Deliverable D8.1: Test data and Scenarios

Deliverable D8.1
Version FINAL

Authors: Pim Stouten(1), Rutger Kortleven(1), Ian Hopkinson(2)
Affiliation: (1) LexisNexis (2) ScraperWiki Ltd

BUILDING STRUCTURED EVENT INDEXES OF LARGE VOLUMES OF FINANCIAL AND ECONOMIC DATA FOR DECISION MAKING

ICT 316404

Grant Agreement No.	316404
Project Acronym	NEWSREADER
Project full title	Building structured event indexes of large volumes of financial and economic data for decision making.
Funding Scheme	FP7-ICT-2011-8
Project Website	http://www.newsreader-project.eu
Project Coordinator	Prof. dr. Piek T. J. M. Vossen VU University Amsterdam Tel. +31 (0) 20 5986466 Fax. +31 (0) 20 5986500 Email: piek.vossen@vu.nl
Document Number	Deliverable D8.1
Status & Version	FINAL
Contractual Date of Delivery	September 2013
Actual Date of Delivery	October 2013
Type	Report
Security (distribution level)	Public
Number of Pages	21
WP Contributing to the deliverable	WP08
WP Responsible	LN
EC Project Officer	Sophie Reig/Susan Fraser
Authors:, Pim Stouten[1], Rutger Kortleven[1], Ian Hopkinson[2]	
Keywords: large volumes of data, sources, test data	

Executive Summary/Abstract

The volume of news data is enormous and expanding. Professional decision-makers who need to respond quickly to new developments and knowledge or who need to explain these developments on the basis of the past are faced with the problem that current solutions for consulting these archives and news streams no longer work. It becomes almost impossible to make well-informed decisions and professionals risk to be held liable for decisions based on incomplete, inaccurate and out-of-date information.

This deliverable (D8.1) describes the project consortium's efforts in the areas of Test Data and Test Scenarios. These two elements are closely intertwined: no working test scenario without proper test data, and no relevant set of test data without a test scenario defining it. Test scenarios are used to evaluate the performance of NewsReader versus tools already in use.

This document outlines selection criteria for the different Scenarios in section 2, describes four unique use cases in section 3 and uses sections 4 and 5 to shed more light on data criteria (4) and specific data for the various uses cases (5). Section 6 covers a closely related area: that of evaluation data.

Where test data covers the full scope of the NewsReader project (data size, document languages), evaluation data covers a slightly different area: controlled data sets made available to the scientific community to evaluate and build upon the knowledge and experience from the NewsReader project.

Table of Revisions

Version	Date	Description and reason	By	Affected Section
0.1	25 Sep 2013	first draft	Pim Stouten, Rutger Kortleven, Ian Hopkinson	all
0.2	05 Oct 2013	review & comments	Marieke van Erp	all
0.3	07 Oct 2013	revised version	Pim Stouten, Rutger Kortleven, Ian Hopkinson	All
0.4	22 Oct	Comments	Ian Hopkinson	All
0.4	30 Oct 2013	Updated with latest status	Rutger Kortleven	5

Contents

List of Tables

1 Introduction

Validation and evaluation are key elements in the NewsReader project, and this deliverable describes relevant test scenarios and test data.

Test data will be used both for machine and human annotation, to allow for benchmarking and comparison. The consortium has been looking for use cases (and the test data that go with them) covering a wide range of topics and potential applications of NewsReader tooling.

The various cases are in different stages of maturity: some test data has already gone through analysis, whereas other use cases have focused on understanding user workflow first.

2 General Selection Criteria for Scenarios

The various Test Scenarios have multiple common denominators, each on its own is crucial to the scope of the NewsReader project:

- source material is a mixture of unstructured and structured data, where structured is mainly used as an add-on or as a reference
- scenario has one or more financial/economical aspects
- scenario is relevant to professional decision-making
- sufficient source data in the project’s languages: English, Italian, Spanish, Dutch

All of the Scenarios (described in the next section) meet all of these criteria. The description of the four individual scenarios will further touch upon their match with the four selection criteria mentioned above.

