# ETL4Social-Data: Modeling Approach for Topic Hierarchy

Afef Walha[1], Faiza Ghozzi[1,2] and Faïez Gargouri[1,2]
[1]*MIRACL Laboratory, Sfax, Tunisia*
[2]*Institute of Computer Science and Multimedia, University of Sfax, Tunisia*

Keywords: ETL, Social Data Warehouse, BPMN, Modeling, Topic Detection.

Abstract: Transforming social media data into meaningful and useful information to enable more effective decision-making is nowadays a hot topic for Social Business Intelligence (SBI) systems. Integrating such data into Social Data Warehouse (SDW) is in charge of ETL (Extraction, Transformation and Loading) which are the typical processes recognized as a complex combination of operations and technologies that consumes a significant portion of the DW development efforts. These processes become more complex when we consider the unstructured social sources. For that, we propose an ETL4Social modeling approach that designs ETL processes suitable to social data characteristics. This approach offers specific models to social ETL operations that help ETL designer to integrate data. A key role in the analysis of textual data is also played by topics, meant as specific concepts of interest within a subject area. In this paper, we mainly insist on emerging topic discovering models from textual media clips. The proposed models are instantiated through Twitter case study. ETL4Social is considered a standard-based modeling approach using Business Process Modeling and Notation (BPMN). ETL Operations models are validated based on ETL4Social meta-model, which is an extension of BPMN meta-model.



## 1 INTRODUCTION

User generated Content (UGC) are created by users of a system or service and made available publicly on that system. UGC most often appears as supplements to online platforms, such as social media websites, and may include such content types as blog posts, wikis, videos, or comments. Companies and institutions worldwide anticipate gaining valuable insights from obtaining access to such data and hope to improve their marketing, customer services and public relations with the help of the acquired knowledge. To this point, social media content has given birth to a novel area of data analysis, namely social media analysis.

Social media users publish their thoughts, activities, and interests in the form of text streams to share them with others in a social network. In this context, most analysis research has primarily focused on sentiment analysis and topic extraction from the text streams. Textual UGC is unstructured or semi-structured, noisy and informal. Thus, it needs a semantic enrichment in order to identify as much features as possible. The social media dataset can be enriched and extended in many ways offering a whole new set of analytical aspects to business analysts.

SDW modeling is giving more attention to the UGC which became a key interest for business owners. The enormous amount of valuable textual clips must be integrated in Decision Support Systems (DSS) which have to be adapted to this type of content. One of the studied issues for this purpose is the topic discovery from a text stream since the topics play a crucial role in user search, friend recommendation, and contextual advertisement. The topic hierarchy is not similar to traditional hierarchies regarding their schemata and instances, which are fluid.

Recently, the emergence of opinion analysis and topic detection of social data has generated much interest in DSS research community. Researchers take care of social data characteristics and propose semantic analysis solutions. These proposals suggest transformation methods to integrate social data in SDW. Furthermore, it is recognized that data transformation is the key role of ETL processes which is a very time-consuming step, it takes about 80% of the total time of the decision-making implementation due to its difficulty and complexity.

Walha A., Ghozzi F. and Gargouri F.
ETL4Social-Data: Modeling Approach for Topic Hierarchy.
DOI: 10.5220/0006588901070118
In *Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KEOD 2017)*, pages 107-118
ISBN: 978-989-758-272-1

Despite that, researchers did not propose a modeling of ETL processes covering social data.

The design and the implementation of an ETL process usually involve the development of very complex tasks imposing high levels of interaction with a vast majority of the components of DW system. Our study is motivated by the fact that the existing ETL modeling approaches didn't cover social activities. This step is complex because Social ETL (S-ETL) modeling is not limited to standard operations (e.g. location detection, time detection), but it integrates semantic processing operations (e.g. topic discovering, sentiment analysis). The difficulty with semantic operations is caused by the dynamicity of social sources and SDW schemas. We propose for that a conceptual modeling approach, called ETL4Social which deals with the complexity of ETL processes in social environment. The main objective of this paper is the design of S-ETL operations including extraction, transformation and loading processes. ETL4Social is considered a standard-based approach because conceptual modeling is based on ETL4Social meta-model which extends BPMN meta-model. The proposed models are instantiated on Twitter social case study to be validated.

The remainder of this paper is organized as follows. Section 2 reviews some related works concerning conceptual ETL modeling approaches and social data integration approaches. Section 3 summarises our proposed modeling approach which describes Extraction, Transformation and Loading processes specific to social data. Section 4 details the ETL4Social operation models, which are validated through examples correspondent to Twitter social data. Section 5 presents the meta-model in which we are based to design social operation models. Finally, Section 6 gives a conclusion and some future research perspectives.

# 2 RELATED WORKS

Our work focuses on the design of social ETL processes and the issues that arise when semantic analysis of social data is adopted for ETL4Social modeling. To this end, this section presents a review of recent studies dealing with the conceptual modeling of ETL processes and the integration of social data in SDW.

## 2.1 Conceptual Modeling of ETL Processes

One of the earliest researches on the ETL field was conducted by (Vassiliadis et al., 2002). The authors proposed a modeling approach based on a non-standard graphical formalism. The special notation devoted to ETL modeling reduces enormously ambiguities for the designer manipulating the graphical symbols. It provides customizable and extensible mechanism so that the designer can enrich it with his particular patterns. Recently, (Petrovic et al., 2016) proposed Domain Specific Languages (DSLs) defining concepts for different aspects including data and control. The activities in ETL process represent the actual data operations (i.e., the data flow), while the control flow represents the execution order of these activities. The aim of DSLs is to provide only a minimal set of domain-specific concepts, with clear and precise semantics, along with a set of strict rules controlling their usage and the way in which they can be composed. Unfortunately, efforts are required to assimilate the non-standard concepts and their graphical symbols.

