NRG4CAST
FP7-2012-NMP-ENV-ENERGY-ICT-EeB
Contract no.: 600074
www.nrg4cast.org

NRG4Cast

Deliverable D3.3

Metadata Generation Prototype

Editor:	Klemen Kenda, JSI
Author(s):	Patrik Zajec, Klemen Kenda, Marko Grobelnik, JSI
Deliverable Nature:	Prototype (P)
Dissemination Level: (Confidentiality)[1]	Public (PU)
Contractual Delivery Date:	November 2014
Actual Delivery Date:	December 2014
Suggested Readers:	Developers creating software components to be integrated into final tool for different users, developers creating models, prediction tools, etc.
Version:	1.0
Keywords:	N-grams, taxonomy, semantics, knowledge acquisition, knowledge extraction, energy markets, energy price prediction.

[1] Please indicate the dissemination level using one of the following codes:
• **PU =** Public • **PP =** Restricted to other programme participants (including the Commission Services) • **RE =** Restricted to a group specified by the consortium (including the Commission Services) • **CO =** Confidential, only for members of the consortium (including the Commission Services) • **Restreint UE =** Classified with the classification level "Restreint UE" according to Commission Decision 2001/844 and amendments • **Confidentiel UE =** Classified with the mention of the classification level "Confidentiel UE" according to Commission Decision 2001/844 and amendments • **Secret UE =** Classified with the mention of the classification level "Secret UE" according to Commission Decision 2001/844 and amendments

Disclaimer

Executive Summary

Fact extraction from the available on-line streams of textual data is a task, which does not yet have an effective solution in the industry. Solution, which is being built through the textual pillar of the NRG4CAST project is a unique one and it offers quite some application and exploitation possibilities.

Fact extraction today is mostly done by human experts. This means that the acquired knowledge is very precise, but only a minor amount of available facts is covered. Our technology is able to extract domain specific knowledge with a great recall. Although the expected precision cannot be compared to the human-based fact extraction, the tool would still produce results of a great value. They could either be used independently or as an efficiency booster for the human experts.

More than 20.000 of World Wide Web publishers are included in the Newsfeed. The raw news data is then clustered, categorized and further enriched within the EventRegistry solution. Such a stream is then used for domain specific factoid extraction in the presented prototype. Factoid templates are extracted from the Google N-Gram database and cross-linked with the Yahoo Financials data to get the intersection of the market stakeholders. Results are presented in a stream of matching events and templates, used for their extraction.

Table of Contents

List of Tables

No table of figures entries found.

List of Figures

Abbreviations

API	Application programming interface
DoW	Description of work document
GUI	Graphical user interface
KB	Knowledge base
WP	Work package, as stated in DoW
Q/A	Question/Answering

1 Introduction

NRG4CAST task T3.3 (Semantic Enrichment) is highly dependent on the results from the task T3.2 (Metadata Generation Prototype). It furthers the work on the textual pipeline of the NRG4CAST project. The pipeline represents a side pillar of the NRG4CAST infrastructure and is depicted in Figure 1.

The pipeline is built on top of EventRegistry[2] data (and that one on top of a NewsFeed[3]), which represents a stream of aggregated news, representing significant events, detected in World Wide Web media. Another important input data is the vast dataset of Google N-Grams [3], which has been used in T3.2 with a combination of Yahoo Financials [2] data to extract the most common phrases (templates), relevant for energy markets.