3 Scenarios

3.1 TechCrunch/Crunchbase

Crunchbase.com[1] is a free database in the form of a wiki, relating to technology companies which anyone can edit. It provides data on companies, people, financial organisations, service providers, funding rounds and acquisitions. The data is created via community editing, and is available via a web site and an API. Crunchbase holds information on over 200.000 persons and 180.000 companies[2].
TechCrunch[3] is a technology news web site related to the Crunchbase database, and contains approximately 45,000 news articles. Data creation is different from CrunchBase, with its wiki-approach:

[1] http://www.crunchbase.com
[2] status as of October 7th, 2013
[3] http://www.techcrunch.com

TechCrunch news documents are written by TechCrunch staff and freelance ICT bloggers and journalists. Founded in June 2005, TechCrunch and its network of websites now reach over 12 million unique visitors and draw more than 37 million page views per month. The TechCrunch community includes more than 2 million friends and followers on Twitter, Facebook, LinkedIn, Google and other social media[4].

TechCrunch is a strongly domain-specific source (opposed to generic newswires/newspapers), and one of the most popular online technology sources for information in the ICT domain..

This scenario anticipates that events in the structured Crunchbase database are reflected in the non-structured TechCrunch articles. Therefore the TechCrunch/Crunchbase system represents an opportunity to test the Newsreader system by deriving events from the news articles in TechCrunch and comparing those events to those found in the Crunchbase database.

The Crunchbase database records the founding date for companies, and for both companies and people so-called Milestones are recorded. These can be thought of as “events”. For acquisitions it records the date, target, acquirer and price. The funding round date, name of the company funded, round (there are various categories of funding) and size are recorded along with the investors. Sometimes the investor information is not available, so there is scope to try and identify investor information from other sources.

Both Company and People records in Crunchbase contain freeform text descriptions alongside more structured data. Therefore, as well as making comparisons between TechCrunch and Crunchbase, Newsreader technology could be applied to the freeform text in Crunchbase to compare against structured elements.

The value of this data is for both investors and technology startups, it provides supporting information for doing business and acquiring funding.

3.2 Dutch House of Representatives

This use case focuses at the specific information needs playing in the ecosystem of parliamentary democracy. The major challenges in this particular ecosystem are known in Big Data jargon as:

- *veracity*: how ‘truthful’ is the information supplied? Does it support or contradict other sources? We will explain this in further detail after the ‘Parliamentary Inquiries’ header.
- *velocity*: with an increasing media coverage of the parliamentary process, reaction speed becomes more and more important for the MPs and their support staff. This requires background information to be available almost in real time.

[4] source: http://techcrunch.com/about/

The Information Department at the Dutch House of Representatives supports all information needs of the Parliamentary organization. Its 60 specialists are tasked with providing the various departments, commissions and parliamentary fractions with information, with roles ranging from information research/information specialist to database manager to systems engineers.

The information supplied by the Department is used as input for government reports, draft bills, parliamentary enquiries, but also for current awareness of MPs and their staff. The information department is complementary to the bureaus of political parties and takes a neutral stance, for example a left wing party has a different view on particular issues than a right wing party.

The Department uses various information providers (Infolook newsletters[5], Nexis.com[6], LexisNexis Publisher[7], Howard's Home[8]) to supply its stakeholders with news information. It also owns and maintains a database with parliamentary documents, and has access to various public data sources.
It uses the Autonomy[9] suite of tools as a document repository, indexing/classification engine and search engine. This delivers Google-style retrieval results, which still require filtering and post processing. The information department is also investigating document classification using GridLine products[10].
The department provides different types of information requests: sometimes they need to provide an informed answer really quickly, sometimes they work on big reports for longer periods of time. The position of the Netherlands and of Dutch politics have changed the past few years, resulting in more debates on current affairs, requiring quick access to information. Thus, there is an increasing pressure on the information department to provide the correct information quickly. Besides, the department sees growing pressure to reduce costs, whilst the amount of available information grows.

