A Framework for Web News Items Analysis in Relation to Company Stock Prices

Robert Max van Essen, Viorel Milea*, Flavius Frasincar
Econometric Institute
Erasmus School of Economics
Erasmus University Rotterdam
P.O. Box 1738, 3000DR Rotterdam
The Netherlands
maxvanessen@gmail.com, {milea, frasincar}@ese.eur.nl

***Abstract*—We present a general approach for Web news items analysis in relation to stock prices. The framework that we introduce provides the ability to study the impact of events extracted from news on stock prices. The relation between events and price is quantified in terms of the i) paired-samples t-test, ii) McNemar's test, and iii) confidence and support. The extraction, representation, and visualization of data are key components of the proposed framework. The validation of the framework is based on three case studies, involving Tesco, Shell, and British Petroleum, and the price reaction(s) to different news events.**

I. Introduction

The Web makes huge volumes of information available to a wide range of users. While the cost of access to information decreases, the cost of processing it poses severe challenges to the academic and business community. Research in this area is now largely focused on representational issues, addressed mainly within the Semantic Web Community. Topics of this research are, for example, the aggregation of knowledge from different Web sources and often in different formats, and the presentation of the aggregated knowledge to users, two of the focal points of Web Engineering.

News agencies have moved to the Web as the channel of choice for reaching customers faster. Companies also make press releases available through their corporate websites, and the same holds true for government agencies. Last, many blogs and opinion websites also populate the Web news landscape. The high number of news sources and the large volume of news on the Web motivate automated approaches for the processing of news. Such approaches are relevant for different areas, but have lately been studied more extensively in the area of Automated Trading, under the common denominator of News Analytics [16].

One of the main research questions at the foundation of News Analytics is how to automatically determine the impact of news messages on companies, as reflected in a company's share price. Regardless of how the impact of a news event on prices is quantified, a validation step will usually involve a study of the quantified effects of news messages and the actual price dynamics. This validation step is often not trivial, and current literature provides no systematic way of performing this analysis.

In this paper we present a framework for Web news items analysis in relation to company share prices. We are not concerned here with the actual annotation and/or sentiment analysis of news items; we assume this has been performed in a previous step. What we focus on is rather bringing together the tools and techniques that are required for such an analysis, and engineering a general framework for assessing the performance of News Analytics techniques.

The outline of the paper is as follows. In Section II we present work related to News Analytics in particular and sentiment- and content analysis in general. In Section III we introduce the framework that we propose for Web news items analysis in relation to company stock prices. Section IV presents experiments and results involving the introduced framework, with a focus on three case studies involving Tesco, Shell, and British Petroleum. We conclude and propose future research directions in Section V.

II. Related Work

Different frameworks have been developed for news (sentiment) analysis, of which some focus on news with an economic content. OASYS [2] is one such framework used for sentiment analysis of documents. The output of systems based on this framework consists of a numerical value that indicates the sentiment of the documents analysed. SemNews [8] uses news items that are automatically collected through Really Simple Syndication (RSS) feeds. The analysis determines news' semantic meaning by using a Natural Language Processing (NLP) engine called OntoSem. Hermes [1] is another framework for news analysis. Hermes is a decision support system that can classify and visualize news using a graphical interface. Hermes, just like SemNews, collects news from the Web using RSS and annotates them using a domain ontology that makes use of WordNet [14] synsets and their lexical representations. This ontology is the main difference between Hermes and SemNews, as the latter does not use a self-defined ontology, and thus has difficulties in dealing with domain specific concepts.

The framework we propose in this paper differs from SemNews and Hermes in the sense that the news stored in the database are already annotated, thus not requiring an algorithm for this purpose. These annotations can be provided by a variety of information extraction tools, as for example GATE [3], an academic product, or ViewerPro [1], an industry product.

[1] http://viewerpro.semlab.nl/

Fig. 1. News related to a company's stock price by Google Finance.

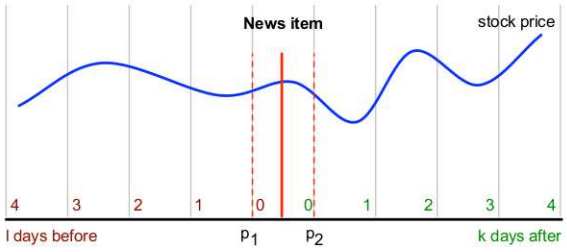

Fig. 2. Interval selection to calculate the difference in price.

