”. Comparing to the last graph, in 2012, only 1.7% of the disturbances caused serious impact on people’s livelihood.

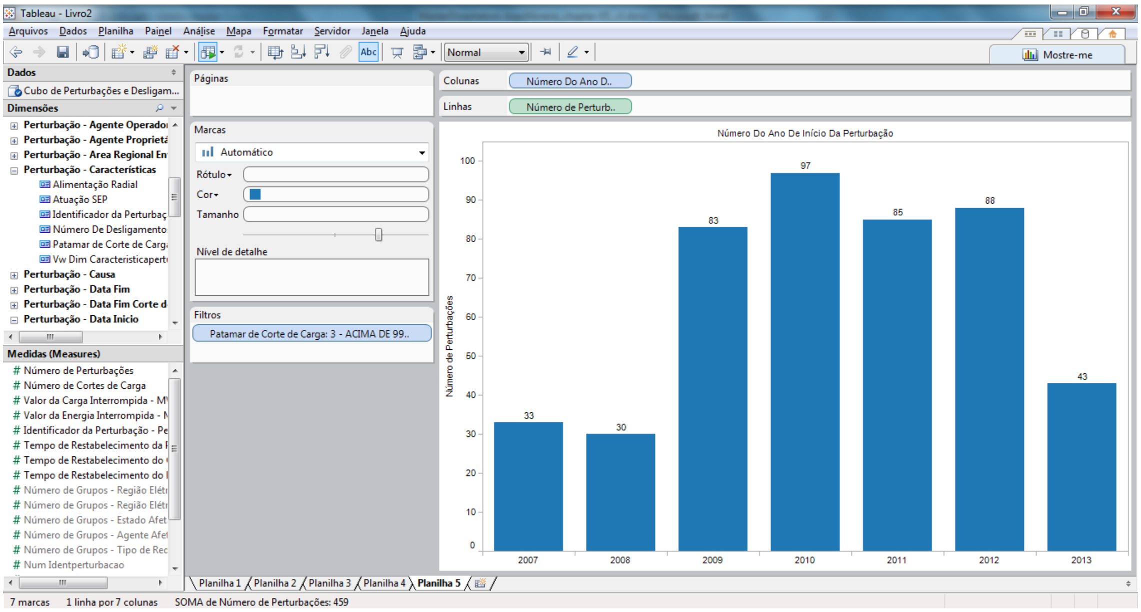


Tableau - Livro2
Arquivos Dados Planilha Painel Análise Mapa Formatar Servidor Janela Ajuda
Normal
Mostre-me
Dados
Cubo de Perturbações e Desligam...
Dimensões
Perturbação - Agente Operador
Perturbação - Agente Proprietá
Perturbação - Area Regional En
Perturbação - Características
Alimentação Radial
Atuação SEP
Identificador da Perturbaç
Número De Desligamento
Patamar de Corte de Carga
Vw Dim Caracteristicapert
Perturbação - Causa
Perturbação - Data Fim
Perturbação - Data Fim Corte d
Perturbação - Data Inicio
Medidas (Measures)
Número de Perturbações
Número de Cortes de Carga
Valor da Carga Interrompida - M
Valor da Energia Interrompida - N
Identificador da Perturbação - Pe
Tempo de Restabelecimento da I
Tempo de Restabelecimento do
Tempo de Restabelecimento do I
Número de Grupos - Região Elétr
Número de Grupos - Região Elétr
Número de Grupos - Estado Afet
Número de Grupos - Agente Afet
Número de Grupos - Tipo de Rec
Num Identperturbacao
Páginas
Marcas
Automático
Rótulo
Cor
Tamanho
Nível de detalhe
Número Do Ano De Início Da Perturbação
Número de Perturbações
100
90
80
70
60
50
40
30
20
10
0
33
30
83
97
85
88
43
2007
2008
2009
2010
2011
2012
2013
Planilha 1
Planilha 2
Planilha 3
Planilha 4
Planilha 5

Colunas: Número Do Ano D..

Linhas: Número de Perturb..

Filtros: Patamar de Corte de Carga: 3 - ACIMA DE 99..

7 marcas 1 linha por 7 colunas SOMA de Número de Perturbações: 459

Figure 0.3: Number of disturbances with load cut level major then 99MW

2. Disturbances by source equipments

A disturbance in SIN always initiates in one or more equipment. For example, when an atmospheric discharge hits a transmission line or a power transformer has a short circuit. Those equipments first affected are called the “source equipments”. As shown in Figure 0.4, the most common source equipments are the generator units, followed by transmission lines, power transformers with three terminals, bank of capacitors and others from outside the network operation.

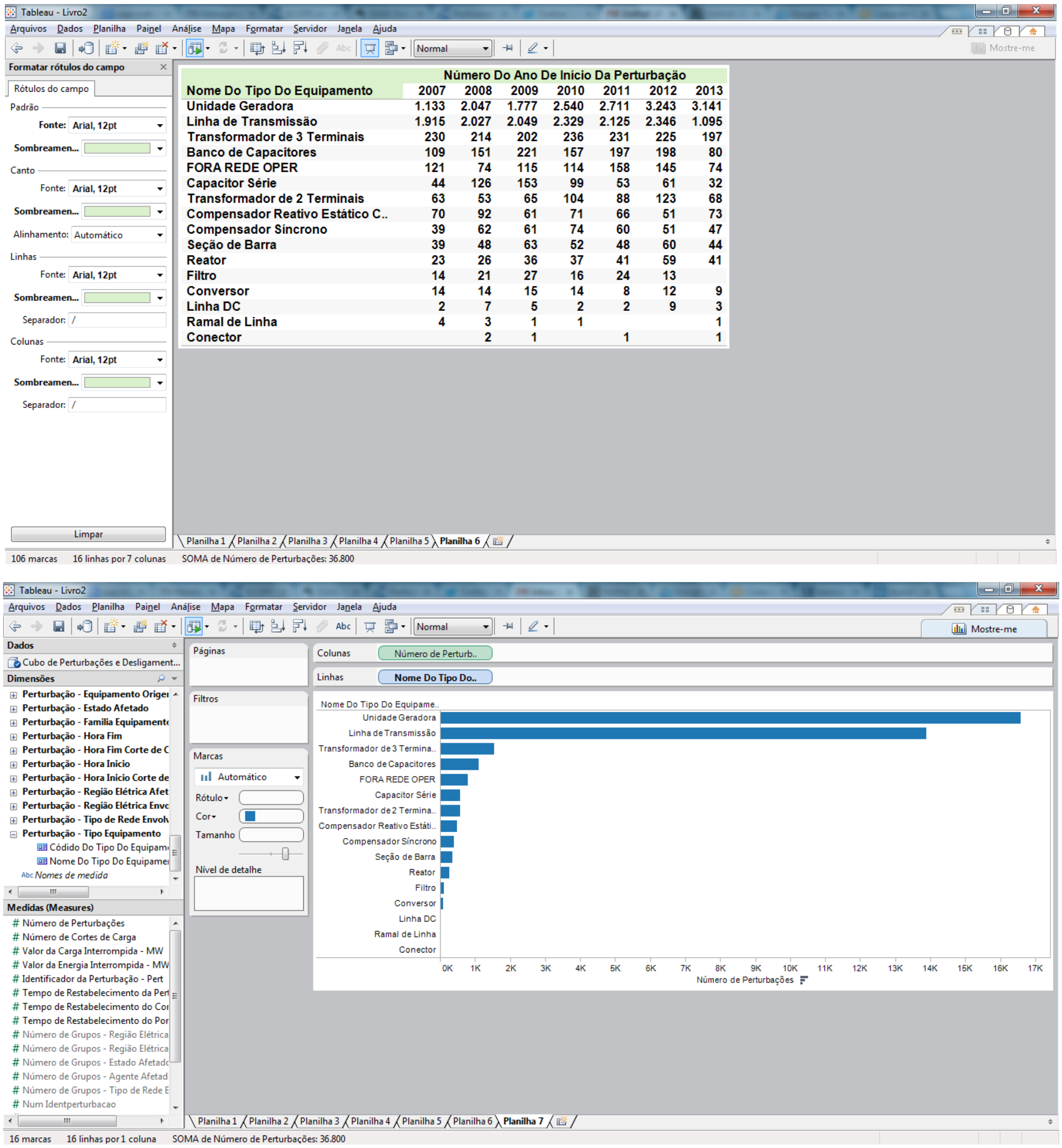


Tableau - Livro2
Arquivos Dados Planilha Painel Análise Mapa Formatar Servidor Janela Ajuda
Normal
Mostre-me
Formatar rótulos do campo
Rótulos do campo
Limpar
Planilha 1 Planilha 2 Planilha 3 Planilha 4 Planilha 5 Planilha 6

Padrão
Fonte: Arial, 12pt
Sombreamen...
Canto
Fonte: Arial, 12pt
Sombreamen...
Alinhamento: Automático
Linhas
Fonte: Arial, 12pt
Sombreamen...
Separador: /
Colunas
Fonte: Arial, 12pt
Sombreamen...
Separador: /

	Número Do Ano De Início Da Perturbação						
Nome Do Tipo Do Equipamento	2007	2008	2009	2010	2011	2012	2013
Unidade Geradora	1.133	2.047	1.777	2.540	2.711	3.243	3.141
Linha de Transmissão	1.915	2.027	2.049	2.329	2.125	2.346	1.095
Transformador de 3 Terminais	230	214	202	236	231	225	197
Banco de Capacitores	109	151	221	157	197	198	80
FORA REDE OPER	121	74	115	114	158	145	74
Capacitor Série	44	126	153	99	53	61	32
Transformador de 2 Terminais	63	53	65	104	88	123	68
Compensador Reativo Estático C..	70	92	61	71	66	51	73
Compensador Síncrono	39	62	61	74	60	51	47
Seção de Barra	39	48	63	52	48	60	44
Reator	23	26	36	37	41	59	41
Filtro	14	21	27	16	24	13	
Conversor	14	14	15	14	8	12	9
Linha DC	2	7	5	2	2	9	3
Ramal de Linha	4	3	1	1			1
Conector		2	1		1		1

106 marcas 16 linhas por 7 colunas SOMA de Número de Perturbações: 36.800

Tableau - Livro2
Arquivos Dados Planilha Painel Análise Mapa Formatar Servidor Janela Ajuda
Normal
Mostre-me
Dados
Cubo de Perturbações e Desligament...
Dimensões
Perturbação - Equipamento Origem
Perturbação - Estado Afetado
Perturbação - Família Equipamento
Perturbação - Hora Fim
Perturbação - Hora Fim Corte de C
Perturbação - Hora Início
Perturbação - Hora Início Corte de
Perturbação - Região Elétrica Afet
Perturbação - Região Elétrica Envo
Perturbação - Tipo de Rede Envolv
Perturbação - Tipo Equipamento
Códido Do Tipo Do Equipame
Nome Do Tipo Do Equipame
Nomes de medida
Medidas (Measures)
Número de Perturbações
Número de Cortes de Carga
Valor da Carga Interrompida - MW
Valor da Energia Interrompida - MW
Identificador da Perturbação - Pert
Tempo de Restabelecimento da Pert
Tempo de Restabelecimento do Cor
Tempo de Restabelecimento do Por
Número de Grupos - Região Elétrica
Número de Grupos - Região Elétrica
Número de Grupos - Estado Afetado
Número de Grupos - Agente Afetad
Número de Grupos - Tipo de Rede E
Num Identperturbacao
Páginas
Filtros
Marcas
Automático
Rótulo
Cor
Tamanho
Nível de detalhe
Colunas
Número de Perturb..
Linhas
Nome Do Tipo Do..
Nome Do Tipo Do Equipame..
Unidade Geradora
Linha de Transmissão
Transformador de 3 Termina..
Banco de Capacitores
FORA REDE OPER
Capacitor Série
Transformador de 2 Termina..
Compensador Reativo Estáti..
Compensador Síncrono
Seção de Barra
Reator
Filtro
Conversor
Linha DC
Ramal de Linha
Conector
0K
1K
2K
3K
4K
5K
6K
7K
8K
9K
10K
11K
12K
13K
14K
15K
16K
17K
Número de Perturbações
Planilha 1 Planilha 2 Planilha 3 Planilha 4 Planilha 5 Planilha 6 Planilha 7

16 marcas 16 linhas por 1 coluna SOMA de Número de Perturbações: 36.800

Figure 0.4: Disturbances by equipment types. (a) values (b) graph

This distribution changes when we filter the number of severe disturbances (load cut level greater then 99MW), evidencing that the transmission line is the main source equipment, followed by the power transformers, representing together almost 60% of the total grave disturbances (Figure 0.5).