Parliamentary Inquiries

The Parliamentary Inquiry is the strongest instrument available to the Dutch House of Representatives to reconstruct a major issue or event, with large impact on the economy, or society as whole. A Parliamentary majority decides to install a commission consisting of several members of Parliament, and also sets the scope of Inquiry.
The commission can call on witnesses to testify under oath. Key sources of information (reports, biographies, analysis, news document) are gathered, assessed and provided by the information department.
We have identified several challenges where NewsReader can play a fundamental role:

- coverage: 'understanding' an event and its key actors and entities helps in quickly retrieving all relevant documents, lowering the risk of missing relevant documents that - at first sight- have no direct value for the Inquiry

[5] http://www.infolook.nl/
[6] http://www.lexisnexis.nl/dutch/products/nexis.page
[7] http://www.lexisnexis.nl/dutch/products/publisher.page
[8] http://w3.howardshome.com/
[9] http://www.autonomy.com/products/
[10] http://www.gridline.nl/gridwalker-thesaurus

- mapping the gaps: identifying which areas have insufficient or no information coverage
- creating networks of events, people and entities (companies, government bodies, …)
- fact checking: fact extraction is one of the fundamentals behind NewsReader; quickly extracting facts and analysing their source of origin would be a great benefit for the Parliament.

NewsReader members (VUA, LN) have performed interviews and a *workflow mapping* workshop, with 11 information specialists to research the ways they gather information for the parliamentary inquiry commissions. We focused on how they retrieve relevant information, what methods and search techniques they use, and how they validate, summarise and distribute information.

NewsReader will use the insights from the aforementioned meetings to draw several proposals for further collaboration with the Parliament's information department, ranging from data supply to the consortium to evaluation of (early) prototypes.

3.3 Global Automotive Industry

NewsReader aims to model the car manufacturing domain in terms of positive and negative news stories about car manufacturers, stock market data, mergers & takeovers, and other key events. The NWR team has chosen this domain to work in as it is truly international, touches upon many socioeconomic issues and is well-represented in the data available from LexisNexis.

The first proof of principle built on the automotive data set (see 5.3) has focused on the changing ownership structures in the VAG automotive consortium over time, with major events such as planned acquisitions and plant closures analysed by early versions of NewsReader.

In this Scenario, NewsReader will help (re)construct:

- complex structures, ie the ownership structure of VAG, or any other automotive conglomerate
- complex events, ie mergers, acquisitions and corporate restructuring in this industry

3.4 Business Intelligence

The Business Intelligence use case relates to gathering information about companies for the purposes of evaluating 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[11].

Bankers Accuity[12] is a business to business publishing company which specialises in this area of business intelligence. ABN Amro[13] carry out similar investigations in their role as a major bank.
Compiling this data currently requires a great deal of manual work. The aim of this use case is to provide initial data to support that manual work by analysing the text content of company websites, including company reports.

We are particularly interested in building the story of a company in terms of key events, and actors in those events. For example, the names of the officers of a bank and their titles, when they arrived in the company, what they did prior to joining the company and whether they have left. Also of interest would be takeover activity and expansion plans.

The focus of Bankers Accuity is the financial sector, in support of this work LexisNexis has provided a list of 1.000 banks from the UK, Netherlands, Italy and Spain, covering the four languages of the Newsreader project.

ScraperWiki will scrape these bank websites for content which will be fed into the Newsreader system. This is a novel application of the Newsreader technology and we will need to establish how to automatically extract the most useful content.

4 General Criteria for Data

This section is a reworked version of Deliverable D1.1: **Definition of Data Sources.**

4.1 Available resources

LexisNexis databases consist of news articles, market reports, company information such as Chamber of Commerce extracts, country reports, market information, public records, legal information and legislation. Most data in the LexisNexis database are owned by publishers and therefore covered by copyright.

ScraperWiki focuses on open sources, in other words: data not covered by copyright as in the LexisNexis case. Open sources are published on the internet. These open data are crawled and retrieved by ScraperWiki. Open data are mostly

[11] Jargon, widely used in the financial world, describing the vetting process before onboarding a customer, or doing business with a supplier, partner or other business relation.
[12] http://www.accuity.com/
[13] http://www.abnamro.com/en/about-abn-amro/index.html

found in the public domain and are provided by for example, governments, municipalities, international intergovernmental organisations and so forth.

