Definition of Data Sources

Deliverable D1.1

Version 0.8

Authors: Rutger Kortleven[1], Ian Hopkinson[2], Marieke van Erp[3], Piek Vossen[3]

Affiliation: (1) LexisNexis, (2) ScraperWiki, (3) VU University Amsterdam

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

ICT 316404

DISCLAIMER

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 D1.1
Status & Version	Final
Contractual Date of Delivery	April 2013
Actual Date of Delivery	July 2013
Type	Report
Security (distribution level)	Public
Number of Pages	33
WP Contributing to the deliverable	WP01
WP Responsible	WP01
EC Project Officer	Sophie Reig
Authors: Rutger Kortleven, Ian Hopkinson, Marieke van Erp, Piek Vossen	
Keywords: data sources, content, XML, open sources, licensed content	

Executive Summary/Abstract

This document summarizes the various data types and data sources used in the NewsReader project. Available data sources will be delivered in the 4 NewsReader languages: English, Dutch, Italian and Spanish. The document gives an overview of all data sources available and relevant to the project and the criteria used to pick the relevant data sources. Furthermore, it provides insight in the way data sources are made available to NewsReader. The data sources will be standardized, tested and documented to be delivered in the project.

Table of Revisions

Version	Date	Description and reason	By	Affected Section
0.1	May 2013	First draft	Rutger Kortleven	
0.2	May 2013	Second draft with input from VUA Amsterdam	Marieke van Erp and Piek Vossen	1-6
0.3	June 2013	Added input from ScraperWiki	Rutger Kortleven and Ian Hopkinson	1-6
0.4	June 2013	Review	Marieke van Erp	
0.5	June 2013	Third draft with based on review VUA Amsterdam	Rutger Kortleven and Ian Hopkinson	2-6
0.6	July 2013	Final draft, removing comments and change markings	Piek Vossen	all
0.7	January 2014	Revision Check	Rutger Kortleven	all
0.8	January 2014	Revision Check	Marieke van Erp	all

Table of Contents

Contents

Executive summary....4

Table of revisions....5

List of Tables....7
Abbreviations....8

1 Introduction....8

2 Available resources....11
2.1 Insight into the amount of documents LexisNexis retrieves....12
2.2 ScraperWiki resources....12
3 Criteria for selection of sources to use in NWR....14
3.1 Distinction between licensed content resources & open data resources..14
3.2 Criteria for selection of sources to use in NWR: licensed content sources and open web source....15

4 Example scenarios and core NWR resources....17
4.1 Example Scenarios....18
4.2 Evaluation Data....20
4.3 Desription of main open sources....23
4.3.1 Wikinews sources....23
4.3.2 Parliamentary documentation....24
4.3.3 European Union Newsroom....24
4.3.4 Finance ministry releases....24
4.3.5 Press releases....25
4.3.6 Discourse database....25
4.3.7 European Central Bank....25
4.3.8 Federal Reserve....26
4.3.9 World Bank....26
4.3.10 World Bank blogs....26
4.3.11 International Monetary Fund (IMF)....26
4.3.12 Securities and Exchange Commision (SEC)....27
4.3.13 Company Annual Reports....27

5 Data formats....29
5.1 Technical description of data format provided to NWR....29

6 Conclusions....31
7 Appendix 1: technical example of data format....32

List of Tables

Table 1: Overview of use case scenarios……18
Table 2: Project data: Car brand ownership……. 18
Table 3: Project data: Economic Crisis……...18
Table 4: Project data: Crunchbase / TechCrunch…….. 19
Table 5: Project data: Dutch Parliament…….19
Table 6: Project data: Bankers Accuity…… 19
Table 7: Licensed content sources for evaluation data……. 21

Abbreviations

XML= Extensible Markup Language
NITF= News Industry Text Format
API=Application Programming Interface
RSS=Really Simple Syndication
NWR=NewsReader
CSV file = Comma Separated Values (database export/import format and file extension)

1. Introduction

The volume of news data is enormous and expanding. Professional decision-makers that need to respond quickly to new developments and knowledge or that 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.