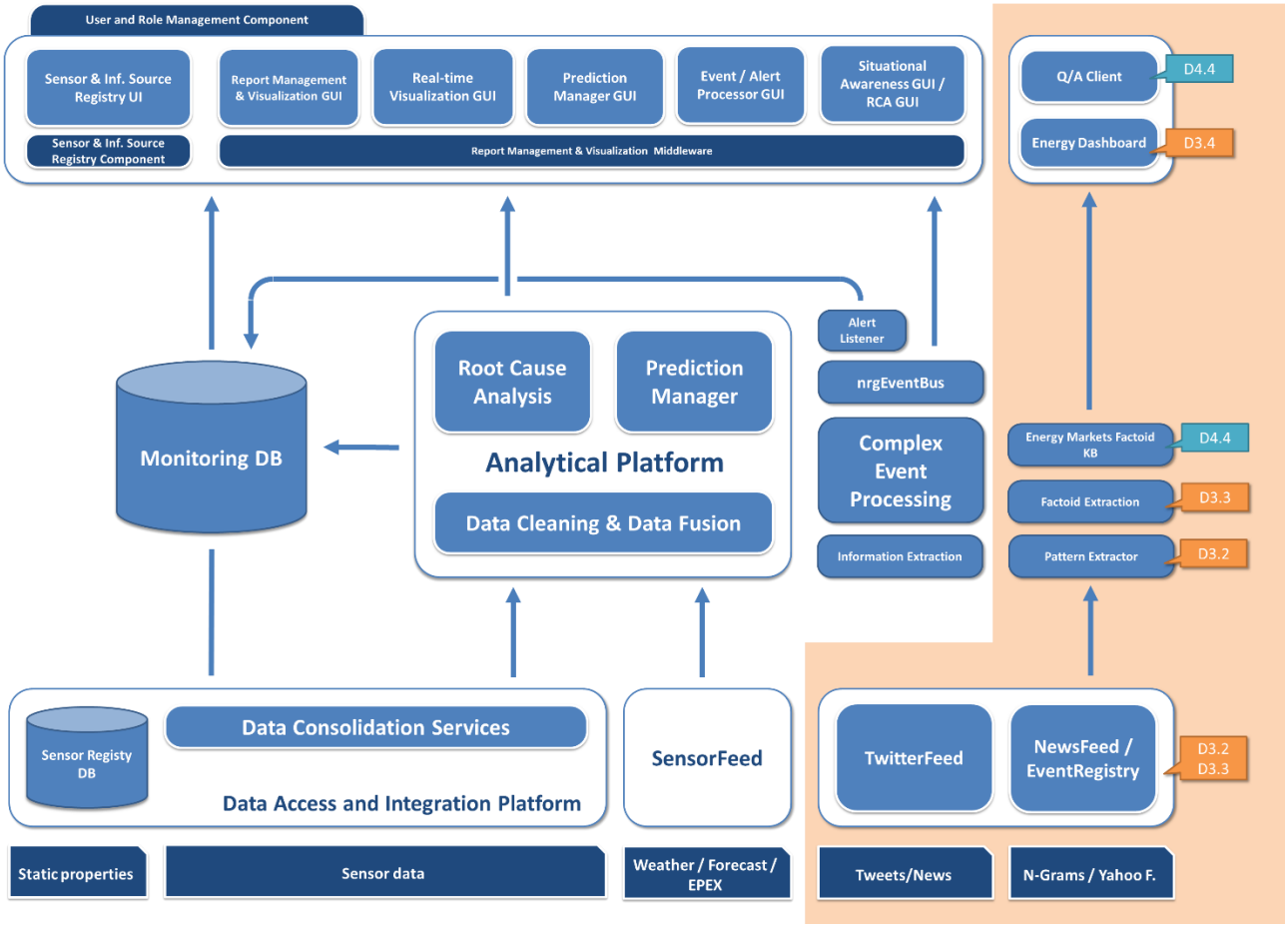

Figure 1: Textual vertical in the NRG4CAST platform.

This work has been done in the task T3.2, but has been significantly improved in the prototype reported in this deliverable. **The efficiency of the template extraction has been improved by a factor of ~10^7.** The new solution includes state of the art algorithms on non-distributed handling of big datasets. This basically means that the solution has been improved, so that we can extract templates for factoid extraction on-the-fly.

We have connected this component to the factoid extraction module, which is able to compare the templates against the stream from the EventRegistry. Extracted factoids will be used for generation of an Energy Market Factoid Knowledge Base in the task D4.4, which will also offer an API and Q/A Client for knowledge representation.

[2] http://eventregistry.org/
[3] http://newsfeed.ijs.si

This prototype is expected to help with improving the prediction of the prices of energy (European energy spot market – EPEX - use-case), which is mostly relevant for the Miren pilot (for real-time energy trading purpose), but also Aachen, Athens, and Turin pilots (electric vehicles and public buildings) could benefit from it. Beside the improvement of energy price predictions, results of factoid extraction could be used within a knowledge base, which could provide a helpful tool for monitoring energy markets.

