Early Demonstrators

Deliverable D1.2

Version first draft

Authors: Pim Stouten[1]
Affiliation: (1) LexisNexis

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	First draft
Contractual Date of Delivery	April 2013
Actual Date of Delivery	July 2013
Type	Report
Security (distribution level)	Public
Number of Pages	24
WP Contributing to the deliverable	WP01
WP Responsible	WP01
EC Project Officer	Sophie Reig
Authors: Rutger Kortleven	
Keywords: data sources, content, XML, open sources, licensed content	

Executive Summary/Abstract

This document summarizes the various information delivery solutions currently available on the market, ranging from free consumer-orientated tools to free tools for specific (professional) users, as well as commercially available solutions.

The document contains an overview of the various providers, their product offering, as well as a summary of the benefits and limitations of each product.

Table of Revisions

Version	Date	Description and reason	By	Affected Section
0.1	22-07-2013	First draft	Pim Stouten	

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make well-informed decisions and professionals risk to be held liable for decisions based on incomplete, inaccurate and out-of-date information.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 tools and products that deliver business-relevant information to end users.

This deliverable consists of a source typology (discerning different types of user-relevant information), listing the main products/tools/solutions used in the market per information type.

An overview of the pros and cons is given per solution and, where relevant, screenshots have been added to illustrate the differences and overlaps between the various tools.

2. Information tools and products: typology

This chapter gives an overview of the different information tools currently available to professional information users.

Typology is based on the goal of a user for the tool at hand: what kind of information are they after, and in which format should it be presented?

Categories are, where relevant, broken down into open source/free on the one hand, and closed source/commercial on the other.

The scope of this overview is limited to those information sources containing **historical** information, i.e. covering information on a certain topic or entity for a longer period of time.

2.1 Web search

Web search is a widely used solution to understand 'the bigger picture'; it gives a good first indication of what goes on about a certain topic, and can help pin-point information on a very specific subject.

Services like www.google.com, www.bing.com, www.baidu.com, www.yandex.ru and www.wolframalpha.com are all available free of charge and are used by a diverse audience, consisting of both consumers and professionals.

Pros:

- Free of charge
- Available in many different languages
- Cover enormous amounts of web pages

- Easy to use: most work
- No access thresholds: works from any device without user credentials

Cons:

- The search engine is a 'black box': with the business models of search suppliers changing towards facilitators of online advertising, there is a visible decline in result quality. Many of the LexisNexis customers interviewed mention this as one of the key reasons to distrust results from 'open' search providers
- No source/data qualification: web pages are indexed, regardless of quality, and results are ranked using algorhitms optimized for advertising. In other words: the 'best result' according to a search engine might not be the *right* result for a specific user.

Wolfram Alpha is the exception in this category, their goal is to understand the user's question first and then find ('compute' in their jargon) the right answer to that question. Wolfram Alpha **does** use source qualification to ensure results have relevance.

2.2 Structured databases

Structured databases are those containing historical data in a highly structured format. This ranges from general statistics to company information to exchange rates, stock prices and more.

The growing trend of *open data* initiatives is opening up more and more public-sector databases for the general public, which also impacts professional users: there is more and more well-structured information available through public sources.

2.2.1 Statistical data

Statistical data (economical, demographical, etc.) is probably the oldest example of structured data, and (see comment above) increasingly become available online through open data initiatives.

Availability of data is especially an issue in developing markets; it is hard to find reliable data, either through open or commercial sources.

Pros:

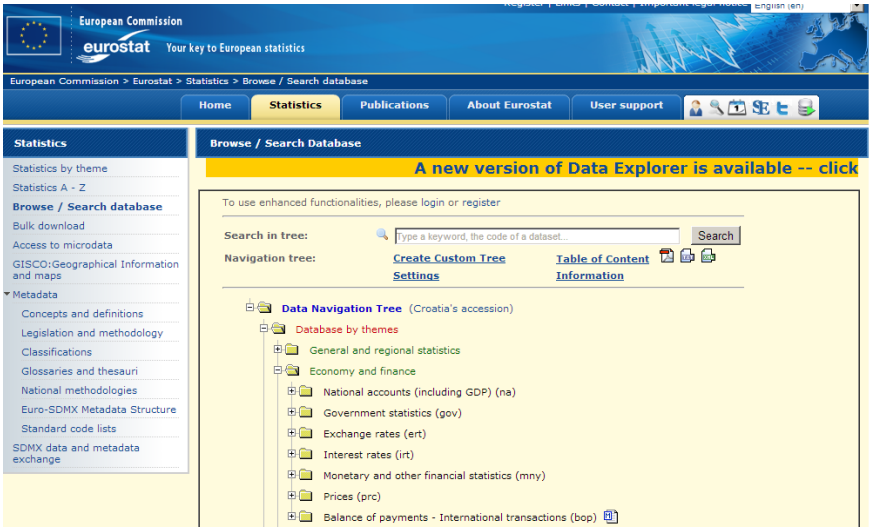
- Structured data makes result interpretation easier
- Growing number of open data (free of charge) sources
- High-quality data from reputed organisations/suppliers
- Many (both open source and commercial) come with integrated search/compare/analysis tools, quickly turning search results into meaningful analysis.