ETL conceptual modeling approaches have been also proposed based on the Unified Modeling Language (UML). (Trujillo and Luján-Mora, 2003) approach extends the class diagram by new UML profiles to allow modeling of ETL operations. Still, (Trujillo and Luján-Mora, 2003) represents multiple important aspects, such as ETL constraints, filters, and data mapping, as UML notes where no restriction on their content is imposed. Despite the advantage of this approach to offer a decomposition of complex ETL to simple processes, the ETL modeling using class diagram cannot represent control flows or temporal restrictions, etc. In response to this limit, (Muñoz et al., 2008) propose a dynamic modeling based on UML activity diagram, which offer a dynamic aspect considering the execution order of control flows and temporal constraints. This approach proposes a decomposition of ETL processes into simple activities (e.g. aggregation, filter, conversion, etc.). An activity is designed by an UML activity diagram. It is characterized by input and output pin(s), a set of actions and parameters. Using activity diagram offers a visual modeling of complex processes and provides interoperabilities with others schemas. (Muñoz et al., 2010) prove the efficiency of activity diagram in representing dynamic aspects of ETL processes through a set of validation experiments. To this end, (Mallek et al., 2014) adopt the ETL design approach proposed in (Muñoz et al., 2008) to

deal with web data. The proposed models are extended to be suitable to data available on the web including web logs, web sites and clickstreams, etc. We notice that the mentioned approaches using activity diagram have the following advantages: standard-based, adaptation/extension of UML for ETL modeling, top-down modeling process. In fact, an ETL process is represented as UML packages which enable the ETL process decomposition into different logical units.

Considering an ETL process as a particular type of business process, (El Akkaoui and Zimányi, 2009) and (Wilkinson et al., 2010) have adopted the standard Business Process Model and Notation (BPMN). From the various BPMN symbols, the authors have proposed a customization that allows selecting a subset of graphical constructs (e.g., ETL palette) for designing ETL processes. For the implementation purpose, the authors have proposed a mapping process from BPMN to Business Process Execution Language (BPEL) which enables the execution of the ETL process. These works were followed by (El Akkaoui et al., 2012), which propose a modeling framework based on a meta-model in the Model Driven Development (MDD) architecture.

For the modeling of ETL process, BPMN notation seems to be a good choice since it can cover a deficit of communication that often occurs between the design and implementation of business process. The ETL modeling solutions based on BPMN standard are rich enough to cover all ETL steps. Otherwise, they are not suitable for social data environments.

## 2.2 Data Integration in Social Data Warehouse

In the context of SBI, the category of UGC that most significantly contributes to the decision-making process in the broadest variety of application domains is the one coming in the form of textual clips. Clips can either be messages posted on social media (such as twitter, facebook, blogs, and forums). Digging information useful for decision-makers from textual UGC requires first crawling the web to extract the clips related to a subject area, then enriching them in order to emerge opinion information from the raw text. This issue has motivated many researchers to propose semantic enrichment solutions of textual UGC in order to be ready for decision analysis. Recent approaches propose frameworks that transform social network data into meaningful and useful information to enable more effective decision-making. They incorporate these data into the existing MD structures.

The majority of approaches establish mappings between incorporate data and their sources. A recent approach to integrating BI with sentiment data was proposed by (García-Moya et al., 2013) where a corporate DW is enriched with sentiment data from opinion posts. As a result, sentiment and business data can be jointly analyzed by means of OLAP tools. (Berlanga et al., 2015) propose an open and dynamic framework, where data can be linked to external sources on demand. They propose a novel semantic data infrastructure for capturing and publishing sentiment data to enable SBI. It also performs automatic extraction of sentiment data from posts, and their linkage to the infrastructure. As a result, decision makers will be able to incorporate opinion related dimensions in their analysis.

Incorporating social opinion data in SDW modeling require a specific definition and modeling of Extraction, Transformation and Loading processes. Modeling such processes is a hot topic for many researchers that proposed a novel approaches for transforming the social data to SDW. (Rehman et al., 2013) describe the process of transforming the original stream into a set of related multidimensional cubes and demonstrate how the resulting data warehouse can be used for solving a variety of analytical tasks. This approach presents the implementation level based on multi-layered system architecture that focuses on the critical stage of transforming the original stream into a structured multidimensional dataset consisting of measures and dimensions. They also elaborated on various options of enriching the dataset and its structure by means of derivation, data mining, linking to additional sources or using external APIs for detecting new features. (Walha et al., 2016) define a lexicon opinion analysis approach that extracts informal texts expressed in the twitter social network, cleans them and then analyzes them in order to derive their sentiment polarity which is a dimension in SDW. This approach proposed a text mining algorithm, called POLSentiment, based on lexical resources to firstly extract opinion words and emoticons from the textual UGC and then analyse it to assign a polarity.

A key role in the analysis of textual UGC is also played by topics, meant as specific concepts of interest within a subject area. Several studies attempted to discover user's topics of interests from a text stream since the topics play a crucial role in user search, friend recommendation, and contextual advertisement. (Shin et al., 2014) study a problem of

detecting the topics of long-term steady interests to a user from a text stream, considering its dynamic and social characteristics, and propose a graph-based topic extraction model. In this model, a topic is represented by a term in a social text stream, and refer to the term representing a long-term persistent topic of interest as a PT (persistently topical) term. PT term is useful in user search, friend suggestion, tag recommendation, and contextual advertisement in microblogging services. The proposed model discovers the PT terms by utilizing the information of topically or socially similar streams based on a graph-based ranking model.

Users are interested in knowing people opinions regarding a specific topic. But, an analysis over the reputation of a single topic is not sufficient to get a clear idea about the global mood. Thus, decision makers need to aggregate different topics according to several criteria for performing efficient analysis. Obviously, topics became a dimension in a Social BI multidimensional cube. In order to enable the aggregation of topics belonging to different levels, a topic hierarchy should be defined. However, this topic hierarchy is not similar to regular hierarchies in terms of their irregularity, dynamicity as well as the existing semantic relationships between topics. Hence, creating this topic hierarchy is a challenging task for the ETL team. (Gallinucci et al., 2013) stated that a topic hierarchy can be approached in the form of an ontology that should follow a well-defined structure, which contains a complete list of topics, their relationships, as well as their corresponding levels. For instance, the topic "Apple" is referring the level "Brand". In (Gallinucci et al, 2013), authors are focused on topic hierarchies and their effective modeling.

We notice that the above-mentioned approaches are high expressive to model topic hierarchies based on same specific requirements: heterogeneity and dynamicity of topic classification. Despites these advantages, ETL activities that describe the execution order of topic detection processes from text stream to topic hierarchy attributes are not modeled. This paper addresses this issue and proposes a conceptual modeling of ETL operations that focus on topic hierarchy building.