1.1 Connection to the Other Parts of the Project

The "textual pillar" in the NRG4CAST architecture is an independent application for fact extraction and knowledge base building. The whole vertical represents a unique application, which is able to gather factoids on different topic with a great recall. The trade-off for recall is, of course, precision. There are a few products in the market (for example Bloomberg tools[4]), which offer similar functionality, but most of these solutions do not offer automatically extracted data (the data is provided by a human). Such data has very good precision, but rather low recall. There certainly is a niche in the market for such a product. It could serve as a standalone solution or as a recommender for the human annotators.

Within the NRG4CAST project the whole textual part is also connected to the prediction manager. With the use-case of European energy spot market (EPEX), where we try to predict energy prices for 1 day ahead, the features, extracted from the text could serve as an additional source of features for modelling. The actual value of factoid extraction for prediction purposes will be tested in the NRG4CAST Deliverable D5.2 (Data Driven Prediction Methods Environment).

The deliverable is strongly connected to its predecessor, NRG4CAST Deliverable D3.2 (Semantic Enrichment) [1]. The work of this deliverable upgrades implementation, done for D3.2, and extends it with a Factoid Extraction Component. The results will be used as input data for the prototype for NRG4CAST Deliverable D4.4 (Knowledge formalization services).

Metadata generation in the sense of the extension of the input features for modelling and reasoning has been reported in the NRG4CAST Deliverable D3.1 (Modelling of the Complex Data Space). Features like working hours, day of the week, holidays, days before holidays, moonrise, moon phase, sunset and similar have been reported.

1.2 Composition of the Deliverable

Section 2 briefly describes the architecture of the overall system, Section 3 discusses innovative methods and approaches used to tackle the problems of "big data" in the task and Section 4 offers some additional implementation details and API descriptions. Section 5 gives an overview of the work done in NRG4CAST Task T3.3 and explains further steps in developing the fact extraction application.

[4] http://www.bloomberg.com/professional/

2 Architecture

The architecture has been explained in detail in the NRG4CAST Deliverable D3.2 [1]. We include a brief summary in this deliverable to ensure completeness.

The aim of the architecture is to build a domain-specific knowledge base from a stream of textual data. In the NRG4CAST project we would like to have a solution, which would build a knowledge base, related to the Energy Market.

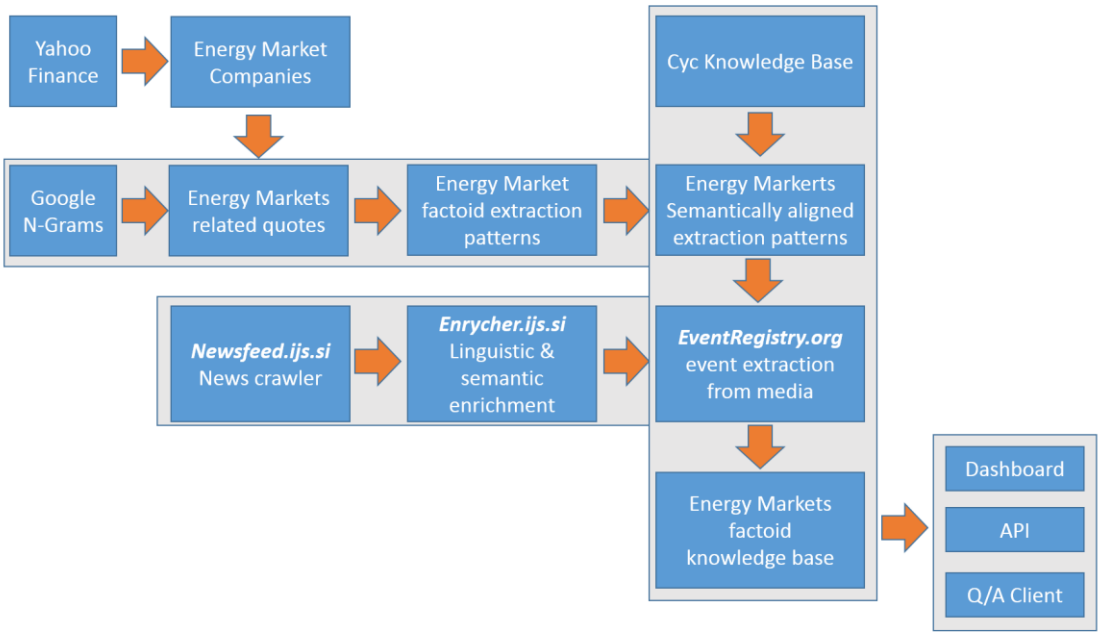

Figure 2: Pipeline of information flow for deriving "Energy Markets" related knowledge factoids.