Cons:

- Coverage: not all countries/regions have good statistical coverage, in those cases secondary data is needed to interpret or extrapolate data values
- Standardization: not all statistics agencies/suppliers use identical terms for similar data elements, calculation of more complex (consisting of multiple data elements) elements can also differ.

2.2.1.1 Open source

One of the best-known open sources is the EU's *Eurostat* database (http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database), which contains multi-year statistics on a plethora of different topics.



Another one is the database from the World Bank, more on that in paragraph 2.2.3.1.

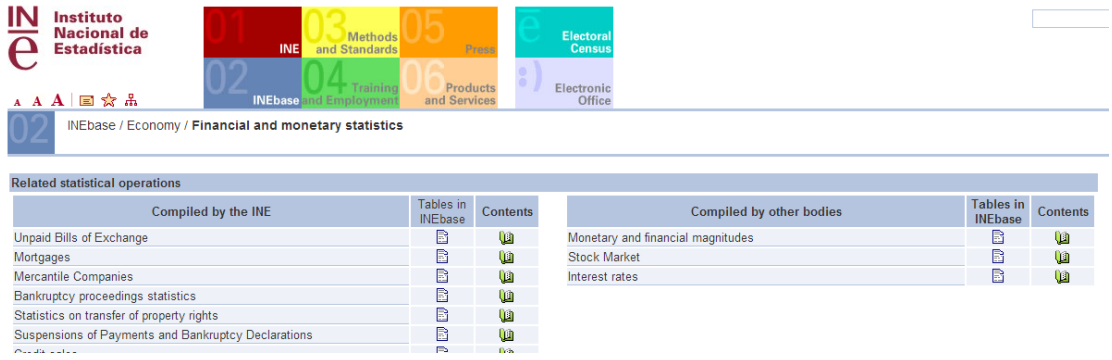
Similar databases are local/country-based, and have similar lay-outs and functionalities. Some project-relevant examples are:

CBS (Netherlands): http://www.cbs.nl/nl-NL/menu/home/default.htm

UK National Statistics: http://www.statistics.gov.uk/hub/index.html

ISTAT (Italy): http://www.istat.it/en/

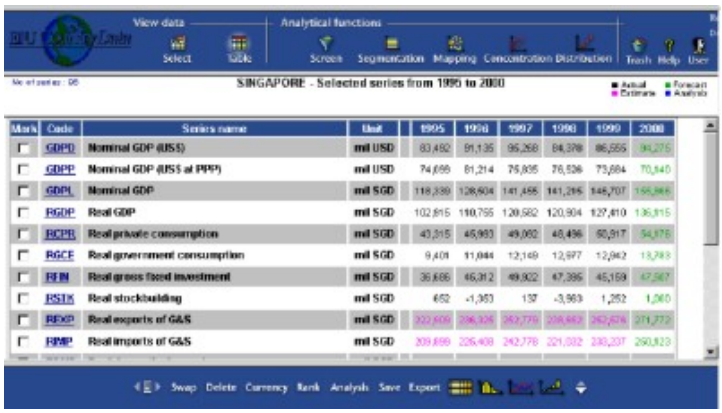
Instituto Nacional de Estadística (Spain): http://www.ine.es/en/



2.2.1.2 Commercial solutions

There are many small suppliers, focusing on niche data sets, i.e. purchasing behavior in a specific industry.

The best-reputed larger solutions are ABI/INFORM® (http://www.proquest.com/en-US/catalogs/databases/detail/abi_inform.shtml) by ProQuest and Countrydata (http://www.eiu.com/site_info.asp?info_name=ps_countryData) by The Economist Intelligence Unit



2.2.2 Company information

Company information, like statistics, is a growing target for the open data movement: the last few years have seen a steady growth of open company data initiatives.

The commercial side of the market has a distinct split between a handful of international players versus a large amount of local/regional suppliers.

Availability of data is especially an issue in developing markets; it is hard to find reliable data, either through open or commercial sources.

Pros:

- Structured data makes result interpretation easier
- Growing number of open data (free of charge) sources
- High-quality data from reputed organisations/suppliers
- Many (both open source and commercial) come with integrated search/compare/analysis tools, quickly turning search results into meaningful analysis

- Different 'depth' levels of information: a user can decide what to retrieve/purchase, ranging from just address data to highly detailed annual reports or ownership trees
- Growing level of standardisation, thanks to e.g. the XBRL reporting standard for annual reports
- Several commercial players (Bureau van Dijk is considered best of breed) are combining national data sets to create a 'full picture' of an organisation's ownership tree

Cons:

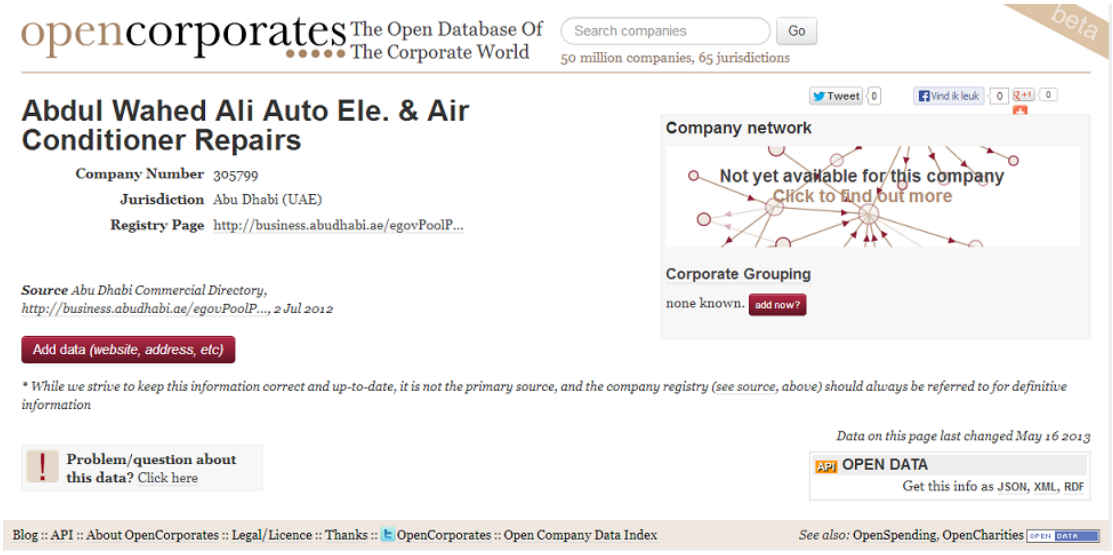
- Coverage: not all countries/regions have good company coverage, in those cases secondary data is needed to interpret or extrapolate data values
- Completeness: local legislation strongly influences data availability. An example is revenue numbers: known for almost the full company population in e.g. Belgium, but only for a small percentage of the registered companies in the Netherlands.
- Globalisation: most company registers are maintained on a national level, creating a natural border to cover much-required information like cross-national company ownership. Identifying the so-called GUO (Global Ultimate Owner) or UBO (Ultimate Beneficial Owner) is a requirement in international anti-money laundering and anti-bribery legislation, but very hard to fulfill with the fragmented data on offer
- Standardisation: XBRL is a step in the right direction for company financials, but other elements of company information (e.g. ownership trees) still lack standardisation, which makes data comparison/integration extremely time-consuming and complex.

2.2.2.1 Open source

Several national company information registrars work with a *freemium* model: basic company information is free, more detailed information is only available on a pay-per-view basis.

A sample from the Dutch Chamber of Commerce (KvK, www.kvk.nl) here, address and other 'basics' are free; paid information is found behind the 'bestel nu' (order now) button:

The best-known completely open source (i.e. no underlying freemium model) solution is opencorporates (opencorporates.com). The have indexed the various local registrars (see comment above), and currently cover over 55 milion companies world-wide. This solution lacks depth as in the example above: a user would still need to cross the local registrar's paywall to access the required information.

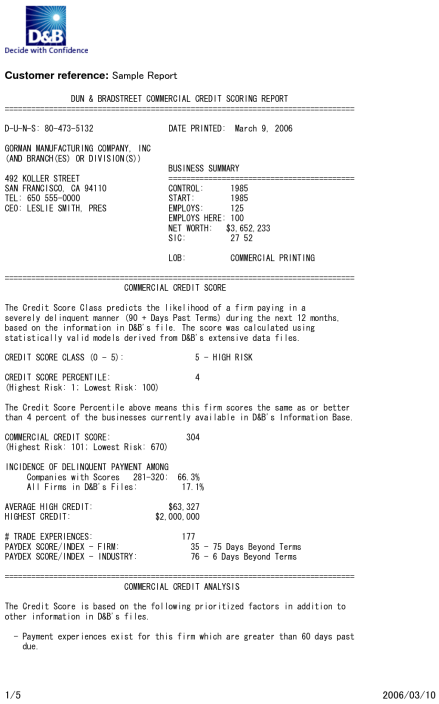


2.2.2.2 Commercial solutions

There are two types of commercial suppliers, as already mentioned in the introduction: operating on either a national/regional or global level. Given the project's international scope we decided to limit this overview to the key international suppliers of company information.

Experian (www.experian.com) has a strong focus on company financials and credit scoring, more 'information delivery' than 'information analysis'.

Dun & Bradstreet, including brands like Hoover's (www.dnb.com) has a similar background and focus as Experian, although there is a gradual development towards more intuitive, dashboard-like solutions. Probably the strongest brand name in the market, with the largest data coverage in this market segment.



