Hybrid Multidimensional Design for Heterogeneous Data Supported by Ontological Analysis: an Application Case in the Brazilian Electric System Operation

João Moreira[1], Kelli Cordeiro[2], Maria Luiza M. Campos[2], Marcos Borges[2]

[1] University of Twente, Services / Cyber-security / Safety (SCS) Group, Netherlands
[2] Federal University of Rio de Janeiro (UFRJ), Knowledge Engineering (GRECO) Group, Brazil
[1] j.luizrebelomoreira@utwente.nl
[2] {kelli,mluiza,mborges}@ppgi.ufrj.br

ABSTRACT

An issue in operating a national electric system is how the corporate image of an Independent System Operator (ISO) can be impacted by disturbances in the system and their related news publications from specialized press. To deal with it, a solution was developed in the context of the Brazilian Electric System National Operator (ONS): an analytical system for disturbances analysis integrating both structured and unstructured data sources. It considers both the daily news publications about the electric sector from ONS clippings website and the details of operational disturbances from the company data marts. We introduce here an adaptation of the hybrid multidimensional (MD) design method, considering heterogeneous data sources during business analysis and design phases. Most important, we illustrate how ontological analysis can enhance the semantic expressiveness of the MD modeling activity through a semi-automatic derivation process. The analytical potential is evidenced by a real scenario case study.

Keywords

Multidimensional design, unstructured, ontology, disturbances.

1. INTRODUCTION

Treating events of the electric sector through the support of database (DB) systems is a critical activity in the operation of multi-owned energy transmission, such as collecting and analyzing disturbances occurrences. Usually, an ISO company is responsible for this activity. In Brazil, ONS is in charge of monitoring the national electric system. A Decision Support System (DSS) based on Business Intelligence / Data Warehousing (BI/DW) architecture [7], coined Disturbances BI, uses structured data for disturbances analysis. Disturbances are most noticed by the population when blackouts occur. Because of their negative consequences, Brazilian press often publishes news on the subject, citing ONS, which may influence its corporate image. News publications regarding the electric sector are collected and made available daily at the clippings website. However, there is still the need of an analytical environment to support decision makers on analyzing the impact of disturbances on ONS institutional image. To achieve this goal, a solution to integrate structured data sources to unstructured data from the clippings in a BI/DW architecture has been a main requirement.

The problem addressed in this paper is the lack of a methodology for BI/DW solutions that considers both types of data sources. Indeed, there are a number of systems and solutions that extract information from text and integrate with existing DBs, but it is still missing a well-defined process to determine how this type of application can be used in a corporative context. We propose an approach for adapting Moss's BI/DW lifecycle methodology [11] so to consider heterogeneous data, addressing semantic problems during the MD design, such as ambiguity and low semantic expressiveness. In this paper a systematic approach is described, extending the hybrid MD design method by considering reverse engineering from text corpora during the source-driven activity. Furthermore, we guide how ontological analysis can be applied as a base for a semi-automatic process for MD schemas derivation, from a well-founded domain ontology that represents information in the data sources.

We also include the application of our approach in the case study of disturbances and news publications joint analysis for ONS corporate image. In the analysis-driven design activity we consulted ONS official glossary, domain engineers and the Common Information Model (CIM). In parallel, during the source-driven design activity, the transactional master DB of ONS and the Disturbances BI solution played the role of structured data sources. Corpora of news publications from clippings website were collected and analyzed by the DW designer as an unstructured data source. Afterwards, the disturbance domain ontology was built considering the scope of the original schemas. It was developed using ontological analysis based on a foundational ontology [4], a high-level category system for a solid grounding of conceptual modeling. The domain ontology had its semantic enriched during the verification and validation activity, where conceptual assertions were made to increase the model quality. Then, a semi-automatic process for MD structures derivation supports the designer in delivering the final MD schema for temporal analysis. The DB design and data cube construction phases were performed so joint analysis examples over the final data cube are presented through reports in an OLAP tool. This paper presents the continuation of the approach introduced in [10], covering the adaptation to consider heterogeneous data in the MD design methodology, supported by the architecture we introduced in [9]. Moreover, the study case is detailed and the ontological approach is depicted.