4.2 Document volumes treated by LexisNexis

LexisNexis handles on average an estimated 1,5,000,000 news documents and 400,000 web pages per weekday. The archive of LexisNexis contains over 25 billion documents spanning several decades.
The LexisNexis database holds approximately 40,000 sources. It includes among others 30,000 different newspapers (with 35,000 issues each day), 85 million company reports, over 60 million manager biographies, several hundreds of thousands market reports.

4.3 ScraperWiki Methodology

ScraperWiki has experience in scraping a wide range of text sources including social media, discussion forums, a wide variety of documents from government sources, parliamentary sources such as Hansard, and the verbatim records of the UN. Applying a logical structure is done on a case by case basis for each document source and depends on there being sufficient formatting information to extract a logical structure. In addition to text sources ScraperWiki also scrapes a wide range of numerical data.

4.4 Distinction between licensed and open data resources

Data sources can be divided in open data sources (for example from the Web) and licensed content. Licensed content are all those data sources for which a licensing agreement is drawn up between a publisher and LexisNexis.

Scraperwiki and LexisNexis crawl the Internet for Web sources, with Internet bots[14] that systematically browse the internet, typically for the purpose of web indexing[15]. Web search engines and some other sites use Web crawling or spidering software to update their web content or indexes of others sites' web content. Specific full text documents for relevant NewsReader web sources are scraped by ScraperWiki.
Open data are freely available to the public. These data can come from any source and are not restricted by copyright. ScraperWiki will crawl and retrieve Open Data. A typical Open Data source is government data, made available through websites to distribute the data they generate and collect.

The NewsReader solution for non-public sources is as follows:

1. All accessible data is downloaded and processed by the NewsReader modules as an internal process.
2. The knowledge extracted is stored in the NewsReader Knowledge Store
3. Anybody can access the knowledge in the Knowledge Store, which is a compact and generalised representation of the content of millions of sources.
4. Every piece of knowledge gives access to the original sources from which it is derived through a URL.

[14] http://en.wikipedia.org/wiki/Internet_bot
[15] http://en.wikipedia.org/wiki/Web_indexing

5. Public content can be retrieved directly from the Knowledge Store through the URL.
6. Non-public content can be accessed outside the system through the URL only, thus limiting access to those possessing the right credentials to access this data

This solution allows us to process the LexisNexis data as well as any other data and represent the cumulated knowledge to the users, while the access to the original sources is left to the providers, e.g. LexisNexis. In this way, LexisNexis or any other provider is free to exploit the traffic coming from the Knowledge Store to their archives and databases.

4.5 Criteria for selection of sources to use in NWR

All licensed content sources and open web sources for the NewsReader project will be gathered based on the following criteria:

1. Data sources in Italian, Spanish, Dutch or English
2. A sufficient number of named entities needs to occur
3. Timespan 10 years 2003-2013 (economic crisis)
4. Licensed content sources meaning those sources for LexisNexis and publisher have an agreement on the usage of such sources.
5. Open web sources (those sources scraped by ScraperWiki).
6. All data sources must be news, events, activities or opinions related to economic or finance issues.
7. Regarding quality and authority, we are looking for definitive sources for example speeches by a finance minister or central banker from quality news sources rather than anonymous individuals on discussion forums.
8. Opinions on blogs and social media are very relevant to NewsReader. As a separate track we therefore seek out opinion-based pieces. These might include blogs, opinion columns from conventional news sources or discussion forums. It may be necessary to create a classifier which would automatically tag pieces as opinion.
9. Quantity, furthermore we will be scraping those data sources which contain large amounts of content. Thus we gain more content from the same effort if we focus on larger sources. This is a practical criterion for the web scraping process rather than a NewsReader criterion.
10. For evaluation of the technical/scientific modules, a small subset of the data needs to be completely free for distribution. These sources will be annotated manually to calibrate and train the software.

4.6 Evaluation data

Besides the project data used internally to feed the NWR system with information, evaluation data will be made available in the NewsReader project. These are to be made available to the research community for verification purposes of research results from the NewsReader project and for end user evaluation.
There are 3 kinds of evaluation data sets:

1. Technical evaluation set that will be used by other researchers and research groups. This dataset must be freely publicly available.
2. User scenario free set. Open data processed by NWR and used in the end user evaluation.