Figure 2 depicts the whole infrastructure. With the NRG4CAST Deliverables D3.2 (Semantic Enrichment) and D3.3 (Metadata Generation Prototype) we have implemented all of the components in the upper 3 levels. The prototype is able to intersect energy markets data with Google N-Grams in an efficient way. We are able to identify most common patterns present in the google N-Grams. Factoid are partially semantically aligned (based on verbs) and the factoid extraction is already being executed on a stream of news data from all over the world on top of EventRegistry stream.

3 Methods and Approaches

3.1 N-Gram Database

The work, reported in this deliverable is based on experience, gained with preparation of the prototype NRG4CAST Deliverable D3.2 (Semantic Enrichment). One of the great weaknesses of the prototype, as reported in D3.2, was the speed of performing queries on the huge dataset of Google N-Grams. The system needed a couple of hours to perform a single set of queries. In this deliverable we have rewritten this component completely. The performance was improved significantly and the current system is able to perform 10^5 queries/second.

The solution – N-Gram Database, which enables such a fast performance, cannot be based on any of the traditional solutions. An alternative for the task could always be an Apache Hadoop based component, but due to simplicity and the fact that the N-Gram Database can nowadays be handled by a single machine, we developed our own database. The trade-off for the speed of our prototype is a big amount of used memory.

The prototype N-Gram Database supports arbitrary wildcard queries. Wildcard query is a query such as "consuming * energy", where the results might include:

1. "consuming thermal energy"
2. "consuming wind energy"
3. "consuming alternative energy"
4. …

To ensure fastest wildcard searching possible, the n-grams need to be indexed as efficiently as possible. A naïve idea to start with would be: For n-grams we would keep $n!$ different prefix trees (also referred to as *tries*[5]). A prefix tree is an ordered tree data structure that can be used to store n-grams. The first sub-node would represent the first word, the second degree sub-node would represent the second word, etc. Results of a particular query would be all the n-grams in the sub-tree of a matched prefix. As we do want a wildcard search, we need to be able to search over all of the permutations. At a first glance, we would need 5! = 120 prefix trees for our n-grams.

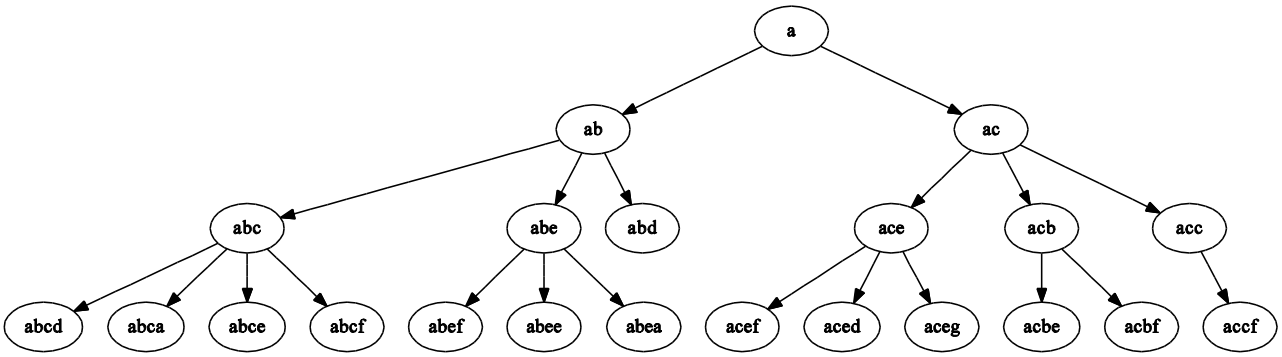

Figure 3: An illustrative trie example (nodes contain words, not just letters).

Figure 3 depicts an example of a trie. For visualization reasons it represents the words (n-letters) and not n-grams, but it illustrates the logic. If this tree would correspond to the permutation (1, 3, 2, 4), it could support, for example, queries of a form "a * b *". All the results would lie in the left side of the figure, under the node "ab" (1, 3).

