DataTourism: designing an architecture to process tourism data

DataTourism: designing an architecture to process tourism data

Fayrouz Soualah-Alila, Mickaël Coustaty, Nicolas Rempulski, Antoine Doucet
L3I laboratory, La Rochelle University
Michel Crépeau avenue, 17000 La Rochelle, France

{fayrouz.soualah-alila, mickael.coustaty, antoine.doucet}@univ-lr.fr
nicolas.rempulski@gmail.com

Abstract

With the rapid diffusion of new technologies in tourism, professionals face new challenges to use efficiently the vast amount of data created by tourists. Nowadays, these information come from multiple and varied sources, as cellular or social networks, touristic location attendance or dematerialized satisfaction surveys, and in huge amount. They are an important resource for the tourism industry, but their heterogeneity makes it difficult to aggregate and analyse them. The key issue for tourism actors, professionals or politics, is to manage and operate tourism information about their territory effectively. The purpose of this paper is to describe synthetically how tourism information are managed under the *Tourinflux* project. We describe in this paper an architecture named *DataTourism* for tourism data management which solve multiple technical locks encountered when working with tourism data: heterogeneity, quality, interoperability, reusability and standardisation.

Keywords: DataTourism; Tourinflux; TIFSem; TimeML; SentiML.

1 Introduction

In today’s rapidly changing world, many data related to tourism area are produced. This is primarily the result of increasing possibilities to digitize growing volumes of data, and the development of open-sources and open-data policies. Likewise, more and more data are being generated by sensors, mobile telephones, and connected devices on the one hand, and the democratization of comparative services dedicated to tourism on the other hand, as *Kayak* or *Yelp* for instance. Most of these data could be collected and used by politics to efficiently assign public funds to increase tourists’ attendance and satisfaction and thus making their territory attractive. But today, they are mostly unused due to a lack of suitable tools.

However, the business sector is already using these data. They are analysed for marketing strategies, predicting trends and also for producing detailed statistics. Tourism professionals are using multiple sources, and fully use the recent development of the World Wide Web and its social services. The web has changed people’s daily life, this is also true for tourism. It has significantly influenced the way information about users are gathered and exchanged in the tourism sector. With the intensive use of social networks and web sites specialized in e-tourism (*TripAdvisor*, *Booking.com*, etc.) web users are no longer passive recipients of contents; they absorb information from the web and in return produce their own new content. But when users collect these information, from professional sources or from other users, they also create their own set of information: tourism goods they are looking for, future date of their vacation, etc. Professional tourism services collect these information while providing information or services to users. We could make the same example with cellular carrier, which are tracking movements of their user. These Information, wherever they are from, are then used to improve the service quality by enhancing employee’s knowledge about customer’s preferences and opinions.

Two main problems occur with tourism data management: their heterogeneity and their volume. As mentioned before, tourism information are continuously enhanced and updated using dedicated websites. These data are contained on web pages that are originally designed to be human-readable, and so, most of information currently available on the web are kept in large collections of textual documents. As the web grows in size and complexity, there is an increasing need for automating time consuming tasks, such as information extraction and interpretation. Some automatic process to annotate and enrich textual information knowledge is needed.

Tourism domain is characterized by a significant information heterogeneity and by a high volume of online data. Data related to tourism are produced by different experts (travel agents, tourist offices, etc.) and by visitors, thus creating an heterogeneous data set from a semantic and typology point of view. Moreover, this set is often incomplete and inconsistent. For instance, these data could contain information related to tourism objects (hotels, concert, restaurant, etc.) with raw information, service description for instance, temporal information, about opening hours or days in the week, and opinions, as users' average satisfaction. There are already multiple taxonomies and catalogues which are designed and used internally by tourism actors to allow them to manage heterogeneous tourism data efficiently. Efforts are now made to generate standards to facilitate inter and intra tourism data exchange.

The *Tourinflux*[1] project falls within this context where it meets a basic need: helping professional and political tourism actors to develop their territory. One way to valorise territories is to generate reports, called dashboards, based on enriched data collected from Tourist Information Systems (TIS) and the web. The emergence dashboards was a consequence of managers needs to monitor a complex subject with indicators clearly showing how a territory's tourism activities are perceived and evaluated. Experts from tourism industry use and need these dashboards to improve their knowledge about the tourist attractiveness of their territory.