3. User scenario closed set. Could be all of LexisNexis data to compare with LexisNexis system.

When it comes to the closed set evaluation data, the usage of these sources is somewhat restricted. Since the licensed sources are to be made available to parties outside the NewsReader project special permission needs to be asked to the publishers concerned.

LexisNexis licensing managers (see section 6) are currently negotiating with a group of targeted publishers willing to provide parts of their data sources to the NewsReader project. This might narrow the amount of sources available for the evaluation data. One of the conditions of providing data sources for evaluation and republication within the research community will be manner in which the data are made available. The data sources will be made available in an intranet-environment where users will have to register with an ID and password.

Evaluation data will gathered based on the same criteria as mentioned under 4.4.2

4.7 Structured data

Data sources can be designated as structured or unstructured data for classification within an organization. The term structured data refers to data that is identifiable because it is organized in a structure. The most common form of structured data -- or structured data records (SDR) -- is a database where specific information is stored based on a methodology of columns and rows.
Structured data is also searchable by data type within content. Structured data is understood by computers and is also efficiently organized for human readers. In contrast, unstructured data, such as free text (most news documents have no solid structure in their body text) has no such identifiable structure.
For NewsReader we have identified the following structured data sources which will be made available within the NWR project:

1. Stock market data restricted: http://www.lexisnexis.com/academic/companydossier/ open: http://eoddata.com/; http://finance.yahoo.com/; http://www.quote.com/
2. LexisNexis Company Dossier: reports covering over 43 million public, private and international companies http://www.lexisnexis.com/academic/companydossier/
3. WorldBank http://www.worldbank.org/
4. Organisation for Economic Co-operation and Development (OECD) http://www.oecd.org/
5. International Monetary Fund (IMF) http://www.imf.org/external/data.htm
6. US Securities and Exchange Commission (SEC) Company Filings http://www.sec.gov/search/search.htm
7. The DBpedia Knowledge Base: http://wiki.dbpedia.org/About. DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web.

5 Data Sources for Scenarios

5.1 TechCrunch/Crunchbase

Output Format:	**SQLite database**
Language:	**English**
Timespan:	**2008 - 2013**
Topic:	**Tech companies**
Result set:	**43,000 articles**
Sources:	**http://techcrunch.com**

table 1: SQLite database Techcrunch

Output Format:	**SQLite database**
Language:	**English**
Timespan:	**2008 - 2013**
Topic:	**Tech companies**
Result set:	**180,000 companies, 200,000 biographies**
Sources:	**http://www.crunchbase.com/**

table 2: SQLite database Crunchbase

5.2 Dutch House of Representatives

Format:	**XML with metadata**
Language:	**Dutch, English**
Timespan:	**10 years: 2003-2013 (economic crisis)**
Topic:	**Parliamentary inquiries and voting behavior by members of Parliament**
Size:	**To be established**
Sources:	**Dutch parliament database, www.tweedekamer.nl www.overheid.nl LexisNexis licensed sources**

table 3: XML-files on Parliamentary inquiries

5.3 Global Automotive Industry

Output Format:	**XML with Metadata**
Language:	**English**
Timespan:	**2003-2013**
Topic:	**Ownership of Car brands (Mergers, acquisitions, corporate restructuring)**
Result set:	**6,171,190 documents**
Sources:	**5,974 English-language news sources licensed by LexisNexis**

Table 4: English language XML files on ownership of car brands

LexisNexis used three complex queries to circumvent data volume limitations to retrieve relevant documents:

(Alfa Romeo or Aston Martin or Audi or BMW or Bentley or Bugatti or Buick or Cadillac or Chevrolet or Chrysler or Daewoo or Daihatsu or Daimler or De Lorean or De Tomaso or Dodge or Ferrari or Fiat or Ford or GMC or Honda or Hummer or Hyundai or Isuzu or Iveco or Jeep or Kia or Koenigsegg or Lada or Lamborghini) and (ownership! or joint venture or merg! or aquisit! or CEO or manager! or managing or director! or (take w/3 over) or takeover or subsidiar! or headquarter or child entity or consolidat! or acquir!)