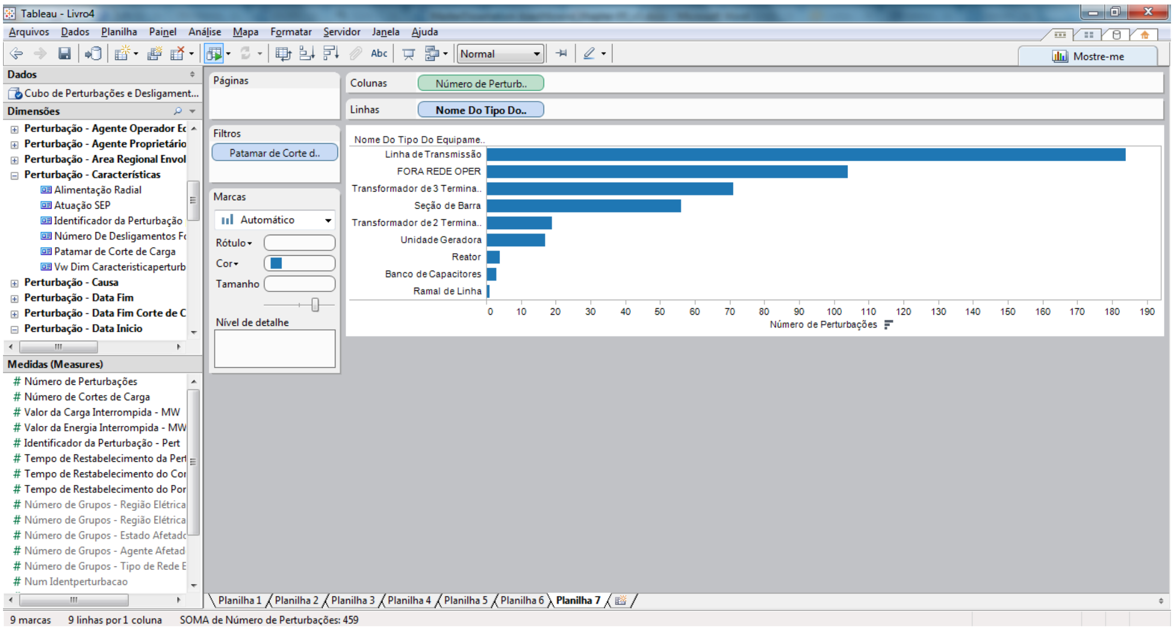


Tableau - Livro4
Arquivos Dados Planilha Painel Análise Mapa Formatar Servidor Janela Ajuda
Abc
Normal
Mostre-me
Dados
Cubo de Perturbações e Desligament...
Dimensões
Perturbação - Agente Operador Ec
Perturbação - Agente Proprietário
Perturbação - Area Regional Envol
Perturbação - Características
Alimentação Radial
Atuação SEP
Identificador da Perturbação
Número De Desligamentos Fc
Patamar de Corte de Carga
Vw Dim Caracteristicaperturb
Perturbação - Causa
Perturbação - Data Fim
Perturbação - Data Fim Corte de C
Perturbação - Data Inicio
Medidas (Measures)
# Número de Perturbações
# Número de Cortes de Carga
# Valor da Carga Interrompida - MW
# Valor da Energia Interrompida - MW
# Identificador da Perturbação - Pert
# Tempo de Restabelecimento da Pert
# Tempo de Restabelecimento do Cor
# Tempo de Restabelecimento do Por
# Número de Grupos - Região Elétrica
# Número de Grupos - Região Elétrica
# Número de Grupos - Estado Afetadc
# Número de Grupos - Agente Afetad
# Número de Grupos - Tipo de Rede E
# Num Identperturbacao
Páginas
Filtros
Patamar de Corte d..
Marcas
Automático
Rótulo
Cor
Tamanho
Nível de detalhe
Colunas
Número de Perturb..
Linhas
Nome Do Tipo Do..
Nome Do Tipo Do Equipame..
Linha de Transmissão
FORA REDE OPER
Transformador de 3 Termina..
Seção de Barra
Transformador de 2 Termina..
Unidade Geradora
Reator
Banco de Capacitores
Ramal de Linha
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
Número de Perturbações
Planilha 1
Planilha 2
Planilha 3
Planilha 4
Planilha 5
Planilha 6
Planilha 7

9 marcas 9 linhas por 1 coluna SOMA de Número de Perturbações: 459

Figure 0.5: Disturbances with load cut level major then 99MW by equipment types

3. Disturbances originated in Transmission Lines

The disturbances occurred in SIN are originated, mostly, in transmission lines because they are extensive and, therefore, more exposed to damages. In Figure 0.6 this predominance is visible when comparing the number of disturbances by year and equipment types (the transmission lines are represented by the biggest bars). As in Brazil much of the electricity is generated by hydroelectric plants, which are located far from the load centres, it is necessary an extensive transmission grid to ensure the supply of energy to all parts of the country.

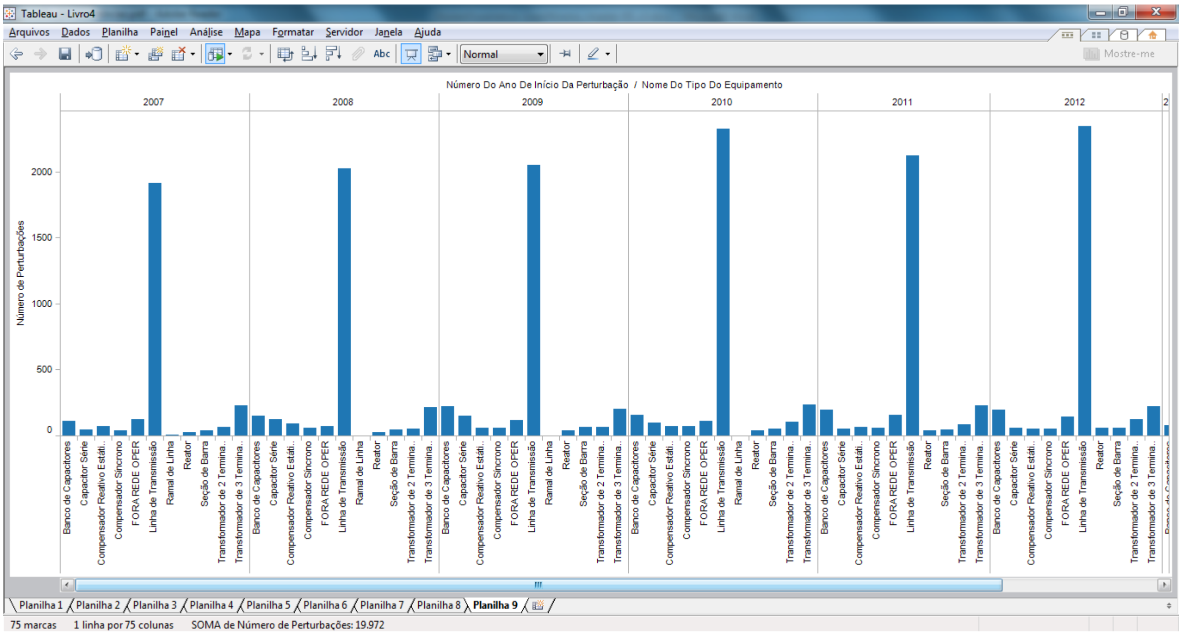


Figure 0.6: Predominance of transmission lines as disturbances source equipment type

The main identified causes of disturbances originated in transmission lines are, in order: natural phenomena; environment issues; human failures; strange objects; accessories and equipments; and protection and control activities. The possible causes are classified by those groups and can be drilled down to specific causes, as illustrated in the following analysis result table in Figure 0.7:

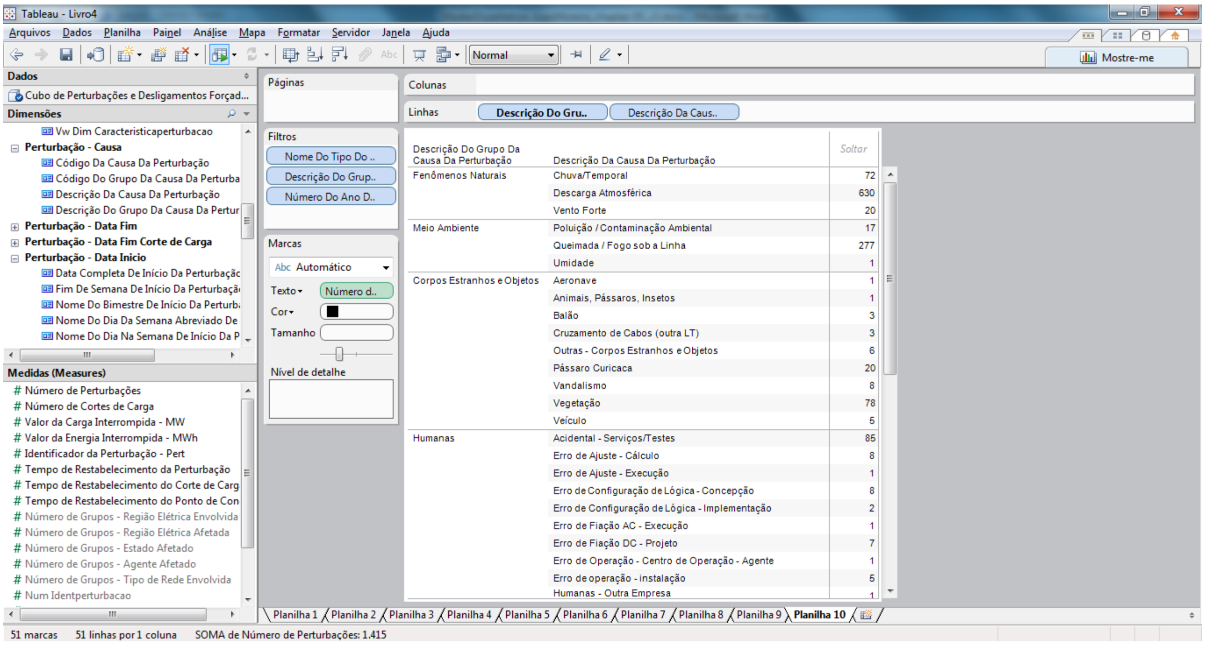


Tableau - Livro4
Arquivos Dados Planilha Painel Análise Mapa Formatar Servidor Janela Ajuda
Normal
Mostre-me
Dados
Cubo de Perturbações e Desligamentos Forçad...
Dimensões
Vw Dim Caracteristicaperturbacao
Perturbação - Causa
Código Da Causa Da Perturbação
Código Do Grupo Da Causa Da Perturba
Descrição Da Causa Da Perturbação
Descrição Do Grupo Da Causa Da Pertur
Perturbação - Data Fim
Perturbação - Data Fim Corte de Carga
Perturbação - Data Início
Data Completa De Início Da Perturbaçã
Fim De Semana De Início Da Perturbaçã
Nome Do Bimestre De Início Da Perturb
Nome Do Dia Da Semana Abreviado De
Nome Do Dia Na Semana De Início Da P
Medidas (Measures)
Número de Perturbações
Número de Cortes de Carga
Valor da Carga Interrompida - MW
Valor da Energia Interrompida - MWh
Identificador da Perturbação - Pert
Tempo de Restabelecimento da Perturbação
Tempo de Restabelecimento do Corte de Carg
Tempo de Restabelecimento do Ponto de Con
Número de Grupos - Região Elétrica Envolvida
Número de Grupos - Região Elétrica Afetada
Número de Grupos - Estado Afetado
Número de Grupos - Agente Afetado
Número de Grupos - Tipo de Rede Envolvida
Num Identperturbacao
Páginas
Filtros
Nome Do Tipo Do ..
Descrição Do Grup..
Número Do Ano D..
Marcas
Automático
Texto
Número d..
Cor
Tamanho
Nível de detalhe
Colunas
Linhas
Descrição Do Gru..
Descrição Da Caus..
Planilha 1 Planilha 2 Planilha 3 Planilha 4 Planilha 5 Planilha 6 Planilha 7 Planilha 8 Planilha 9 Planilha 10
51 marcas 51 linhas por 1 coluna SOMA de Número de Perturbações: 1.415