2. EVENTS IN ELECTRIC SYSTEMS

Reliable and sustainable electric systems depend on the ability of monitoring and responding, or even predicting, occurrences in the electric sector. The treatment of events in the transmission grid is a crucial activity, like responding to the shutdown of transmission lines or other equipment, i.e. electrical disturbances. Such treatment is made possible by solutions that handle large amount of data, providing high quality information for decision makers in adequate time. BI/DW architecture is a consolidated and usual way to deliver information originated in structured data sources.

2.1 BI/DW Solution for Disturbances Analysis

ONS is a non-profit ISO, unique in Brazil, performing its duties under the supervision and regulation of the national electric energy agency. Its mission is to operate the Integrated National System (SIN), a large hydrothermal system responsible for 97% of the Brazilian electricity supply. It has a strong predominance of hydroelectric plants with multiple owners and their facilities and equipment, such as power plants, transmission lines and power transformers. Equipment in SIN is subject to faults and failures of various natures, causing forced shutdowns of one or more devices. This can interrupt the power supply to consumers depending on the resulting load cut level. These events are known as electric disturbances and may be caused by atmospheric electric discharges, floods, fires or even human failures. ONS official glossary defines an electric disturbance 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 fulfill the coordination and the control of SIN operation are based on technical procedures, rules and criteria defined in normative documents. Information systems were developed by ONS, e.g. Disturbances Integrated System (SIPER), to support the registration of disturbances as abnormalities, undesirable events or unsatisfactory performance. They are integrated through ONS master DB, which stores the core transactional data from the electric system. Analytical processes of disturbances are supported by a BI/DW solution, coined Disturbances BI. It consolidates data from transactional systems and a historical DB. The integration is made through a conventional ETL process over structured data, being available in a disturbances data cube. The users can navigate and generate reports over the data cube through OLAP tools.

2.2 Impact of Disturbances on ONS Image

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 opportunities, threats and competences. ONS provides a daily summary of news related to the electricity sector in its intranet home page, the clippings website. Its main purpose is to provide to ONS collaborators news publications from the specialized press, quoting the organization when it is mentioned. The result of a disturbance in the system can lead to power supply cuts, popularly known as blackouts. This situation has a direct relation to the load cut level measure of disturbances fact in Disturbances BI. The negative consequences of a blackout to the population are numerous, generating large financial losses in all sectors of the economy. The press gives great focus to the subject, often citing ONS when such situation occurs, which may influence its corporate image. Among ONS main concerns in the electric security domain, the analyses of faults caused by disturbances in the system and their impact on users' lives, reflected in the media, is much relevant for decisions related to the corporate marketing. Current information systems present the information of disturbances and news about SIN independently. 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 can address this need. Existing DSS solutions were mapped: Disturbance BI and clippings website. They represent operational data sources captured in business case assessment.

3. APPROACH PROPOSAL

To address the methodological support needed for a disturbances and clippings integration solution we propose adaptations of Moss's methodology [11] to consider heterogeneous data. This BI/DW lifecycle resembles Kimball's [7] and Malinowski's [8] approaches, but it adds a metadata repository. Even being called a lifecycle, it lacks operations and decommissioning phases after deployment. However, it presents a balanced approach, considering complexity and practice. Each activity is part of a specific phase, as depicted in Figure 1 (left). Business analysis and design phases are considered the most important activities because they guide the BI/DW solution development. The efficacy of MD modeling is directly related to future costs in maintaining the BI/DW solution. It can be increased by avoiding conceptual mistakes through the application of a common understanding formalization [16]. Here we focus in the MD design activity as illustrated in Figure 1 (right). Our method is based on [8], but we consider an ontological approach.