D&B
Decide with Confidence

Customer reference: Sample Report

DUN & BRADSTREET COMMERCIAL CREDIT SCORING REPORT

D-U-N-S: 80-473-5132

DATE PRINTED: March 9, 2006

GORMAN MANUFACTURING COMPANY, INC
(AND BRANCH(ES) OR DIVISION(S))

492 KOLLER STREET
SAN FRANCISCO, CA 94110
TEL: 650 555-0000
CEO: LESLIE SMITH, PRES

BUSINESS SUMMARY

CONTROL: 1985
START: 1985
EMPLOYS: 125
EMPLOYS HERE: 100
NET WORTH: $3,652,233
SIC: 27 52

LOB: COMMERCIAL PRINTING

COMMERCIAL CREDIT SCORE

The Credit Score Class predicts the likelihood of a firm paying in a severely delinquent manner (90 + Days Past Terms) during the next 12 months, based on the information in D&B's file. The score was calculated using statistically valid models derived from D&B's extensive data files.

CREDIT SCORE CLASS (0 - 5): 5 - HIGH RISK

CREDIT SCORE PERCENTILE: 4
(Highest Risk: 1; Lowest Risk: 100)

The Credit Score Percentile above means this firm scores the same as or better than 4 percent of the businesses currently available in D&B's Information Base.

COMMERCIAL CREDIT SCORE: 304
(Highest Risk: 101; Lowest Risk: 670)

INCIDENCE OF DELINQUENT PAYMENT AMONG
Companies with Scores 281-320: 66.3%
All Firms in D&B's Files: 17.1%

AVERAGE HIGH CREDIT: $63,327
HIGHEST CREDIT: $2,000,000

TRADE EXPERIENCES: 177
PAYDEX SCORE/INDEX - FIRM: 35 - 75 Days Beyond Terms
PAYDEX SCORE/INDEX - INDUSTRY: 76 - 6 Days Beyond Terms

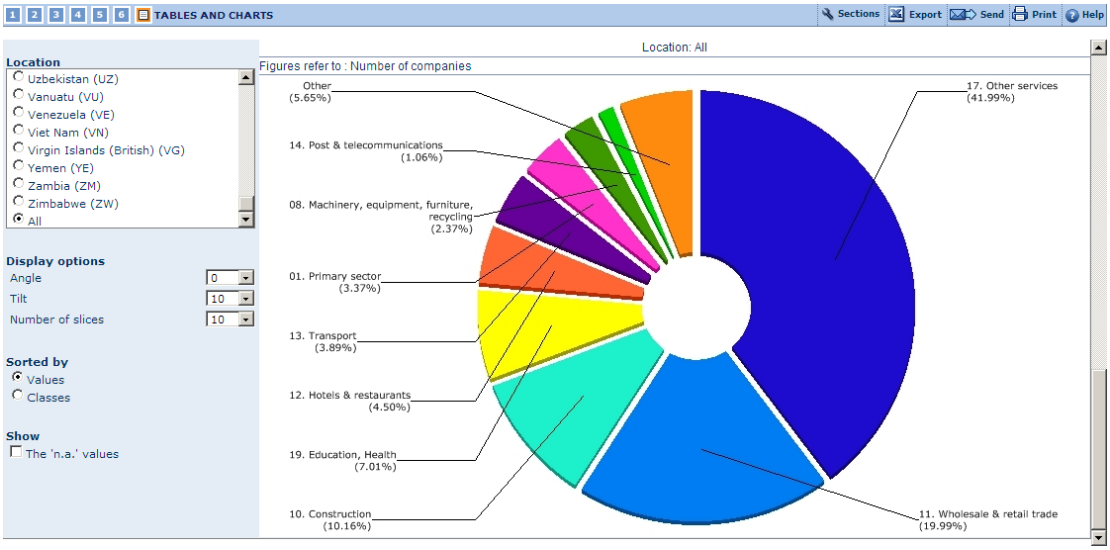
COMMERCIAL CREDIT ANALYSIS

The Credit Score is based on the following prioritized factors in addition to other information in D&B's files.

- Payment experiences exist for this firm which are greater than 60 days past due.

1/5 2006/03/10

Bureau van Dijk (www.bvdinfo.com) has focused on data normalisation and analysis, which has resulted in a best-of-breed set of ownership information, as well as integrated visualisation and analysis tools.



2.2.3 Financial/trading information

Financial information and company information are sometimes considered to be overlapping areas, from our perspective company financials are NOT financial information. This category contains elements where often timeliness is key: some trading systems express latency in milliseconds, whereas other information suppliers measure in minutes, hours or even days.

We focus at financial information that has relevance to many different users in different situations, namely: interest and exchange rates and share/commodity prices.

