PUBLIC

Feedback mechanisms Across the Lifecycle for Customer-driven Optimization of iNnovative product-service design
Acronym: FALCON Project No: 636868
FoF-05-2014 Duration: 2015/01/01-2017/12/31

PROJECT DELIVERABLE 2.2:

Conceptual approach for knowledge acquisition from social media related to product usage information

Content: This deliverable presents a conceptual approach for knowledge acquisition from social media related to product usage information. The concept is described as a social media wrapper as a part of the FALCON VOP. The presented conceptual approach is derived from requirements, which address the interoperability with FALCON-specific social media data sources. The interoperability approach presented focusses on the extraction of information from continuous, unstructured text and the transformation of the knowledge contained within it into an ontology.

Versioning and contribution history

Version	Description	Contributors
0.1	Draft outline	Marco Franke (BIBA)
0.2	First complete draft	Quan Deng, Marco Franke (BIBA)
0.3	Adapted according to review results (Karl Hribernik)	Quan Deng, Marco Franke (BIBA)
1.0	Adapted according to review results (Panagiotis Gouvas)	Quan Deng, Marco Franke (BIBA)

Reviewer

Name	Affiliation
Panagiotis Gouvas	UBITECH LTD
Karl Hribernik	BIBA

Table of contents

List of Figures

List of Tables

List of Source Codes

Acronyms and Abbreviations

Falcon VOP	FALCON Virtual Open Platform

1 Introduction

1.1 Objective of T2.2

The overall WP2 is in order to reflect product usage information (PUI) from the middle-of-life (MOL) phase of the product-service lifecycle to the design phase. Two types of PUI sources are considered in this work package: sensor-related information from PEIDs (Product Embedded Information Devices) and in social media.

D2.2 documents the results of Task 2.2 and comprises concepts for acquiring product-service-related consumer feedback from social media and other channels such as helpdesk communications, in addition to concepts to filter and organize the acquired information. The extraction of relevant product usage information typically deals with retrospective user feedback from the actual usage of a product, service or product-service. While individual examples of retrospective user feedback in social media are in most cases of a subjective nature, the amount of users in social media allows the extractions of more sophisticated feedback, if filtered properly. For example, filters can ensure that PUI from only highly recommended users is taken into consideration. Such filter mechanisms are already well established in IT, usually covered under the topic of collective intelligence. However, the corresponding techniques need to be evaluated and adapted to the FALCON requirements, especially those generated by the industrial end-users. In addition, interrelations between the PUI identified and the semantic model of the PSS domain and the end-users' specific knowledge domains have to be identified.

The conceptual approach documented here will be implemented as Social Media Wrapper in Task 2.3 prototypically and will be delivered as D2.3. This wrapper will allow the acquisition of data and semantic interoperability with social media data sources. The main goal of the Social Media Wrapper is to extract PUI from social media and to map it onto an ontology, in FALCON specifically onto the FALCON Ontology (as documented in D3.2). The Social Media Wrapper will instantiate Abox of the FALCON ontology with the PUI extracted from social media.

1.2 Content of the document

The content of this document includes an investigation into potential sources of PUI in social media, the state-of-the-art in knowledge extraction from social media, the functional requirements for the Social Media Wrapper and the resulting conceptual approach. The conceptual approach for the Social Media Wrapper takes into consideration the state-of-the-art as well as the functional requirements of the FALCON business cases and presents an innovative approach to the semantic interoperability with PUI which is in line with the overall FALCON approach and tailored to meet the requirements of the (re-)design of PSS in industry.

Chapter 2 contains a brief description of social media including a definition, different kinds of social media data sources as well as available meta-information. The meta-information describes the context of the social media content which is of vital importance to the semantic interoperability process. The state-of-the-art according to the extraction of knowledge from social media is presented in Chapter 3. Subsequently, Chapter 4 contains the classification of social media data sources in the FALCON business cases. On the basis of the classification results, the requirements for the Social Media Wrapper are presented. The conceptual approach of the Social Media Wrapper is described in detail in Chapter 5 considering the state-of-the-art (Chapter 3) and the requirements (Chapter 4). Finally, a conclusion is given in Chapter 6.

1.3 Motivation

Social media is an ongoing trend in which consumers can publish information of their daily life on the Internet. This increasing amount of information includes such that can be considered consumer feedback about products. This kind of information is a high-value information source for the manufacturer. The following facts demonstrate the potential of this kind of data sources.

Kaplan pointed out that *"...According to Forrester Research, 75% of Internet surfers used "social media" in the second quarter of 2008 by joining social networks, reading blogs, or contributing reviews to shopping sites; this represents a significant rise from 56% in 2007. The growth is not limited to teenagers, either; members of Generation X, now 35–44 years old, increasingly populate the ranks of joiners, spectators, and critics. It is therefore reasonable to say that social media represent a revolutionary new trend that should be of interest to companies operating in online space—or any space, for that matter...."* [Kaplan & Haenlein 2010]. The term 'blogosphere' defined Brooks as "…the most commonly-used term for the space of blogs as a whole…" [Brooks et al. 2006] The current available 'blogosphere' is more than 100 million blogs and their interconnections have become an important source of public opinion [Kietzmann et al. 2011]. In the application of micro-blogging, the leading company Twitter has achieved more than 145 million users who send more 90 million 'tweets' per day, each consisting of 140 characters or less [Madway 2010].

Consequently, a huge amount of information is available and is related to a wide range of consumer products and corresponding services. The data to be used contains product specific feedbacks including usage scenarios, technological hurdles from the perspective of the daily usage, best practices or the quality of product-specific services like individualisation or maintenance. Thus, the data and the knowledge contained in it has a high value for manufacturers and could be used to improve not only the product but also the overall product service system. To achieve such an impact, the relevant knowledge must be found, extracted and aggregated. It must also be provided in a suitable way to the processes and IT systems involved in PSS (re-)design. This means that PUI from social media needs to be made interoperable with these knowledge domains and IT systems. A corresponding approach to the interoperability of PUI with these processes and IT systems should be automated to be useful, because manual search and extraction by employees would require a lot of effort which is not economically feasible. By using PUI coming from PEIDs and social media in processes throughout the product lifecycle, low-cost usability test results for the short and long-term usage of consumer products can be facilitated. The evaluation of this PUI will enable the confirmation or falsification of assumed user roles, user behaviour, and product usage scenarios by the designers. The extracted PUI could be used to develop products which are closer to how the customer will use them in reality and therefore, to develop products which are designed to be more attractive for the customers.

The extraction of knowledge from social media is a challenging task, because the sources contain a wide range of different kinds of data structures and types ranging from continuous text to multimedia. The extraction of knowledge from these set of data sources and the corresponding transformation into formalised knowledge is challenging, but could be applied to a wide range of analysis, simulation and optimization methods. The proposed Social Media Wrapper approach will implement the aforementioned knowledge extraction capabilities.

2 Social Media

2.1 Social Media Overview

The technical foundation of social media is Web 2.0 contains e.g. the key technologies, HTML 5, RSS, AJAX etc. [Kaplan & Haenlein 2010]. The listed technologies help the integration of user-generated content (UGC) in web sites. The UGC also contains the user experiences according to the expectations and the usage of a product. The latter one are PUI and are in the focus of the FALCON specific data acquisition. Thus, an average internet user is able to add his knowledge according to facts and his experiences in web sites and to share it with other users. Sharing and consuming blogs, tweets, Facebook entries, movies, pictures, and so forth are common in social media channels. Kaplan summarized the different types of social media into the following four groups, which are illustrated in Table 1 and described in the following.

Table 1 Classification of social media by social presence/media richness and self-presentation/self-disclosure [Kaplan & Haenlein 2010]

		Social presence/ Media richness		
		Low	Medium	High
Self-presentation/Self-disclosure	High	Blogs	Social network sites (e.g. Facebook)	Virtual social worlds (e.g. Second Life)
	Low	Collaborative projects (e.g. Wikipedia)	Content communities (e.g. YouTube)	Virtual game worlds (e.g. World of Warcraft)

Collaborative projects

Collaborative projects enable the creation of content by many end-users. Thus, the UGC covers not only the creation but also the evaluation and improvement of the contained content. Examples of collaborative projects are Wikipedia offering an encyclopaedia or Schema.org offering common ontologies. The specific content varies from data source to data source and therefore, it could contain any kind of data covering different degrees of formality.

Blogs

Kaplan mentioned, *"Blogs, which represent the earliest form of social media, are special types of websites that usually display date-stamped entries in reverse chronological order"*[Kaplan & Haenlein 2010]. Each added blog entry contains the user comment as continuous text, which could be linked to other kind of media sources like video, sound or images. The continuous text is represented as natural language and offers the lowest degree of formality. The semantic content of the continuous text is not reviewed or checked according to spelling and grammar. In consequence, the information quality and correctness is not guaranteed in blogs. Kaplan mentioned that most of the content relies on complaints according to a specific product or to a company. In the latter case, employees or ex-employees are mostly the corresponding authors [Kaplan & Haenlein 2010].

Content communities

The main goal of content communities is to share media content between users. The kind of media content depends on the specific data source and could range from continuous texts (book sharing), sound files (music sharing), semi-structured/structured text (specifications sharing), images (fashion trends sharing) or videos (sharing films, home videos, etc.). The primary focus relies on the media sharing functionality. Therefore, the authors profile and additional information (meta-information to the content) are available which are not always reviewed and not so detailed. The extraction of the contained knowledge carries the risk of being used as platforms for the sharing of copyright-protected materials.

Social networking sites

Social networking sites focus on the representation of user information instead of representing facts like in blogs or collaborative projects. Moreover, the communication with other persons lies in the foreground. For that purpose, messages and e-mails are the main communication channels. The user is allowed to create comments and the classification in like/don’t like to any topic. These kinds of information and the user profile information are available. Therefore, this kind of information is in the focus of a knowledge extraction.

The group **Virtual game worlds** and **Virtual social worlds** will not be presented in detail, because the groups contain no PUI and therefore, have no relevance for the proposed FALCON business cases. It consequence, neither have any impact on the Social Media Wrapper conceptual approach. The above-mentioned social media groups contain information in any kind of representation form and degree of formality formalism.

The aforementioned social media related data sources differ not only in the representation of the knowledge and in the different degree of formality but also in additional properties. The Table 2 summarizes the additional properties of the social media groups, which are important for the knowledge extraction process of the Social Media Wrapper.

Table 2 Properties of social media groups

Blog group	Representation	Type of Media	Chronological ordering	Underlying terminology	Syntax checking (spelling, grammar)	Semantic checking (Review)
Collaborative projects	Informal - formal	Text	Only for Versioning	- domain specific - consistent	Y	Y
Blogs	Informal - formal	Mainly text + other media	Y	- no - not consistent	N	N
Content communities	Informal-formal	All	Y	-no -not consistent	N	N
Social networking sites	Informal-semi-formal	Mainly text + other media	Y	-no -not consistent	N	N

Table 2 summarizes that not all data sources are reviewed or use a common terminology within the data source. Both criteria are important to enable the extraction of facts (not only assumptions) and enable a corresponding unambiguous interpretation. Achieving an unambiguous interpretation is not straightforward, if different users apply the same concept for different things. In such a case, additional context-sensitive knowledge is necessary. In general, the extraction of knowledge from social media is only possible completely if all necessary information is contained in the social media data source on its own or could be included from external sources. Table 3 summarizes for each kind of common representation form whether the semantic description is completely contained in the social media related data sources. Text-based content is listed, because available PUIs are represented mainly as snippets of natural language in social media. Especially, product reviews, complaints or usage scenarios are written in his/her natural language.

Table 3 Kinds of social media data sources and their relation to semantic descriptions

Kind of data source	Semantic understanding[1]	Semantic understanding[2]	Structural Coverage [3]	Common data source
Tables/Database (Relational) (e.g. Product specification)	Y	N	Nothing	Y
Plain text as part of web page (e.g. Complaints of a product)	Y	N	Nothing	Y
Ontology represented as files or in RDF database (e.g. Open Linked Data)	Y	Y	Hierarchical, taxonomical object oriented data structures including primitive data types and instance specific data types	N
Source Code in a programming language (e.g. Forum for developers, GitLab)	Y	Y	Hierarchical, taxonomical object oriented data structures including primitive data types and	Y
Web Service (e.g. Access to sensors, Social networks (like Facebook))	Y	Y, the WSDL contains schema	Hierarchical Object Oriented data structures including primitive data types und simple relationships between data structures	Y
XML files in a repository (e.g. Shared configurations, models)	Y	Y, if the XSD schema is also available	Hierarchical Object Oriented data structures including primitive data types	N

Note:
[1]: Semantic understanding is possible for syntax of the language
[2]: Semantic understanding is possible for modelled in-formation in the language
[3]: Structural Coverage of semantics for information modelling

Apart from ontologies, not all of the data sources listed are semantically described, or a semantic description is unavailable. To transform continuous text from social media external information sources containing the missing semantic descriptions are necessary. Moreover, context sensitive meta-information is necessary to determine the correct interpretation of continuous text's token. An opportunity as well as a hurdle are the author's properties of social media content, which are presented in the following.