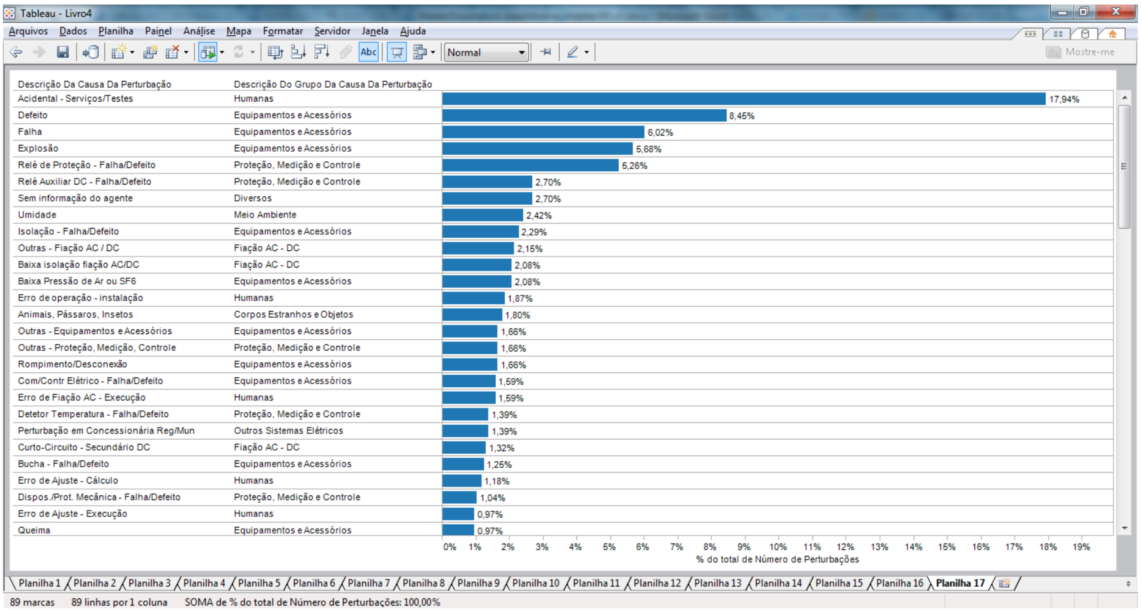
Descrição Do Grupo Da Causa Da Perturbação	Descrição Da Causa Da Perturbação	Soltar
Fenômenos Naturais	Chuva/Temporal	72
	Descarga Atmosférica	630
	Vento Forte	20
Meio Ambiente	Poluição / Contaminação Ambiental	17
	Queimada / Fogo sob a Linha	277
	Umidade	1
Corpos Estranhos e Objetos	Aeronave	1
	Animais, Pássaros, Insetos	1
	Balão	3
	Cruzamento de Cabos (outra LT)	3
	Outras - Corpos Estranhos e Objetos	6
	Pássaro Curicaca	20
	Vandalismo	8
	Vegetação	78
	Veículo	5
Humanas	Acidental - Serviços/Testes	85
	Erro de Ajuste - Cálculo	8
	Erro de Ajuste - Execução	1
	Erro de Configuração de Lógica - Concepção	8
	Erro de Configuração de Lógica - Implementação	2
	Erro de Fiação AC - Execução	1
	Erro de Fiação DC - Projeto	7
	Erro de Operação - Centro de Operação - Agente	1
	Erro de operação - instalação	5
	Humanas - Outra Empresa	1

Figure 0.7: Number of disturbances by cause

The main specific cause of disturbances in natural phenomena is the atmospheric discharges, followed by storms and strong winds. When disturbances are caused by environment issues, the specific causes are fires, followed by pollution. Human failures that cause disturbances in transmission lines are mostly classified as accidental (70.8%). An interesting fact occurs when we filter this analysis by the level cut grater then 99MW and compare the representativeness of the disturbances caused by fires and accidental human failures. They increase considerably from an average (from 2007 to 2013) of 2.4% to 9.6% and 3.1% to 7.4%, respectively. However, independently from the load cut level, the atmospheric discharges keep being the most common cause of disturbances when they are originated in transmission lines.

4. Disturbances originated in Power Transformers

As cited before, the second most common equipment source of disturbances is the power transformer, which represents 14.1% of the total. In the disturbances with load cut level greater then 99MW this tax increases to 26.2%. As these devices are usually located on the border of the system to supply the consumers (by distribution companies), disturbances involving power transformers typically cause load interruption. Figure 0.8 illustrates the main causes of disturbances originated in power transformers, mostly accidental human failures (17.9%) followed by defects (8.4%) and failures (6.0%).



Descrição Da Causa Da Perturbação	Descrição Do Grupo Da Causa Da Perturbação	
Acidental - Serviços/Testes	Humanas	17,94%
Defeito	Equipamentos e Acessórios	8,45%
Falha	Equipamentos e Acessórios	6,02%
Explosão	Equipamentos e Acessórios	5,68%
Relé de Proteção - Falha/Defeito	Proteção, Medição e Controle	5,26%
Relé Auxiliar DC - Falha/Defeito	Proteção, Medição e Controle	2,70%
Sem informação do agente	Diversos	2,70%
Umidade	Meio Ambiente	2,42%
Isolação - Falha/Defeito	Equipamentos e Acessórios	2,29%
Outras - Fiação AC / DC	Fiação AC - DC	2,15%
Baixa isolação fiação AC/DC	Fiação AC - DC	2,08%
Baixa Pressão de Ar ou SF6	Equipamentos e Acessórios	2,08%
Erro de operação - instalação	Humanas	1,87%
Animais, Pássaros, Insetos	Corpos Estranhos e Objetos	1,80%
Outras - Equipamentos e Acessórios	Equipamentos e Acessórios	1,66%
Outras - Proteção, Medição, Controle	Proteção, Medição e Controle	1,66%
Rompimento/Desconexão	Equipamentos e Acessórios	1,66%
Com/Contr Elétrico - Falha/Defeito	Equipamentos e Acessórios	1,59%
Erro de Fiação AC - Execução	Humanas	1,59%
Detetor Temperatura - Falha/Defeito	Proteção, Medição e Controle	1,39%
Perturbação em Concessionária Reg/Mun	Outros Sistemas Elétricos	1,39%
Curto-Circuito - Secundário DC	Fiação AC - DC	1,32%
Bucha - Falha/Defeito	Equipamentos e Acessórios	1,25%
Erro de Ajuste - Cálculo	Humanas	1,18%
Dispos./Prot. Mecânica - Falha/Defeito	Proteção, Medição e Controle	1,04%
Erro de Ajuste - Execução	Humanas	0,97%
Queima	Equipamentos e Acessórios	0,97%

89 marcas 89 linhas por 1 coluna SOMA de % do total de Número de Perturbações: 100,00%

Figure 0.8: Main causes of disturbances originated in power transformers

5. Disturbances caused by atmospheric discharges

As seen in previews analysis over transmission lines, the atmospheric discharge is the main disturbances cause, they represent 29% of the total from 2007 to 2013. One relevant characteristic is that there is a direction relation to climatic seasonality, once it is possible to identify the periods of the major incidence of disturbances caused by atmospheric discharges. Figure 0.9 bellow shows that in the same periods of severe rains and electrical storms in Brazil, i.e. from January to March, the disturbances caused by lightening are more often than in other months. In addition, whilst in months with high level rains there are more disturbances caused by atmospheric discharges, in months that are lower levels there are less disturbances caused by atmospheric discharges, such as July and August.

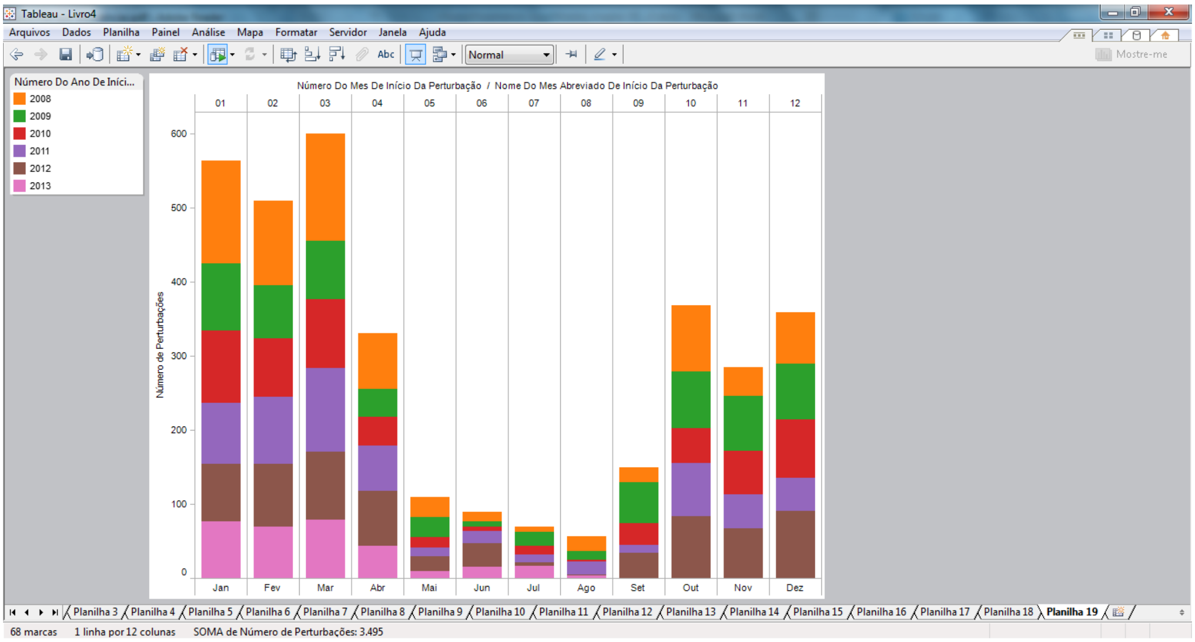
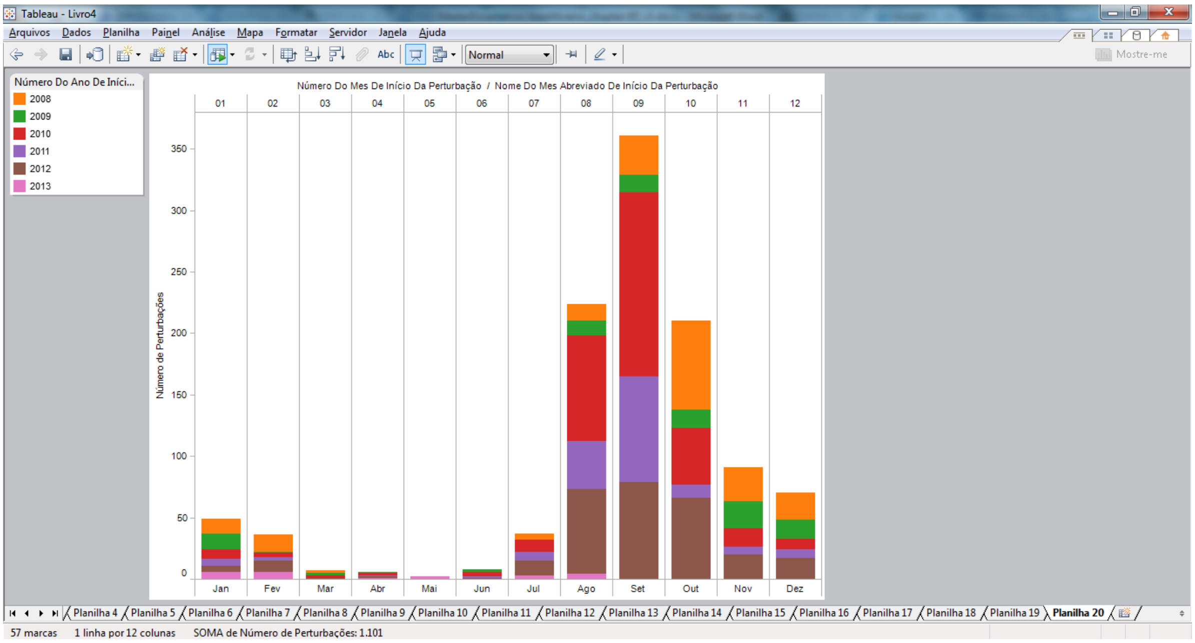


Figure 0.9: Disturbances caused by atmospheric discharges by month

6. Disturbances caused by fires

With similar characteristics of atmospheric discharges when comparing to climate seasonality, the fires are also one of the main causes of disturbances, occurring mainly in dry periods. It coincides with periods of high energy transfers between Brazilian north with mid-west and north-south with north-northeast. Thereafter, the disturbances occurred in those interconnections commonly entail in severe consequences for the affected regions.