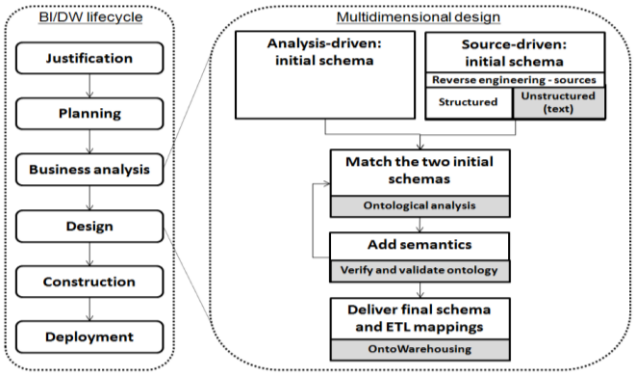

Figure 1. Hybrid multidimensional design activity adapted.

3.1 Hybrid Design for Heterogeneous Data

In our approach we consider unstructured data sources as text and ontological analysis to increase the semantic expressiveness of the MD modeling activity. The semantic expressiveness (or semantic power) is the quality of how precise a model is on representing the reality [4]. It considers both supply-driven and analysis-driven strategies running in parallel for deriving the initial schemas. In the analysis-driven schema derivation, the designer can use the domain knowledge from domain experts, existing procedures, glossaries, taxonomies or other terminological standards. In the supply-driven one, both structured and unstructured data sources can be analyzed. Then, the matching of the schemas sketches, i.e. their merging, is supported by ontological analysis [4], representing the business concepts in a domain ontology, categorized by top-level categories. Then, the domain ontology is verified and validated, increasing its quality in a cyclical way. Afterwards, rules defined as in a prior work, coined OntoWarehousing [10], can be applied to derive possible MD structures, used by DW designers in the final

definitions of MD schemas. Adaptations of the activities to cope with unstructured data and ontological engineering are described below.

3.1.1 Analysis-Driven Design

Each domain concept should be correctly named, uniquely identified and validated by business people who will access the data. Therefore, ontological analysis can be applied to support common understanding. A domain ontology can be sketched based on interviews with the main stakeholders and business official vocabularies, such as glossaries and standards. The ontology should be independent of technologies, not being influenced by any type of software or hardware. Current business processes should be understood, so the behavior of the concepts is mapped to the ontology, e.g. their creation or modification.

3.1.2 Source-Driven Design

In our approach we divided the reverse engineering in two main activities: from transactional DBs (usual), as structured sources, and, from textual sources. The result artifact from this activity is a sketch of the domain ontology from the point of view of the data sources. In both reengineering processes, making annotations about the origins of the data is crucial for the ETL design. Reverse engineering from structured data sources is widely addressed by supply-driven related works, e.g. AMDO [15]. It checks functional dependencies among tables by verifying relationships, cardinalities and constraints. Then, MD structures can be derived automatically based on a set of heuristics. It can capture important business rules and policies that may not be gathered during interview sessions. Some CASE tools implement this capability.

Reverse engineering from text generates representations of entities and their relations from unstructured data sources. This can be made manually or automatically. In both cases a set of text corpora is selected with support of business experts. Then, its content is analyzed. Automatic approaches consider Natural Language Processing (NLP) and Information Retrieval (IR) techniques applied to the corpora, resulting in suggestions of models. Entity and relations recognition techniques play an important role on ontology generation. 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. In our approach we do not choose one specific technique or tool. Instead, we guide the designer to first check the existing NLP and IR solutions.