2.2 User Context as Barrier towards Knowledge Extraction

The extraction of information from legacy systems or company specific documents focus on the extraction of codified information in the documents. The common extraction methods have the assumption that the found knowledge is true and meaningful for a given context. Neither of these assumptions is reliable in the context of social media. The intention of publishing something is wider than just information sharing. To consider this, Kietzmann mentioned the honeycomb of social media that is illustrated in Figure 1 [Kietzmann et al. 2011].

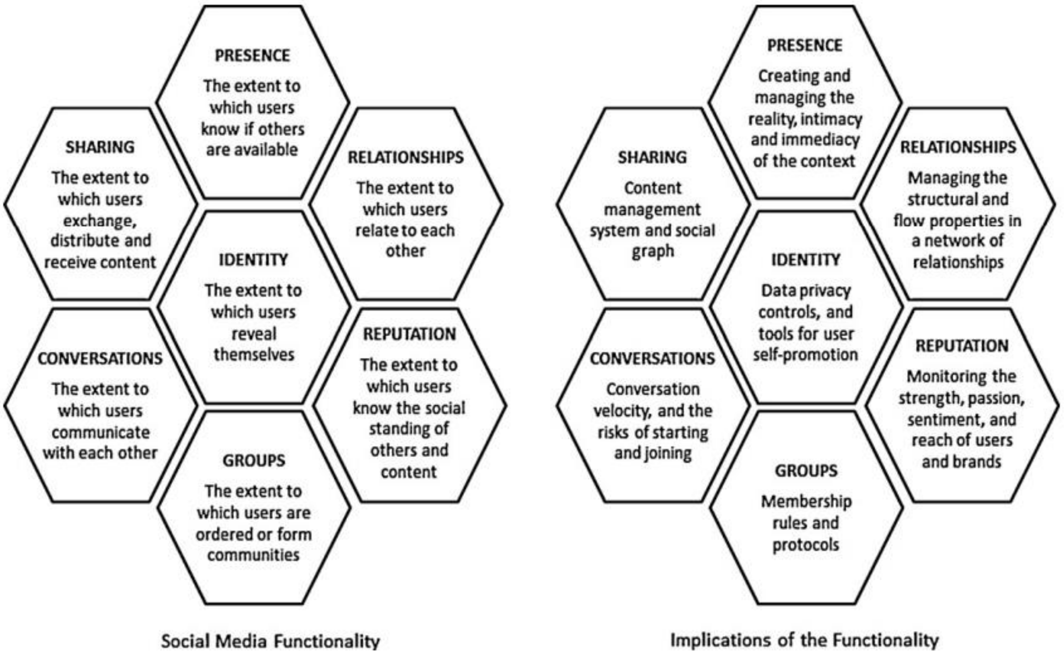

PRESENCE
The extent to which users know if others are available
SHARING
The extent to which users exchange, distribute and receive content
RELATIONSHIPS
The extent to which users relate to each other
IDENTITY
The extent to which users reveal themselves
CONVERSATIONS
The extent to which users communicate with each other
REPUTATION
The extent to which users know the social standing of others and content
GROUPS
The extent to which users are ordered or form communities
Social Media Functionality
PRESENCE
Creating and managing the reality, intimacy and immediacy of the context
SHARING
Content management system and social graph
RELATIONSHIPS
Managing the structural and flow properties in a network of relationships
IDENTITY
Data privacy controls, and tools for user self-promotion
CONVERSATIONS
Conversation velocity, and the risks of starting and joining
REPUTATION
Monitoring the strength, passion, sentiment, and reach of users and brands
GROUPS
Membership rules and protocols
Implications of the Functionality

Figure 1 Honeycomb of social media [Kietzmann et al. 2011]

Each honeycomb must be considered if the verisimilitude of an extracted information item is executed. Each honeycomb summarizes how the extracted information could be influenced by the user's background. The impact of the user's background could be included by the user implicitly which means that he is not aware to have added a subjective and not objective knowledge. To transform the subjective knowledge to an objective one, the information of these 7 honeycombs must be applied to reinterpret or at least mark the intention of the shared knowledge. The access to the user's context is not possible in a complete manner.

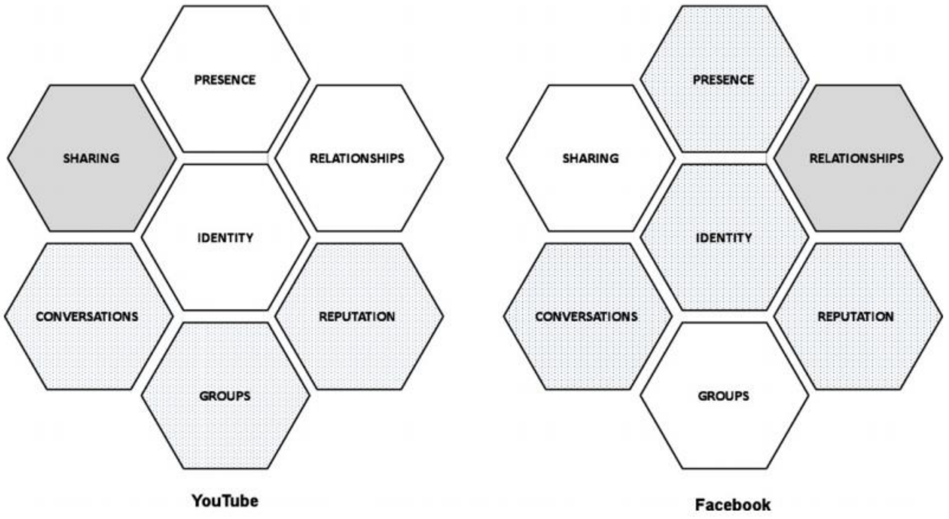


Figure 2 Honeycombs of Youtube and Facebook [Kietzmann et al. 2011]

Each social media data source focus on specific honeycombs and therefore, the completeness is not given. An example from Kietzmann is given in Figure 2 [Kietzmann et al. 2011].

3 State-of-the-art

The principle idea behind acquiring product related consumer feedback from social media or sources similar to social media such as a helpdesk is to automatically extract specific information and to transform it into a higher degree of formality. The higher the degree of formality, the more downstream analysis methods are applicable. Since social media has been popular for customers to share their opinions towards products, to unlock the products associated facts as well sentiment information embedded in the amount of social media data is critical for enterprise in the transformation process. A common target representation form is an ontology. The extraction and transformation can be realised through methods of Natural Language Processing (NLP). Various sub-research areas such as Ontology Based Information Extraction (OBIE), Opinion Mining and semantic interoperability are related. For example, if information regarding customers' sentiment towards a product are required, methodologies of opinion mining should also be considered. In the following, the state-of-the-art analysis focusses on ontology-based information extraction, opinion extraction and how the approaches are combined to gain a holistic approach to extract both factual and subjective information for products.

3.1 Ontology-Based Information Extraction

Information extraction (IE) is in principle a subfield of NLP. The general process of IE is based on automatically retrieving certain types of information from natural language text [Wimalasuriya & Dou 2010]. According to Riloff, IE is a form of natural language processing in which certain types of information must be recognized and extracted from text [Riloff 1999]. While the recently emerged term OBIE is once again a subfield of IE, it is described as a system that processes unstructured or semi-structured natural language text through a mechanism guided by ontologies to extract certain types of information and presents the output using ontologies [Wimalasuriya & Dou 2010]. According to these definitions, the main differences between OBIE and traditional IE is the former's linkage to semantic ontology.

Researches in NLP and in particular in the subfield of IE have adopted various approaches including machine learning approaches like Latent Dirichlet Allocation (LDA) [Blei, Ng & Jordan 2003] and syntactic as well as rule/pattern-based approaches to build systems that extract certain information from natural language documents [Muslea 1999, Grishman 1997, Cowie & Lehnert 1996]. These approaches often rely on linguistic processing techniques for understanding the linguistic structures of sentences. The pre-processing methods listed in Table 4 are common in the IE systems. Corresponding NLP tools are: GATE[1], UIMA[2], RapidMiner[3], etc.

[1] https://gate.ac.uk/

[2] https://uima.apache.org/

[3] https://rapidminer.com

Table 4 Pre-Processing methods for NLP

Pre-Processing Method	Description
Tokenizer	Ensures the segmentation of the input texts into simple tokens such as words, punctuation and numbers
Sentence Splitter	This method splits the input texts into sentences
Stemming and lemmatization	Used to reduce inflectional forms and derive a word's common base form. This is necessary to achieve the similarity of inflectional forms and therefore, to enable the similarity analysis of sentences using different grammar.
Part of Speech (POS) Tagger	This process associates each word form to its corresponding particular part of speech to allow preliminary recognition of the phrasal units. This is necessary to enable the detection of relations between subject and objects in sentences. Moreover, the extraction of the adjectives and preposition enables the sentiment analysis.

Apart from aforementioned pre-processing, for the information extraction in the field of OBIE, many OBIE systems have been developed over the last years. One of the rule-based approaches was established by Embley et. al. which is based on the formulization of rules to extract constants and context keywords from unstructured documents based on domain specific ontology [Embley et al. 1998]. Another more sophisticated approach was the KIM semantic annotation platform, which provides services and infrastructure for information extraction as well as semantic annotation, indexing and retrieval based on GATE [Popov et al. 2004]. One additional example is the PANKOW system which semantically annotates given web pages using an unsupervised, pattern-based approach to categorize instances with regard to an ontology [Cimiano, Handschuh & Staab 2004]. SOBA is another solution which focusses on an ontology-based information extraction

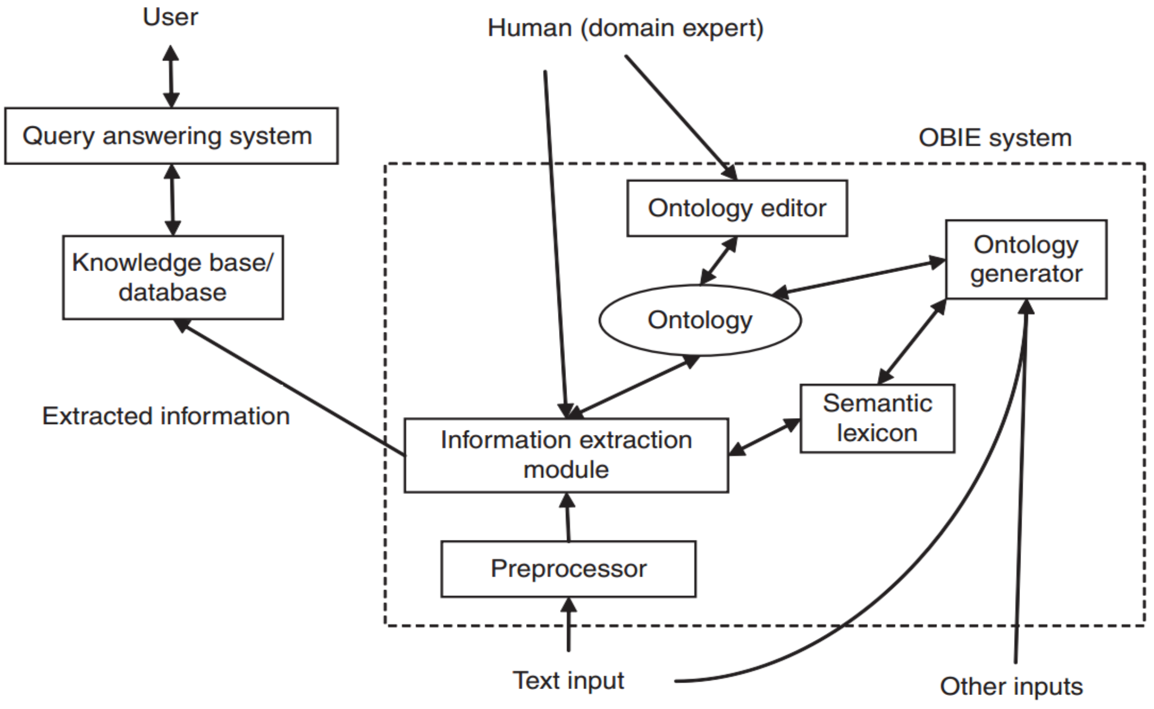


Figure 3 General Architecture of OBIE system by Wimalasuriva and Dou

mechanism from soccer web pages. The extraction goal focussed on an automatic population of a knowledge base that can be used for domain specific question answering. The above mentioned tools demonstrated the wide applicability for the ontology and rule based information extraction. A systematic survey on the OBIE field including a general architecture of OBIE system was given by Wimalasuriya and Dou, which is shown in Figure 3 [Wimalasuriya & Dou 2010]. As an increasing number of existing OBIE system may cause a problem with selection the most suitable solution, KONYS have provided an approach for OBIE system selection and evaluation [KONYS 2015].