Moreover, the fires have a direct relation to the harvest period in the areas of sugarcane cultivation (August to October), occurring frequently in northeast region, where the sugarcane stills burned before manual picking. The graph depict in Figure 0.10 demonstrate both situations.



57 marcas 1 linha por 12 colunas SOMA de Número de Perturbações: 1.101

Figure 0.10: Disturbances caused by atmospheric discharges by month

7. Disturbances caused by human failures

The disturbances caused by human failures are fairly representative within the universe of occurrences in transmission lines and power transformers, often causing severe impact to the system and for energy consumers. Its origin is external to equipment and is associated with maintenance teams and electrical system operation. Therefore, it is believed that there are ways to minimize these events. The vast majority incidents involving human failures (68.7%) occurs during the execution of interventions and equipment testing in SIN, caused by companies' maintenance teams. In general, with expansion increasing of the electrical system, the main causes of disturbances originating from human failures are: difficulty of skilled human resources for maintenance performance, low skills of the teams involved in the maintenance and operation of the system; increased outsourcing services by companies from the development and implementation of projects of new facilities. Figure 0.11 bellow illustrates this analysis.

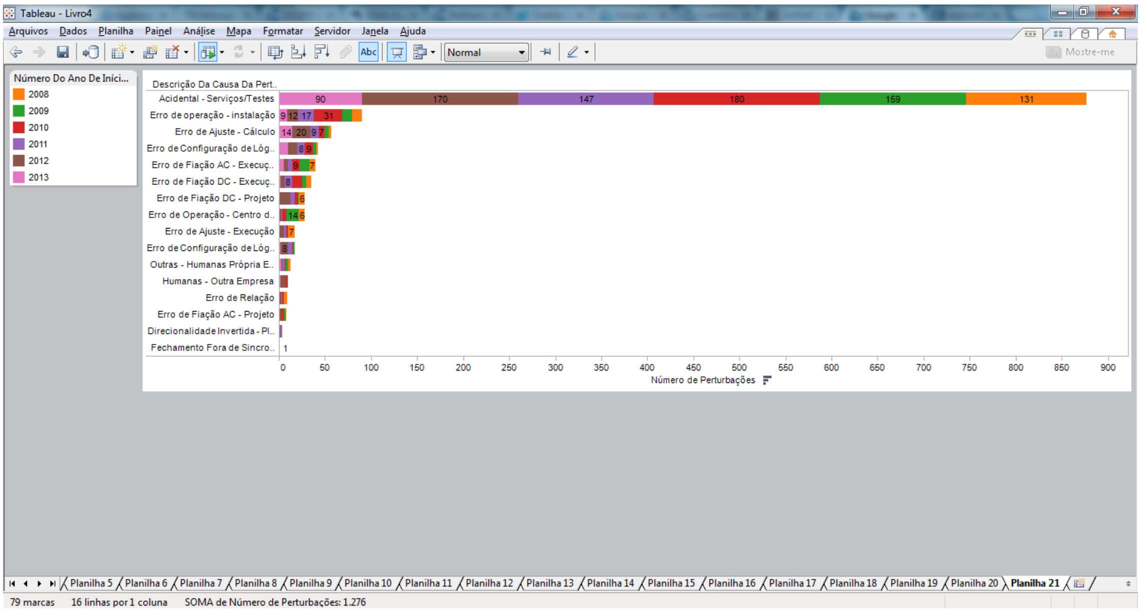


Figure 0.11: Disturbances by the most common human failures

8. Analysis Conclusions

To reduce the number of disturbance occurrences and minimize impacts to the electrical system and consumers is crucial. Based on the presented analysis, several actions can be taken by ONS and the involved companies. Considering the disturbances caused by atmospheric discharges, it is possible to identify the specific equipment with high incidence of events caused by lightning by working aside company owners of such equipment. The investment in tools to advance climate prediction is a necessity to support operation electro-energetic planning. Another action would be the upgrade of atmospheric discharges detection used by ONS operation centres, including the investment in transmission lines geo-referencing systems, with the objective of preparing the system to support possible contingences and minimize the impact.

Considering the disturbances caused by fires, an action would be to intensify the campaigns of fires prevention in the areas nearby the most affected transmission lines. Also, stimulate the work together with the company owners of the transmission lines most frequently affected to avoid new occurrences. Moreover, to improve the detection of fires and hotspots tools used in ONS Operation Centres, including investment with the companies in geo-referencing of the transmission lines in order to prepare the system to support the possible contingencies.

Regarding human failures, actions related to quality of work should be taken, such as the requirement for certification of maintenance and operation teams of operating companies. In addition, intensify surveillance in operating companies that perform maintenance on equipment of their own. For last, to prepare the executive planning schedules of interventions performed by operating companies with a focus on reducing the probability of human failures during execution of services is important.

APPENDIX C – EXPERIMENTAL ENVIRONMENT AND ETL DEVELOPMENT

There were two computational environments used in this study case. The first one for the Textual ETL Framework (Java, Postgres SQL, TreeTagger) was a Virtual Machine with Windows XP operational system, 1024 MB RAM, 1 core (1Ghz), 32 bits. The second for modeling (EA/OntoUML, OLED) and implementing (.NET 4.0 and SQL Server BI 2012).

As cited previously, a web crawler was built, it was used the .NET environment and a WebForm project (C# language) using the HTTPRequest class, so programmatically we could access the Clippings website and download each news article text from January 2011 to March 2013, resulting in a total of 20,306 text (.txt) document files (around 68MB). Thereafter, the Textual ETL Framework (ALMEIDA e SILVA, 2009) with IR functionalities was applied in the set of documents extracted from the Clippings. It was responsible for extracting, transforming and loading the terminology contained in the clippings to its relational DB structure in Postgres DBMS. The total execution of this step for all the news articles was around 120 hours of processing. This process had generated a relational DB, which represents the terminological treatment of the terms, containing records of the documents processed, terms, synonyms, stop words, tokens and stems. Below is listed each table with its description and rows count resulted from the process:

Table 2.9: Result data tables and rows count from Textual ETL terminology extraction task

Table	Description	Rows
TB_ARQUIVO	Stores text files information (name, physical path, published date, etc).	14,250
TB_TERMO	Stores different terms found in the files.	117,186
TB_RADICAL	Stores radicals (stems) of the terms.	82,513

TB_CLASSIF_TERMO	Stores morphological terms classifications.	17
TB_SINONIMO	Stores terms synonymous found in files.	17,064
TB_STOPWORD	Stores the stop words to be avoided.	331
REL_ARQUIVO_STOPWORD_CLASSIF	Stores all relationships between files, terms and stop words.	1,335,068
REL_ARQUIVO_TERMO_CLASSIF	Stores all relationships between files, terms and their classifications.	1,931,958
REL_TERMO_SINONIMO	Stores all relationships between terms and their synonymous.	49,956

Subsequently to the Textual ETL terminology extraction task, a manually activity was made, where each table was exported from the Postgres SQL DBMS to semi-structured files (.csv), being saved in a shared folder. A conventional ETL was applied to bring the data contained in those files to the Textual ODS implemented in the SQL Server DBMS. This process was implemented in SQL Server Integration Services (SSIS) ETL tool and is depict in Figure 0.12. At first the ODS_Textual DB is clear (using full load technique), then each ".csv" file has its data is extracted and loaded. Finally, a transformation task is executed to update the publication date and press company from each news article file. This last task was done in this phase but, for performance enhancement, the better phase to extract this information would be during the Textual ETL framework execution when it writes each text file in "TB_ARQUIVO" table. We did not implement like that because we did not want to change its code. Moreover this change in the framework would make it dependent on this application, changing one of its basic characteristics (to be domain independent).

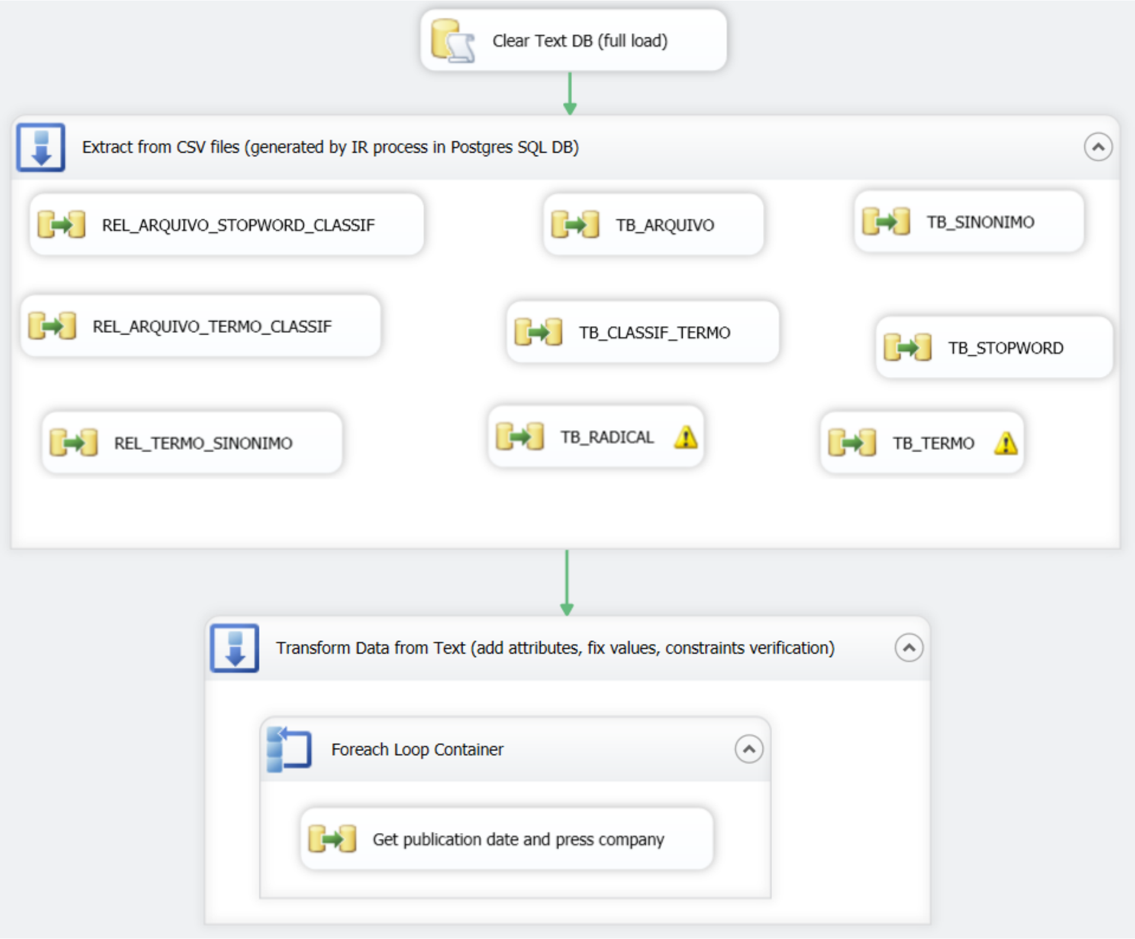

Clear Text DB (full load)
Extract from CSV files (generated by IR process in Postgres SQL DB)
REL_ARQUIVO_STOPWORD_CLASSIF
TB_ARQUIVO
TB_SINONIMO
REL_ARQUIVO_TERMO_CLASSIF
TB_CLASSIF_TERMO
TB_STOPWORD
REL_TERMO_SINONIMO
TB_RADICAL
TB_TERMO
Transform Data from Text (add attributes, fix values, constraints verification)
Foreach Loop Container
Get publication date and press company

Figure 0.12: ETL process to load the Textual ODS

Parallel to the previews ETL step, a conventional ETL was also built to extract the electric domain entities, such as the disturbances characteristics, source transmission lines and other equipments (participants), companies, etc, i.e. all necessary information from the existent Disturbance Data Mart and BDT. Figure 0.13 illustrates this process, which begins with a full load of the ODS DB (also in SQL Server DBMS). Then, all necessary entities mapped during the MD task are loaded in the ODS, followed by the acquisition of the existent fact from the Disturbance Data Mart, which brings all relations between the entities loaded. Note that in this phase we did not yet load the dimensions which we are going to use in the temporal relation multidimensional schema, only the entities that will form them.