ETL modeling approaches above-reviewed in section 2.1 provide very interesting models. Otherwise, they are not adapted to social data characteristics. The objective of ETL4Social data modeling approach is to extend the proposed models by semantic enrichment specific to textual social data sources. In this work, we focus on ETL modeling of the dynamic topic hierarchy.

# 3 ETL4Social-Data MODELING APPROACH

Social data are unstructured and heterogeneous in comparison with business data sources. Due to these data characteristics, existing ETL operations cannot be used by ETL designers to map social sources into SDW. This problem requires the redefinition of Extraction, Transformation and Loading processes. ETL4Social modeling approach reply to this need through social operations design.

## 3.1 Overview of ETL4Social Modeling Approach

ETL4Social modeling helps ETL designers to correctly use the adequate social operations resolving the problem of complexity related to social data. This complexity is due to the integration of unstructured, informal and heterogeneous opinion data extracted from social network sources (e.g. twitter).

The asset of ETL4Social-Data is the definition and modeling of extraction, transformation and social operations applied on SDS in order to be loaded into the SDW. Figure 1 shows an overview of the proposed approach. It is organized in four main processes: media clips crawling, data preprocessing, social ETL (S-ETL) operations modeling and data loading.

Social networks hold a huge volume of data about many topics of interests related to diverse contexts of study. To keep only valuable information replying to a specific study context, the data crawling step is realized by crawling services based on a set of keywords describing the ETL designer requirements. During this step, useless media clips data such as very short or non-English text clips should to be also discarded before their storing in the Data Staging Area (DSA).

In traditional ETL processes, the extraction step is followed by transformation of data sources to DW elements. In social environment, media clips are a user generated content (exp. tweet text) that may be informal and unstructured. So, the crawling of media clips must be followed by a preprocessing step that defines a set of cleaning operations specific to social data. Among these operations, some are standard, which are approved by many research communities, and others are specific to semantic analysis objectives. For example, cleaning operations before topic detection differ from those defined before sentiment polarity detection. At this point,

preprocessed media clips data stored in the DSA are ready to the next step: Transformation.

The transformation process is realized through the execution of a set of operations. This process is defined by transformation rules that map data sources to SDW. Existing ETL approaches designed standard transformation operations which are not adapted to social data characteristics and not deal with semantic analysis aspect. In ETL4Social, two types of social operations are proposed: (1) standard and (2) semantic. In type (1), only data stored in DSA are used to determine SDW elements. Otherwise, type (2) uses other external resources (e.g. domain ontology) to semantically enrich data sources in order to be transformed into target data (e.g. topic detection, semantic polarity).

The final process of our approach is the loading of data into SDW. It consists on data verification steps before its insertion in the SDW.

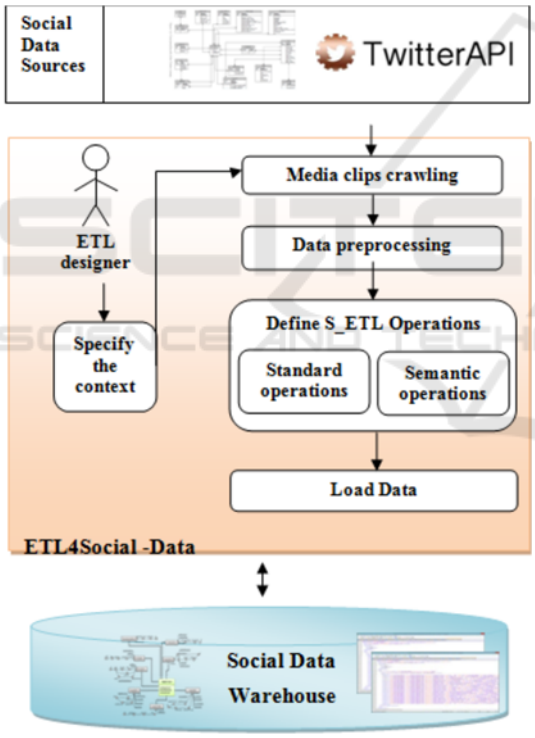


Figure 1: Overview of ETL4Social Modeling Approach.

The literature indicates that the BPMN standard is well adapted to model ETL activities. Also, it allows covering a deficit of communication that often occurs between the design and implementation of business processes. We choose this standard to design social extraction, transformation and loading operations as business processes.

## 3.2 Definition of DWB Schema from Social Sources: Case of Twitter

In the literature, many approaches are interested to decision analysis of twitter social data and they propose many multidimensional views of twitter SDW schema. These schemas were designed based on the Twitter Data Warehouse schema presented in (Rehman et al., 2013). In this schema, **TweetFact** and **UserFact** are defined as facts according the dimensions **Time**, **Date**, **Tweet**, and **User**.

A tweet object is known as the basic atomic building block of all things in Twitter. Tweet is also known as a “status update”. In fact, it is a message which may include texts, links, photos, or videos, that may be embedded, replied to, liked, or deleted. Thus, a tweet may be a rich source of information through which users express their feelings and opinions on every topic of interest. A tweet object stores several data fields that can be assigned directly or by semantic enrichment as fact/measures, dimensions or dimensional hierarchies. We think on extending the described twitter DW schema by semantic analysis elements. This enrichment integrates semantic attributes specific to new analytical aspects including sentiment analysis, study context and topics of interest. Figure 2 shows the resulted semantic twitter SDW schema, in which **SentimentDim**, **ContextDim** and **TopicDim** are defined as semantic dimensions attached to the analysis subject **TweetFact**. **SentimentDim** attributes contains information about tweet sentiment (e.g. high positive, positive, low positive) and its polarity, i.e. positive, negative or neutral. **ContextDim** attributes are related to the context of study and its relevance. As shown in our proposed semantic twitter SDW schema, **TopicDim** is different from other dimensions. It is presented without attributes because it has dynamic hierarchy structures that can be designed according to the study context specified by ETL designer (Gallinucci et al., 2013).