In the following, the main information extraction methods for the general architecture of OBIE are described in detail.

Gazetteer lists

This technique is used to identify individual entities of a particular semantic class usually directly via look-up from a word list, known as gazetteer list. For example, gazetteer list can be used to recognize all countries of the world. It is widely used in information extraction systems for named entity recognition.

Linguistic rules

The general idea behind this technique is to specify various grammar rules. A grammar rule is defined on basis of regular expressions, gazetteer list and the part of speech to extract specific ontological aspects.

Analysing tags

This technique makes the usage of tags such as html/xml tag from the document to extract specific information. It is very useful to extract information from texts with a number of meaningful tags. The mapping of XML tags to ontology aspects is a common approach to transform xml files to Abox.

Machine Learning Techniques

Various machine learning techniques have been used for the purpose of information extraction, such as support vector machines. For supervised machine learning approaches, many different linguistic features such as POS tags and syntactic dependencies could be selected and used for the training. For example, the KYLIN OBIE system uses a large set of features including POS tags for the training [Wu & Weld 2007].

In practice, it depends on various factors which kind of rules could be applied in which application scenario. For instance, for the extraction of attribute values of a hotel ontology, Anantharangachar et al. proposed an extraction methodology which is in principle based on handcrafted extraction rules which are based on regular expression [Anantharangachar, Ramani, & Rajagopalan 2013]. While Gutierrez et al. proposed a hybrid OBIE system which consists both extraction rule based information extractor and machine learning based information extractor [Gutierrez et al. 2015].

The Social Media Wrapper approach foresees the application of the mentioned OBIE rule types. It will enable a rule system in which different kind of rules could be combined to satisfy the case specific heterogeneity as well as the different requirements. The application of OBIE rules will enable the extraction to the level of single word, which has a specific grammatical meaning and a defined relation to other parts of the sentence. Moreover, the linkage of extern knowledge will enable the detection of predefined meanings for acronyms, ids and other unique identification.

3.2 Opinion Mining

While the focus of OBIE is the extraction of factual information from text, opinion mining is also known as sentiment analysis. Here, methodologies are employed to meet the demands for detecting opinions and sentiments. Opinion mining has become a popular topic in recent years, and much effort has been invested in this research area. An early extensive survey on opinion mining and sentiment analysis is performed by Pang and Lee, covering techniques and approaches that promise to directly enable opinion-oriented information-seeking systems [Pang & Lee 2008]. Lee presented another more recently survey of all important research topics in this field, in this survey a formal definition of the objective of sentiment analysis is defined [Liu 2012]. Following the definition, the necessary tasks as well as related techniques to be performed to achieve the objective are mentioned. Medhat et al., Schouten and Frasincar have given short surveys illustrating the new trends in opinion mining [Medhat, Hassan & Korashy 2014, Schouten & Frasincar 2016].

With regards to the dimension of granularity, the following three levels of sentiment analysis are often differentiated: document-level sentiment analysis, sentence-level sentiment analysis and feature-based sentiment analysis. Document-level sentiment analysis considers a whole document as an information unit. It is usually under the assumption, that whole document focuses on a single product, and the analysis is supposed to get opinions towards the product as general. If more than one product is mentioned in the input text (e.g. blogs), the analysis would be unable to detect the sentiment towards an individual product. Sentence-level sentiment analysis takes a sentence as an information unit and tries to find out the sentiment expressed in the sentence. It is not suitable when more product features are mentioned in a single sentence. Feature-based sentiment analysis is in a more fine-grained level, which is dedicated to extracting opinions about an individual product feature.

For the sentiment analysis in different granularity level, various sentiment classification methodologies are preferred. In general, two main research directions, i.e. machine learning approaches and lexicon based approaches, have been focused on in the recent years. Machine learning approaches are categorized into unsupervised learning approaches and supervised learning approaches which can be then detailed into the different learning algorithms and various text features. With supervised learning approaches, the input text is then based on the text features and trained models classified into positive, negative or neutral. Saleh et al. carried out sentiment analysis applying Support Vector Machines (SVM) in different domains with different features. In lexicon-based approaches, the analysis is often based on lexicons such as SentiWordNet [Baccianella, Esuli & Sebastiani 2010] or WordNet-Affect [Valitutti & Strapparava 2004] and some defined opinion rules. For example, when there is a negation word “no” before positive word “good”, the sentiment of the expression is then negative. Salas-Zarate et al. proposed an approach for feature-based opinion mining in financial news based on SentiWordNet lexical resource, in which polarity of the features in each document is calculated by taking into account the words from around the linguistic expression of the feature [Salas-Zarate et al. 2016]. In many cases, both machine learning approaches and lexicon based approaches could be employed. For example, for the sentiment analysis on microblogs e.g. Twitter, Li et al. proposed an approach using SVM classifier to classify sentiment and finally derive market intelligence from microblogs [Li & Li 2013]; while Zhang et al. presented an approach combined both lexicon-based and learning-based Methods for twitter sentiment analysis [Zhang et al. 2011].

The Social Media Wrapper approach foresees the extension of the mentioned OBIE rule types with the lexicon-based approach. The combination of both rules engines will allow the Wrapper to derive facts and evaluate the verisimilitude on the basis of sentiment analysis. In consequence, it will enable a rule system in which different kind of rules could be combined to satisfy the case specific user heterogeneity as well as the different requirements.

3.3 Integration of OBIE and Opinion Mining

Both OBIE and opinion mining are highly active research fields. Considerable research has been conducted in recent years to take advantage of ontology to help feature-based opinion mining. Ontologies provide a formal structure knowledge representation as well as common vocabularies for a domain. It helps the identification of product features. In the movie domain, Zhao and Li proposed an ontology-based approach for opinion mining, in which they use the ontology structure as an essential part of the feature extraction process, and then use lexicon-based approach for sentiment analysis [Zhao & Li 2009]. In financial domain, Salas-Zarate et al. has been addressed the problem of feature-based opinion mining with the help of financial ontology [Salas-Zarate et al. 2016]. For the language Portuguese, Freitas et al. have conducted a research to identify polarity in Portuguese user generated reviews according to features described in domain ontologies [Freitas & Vieira 2013]. Considering domain independent approaches, Peñalver-Martinez et al. presented another feature-based opinion mining approaches which leverage knowledge technologies that supposed to be applied to different languages (i.e. English and Spanish) and in different domains [Peñalver-Martinez et al. 2014]. In the approach, features are identified with the help of ontology, four different configurable methods e.g. "N_GRAM Around" are introduced to get the words that are close to the feature, these words are then used to calculate sentiment polarity of the feature based on SentiWordNet.

Research has also been conducted into combining information extraction with opinion mining to extract both factual and subjective information. Saggion et al. provide a practical solution to track the reputation of a company by identifying factual and subjective information for business intelligence [Saggion & Funk 2009]. They use information extraction technology to extract company facts from multiple sources and opinion mining techniques based on supervised machine learning technology to identify positive and negative texts and fine-grained sentiment classification. In the domain of restaurant reviews, a corpus-based information extraction and opinion mining method are proposed for Russian, it uses machine learning techniques and is based on elaborate corpus analysis and automatic classifier selection [Pronoza , Yagunova & Volskaya 2014]. However, little research has looked into integrating OBIE and ontology-based feature-level opinion mining to extract both factual and subjective information to fill Abox entries for a given ontology which can be applied in various data sources from different domains (i.e. business domains in FALCON) and to different languages (i.e. English, Germany and Turkish).

The Social Media Wrapper approach foresees the same strategy to combine the factual and the sentiment extraction and therefore, to enable a holistic data integration approach including evaluation mechanism.

4 Requirement Analysis

The goal of this chapter is to derive functional requirements for the Social Media Wrapper, which cover the FALCON-specific social media data sources as well as the general properties of social media. In the following, the requirements according to the Business Scenarios are presented in detail.

The first stage in developing the FALCON VOP consists of all the activities aimed to the identification of user domain context and needs, in the analysis of these user needs to drive additional requirements, and in the documentation and validation of the requirements as specification document. The result of this requirement analysis are listed in the deliverables 5.1 – 8.1 which are specialized from the requirement FALCON_MOL8. In the following, the identified relevant data sources and corresponding social media specific types are listed in Table 5. The listed data sources and their properties will be evaluated to define the functional requirements of the Social Media Wrapper concept. The cross-data source requirement is that PUI should be extractable and could be added as information for the design phase of a product. Therefore, PUI lies in the focus of the extraction process of the Social Media Wrapper.

Table 5 FALCON related social media Data Sources

Data Source	Kind of Data Source	Information to be Extracted	Access Scenario	Requirements (D5.1-D8.1)
White Goods Scenario				
Brown Goods Scenario				

Customer Services Database	Database	Call centre id, customer e-mails, technical service, field testing and comments of social media (complaints)	Web service is possible but data is confidential	FALCON_MOL8
Facebook	Social Networking Site	Complaints, user information	Facebook API	FALCON_MOL8
Instagram	Content Community	Meta information about images	Instagram API	FALCON_MOL8
Sikayetvar	Website (Blog)	PUI (Complaints + usage data)	Website crawling	FALCON_MOL8
May be Amazon	Website (Blog)	PUI (Complaints + Usage Data)	Amazon API	FALCON_MOL8
Twitter	Social Networking Site	PUI (Complaints + usage data)	Twitter API	FALCON_MOL8
Website	Website (Blog) containing German complaints	PUI (Complaints + usage data)	t.b.d	FALCON_MOL8
Healthcare Scenario				
Repository of Log files	Log file (Blog)	PUI (Usage data)	File based Access like FTP/Samba	1. FALCON 0.6.1 2. FALCON 0.6.2.3
Clothing Textiles Scenario				
Facebook	Social Networking Site	PUI (Complaints), user information	Facebook API	D1.1.2.1
Instagram	Content Community	Meta information about images	Instagram API	D1.1.2.1
Twitter	Social Networking Site	PUI (Complaints + usage data)	Twitter API	D1.1.2.1
Pinterest	Blog	PUI (Complaints + usage data)	Pinterest API	D1.1.2.1
Lookbook	Blog/Content Community	PUI (Complaints + usage data)	Website crawling	D1.1.2.1
High-tech Products Scenario				

Help Desk	(Blog)	PUI (Complaints + usage data)	Not defined yet	DP1.2.1.1

All above listed social media related data sources represent the knowledge in continuous text which is linked to images and sometimes other kind of media. In consequence, the top-level requirement (FALCON_SMW) is the extraction and interpretation of PUI on basis of continuous text from the social media data sources. This top-level requirement (FALCON_SMW) is divided into three categories of sub requirements:

- FALCON_SMW.2 Data Acquisition
- FALCON_SMW.3 Data Pre-Processing
- FALCON_SMW.4 Data Transformation

The overview of the social media wrapper concept related requirements are shown in Figure 4.

The first requirement category addresses the functional requirements that ensure the data access to the relevant data. The relevant data sources will offer a web service based access, a file based access or the relevant content included in an html file. In consequence, the derived functional requirements are listed in Table 6.