In the experiments performed in this paper we have used events extracted using ViewerPro. Additionally, we focus on analysis of news items in terms of the possible relationships between a company's stock price and events.

Several online Web applications let the user view news together with stock prices. Known examples of such applications are Google Finance [5] and Yahoo! Finance [7]. Google Finance displays news items related to a company's stock price as depicted in Figure 1. When a news item is selected (such as A or B in the example depicted in Figure 1) the news item is shown on the right-hand-side of the page. The Google Finance interface allows for fast and efficient browsing through stock quotes and news items. However, no automatic analysis of (pre-) annotated news is provided. Also, the user cannot select the type(s) of news to be displayed, e.g., acquisitions, or rating changes. The framework that we propose differs from Google Finance in that it provides automated methods for computing relationships between events and the stock price. Also, we aim for any type of event to be filtered for use in the analysis.

III. The Framework

The framework that we propose focuses on the analysis of relationships between a company's *stock price* and *events*. An *event* is defined here as a type of annotation that describes economic facts or developments regarding a company. We rely on a domain ontology that defines different types of events. This ontology is made by domain experts in the field of News Analytics from SemLab [2]. One type of event defined in the ontology is, for example, *Company shares up*. This event describes an observation, in this case the fact that the shares of a given company have increased in price. An example sentence from a news message states: "Europe's biggest bank, HSBC, rose 1 percent, despite profits rising by slightly less than analysts had expected". Even though no exact reference is made mentioning an increase in share price, "rose 1 percent" is annotated and understood by a computer as the meaning behind the text. Other examples of events defined in the ontology are *Company sales up*, *Company revenue record*, *Company CEO new*, and *Company CEO resigns*. The framework we propose is not limited to certain event types, as any event can be recorded and described in the underlying domain ontology. For the remaining part of the paper, we assume that the news items considered relevant for the analysis are annotated.

We do allow news items to be gathered from various news RSS feeds. As we focus on financial news in this paper, we filter these news based on the finance channels the the original feeds provide as annotation. The events discovery is based on the ViewerPro tool, which uses lexico-semantic regular expressions for event detection.

[2] http://semlab.nl/

A. Analysis of events

In order to determine whether an event has impact on a company's stock price, every news item in which a selected event occurs is analysed. For this, a set of news items referring to the selected event are selected, as set S. For each news item in set S, we select two stock prices, one before the news item is posted and one after.

The first price, p_1, is the opening price on the day the news item is posted (or l days before that day). The second price, p_2, is the closing price on the day the news item is posted (or k days after that day). Figure 2 illustrates this news item-price relationship.

Given these two prices, a number of statistical tests can be used. For the framework we introduce, we consider two such tests: the matched- or paired-samples t-test, and McNemar's test. Finally, we analyze our results in terms of confidence and support.

1) Paired-samples t-test: The paired-samples t-test is a parametric test which is chosen because the measurement of two values of a stock price are dependent (we are measuring the stock price twice). The data used for this test is of ration type. The general formula of the paired-samples t-test is:

$$t = \frac{\overline{D}}{S_D/\sqrt{n}} \tag{1}$$

where

$$\overline{D} = \frac{\sum D}{n} \tag{2}$$

$$S_D = \sqrt{\frac{\sum D^2 - \frac{(\sum D)^2}{n}}{n-1}} \tag{3}$$

with $\overline{D}$ being the average difference in stock price and n being the sample size (in our case the total number of news items where the considered event appears). When applying this to each news item where the event is present, the average difference in stock price of l days before and k days after the news item took place are compared. We use a one-tailed test to obtain the critical value for which we compare the t-value. As a result of using a one-tailed test, the user must be able

	AFTER	
BEFORE	Stock down	Stock up
Stock up	A	B
Stock down	C	D

TABLE I. FREQUENCY TABLE FOR THE MCNEMAR'S TEST.

to select the direction in which he thinks the stock is going (up or down). Once the direction of the test is established, the critical value can be obtained and compared to the calculated t-value. If the t-value is smaller/bigger than the critical value we state that the change in price is insignificant/significant.