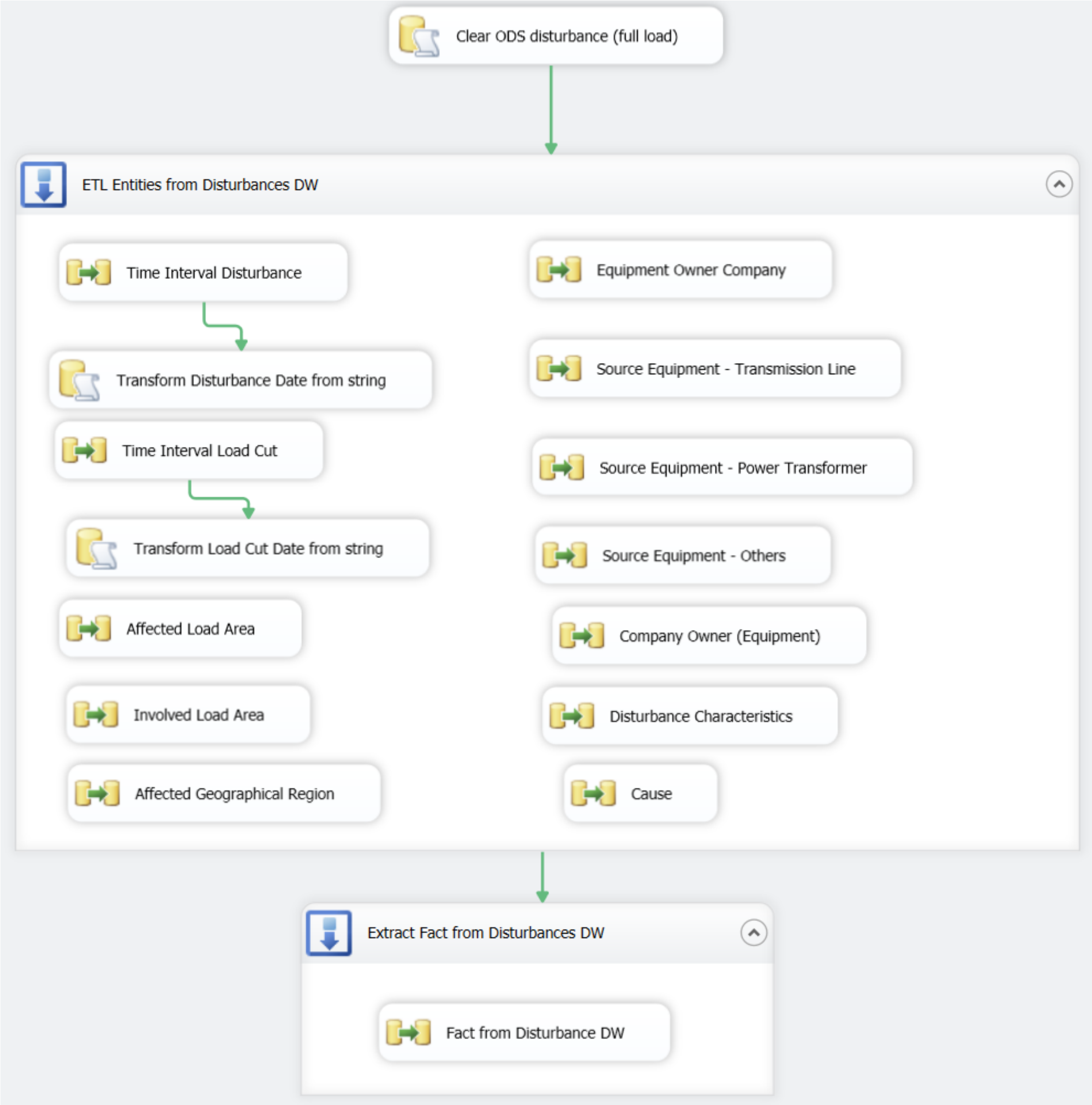

Clear ODS disturbance (full load)
ETL Entities from Disturbances DW
Time Interval Disturbance
Transform Disturbance Date from string
Time Interval Load Cut
Transform Load Cut Date from string
Affected Load Area
Involved Load Area
Affected Geographical Region
Equipment Owner Company
Source Equipment - Transmission Line
Source Equipment - Power Transformer
Source Equipment - Others
Company Owner (Equipment)
Disturbance Characteristics
Cause
Extract Fact from Disturbances DW
Fact from Disturbance DW

Figure 0.13: ETL process to load the ODS with the domain entities

After finishing the load of the textual ODS and conventional ODS (with disturbances structured data), it is the moment to run the Linkage process to load the multidimensional schema. At first, as in previews ETL packages, a full load occurs where all dimensions and the fact table are truncated. Afterwards, two parallel processes are executed: one for the construction of the dimension from the structured data and another from the unstructured data. As seen during the MD task, each of those dimensions represents an event. From the structured data, it is built the disturbance dimension and the hierarchies representing the participations with their relations. Figure 0.14 below depicts this process.

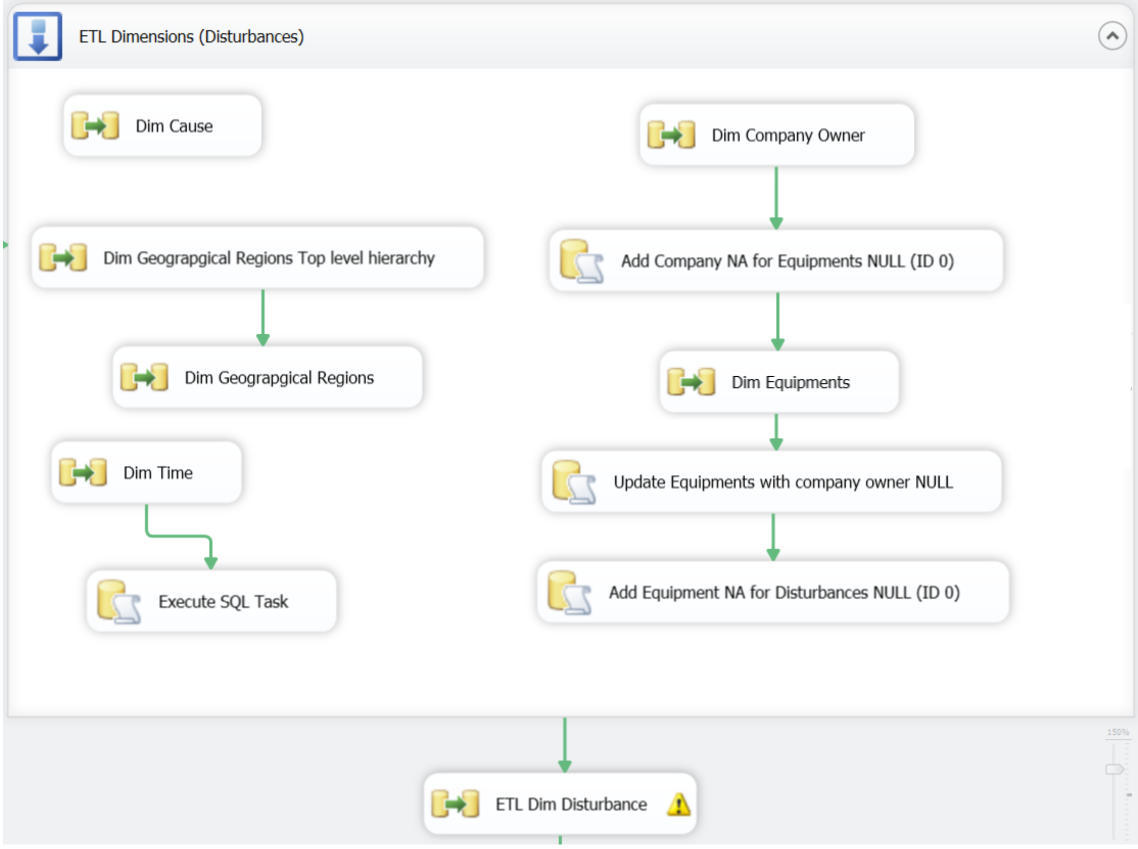

ETL Dimensions (Disturbances)
Dim Cause
Dim Company Owner
Dim Geograpgical Regions Top level hierarchy
Add Company NA for Equipments NULL (ID 0)
Dim Geograpgical Regions
Dim Equipments
Dim Time
Update Equipments with company owner NULL
Execute SQL Task
Add Equipment NA for Disturbances NULL (ID 0)
ETL Dim Disturbance

Figure 0.14: ETL process to create disturbance dimension and its hierarchies

From the textual information sources (the clippings), the ETL to build the news article terms publication dimension and its hierarchies is presented in Figure 0.15. At first all top level hierarchies are processed, following to the next levels in the hierarchies by the functional dependencies presented in the DB relationships. For example, the term classification hierarchy – which presents the facets, categories, terms and synonymous – followed the dependency existent from facet to category, from category to term, from term to synonymous. This last hierarchy level is an addition to the pattern presented in the fact term occurrences, where we grouped the terms by their synonymous. Moreover, the facet level was chosen to represent a top level domain classification, such as “Terms in Official Glossary” to evidence their semantic to the business. In the category level it was chosen to load the morphological term classes presented in the “TB_CLASSIF_TERMO” original table. Examples are “verbs”, “adjectives”, “preposition” and “adverbs”. This choice was made to enable the analysis capability by the terms grammar classifications.