Table 6 Data Acquisition Requirements

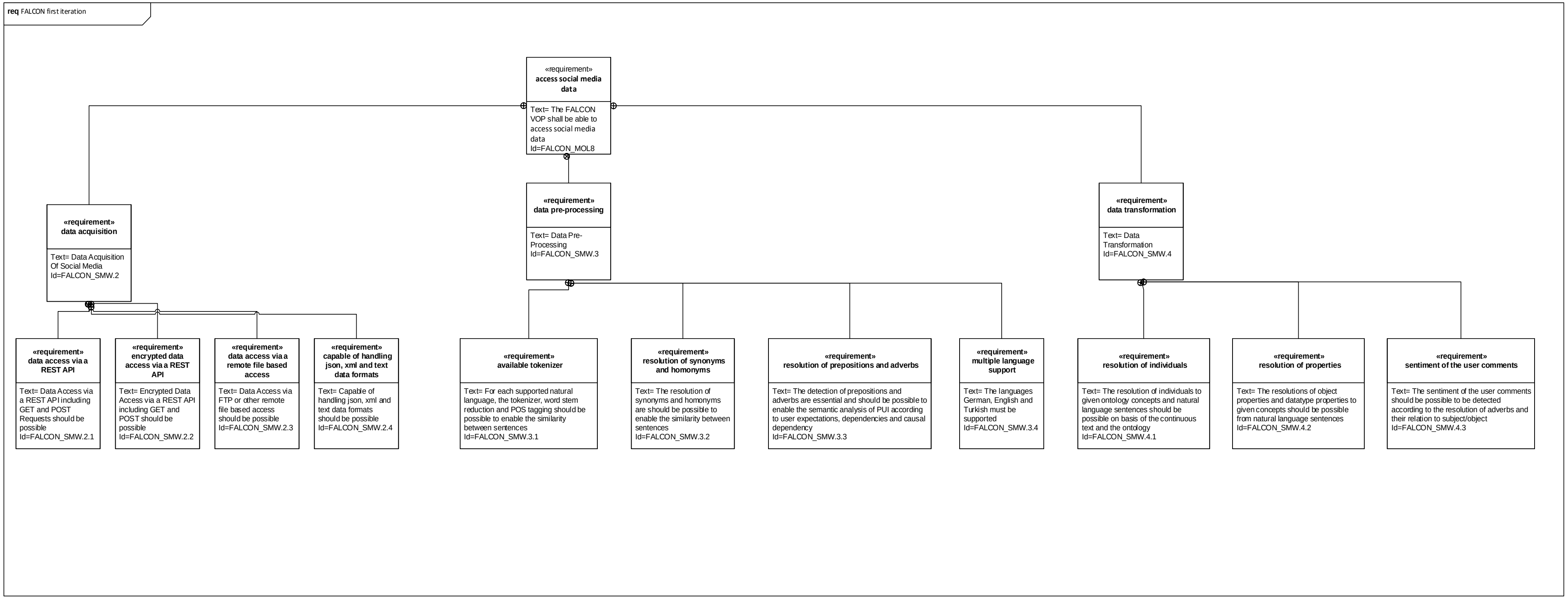
REQ ID	Description
FALCON_SMW.2.1	Data Access via a REST API including GET and POST Requests should be possible
FALCON_SMW.2.2	Encrypted Data Access via a REST API including GET and POST Requests should be possible
FALCON_SMW.2.3	Data Access via FTP or other remote file based access should be possible
FALCON_SMW.2.4	Capable of handling json, xml and text data formats should be possible

The second requirement category addresses the functional requirements that ensure the pre-processing of continuous text from the social media related data sources. In so doing, the pre-processing must transform the continuous text from social media type Blog which includes slang and no common terminology. Consequently, the transformation must be so extensive that continuous text from different data sources containing different slang and different languages could be harmonized to enable a semantic transformation process. In consequence, the derived functional requirements are listed in Table 7.

Table 7 Data Pre-Processing Requirements

REQ ID	Description
FALCON_SMW.3.1	For each supported natural language, the tokenizer, word stem reduction and POS tagging should be possible to enable the similarity between sentences
FALCON_SMW.3.2	The resolution of synonyms and homonyms should be possible to enable the similarity between sentences
FALCON_SMW.3.3	The detection of prepositions and adverbs are essential and should be possible to enable the semantic analysis of PUI according to user expectations, dependencies and causal dependency

FALCON_SMW.3.4	The languages German, English and Turkish must be supported



req FALCON first iteration

«requirement»
access social media data
Text= The FALCON VOP shall be able to access social media data
Id=FALCON_MOL8

«requirement»
data acquisition
Text= Data Acquisition Of Social Media
Id=FALCON_SMW.2

«requirement»
data pre-processing
Text= Data Pre-Processing
Id=FALCON_SMW.3

«requirement»
data transformation
Text= Data Transformation
Id=FALCON_SMW.4

«requirement»
data access via a REST API
Text= Data Access via a REST API including GET and POST Requests should be possible
Id=FALCON_SMW.2.1

«requirement»
encrypted data access via a REST API
Text= Encrypted Data Access via a REST API including GET and POST should be possible
Id=FALCON_SMW.2.2

«requirement»
data access via a remote file based access
Text= Data Access via FTP or other remote file based access should be possible
Id=FALCON_SMW.2.3

«requirement»
capable of handling json, xml and text data formats
Text= Capable of handling json, xml and text data formats should be possible
Id=FALCON_SMW.2.4

«requirement»
available tokenizer
Text= For each supported natural language, the tokenizer, word stem reduction and POS tagging should be possible to enable the similarity between sentences
Id=FALCON_SMW.3.1

«requirement»
resolution of synonyms and homonyms
Text= The resolution of synonyms and homonyms are should be possible to enable the similarity between sentences
Id=FALCON_SMW.3.2

«requirement»
resolution of prepositions and adverbs
Text= The detection of prepositions and adverbs are essential and should be possible to enable the semantic analysis of PUI according to user expectations, dependencies and causal dependency
Id=FALCON_SMW.3.3

«requirement»
multiple language support
Text= The languages German, English and Turkish must be supported
Id=FALCON_SMW.3.4

«requirement»
resolution of individuals
Text= The resolution of individuals to given ontology concepts and natural language sentences should be possible on basis of the continuous text and the ontology
Id=FALCON_SMW.4.1

«requirement»
resolution of properties
Text= The resolutions of object properties and datatype properties to given concepts should be possible from natural language sentences
Id=FALCON_SMW.4.2

«requirement»
sentiment of the user comments
Text= The sentiment of the user comments should be possible to be detected according to the resolution of adverbs and their relation to subject/object
Id=FALCON_SMW.4.3

Figure 4 FALCON requirement overview

The third requirement category addresses the functional requirements which addresses the semantic transformation of pre-processed continuous text into triples of an ontology. In consequence, the derived functional requirements are listed in Table 8.

Table 8 Data Transformation Requirements

REQ ID	Description
FALCON_SMW.4.1	The resolution of individuals to given ontology concepts and natural language sentences should be possible on basis of the continuous text and the ontology
FALCON_SMW.4.2	The resolutions of object properties and datatype properties to given concepts should be possible from natural language sentences.
FALCON_SMW.4.3	The sentiment of the user comments should be possible to be detected according to the resolution of adverbs and their relation to subject/object

5 Conceptual Approach

Following the methodology described in the previous section, this chapter presents the conceptual approach performed in the context of WP2, which builds upon the results of the elicitation and analysis activities performed respectively in work packages 3, 5, 6, 7 and 8. To achieve the semantically extraction of information from the social media relevant data sources, the interpretation of data beyond a data source is necessary. The unambiguity interpretation can be achieved on different levels. Oren et al. [Oren, Ghassam-Aghaee & Yilmaz 2007] defined different levels of understanding, which are

- Lexical understanding
- Syntactical understanding
- Morphological understanding
- Semantic understanding
- Pragmatic understanding

The lexical, syntactical and morphological understanding enables the recognition of the structure of the information and the grouping of relevant entities together, but the meaning is still unclear. The semantic understanding is the key to understand the meaning of the data and enables first the application of common data integration approaches to achieve the interoperability. The enforcement of data integration approaches in daily business requires both automated data integration processes and the maintenance of the established data sources. The Social Media Wrapper approach enables the data integration on the semantic understanding level. In cases, which don't enable an unambiguity extraction, the information will not be extracted to ensure the quality of information for the downstream FALCON analysis methods.

The overall approach foresees three steps proceeding, which enables the transformation from the data into an ontology for a specific social media data source. The three steps proceeding foresees the application of a rule based approach like usual in OBIE and Opining Mining approaches. For each new social media data source, the following 3 steps must be adapted. A new social media source is given if the physical access to the data source

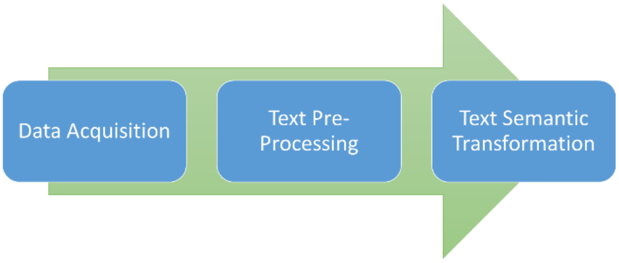


Figure 5 Social Media Wrapper Approach

has changed or the internal content changed so much that the configured pre-processing and rules based transformation approach are not applicable any longer. The 3 steps proceeding is illustrated in Figure 5.

A more detailed architecture for the approach is shown in Figure 6. In the data acquisition phase, social media related data sources resources such as tweets which are related to the given domain ontology are accessed. Afterwards, in the text pre-processing module, common NLP text pre-processing methods e.g. tokenizing are applied on the unstructured texts to provide a basis for further text analysis i.e. semantic transformation. The text pre-processing relies highly on the NLP resources of the target language. That means, when respective NLP resources for a specific target language are lacking, high quality pre-processing result is hard to be achieved. In the semantic transformation process, different pieces of information including both factual information as well as subjective information are extracted. The extracted pieces of information are further associated together to fill out Abox and then, to be uploaded as triples in the Triple-Store. The proposed semantic transformation process is driven by a given domain ontology. Factual information is extracted in principle based on Gazetteers and various extraction rules. Product features as well as sentiment towards the respective product feature are obtained respectively through the product feature identification module and sentiment analysis module. The module Aggregator concerns the association and aggregation of the extracted pieces of information.

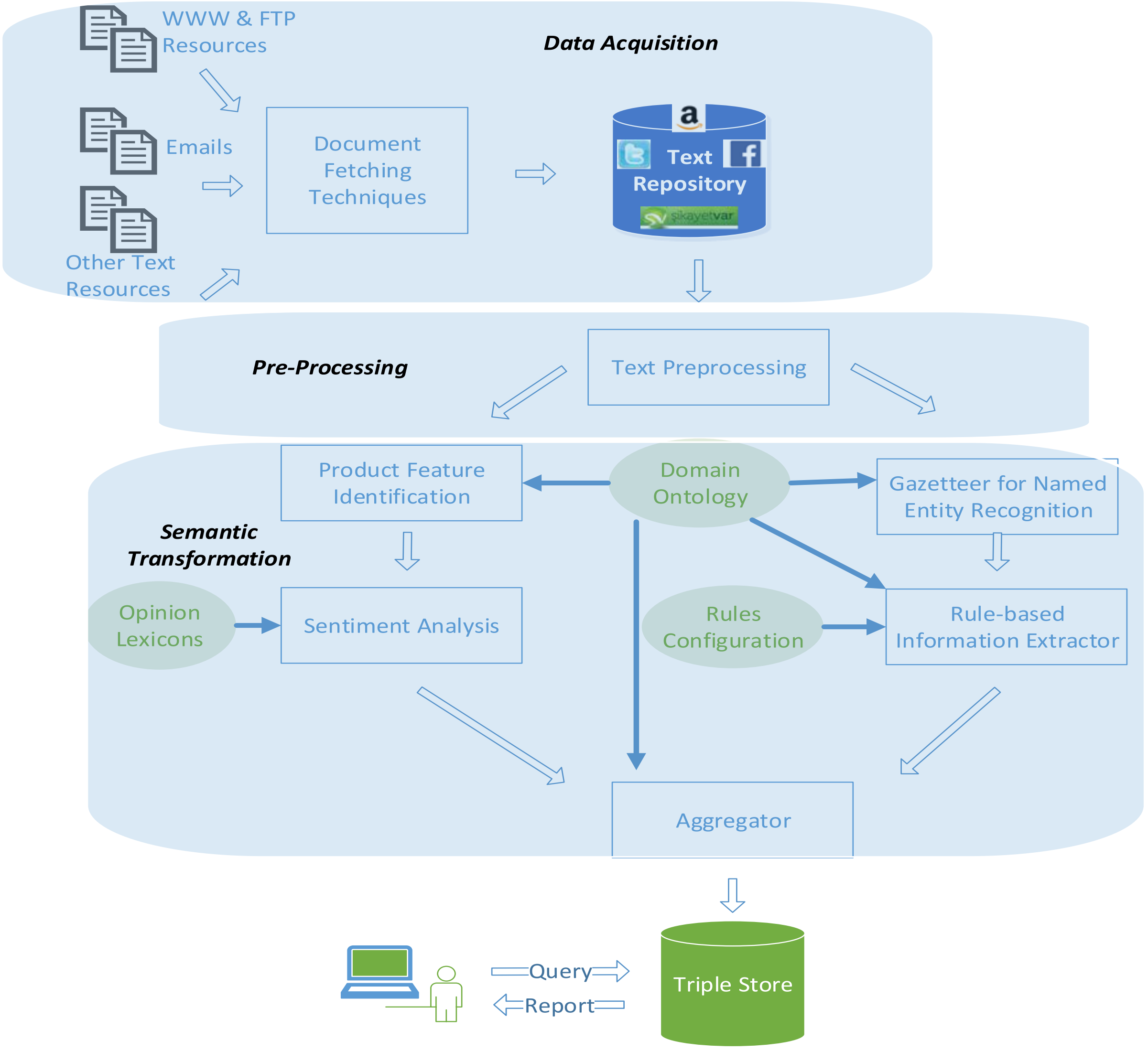


Figure 6 Social Media Wrapper Architecture

5.1 Approach to Data Acquisition

The data acquisition approach must be capable of accessing the file based repositories as well as the web resources (FALCON_SMW.2.1 - FALCON_SMW.2.4), which are shown in Figure 7.