Pros:

- Structured data makes result interpretation easier
- Growing number of open data (free of charge) sources
- High-quality data from reputed organisations/suppliers
- Many (both open source and commercial) come with integrated search/compare/analysis tools, quickly turning search results into meaningful analysis

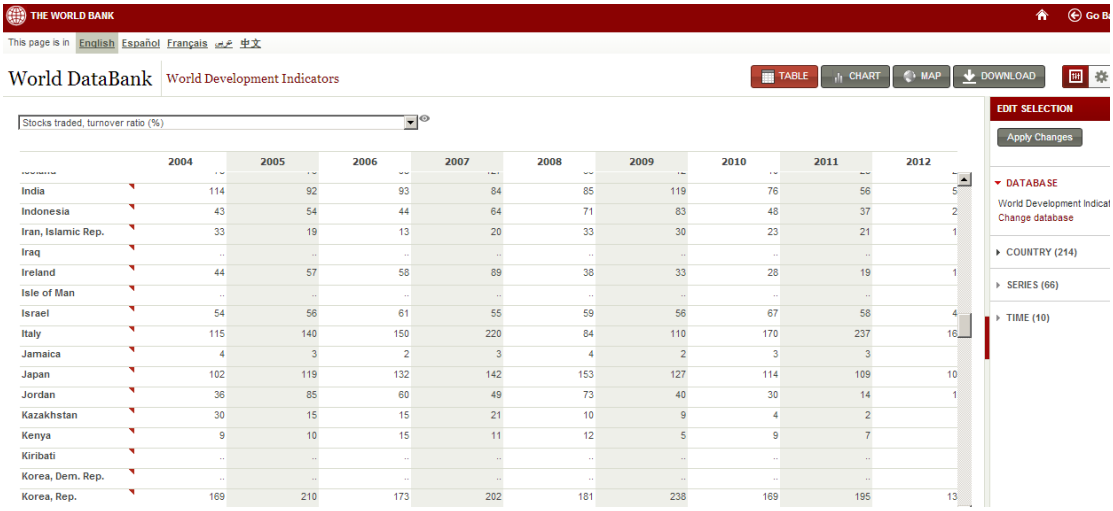
Cons:

- Coverage: not all countries/regions have reliable coverage, in those cases secondary data is needed to interpret or extrapolate data values.

- Timeliness: especially in the trading business, swift information delivery is key. Not all suppliers offer similar quality in this area.
- Costs: whereas much of the other information types have decent or 'good-enough' free alternatives, the tools used in the trading market come with price tags that easily reach several tens of thousands of € per user, creating a threshold for non-profit or startup users of this information.

2.2.3.1 Open source

The World Bank (http://data.worldbank.org/), already mentioned in the paragraph on statistical information sources (2.2.1.1) also stores specific information on the financial sector, and is the best-reputed independent source on a global scale for key financial/economical indicators.

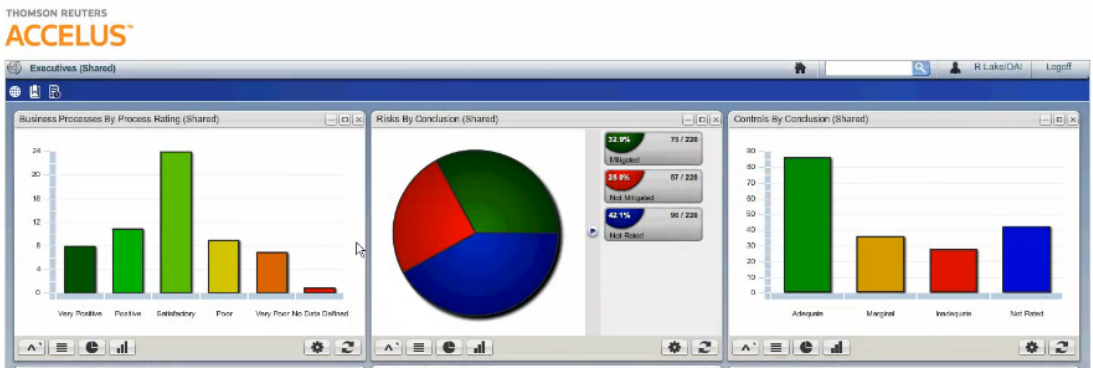


Due to its more statistical nature, World Bank data is 'slower' than the feeds offered by the for-profit suppliers (see paragraph below).

2.2.3.2 Commercial solutions

ThomsonReuters (www.thomsonreuters.com) has multiple brands and specific solutions, focusing on various niches and job roles within the finance industry, e.g. investment banking, M&A. They have recently

started working on a portfolio for the risk/compliance market specifically, bundled under the brand name Accelus.



Bloomberg (www.bloomberg.com) focuses –especially outside of the US- on the trading niche, where realtime and near-time information is key. Many Bloomberg solutions still use bespoke interfaces and keyboards (the combination is known in the market as a 'Bloomberg Terminal'), setting it apart from virtually all other suppliers in this overview.

2.3 (Financial) News

News sources in the financial/economical field are plentiful; we therefore made the conscious choice to limit scope (as with some of the previous paragraphs) to the international suppliers in this category.

This creates a rather strange situation: most of the organisations listed below do not **produce** news (like a newspaper or television network would), but aggregate data from the organisations and individuals that create the news stories.