NewsReader will process news in 4 different languages when it comes in. It will extract what happened to whom, when and where, removing duplication, complementing information, registering inconsistencies and keeping track of the original sources. Any new information is integrated with the past, distinguishing the new from the old and unfolding story lines in a similar way as people tend to remember the past and access knowledge and information. The difference being that NewsReader can provide access to all original sources and will not forget any details. We will develop a decision support tool that allows professional decision-makers to explore these story lines using visual interfaces and interactions to exploit their explanatory power and their systematic structural implications. Likewise, NewsReader can make predictions from the past on future events or explain new events and developments through the past. The tool will be tested by professional decision makers in the financial and economic area.

This deliverable provides an overview of available data sources in the NewsReader project. Besides a quantitative overview of the available data from both the in-house LexisNexis data as well as the publicly scraped data gathered by ScraperWiki, a qualitative survey is provided of the core data sources that will be used in the project.

The quantitative survey provides an overview of the total data volumes available for decision makers, plus gives a deeper insight of the actual data volumes digested by the different user profiles.

The qualitative survey consists of the selection and detailed descriptions of the data that will be used in NewsReader, a technical description of the data formats, and example-driven explanations on the importance of these specific data streams to decision makers in various economical segments.

This deliverable is divided into 4 parts:

1. Quantitative survey on available resources (Section 2)
2. Criteria for the selection of sources to use in NewsReader (Section 3)
3. Qualitative survey on core NWR resources and example scenarios (Section 4)
4. Data formats (Section 5)

We do not exclude the option of adding additional data sources to the NewsReader data set in future. As we gain deeper insights into our users' needs and new data sets become available online, we may include other relevant data sources. However, this deliverable gives an overview of the currently available data and the core data sets that we will focus on throughout the duration of NewsReader.

2. Available resources

LexisNexis databases consists 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 therefor covered by copyright.

ScraperWiki focuses on open sources. Open sources are published on the internet. These open sources are crawled and retrieved by ScraperWiki. Open data are mostly found in the public domain and are provided by for example, governments, municipalities, international intergovernmental organizations and so forth.

2.1 Insight into the amount of documents LexisNexis retrieves

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. These documents are either so called licensed sources or open websources. Licensed sources are both pulled and sent from the databases of the publishers on a daily basis. The documents are stored in the LexisNexis databases which are located in Dayton, US. The open sources are crawled from the web on a daily basis and stored for on the LexisNexis databases for a period of no longer than 90 days due to legislation.

The LexisNexis database holds approximately 40,000 sources. It includes amongst others 30,000 different newspapers (with 35,000 issues each day), 85 million company reports, over 60 million manager biographies, and several hundreds of thousands market reports. Licensing managers are continuously working on adding relevant sources for LexisNexis clients. Licensing managers approach publishers for specific sources to be added to the catalog. These sources can be added to the database by client request or because they will bring added value to the content catalog. Once LexisNexis and a publisher have come to an agreement a Licensing agreement is drawn up. Authors' rights are not transferred in such an agreement, these agreements concern merely exploitation rights.

The usage of the content is settled and described in such an agreement. Usually publishers will receive royalties based on the usage of their content by LexisNexis' clients. Content is made available digitally in

LexisNexis products in closed environments that are protected by username and password. The sources cannot be reproduced by customers in any way or form. Only the headlines can be republished, albeit only in internal client tools such intranet or newsletters. These documents are made available through our products to the customers of LexisNexis on a daily basis.

Example of a Dutch Bank:

For a Dutch bank LexisNexis retrieves relevant sector information through queries on topics relevant to this bank. Per month an average of 90,000 documents are retrieved on topics relevant to the bank. These topics are market sectors such as agriculture, automotive, leisure, construction, food, wholesale, retail, transport & logistics and business services. These documents are processed by their Business Information department which manually selects the most relevant news articles per market sector. These news articles are sent out through newsletters to all stakeholders for such a particular market sector.

2.2 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. These include sources published as PDF documents which are often seen as the most challenging resources to scrape since PDF is a "page description format" which describes the location of visual elements on the page rather than the logical structure of the document. 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.