As shown in Figure 7, all addressed FALCON-specific social media related data sources are categorised into three groups, which are described in the following. Each category will enable both a similar access scenario and applicable access technology.

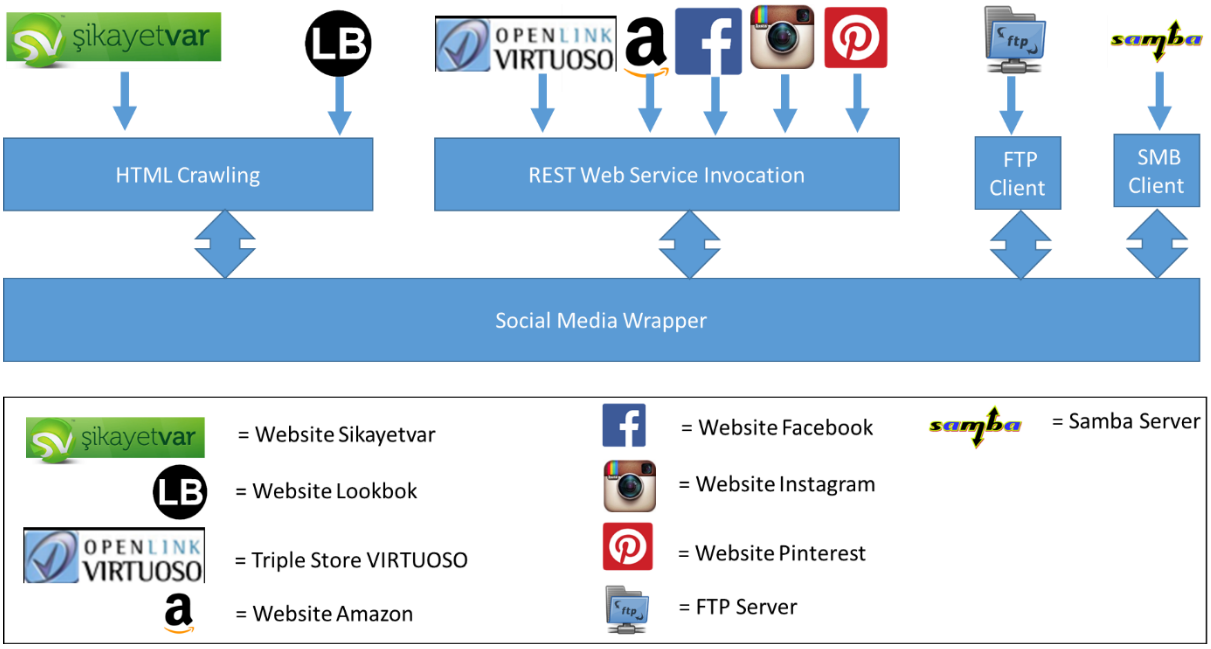


Figure 7 Data Access Layer for Social Media Wrapper

The first group contains websites, which offers no public API. The data acquisition is going to use HTTPGET requests and a downstream interpretation of the HTML content. In so doing, the data acquisition will follow included links to a specified configured depth. This feature is necessary to collect more detailed information of products, which are mentioned, not only on the main page. Apart from Lookbook and Sikayetvar (FALCON_SMW.2.1, FALCON_SMW.2.2), the other social media related web sites offers an API that belongs to the second group. The second group offers REST web services for third party tools (FALCON_SMW.2.1, FALCON_SMW.2.2). Thus, the data acquisition is going to implement HTTPGET and HTTPPOST requests and a downstream interpretation of JSON and XML content (FALCON_SMW.2.4). The data acquisition from the FALCON Triple Store (VIRTUOSO) will also be realized via REST web service. The remaining group provides data access via remote file repositories. The access to these repositories will be establish by FTP, SMB or other necessary protocols via pre-existing clients. For that purpose, open source java libraries will be applied to support the necessary protocols. The files included in the remote repositories will be handled as text files.

The above-mentioned data acquisition methods will be capable to collect continuous text from web resources as well as from remote files repositories. The data acquisition methods extract from all different data sources continuous text and forward it to the pre-processing module. This text will be handled as natural language. The data acquisition approach foresees the update capability to extract only new and not already acquired data.

5.2 Approach for Text Mining: Pre-Processing

The pre-processioning approach of social media includes the common pre-processing steps of text mining, which resulted in a 4-step approach. The overall approach assumes that the continuous text is represented as natural language (FALCON_SMW.3.4). Each step focusses on a specific step in natural language processing and will be executed when the previous step is finished. The overall 4-step pre-processing approach is shown in Figure 8 and is described in the following.

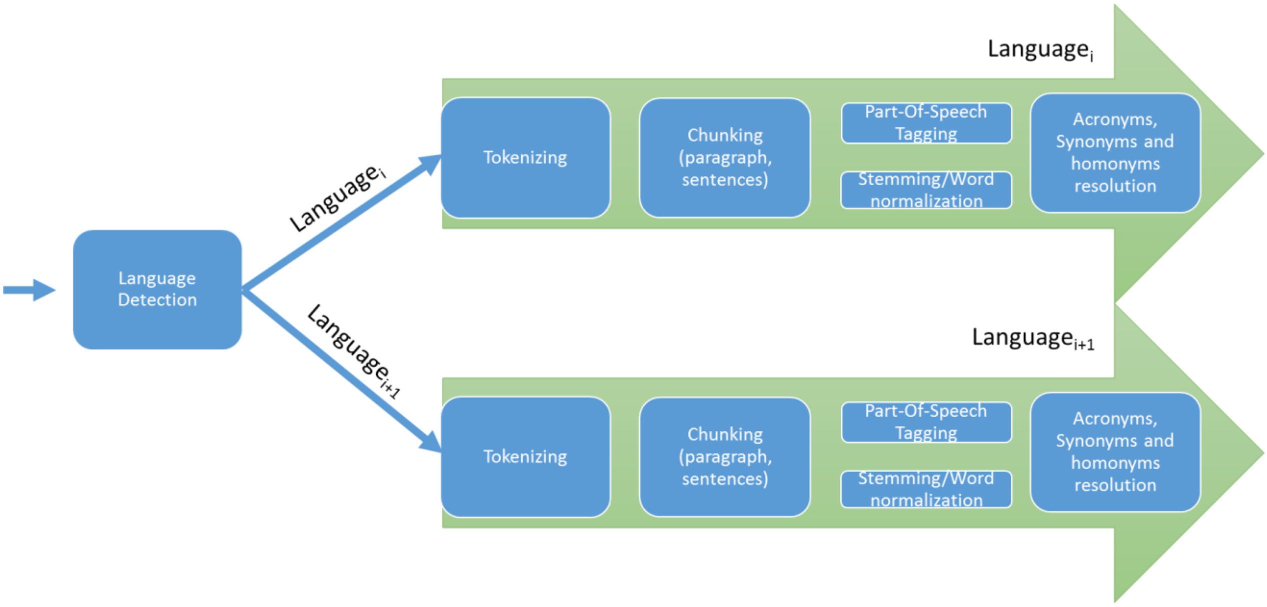

Language Detection
$Language_i$
$Language_{i+1}$
$Language_i$
Tokenizing
Chunking (paragraph, sentences)
Part-Of-Speech Tagging
Stemming/Word normalization
Acronyms, Synonyms and homonyms resolution
$Language_{i+1}$
Tokenizing
Chunking (paragraph, sentences)
Part-Of-Speech Tagging
Stemming/Word normalization
Acronyms, Synonyms and homonyms resolution

Figure 8 Pre-Processing of Continuous Text

The input of the 4-step pre-processing method is triggered by the Language Detection. This method detects the chosen language of the continuous text and choose the corresponding language specific 4-step pre-processing method. The syntactical and semantic differences between natural languages are too high (reading direction, word boundaries, grammar, etc.) to have one general pre-processing (FALCON_SMW.3.1).

The first step of the language specific pre-processing is the tokenizing of the continuous text. In so doing, a sequence of characters is converted into a sequence of tokens. A token could be a word, phrase, symbol, or other meaningful elements called token. The boundary of a token will be detected by the usage of language specific grammatical characters which represent e.g. the end of a word, the end of a sentences or the end of n-gram (FALCON_SMW.3.1).

The second step is to define and extract chunks of the continuous text. A “chunk” is a continuous non-overlapping sequence of words. The proposed chunking approach will detect sections, paragraphs and sentences. The opportunity to detect specific chunks within sentences through chunk rules (like (S: (NP: I) saw (NP: the Chinese woman) is not covered (FALCON_SMW.3.1).

The third step is to extract the grammar as well as stemming/word normalization. The part-of-speech tagging (POS tagging) analyses a word within a chunk and assign to each word a particular part of speech. For that purpose, a part of speech is a category of words, which represents the similar grammatical meaning. Common groups of part of speech are in English and German noun, verb, adjective, adverb, etc. (FALCON_SMW.3.3) The POS tagging must be applied in parallel to the stemming/word normalization, because these methods remove the information to classify words into parts of speech. The objective of the stemming/word normalization is to transform all words into its infinitive. The transformation is necessary to enable the similarity of words for the semantic transformation (see. 5.3). Common approaches for the reduction to the infinitive form are the application of stemming algorithms or the lookup in dictionaries. While the stemming of English natural language results a high quality, stemming is not applicable for the German continuous text. The applicability of stemming algorithms for Turkish continuous text is an open research question and will be reported in D2.3 (FALCON_SMW.3.1).

The fourth and therefore last step is to resolve the acronyms, synonyms and homonyms in the chunks. The approach foresees the replacement of acronyms with their long terms, the reduction of synonyms with one favourite version and to replace homonymies through more precise words (FALCON_SMW.3.2). All these steps are necessary to reduce the ambiguity as well as to enable the similarity of chunks. For that purpose, look up tables for acronyms, synonyms and homonyms are necessary for all supported natural languages.

The result of the 4 step pre-processing is the transformed continuous text into chunks whereby each chunk contains a set of word. Each word is provided as its infinitive version and is assigned to a part of speech group. Moreover, grammatical specific characters are not contained any longer in the chunks. The necessary pre-processing approach of words, which are not detectable in dictionaries is still an open research question and first evaluation results will clarify it. The final approach will be reported in D2.3.

5.3 Approach for Text Mining: Semantic Transformation

The presented approach foresees an information transformation chain to transform pre-processed unstructured texts from the FALCON business cases to Abox of the FALCON ontology (FALCON_SMW.4.1-FALCON_SMW.4.3). The task of this component is defined as follows:

The input for the information transformation chain is a given an ontology Tbox (FALCON ontology) and a pre-processed unstructured text. The output is filled out Abox in which the information of the given pre-processed unstructured text are inserted. For that purpose, the ontology properties for both factual and sentiment information should be included in the ontology. Without these predefined properties, no mapping could be defined for the extraction process.

As an example, companies could have interest to know the sentiment polarities of customers towards a specific product feature e.g. “price”. In this approach, both factual information and sentiment information on product features must be contained as properties in the ontology and would be extracted. The extracted pieces of information will be associated/aggregated together and converted to triples, then finally kept in triple store for further analysis. However, this approach is not intended to build a “global problem solver” which is applicable in any business domain and for any unstructured texts. It will focus on the business domains in the FALCON project. Due to the limitations described in the following sub-modules, information for some concepts would be unable or not precisely extracted.

The approach presented in this section can be seen as information transformation chain to transform pre-processed unstructured texts in FALCON business domains to structured semantic annotated triple entries. The task of this component is defined as following: Given an ontology Tbox and a pre-processed unstructured text, Ontology Abox are supposed to be automatically filled out using the information extracted from the given pre-processed unstructured text (FALCON_SMW.4.1-FALCON_SMW.4.3). As in the Tbox of a domain ontology, beside properties for factual information, properties asking for sentiment information could be as well present. Companies would have interest to know the sentiment polarities of customers towards a specific product feature e.g. “price”. In this approach, both factual information and sentiment information on product features will be extracted. It focusses on the business domains in FALCON project. Due to the limitations described in the following sub-modules, information for some concepts would not be precisely extracted.

5.3.1 Extracting Factual Information

Extracting factual information consists of the identification and extraction of pieces of factual information for the properties/concepts in a domain ontology to a given unstructured text (FALCON_SMW.4.1- FALCON_SMW.4.2). In order to identify the applicable ontology concepts to a specific chunk various information extraction techniques could be applied. In general, following four techniques are often used: Gazetteer lists, linguistic rules, analyzing tags, machine learning techniques. In our approach, the approach Gazetteer lists and linguistic rules or patterns matching technique are mainly investigated and will be combined within transformation rules. In the following, each of the techniques are presented.

