PROJECT PERIODIC REPORT
PUBLISHABLE SUMMARY

Grant Agreement number: ICT 316404
Project acronym: NewsReader
Project title: Building structured event indexes of large volumes of financial and economic data for decision making
Funding Scheme: FP7-ICT-2011-8
Date of latest version of Annex I against which the assessment will be made:
Periodic report: 1st ☒ 2nd ☐ 3rd ☐ 4th ☐
Period covered: from 1/1/2013 to 1/1/2014
Name, title and organisation of the scientific representative of the project's coordinator :
Prof. dr. Piek Vossen , Faculty of Arts, VU University Amsterdam
Tel. + 31 (0) 20 5986466
Fax: Fax. + 31 (0) 20 5986500
E-mail: piek.vossen@vu.nl
Project website address: `http://www.newsreader-project.eu`

1 Project summary and objectives

NewsReader develops software that automatically reads daily streams of news in 4 language (English, Spanish, Italian and Dutch). For each news article, it determines *what* happened, *where* and *when*, and *who* was involved. NewsReader will not just read a single newspaper each day but massive amounts of news articles coming from thousands of different sources. Furthermore, it will merge the news of today with previously stored information, creating a long-term history rather than storing separate events. Whereas news comes in every day in piecemeal fragments, the **history recorder** developed in NewsReader eventually creates a single condensed story from all these fragments, very much in the same way that humans do by putting information together and de-duplicating redundant information. Unlike humans, however, NewsReader will not forget any detail, be able to recall the complete story as it was told, know who told what part of the story, and identify what sources contradict each other. Unlike humans, NewsReader can process millions of articles in different languages from a large variety of sources.

The extracted information is stored in a KnowledgeStore that supports formal reasoning and inferencing on the knowledge. Over time, NewsReader will record the history as it was told and perceived in the media in the KnowledgeStore. Since this **history recorder** keeps track of all the origins of information, it provides valuable insights into how the story was told. This will tell us about the different perspectives from which different media sources present the news, both the news of today and the news of the past. The data produced in NewsReader is extremely large but also complex: exhibiting the dynamics of news coming in as a stream of continuously updated information with changing perspectives. A dedicated decision support tool suite is developed that can handle the volume and complexity of this data and allows professionals to interact through visual manipulation and feedback and new types of representation.

NewsReader is a collaboration of three European research groups and three companies: LexisNexis, ScraperWiki and SynerScope. The project started on January 2013 and will last 3 years. NewsReader will be tested on economic-financial news and on events relevant for political and financial decision-makers. About 25% of the news is about finance and economy. LexisNexis estimates the total volume of daily news items for this domain on about five-hundred-thousand.

2 Progress in the first year

In the first year of the project, we carried out a full project cycle consisting of the following steps:

1. user-requirements analysis
2. system design and architecture
3. development of benchmark data

4. software development
5. data processing
6. end-user application development
7. end-user evaluation pilot

The user-requirements analysis was carried out through intense collaboration between the industrial and academic partners, involving interviews with the specialists at LexisNexis that directly deal with the end-user customers and a range of smaller project meetings focusing on the user perspectives. We defined 4 use-cases:

1. TechCrunch/Crunchbase: a well-structured wiki-like database (Crunchbase) and news documents (TechCrunch) on the same topic: information technology. We anticipate that events in structured Crunchbase data will be reflected in non-structured TechCrunch articles;
2. the Dutch House of Representatives: focusing on information-intensive Parliamentary Inquiries, we identified several challenges:
   (a) coverage: 'understanding' an event, its key actors and entities
   (b) mapping the gaps: identifying areas with insufficient information coverage
   (c) creating networks of events, people and entities (companies, government bodies)
   (d) fact checking
3. Global Automotive Industries: using a large, multilingual data set on the automotive industry, NewsReader will help (re)construct:
   (a) complex structures, i.e. the ownership structures of automotive conglomerate
   (b) complex events, i.e. mergers, acquisitions and corporate restructuring in this industry
   (c) Business Intelligence: gathering information about companies to evaluate them as potential business partners. This evaluation may relate to a business' ability to repay a loan, to comply with anti-money laundering legislation or to carry out regular due diligence investigations.