Our process is to identify with the client the resources they require to scrape, typically a root URL on a web page which will lead to further webpages which our software will access sequentially, as required. We also discuss with clients the precise content they wish to scrape, and the format in which they require output; this may be to a database, as HTML files, XML as specified by the client, CSV files and so forth. The scale of a job depends on the number of distinct sources which need to be scraped,

typically defined by the number of root URLs and the complexity of the documents to be scraped. The number of documents to be scraped is not a major issue since once the scraper is written the process is automated, larger numbers of documents simply take longer to scrape.

Our scrapers are written in the Python programming language which gives a access to a wide range of programming library resources which SCW has supplemented with it's own specialist libraries.

Recently, SCW completed work for the UK Cabinet Office who led the UK government project to migrate the content from 34 departmental websites with many and varied designs and underlying content management systems to a single central domain gov.uk. This was complex work carried out to tight deadlines.

In addition to scraping SCW also carry out analysis and visualization work for clients. These include visualizations for the Channel 4 Dispatches TV programme which allowed viewers to explore the UK government's National Asset Register which was originally published as a large collection of PDF files, and a tool for exploring the Devon and Somerset Fire Incident Data.

3. Criteria for selection of sources to use in NWR

For the NewsReader project, LexisNexis makes project data available to be used for research purposes within the project. For the project data NewsReader can draw on the 40,000 available sources since the data will stay within the confines of the project. Depending on the specific use case and topics chosen any number of documents can be retrieved. To come to relevant sets of data for the NewsReader-project we have drafted a list of criteria that the data should meet in order to be used for NWR. In section three we explain what those criteria are.

3.1 Distinction between licensed content resources 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 for example LexisNexis.

Both Scraperwiki and LexisNexis crawl the Internet for Web sources. A Web crawler is an Internet bot that systematically spiders the World Wide Web, typically for the purpose of web indexing. 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. Parts of the Web sources that are crawled can be used for LexisNexis products. Due to legislation and regulations in the publishing sector LexisNexis cannot archive full text documents for clients. Therefor full text Web sources are not available to NewsReader through LexisNexis. However a URL and short description of the document can be stored and made available to the NewsReader by LexisNexis. Specific full text documents for relevant NewsReader web sources are scraped by ScraperWiki.

On the internet however open data are published. Open data are freely available to the public. These data can come from any source and are not restricted by copy right. Open data are crawled and retrieved by ScraperWiki. Open data are currently to be found mostly in the public domain, such as data sets provided by governments, municipalities, international organizations and so forth. A typical open data source is government data which are made available by many countries 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
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

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.

3.2 Criteria for selection of sources to use in NWR: licensed content resources and open web sources

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

1. Data sources must be 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 (meaning those sources that will be 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 webblogs 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, we will be scraping those data sources which contain large amounts of content because it is as easy to scrape ten articles from a website as it is to scrape several hundred. Therefore we gain more content for the same effort if we focus on larger sources. This is a practical criterium for the web scraping process rather than a NewsReader critera.The list of sources in section 4 are shown in order of priority with size, quality and relevance driving the ordering. Considering this collection of datasets as a whole: there are relatively few non-governmental sources which appear appropriate to scrape the exception being WikiNews, and possibly press release aggregators. Proceedings and publications of parliaments and ministries are plentiful but the material may only be relevant under certain circumstances. Additionally, these types of source will increase the number of sources to scrape rapidly without necessarily adding to the depth of material. See section 4, table 7 for a list of all open data sources that will be retrieved.
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. Example scenarios and Core NWR sources

We have defined 4 use cases for NewsReader to focus on. In this section, we explain the use cases after which we detail what data sources will be used for each use case.

Use case scenarios:

1. **Car use case**: NWR aims to model the car manufacturing domain in terms of positive and negative news stories about car manufacturers, stock market data, take-overs etc. The NWR team has chosen this domain to work in as it is international, touches upon many socio-economic issues and is well-represented in the data available from LexisNexis.
2. **TechCrunch / CrunchBase**: TechCrunch is an online news resource specialised in the technology domain. CrunchBase is a wiki-style database derived from it, describing companies in terms of size, growth and funding. It is an interesting use case for NWR to start with in Year 1 because the data is easily available and the overlap between TechCrunch and CrunchBase enables us to kickstart the evaluation of our extraction tools.
3. **Bankers Accuity / ABN Amro**: Bankers Accuity are interested in providing information on banks. This use case scenarios envisages scraping bank websites, and their company reports, for content and structuring using NewsReader such that questions such as “Who are the officers of this bank?”, “When did they join the bank?”, “Who did they work for previously?”.
4. **Dutch Parliament:** NWR is negotiating with the Dutch parliament to build a test case around the daily information need in the Dutch policy making domain. NWR will for example model the types of information needed for parliamentary enquiries and motions. In years 2 and 3, we aim to expand the parliamentary use case to also the Italian, Spanish and UK parliaments.