(Lancia or Land Rover or landrover or Lexus or Lincoln or Lotus or Maserati or Maybach or Mazda or McLaren or Mercedes-Benz or Mercury or Mitsubishi or Nissan or Oldsmobile or Opel or PAZ or Peugeot or Pontiac or Porsche or Proton or Renault or Rolls-Royce or SEAT or Saab or Scania or SsangYong or Steyr or Subaru or Suzuki or TVR or Texmaco or Toyota or Vauxhall or Volkswagen or Volvo or Yamaha or Yugo or Zephyr or Skoda) and (ownership! or joint venture or merg! or aquisit! or CEO or manager! or managing or director! or (take w/3 over) or takeover or subsidiar! or headquarter or child entity or consolidat! or acquir!)

(Citroen or DAF or Jaguar or MG or Mercedes or Morgan or Morris or Plymouth or Puch or Rover or Saturn or Smart) and (car or auto or automobile or automotive) and (ownership! or joint venture or merg! or aquisit! or CEO or manager! or managing or director! or (take w/3 over) or takeover or subsidiar! or headquarter or child entity or consolidat! or acquir!)

Output Format:	**XML with Metadata**
Language:	**Spanish, Italian, Dutch**
Timespan:	**2005-2010**
Topic:	**Ownership of Car brands (Mergers, acquisitions, corporate restructuring)**
Result set:	**2,158 Italian language documents** **2,474 Spanish language documents** **1,116 Dutch language documents**

Table 5: Italian, Spanish and Dutch language XML-files on ownership of car brands

LexisNexis used three complex queries to circumvent data volume limitations to retrieve relevant documents in Italian, Spanish and Dutch language:

Italian:

(Alfa Romeo or Aston Martin or Audi or BMW or Bentley or Bugatti or Buick or Cadillac or Chevrolet or Chrysler or Daewoo or Daihatsu or Daimler or De Lorean or De Tomaso or Dodge or Ferrari or Fiat or Ford or GMC or Honda or Hummer or Hyundai or Isuzu or Iveco or Jeep or Kia or Koenigsegg or Lada or Lamborghini) and (empresa! Or casa Or fabricante!) w/3 (automovil! Or coche! Or automocion)) AND (propriedad or joint venture or empresa! Conjunta! Or alianza! Estratégica! Or alianza! Commercial! Or fusion or consolidacion or consorcio Or adquisicion! Or adquir! Or director! Ejecutiv! Or CEO or president! Ejecutiv! Or consejer! Delegad! Or jefe de operaciones or director de operaciones or director financier or director de finanzas or oferta de compra or oferta publica de adquisicion or sede or filial!)*

(Lancia or Land Rover or landrover or Lexus or Lincoln or Lotus or Maserati or Maybach or Mazda or McLaren or Mercedes-Benz or Mercury or Mitsubishi or Nissan or Oldsmobile or Opel or PAZ or Peugeot or Pontiac or Porsche or Proton or Renault or Rolls-Royce or SEAT or Saab or Scania or SsangYong or Steyr or Subaru or Suzuki or TVR or Texmaco or Toyota or Vauxhall or Volkswagen or Volvo or Yamaha or Yugo or Zephyr or Skoda) and (empresa! Or casa Or fabricante!) w/3 (automovil! Or coche! Or automocion)) AND (propriedad or joint venture or empresa! Conjunta! Or alianza! Estratégica! Or alianza! Commercial! Or fusion or consolidacion or consorcio Or adquisicion! Or adquir! Or director! Ejecutiv! Or CEO or president! Ejecutiv! Or consejer! Delegad! Or jefe de operaciones or director de operaciones or director financier or director de finanzas or oferta de compra or oferta publica de adquisicion or sede or filial!)*

(Citroen or DAF or Jaguar or MG or Mercedes or Morgan or Morris or Plymouth or Puch or Rover or Saturn or Smart) AND (empresa! Or casa Or fabricante!) w/3 (automovil! Or coche! Or automocion)) AND (propriedad or joint venture or empresa! Conjunta! Or alianza! Estratégica! Or alianza! Commercial! Or fusion or consolidacion or consorcio Or adquisicion! Or adquir! Or director! Ejecutiv! Or CEO or president! Ejecutiv! Or consejer! Delegad! Or jefe de operaciones or director*

de operaciones or director financier or director de finanzas or oferta de compra or oferta publica de adquisicion or sede or filial!)