3.1.3 Domain Ontology for Initial Schemas

The inputs for this phase are the model sketches from prior activities. Common concepts found in these representations should be matched or associated, by annotating their data structures origins. This information will be necessary to build the linkages between the structured and the unstructured universes for designing the ETL process. After matching all entities, annotating their origins, the consolidated domain ontology should be built. We propose the use of the OntoWarehousing [10] ontological approach to increase the semantic expressiveness of the MD design. It presents a systematic and semi-automatic derivation process to suggest MD structures from categories of a foundational ontology, a high-level category system that represents concepts such as endurants (things that are in time) and perdurants (events or things that happen in time) [3] – refer to section 5 for a more detailed explanation. The output of this activity is the consolidated well-founded domain ontology.

3.1.4 Add Semantics: Verify and Validate

In this phase the designer analyzes the foundational constructs and checks if the entities and relations from the model are semantically consistent, also verifying business rules violations. Domain experts can support the quality improvement of the domain ontology, ensuring that the model is semantically correct, covering the main entities involved in the business requirements, avoiding ambiguity. This is a cyclical process: when the designer finds an inconsistency, he fixes the model and validates it again. Verifying and validating (V&V) ontologies with many concepts can be unfeasible for humans because of their size and complexity. Thus, a common practice is to choose ontology sub-domains, validate each separately and merge them. The resulting artifact is the valid well-founded domain ontology.

3.1.5 Deliver Final Schema: Derivation Process

The final MD schemas are designed based on the well-founded domain ontology and other existing MD schemas. In common methodologies [7,8] this task depends purely on informal guidelines for MD designer decisions. It depends on tacit knowledge, being error prone. In OntoWarehousing [10], we defined a set of mapping rules to derive possible MD structures from the well-founded domain ontology. The MD designer can use this method to increase its assertiveness. 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 modeler the possible MD structures inferred. Thereafter, the derived MD concepts are conformed to other MD schemas, providing new analytical possibilities. At last, the designer defines the final MD schemas with data sources annotations as comments in natural language to serve as specification for the ETL processes. Other ontological approaches for MD design can be used in parallel, combining the final MD concepts produced such as in [16].

4. APPLICATION CASE

To handle disturbances and clipping s integration we have applied our approach in the construction of a BI/DW solution. Business analysis, design and construction phases were performed and some analysis examples over the final data cube could be made. To support the BI/DW lifecycle, Enterprise Architect (EA: http://www.sparxsystems.com.au/) CASE tool was chosen, which provides a full-set of capabilities for requirements formalization, conceptual models design and behavioral aspects representations. In addition, OntoUML Lightweight Editor (OLED: https://code.google.com/p/ontouml-lightweight-editor/) software and its plug-in to EA supported the V&V process.

4.1 Business analysis

4.1.1 Analysis-Driven Design

The analysis driven design was supported by ONS domain experts, the official glossary, and the CIM IEC 61970 (http://www.iec.ch/smartgrid/standards/). ONS official glossary contains the definitions of the main terms used in the electric sector. It serves as main business concepts source for common understanding among ONS and other agents. It helped in asserting the initial domain representations. When a specific term was not encountered in the glossary or there was ambiguity, the domain experts were consulted, mostly power systems engineers. They asserted specific rules, such as the part-whole relation between a disturbance and a forced shutdown, where a forced shutdown is part of one unique disturbance and it is existentially dependent of the disturbance. The CIM IEC 61970 is an international standard built by the electric power industry and it was adopted to support

information systems interoperation and common concepts agreement. Particularly, the main part of this standard was chosen, the IEC-61970 for energy management. It brings the representations of core concepts of electric power transmission and distribution domain, such as equipment (e.g. power transformer) and its sets (equipment containers) as power system resources. As a practical advantage, this standard is available and extensible in the EA tool as a UML class structural package.