The project achieved ground-breaking results by designing shared representation formats for both representing Natural Language Processing (NLP) output (the NLP Annotation Format, NAF)[1] and Semantic Web content related to events (the extended Simple Event Model, SEM+).[2] This allows us to make a fundamental distinction between the

[1] `http://wordpress.let.vupr.nl/naf/`
[2] `http://wordpress.let.vupr.nl/sem/`

semantic representation of mentions of events and participants in the text and instances of events and participants in the formal triple representation in SEM+. We launched the Grounded Annotation Framework (GAF)[3] to establish the relation between the two paradigms. This also enabled us to include a provenance model according to the principles defined in PROV-O,[4] the W3C community recommendation for modeling provenance. This elaborate modeling, which is unique in its kind at this scale and complexity, enabled us to create the complex NLP pipelines for processing the textual data, aggregating the textual representation to an instance-based representation in SEM+ as well as designing the KnowledgeStore for storing the result.

The system design and architecture provided the roadmap for the successful implementation of a range of 15 NLP modules for English but also for Dutch, Spanish and Italian, and the implementation of a database, the so-called KnowledgeStore, for storing the results. We were able to package the NLP modules in a Virtual Machine (VM) that was shared across the partners. We deployed 8 parallel copies of the VMs to process over 60K news articles on the car industry and 40K articles from TechCrunch. The processed data have been loaded into the KnowledgeStore which now provides access to 40 million triples (statements) on the car industry covering a period of 10 years (2003 till 2013).

Visualisation of this complex data and the relation is extremely important. Figure 1 shows an example of such a visualisation in a time line. It is automatically generated from data produced for documents on Volkswagen and its factories in the Belgium city Vorst. Around 2006, Volkswagen decided to move production factories from Pamplona to Vorst and a few years later from Vorst to Germany and other countries. This had a big social-economic impact on Vorst. The image shows participants connected through events as story lines in time, as extracted from the news. This visualisation is based on about 50 documents with shallow processing. When large volumes of data are processed, more complex graph-visualisations are needed.

Figure 2 shows the big-data visualisation through the Marcato tool developed by SynerScope and applied to the Crunchbase data. It shows a table view of investors in an IT company but also a graph view on all connections. Through filtering, selection and expansion, users can find correlations and patterns.

In January 2014, we carried out a first task-based user evaluation. We asked 15 participants to answer questions using the Marcato tool on the Crunchbase data, using the LexisNexis standard interface to do filtered search on their archives and using free Google search. The results of this pilot evaluation are preliminary but they will help the design and organisation of a larger evaluation in the 2nd year of the project. The evaluation already provided useful input for improving the Marcato tool.

The project had some great successes in terms of dissemination and extension. We launched the concept of a *history recorder*, a machine that can read all the news in different languages and reconstruct the history over longer stretches of time as it has been told in the media without loosing any details. This turned out to be an inspiring concept that

[3] `http://groundedannotationframework.org`

[4] `http://www.w3.org/TR/prov-overview/`

resulted in a lot of media attention for the general public (see the interviews and talks given) but also from a broader e-humanities perspective (e.g. historians). The concept of a 'history recorder' contributed to winning two prices in 2013:

1. Piek Vossen received the NWO Spinoza award in 2013, which is the highest Dutch award in science. NWO awards the prize to Dutch researchers who rank among the absolute top of science. The prize was awarded for his ground-breaking work on wordnets and NewsReader, specifically the concept of a 'history recorder'.[5]

2. The consortium won the Enlighten Your Research (EYR) prize of 2013 for their proposal "Can we handle the news?"[6]

The EYR prize allows us to turn NewsReader into a production environment and carry out additional performance research using the best hardware infrastructure available, with support from a team of experts in big data processing. The Spinoza award is used for a number of follow-up PhD projects that look deeper into the concepts of storylines, world views and opinions expressed in news and contextual reasoning and interpretation processes to support NLP.

3 Final results, use and impact

The project has large exploitation potentials coming from *the data* we produce, the *technology* we develop and the *framework* we have designed. The concept of a *history recorder* raised a lot of interest and publicity. The data in this recorder can be used to investigate how different sources report on events, people and companies in the news. Research can be carried out on how sources select what they publish and what they say about it providing insight into how news influences our perspective on events.

The methodology of processing large volumes of textual data can be applied to a wide range of sources, not only news. It can be seen as a new way of indexing and compacting textual data that report on changes in the world over longer periods of time. Our technology can be easily extended to many more languages, which enables large-coverage comparison of the ways people write about these changes across languages and cultures. It can thus be used to support a wide range of research in humanities.