Spanish:

(Alfa Romeo or Aston Martin or Audi or BMW or Bentley or Bugatti or Buick or Cadillac or Chevrolet or Chrysler or Daewoo or Daihatsu or Daimler or De Lorean or De Tomaso or Dodge or Ferrari or Fiat or Ford or GMC or Honda or Hummer or Hyundai or Isuzu or Iveco or Jeep or Kia or Koenigsegg or Lada or Lamborghini) and (societa or aziend! or impres! or casa or case) w/3 automobilistic!) AND (proprieta or joint venture or merger or partnership or fusion! Or raggrupa! Or consorz! Or acquisizion! Or CEO or amministratore delegato or direttore generale or takeover or (ofert! w/acquisto) or take over or direttore finanziario or direttore operativo or sede or quartier generale or filial! Or consolida!)

(Lancia or Land Rover or landrover or Lexus or Lincoln or Lotus or Maserati or Maybach or Mazda or McLaren or Mercedes-Benz or Mercury or Mitsubishi or Nissan or Oldsmobile or Opel or PAZ or Peugeot or Pontiac or Porsche or Proton or Renault or Rolls-Royce or SEAT or Saab or Scania or SsangYong or Steyr or Subaru or Suzuki or TVR or Texmaco or Toyota or Vauxhall or Volkswagen or Volvo or Yamaha or Yugo or Zephyr or Skoda) and (societa or aziend! or impres! or casa or case) w/3 automobilistic!) AND (proprieta or joint venture or merger or partnership or fusion! Or raggrupa! Or consorz! Or acquisizion! Or CEO or amministratore delegato or direttore generale or takeover or (ofert! w/acquisto) or take over or direttore finanziario or direttore operativo or sede or quartier generale or filial! Or consolida!)

(Citroen or DAF or Jaguar or MG or Mercedes or Morgan or Morris or Plymouth or Puch or Rover or Saturn or Smart) and (societa or aziend! or impres! or casa or case) w/3 automobilistic!) AND (proprieta or joint venture or merger or partnership or fusion! Or raggrupa! Or consorz! Or acquisizion! Or CEO or amministratore delegato or direttore generale or takeover or (ofert! w/acquisto) or take over or direttore finanziario or direttore operativo or sede or quartier generale or filial! Or consolida!)

Dutch:

(Alfa Romeo or Aston Martin or Audi or BMW or Bentley or Bugatti or Buick or Cadillac or Chevrolet or Chrysler or Daewoo or Daihatsu or Daimler or De Lorean or De Tomaso or Dodge or Ferrari or Fiat or Ford or GMC or Honda or Hummer or Hyundai or Isuzu or Iveco or Jeep or Kia or Koenigsegg or Lada or Lamborghini) and (autoproducent OR automaker or autofabrikant) AND (fusie! OR overname! OR aankoop! OR acqui! OR joint venture OR CEO OR CFO OR COO OR (managing w/2 director) OR (general w/2 manager) OR consolid! OR hoofdkwartier! OR dochtermaatschappij! OR nevenvestiging!)

(Lancia or Land Rover or landrover or Lexus or Lincoln or Lotus or Maserati or Maybach or Mazda or McLaren or Mercedes-Benz or Mercury or Mitsubishi or Nissan or Oldsmobile or Opel or PAZ or Peugeot or Pontiac or Porsche or Proton or Renault or Rolls-Royce or SEAT or Saab or Scania or SsangYong or Steyr or Subaru

or Suzuki or TVR or Texmaco or Toyota or Vauxhall or Volkswagen or Volvo or Yamaha or Yugo or Zephyr or Skoda) and (autoproducent OR automaker or autofabrikant) AND (fusie! OR overname! OR aankoop! OR acqui! OR joint venture OR CEO OR CFO OR COO OR (managing w/2 director) OR (general w/2 manager) OR consolid! OR hoofdkwartier! OR dochtermaatschappij! OR nevenvestiging!)