The technique "analyzing tags" is based on html/xml tags, which is usually not available in unstructured texts. In case of their occurrences, a rule based approach for detecting and extracting information from tags can then be applied. Machine learning based information extractor would not be taken into consideration at this version, because for the extraction of each concept/property, considerable effort for the preparation on the training model e.g. data preparation/labeling is requested. When considering hundreds of concepts are presented in a domain ontology, a supervised approach requiring thousands of examples for the model training seems quite unfeasible. Moreover, when a new concept is added in the domain ontology, the user has to then prepare amount of training texts for the extraction of the single new concept. As already shown, the elements "Gazetteers for Named Entity Recognition" and "Rule-Based Information Extractor" are two core modules for the extraction of factual information from texts.

5.3.1.1 Gazetteers for Named Entity Recognition

This module involves the identifying individual entities for a particular ontology concept. It concerns typically on a limit number of concepts such as PERSON, LOCATION. Various techniques e.g. supervise learning can be applied in this task. Due to the drawback of requiring amount of labeled training data in supervised technique, the approach presented here would focus on usage of gazetteer list. The gazetteer lists are used to find occurrences of entities in text. It is clear that in this approach, the quality of the named entity recognition is highly depended on the gazetteer resource. In order to identify a concrete ontology concept, this technique requires a carefully prepared gazetteer list containing all instances of the concept. It is not appropriate for the extraction on the concept that the gazetteer resource is not available. The information for the gazetteer list can either obtained with the help of domain ontology or from other resources. For example, in case we want to use this technique to identify the concept" Product Model Number" for Arçelik Washing Machine, the gazetteer information can be provided by Arçelik, or automatically generated based on the given ontology in case all product model numbers are listed in the ontology.

5.3.1.2 Rule-Based Information Extractor

This module concerns the extraction of specific information from text using rule matching technique. The rules are in general based on regular expression combining with the information from Tokenizer, POS tags and Gazetteer etc. Obviously, there is no general rule which can be applied to recognize information for all concepts, individual rules are necessary for specific concepts. In the case of extraction rules such as regular expression is already combined with the ontology elements in a given ontology, the specified rules can be employed to get the desired information. Otherwise, suitable extraction rules have to be manually figured out after learning on sample documents. Extraction rules can be written in the language Java Annotation Patterns Engine (JAPE), which is a finite state transductor allowing the creation of complex extraction rules over annotations.

For the extraction of different ontology elements, various approaches could be applied. Meanwhile, for the same ontology element, depends on the characteristics of text sources, different approaches could be required. For example, the concept "product name" from the amazon reviews can be easily retrieved through APIs, because the reviews in amazon belong usually to a specific product, while the concept "product name" is not so easy to be identified from twitter texts. The extraction methodology or rules can only be clearly defined, when the to be recognized ontology element as well as related resource are specified. The concept" Product Model Number" for Arçelik Washing Machine can be again taken as an example, considering the following four cases: 1) All valid product model numbers are provided by Arçelik, but this kind of information is not modelled in the ontology. In this case, a gazetteer list with the information from Arçelik could be used for the recognition. 2) All valid product model numbers are already listed in the given ontology. In this case, a gazetteer list can be derived based on the ontology and then employed for information extraction. 3) Regular expression for extraction the product model number is specified in the ontology. In this case, the regular expression from the ontology would be parsed to match the needed information. 4) Neither valid product model numbers are listed nor extraction rule is given. In this case, extraction rule must be created manually to identify potential valid product model numbers. However, the manual creation of extraction rules is a time consuming exercise, and for some concepts it would be very difficult to manually figure out suitable extraction rules. When the extraction rule is not available or not suitable enough, the rule-based information extractor can only get poor results.

5.3.2 Extracting Feature-based Sentiment Information

The process is responsible for identifying product features and obtaining the respective sentiment (FALCON_SMW.4.3) with the help of domain ontology. As shown in Figure 6, it consists of two modules: product feature identification and sentiment analysis.

5.3.2.1 Product Feature Identification

The main idea behind this component is to use formal define ontology terminology for the identifying of product features. It takes the pre-processed text and domain ontology as input to recognize the product features that are present in the text. This would put requirements on the ontology developer to model ontologies that can a) list all to be identified product features b) distinct elements representing product features from other ontology elements. With such kind of ontology, we can identify related sentence which contains the product features, and further easily extract the product features from the sentences. Furthermore, consumers can write the name of the same product feature in many ways, for example with synonyms or abbreviation. For that reason, the synonyms or abbreviation of the to be identified product features need to be handled with the help of lexicons to increase recognition quality.

5.3.2.2 Sentiment Analysis

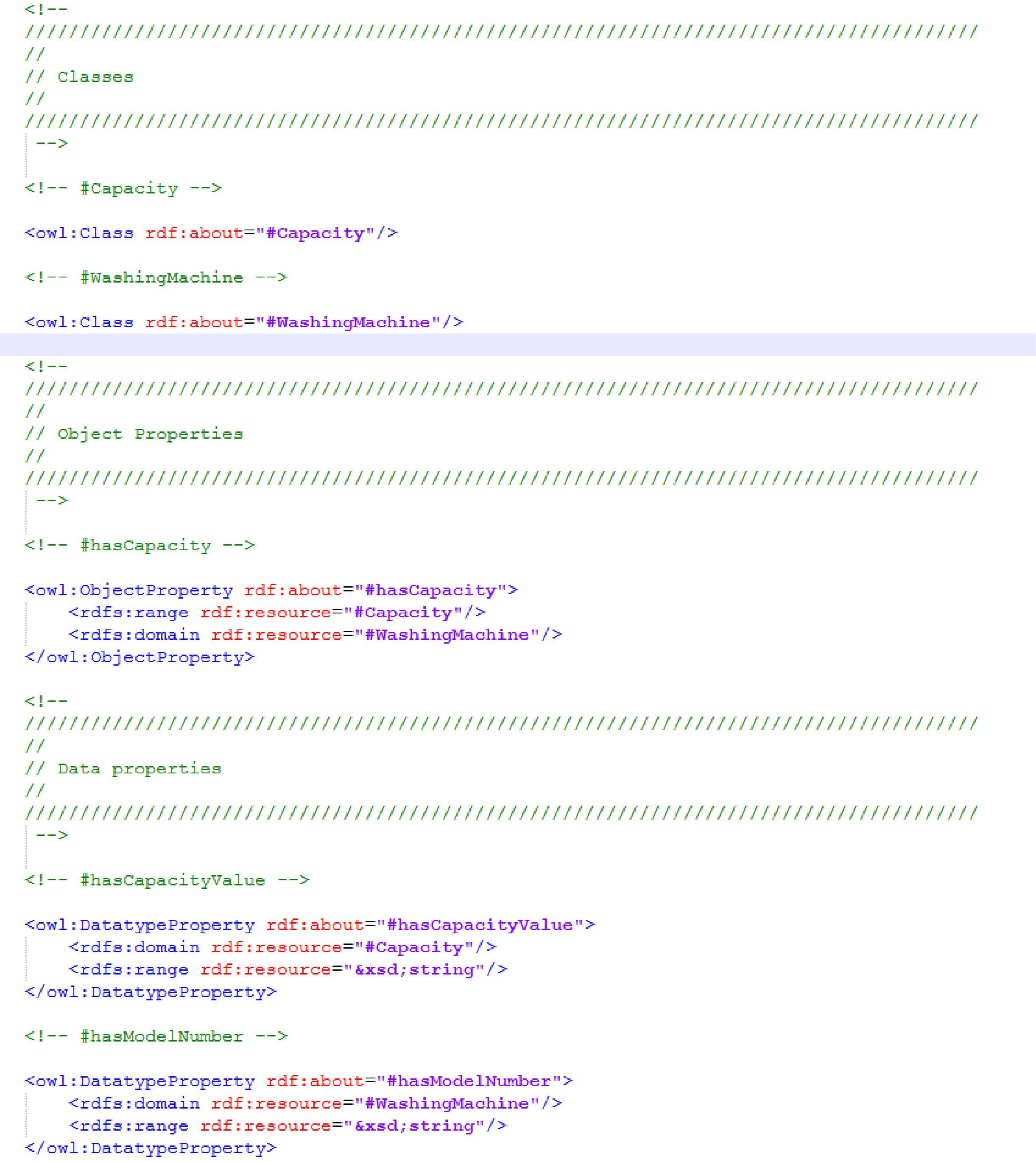
This module concerns the identification of the sentiment polarity towards target features. In this approach, sentiment analysis will be done using generic lexicons and set of basic opinion rules. To identify the words which express opinions towards a target feature, the same strategy presented by Martinez et al. [Peñalver-Martinez et al. 2014) can be employed: 'N_GRAM After', 'N_GRAM Before', 'N_GRAM Around' and 'All_Phrase'. The sentiment polarity of the target feature is then calculated based on the sentiment values of each opinion word. The sentiment values of each word can be derived from lexical resources. For example, for the English language, SentiWordNet is a lexical resource which assigns three sentiment scores (i.e. positive, negative, objectivity) to each synset of WordNet. For the language Germany or Turkish, other respective lexicon resources are required. Meanwhile, the negation words such as "not" need to be specially handled, because this kind of words would usually change or even reverse the sentiment orientations.

5.3.3 Aggregator

This module ensures that different pieces of extracted information are merged together to have a holistic view of a single ontology Abox instance node. For example, an instance of washing machine is associated with its product features "energy consumption", "capacity" as well as consumers' sentiment towards the product features. The task is performed by Social Media Wrapper and Data Federation Module. Once an instance is identified, Social Media Wrapper will glue other extracted attribute values to the instance. Data Federation Module takes the advantage of ontology reasoning ability for further post-processing or other kind of information merging. The Abox instances are temporally managed within Data Federation Module, and finally converted and kept into Triple-Store through SPARQL.

In order to guide the semantic transformation process, a mapping file will be configured. In the mapping file, information extraction related items are configured for each ontology concept as well as ontology object property and datatype property. Under the guide of the configurations, the semantic transformation process extracts respective information for each ontology element and glues them together into ontology Abox instances. In the following, a short example is shown to depict the principle idea.

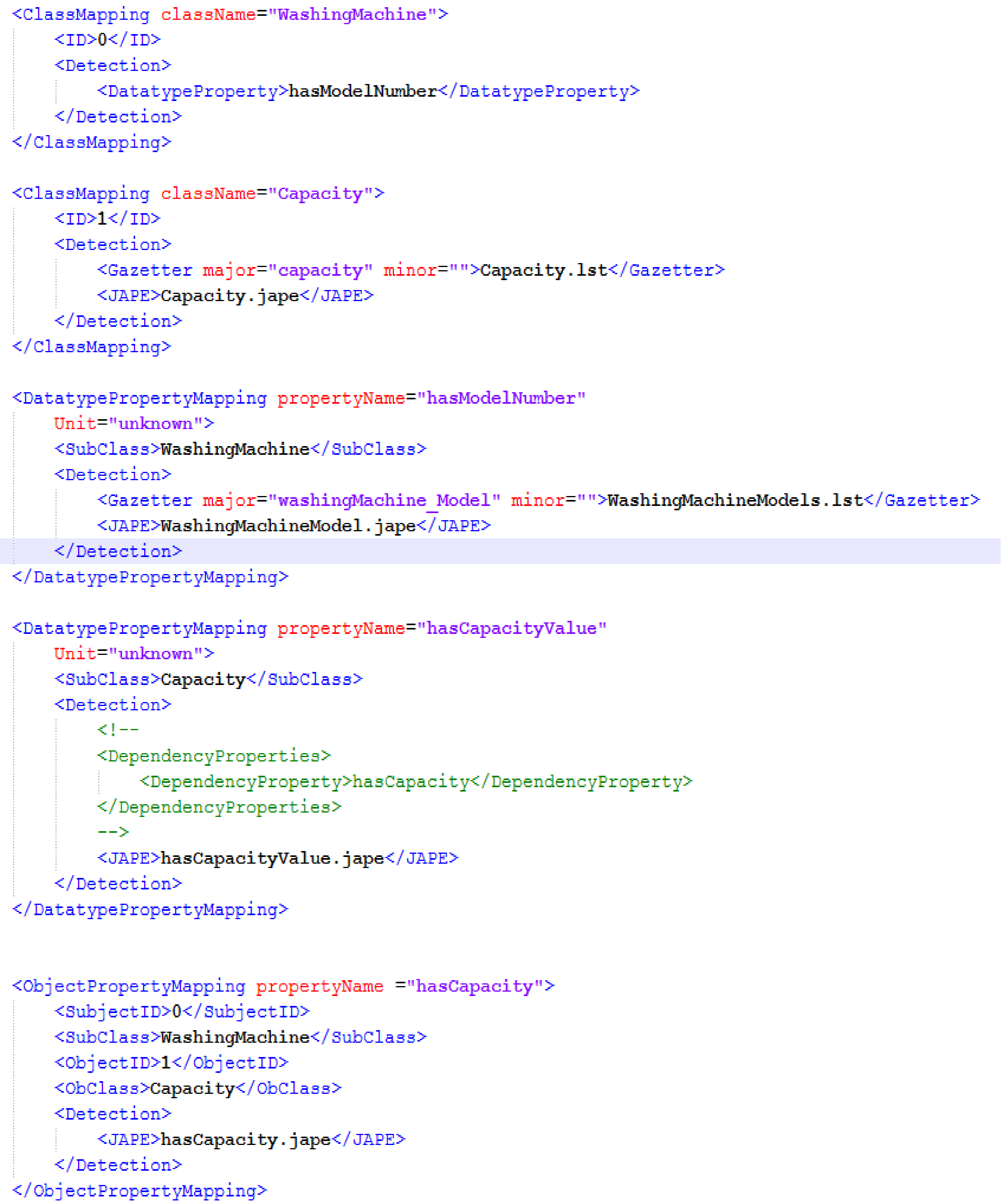
Consider a sentence “The Maytag MHW7100DC has a good capacity of 4.5 cubic feet.” and a given washing machine ontology shown in the Source Code 1.



```
<!--
//////////////////////////////////////////////////////////////////////////////////////
//
// Classes
//
//////////////////////////////////////////////////////////////////////////////////////
 -->

<!-- #Capacity -->

<owl:Class rdf:about="#Capacity"/>

<!-- #WashingMachine -->

<owl:Class rdf:about="#WashingMachine"/>

<!--
//////////////////////////////////////////////////////////////////////////////////////
//
// Object Properties
//
//////////////////////////////////////////////////////////////////////////////////////
 -->

<!-- #hasCapacity -->

<owl:ObjectProperty rdf:about="#hasCapacity">
    <rdfs:range rdf:resource="#Capacity"/>
    <rdfs:domain rdf:resource="#WashingMachine"/>
</owl:ObjectProperty>

<!--
//////////////////////////////////////////////////////////////////////////////////////
//
// Data properties
//
//////////////////////////////////////////////////////////////////////////////////////
 -->

<!-- #hasCapacityValue -->

<owl:DatatypeProperty rdf:about="#hasCapacityValue">
    <rdfs:domain rdf:resource="#Capacity"/>
    <rdfs:range rdf:resource="&xsd;string"/>
</owl:DatatypeProperty>

<!-- #hasModelNumber -->

<owl:DatatypeProperty rdf:about="#hasModelNumber">
    <rdfs:domain rdf:resource="#WashingMachine"/>
    <rdfs:range rdf:resource="&xsd;string"/>
</owl:DatatypeProperty>
```