4.1 Example scenarios

Table 1: overview of use case scenarios

Use case	News data sources	Structured data benchmark
Car ownership use case	News articles (LN)	
TechCrunch	Eponymous	CrunchBase
Bankers Accuity & ABN AMRO	Bank websites	LexisNexis Company Data, Bankers Acuity database
Dutch Parliament	Parliamentary data, WikiNews	Original parliamentary enquiry & Parliamentary

Project data: Car brand ownership

Table 2: Car brand ownership

Format:	XML with Metadata
Language:	English
Timespan:	2003-2013
Topic:	Ownership Car brands (Take-over wars)
Size:	6,171,190 documents
Sources:	5,974 English licensed content sources (see Appendix 2 ownership car brands)

Project data: Economic Crisis

Table 3: *Economic Crisis*

Format:	XML with Metadata
Languages:	Italian, Spanish, Dutch, English
Timespan:	10 years: 2003-2013 (economic crisis)
Topics:	Economic crisis related
Size:	To be established
Sources:	LexisNexis licensed news sources & Company data, Scraped websources by ScraperWiki

Project data: CrunchBase/TechCrunch

Table 4: Crunchbase / TechCrunch

Format:	XML with metadata
Language:	English
Timespan:	2005 -present
Topic:	TechCrunch is a leading technology media property, dedicated to profiling startups, reviewing new Internet products, and breaking tech news. CrunchBase, TechCrunch's open database about start-up companies, people and investors, has become the leading statistical resource for technology companies and transactions.
Size:	To be established
Sources:	www.techcrunch.com www.crunchbase.com

Project data: Dutch parliament*

Table 5: Dutch Parliament

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

*The use case for Dutch Parliament can be duplicated for the parliaments of Italy, Spain and the United Kingdom.

Project data: Bankers Accuity/ ABN AMRO

Table 6: Bankers Accuity

Format:	XML with metadata
Language:	English
Timespan:	10 years: 2003-2013 (economic crisis)
Topic:	Company data and ownership in Banking
Size:	To be established
Sources:	LexisNexis Conpany Data, Bankers Accuity database, scraped bank websites

4.2 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. Evaluation data 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 2 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.

When it comes to the 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 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 in an intranet-environment where users will have to register with an ID and password. See Table 1 for a list of targeted sources for the evaluation data. Based on the negotiations so far and due to the tight copyright laws in the United Kingdom at this point it doesn't seem likely that LexisNexis will be able to provide UK data sources for the evaluation data set.

Evalution data will gathered based on the following criteria:

1. Data sources must be 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. Topics should be related to the usecases
5. Topics should have an international attention so we can trace stories through different countries

6. Mix of long and short term documents
7. Mix of perspectives (leftwing vs. rightwing. International vs. national view, Northern Europe (United Kingdom) vs Southern Europe (Spain, Italy)
8. Topics should be sufficiently long to be able to trace: minimum of 5 years)
9. Minimum size of the data set in each language must be 5,000 articles