Semantic dimension attributes are derived by the execution of ETL operations on the textual social data source. Due to the unstructured and informal texts available in social networks, their extraction, transformation and loading into SDW attributes is a complex task. Our objective is to define social ETL operations. In this paper, we insist mainly on those specific to topic dimension attributes and hierarchies. This work is an extension of our previous work (Walha et al., 2016), which describes a lexicon approach that determines sentiment

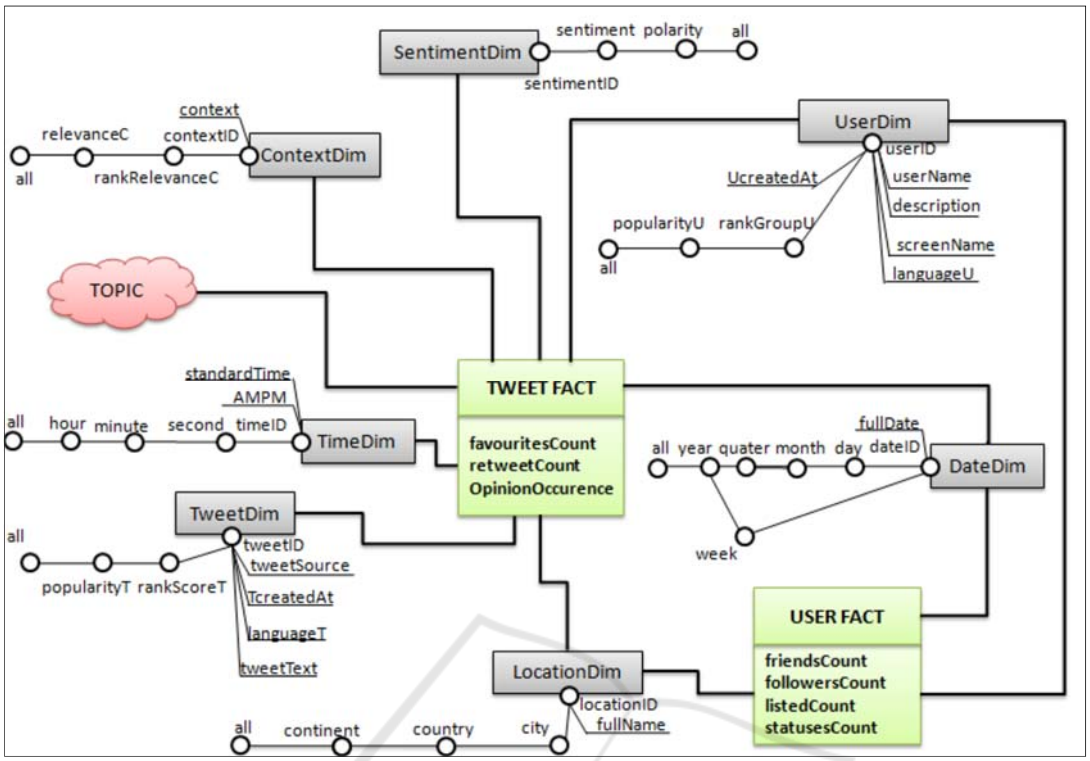


Figure 2: Multidimensional schema for Twitter SDW.

dimension attributes. ETL4Social operations will be detailed in the next section.

# 4 ETL4Social-Data OPERATIONS MODELING

ETL4Social-Data handles the modeling of ETL operations applied on SDS in order to be loaded into the SDW. These operations deals with data crawling from social sources (e.g. twitter). Data crawling is realized according to the analysis context (e.g. internet technology), which is specified by ETL designer. Data sources are then cleaned by means of preprocessing tasks. To be adapted to SDW, a set of standard and semantic ETL operations are applied on data. Finally, they are loaded into the SDW.

## 4.1 Extract Media Clips Data

This process aims to crawl the web for media clips extraction according to the analysis domain specified by ETL designer. Based on keywords related to the study context, a set of queries are then formulated and sent to web crawlers, whose goal is to retrieve media clips in the scope of the subject area. As response to these queries, we obtain a huge number of media clips. Otherwise, these data may include a set of useless media clips, which will be discarded before its storing in the DSA. Figure 3 summarizes media clips crawling operation tasks: query formulation, query execution, discarding useless clips and saving.

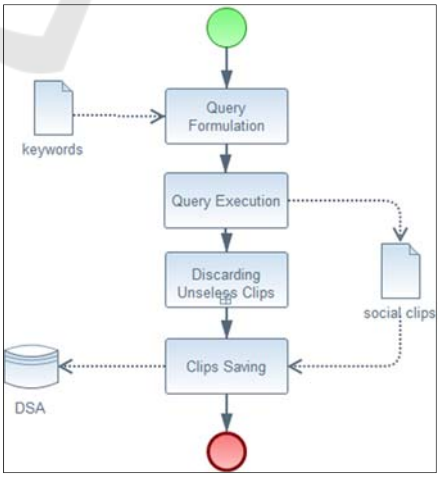


Figure 3: BPMN Model of Media Clips Crawling Operation.

**Example: Twitter Statuses Crawling** is the Media Clips Crawling operation for twitter. Twitter Streaming API is a social crawler used to search for ETL designer queries, a list of statuses that could store information about user, his friends and followers, tweets published by a user, geo-location and date information, etc. The crawling results may store useless information. At this level, we offer to ETL designer tasks that allow him to discard useless statuses. Useless Data Removing activity is presented in figure 3 by a BPMN subprocess element. Figure 4 depicts some tasks including discard retweet statuses or those having very short or non-English text. The semi-structured statuses are then transformed into a structured form and then stored in the DSA.

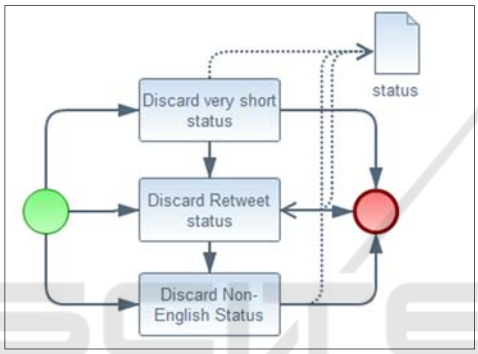


Figure 4: BPMN Expanded Subprocess for Discarding Uselss Statuses Activity.

## 4.2 Preprocessing of Media Clips Data

Detection valuable semantic information from unstructured and informal text requires its cleaning. Preprocessing is a crucial step before performing any standard or semantic processing.