Source Code 1 Example for Washing Machine Ontology

An example configuration file to link unstructured text and ontology is shown in Source Code 2. The XML element “Detection” specify methodologies for the extracting of a specific element. For instance, as stated in the mapping file, a gazetteer list “WashingMachineModels.lst” is used to recognize the property “hasModelNumber” and further for the instance “WashingMachine”. For the identifying of the datatype property “hasCapacityValue”, rules specified in the JAPE file “hasCapacityValue.jape” are applied. In the rules, regular expressions as well as results from the other components e.g. tokenizer, gazetteers are applicable.



```
<ClassMapping className="WashingMachine">
    <ID>0</ID>
    <Detection>
        <DatatypeProperty>hasModelNumber</DatatypeProperty>
    </Detection>
</ClassMapping>

<ClassMapping className="Capacity">
    <ID>1</ID>
    <Detection>
        <Gazetter major="capacity" minor="">Capacity.lst</Gazetter>
        <JAPE>Capacity.jape</JAPE>
    </Detection>
</ClassMapping>

<DatatypePropertyMapping propertyName="hasModelNumber"
    Unit="unknown">
    <SubClass>WashingMachine</SubClass>
    <Detection>
        <Gazetter major="washingMachine_Model" minor="">WashingMachineModels.lst</Gazetter>
        <JAPE>WashingMachineModel.jape</JAPE>
    </Detection>
</DatatypePropertyMapping>

<DatatypePropertyMapping propertyName="hasCapacityValue"
    Unit="unknown">
    <SubClass>Capacity</SubClass>
    <Detection>
        <!--
        <DependencyProperties>
            <DependencyProperty>hasCapacity</DependencyProperty>
        </DependencyProperties>
        -->
        <JAPE>hasCapacityValue.jape</JAPE>
    </Detection>
</DatatypePropertyMapping>


<ObjectPropertyMapping propertyName ="hasCapacity">
    <SubjectID>0</SubjectID>
    <SubClass>WashingMachine</SubClass>
    <ObjectID>1</ObjectID>
    <ObClass>Capacity</ObClass>
    <Detection>
        <JAPE>hasCapacity.jape</JAPE>
    </Detection>
</ObjectPropertyMapping>
```

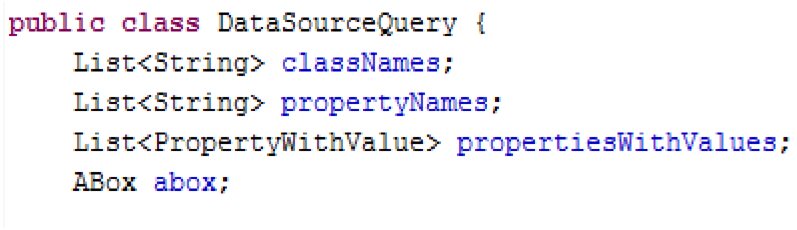
Source Code 2 Example for mapping file between unstructured text and ontology element

5.4 Integration of Social Media Wrapper into the FALCON VOP

The general wrapper approach of [Klein et al. 2014, Franke, Pirvu & Lappe 2014], the specialization of this approach for web services of [Franke et al. 2014] and the integration of wrappers as part of a mediator oriented data integration solution like proposed in [Klein et al. 2014] are the cornerstones of the Data Federation Module of the FALCON VOP. In consequence, the Social Media Wrapper will be a part of the Data Federation Module and therefore, must be compatible with the predefined wrapper approach. In the following, the predefined functional requirements of the general wrapper approach must be considered. The chosen wrapper approach foresees the availability of transformation rules and an ontology for each data source as part of the configuration. The proposed architecture of a wrapper requires an ontology and transformation rules as input which must be aligned with "global-as-view" (GAV) approach of the Data Federation Module. The GAV approach enables information requests and fast information aggregation over multiple data source specific ontologies.

The data source specific ontologies will be subsets of the overall FALCON Ontology. The chosen GAV integration approach *"….is effective whenever the data integration system is based on a set of sources that is stable…"* [Lenzerini 2002]. A subset of the FALCON ontology will be stored as a RDF XML file and added as part of the wrapper configuration.

Apart from the general architecture, the integration of the Social Media Wrapper as a Data Federation Modulus's wrapper has to satisfy strict requirements according to the data structure of the input and result parameters. The input for any kind of wrapper is a parameter including the data structure DataSourceQuery which is illustrated in Source Code 3.



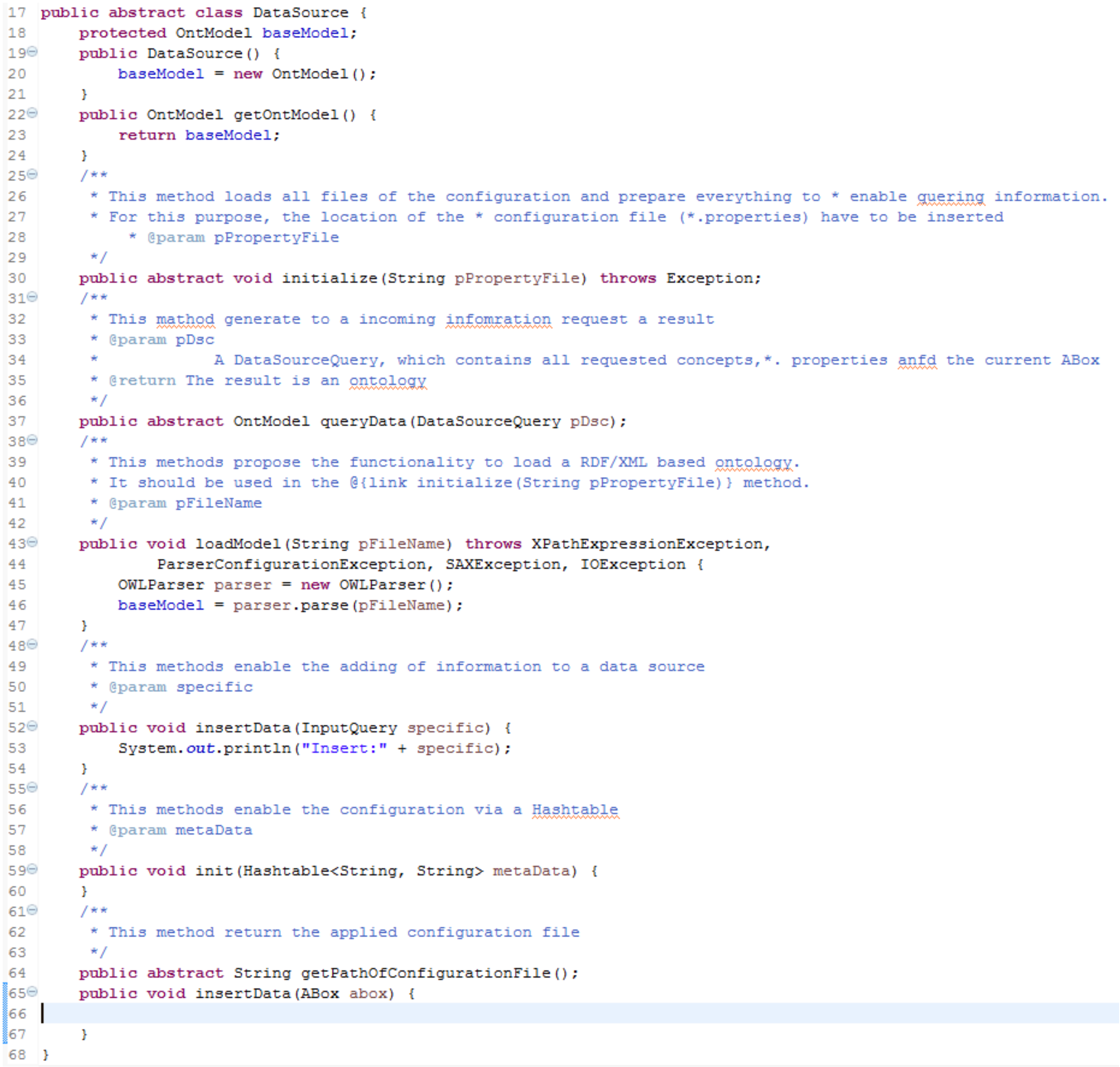
```
public class DataSourceQuery {
    List<String> classNames;
    List<String> propertyNames;
    List<PropertyWithValue> propertiesWithValues;
    ABox abox;
```

Source Code 3 Overview of DataSourceQuery

The data structure DataSourceQuery contains the list of concepts, list of datatype and object properties, a list of constant values and the current available Abox. Apart of the defined input parameter, a wrapper must return the generated Abox as an OntModel. Thus, the OntModel is a java class which implements the functionality of an ontology's Abox and Tbox. Apart of the specific data structures, each possible wrapper implementation of the Data Federation Module must implement at least the Java interface DataSource which is illustrated in following Source Code 4.

The interface DataSource defines the JAVA methods related to how the wrapper is initialized, how configuration-related information is requested, how SPARQL queries are invoked and how data is inserted into the linked data source.