Table 7: licensed content sources for evaluation data

ITALIAN	**SPANISH**	**DUTCH**
AFX - PMF*	ABC	NRC Next
ANSA Business News	Agence France Presse – Spanish	Nrc Handelsblad
ANSA Financial News	Aseguranza	Het Financieele Dagblad
ANSA Notiziario Generale in Italiano	Cinco Dias	Trouw
ANSAmed – Italian	Deutsche Presse-Agentur (Spanish)	de Volkskrant
AWP Premium Swiss News (Italian)	Diario Cordoba	Algemeen Dagblad
Business Wire Italiano	Diario Montas	Het Parool
Corriere della Sera (Italy)	Diario Vasco	
EuroNews - Versione Italiana	EFE Newswire - Albacete (Spain)	
GlobalAdSource (Italian)	EFE Newswire - Almeira (Spain)	
Hugin – Italian	EFE Newswire - Americas in Focus	
Il Giorno (Italy)	EFE Newswire - Asturias (Spain)	
Il Resto del Carlino (Italy)	EFE Newswire - Barcelona (Spain)	
ItaliaOggi	EFE Newswire - Bienestar Social (Spain)	
ItaliaOggi7	EFE Newswire - Biotecnologia (Spain)	
ItaliaOggi.it	EFE Newswire - Cadiz (Spain)	
La Gazzetta dello Sport (Italy)	EFE Newswire - Casa Real Espana (Spain)	
La Nazione (Italy)	EFE Newswire - Ciencia y Tecnologia (Spain)	
La Stampa	EFE Newswire - Coruna (Spain)	

Lavoroggi*	EFE Newswire - Cuenca (Spain)	
Marketing Oggi	EFE Newswire - Cultura (Spain)	
MF	EFE Newswire - Economia de Espana (Spain)	
MFFashion	EFE Newswire - Empresas (Spain)	
Milano Finanza	EFE Newswire - Madrid (Spain)	
News Aktuell Svizzera	EFE Newswire - Relevantes de Central America	
PC Magazine (VNU)	EFE Newswire - Relevantes de Deportes*	
PR Newswire Europe (Italian)	EFE Newswire - Relevantes de Economia*	
SDA - Servizio di base in Italiano	EFE Newswire - Relevantes de Estados Unidos	
Turismo Oggi*	EFE Newswire - Relevantes de LatinoAmerica	
	EFE Newswire - Relevantes del Mundo	
* archive only	EFE Newswire - Relevantes Hispanos	
	EFE Newswire - Toledo (Spain)	
	EFE Newswire - Turismo Cultural (Spain)	
	EFE Newswire - Valencia (Spain)	
	EFE Newswire - Vizcaya (Spain)	
	EFE Spanish Language Newswires	
	El Comercio	
	El Correo	
	El Mundo	
	El Norte de Castilla	
	El Pais	
	El Periodico de Aragon (Grupo Zeta)	
	El Periodico de Catalunya	
	El Periodico de Catalunya – Castellano	
	El Periodico Extremadura	
	El Periodico Mediterraneo	
	EuroNews - Version Espanola	
	Expansion (MADRID)	
	Global Broadcast Database - Espaol*	
	GlobalAdSource (Spanish)	
	Hoy	

	Hugin – Spanish	
	Ideal	
	La Rioja	
	La Verdad	
	La Voz de Cdiz	
	PR Newswire Europe(Spanish)	
	Spanish Newswire Services (Efe News Services)	
	Sur	
	Thomson Financial News Espana Super Focus*	

4.3 Description of main open sources

In addition to the licenced sources, there are many interesting open sources that we are considering for inclusion in the NWR evaluation data set. These will generally be obtained by scraping the web sources. The sources are listed below, note that it can be difficult to establish the quantity of material available in these sources prior to carrying out scraping.

4.3.1 Wikinews sources

Wikinews are licensed news articles going back to 2005. This source is a high priority because it is fits the requirements in terms of content but also from a technical point of view it is straightforward to use.

The data set is large, uniform, and available as a single data dump rather than requiring scraping. However, there will still be a need to transform the data to a suitable format. This source does not simply supply data that is specifically related to economic matters but it covers major economic events. In terms of authority, it is not a definitive source i.e. originating from a body with regulatory authority and its authors are essentially anonymous. However, work on Wikipedia suggests that it can generally be treated as an accurate source. The bulk downloads contain approximately 600,000 pages in the English language version. There are also versions in Spanish and Italian & Dutch although they are smaller.

Source URL
http://archive.org/details/wikimediadownloads

4.3.2 Parliamentary documentation

Parliaments generally provide some sort of verbatim record of their deliberations, some of this will relate directly to economic matters. In addition they also provide a variety of other reports and documents. This material is available in the four project language Dutch, English, Spanish and Italian from the mother country. Parliamentary documentation in Spanish and English will be available from the Americas. There is an international network of "Parliamentary Informaticians" from whom we will seek help in sourcing material. In addition to this we have good direct contacts with the Dutch parliament, and experience with scraping the UK Parliament.