The technical architecture we created in *Tourinflux* is aimed at providing the tourism industry with a set of tools (1) allowing them to handle both their internal data, and the information available on the web, and (2) allowing to improve the display information available about their territory on the web. In this paper we describe synthetically how tourism data, composed of information related to tourism objects (hotels, concert, restaurant, etc.), temporal information and opinions, are managed under the *Tourinflux* project. We present an architecture named *DataTourism* for the management of tourism data that allows solving different locks: heterogeneity of tourism data sources, quality of these data, interoperability, reusability and standardisation.

2 Designing touristic dashboards

A touristic dashboard is a set of management indicators, built periodically, for a tourist actor or a group of tourist actors, in order to guide their decisions and actions to achieve performance goals. A touristic dashboard is considered as:

- An instrument of control and comparison: It allows tracking the evolution of tourist offers;
- A tool for taking decision: It communicates key information about the situation of a touristic activity;
- A communication tool: It provides a permanent communication between the various tourism actors and between different hierarchical levels;
- A standby tool: It allows to identify emerging opportunities and risks.

Figure 1 shows an example of dashboard.

[1] http://tourinflux.univ-lr.fr/index.php/component/content/?view=featured

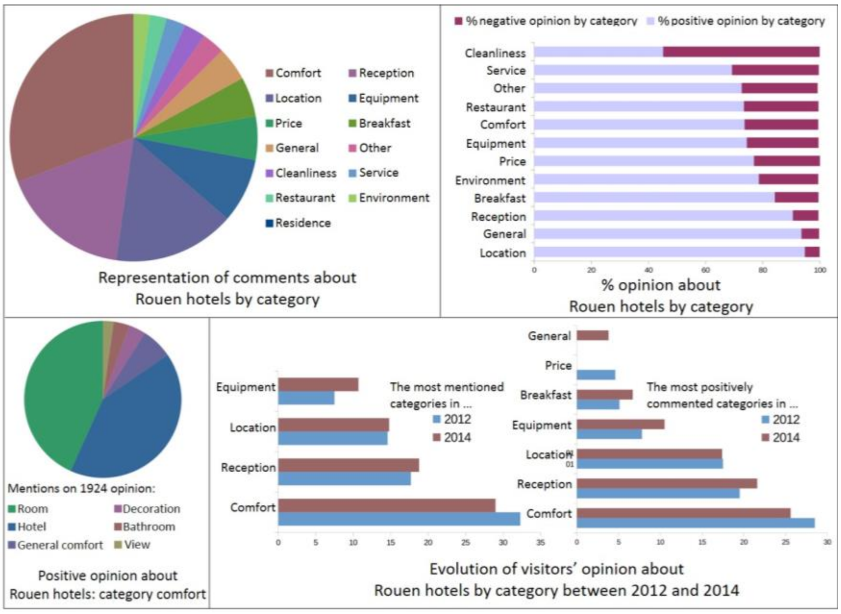

Fig. 1. Example of dashboard

In France, we distinguish five main institutional publishers of tourist dashboards: Tourist Offices (TO), Departmental Tourist Committees (DTC), Enterprises General Direction (EGD), National Institute for Statistics and Economic Studies (Insee) and Atout-France[2]. Today, each of these publishers has independently developed various techniques to assess a territory (a city, a department, a region, the whole country) and despite all the efforts made so far in developing their own dashboards, these dashboards remain insufficient to fulfil goals described above:

- They are not sufficiently representative of tourism territory. They focus more on accommodations. An objective is to integrate other sources such as content about opinions and intentions of visitors.
- They are limited to the territory scale they are developed for. It is impossible to generate dashboards at all hierarchical levels (department, region, country) or make a comparison between territories. This is mainly due to the heterogeneity of TIS.

In order to generate rich dashboards, it is necessary to exploit optimally all information available. The data set used has to be the most complete and varied as possible to reflect faithfully the touristic activity of a territory.

In the next section we describe our DataTourism architecture for the aggregation of tourism information from different data sources. The goal of the architecture is to provide a framework that allows the aggregation of tourism information following an ontological based approach. The framework must access tourism data sources, extract

[2] http://atout-france.fr/

their information combine them, as different as they are, and present them to the tourist professionals in the form of dashboards.