The lifecycle of the Social Media Wrapper is dependent on the Data Federation Module. That means, that the Data Federation Module instantiates the wrapper and invoke its querying capabilities. Other FALCON VOP modules are not allowed to invoke the wrapper.



```java
public abstract class DataSource {
    protected OntModel baseModel;
    public DataSource() {
        baseModel = new OntModel();
    }
    public OntModel getOntModel() {
        return baseModel;
    }
    /**
     * This method loads all files of the configuration and prepare everything to * enable quering information.
     * For this purpose, the location of the * configuration file (*.properties) have to be inserted
         * @param pPropertyFile
     */
    public abstract void initialize(String pPropertyFile) throws Exception;
    /**
     * This mathod generate to a incoming infomration request a result
     * @param pDsc
     *             A DataSourceQuery, which contains all requested concepts,*. properties anfd the current ABox
     * @return The result is an ontology
     */
    public abstract OntModel queryData(DataSourceQuery pDsc);
    /**
     * This methods propose the functionality to load a RDF/XML based ontology.
     * It should be used in the @{link initialize(String pPropertyFile)} method.
     * @param pFileName
     */
    public void loadModel(String pFileName) throws XPathExpressionException,
            ParserConfigurationException, SAXException, IOException {
        OWLParser parser = new OWLParser();
        baseModel = parser.parse(pFileName);
    }
    /**
     * This methods enable the adding of information to a data source
     * @param specific
     */
    public void insertData(InputQuery specific) {
        System.out.println("Insert:" + specific);
    }
    /**
     * This methods enable the configuration via a Hashtable
     * @param metaData
     */
    public void init(Hashtable<String, String> metaData) {
    }
    /**
     * This method return the applied configuration file
     */
    public abstract String getPathOfConfigurationFile();
    public void insertData(ABox abox) {

    }
}
```

Source Code 4 Generic Wrapper Interface

5.5 Configuration Process of Social Media Wrapper for each Business Case

The Social Media Wrapper will be capable to connect to a specific social media data source and to extract a specific set of use case specific amount of information. Before, the Social Media Wrapper can be used in such a way within a business process, a three step configuration process is necessary which is shown in Figure 9. In the first step, the amount of knowledge that shall be extracted must be configured. For that purpose, an ontology must be created covering all concepts and properties of interest. The created ontology may only contain concepts and properties which are also contained in the FALCON ontology. Any other concepts can’t be applied in the FALCON services and therefore are useless. The created ontology will be represented as RDF/XML and contains only language elements of OWL DL and the inverse functional property of OWL Full.

In the second step the linkage between the social media related data and the ontology is configured. For that purpose, a mapping file must be created. This file contains the transformation rules. Each rule defines the transformation of the social media data into an Abox. The implemented GAV approach requires a transformation rule for each property as well as for each addressed concept. The mapping file will be represented as a XML file which guarantees the readability by humans as well as by computer programs.

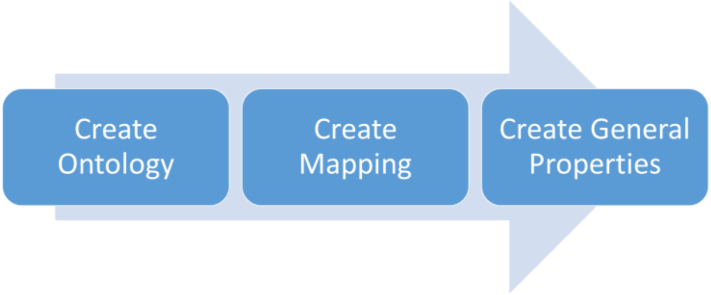


Figure 9 Configuration Process

The last configuration step contains the definition of social media specific access information like credentials and the chosen extraction method to be applied. The location of the ontology file and the mapping file will be also listed in the configuration file. The representation form of this file will be as an Eclipse property file like the configuration files of the Legacy System Wrappers.

The configuration files will be included in a Java project considering the Data Federation Module Configuration and uploaded as library to the Nexus server. Then, the configuration would be available the next time the solution would be build and provided for the business case. This configuration approach is preliminary till the FALCON VOPs provide containers for project specific settings. Then, a configuration of the Social Media Wrapper including the three mentioned files will be stored and accessed via the provided containers.

6 Conclusion

This deliverable documents the development of the conceptual approach of the FALCON Social Media Wrapper. It presents a general overview over the term social media and the corresponding FALCON specific social media data sources including the relation with business scenario work packages (WP5, WP6, WP7, WP8). Then, the state-of-the-art regarding natural language processing/information extraction and opinion mining was presented. On the basis of the state-of-the-art and the FALCON specific social media data sources, a set of functional requirements has been derived and presented. REST services and remote file repositories are in the focus of the data acquisition. The derived functional requirements and the FALCON VOP specific integration requirements form the cornerstones of the developed Social Media Wrapper conceptual approach which is presented in detail in Section 5. The presented conceptual approach will enable the extraction of opinions and facts. Therefore, it is going to apply the transformation capabilities of the semantic data integration whereby continuous text, which is represented in a natural language, could be transformed into an Abox of the FALCON ontology. In so doing, a generated Abox won’t be added directly to the Triple Store of FALCON VOP. The Data Federation Module of WP3 will take the responsibility to integrate the Social Media Wrapper into the FALCON VOP. Moreover, the Data Federation Module is the only FALCON VOP module, which is going to communicate with it. The current Social Media Wrapper approach does not support the inclusion of the user context information to determine the difference between facts, irony and targeted misstatements. These issues will be clarified in D2.3

7 References

[Anantharangachar, Ramani, & Rajagopalan 2013] Anantharangachar, Raghu; Ramani, Srinivasan; Rajagopalan, S., (2013): Ontology Guided Information Extraction from Unstructured Text. In: IJWesT 4 (1), S. 19–36. DOI: 10.5121/ijwest.2013.4102.

[Baccianella, Esuli & Sebastiani 2010] Baccianella S, Esuli A and Sebastiani F (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. European Language Resources Association. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation, May 2010, pp. 2200–2204.

[Blei, Ng & Jordan 2003] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3,993-1022

[Brooks et al. 2006] Brooks, C. H., & Montanez, N. (2006, May). Improved annotation of the blogosphere via autotagging and hierarchical clustering. In Proceedings of the 15th international conference on World Wide Web (pp. 625-632). ACM.

[Cimiano, Handschuh & Staab 2004] Cimiano, P., Handschuh S., & Staab S.. 2004. Towards the self-annotating web. In Proceedings of the 13th international conference on World Wide Web (WWW '04). ACM, New York, NY, USA, 462-471.

[Cowie & Lehnert 1996] Cowie, Jim; Lehnert, Wendy. (1996). Information extraction. Commun. ACM 39, 1 (January 1996), 80-91. DOI=http://dx.doi.org/10.1145/234173.234209

[Embley et al. 1998] Embley, David W.; Campbell, Douglas M.; Smith, Randy D.; Liddle, Stephen W. (1998): Ontology-based extraction and structuring of information from data-rich unstructured documents. In: Niki Pissinou, Charles Nicholas, James French, George Gardarin, K. Makki und L. Bouganim (Hg.): the seventh international conference. Bethesda, Maryland, United States, S. 52–59.

[Franke, Pirvu & Lappe 2014] Franke, M., Pirvu B.-C., Lappe D. and other Interaction Mechanism of Humans in a Cyber-Physical Environment. Proceedings of the 4th International Conference on Dynamics in Logistics. International Conference on Dynamics in Logistics (LDIC-14), 4th, February, Bremen, Germany, LDIC.10-14; 2014

[Franke et al. 2014] Franke, M., Klein, K., Hribernik, K., Lappe, D., Veigt, M., & Thoben, K. D. (2014) Semantic Web Service Wrappers as a Foundation for Interoperability in Closed-loop Product Lifecycle Management. Procedia CIRP, 22, 225-230; 2014

[Freitas & Vieira 2013]de Freitas, L. A., and Vieira, R. (2013). Ontology based feature level opinion mining for portuguese reviews. In Proceedings of the 22nd International Conference on World Wide Web (WWW '13 Companion). ACM, New York, NY, USA, 367-370.

[Grishman 1997] Grishman , Ralph (1997) Information Extraction: Techniques and Challenges. In International Summer School on Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology (SCIE '97), Maria Teresa Pazienza (Ed.). Springer-Verlag, London, UK, UK, 10-27.

[Gutierrez et al. 2015] Gutierrez, F.; Dou, D.; Fickas, S.; Wimalasuriya, D.; Zong, H. (2015): A hybrid ontology-based information extraction system. In: Journal of Information Science.

[Kaplan & Haenlein 2010] Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. Business horizons, 53(1), 59-68.

[Kietzmann et al. 2011] Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). social media? Get serious! Understanding the functional building blocks of social media. Business horizons, 54(3), 241-251.

[Klein et al. 2014] Klein, K., Franke, M., Hribernik, K. A., & Thoben, K. D. Identification of Interface Information for a Virtual Data Integration. In Enterprise Interoperability VI. Springer International Publishing. 89-99; 2014

[Kleiza, Franke & Thoben 2010] Kleiza, K., Klein, P., Franke, M., & Thoben, K. D. (2010). Integrated Semantic Search in the Product Development Phase. In 16th International Conference on Concurrent Enterprising (ICE2010); Proceedings; Lugano, Switzerland.

[KONYS 2015] KONYS, Agnieszka (2015): An Approach for Ontology-Based Information Extraction System Selection and Evaluation. In: ELECTROTECHNICAL REVIEW 1 (11), S. 207–211. DOI: 10.15199/48.2015.11.49.

[Lenzerini 2002] Lenzerini, M. Data integration: A theoretical perspective. In Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems ACM. 233-246. 2002

[Li & Li 2013] Li, Yung-Ming, Li, Tsung-Ying (2013), Deriving market intelligence from microblogs, Decision Support Systems, Volume 55, Issue 1, April 2013, Pages 206-217, ISSN 0167-9236

[Liu 2012] Liu, B. (2012), Sentiment Analysis and Opinion Mining, Morgan and Claypool, 2012

[Madway 2010] Madway, G. (2010, September 14). Twitter remakes website, adds new features. Retrieved November 5, 2010, from http://www.reuters.com/article/idUSN1411135520100915

[Medhat, Hassan & Korashy 2014] Medhat, Walaa; Hassan, Ahmed; Korashy, Hoda (2014): Sentiment analysis algorithms and applications. A survey. In: Ain Shams Engineering Journal 5 (4), S. 1093–1113

[Muslea 1999] Muslea, I. (1999) Extraction patterns for information extraction tasks: a survey. In Proceedings of AAAI '99: Workshop on Machine Learning for Information Extraction

[Oren, Ghassam-Aghaee & Yilmaz 2007] Oren, T., Ghassam-Aghaee N., and Yilmaz, L. (2007). An Ontology-based Dictionary of Understanding as a Basis for Software Agents with Understanding Abilities. Proceedings 2007 Spring Simulation Multiconference, IEEE Press

[Pang & Lee 2008] Pang, B., Lee, L. (2008), Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2), pp. 1–135, 2008

[Peñalver-Martinez et al. 2014] Peñalver-Martinez, Isidro; Garcia-Sanchez, Francisco; Valencia-Garcia, Rafael; Rodríguez-García, Miguel Ángel; Moreno, Valentín; Fraga, Anabel; Sánchez-Cervantes, Jose Luis (2014): Feature-based opinion mining through ontologies. In: Expert Systems with Applications 41 (13), S. 5995–6008.

[Popov et al. 2004] Popov B, Kiryakov A, Kirilov A, Manov D, Ognyanoff D & Goranov M (2004).KIM–semantic annotation platform. Natural Language Engineering, 10(3–4),375–392.

[Pronoza , Yagunova & Volskaya 2014] Pronoza E., Volskaya S., Yagunova E. Corpus-based Information Extraction and Opinion Mining for the Restaurant Recommendation System. Proceedings of the 2nd Statistical Language and Speech Processing. L. Besacier et al. (Eds.): SLSP LNAI 8791, pp. 272–284, 2014.

[Riloff 1999] Riloff, E. (1999) Information extraction as a stepping stone toward story understanding. In: A. Ram and K. Moorman (eds), Understanding Language Understanding: Computational Models of Reading (MIT Press, Cambridge, MA, 1999)

[Saggion & Funk 2009] Saggion H, Funk A (2009) Extracting Opinions and Facts for Business Intelligence. RNTI E-17:119–146

[Salas-Zarate et al. 2016] Salas-Zarate, M. d. P.; Valencia-Garcia, R.; Ruiz-Martinez, A.; Colomo-Palacios, R. (2016): Feature-based opinion mining in financial news. An ontology-driven approach. In: Journal of Information Science.

[Schouten & Frasincar 2016] Schouten, Kim; Frasincar, Flavius (2016): Survey on Aspect-Level Sentiment Analysis. In: IEEE Trans. Knowl. Data Eng. 28 (3), S. 813–830.

[Strapparava & Valitutti 2004] Strapparava, Carlo and Valitutti, Alessandro.(2004) WordNet-Affect: an Affective Extension of WordNet, in Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, May 2004, pp. 1083-1086.

[Wimalasuriya & Dou 2010] Wimalasuriya, D. C.; Dou, Dejing(2010): Ontology-based information extraction. An introduction and a survey of current approaches. In: Journal of Information Science 36 (3), S. 306–323. DOI: 10.1177/0165551509360123.

[Wu & Weld 2007] Wu F. and Weld DS (2007). Autonomously semantifying Wikipedia. In: ACM Conference on Information and Knowledge Management, 2007, pp. 41–50

[Zhang et al. 2011]Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., and Liu,B. (2011). Combining lexicon-bsed and learning-based methods for twitter sentiment analysis. HP Laboratories, Techncal Report HPL-2011, 89.

[Zhao & Li 2009] Zhao, L. and Li, C.(2009) Ontology based opinion mining for movie reviews. In 3rd International Conference Knowledge, Science, Engineering and Management,2009