According to the literature, the majority of preprocessing techniques applied on formal text insist mainly on cleaning tasks, i.e. remove diacritics (e.g. accents, unknown characters), URLs, and punctuations and stops words, etc. Otherwise, UGC available on social data is informal and may contain other information that may later cause problems in semantic detection process. For preprocessing operation of social data clips, the existing cleaning techniques are insufficient and should be extended by other tasks according to ETL modeling objectives. **Media Clips Preprocessing** operation is then summarized by the following tasks: Get text field, clean text, remove useless data and save preprocessed text. Figure 5 depicts the generic BPMN model specific to this operation.

**Example: Tweet Preprocessing Operation before Topic Detection.** In order to transform a tweet text into topics in the SDW, we apply the well-known existing cleaning techniques on the tweet text followed by the removing of opinion words, their modifiers and emoticons, which are useless for topic detection. Figure 6 illustrates the tweet text preprocessing tasks before topic analysis process. The preprocessing steps of textual media clips provide structured information enriched by semantic data. The following section details transformation operations in order to reach SDW elements.

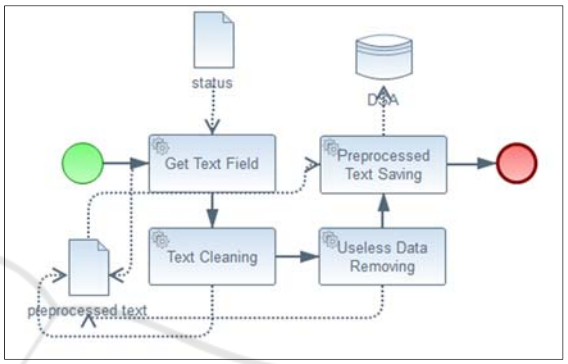


Figure 5: BPMN Model for Media Clips Preprocessing Operation.

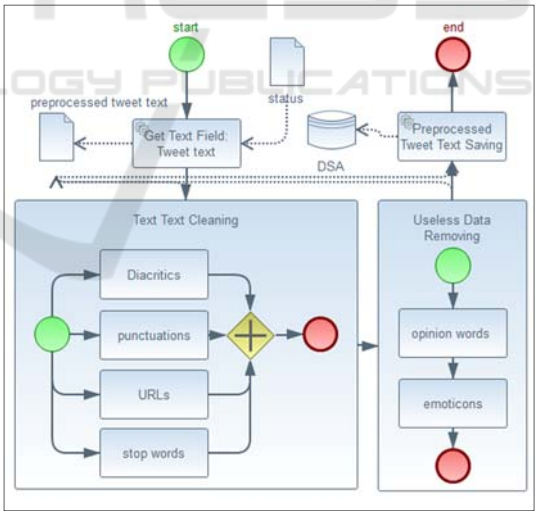


Figure 6: BPMN Model for Tweet Preprocessing Operation before Topic Detection.

## 4.3 ETL4Social Data Transformation Processes

Transforming data from their social sources to the correspondent DWB elements is a complex task. This complexity increases when we think about

transforming social text, considering their dynamic and informal characteristics into semantic dimensions. In this section, we detail the definition and modeling of data transformation operations specific to ETL4Social modeling. Operations are distinguished into standard and semantic operations. Standard means applying atomic or composite transformation rules on data sources to be associated to SDW. For this transformation type, input data are sufficient to determine the target elements. Otherwise, semantic operations are determined by using external resources including domain ontology, and lexical dictionary, which ensure the semantic enrichment of SDS in order to handle SDW opinion elements, such as sentiment polarity and topics.

### 4.3.1 Standard Operations

The objective of standard operation is to map data sources fields to SDW elements. We propose a definition and modeling of standard operations specific to social data.

**Social Standard Transformation** operation aims to associate each stored status fields into the correspondent SDW attribute(s) without the use of external resources. It starts by getting input field(s) and output attribute(s), which are already defined in the correspondence table (CT), ensures the mapping of data sources to the target attributes based on a set of transformation rules and finally it inserts data in the DSA. These steps are shown in figure 7 as the succession of the following activities: get input field(s), get output attribute(s), field-attribute mapping and data insertion.

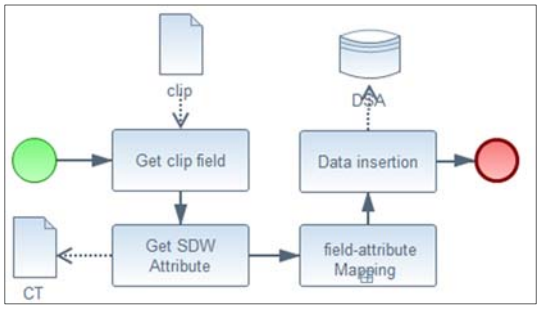


Figure 7: BPMN model for Social Standard Transformation Operation.

**Example: Determine Tweet Popularity** is a social standard transformation operation. It is applied on tweet text field (source) to determine the values of **TweetDim** attributes: **rankScoreT** and **popularityT**. These attributes are absent in the source table (tweet). To detect their values, we define a set of tasks applied on tweet text. Figure 8 details the execution order of these tasks. The value of rank is then calculated based on **favoriteCount**, and **retweetsCount** values following the score formula (1). According to the resulted score, the popularity (pop) is derived based on a set of criteria defined in (Rehman et al., 2013). rank and pop are respectively associated to **rankScoreT** and **popularityT** attributes of the dimension **tweetDim**.

**rankScore= RetweetCount∗80+FavoriteCount∗20** (1)

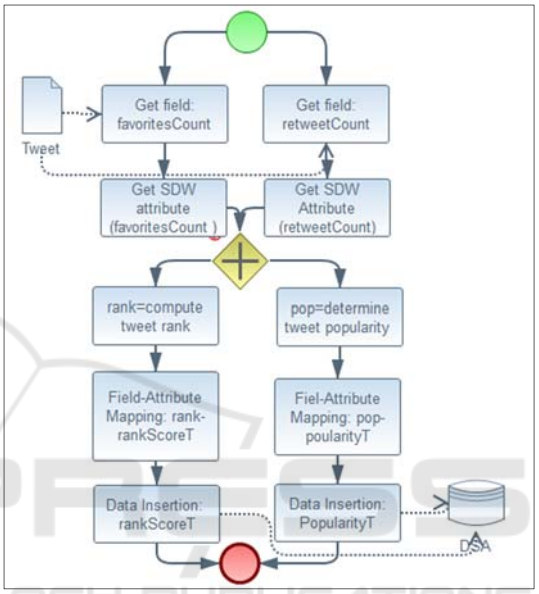


Figure 8: BPMN Model for DetermineTweetPopularity Operation.