Our framework (GAF) also has the potential of integrating non-textual sources into the same model. The interpretation of events can thus be extended to sensors, databases and multimedia recordings.

[5] http://www.nwo.nl/en/research-and-results/programmes/spinoza+prize
[6] http://www.surfsites.nl/eyr/

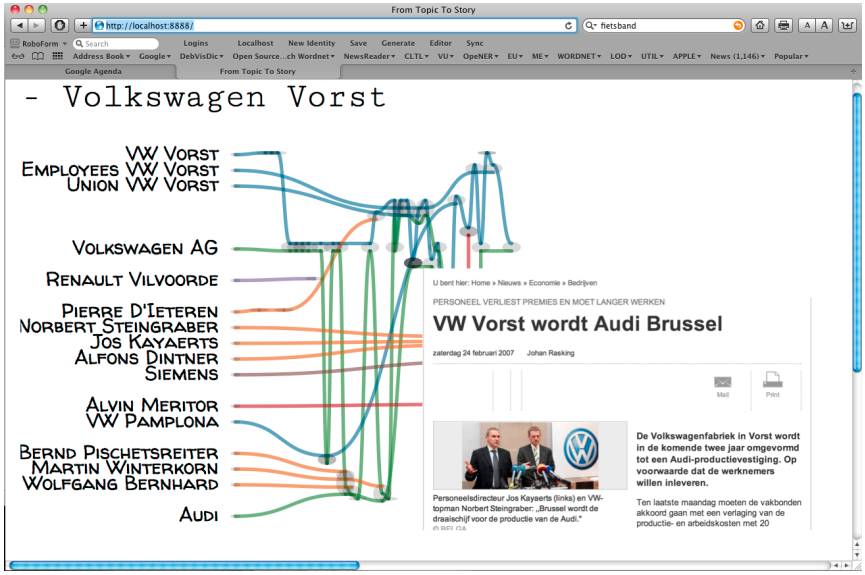

Figure 1: Storyline visualisation generated from automatically processed data on Volkswagen factory in Vorst Belgium

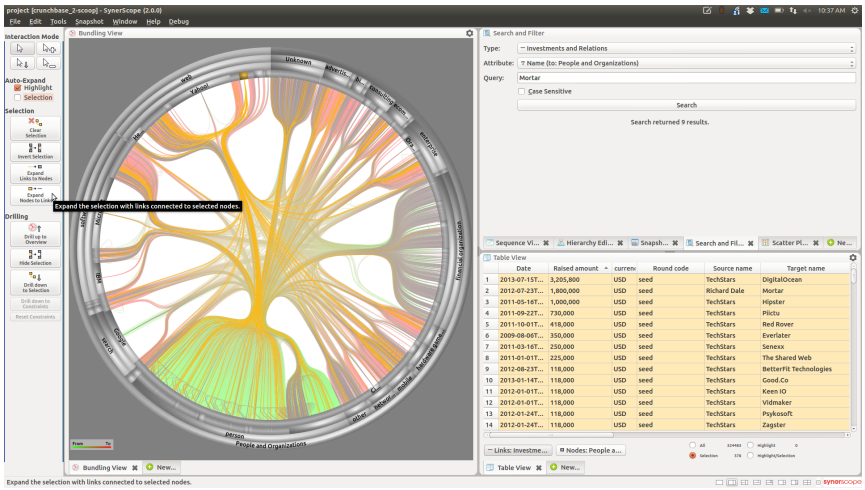

Figure 2: Screen dump of the Marcato tool for analysing complex graphs

4 Consortium and contacts

Partner	Country	Contact	Email
(Coordinator) Faculteit of Arts, Vrije University Amsterdam	Netherlands	Piek Vossen	piek.vossen@vu.nl
Euskal Herriko Unibertsitatea San Sebastian	Spain	German Rigau	german.rigau@ehu.es
Fondazione Bruno Kessler Trento	Italy	Luciano Serafini	serafini@fbk.eu
LexisNexis, Amsterdam	Netherlands	Pim Stouten	pim.stouten@lexisnexis.com
ScraperWiki, London	United Kingdom	Aidan Mcguire	aidan@scraperwiki.com
SynerScope, Eindhoven,	Netherlands	Willem van Hage	willem.van.hage@synerscope.com

Table 1: Consortium members and contacts