2) McNemar's test: The McNemar's test is a non-parametric method and can be applied to the data that we use. The data used for this test is of nominal type. The test is especially useful when comparing before/after values for a certain event. For this test we set up a frequency table as shown in Table I.

The McNemar test uses the following transformation of the χ^2 test:

$$\chi^2 = \frac{(|A - D| - 1)^2}{A + D} \tag{4}$$

with d.f.=1 (d.f. stands for degrees of freedom).

In order to make predictions for future data, we also introduce two measurement values used in data mining: confidence and support. We use these measures to create association rules, e.g., if a news item with event *CEO resigns* annotated occurs then the stock goes down. Or, in general: when a news item with event E annotated occurs then company stock goes significantly in a certain direction d.

3) Confidence & Support: To compute the confidence we first define N_A as the number of times an event E occurs and N_B as the number of times the stock goes up (or down, depending on the direction of the test). Now we can define the confidence as the number of times that N_A and N_B appear together divided by N_A. Or, more formally:

$$confidence = \frac{N_{A,B}}{N_A} \tag{5}$$

where $N_{A,B}$ = number of times A and B happen together.

The confidence is an indicator of how strong the relationship is between an event and the direction of movement of the stock price. To compute the support we introduce the number of days the selected period contains, N. Now we can define the support as the number of times that N_A and N_B appear together divided by N, or:

$$support = \frac{N_{A,B}}{N} \tag{6}$$

If the support is large, we can say that there is a solid base we draw our conclusions on. If the support is small, the conclusion is only valid for a small portion of the data so we should be careful with generalizing the conclusion.

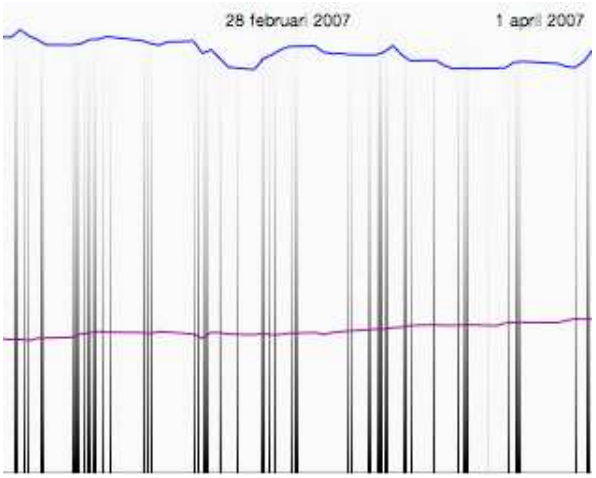

Fig. 3. Displaying news items related to stock prices.

B. Parameters

The following parameters can be selected by the user in order to perform the analysis by using the framework that we introduce:

- The timeframe;
- A company which is quoted on a stock exchange;
- The event to analyse;
- The direction of the tests;
- The l and the k number of days for the calculation of p_1 and p_2.

C. Visualizing news items

To visualize the news items, the open source application Timeplot [9] is used, a tool developed within the SIMILE Project [15]. SIMILE focusses on creating robust, open source tools for visualizing time series and events. In Timeplot the news items are being displayed on a time-scale using vertical bars. Every bar represents a single news item, as shown in Figure 3. The user is able to click on a bar in order to read the whole news item. When viewing the text, every annotated event is highlighted, as shown in Figure 4. Additionally, the Uniform Resource Identifier (URI) of the events is shown as well.

D. User interface

The main part of the GUI is the Timeplot which displays the news items. As described earlier, news items are represented as vertical bars. The user can click on each bar to display the text of the news item. Additionally, the Timeplot displays the stock price of a company. The stock data is gathered online from Yahoo! Finance. For most companies available at Yahoo! Finance, historical data can be downloaded as a comma-separated (CSV) file. We use the daily closing prices to create a graph for analysing the impact of events on the stock price. Once the data is fetched, the graph is laid over the news items to create a visualization of news related to stock prices.

Fig. 4. Annotated events are highlighted when viewing the news item.

Figure 3 displays a selection of news items related to pharmaceutical AstraZeneca. As an additional feature, the user can type in a stock quote of a random company and display it as a second line on the Timeplot. In this case, another pharmaceutical, King Pharmaceuticals, is selected. This can be useful when analysing the effect an event relating to a company has on, for example, the company's competitors. When the users hovers over the graphs, the price of the stock and the current date are displayed.