### 4.3.2 Semantic Operations

Textual clips available in social data sources are unstructured in comparison to structured fields in business data sources, but it can provide valuable semantic opinion information. In twitter SDW (figure 2), we have three semantic dimensions: **ContextDim**, **TopicDim** and **SentimentDim**. These dimensions data are determined from tweet text, after applying semantic transformation operations.

**Semantic_Social_Mapping** is a transformation operation that occurs when we want to associate semantic data derived from input text to semantic dimensions attributes. These data are determined by querying external resources, like domain ontologies, dictionaries, etc. **Semantic_Social_Mapping** is modeled by the following activities (presented in figure 9):

- Get input text
- Get dimension attributes

- Semantic enrichment: clips enrichment is realized through asking external resources, which detect semantic information of the input text.
- Semantic social mapping: applying transformation rules that determine the instances of dimension attributes.
- Data insertion: associate the obtained data to the correspondent attributes.

In this paper, we propose to define and model the main Semantic Social Mapping operations of the dimension **TopicDim**, which is a dynamic dimension (as described in section 3.2) in comparison to other semantic dimensions. Indeed, many possible hierarchies may be defined in **topicDim**. The absence of this dimension attributes in twitter SDW schema is justified by their variation from one study context to another one. For example, in a commercial context, attributes may be company, product, type product, category product, etc. In politic context, none of those attributes mean anything. In this case, a topic may be people or political event, etc.

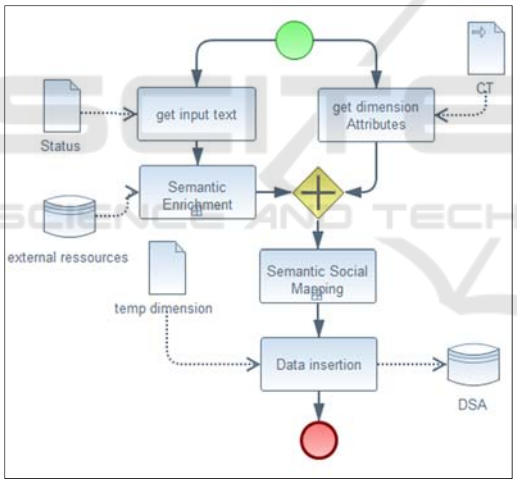


Figure 9: BPMN Model of Semantic Social Mapping operation.

**Topic Semantic Social Mapping** starts by entities detection. After studying the existing platforms, we choose Thomson Reuters Open Calais service (Calais, 2016), which is sophisticated to automatically analyze and tags input text to pinpoint relevant data contained within the text, including products, people, companies, political events, city and country. Topic detection is followed by entities data transformation in order to be affected to **topicDim** attributes, which are based on a Topic Correspondence Table (TCT). Its role is to associate each entities field to its correspondent attribute in the SDW. Table 1 is TCT example that enumerates some detected entities specific to the context "internet technology" and their correspondent elements in **topicDim**.

In the following, we will detail Semantic_Social_Mapping transformation operations for two examples of Topic hierarchies: Product_Brand and Product_Category.

Table 1: TCT example for "internet_technology" context.

| Entity Field | SDW Attribute | Example |
|---|---|---|
| productName | Product | "iphone 5" |
| productType | Type | "Electronics" |
| companyName | Brand | "Apple" |
| operatingSystem | Operating system | "Paris" |

**Example 1: Topic Semantic_Social_Mapping for the hierarchy "Product_Brand".**

According to the hierarchy definition proposed in (Gallinucci et al., 2013), Product_Brand hierarchy is defined as follows:

- Hierarchy Name: Product_Brand
- Levels={Product, Brand} with Product>Brand
- Relation(Product, Brand, hasBrand)

BPMN model specific to **Topic Semantic_Social_Mapping** (figure 10) illustrates the execution order of tasks specific to the hierarchy Product_Brand. The first step is to determine from TCT the entity fields correspondent to Product_Brand attributes, i.e., product and brand. According to these fields, a query is then formulated. OpenCalais service executes this query and returns a list of entities. Tables 2 and 3 respectively enumerate entities examples specific to "product" and "company" types for the study sub-context "smartphone". The next step consists on realizing the mapping between entity fields and hierarchy attributes. For Product_Brand hierarchy, the principle is to search for each product type its correspondent company name, by querying DBpedia ontology. Yet, if the relation "hasBrand" exists between productName and companyName instances, the join between the two entities product and company is recommended. Finally, the next step is the mapping between entity fields and Product_Brand hierarchy attributes. ProductName and companyName are respectively mapped to the attributes **Product** and **Brand**.

Table 2: Examples of entities of type product.

| EntityName | EntityType | ProductType |
|---|---|---|
| iphone 5 | Product | Electronics |
| samsung galaxy J7 | Product | Electronics |
| nokia 3 | Product | Electronics |

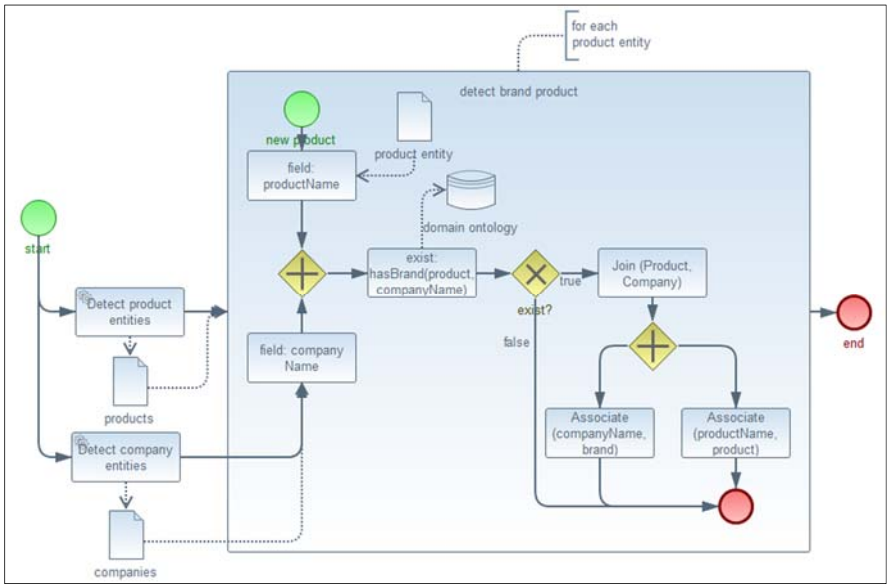


Figure 10: BPMN model for Semantic Social Mapping for Product_Brand.