Source URLs
UK: http://www.parliament.uk/business/publications/
Netherlands: http://www.tweedekamer.nl/
Spain: http://www.congreso.es/portal/page/portal/Congreso/Congreso
Italy: http://www.senato.it/home

4.3.3 European Union Newsroom

We will likely not use the verbatim records of the European parliament since these are typically in the mother tongue of the speakers and thus highly complex. However, there is a European Union newsroom including speeches by ministers, announcements of policy decisions and events. The database of releases goes back to 1975 containing 60,000 entries each of them appears to be approximately one page long.
This source is relevant in that we can anticipate that announcements of EU matters will have relevance to economic matters, and authoritative in the sense that their announcements can have regulatory weight.

Parallel releases are made in multiple languages.

Source URLs
http://europa.eu/newsroom/index_en.htm
http://europa.eu/rapid/search-result.htm?query=18&locale=en

4.3.4 Finance ministry releases

The finance ministries across Europe also make announcements. These typically contain speeches from the finance ministers, press releases and news stories. They are not particularly large sources of material, for example the UK Treasury appears to have only made 659 announcements

since May 2010. Output from finance ministries is available in the four project languages.

Source URLs
UK:
https://www.gov.uk/government/announcements?keywords=&announcement_type_option=all&topics%5B%5D=all&departments%5B%5D=hmtreasury&world_locations%5B%5D=all&direction=before&date=2013-06-01
Netherlands: https://data.overheid.nl/
Spain: http://www.minhap.gob.es/en-GB/Paginas/Home.aspx
Italy: http://www.dt.tesoro.it/
Ireland: http://www.finance.gov.ie/ (Back to 1997)

4.3.5 Press releases

Press releases are materials that people want to see published. There are a couple of large scale aggregators (~5,000,000 documents) of freely available press releases. These sources do not fit strictly within the original remit of the project but if large scale resources are required then it may be a useful route to pursue.

Source URLs
http://www.prweb.com/
http://www.prnewswire.com/

4.3.6 Discourse DB

Discourse DB is a hand-curated database of opinion pieces from journalists; "opinion roundups" are generated, pulling together multiple different pieces on a topic.

Source URL
http://www.discoursedb.org/wiki/Main_Page

4.3.7 European Central Bank

The European Central Bank makes a range of press releases that are categorised into ECB, monetary policy, statistics, payments and securities, financial stability and supervision, international and European cooperation, bank notes & coins, legal and other. There are also weekly financial statements which are presented in consistent format back to

2005, where upon they switch to PDF format. These financial statements could be scraped to provide structured data.

Source URL
http://www.ecb.int/press/pr/date/2013/html/index.en.html

4.3.8 Federal Reserve

The US Federal Reserve makes a range of press releases under the headings monetary policy, orders on banking applications, banking and consumer regulatory policy, enforcement actions and other announcements.

Source URL
http://www.federalreserve.gov/newsevents/press/all/2013all.htm

4.3.10 World Bank

The World Bank is concerned with reducing poverty and encouraging development. The "all news" page contains the following types of news article.

- Press Releases
- Feature Stories
- Opinions
- Speeches & Transcripts
- Loans & Credits
- Issue Briefs
- Results

4.3.11 World Bank blogs

When searching for all items since 1900, there are 6,833 items, the earliest of which was from December 2004. In addition there are 35 blogs hosted on the World Bank website. The first three blogs had an archive of several hundred blog posts.

Source URLs
http://www.worldbank.org/en/news/all
http://blogs.worldbank.org/blogs

4.3.12 IMF

The International Monetoary fund helps ensure stability in the international system. It does so in three ways: keeping track of the global economy and the economies of member countries; lending to countries with balance of payments difficulties; and giving practical help to members. IMF News page lists 10,418 historical articles in the following sections:

- Communiqués (123)
- Mission Concluding Statements (680)
- News Briefs (610)
- Press Releases (4,419)
- Public Information Notices (2176)
- Speeches (988)
- Statements at Donor Meetings (68)
- Transcripts (1,008)
- Views and Commentaries (365)

Source URL
http://www.imf.org/external/news/default.aspx

4.3.13 Securities and Exchange Commission (SEC)

The US Securities and Exchange Commission regulates securities markets in the US. Their archive of news releases dates back to 1997 and appears to have around 250-300 press releases per year, totalling around 4,500 articles.