3 Sources and types of tourism data

Tourism in its nature is an industry strongly dependent on data exchange. In the last decade, more and more data have been becoming available for research and development. These data are from different sources. The main sources of tourism data which we consider as part of this project are: the data available in the different TIS, the data available on the web and Open-Data. This data could be composed of (1) information related to tourism objects (hotels, concert, restaurant, etc.), (2) temporal information and (3) events and opinions. Our architecture to generate dynamic dashboards has four major phases: integration of information related to tourism objects in the system, annotation of temporal information and opinions in web pages, enhancement of the tourism objects from the annotated information (temporal and opinions), and finally dynamic generation of dashboards. All the components that are used in each phase are illustrated in Figure 2.

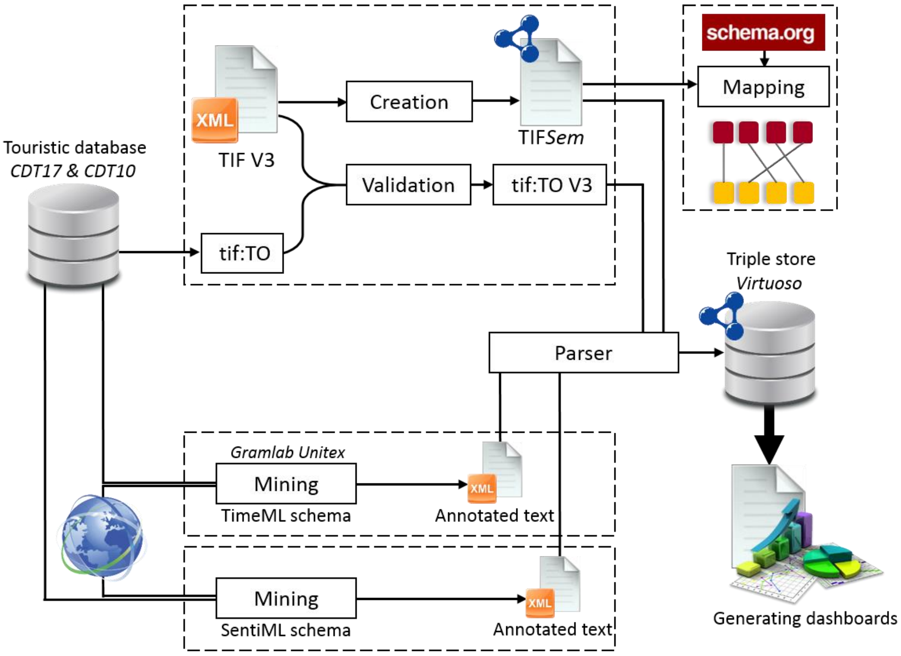

Fig. 1. DataTourism general architecture

We describe briefly in the following sections how we modelled tourism objects in our architecture and how we proceed to annotate temporal information and opinions in web pages to complete the description of the tourism objects.

3.1 Tourism objects modelling

The interoperability of TIS is a major challenge for the development of tourism domain. Several national, European and international institutional initiatives have proposed different standards to meet the specific needs of tourism professionals. In France, the major TIS such as *Raccourci Interactive*[3], *TourinSoft*[4] and *Sitra*[5], have adopted the *TourInFrance* standard (TIF). Since its creation in 1999 by the tourism ministry, TIF is used today by more than 3000 tourist offices in France, by DTC and by different tour operators, to facilitate data exchange between these different actors. In 2004, the TourInFrance Technical Group (TIFTG) approved the new version of the standard, TIF V3. In this version, the standard has evolved towards XML technologies to facilitate publishing information on the web and information exchange between systems, and is accompanied by several thesaurus. Since 2005, this standard stopped evolving. As a result, tourism professionals have adapted the standard to their own needs (different tags syntaxes, new tags added, etc.) and proposed their own evolution in an unorganized way, producing TIS that are not really interoperable among themselves and that cannot directly be shared using international standard. Accordingly, the exploitation of tourism information is trapped in its own territory, and so it is impossible to aggregate these information (Bittner and al., 2005).