Table 3: Examples of entities of type company.

| EntityName | EntityType | FullName |
|---|---|---|
| apple | Company | apple Inc. |
| samsung | Company | samsung electronics CO.LTD |
| motorola | Company | motorola solutions, Inc. |

**Example 2: Topic Semantic Social Mapping for the hierarchy "Product_Category"**. The Product_category hierarchy is defined as follows:

- Hierarchy Name: Product_category
- Levels={Product, Category, Type} with Product>Category, Category>Type.
- Relation{R1, R2}:
  R1{(Product, Category), hasCategory}
  R2{(Category, Type), isOfType}

The goal of **Semantic Social Mapping of Product_category** is to associate data detected from tweet text to **Product_category** attributes. The BPMN model proposed for this operation is illustrated in figure 11. As in the example 1, it starts by detecting product entities. Table 2 mentions instances of products. As input, a product is characterized by its name and its type. For the **Product_category**, a product is characterized by three parameter levels, i.e., product, Category and Type. Table 4 mentions some instances of the product_category hierarchy table.

**Semantic_Social_Mapping for the Product_category** aims to realize the mapping between instances mentioned in table 2 (source) and those of table 4 (target). By way of comparison between contents of the two tables, it may be noted that Category information is absent in the first one. This information may be stored in a domain ontology and accessible by applying the relation R1, i.e., hasCategory, to the productName. The resulted value is associated to the parameter InterCat. Then, R2, i.e. isOfType, is then verified between InterCat and ProductType. In case of existence, the mappings between productName and product, InterCat and Category, productType and Type are then permitted.

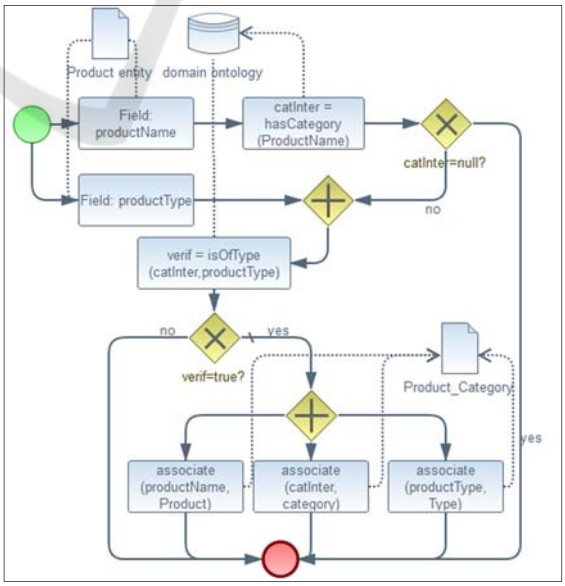


Figure 11: BPMN Modeling of Semantic Social Mapping for Product_category.

Table 4: Example of Product_category hierarchy table.

| Product | Category | Type |
|---|---|---|
| iphone 5 | smartphone | electronics |
| ipad | tablet_computer | electronics |
| samsung galaxy J7 | smartphone | electronics |
| moto 5 | smartphone | electronics |

# 5 ETL4Social OPERATIONS META MODELING

ETL4Social data is a set of Extraction, Transformation and Loading processes applied on social data to reach SDW. A process is composed of data operations. In the previous sections, ETL4Social operations are modelled by BPMN diagrams. Figure 12 depicts the meta-model for ETL processes, which extends BPMN 2.0 meta-model. A BPMN diagram is designed by BPMN base elements that are classified into Flow objects, Connecting objects, Swimlane and Artifacts (BPMN, 2011).

ETL4Social operations are modelled by succession of activities started by a start event and finalized by end event. Activities are categorized into atomic and compound. An atomic activity (e.g. map, convert, join, etc.) are presented as task in BPMN diagram. Otherwise, a compound activity (e.g. Discarding Useless Clips defined in figure 3) is illustrates by a sub-process, which is an **Activity** that encapsulates a **Process** that is in turn modeled by **Activities**, **Gateways**, **Events**, and **Sequence Flows.** In ETL4Social operations models, activities are inter-related through sequence flows. A Sequence Flow is used to show the order in which activities will be performed in a process. Many sequence flows may be related to an activity. The gateway is then used to control the divergence and convergence of sequence flows in an operation process. Thus, it may determine branching, forking, merging, and joining of paths. Gateway internal markers indicate the type of behaviour control, including parallel, exclusive, inclusive, based-event, and complex gateways. During the execution of an ETL4Social operation activity, events are something that may happen during. These Events affect the flow of the model and usually have a cause (*trigger*) or an impact (*result*). There are three of events: start, beginning and end.

In ETL4Social operation, processes often require data in order to be executed. In addition they may produce data during or as a result of execution. Data requirements are captured as **Data Inputs** and **Data Outputs**. In addition to these two types, Data may be data object or data store. **Data Objects** are the primary construct for modeling data within the process flow. **Data Store (e.g. domain ontology)** provides a mechanism for **activities** to retrieve or update stored information that will persist beyond the scope of the process. Data are associated to activities through **Data association** elements, which are connecting objects. Another type of connecting object is association that links information and artefacts with BPMN graphical elements.

# 6 CONCLUSION AND PERSPECTIVES

In this work, we are interested to model ETL operations specific to social environment in which social data are unstructured and informal. This approach offers to ETL designer the opportunity to easily integrate social data into SDW.

ETL modeling is based on the explicit knowledge stored in the design step. The ETL4Social data processes modeling as detailed in this paper greatly minimize the complexity and the time allocated to the implementation of extraction, transformation and loading of social operations.

As compared to existing methods which integrate social data in SDW, the power of ETL4Social approach is demonstrated by the proposed generic models that illustrate a succession of operations related to extraction, transformation and loading of text stream to be integrated in semantic dimensions including topic dimension, sentiment and context dimensions.