4.1.2 Source-Driven Design

The involved data sources were listed as: the SIPER cut-off of ONS master DB, an entities mapping between the master DB and CIM models, Disturbances BI and clippings website (news publications). The physical data model was used to check tables, attributes, relationship integrities and constraints that implement the domain behavior. This type of information was included when representing the company concept. The entities mapping specification between the master DB and CIM describes the equivalence between the data structures from ONS master DB and the classes and relations of CIM meta-model. This document was previously built and used for the development of an ETL process, which extracts data from ONS master DB, transforms and load it in a CIM file representation. Thus, the domain could be designed in English terms, reusing the existing knowledge. The available ETL processes of the Disturbances BI were analyzed. We checked the ETL process to load the fact disturbance, which has associations to dimensions such as owner agent, source equipment, cause, begin/end time, among others. The clippings website was checked and the news publications sub-domain modeled, as textual information source. At first, a web crawler was built to download news publications from January 2011 to March 2013. A textual ETL process from a prior work [9] was applied in these corpora selected. It resulted on a data repository, named terminological DB, which stores the terms and their lexical and semantic categories, supported by IR.

4.1.3 Domain Ontology for Initial Schemas

The domain ontology was built in the EA tool supported by OLED plug-in. As starting point, the SIN domain package was composed by five sub-domains: companies, facilities, equipment and geographical region (structural aspects); and disturbances and news publications (dynamic aspects). The most important relation to link disturbances and news publications was defined through the temporal formal relation "before" at the conceptual level, meaning that a disturbance that occurs before a news publication can be somehow related to it. Disturbance and news publication are both classified as complex events, inheriting a series of properties, such as their beginning and ending time points, composition by other events, etc. There is a practical implication in this representation that was found during the construction phase regarding how long a disturbance occurred before a publication about it. For instance, if a disturbance occurred in 2010 and some news are published in 2013, if even this relation respects the "before" relation, it is most unlikely that they are related. Therefore, we stated a threshold of ten days based in prior experimentation [9]. Initial analysis evidenced the increase in publications after a severe disturbance and a decrease on subsequent days, reaching the publications average in ten days.

4.1.4 Add Semantics: Verify and Validate

This activity was supported by OLED software. After the first time designing the main concepts in the domain ontology (in EA tool), we exported it as an XMI file and imported into the OLED tool. During the import process, the tool provides a selection of classes and relations that the user would like to validate. Cut offs were made for V&V each part of the domain ontology. We could validate the domain ontology by examples and counterexamples, simulating instances of classes and their relations through the visual capability of Alloy analyzer provided in OLED. Moreover, OCL check statements were written as business rules representations. The result of this activity was the well-founded domain ontology considering disturbances and news publications.

4.1.5 Deliver Final Schema: Derivation Process

The final MD schema was designed based on the well-founded domain ontology. The OntoWarehousing approach was applied to discover MD structures, implemented through a prototype, which was executed in the domain ontology (refer to [9]). The MD schema to analyze the "before" relation could be designed by the derived MD structures from the proposed rules and conciliated with the existing disturbances MD schema. Moreover, from the axiomatization of the temporal operator "before", the constraint for the WHERE clause of the SQL to load the fact at the end of the ETL process was derived. Each event is considered a data structure (e.g. table, view, procedure) having the columns of start and end time points as date/time fields, where "before" is represented when the end of the first event is lower than the begin of the second event. As a result, the domain ontology, the MD schema specification, the requirements document and a high-level design of the ETL process were produced.

4.2 Construction and Deployment

A DB was physically created reflecting the MD schema specification. It supported the ETL process construction based on the ETL design and the domain ontology. It considered an ETL integration architecture coping with textual ETL, termed JointOLAP [9]. It uses IR and NLP techniques for the extraction process from text files and loads the terminological DB. The high-level data flow design is illustrated in Figure 3, having each activity supported by a set of tools. It begins with parallel activities: the conventional ETL process execution of disturbances Operational Data Store (ODS) and the textual ETL downloading news publications through a web crawler. Then, JointOLAP is performed in these documents, populating the terminological DB with all news articles content. It checks patterns in headers (e.g. title, publication date and press company), structuring this information in the DB.