To overcome these limitations we propose to evolve the TIF standard to share the knowledge it contains and to ensure data interoperability, by applying the concept of ontology to represent the standard terminology. An ontology is considered as *the specification of a conceptualization,* in other words, *a specific artifact designed with the purpose of expressing the intended meaning of a shared vocabulary* (Hirst, 2004). Having a common semantic base compensates the interoperability problem (Fodor, 2005) that comes along with the integration of heterogeneous data sources by converting existing heterogeneous unstructured tourism data into structured ontological data. In the tourism area, some researches have already tackled the design of ontologies. Several available tourism ontologies show the current status of the efforts: the OTA (Open Travel Alliance) (OTA, 2000), the Harmonise ontology (Dell'Erba and al., 2002), the Hi-Touch ontology (Legrand, 2004), the QALL-ME ontology (Ou, 2008), the Tourpedia catalogue (Cresci, 2014), etc. These models focus on different areas of tourism domain but there is not exist one single ontology which matches all the needs of different tourism related applications. We then propose in (Soualah-Alila and al., 2015) an ontology called TIFSem (Semantic TourInFrance) to globally describe tourism objects (TO) mixing heterogeneous content. Concepts included in the defined ontology will allow us to describe information sources on tourism. This enriched information can be used: (1) from the user side, to match tailored package holidays to client preferences for instance, and (2) from tourism experts' point of view, to analyse and better manage online data about their territory.

In order to elaborate the TIFSem ontology, we have consulted different kinds of sources to enable the understanding and the creation of concepts related to the specialized domain of tourism, and the corresponding vocabulary. Sources coming from Departmental Tourism Committee of the Charente Maritime[6] (CDT17) and

[3] http://www.raccourci.fr/

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

[5] http://www.sitra-tourisme.com/

[6] http://www.charente-maritime.org/

Departmental Tourism Committee of the Aube[7] (CDT10) were consulted. We are also in the process of extending the TIFSem ontology by collecting contents about more touristic service providers.

In a second step we are interested on facilitating the process of researching and publishing tourism data on the Web. Since TIF standard is unable to easily share and interoperate with global Web standards, we propose to enrich the TIFSem ontology with the Schema.org[8] model. This initiative started a few years ago, led by Bing, Google and Yahoo to standardize structured data format in the Semantic Web. Schema.org tends to become a de-facto norm to easily share semantic content. Launched in 2011, Schema.org aims to create and to support a common set of schemas for structured data mark-up Web pages. The purpose of Schema.org is to provide a collection of schemas (HTML tags) that webmasters can use to mark-up HTML pages in a way recognized by major search providers, and that can also be used for structured data interoperability (RDFa, JSON-LD, etc.). When these tags are used in a website, search engines can better understand the meaning of embedded resources(text, image, video) of that website; which in turn would allow them to return better search results when a user is looking for a specific resource through a search engine (Toma, 2014). Schema.org is a very large vocabulary counting hundreds of terms from multiple domains. In our approach we want to spread enriched semantic tourist data that can be easily indexed by search engines. We propose to match terms of TIFSem with terms of Schema.org by using semantic relations, and work with Schema.org community to extend the schema, either formally by adding new terms or informally by defining how Schema.org can be combined with some additional vocabulary terms. This method is complex to implement because it requires to find matching between the terms of TIF and those of Schema.org and seeking what is lacking in each model.

After collecting the concepts and corresponding lexical items from our sources, we started structuring the first version of the ontology, selecting classes and subclasses. The mapping with Schema.org enhanced the ontology to make it more complete, up to date and coherent. We now want to expand the TIFSem model with other types of data, about time and opinions, to provide semantic and contextual answers to queries. In the following sections we describe how in our architecture we integrate information about time and opinions to complete the TIFSem model.

3.2 Tourism temporal data

In the Tourinflux project, we analysed how temporal data could be collected from the web and integrated into DataTourism. Temporal data are pieces of information frequently encountered in tourism web pages. Most tourism objects (events, hotels, restaurants, etc.) on the web are associated to periods and events and are characterized by different timestamps like date, duration, opening hours, opening conditions, frequency, etc. Textual tourism data on the web is a rich body of phenomena for linguistic analysis. The automatic recognition of temporal and event expressions in

[7] http://www.aube-champagne.com/

[8] https://schema.org

natural language text has recently become an active area of research in computational linguistics and semantics.

In our proposed model, events will be annotated according to the TimeML language, a robust specification language for the challenging task of annotation of temporal information over natural language text (Pustejovsky and al., 2005). TimeML has been developed in the context of AQUAINT workshops and projects. The 2002 Time and Event Recognition for Question Answering Systems (TERQAS) workshop set out to enhance natural language question answering systems to answer temporally-based questions about the events and entities over free text on the Web. The first version of TimeML was defined and the TimeBank corpus was created as an illustration. In 2003, TimeML was further developed in the context of the TimeML Annotation Graphical Organizer (TANGO) workshop. In 2009 TimeML has been developed into an ISO standard (ISO WD 24617-1:2007).