(Citroen or DAF or Jaguar or MG or Mercedes or Morgan or Morris or Plymouth or Puch or Rover or Saturn or Smart) and (autoproducent OR automaker or autofabrikant) AND (fusie! OR overname! OR aankoop! OR acqui! OR joint venture OR CEO OR CFO OR COO OR (managing w/2 director) OR (general w/2 manager) OR consolid! OR hoofdkwartier! OR dochtermaatschappij! OR nevenvestiging!)

5.4 Business Intelligence

Output Format:	**XML with Metadata (to be confirmed)**
Language:	**English, Italian, Spanish, Dutch**
Timespan:	**Current (documents are static: sites usually hold latest version of a document only)**
Topic:	**Banks**
Result set:	**1,000 documents**
Sources:	**1,000 bank websites based on a list provided by LexisNexis**

Table 6: XML of scraped content from Bank websites

6 Data Available for Evaluation

Also see section 4.6: we distinguish different types of evaluation data, based on their availability to the general public. The main reason for this distinction is copyrighted data: some data sources are copyrighted, some are not.

6.1 Publicly re-usable data

There are several open-source data repositories that are not copyrighted, and can be freely used for evaluation. Sections 4.7 and 5.1 give an insight in the sources the consortium uses in the open-source area.

6.2 Access to copyrighted data

LexisNexis' regular data licensing contracts do not allow for republication through any other means than LexisNexis products and services. This is a limitation to the impact of the project, since we want to make (see 4.6) relevant evaluation available **outside** the project consortium as well.

LexisNexis are working around this barrier by explicitly asking publisher for a written agreement, a de facto copyright waiver for a limited set of evaluation data. We will limit the scope of the data needed by imposing a specific time span covered, as well as limiting the data set to the specific topics covered by the use cases.

A status overview per language/region:

- Netherlands: de Persgroep (AD/Algemeen Dagblad, de Volkskrant, Trouw, het Parool) and NRC Media (NRC Handelsblad, NRC next) have agreed to make their data available. Negotiations with business newspaper Het Financieele Dagblad are still under way.

- Spain: Unidad Editorial (El Mundo) has agreed to make its data available, Grupo Zeta (El periódico de Catalunya; El periódico de Aragón; El periódico de Extremadura; El periódico Mediterráneo; Diario de Córdoba) has not responded yet.

- Italy: Monrif (Quotidiano Nazionale; Il Resto del Carlino; La Nazione and Il Giorno) does not want to supply its data; RCS Quotidiani (Corriere della Sera, Gazzetta dello Sport) just confirmed that it is willing to make data available.

- UK/English: most UK publishers have responded negatively to our request. This means we are looking for alternatives: English-language content originating from different countries/regions.We are waiting for a response from Emirates-based SyndiGate/Al Bawaba, a news aggregation organisation with dozens of English-language publications in its portfolio.

There are additional sources available beyond the ‘closed’ data sources mentioned above:

- corpora commonly used within the NLP community like the *Reuters RCV1*[16] corpus.
- open sources such as Wikinews[17] and Euronews[18].

7 Conclusions

Defining the right criteria to select use cases and their accompanying data sets was a relatively straightforward process, since we could build on knowledge gathered in earlier stages of the project.
Finding, describing and developing use cases that would encompass as many different aspects of the NewsReader deliverables drove strong interaction between the different

[16] http://about.reuters.com/researchandstandards/corpus/
[17] http://en.wikinews.org/wiki/Main_Page
[18] http://www.euronews.com/

consortium members, resulting in use cases like Business Intelligence and the Dutch Parliament.

The next steps in these use cases will focus on data gathering as well as workflow mapping: getting a better understanding of (potential) users of the solutions that NewsReader might offer.

The discussion around copyrighted evaluation data sets was an greater than foreseen barrier, that took a substantial amount of time to solve. The current situation, with several ‘copyrighted’ publishers agreeing to offer their content, combined with the wide range of open source data repositories at hand means that we are very close to coming to a satisfactory solution with regards to providing evaluation data.