Pros:

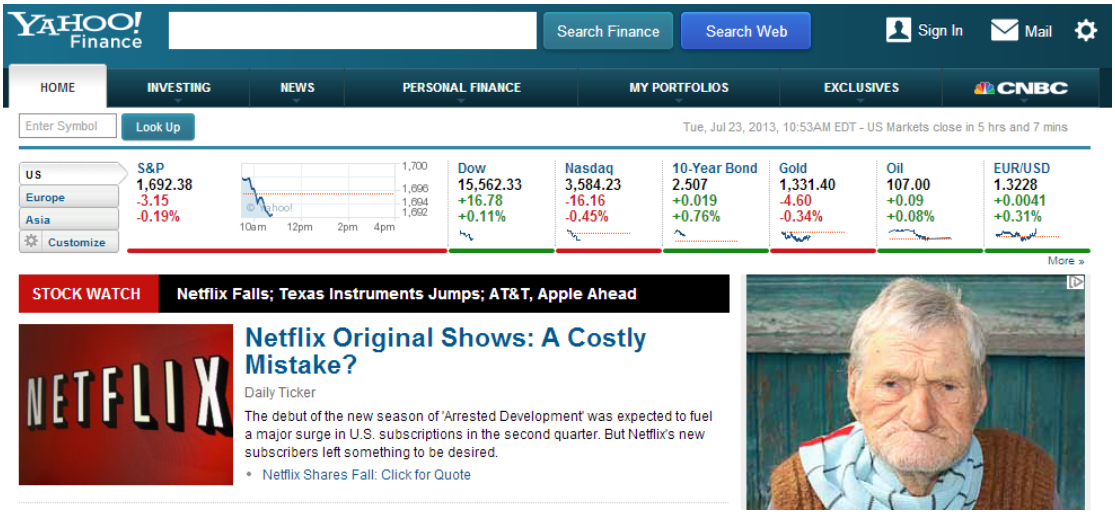
- Web scraping technology has become a commodity: it's easier than ever to 'scrape' data from different web sites
- High-quality data from reputed organisations/suppliers
- Much more efficient than having to search/visit the individual sources for the same information
- "You get the full picture": aggregating news gives a user different opinions/backgrounds on the same subject

Cons:

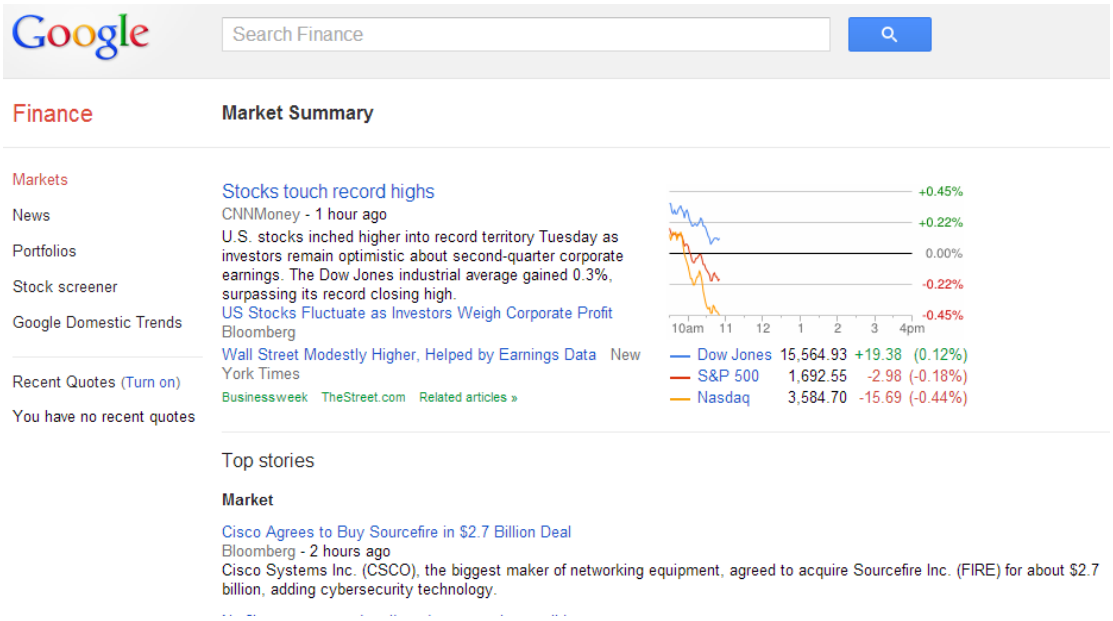
- Unstructured data, requires strong metadata, substantially more risk of 'noise' results than with other information types.
- Coverage: not all countries/regions have reliable coverage, because of low web presence of their news media. The only workaround in those cases is to scan/digitise the hardcopy publications.
- Source qualification: no or weak qualification of sources/authors will also cause noise results, or even worse: incorrect information presented as being true.
- Growing number of paywalls (e.g. Financial Times, New York Times) decrease the value of free solutions, since they can't aggregate data beyond the paywall.