TimeML includes four major data structures: EVENT, TIMEX3, SIGNAL, and LINK. In TimeML, *events are situations that occur or happen, or predicates that describe states or circumstances in which something obtains or holds the truth* (Pustejovsky and al., 2003). Events in TimeML are annotated with the tag EVENT. TIMEX3 is used to tag explicit temporal expressions, such as time, dates, and durations. SIGNAL is used to annotate sections of text, typically function words that indicate how temporal objects are related to each other (when, during, before, etc.). Finally, LINK encode various relations that exist between the temporal elements of a content. Each of these tags are associated to attributes to integrate temporal expressions. As the description of TimeML is not the aim of this paper, a complete description of the language is given in the manual (Sauri and al., 2009).

Within the Tourinflux project, in order to facilitate the extraction of temporal data, a corpus of tourism Web pages to be analysed is first constituted. This corpus is constituted of:

- A free text corpus containing festivals and events description, provided by the Local Action Group Othe Armance[9]. This corpus is available under LGPL/LR license (Lesser General Public License For Linguistic Resources).
- A corpus provided by the CDT10. This corpus contains descriptions of Places of Interest (POI): hotels, restaurants, etc. containing in particular information about opening and closing dates, opening and closing times, etc.
- Open data, including data concerning national museums.

In our case, annotation of temporal expressions of our corpus is performed by a set of finite state transducers, developed with the Gramlab Unitex[10] corpus processor. Unitex is a corpus processing system for processing natural language texts using electronic resources such as electronic dictionaries and grammars, and Gramlab is an integrated development environment, based on Unitex software components, designed for industrial project management purpose. Before applying the transducers, Unitex performs some pre-processing that consist on cleaning the text. It (1) normalizes apostrophes, quotes etc., (2) segments the text into sentences and tokenizes it and (3)

[9] http://www.tourisme-othe-armance.com/

[10] http://www-igm.univ-mlv.fr/~unitex/

applies a number of built-in lexical resources, such as dictionaries to identify, for instance, compound word forms, proper names, etc. (Paumier, 2008). Once the text is cleaned, temporal expressions are tagged according to their TimeML type. The tagger performs identification of events. Then Unitex detects and annotates temporal expressions and calculates the value attribute for each of the tags as specified by the TimeML guidelines. The tagger also detects certain relation markers, such as temporal prepositions like before, after, etc. The last spot of the tagger is to determine the links between the different annotations. The resulting output of Gramlab is the original corpus annotated with EVENT, TIMEX3, SIGNAL and LINK tags, whose values can be after integrated in the TIFSem model.

3.3 Tourism opinion data

An important part of our information-gathering behaviour has always been to find out what people think about their touristic experience. Opinions help to analyse a situation from different aspects and take an appropriate decision. The opinion of one individual may influence another individual opinion and hence the concept of public opinion is generated. Public opinion is considered important in tourism area.

The amount of opinionated data on tourism websites has exponentially increased especially after the rapid growth of online social networks. With the availability and popularity of rich opinion resources, we need to have reliable mechanisms which help identifying all aspects of opinions in a text and to extract useful information from it. Thus, we introduce the concept of opinion mining.

Opinion Mining is the process of automatic extracting opinions from textual segments (Liu, 2012). In the literature, it has commonly been referred as sentiment analysis or sentiment classification and sometimes as subjectivity analysis (Cambria and al, 2013). There are many related sub-tasks of opinion mining and semantic annotation of opinions is one of them. Semantic annotations are very important for preparing data for machine learning and evaluation of opinion mining approaches. Besides this, it helps the automatic extraction of opinions. Some annotation schemas have been proposed by the research community such as SentiML, OpinionMining-ML and EmotionML. A detailed comparison of SentiML with other existing annotation schemas is also presented in (Malik and al, 2014).