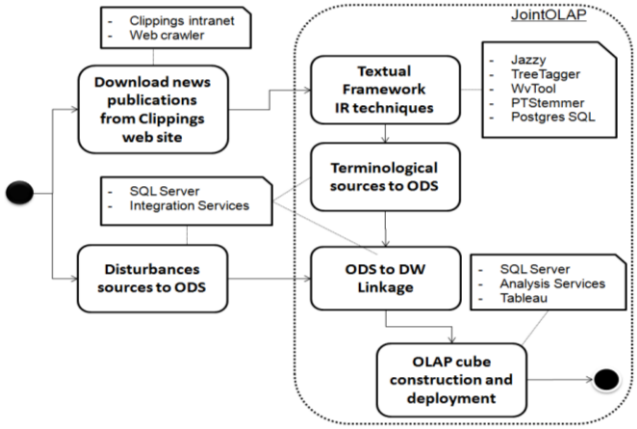

Figure 3. High-level ETL process workflow design.

The framework applies orthographic correction, case sensitive elimination, tokenizing and morphological classification, stop word and punctuation elimination, synonymous resolution and stemming. It indexes the terms, their stems, morphological

classification (e.g. verb, adjective, preposition and adverb). A list of business terms was created based on ONS official glossary and the terms of news articles were marked. An ETL process was created to execute the linkage activity, integrating data between disturbances and terminological ODSs and loading the final MD schema. It was used to build a data cube, making data accessible by OLAP tools.

The impact of blackouts on ONS corporate image analysis was supported through the OLAP tool connected to the data cube. The navigation was made possible through the dimensions and their hierarchies, enabling the exploration of the measures within the fact and possible aggregations with drill-down and roll-up operations, as well as graphics and reports generation. An analysis example is the number of terms published in news by the load cut level measure of the related disturbances. Average terms occurrences by disturbances is 5,720. Analyzing this measure by the severity of the disturbances makes it possible to verify a direct relation with the number of news publications. The more severe are the disturbances, more terms are published. In average the "blackout" term is, considering synonymous, the 27th most common term, but when load cut level is greater than 99MW it jumps to 1st. When it is lower than 49MW it becomes the 52nd of the blackout terms. News publications by disturbance month in 2011 revealed a significant variance of terms published after the disturbances occurred in February, which caused an enormous blackout in northeast. The number of publications increases considerably in March and April, then, it decreases back again to the standard baseline. These analyses are evidences that the expressivity enrichment of the MD design is a differential of our proposal, having the relation "before" connecting disturbances and news publications. The counting of mentions to certain words emphasized the press terminology when severe disturbances happened, addressing the main requirement of the solution.

5. RELATED WORK

Energy data management is a knowledge area that addresses the techniques for collecting, storing and analyzing huge amounts of data from the energy sector through IT solutions. Common definitions of data and information concepts by ontological approaches are still open topics [14]. Ontologies may be applied for the representation of portions of reality to understand, communicate and reason about the domain. In software engineering it is commonly built as UML class diagrams. In artificial intelligence it is commonly built as semantic networks and in DB area as ER diagrams. All these models seek to represent entities, relationships, properties, rules and restrictions of the involved domain. It can be considered formal when it is machine-processable, enabling automated reasoning by the semantics described in formal logic [4]. An example of an ontological solution for real-time data sources integration is the smart grid domain ontology introduced in [2]. It presents representations of event types, such as electrical appliances, weathers, storages and generators.