The search problem is therefore reduced to a descent down the trie and can be described in three steps:

1. Every wildcard query is translated into a sequence of a non-wildcard grams.
2. According to the selected permutation (of non-wildcard grams) we select a proper tree.

[5] http://en.wikipedia.org/wiki/Trie

3. We descent down the tree as long as all n-grams are used; requested results are located in a sub-tree below.

Such a technique would enable very fast search over n-grams, but it would take too much space (the index size would be 120-times the size of the n-gram dataset). Careful analysis shows that for representing all the n-grams from (1-grams to 5-grams) we only need 10 different prefix trees, as not all the permutations are needed.

In the NRG4CAST n-gram database the trie-s are implemented as ternary search trees [4], so they can efficiently be represented in the memory.

The prototype index uses 10 ternary search trees, where each tree represents an operation. Index building took 15 hours on a local NRG4CAST machine (with 512 Gb of RAM). Additionally, various compression methods had been used on the index, decreasing its size to around 280 Gb, which can be fully loaded into the memory at runtime. This represents approximately 3.25-times the size of the raw corpus.

Some ideas for the implementation were also taken from [5].

For a given wildcard query, the database is currently capable of ***returning around 10^5 hits per second*** *with constant memory usage. The database runs on one of our server machines with a constant memory usage of around 280 Gb.*

3.2 Factoid Extraction

The N-Gram database is compared against Yahoo Finance names and most common phrases are extracted. Current prototype uses phrases that are built from a “company name” and a verb. The WordNet [6] (large lexical database of English) has been used for identification of types of words in the phrases.

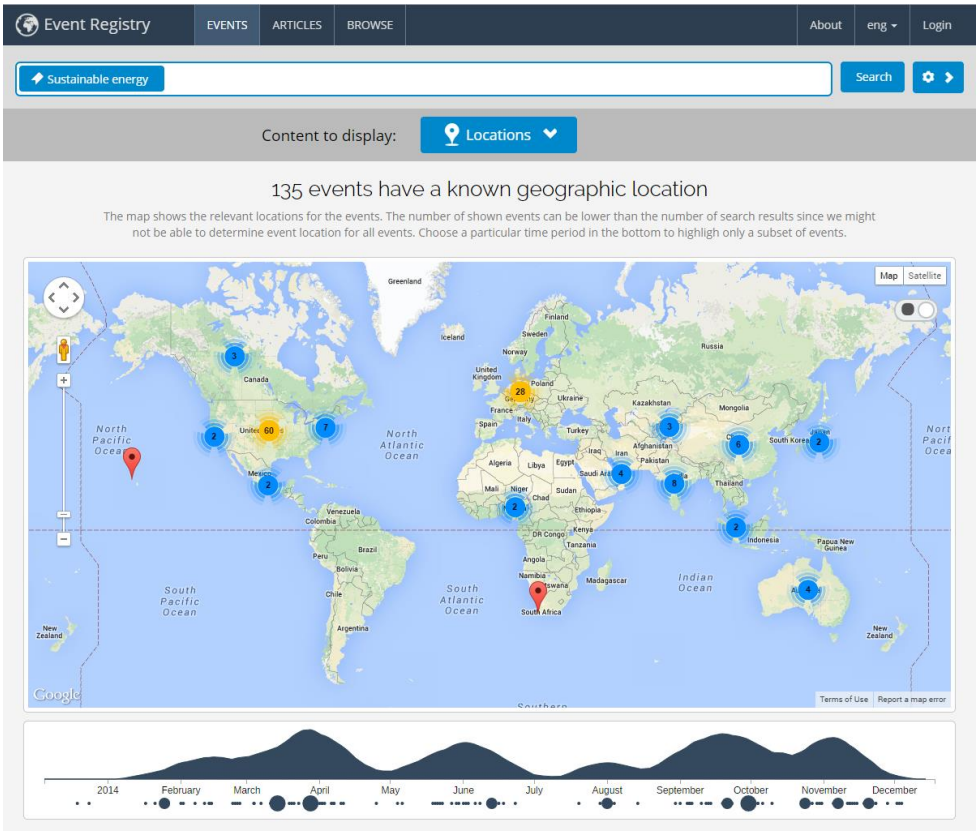

Figure 4: Screenshot from the EventRegister frontend.