In our case, we used SentiML for annotating opinion data. In SentiML we talk about sentiments rather than opinions. The goal of SentiML is to identify and classify sentiment groups (positive and negative) at the sentence level. In order to do this, the schema focus on three categories: target (expression the sentiment refers to) and modifier (expression conveying sentiment). A target is any entity (object, person or concept) that is implicitly or explicitly regarded as positive or negative by the author of the text. A modifier is what modifies the target. It can be an adjective, a verb, an adverb or a noun. However, SentiML also adds in its vocabulary a much needed appraisal type tag. An appraisal group represents an opinion on a specific target. For this reason, it is defined as the link between the target and the modifier (example, link a noun with an adjective, a verb with an adverb, etc.). Besides this, SentiML is based on Appraisal Framework (AF) which is a strong linguistically-grounded theory. AF helps to define appraisal types (affect, judgments and appreciation) within the modifier tag.

References

Bittner, T., Donnelly, M. & Winter, S. (2005). Ontology and Semantic Interoperability. Zlatanova, S & Prosperi, D. (Eds), *Large-Scale 3D Data Integration: Challenges and Opportunities*, 139-160.

Hirst, G. (2004). Ontology and the lexicon. Staab, S. & Studer, S. (Eds), *Handbook on Ontologies*: Springer-Verlag, 209-229.

Fodor, O. & Werthner, H. (2005). Harmonise - a Step Towards an Interoperable e-Tourism Marketplace. *International Journal of Electronic Commerce*.

OTA (2000). Opentravel Alliance. Opentravel Alliance message specications. Specications Document 1.

Dell'Erba, M., Fodor, O., Ricci, F. & Werthner, H. (2002). Harmonise: A Solution for Data Interoperability. *Towards the Knowledge Society: eCommerce, eBusiness, and eGovernment, the Second IFIP Conference on E-Commerce, E-Business, E-Government*, 433-445.

Legrand, B. (2004). Semantic Web Methodologies and Tools for Intra-European Sustainable Tourism. White paper, Paris, Mondeca.

Cresci, S., D'Errico, A., Gazze, D., Duca, A. L., Marchetti, A. & Tesconi, M. (2014). Towards a dbpedia of Tourism: the Case of tourpedia. *International Semantic Web Conference*, 129-132.

Ou, S., Pekar, V., Orasan, C., Spurk, C. & Negri, M. (2008). Development and Alignment of a Domain-Specic Ontology for Question Answering. Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S. & Tapias D. (Eds), *the Sixth International Language Resources and Evaluation Conference*, 2221-2228.

Toma, I., Stanciu, C., Fensel, A., Stavrakantonakis, I. & Fensel, D. (20014). Improving the Online Visibility of Touristic Service Providers by Using Semantic Annotations. *The Semantic Web: ESWC 2014 Satellite Events*, Anissaras, Crete, Greece, 259-262.

Pustejovsky, J., Ingria, B., Sauri, R., Castano, J., Littman, J., Gaizauskas, R., Setzer, A., Katz, G. & Mani I. (2005). The Specification Language TimeML. *The Language of Time: a Reader*, 545-557.

Pustejovsky, J., Castano, J., Ingria, R., Sauri, R., Gaizauskas, R., Setze,r A. & Katz, G. (2003). TimeML: Robust Specification of Event and Temporal Expressions in Text. *IWCS-5 Fifth International Workshop on Computational Semantics.*

Sauri, R., Knippen, R., Verhagen, M. & Pustejovsky, J. (2005). Evita: A Robust Event Recognizer for QA Systems. *HLT/EMNLP 2005*, 700-707.

Sauri, R., Goldberg, L., Verhagen, M. & Pustejovsky, J. (2009). Annotating Events in English, TimeML Annotation Guidelines, Version TempEval-2010.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, Morgan & Claypool Publishers.

Cambria, E., Schuller, B., Xia, Y. & Havasi, C. (2013). New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intelligent Systems*, 28(2), 15-21.

Malik, M., Missen, S., Attik, M., Coustaty, M., Doucet, A. & Faucher, C. (2014). SentiML ++: An Extension of the SentiML Sentiment Annotation Scheme. *The 12th Extended Semantic Web Conference (ESWC2015)*.

Robaldo, L. & Caro, L. D. (2013). OpinionMining-ML. *Computer Standards & Interfaces*, 35 (5), 454-469.

Schrder, M., Baggia, P., Burkhardt, F., Pelachaud, C., Peter, C. & Zovato, E. (2011). Emotionml - an Upcoming Standard for Representing Emotions and Related States. *Affective Computing and Intelligent Interaction*, Springer.