E. Data retrieval

To provide a fast and responsive user-interface, Asynchronous JavaScript and XML (AJAX) is used. The advantage of using AJAX is that every communication with the server is done in the background and the page does not have to reload in order to display new information. While the technique behind AJAX originated in 1998, the term was introduced in 2005 [4] and is now a W3C Working Draft [21]. The Ajax on the client-side of the application communicates with an Apache Tomcat [20] server. Apache Tomcat allows using Java classes, for which we use the Sesame [17] database and API to retrieve the data. The data is stored as persistent RDF data and is based on RDF triples.

To search through the data for companies, events, and news items, we rely on SPARQL queries [19]. For example, the SPARQL query of a search for *ECB* (European Central Bank) looks as follows:

```
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX icb:<http://www.semlab.nl/ICB.owl#>
SELECT distinct ?uri ?name
WHERE {
    ?uri icb:hasSynonym ?syn .
    ?uri rdfs:label ?name .
    FILTER regex(str(?syn), "ECB", "i")
}
```

For each company in the ontology there are synonyms of company names defined. This query deliberately searches through the collection of synonyms, so that if you search for *ECB*, the results also include matches for *European Central Bank* and other synonyms.

The user can browse through the data by applying filters on events. Any event in the ontology can be selected. If a user wants to search for event *Company shares up* in the year 2007, the SPARQL query is defined as follows:

```
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX point: <http://semlab.nl/1.0/point#>
PREFIX event: <http://www.semlab.nl/1.0/event.owl#>
SELECT distinct ?p ?date ?title {
    ?y event:hasEvent event:companySharesUp_i .
    ?y point:hasNewsItem ?p .
    ?p dc:date ?date .
    ?p dc:title ?title
    FILTER (?date >= "2007-01-01T00:00:00.000
     +01:00"^^xsd:dateTime
         && ?date <= "2007-12-31T23:59:59.000
          +01:00"^^xsd:dateTime)
} ORDER BY ?date
```

The result of the query above consists of all the news items in the given time period in which have the event (in this case an event of type *Company Shares Up*) annotated in its text.

F. Event impact

To analyze events for a possible relationship with a company's stock price, an *event analyzer* is implemented. The user can analyse a random event from the domain ontology over a certain time period. For every news item on time t, the application looks at the opening price of that day and the user is able to select the number of days that the system should look ahead to calculate the difference in price. Once this time-frame is specified, the system calculates if there has been a significant change in stock price. To draw a conclusion, some statistics are applied to the data. Currently implemented are the paired-samples t-test, the McNemar's test, and confidence and support (for detailed explanations of these terms see Section III-A).

IV. Experiments & Results

Hypothesis testing is built within the framework for the different tests described in Section III-A. When the user selects, through the interface, the *Compute Event Impact* button, the user must first indicate the direction of the test (stock going up or down). The system then draws and displays a conclusion based on the following hypothesis:

H_0: There is no difference in the stock price of a company when event E occurs.

H_A: The stock price of company C significantly changes when event E occurs.

For all statistical tests we chose an α of 0.01, a more conservative value than the usual 0.05.

For drawing conclusions, we aim for support from at least the t-test or McNemar test and high level of confidence (support plays a lesser role here as this depends very much on the chosen sample).

A. Tesco Case Study

The first case study relates to a supermarket retailer in the UK called Tesco. We are interested in the performance of the stock price of this company. For the time-frame we choose the first four months of 2007 and analyse the event *Company Shares Up*. We choose *stock going up* as the direction of the test since we expect the stock to go up following events of this class. We select the l and the p values as 0 days for both values, meaning that we compare the opening and the closing price of the same day.

We obtain the following results for the t-test: the calculated value t is 3.53 and the critical value is 2.45. The calculated

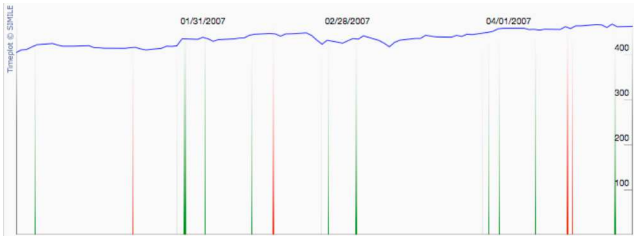

Fig. 5. News items are colored red (dark gray) when the stock went down and green (light gray) when it went up.

value t is bigger than the critical value so we reject the null hypothesis and conclude that the upward change in the stock is significant. The McNemar's test shows that the calculated value is 4.36 with a critical value of 6.64. According to this test we conclude that the change in price is not significant.