The prototype can detect any templates of this form in the N-Gram Database.

<company>"basic materials\oil & gas drilling & exploration\seadrill limited"</company><verb>"associated"</verb>

<company>"basic materials\independent oil & gas\conocophillips"</company><verb>"claims"</verb>

<company>"financial\investment brokerage - national\morgan stanley"</company><verb>"predicts"</verb>

<company>"basic materials\oil & gas drilling & exploration\seadrill limited"</company><verb>"reports"</verb>

<company>"basic materials\independent oil & gas\exxon mobil corp"</company><verb>"fell"</verb>

<company>"basic materials\major integrated oil & gas\chevron corp"</company><verb>"was"</verb>

<company>"basic materials\major integrated oil & gas\petrobras"</company><verb>"shares"</verb>

<company>"industrial goods\heavy construction\ideal"</company><verb>"buying"</verb>

<company>"basic materials\major integrated oil & gas\petrobras"</company><verb>"has"</verb>

<company>"basic materials\major integrated oil & gas\petrobras"</company><verb>"were"</verb>

<company>"basic materials\major integrated oil & gas\petrobras"</company><verb>"board"</verb>

<company>"basic materials\major integrated oil & gas\petrobras"</company><verb>"could"</verb>

<company>"basic materials\independent oil & gas\conocophillips"</company><verb>"has"</verb>

<company>"basic materials\major integrated oil & gas\petrobras"</company><verb>"was"</verb>

<company>"basic materials\major integrated oil & gas\petrobras"</company><verb>"reveals"</verb>

<company>"basic materials\oil & gas drilling & exploration\seadrill limited"</company><verb>"reports"</verb>

Figure 5: Example from the list of extracted templates.

Templates are then stored and pushed to the factoid extraction module. The module is connected to the EventRegistry (see Figure 4) streaming API and templates are compared against the aggregated streaming news data. The result is composed of a phrase and the part of the text where the phrase was detected. More information on the output is available in the Section 4.

4 Implementation Details

The application is implemented as a simple console application, that takes one parameter as input and this parameter is the sector of the companies in the Yahoo Financials register (e. g. basic materials\independent oil & gas). Results are then composed of the hits of the extracted templates from those companies.

4.1 Input API

EventRegistry is connected via a Python API[6]. The Python script is executed within the C++ application with a Python library which executes the interpreter. New articles are collected every 10 seconds.

4.2 Output API

Results of a query are returned via a file interface in the form, presented in Figure 6. It provides a simple XML structure.

```
<article 4>
        <match1>
                <company>"basic materials\oil & gas drilling & exploration\seadrill
                limited"</company><verb>"associated"</verb>
                <context>"seadrill limited associated companies after the third-quarter 2014 results: 1 - seadrill partners, llc
                (nyse:sdlp): sdrl owns 39,635,400 shares or 53.2% (majority holder.) 2 - north atlantic drilling (nyse:nadl): sdrl
                owns 169,663,723 shares or 70.36% (majority holder.) this situation may change in q2 2015 with a potential
                new partnership with rosneft (otc:rnftf) put on hold by the sanctions of the eu and the usa against russia.
                "</context>
        </match1>
</article 4>
<article 28>
        <match1>
                <company>"basic materials\independent oil & gas\conocophillips"</company><verb>"claims"</verb>
                <context>"in papers filed in harris county court in texas, conocophillips claims that "pdvsa seeks to liquidate
                and repatriate assets out of the current jurisdictions and to venezuela or elsewhere to hinder collection on the
                claims against it because venezuela will not recognize the arbitration awards or judgments recognizing or
                confirming them." e-mail start a free trial bankers representing venezuela have set a date for late december for
                prospective buyers to submit revised offers for citgo, according to a reuters report. "</context>
        </match1>
</article 28>
<article 37>
        <match1>
                <company>"financial\investment brokerage - national\morgan stanley"</company><verb>"predicts"</verb>
                <context>"morgan stanley predicts apple will sell 30 million units. "</context>
        </match1>
</article 37>
<article 57>
        <match1>
                <company>"basic materials\copper\freeport-mcmoran"</company><verb>"copper"</verb>
                <context>"sector % total country % total analysis assets analysis assets integrated oil 34.2 global 33.1
                diversified 18.7 usa 20.3 exploration & production 15.1 canada 18.8 copper 7.5 europe 11.1 oil sands 4.7 latin
                america 6.5 nickel 4.6 asia 4.4 distribution 3.6 china 2.2 oil services 3.5 africa 2.0 gold 2.4 australia 2.0 coal 2.2
                current liabilities (0.4) iron ore 1.4 ----- silver 1.3 100.0 diamonds 0.5 ===== platinum 0.4 fertilizers 0.3 current
                liabilities (0.4) ----- 100.0 ===== ten largest equity investments % total company region of risk assets chevron
                global 6.9 exxonmobil global 6.0 bhp billiton global 5.5 rio tinto global 3.8 royal dutch shell global 3.6 eni europe
                3.6 enbridge income canada 3.6 freeport-mcmoran copper & gold asia 3.5 conocophillips usa 3.3 glencore global
                3.2 commenting on the markets, olivia markham and tom holl, representing the investment manager noted:
                both the energy and mining sectors came under significant downward pressure during the month as commodity
                prices, particularly on the energy side, fell sharply. "</context>
```