2.3.1 Open source

Yahoo! Finance (finance.yahoo.com) is a spin-off from Yahoo!'s search engine activities, aggregating news in the financial/economical field. Beyond bringing headlines from different sources into one place, Yahoo! also offers graphs and charts, basic analysis features and alerting functionality, which sends users an alert when a story of interest is detected.

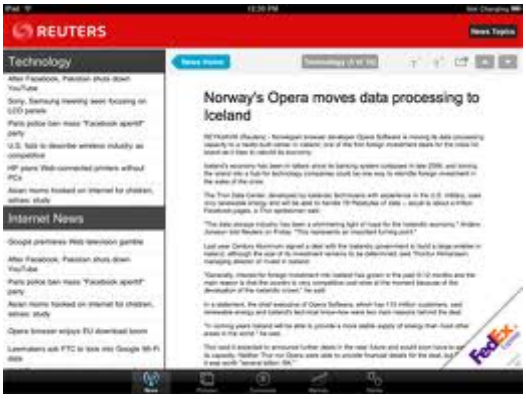


Google News/Google Finance (https://news.google.com and https://www.google.com/finance) offer similar functionalities as Yahoo!, with more filtering/personalization possibilities, but (based on customer feedback gathered by LexisNexis) weaker matching algorithms, resulting in more noise results.



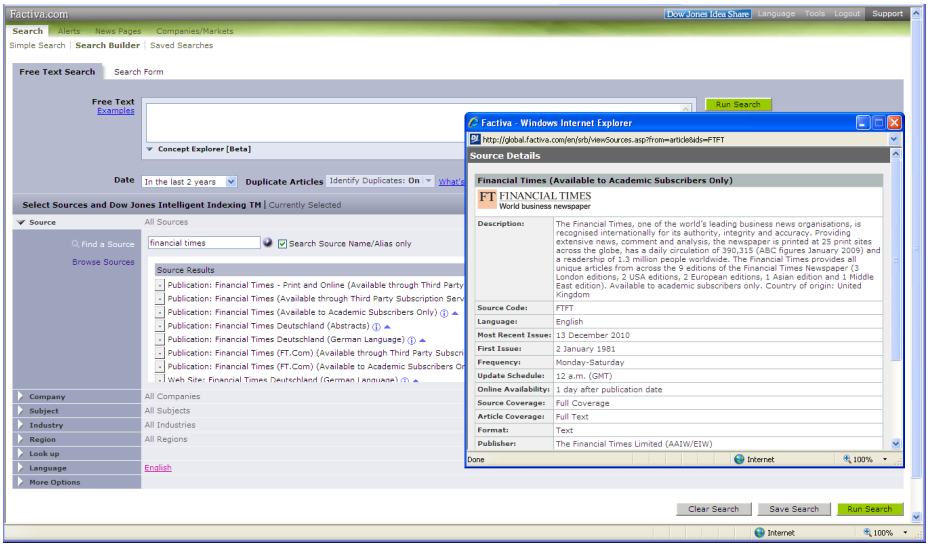
2.3.2 Commercial solutions

Reuters (www.reuters.com) started as a news agency, and has developed beyond that, by aggregating news from other sources and expanding into ‘non-news’ areas, like company information.



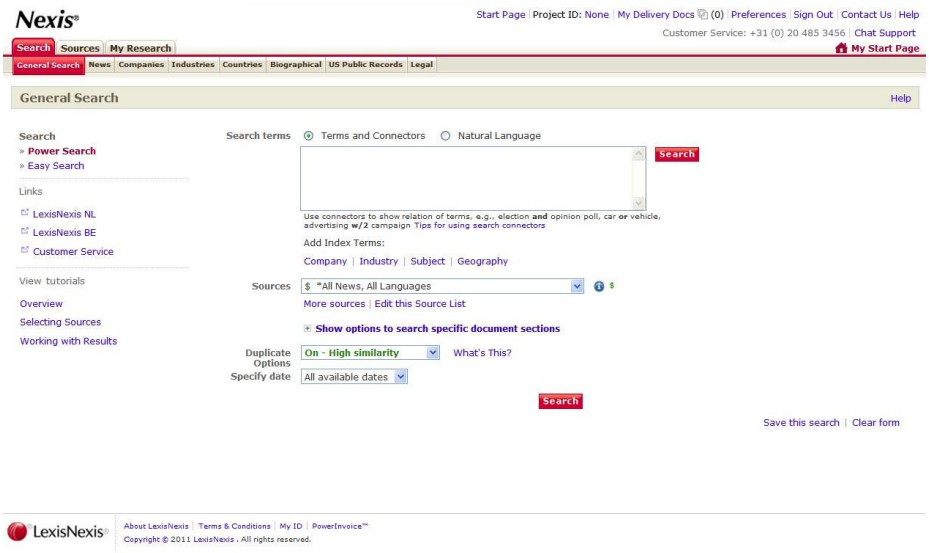
Dow Jones/Factiva (www.dowjones.com and www.factiva.com) has a similar background, but has invested substantially more in data aggregation, licensing information from other parties, like newspaper publishers and company information providers. Many of the aforementioned suppliers appear as a source in Factiva’s (and LexisNexis’, for that matter) solutions.