Source URLs
http://www.sec.gov/news/press.shtml

4.3.14 Company Annual Reports

Publically owned companies with shareholders are legally required to produce a company report. These are generally available on the company website but there is no freely available central list leading to these reports. The Institute of Chartered Accountants in England and Wales provides some suggested routes for obtaining these reports. For the narrower question of company reports for banks it may be possible to reach these via the Bank of International Settlements which lists all central banks who should in turn list all the banks which they regulate.

Source URLs
http://www.icaew.com/en/library/company-research/company-reports-and-profiles/annual-reports
http://www.bis.org/cbanks.htm

4.4 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 has no easy machine identifiable structure.

The term "unstructured data" refers to any data that has no identifiable structure. For example, images, videos, email, documents and text are all considered to be unstructured data within a dataset.While each individual document may contain its own specific structure or formatting that based on the software program used to create the data, unstructured data may also be considered "loosely structured data" because the data sources do have a structure but all data within a dataset will not contain the same structure.

For NewsReader we have identified the following structured data sources which will be made available to the NWR project:

1. Stock market data
2. LexisNexis Company Dossier:reports covering more than 43 million public, private and international companies
3. Worldbank
4. Organisation for Economic Co-operation and Development (OECD)
5. International Monetary Fund (IMF)
6. US Securities and Exchange Commission (SEC) Company Fillings

7. The DBpedia Knowledge Base[1]. DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web. DBpedia allows you to ask against sophisticated queries Wikipedia, and to link the different data sets on the Web to Wikipedia data.

5. Data formats

LexisNexis provides XML (Extensible Markup Language) to the NewsReader project. XML is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable. It is defined in the XML 1.0[2] Specification produced by the by the World Wide Web Consortium (W3C), and several other related specifications, all open standards.

Many application programming interfaces (APIs) have been developed to aid software developers with processing XML data, and several schema systems exist to aid in the definition of XML-based languages.

As of 2009, hundreds of document formats using XML syntax have been developed, including RSS, Atom, SOAP, and XHTML. XML-based formats have become the default for many office-productivity tools, including Microsoft Office (Office Open XML), OpenOffice.org and LibreOffice (OpenDocument), and Apple's iWork. XML has also been employed as the base language for communication protocols, such as XMPP.

5. 1 Technical description of data formats provided to NWR

The XML LexisNexis provides to the NewsReader project is News Industry Text Format (NITF), an XML specification published by the International Press Telecommunications Council that is designed to standardize the content and structure of individual text news articles. The NITF specification defines a standard way to mark up an article's content and structure, as well as a wide variety of metadata that different organizations may choose to use.

[1] http://wiki.dbpedia.org/About
[2] http://www.w3.org/TR/REC-xml/

The format is widely used across the news industry. Newspapers, news agencies and archival services use NITF for inter-agency transmission of news as well as internal transmission and storage. An example document is provided in Appendix 1[3].

ScraperWiki will gather open sources on the internet. The datasets gathered from the internet can comprise of data formats different from the News Industry Text Format (NITF). In those cases where other data formats such as HTML, PDF or some other form of XML are being scraped from the internet, the data will be transformed. Therefor all sources used for NewsReader are made up of the same common data format. This form of standardization and structuring of data formats supports an easily accessible database available to the NewsReader-project.

[3] http://www.iptc.org/std/NITF/3.4/documentation/nitf-documentation.html

Conclusions

In this deliverable, we provided a survey of available resources in the four NewsReader languages. Both open sources scraped from the internet and licensed sources from LexisNexis are avaible to the project. As the use cases develop and the interviews provide us with more customer insights we can include (or if needed exclude) specific data sources.

As the adding of new data sources to the LexisNexis Database is a continuous process, new data sources will be become available during the course of the NewsReader-project. When relevant, these will be added to the existing NewsReader data sources.

The usage of evaluation data is currently being discussed with the Publishers.We aim for the negotiations with regards to the evaluation data to finish in the third quarter of 2013.