[6] https://github.com/gregorleban/event-registry-python

```
        </match1>
</article 57>
<article 68>
        <match1>
                <company>"basic materials\oil & gas drilling & exploration\seadrill limited"</company><verb>"reports"</verb>
                <context>"the seadrill group* on a consolidated basis reports ebitda of us$842 million, a year over year
                increase of 27% seadrill limited reports third quarter 2014 ebitda* of us$635 million seadrill limited reports
                third quarter 2014 net income of us$190 million and earnings per share of us$0.31 the seadrill group on a
                consolidated basis maintains orderbacklog of approximately us$20 billion seadrill limited is suspending dividend
                distributions and focusing on debt reduction and value creating opportunities due to significant deterioration in
                the broader markets seadrill receives a commitment for a us$1.35 billion credit facility to refinance the credit
                facilities secured by the west pegasus, west gemini, and west orion. "</context>
                </match1>
</article 68>
<article 131>
        <match1>
                <company>"basic materials\independent oil & gas\exxon mobil corp"</company><verb>"fell"</verb>
                <context>"exxon mobil corp fell 1.2 percent to $94.61 while chevron corp was off 1.5 percent to $115.87.
                "</context>
        </match1>
        <match2>
                <company>"basic materials\major integrated oil & gas\chevron corp"</company><verb>"was"</verb>
                <context>"exxon mobil corp fell 1.2 percent to $94.61 while chevron corp was off 1.5 percent to $115.87.
                "</context>
        </match2>
</article 131>
```

Figure 6: Example of the output of Factoid Extractor component.

5 Conclusions and Future Work

With the prototype we have added another building block to the overall textual infrastructure for building a domain-specific knowledge base from real-time streaming news. We have improved extraction of factoid templates by approximately 10^7, so domain specific templates can now be built in a matter of milliseconds, while previously the time needed for this was a couple of hours. These patterns are used on the stream of semantically enriched news data from the EventRegistry and results are available in real time in a structured XML output.

The goal for the 3rd year of the project is to align the extracted factoids with the Cyc KB and to create an environment for domain-specific knowledge base building. The following tools for exploitation of KB will be added: Q/A client and API. The possibility of using the aggregated factoid extraction data for improvement of modelling of energy prices will also be tested.

References

[1] Novalja I. et al., NRG4CAST Deliverable D3.2 – Semantic Enrichment Prototype, November 2014.

[2] Yahoo Finance Taxonomy, available on: http://biz.yahoo.com/p/s_conameu.html

[3] Web 1T 5-gram Version 1, description available on: https://catalog.ldc.upenn.edu/LDC2006T13

[4] Jon L. Bentley, R. Sedgewick, Fast Algorithms for Sorting and Searching Strings, SODA '97 Proceedings of the eight annual ACM-SIAM symposium on Discrete Algorithms, 360-369, 1997.

[5] M. Flor, A Fast and Flexible Architecture for Very Large Word N-gram Datasets, Natural Language Engineering, Volume 19, Issue 01, January 2013, pp 61-93. Cambridge University Press, 2012.

[6] http://wordnet.princeton.edu/ (accessed: 24. 11. 2014)