To fulfill analysis requirements in a BI/DW project, the MD and ETL design activities are supported by ontological solutions addressing the lack of semantic expressivity in MD models [1,12]. We introduced OntoWarehousing approach [10] where ontological analysis is applied in MD design based on formal ontology discipline. Analogous to formal logic, which contemplates logic formal structures, such as truth, validity and consistence [4], formal ontology is founded on mathematical disciplines of mereology (part-wholes), theory of dependence and topology and principles of identity and unity. In our approach we used the Unified Foundational Ontology (UFO) [3] to enrich the domain representation. It is a high-level category system, a top-level ontology, which presents these philosophical concepts interpreted, describing the most general concepts, such as space, time, matter, object, event and action, concepts independent of a domain or a particular problem. In OntoWarehousing, the domain ontology is semantically enriched by these top-level formalizations, e.g. domain concepts classified as events, participations, temporal relations, roles, among others. These high-level categories are used during the derivation process, which is based on a set of mapping rules from UFO categories to MD structures. The idea of such interpretation mapping from a foundational ontology to MD concepts was first discussed in [12].

A survey [1] summarized semantic web technologies (e.g. RDF and OWL) applied in BI/DW, discussing advantages, disadvantages and cases. Description logic can be used to assist data aggregation processing by reasoning services over rules. To enforce the semantics in MD design, Romero et al. [15] proposed the AMDO approach for conceptual modeling in BI/DW solutions based on end-user requirements elicitation and hybrid MD design. It uses a supply-driven mechanism where a set of rules formalized in first order logic derive MD structures (facts, measures, dimensions, hierarchies and attributes). The GEM approach [16] operationalizes the whole process, automating the identification of potential MD concepts by analyzing the domain ontology and the semantic annotations represented in OWL-DL. The ORE module [6] evolves GEM considering the complexity of frequent changes in MD design, integrating each new analytical requirement. These tools are mature enough, but they still lack some common understanding, which can be provided by a foundational ontology.

The need of considering unstructured data in BI/DW solutions is fundamental for business analytics. Even so, most of BI/DW methodologies are based on structured data. Analyzing and exploring data from heterogeneous natures, jointly, can enhance the potential of analytical applications offered to decision makers [5]. Several works are being proposed to consider the unstructured data sources by applying IR and NLP techniques as listed in [13]. We introduced the architecture JointOLAP [9] as a solution for joint exploration. It takes advantage of semantic treatment mechanisms for the unstructured content.

6. CONCLUSIONS AND FUTURE WORK

We introduced our approach as an adaptation of Moss's BI/DW methodology to consider heterogeneous data by making specific changes in the hybrid MD design activity. It takes advantage of IR and NLP techniques during the source-driven analysis phase to derive complementary analytical elements and associate data from structured and unstructured sources. In addition, we increased the MD design semantics by applying ontological engineering supported by a foundational ontology. A case study regarding ONS joint analysis of distribution consumption energy affected by press publications was described. The MD schema was derived considering the "before" temporal formal relation between disturbances and news publications. This case study is a work-in-progress, being considered as an original research and an industrial-strength solution for energy data management. Its main contribution is the integrated OLAP specification for ONS corporate image impact analyses built based on our approach. Indeed, the correlation between disturbances and news publications is not surprising. However, our case study could materialize this relation and its exploration using real data.

Lessons learned from our approach application on hybrid MD design activity include: (i) unstructured data sources proved to be essential information for MD conceptual design; (ii) ontological engineering seems to be an adequate method to improve knowledge acquisition and its design through a well-founded domain ontology; (iii) we believe this method may increase the productivity in business analysis and design phases of BI/DW projects. Some limitations are: (i) the derived ETL process did not considered implementation issues such as surrogate keys treatment and indexing management; (ii) to simplify, we considered a 1:1 relation between terms and categories, restricting the terms classification; (iii) the reverse engineering from text can be unfeasible depending on the amount of data; and (iv) the choice of MD concepts for the resulting MD schemas continues to be a tacit activity depending on the designer's decisions. Future work includes: (i) to enhance the textual ETL for news publications with new text treatment techniques, considering distributional models; (ii) to apply sentiment analysis techniques to discover the polarity of the sentiment around the events (e.g. positive, negative and neutral); (iii) to predict how quickly after an event the sentiment for or against ONS changes in the news and also to incorporate sentiment from crowd sources; (iv) we believe that entity recognition and relation extraction activities can consider categories of a foundational ontology; (v) regarding the involved tools, we believe that the prototype should be developed as an extension of OLED integrated to GEM/ORE.