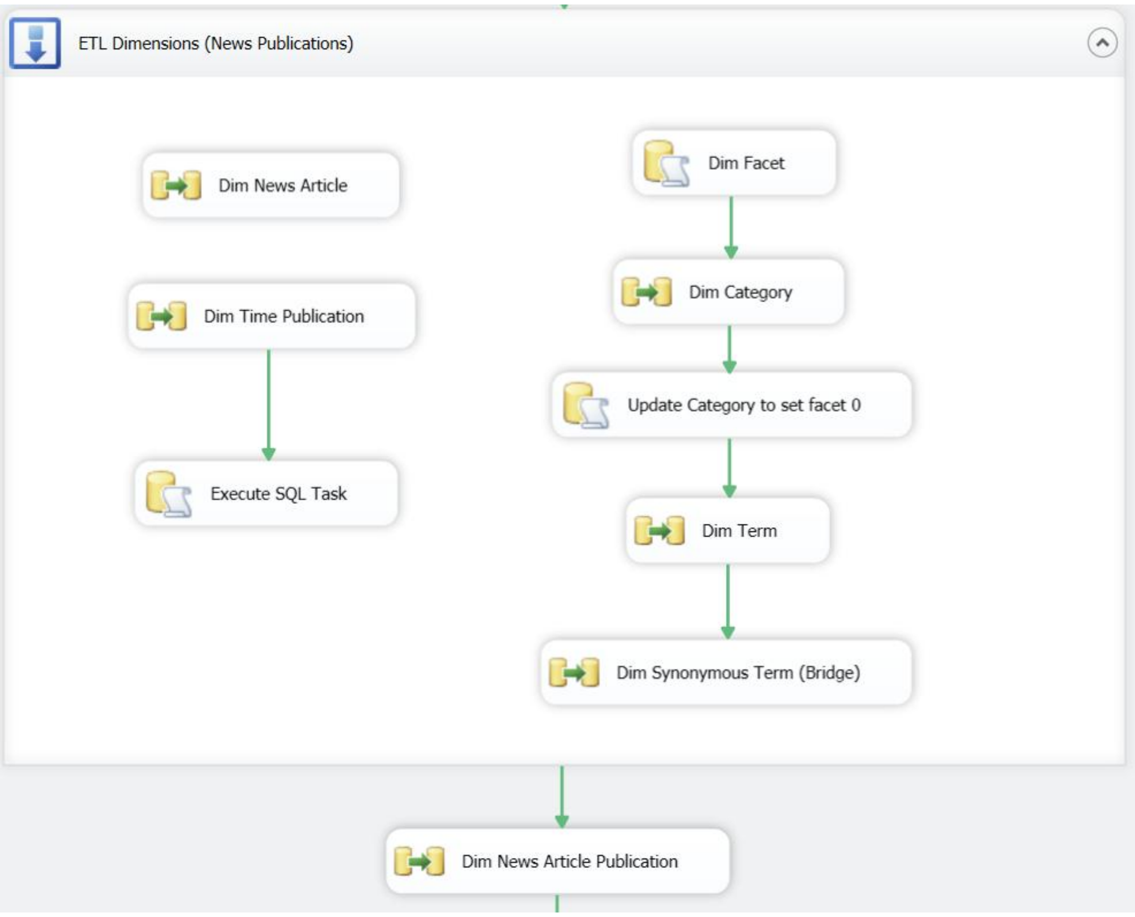

ETL Dimensions (News Publications)
Dim News Article
Dim Time Publication
Execute SQL Task
Dim Facet
Dim Category
Update Category to set facet 0
Dim Term
Dim Synonymous Term (Bridge)
Dim News Article Publication

Figure 0.15: ETL process to create news article publication dimension and its hierarchies

After those ETL parts are executed to construct both dimensions (and their hierarchies) it's the moment to build the fact. This is the fundamental step in the process because it's where the Linkage is built, relating the structured and unstructured universes by the temporal relation fact pattern resulted from the rule 3 in MD. A SQL query is responsible to relate them, following the constraints from the temporal relations (Allen's operators) represented in the domain ontology. In this case, the "before" relation was set from a disturbance to a news publication, i.e. a news article publication can be related to a disturbance if it occurred before. In the case of "before" relationship it is important to set a delta time to relate the events. That is, a disturbance can be mentioned (or related somehow) on a news article if the news is published until X days after the disturbance. At first this X variable should be set based in a guess from a domain expert. After running several variances of this value it is possible to set a more specific value. In our experiment, at the beginning, we chose the value of 30 days. The initial analysis evidenced the increasing of the number of publications after a severe disturbance occurred (which caused a long and extensive load cut). The publications number decreases on subsequent days, reaching the regular publications average after ten days (in average). So, based in this information we could refine this delta time for the “before” relationship. An important issue about this choice is that: the higher this value is, the number of rows in the fact grows exponentially. Therefore, the SQL query for joining the disturbance dimension to the news article publication dimension by the inner join clause to load the fact is illustrated as following:

```
select top publ.id_newsarticlepublication, disturb.id_disturbance,
       DATEDIFF(day, max(publ.date), max(disturb.date)) num_difdays,
       max(f.val_cargainterrompida)   val_cargainterrompida,   max(f.num_cortescarga)
num_cortescarga,                          max(f.val_temporestabelecimentoperturbacao)
val_temporestabelecimentoperturbacao
from
(
       select tp.date, tp.id_timepublication, np.id_newsarticlepublication
       from DimNewsArticlePublication np inner join
              DimTimePublication tp on np.id_timepublication = tp.id_timepublication
) publ inner join
(
       select te.date, te.id_time_end, te.date_str, d.id_disturbance
       from DimDisturbance d inner join
              DimTimeEnd te on d.id_time_end = te.id_time_end
) disturb on
(DATEDIFF(day, publ.date, disturb.date) <10
       and DATEDIFF(day, publ.date, disturb.date) >0)
```

Notice the highlighted part of the query above, where the relationship is defined to be less than ten days. Also, it is necessary to set the lower boundary for the relationship to zero. It represents that a news publication from a prior day of a disturbance should not be considered in the relation, so the delta time cited should be from zero to ten. With this value set we executed the fact load and it resulted in a huge volume of 120,932,538 rows. This is the number of rows of the relation of the disturbances occurred ten days after the terms published by the media from January 2011 to February 2013.

At the end of this process we could load all data in the relational multidimensional schema. Although it is the sufficient structure to make the required analysis, by SQL queries, there is a last process to build and deploy the OLAP cube, described in next section.

The ETL built present some limitations and miss usual techniques. At first, the ETL built was made without several concerns about ETL design issues and implementation – such as surrogate keys treatment or indexing management – due to the scope of this work. Moreover, to simplify the study case we choose to set a 1:1 relation between terms and categories. This leads to the restriction that a term is only classified as one type of grammar category, i.e. a term is a verb or an adjective or an adverb and so on.

APPENDIX D – DATA CUBE DEVELOPMENT AND OLAP ANALYSES

The preparation of disturbances and clippings *data cube* was based in the data source view of the MD schema. At first the `Disturbance` *Dimension* was manipulated to represent its *Hierarchies*, such as the `date/time begin and end`, the `grouping causes`, the `affected regional and load areas`, the `involved load areas`, the `source transmission lines` and `source power transformers`.

Thereafter, the dimension `article news terms published` was also defined for the data cube, with basically two hierarchies: `publication date/time` and `terms classification (facet, category, term and synonymous)`. Other configurations were made in the cube, such as measures to be used and dimension use for them. The `disturbance Load Cut value` (in MW) is a semiadditive measure because it can not be aggregated by the `News Article Publication` dimension, only by the `Disturbance` dimension. The same behaviour is expected by the `number of Load Cuts` and the `restoration time value`. The measure `difference time between the events` is also a semiadditive measure because it retrieves the highest value for all children members. It was set as maximum aggregation function, so when exploring a set of disturbances with a set of news publications the maximum difference days between then will be presented to user. Those configurations were made in SSAS by particular capabilities of the tool, such as the semiadditive behavior definition wizard and the creation of calculated measures with MDX language.

As cited, the presentation layer used in this study case was an OLAP tool, the Tableau software (a top leader OLAP tool pointed by Gartner). By configuring the access to the cube, with both dimensions, their hierarchies and the measures, its structure is available for the business analyst to explore the data in the left layer. Figure 0.16 illustrates this analytical environment, following the common patterns of OLAP tools.