The support and the confidence are also calculated. The support is 0.19 and the confidence is 0.70. The confidence tells us that in 70% of the news items with the event selected the stock went up. The support tells us that in 19% of the total number of days there is a news item with the event annotated in which the stock went up. Despite this value being small, Figure 5 shows that the news items are well spread over the time-period so we do not find a low support very important in this case. By looking at the t-test and the confidence we can state that the impact of this type of event on the stock price of Tesco is significant and define that when this event happens, the stock price does indeed go up.

B. Royal Dutch Shell Case Study

The second scenario focusses on Royal Dutch Shell. We choose the same system setting as in the previous case study (first four months of 2007 and 0 for the l and p values). For this scenario we select the event *Company Shares Down*. Because of the nature of this event we expect that the stock goes down. This event is also an observation so we expect this event to have a significant (negative) impact on the stock price of Royal Dutch Shell. The calculated value that we obtain for t is -9.34. This value is smaller then the critical value (which is -2.37) so we reject the null hypothesis and conclude that the impact is significant.

The support is 0.68 and the confidence is 0.82. These numbers are high enough to allow us to draw the conclusion that the event *Company Shares Down* has a significant impact on the stock price of Royal Dutch Shell and we can state that when this event occurs the stock price will most likely go down.

By using the same input values as before, we also found some other interesting results. For example, looking at Royal Dutch Shell in combination with the event *Joint Venture* and *Collaboration Consideration*, we found the values shown in Table II.

While the changes in both cases are not significant, we can clearly see a negative relationship between the stock price and the events. While the confidence of *Joint Venture* is not very high, the confidence of *Collaboration Consideration* reaches a relatively high level. Based on these values we can state that the event *Collaboration Consideration* has a negative impact on the stock price of Royal Dutch Shell.

	Joint Venture	Collaboration Consideration
t-value	-1.66	-2.47
t-critical	-2.52	-3.75
McNemar	0.41	2.25
McNemar critical	6.64	6.64
Support	0.11	0.03
Confidence	0.59	0.80

TABLE II. SHELL RESULTS.

	Cost Cut	Collaboration Consideration
t-value	-1.26	-5.12
t-critical	-3.75	-3.36
McNemar	0.80	4.17
McNemar critical	6.64	6.64
Support	0.03	0.05
Confidence	0.80	1.00

TABLE III. BP RESULTS.

C. British Petroleum Case Study

Finally, we take a closer look at another oil company: British Petroleum (BP). We use the same settings as before and we see some interesting results for the events *Cost Cut* and *Collaboration Consideration*, as shown in Table III.

We observe again two events that have a negative impact on the stock price. The impact of *Cost Cut* is not significant, but looking at the confidence we can see that in 80% of all cases the stock indeed went down. Looking (again) at the *Collaboration Consideration* event, we note that there is a significant change in stock price and a 100% confidence. We conclude that the events *Cost Cut* and *Collaboration Consideration* both have a negative impact on the stock price of BP, and when these events arise in a news item the stock price will go down more often than not.

V. CONCLUSIONS & FUTURE WORK

In this paper we presented a framework for Web news items analysis in relation to stock prices. Based on different input data sources, our framework enables visualization and statistical analysis of the impact of events on prices. The three case studies presented in this paper, involving Tesco, Shell, and British Petroleum, show the ease of use of the introduced framework and its usefulness within News Analytics approaches. The framework is based on AJAX technology to enable fast communication between the client and the server and various statistical tests to support the experiments to be carried.

The framework can further be extended to include different event extraction methods. This would enable running, testing, and comparing these methods against each other on different datasets. In this way, it could be assessed for which particular cases different event extraction algorithms are more suited. For example we could replace the lexico-syntactic rules with the more advanced lexico-semantic rules that we have recently developed [6]. Sentiment- and content analysis approaches within News Analytics, such as [10], [13], could also be implemented within the framework. This will allow furhter comparison of different methods and the ability to aggregate different approaches for the purpose of trading signal generation.