This deliverable on data sources provides an overview of the enormous amounts of data available on the internet and within the LexisNexis databases. To come to the datasources relevant to NWR we have given a set of criteria these data should meet. These large amounts of newsdata resemble the kinds of data volumes decisionmakers have to digest on a daily basis. Providing and structuring the relevant data sources in the same manner to the NWR-project allows research to be done for the identified use cases as well as delivering data for evaluation purposes.

Appendix 1 Technical example of data format

All sources in the NewsReader project will be structured in the same manner using the same metadata. See below for an example of the NITF XML for the NWR project:

<?xml version="1.0" encoding="UTF-8"?><!DOCTYPE nitf PUBLIC "-//IPTC//NITF DTD 3.0//EN" "nitf-3-0.dtd"><nitf baselang="EN-US"><head><meta content="FULL" name="_lnbillview"/><meta content="00005" name="_lnminrev"/><meta content="November 24, 2011" name="_loaddate"/><meta content="November 24, 2011" name="_eoptdate"/><meta content="07:20:45 EST" name="_eopttime"/><meta content="e6874d82b9c4d010f7292aaa48e5d02945a985aabb3a1ea6ec8fa921675cc58deabc2074a13abfce5dd1ac87dc480217d261f9cf18d5dfd4e02ec535b2c367e324d29f061943830611aa1dbfa688ec132afad0e41dece924df18e415fa7b33f729405734efd3fddd43c9fc11585b959277397b6dee35e52ee0b31093c88289d3" name="documentToken"/><docdata><doc-id id-string="54B2-BBY1-JD34-P1B9" regsrc="lexisnexis.com"/><date.issue norm="20120101T000000Z"/><doc.copyright holder=" Copyright 2012 Business Monitor International Ltd. All Rights Reserved"/></docdata><pubdata unit-of-measure="word" item-length="181" name="Argentina Autos Report"/></head><body><body.head><hedline><hl1>Peugeot Citroën Argentina (PCA) - Q1 2012</hl1></hedline></body.head><body.content><block class="publogo"><media media-type="image"><media-metadata name="attachmentId" value="LNCDBE032A334E6199C977FF087B5CAE965F5D50C17247B097"/><media-reference source="unavailable" alternate-text=" Argentina Autos Report " mime-type="application/octet-stream"/></media></block><block class="lead"> Market Position </block> <block>
 <table cols="1"><col align="left" width="556"/><tbody><tr><td> Peugeot Citroën Argentina (PCA) </td></tr><tr><td>Company Data

Domestic sales, 2010: 78,514 CBUs

No. of employees: 1,560</td></tr></tbody></table> In 2010, PCA sold 78,514 units in Argentina. This put PSA in fourth position with a market share of 15.8%. In terms of production, however, PSA emerged as Argentina's second biggest vehicle producer, building 126,968 units during the year. Products PCA operates two manufacturing facilities

in Buenos Aires and Jeppener. The Buenos Aires plant, which has an annual production capacity of 140,000 units, produces the Peugeot 206, Partner and Citroën Berlingo models. The 307 joined the local line-up in 2004 as PCA introduced its mid-range platform to the country. The 307 model is produced mainly for export. The Jeppener facility focuses on components, producing mechanical, power-train, transmission and front suspension components for all of the models produced locally, as well as the 307. In addition, the plant manufactures engines for the 206, 307, Partner, Citroën Xsara, and Berlingo. The company began producing a new Citroën model in its Buenos Aires plant in 2007. </block><note class="subject" type="pa"><body.content> AUTOMAKERS (78%); MANUFACTURING OUTPUT (76%); MANUFACTURING FACILITIES (75%); PLANT CAPACITY (74%); MARKET SHARE (56%) </body.content></note> <note class="company" type="pa"><body.content> PSA PEUGEOT CITROEN SA (95%) </body.content></note><note class="ticker" type="pa"><body.content> UG (PAR) (95%) </body.content></note><note class="city" type="pa"><body.content> BUENOS AIRES, ARGENTINA (91%) </body.content></note> <note class="country" type="pa"><body.content> ARGENTINA (98%); LATIN AMERICA (79%) </body.content></note> <note class="publication-type" type="pa"><body.content> Newsletter </body.content></note></body.content></body></nitf>