7. ACKNOWLEDGMENTS

Our thanks to CAPES PhD scholarship (process BEX 1046/14-4).

8. REFERENCES

[1] Berlanga, R., Romero, O., Simitsis, A., Nebot, V., Pedersen, T. B., Abelló, A., Aramburu, M. J. 2011. Semantic Web Technologies for Business Intelligence. *BI Applications and the Web - Models, Systems, and Technologies.* pp. 310-339.

[2] Gillani, S., Laforest, F. e Picard, G. 2014. A Generic Ontology for Prosumer-Oriented Smart Grid. *3rd workshop on Energy Data Management (EnDM) - EDBT*.

[3] Guizzardi, G., Wagner, G., Falbo, R. A., Guizzardi, R. S. S., Almeida, J. P. A. 2013. Towards Ontological Foundations for the Conceptual Modeling of Events. *Conceptual Modeling, Lecture Notes in Computer Science.* pp. 327-341.

[4] Guizzardi, G.; Lopes, M.; Baião, F.; Falbo, R. 2010. On the importance of truly ontological representation languages. *Journal of Info. Systems Modeling and Design (IJISMD).*

[5] Inmon, W. H., Strauss, D. e Neushloss, G. 2008. *DW 2.0 - The Architecture for the Next Generation of DW*.

[6] Jovanovic, P., Romero, O., Simitsis, A., Abelló, A., Mayorova. 2014. A requirement-driven approach to the design and evolution of data warehouses. *Information Systems 44, 94-119*.

[7] Kimball, R. e Ross, M. 2013. *The Data Warehouse Toolkit: The definitve Guide to Dimensional Modeling.* Wiley.

[8] Malinowski, E. e Zimányi, E. 2009. *Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications.* Springer.

[9] Moreira, J., Cordeiro, K. F. e Campos, M. L. M. 2013. JointOLAP – Sistema de Informação para Exploração Conjunta de Dados Estruturados e Textuais - Um estudo de caso no setor elétrico. *SBSI.*

[10] Moreira, J., Cordeiro, K. F., Campos, M. L. M., Borges, M.. 2014. OntoWarehousing – Multidimensional Design Supported by a Foundational Ontology: A Temporal Perspective. *16th International Conference on Data Warehousing and Knowledge Discovery (DaWaK/DEXA).*

[11] Moss, L. T. 2003 Business Intelligence Roadmap - The Complete Project Lifecycle for Decision-Support Applications. 1st. s.l. : *Addison-Wesley Professional*.

[12] Pardillo, J. and Mazón, J. N. 2011. Using Ontologies for the Design of Data Warehouses. *International Journal of Database Management Systems (IJDMS).* May, pp. 73–87.

[13] Park, B. K. and Song, I. Y. 2011. Toward Total Business Intelligence Incorporating Structured and Unstructured Data. *Proceedings of the 2nd International Workshop on Business intelligence and the WEB (BEWEB).* pp. 12-19.

[14] Pedersen, T. B. 2014. Energy Data Management: Where Are We Headed? *Panel paper, 3rd workshop on Energy Data Management (ENDM) - EDBT.*

[15] Romero, O. e Abelló, A. 2010. A framework for multidimensional design of data warehouses from ontologies. *Data & Knowledge Engineering Journal. Elsevier Science Publishers B. V.* Vol. 69, pp. 1138-1157.

[16] Romero, O., Simitsis, A. e Abelló, A. 2011. GEM: Requirement-Driven Generation of ETL and Multidimensional Conceptual Designs. *13th International Conference on Data Warehousing and Knowledge Discovery (DaWaK).* August, pp. 80-95