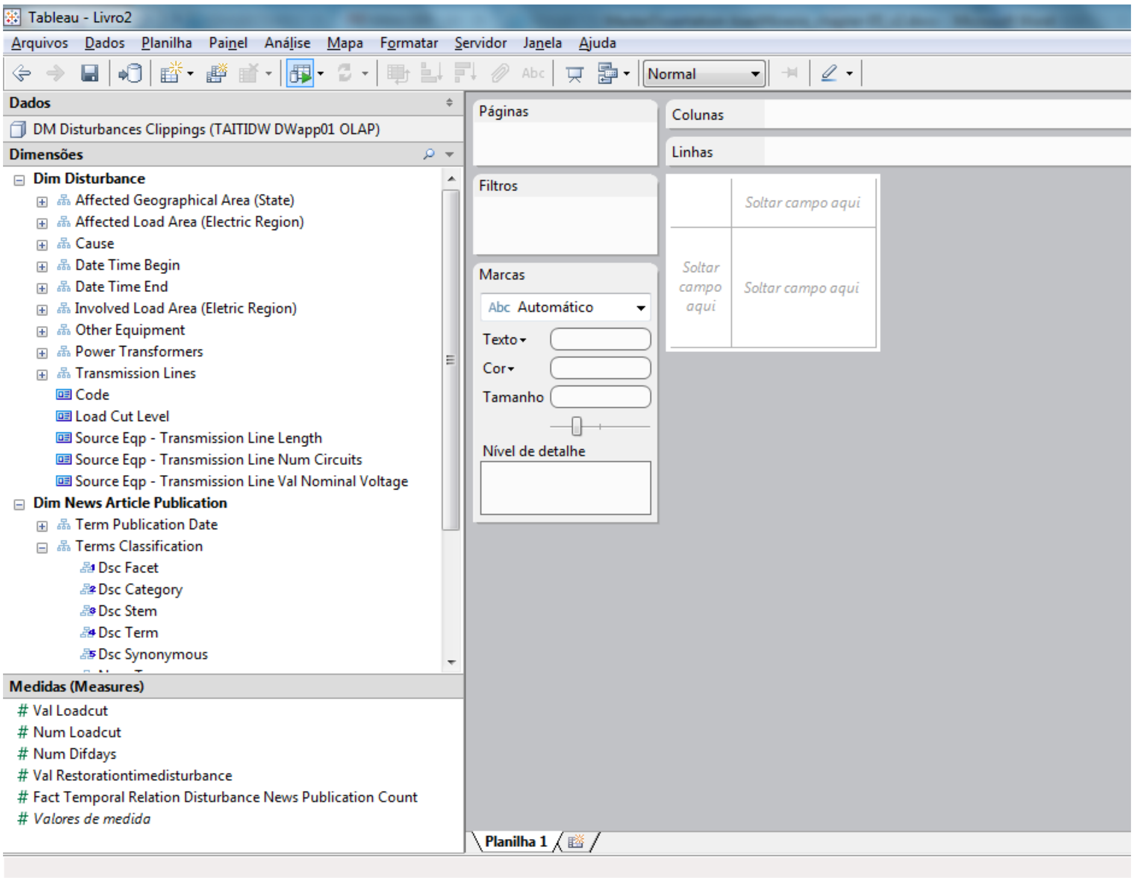


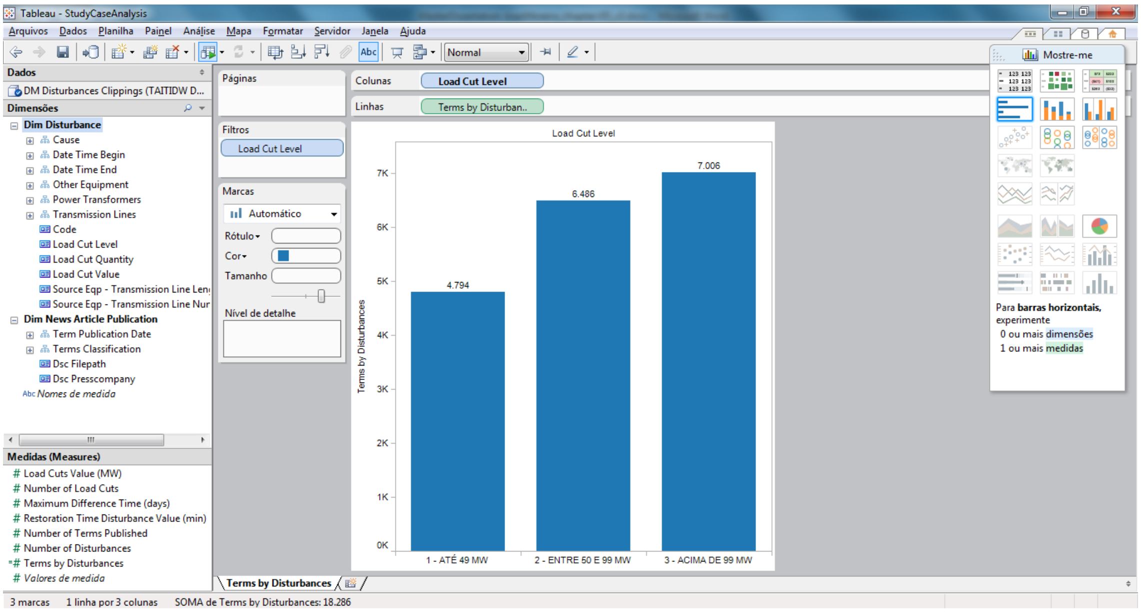
Tableau - Livro2
Arquivos Dados Planilha Painel Análise Mapa Formatar Servidor Janela Ajuda
Normal
Dados
DM Disturbances Clippings (TAITIDW DWapp01 OLAP)
Dimensões
Dim Disturbance
Affected Geographical Area (State)
Affected Load Area (Electric Region)
Cause
Date Time Begin
Date Time End
Involved Load Area (Eletric Region)
Other Equipment
Power Transformers
Transmission Lines
Code
Load Cut Level
Source Eqp - Transmission Line Length
Source Eqp - Transmission Line Num Circuits
Source Eqp - Transmission Line Val Nominal Voltage
Dim News Article Publication
Term Publication Date
Terms Classification
Dsc Facet
Dsc Category
Dsc Stem
Dsc Term
Dsc Synonymous
Medidas (Measures)
# Val Loadcut
# Num Loadcut
# Num Difdays
# Val Restorationtimedisturbance
# Fact Temporal Relation Disturbance News Publication Count
# Valores de medida
Páginas
Colunas
Linhas
Filtros
Soltar campo aqui
Soltar campo aqui
Soltar campo aqui
Marcas
Abc Automático
Texto
Cor
Tamanho
Nível de detalhe
Planilha 1

Figure 0.16: Tableau software connected to the Disturbances Clippings cube

Some representative analysis in energy supply security domain and the impact of blackouts in the ONS corporate image could be executed by the integrated analysis environment. It is important to remember that, as it was defined in the domain ontology, in this analytical solution we set that the disturbances occurs before (until ten days) of the terms occurrences in news publications. Based in this assumption, a set of analysis examples was made and are presented as follows.

1. Number of terms published in News

By selecting the Total Number of Disturbances (36,778) and the Total Number of Terms occurrences (210,373,632), the calculated measure Terms by Disturbances presents the average of 5,720 term occurrences per disturbance. When navigating through disturbance dimension selecting Load Cut Level attribute and analyze the Terms by Disturbances measure it is possible to see a direct relation between the severity of the disturbances and the number of news. Figure 0.17 illustrates this analysis: as most grave are the disturbances, more terms are published. When load cut level is minor than 49MW the average is 4,794 terms/disturbances; between 50MW and 99MW is 6,486 terms/disturbances; greater then 99MW is 7,006 terms/disturbances.



Colunas: Load Cut Level
Linhas: Terms by Disturban..
Filtros: Load Cut Level
Terms by Disturbances
3 marcas 1 linha por 3 colunas SOMA de Terms by Disturbances: 18.286

Figure 0.17: Number of terms published by load cut level

The most common term encountered in the news is energy (“energia”). It is followed by agreement (“acordo”), government (“governo”), plan (“plano”), capacity (“capacidade”), hardiness (“robustez”) and potency (“potencia”). Blackout (“apagão”) term is in 27th position in the most published. An interesting analysis when analyzing the Terms by Disturbances measure by the load cut level, when it is greater than 99MW there is an increase in the percentage of the blackout (“apagão”) term occurrence in the published news. When it is lower than 49MW the tax decreases substantially, as illustrated in Figure 0.18. This type of analysis takes into account the synonymous of the terms, which is considered in the result.

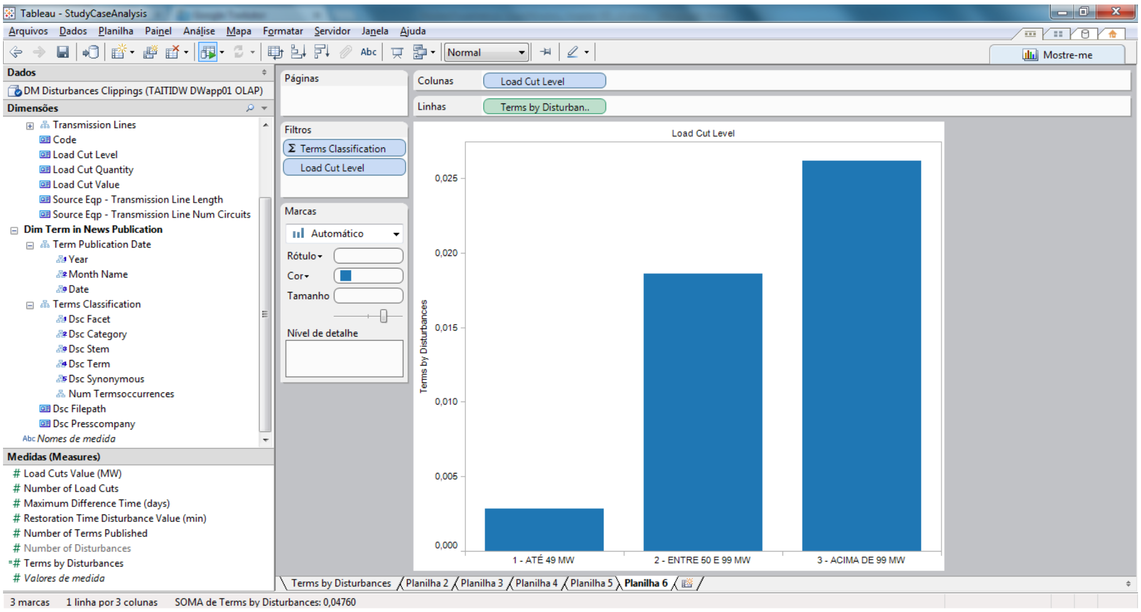


Tableau - StudyCaseAnalysis

Colunas: Load Cut Level

Linhas: Terms by Disturban..

Filtros: Terms Classification; Load Cut Level

Marcas: Automático

3 marcas 1 linha por 3 colunas SOMA de Terms by Disturbances: 0,04760

Figure 0.18: Number of terms blackout published by load cut level

2. Press companies most common terms

The Press Companies that most publish terms in their news about the electrical sector is presented in Figure 0.19. In the first place is “Valor Econômico” with almost the double of the terms from the second, “O Estado de S. Paulo”. It is followed by “O Globo”, “Jornal do Comercio”, “Brasil Econômico” and “Folha de S. Paulo”.

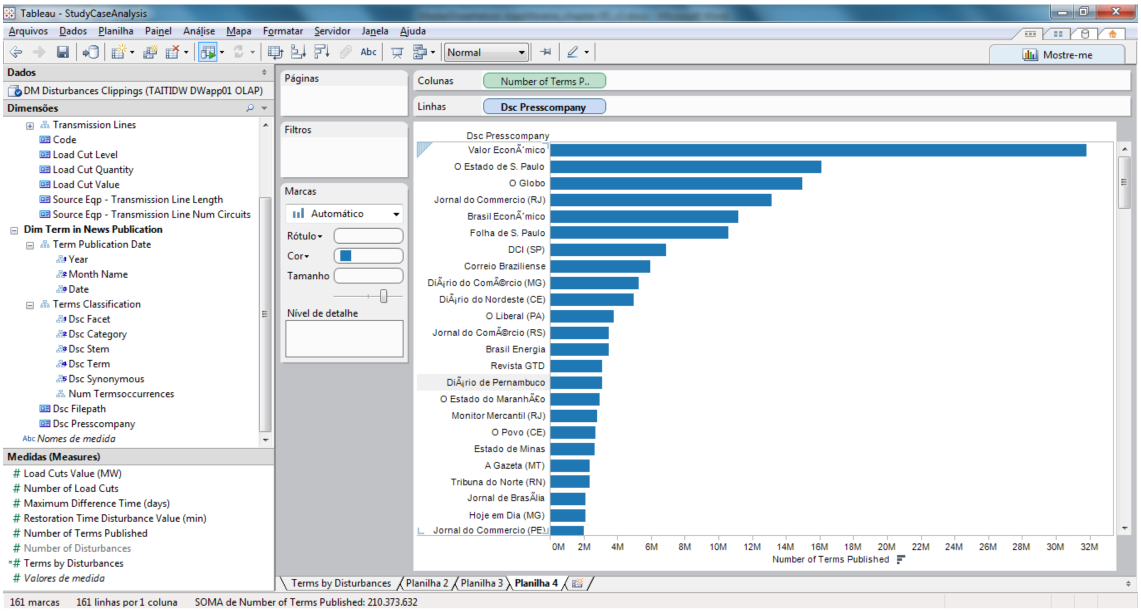


Tableau - StudyCaseAnalysis

Colunas: Number of Terms P..

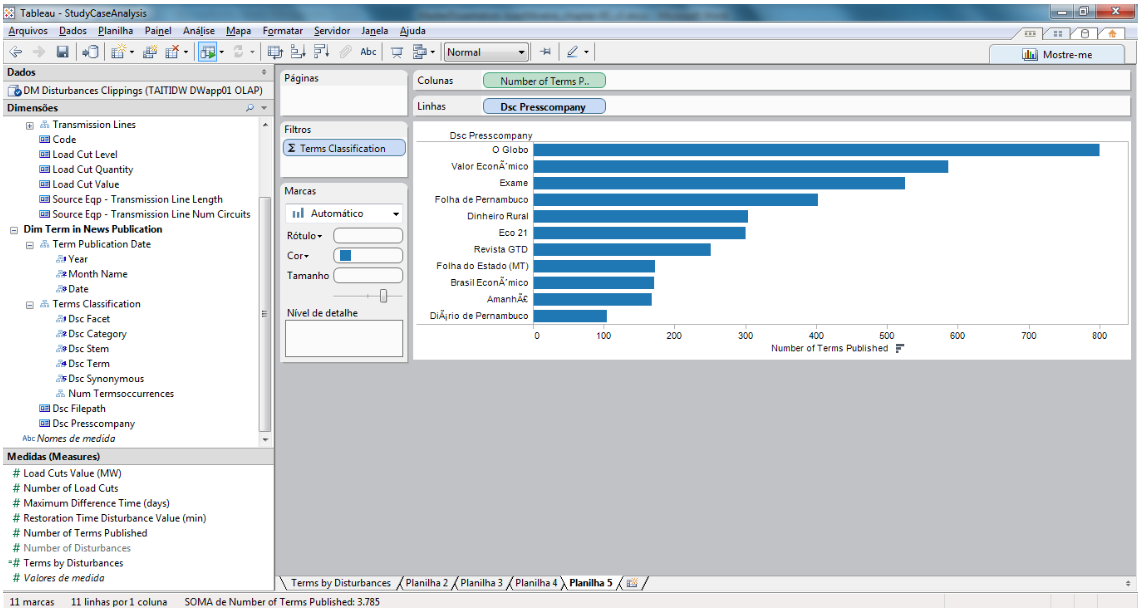
Linhas: Dsc Presscompany

Marcas: Automático

161 marcas 161 linhas por 1 coluna SOMA de Number of Terms Published: 210.373.632

Figure 0.19: Press Companies with more terms published about the electrical sector

Any term can be filtered and analyzed by the Press Company publications, for example the term fire (“fogo”) was much more used by “O Globo” than by “Valor Econômico”, followed by “Exame” and “Folha de Pernambuco” (Figure 0.20):