Taking into account the complexity of ETL processes, the various aspects of ETL processes (control flow, data flow) should be modeled separately. We aim to handle ETL4Social processes as a combination of control and data processes, where control processes manage the coarse-grained groups of tasks and/or sub-processes, and data processes operate at finer granularity, detailing how input data are transformed and output data are produced.

ETL4Social-Data handles the modeling of ETL operations applied on social data sources in order to be loaded into the SDW having static elements. Yet, a SDW schema may contain dimensions having dynamic and fluid structures such as the topic dimension presented in the twitter case study. In the near future, we aim to propose an ETL4Social-schema that ensures the conceptual modeling of ETL operations dealing with SDW schema transformation.

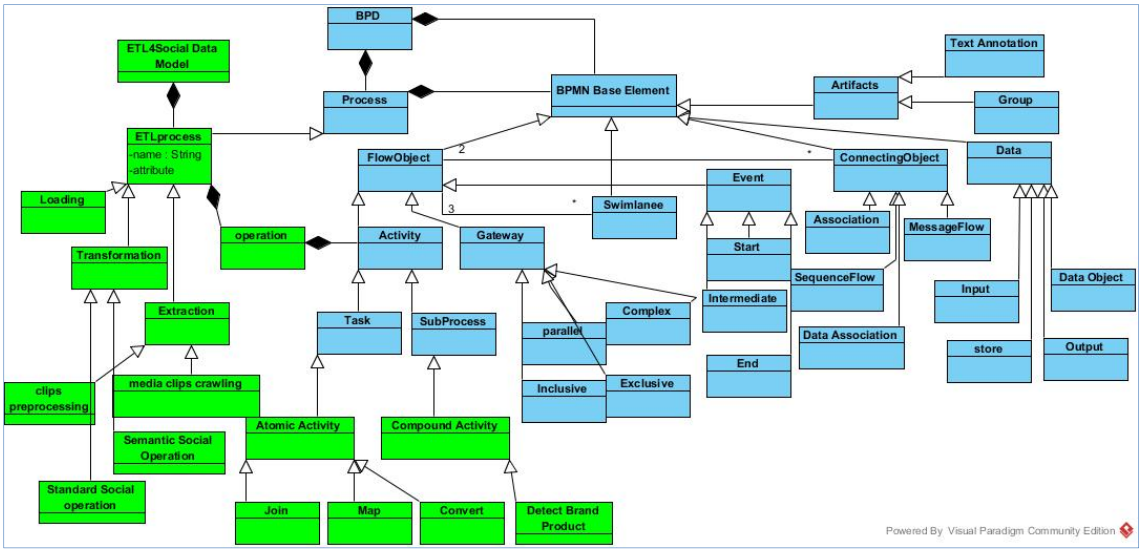


Figure 12: Class Diagram of ETL4Social Operations Meta-Model.

# REFERENCES


Berlanga, R., García-Moya, L., Nebot, V., Aramburu, M.J., Sanz, I., Llidó, D.M., 2015. SLOD-BI: an open data infrastructure for enabling social business intelligence. In *International Journal on Data Warehousing and Data Mining*.

BPMN, Business Process Modeling and Notation, http://www.omg.org/spec/BPMN/2.0, January 2011.

Calais, Thomson Reuters Open Calais, User Guide, Mars 2016.

El Akkaoui, Z., Zimányi, E., 2009. Defining ETL worfklows using BPMN and BPEL. In *DOLAP'09, ACM Twelfth International Workshop on Data Warehousing and OLAP*.

El Akkaoui, Z., Mazon, J.N., Vaisman, A., Zimanyi, E., 2012. BPMN-based conceptual modeling of ETL processes. In *DaWaK'12, 14th International Conference on Data Warehousing and Knowledge Discovery. Information technology*.

Gallinucci, E., Golfarelli, M., Rizzi, S., 2013. Meta-Stars: Multidimensional Modeling for Social Business Intelligence. In *DOLAP'13, sixteenth international workshop on Data warehousing and OLAP*.

García-Moya, L., Kudama, S., Aramburu, M.J., Berlanga, R., 2013. Storing and analysing voice of the market data in the corporate data warehouse. In *Information Systems Frontiers*.

Jovanovic, P., Romero, O., Simitsis, A., Abell´o, A., 2012. Requirement-driven creation and deployment of multidimensional and ETL designs. In *ER'12 Workshops. 31st International Conference on Conceptual Modeling*.

Mallek, H., Walha, A., Ghozzi, F., Gargouri, F., 2014. ETL-web process modeling. In *ASD'14, Advances on Decisional Systems Conference.*

Muñoz, L., Mazón, J.N., Pardillo, J., Trujillo, J., 2008. Modeling etl processes of data warehouses with uml activity diagrams. In *On the Move to Meaningful Internet Systems: OTM 2008*, Lecture Notes in Computer Science, Springer.

Muñoz, L., Mazón, J.N., Trujillo, J., 2010. A family of experiments to validate measures for uml activity diagrams of etl processes in data warehouses. In *Information & Software Technology*.

Petrovic, M., Vuckovic, M., Turajlic, N., Babarogic´, S., Anicic´, N., Marjanovic´, Z., 2016. Automating ETL processes using the domain-specific modeling approach. In *ISBM'16, Information Systems and e-Business Management.*

Rehman, N., U., Weiler, A., Scholl, M.H, 2013. OLAPing Social Media: The case of Twitter. In *ASONAM'13, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*.

Shin, Y., Ryo, C., Park, J., 2014. Automatic extraction of persistent topics from social text streams. In *WWW'14, 23rd International World Wide Web Conference.*

Trujillo, J., Luján-Mora, S., 2003. A UML Based Approach For Modeling ETL Processes in Data Warehouses. In *ER'03, 22nd International Conference on Conceptual Modeling*, pages.

Vassiliadis, P., Simitsis, A., Skiadopoulos, S., 2002. Conceptual modeling for etl processes. In *CIKM'02, 5th ACM international workshop on Data Warehousing and OLAP.*

Walha, A., Ghozzi, F., Gargouri, F., 2016. A Lexicon approach to multidimensional analysis of tweets opinion. In *AICCSA'16, 13th ACS/IEEE International Conference on Computer Systems and Applications.*

Wilkinson, K., Simitsis, A., Castellanos, M., Dayal, U., 2010. Leveraging business process models for etl design. In *ER'10 WorkShop, 29st International Conference on Conceptual Modeling.*