Factiva’s traditional products focus at delivering documents to users, only a small range of solutions (see next paragraph) delivers ‘insight’ beyond just document text.



LexisNexis (www.lexisnexis.com) has a different background from the other two, and started as a legal database in the 1970s (Lexis). With an increasing need for non-Legal information, Nexis was launched, focusing at business information: news, company, market, country, patent information, using a similar model as Factiva by licensing information from their publishers/creators.

LexisNexis' traditional products focus at delivering documents to users, only a small range of solutions (see next paragraph) delivers 'insight' beyond just document text.



2.4 Portals & dashboards

Portals and dashboards aggregate and integrate much of the information mentioned above, and turn textual/numerical data into visuals, helping users make their decisions as quickly and efficiently as possible.

Most providers fall into an 'or-or' camp: they either deliver the software to develop these front ends, or they deliver the information streams that feed these dashboards. The first camp mainly consists of companies in the software field (BI, analytics, statistics), the second camp is populated by the companies mentioned in the previous paragraphs of this document.

To limit scope to manageable numbers we decided to only list organisations that A) have an international focus and B) offer both content AND front end.

Pros:

- Web scraping technology has become a commodity: it's easier than ever to 'scrape' data from different web sites
- High-quality data from reputed organisations/suppliers
- Much more efficient than having to search/visit the individual sources for the same information
- "You get the full picture": aggregating news gives a user different opinions/backgrounds on the same subject
- The visual layer makes it much easier to handle large data volumes: i.e. you look at a change in a graph, rather than having to read the thousands of individual documents that created the graph
- The distinction between data and front end creates room to build in business rules and processes in between, ensuring that the visual output matches the rules and logic of a specific user's organisation or department.
- Pro-active: these solutions 'push' information to their users

Cons:

- Unstructured data, requires strong metadata, substantially more risk of 'noise' results than with other information types.
- Coverage: not all countries/regions have reliable coverage, because of low web presence of their news media. The only workaround in those cases is to scan/digitise the hardcopy publications.
- Source qualification: no or weak qualification of sources/authors will also cause noise results, or even worse: incorrect information presented as being true.
- Growing number of paywalls (e.g. Financial Times, New York Times) decrease the value of free solutions, since they can't aggregate data beyond the paywall.
- Analysis is as strong as the weakest algorithm driving it: 'black box' solutions come with the risk of getting results without understanding where they came from.

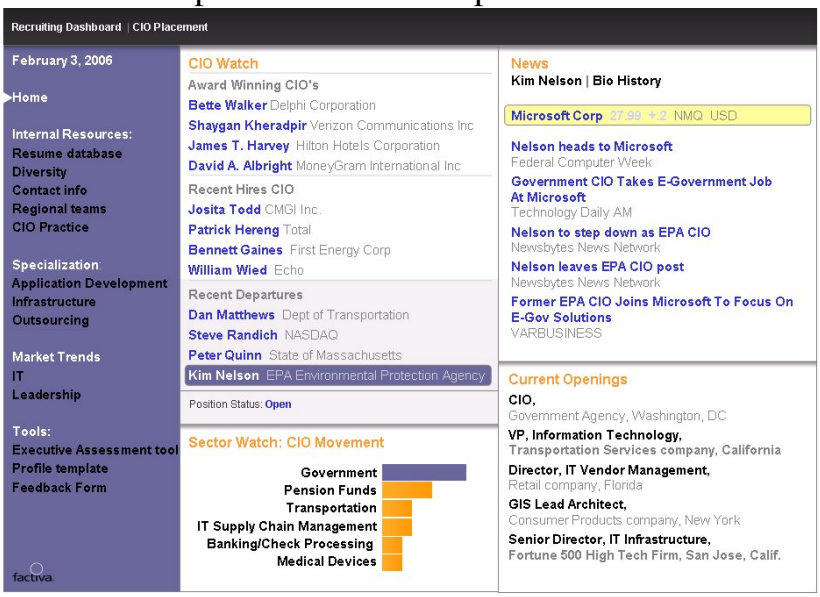
2.4.1 Open source portals & dashboards

EMM (http://emm.newsbrief.eu/NewsBrief/clusteredition/nl/latest.html) is a joint project by EU's DC CCO and DC COMM.

It processes, analyses and visualises newsfeeds from all over the globe, categorising them by e.g. theme, country or entity.

2.4.2 Commercial portals & aggregators

Factiva (www.factiva.com): see previous paragraph for company backgrounds. Factiva offers various widgets, as well as powerful APIs to deliver a complete dashboard experience.



LexisNexis (www.lexisnexis.com): see previous paragraph for company backgrounds. LexisNexis offers widgets and APIs, plus various toolkits which can be used to build dashboards from scratch, using information from different sources. These toolkits are mostly used to develop **for** customers, whereas the widgets and APIs are used to develop bespoke portals and dashboards **by** customers.