Other extensions can be thought of, such as using different knowledge representation languages, e.g., OWL [18] or tOWL [11], [12], for building a common knowledge base of events, impacts of events, and share prices. Modelling the temporal constraints between the entities participating in the experiment would allow us to better specify the temporal context and results of the performed experiments.

ACKNOWLEDGEMENT

We would like to thank SemLab, a small but innovative company in the Netherlands working, amongst others, on News Analytics, for their support during this research. We are especially grateful to dr. Mark Vreijling for the very fruitful discussions and for his help in obtaining the presented results.

REFERENCES

[1] J. Borsje, L. Levering, and F. Frasincar. Hermes: a semantic web-based news decision support system. In *23rd Annual ACM Symposium on Applied Computing (SAC 2008)*, pages 2415–2420. ACM, 2008.

[2] C. Cesarano, B. Dorr, A. Picariello, D. Reforgiato, A. Sagoff, and V. Subrahmanian. Oasys: An opinion analysis system. In *AAAI 2006 Spring Symposium on Computational Approach to Analyzing Weblogs*, pages 21–26. The AAAI Press, 2006.

[3] H. Cunningham. Gate, a general architecture for text engineering. *Computers and the Humanities*, 36(2):223–254, 2002.

[4] J. J. Garrett. AJAX: A new approach to Web applications, 02 2005. http://www.adaptivepath.com/ideas/essays/archives/000385.php.

[5] Google. Google Finance, 2013. See http://finance.google.com, last visited November 2013.

[6] W. IJntema, J. Sangers, F. Hogenboom, and F. Frasincar. A Lexico-Semantic Pattern Language for Learning Ontology Instances from Text. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, 15(1):37–50, 2012.

[7] Y. Inc. Yahoo finance, 2013. See http://finance.yahoo.com/, last visited July 2012.

[8] A. Java, T. Finin, and S. Nirenburg. Semnews: A semantic news framework. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI 06)*, pages 1939–1940. American Association of Artificial Intelligence, 2006.

[9] S. Mazzocchi and SIMILE project. Timeplot, 2013. See http://simile.mit.edu/timeplot/, last visited November 2013.

[10] V. Milea, R. Almeida, U. Kaymak, and F. Frasincar. A fuzzy model of a european index based on automatically extracted content information. In *Computational Intelligence for Financial Engineering and Economics (CIFEr), 2011 IEEE Symposium on*, pages 1–8. IEEE, 2011.

[11] V. Milea, F. Frasincar, and U. Kaymak. tOWL: A temporal web ontology language. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 42(1):268–281, 2012.

[12] V. Milea, F. Frasincar, U. Kaymak, and G. Houben. Temporal optimisations and temporal cardinality in the towl language. *International Journal of Web Engineering and Technology*, 7(1):45–64, 2012.

[13] V. Milea, N. Sharef, R. Almeida, U. Kaymak, and F. Frasincar. Prediction of the MSCI EURO index based on fuzzy grammar fragments extracted from european central bank statements. In *International Conference of Soft Computing and Pattern Recognition (SoCPaR 2010)*, pages 231–236. IEEE, 2010.

[14] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244, 1990.

[15] MIT Libraries and MIT CSAIL. The SIMILE Project, 2013. See http://simile.mit.edu/, last visited Movember 2013.

[16] G. Mitra and L. Mitra. *The handbook of news analytics in finance*. John Wiley & Sons, 2011.

[17] openRDF.org. Sesame, 2013. See http://www.openrdf.org/, last visited November 2013.

[18] P. Patel-Schneider, Hayes, and I. P., Horrocks. Web ontology language (OWL) abstract syntax and semantics. *W3C Recommendation*, 2004.

[19] E. Prud'hommeaux and A. Seaborne. SPARQL Query Language for RDF. W3C Recommendation, 2008.

[20] The Apache Software Foundation. Apache tomcat, 2013. See http://tomcat.apache.org/, last visited November 2013.

[21] A. van Kesteren. The XMLHttpRequest Object. W3C Working Draft 15 April 2008, April 2013. http://www.w3.org/TR/XMLHttpRequest/.