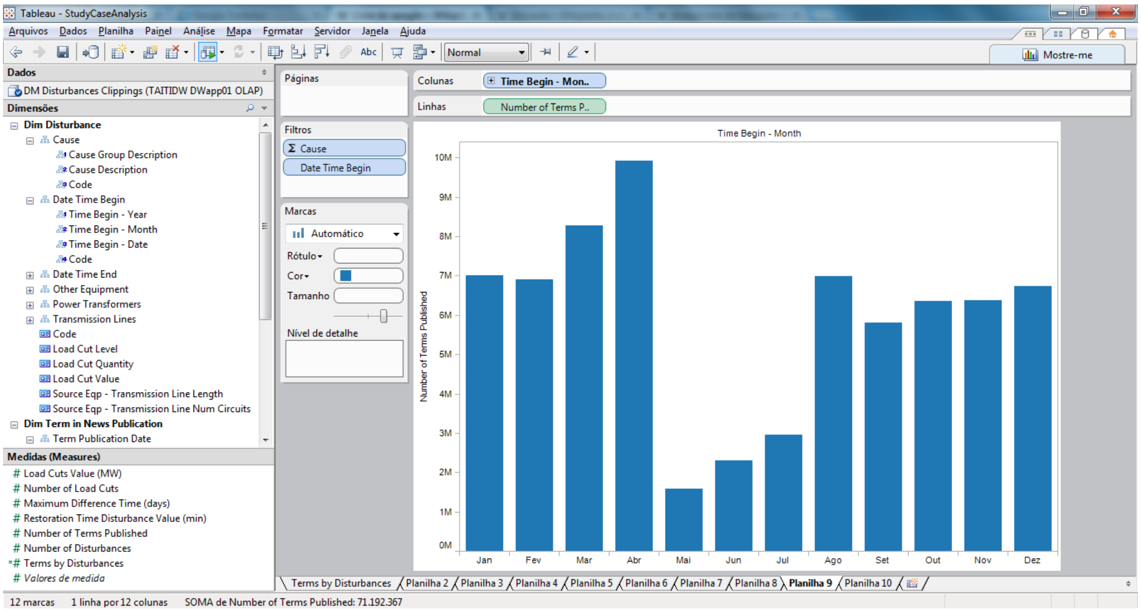


11 marcas 11 linhas por 1 coluna SOMA de Number of Terms Published: 3.785

Figure 0.20: Press Companies with more “fire” term occurrences published

3. News publications by Disturbance Time

When analyzing the Number of Terms Published by month in 2011 it is visible a significant variance of terms published after the disturbances occurred in February, they increase considerably in March and April, then, it decreases, as illustrated in Figure 0.21. This can be explained by the enormous blackout that happened in the electrical network in northeast February 3rd when a problem in the protection components occurred in Luiz Gonzaga substation (in Pernambuco), which caused the shutdown of a transmission line.



Colunas: Time Begin - Mon..
Linhas: Number of Terms P..
Filtros: Cause; Date Time Begin
Marcas: Automático

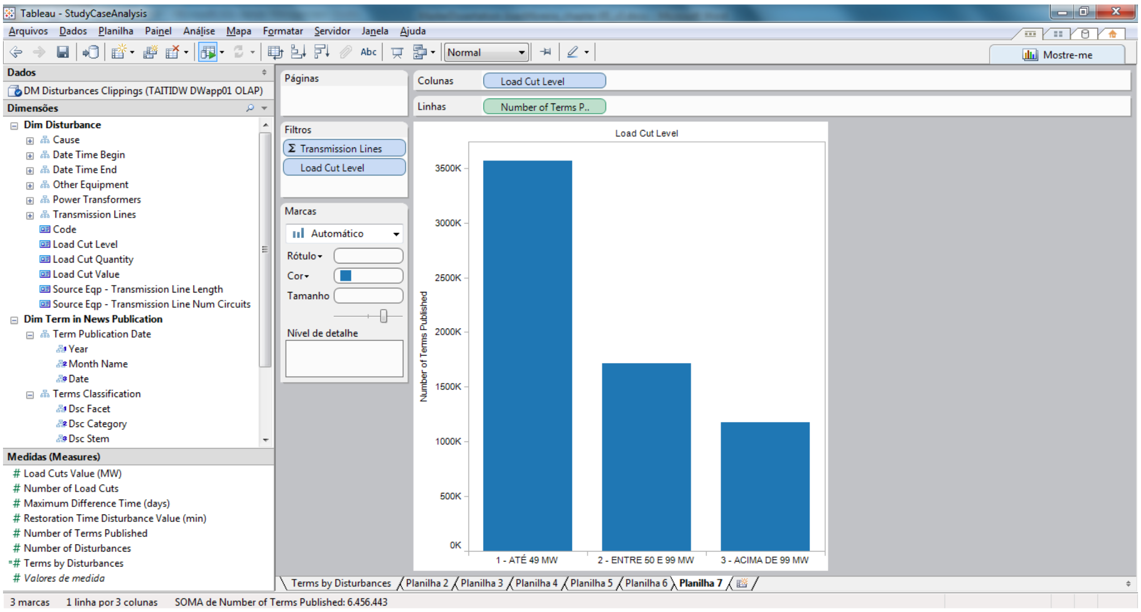
12 marcas 1 linha por 12 colunas SOMA de Number of Terms Published: 71.192.367

Figure 0.21: Average of terms published by disturbances

When analyzing in time perspectives the Maximum Difference Time in days measurement gives a weight of how the terms can be published about the disturbances related.

4. News publications and disturbances originated in transmission lines

A curious fact occurs when the disturbances are originated in transmission lines, the number of terms published decreases as the load cut level raises (Figure 0.22).



Colunas: Load Cut Level
Linhas: Number of Terms P..
Filtros: Transmission Lines; Load Cut Level
Marcas: Automático

3 marcas 1 linha por 3 colunas SOMA de Number of Terms Published: 6.456.443

Figure 0.22: Number of terms published by Load Cut Level when originated in Transmission Lines

5. Disturbances Causes and Terms Publications

The number of terms published in news by disturbance causes has equivalence to the number of disturbances by causes, presenting a curious fact that there is almost none terms published about disturbances caused by “Other Electrical Systems”. Figure 0.23 evidenced this comparison. The top disturbance cause is the “Natural Phenomena”.

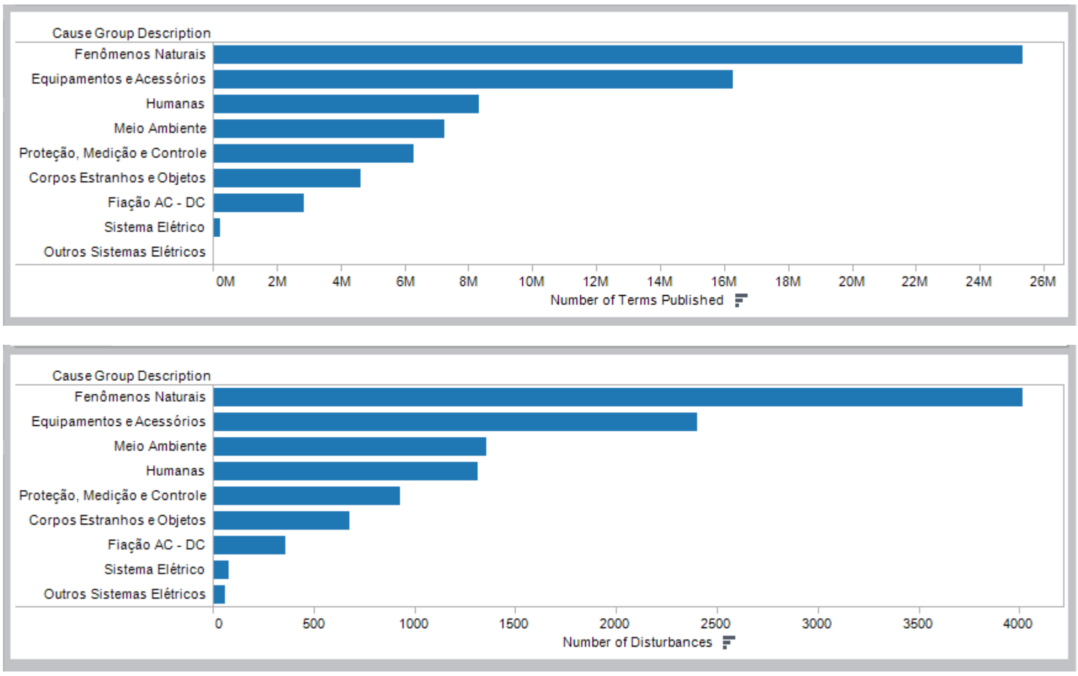
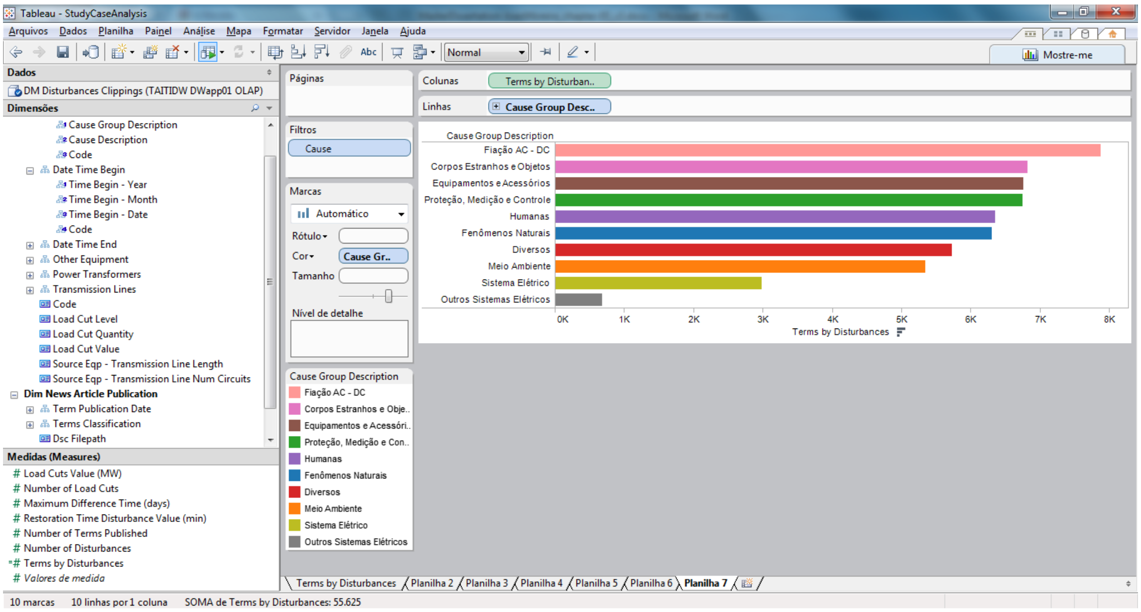


Figure 0.23: Comparison by Causes: (a)Terms Published (b) Disturbances

However, when analyzing the Terms by Disturbances measure, a curious result is found, where the “AC-DC Wires” cause has the greatest average, followed by “Strange Bodies and Objects”, “Equipments and Accessories”, “Protection, Measuring and Control” and “Humans”. The graph in Figure 0.24 demonstrates this analysis.



Normal

10 marcas 10 linhas por 1 coluna SOMA de Terms by Disturbances: 55.625

Figure 0.24: Comparison by Causes: (a) Terms Published (b) Disturbances

6. Disturbances caused by fires and most common terms published

Using the information from the structured universe, the user, when analyzing the average reestablishment time of the disturbances with load cut greater then 99MW in October and November 2009, could see a significant increase of about 1000% from the first to the second (Figure 4 - Navigation 1). The impact of this fact on the news of the electricity sector could be seen crossing the result with the terms and their occurrences in the clippings where most publications mentioning the terms "blackout" and "operation error" in November, citing the organization for various times. As the semantic treatment was performed, the query result includes the blackout term and its synonyms.

The amount of times that the term "blackout" was quoted in the news the next day and the disturbance in the following 15 days, and that on the fourth day, the news is not so prominent (Figure 4 - Navigation 2). From there, managers can organize themselves to intense activity by the press, with knowledge about failures in four days subsequent thereto. And with navigation to the source of the news, you can know which newspapers talk more about it, including access to the document of the report.