## DATA CRAWLERS, ADAPTORS AND EXTRACTORS

<table>
<thead>
<tr>
<th>Work package No.</th>
<th>WP2</th>
<th>Work package Title</th>
<th>Linked Data Life Cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task No.</td>
<td>T2.1</td>
<td>Task Title</td>
<td>DATA HARVESTING, EXTRACTION AND ASSESSMENT</td>
</tr>
<tr>
<td>Milestone No.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organization name of lead contractor for this deliverable</td>
<td>SWC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Editor</td>
<td></td>
<td></td>
<td>Heidelinde Hobel (SWC)</td>
</tr>
<tr>
<td>Contributors</td>
<td></td>
<td></td>
<td>Tomas Knap, Artem Revenko (SWC), Ioannis Pragidis, Eirini Karapistoli, Antonios Sarigiannis, Panagiotis Tsintzos (DUTH), George Panos, Christoforos Bouzanis (UoG), Anna Satsiou, Ioannis Kompatsiaris (CERTH)</td>
</tr>
<tr>
<td>Reviewers</td>
<td></td>
<td></td>
<td>Artem Revenko (SWC), Peter Hanečák (EEA)</td>
</tr>
<tr>
<td>Status (F: final; D: draft)</td>
<td></td>
<td></td>
<td>F</td>
</tr>
<tr>
<td>Nature</td>
<td></td>
<td></td>
<td>R – Report</td>
</tr>
<tr>
<td>Dissemination Level</td>
<td></td>
<td></td>
<td>PU – Public</td>
</tr>
<tr>
<td>Project start date and duration</td>
<td></td>
<td></td>
<td>January 2016, 36 Months</td>
</tr>
<tr>
<td>Due Date of Deliverable:</td>
<td></td>
<td></td>
<td>December 30, 2016</td>
</tr>
<tr>
<td>Actual Submission Date:</td>
<td></td>
<td></td>
<td>December 30, 2016</td>
</tr>
</tbody>
</table>
## REVISION HISTORY

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Modified by</th>
<th>Changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>08-11-2016</td>
<td>H. Hobel (SWC)</td>
<td>First version of ToC</td>
</tr>
<tr>
<td>0.2</td>
<td>09-11-2016</td>
<td>H. Hobel (SWC)</td>
<td>Started unstructured data section</td>
</tr>
<tr>
<td>0.3</td>
<td>16-11-2016</td>
<td>H. Hobel (SWC)</td>
<td>Unstructured data section: RSS feeds</td>
</tr>
<tr>
<td>0.4</td>
<td>17-11-2016</td>
<td>H. Hobel (SWC)</td>
<td>Revision, unstructured data section</td>
</tr>
<tr>
<td>0.5</td>
<td>18-11-2016</td>
<td>H. Hobel (SWC)</td>
<td>Social Media section</td>
</tr>
<tr>
<td>0.6</td>
<td>21-11-2016</td>
<td>H. Hobel (SWC)</td>
<td>Started structured data section</td>
</tr>
<tr>
<td>0.7</td>
<td>23-11-2016</td>
<td>H. Hobel (SWC)</td>
<td>Finalized structured data section</td>
</tr>
<tr>
<td>0.8</td>
<td>28-11-2016</td>
<td>H. Hobel (SWC)</td>
<td>Included feedback provided by I. Pragidis, Revisions, Introduction, Conclusion, Background</td>
</tr>
<tr>
<td>0.9</td>
<td>05-12-2016</td>
<td>H. Hobel (SWC)</td>
<td>Revisions</td>
</tr>
<tr>
<td>1.0</td>
<td>12-12-2016</td>
<td>H. Hobel (SWC)</td>
<td>Incorporated feedback</td>
</tr>
</tbody>
</table>
List of Figures

1. Overview page for the retrieval of articles .............................................. 16
2. Semantic data model for resource investing.com .................................... 17
3. Breakdown of the collected articles from investing.com ............................ 18
4. Generic pipeline for RSS feed extraction .................................................. 20
5. Single resource pipeline for RSS feed extraction without graph name in- 
   ference ........................................................................................................ 20
6. Semantic data model for RSS feeds enriched with full-text ...................... 21
7. Breakdown of retrieved RSS feeds .............................................................. 24
8. Pipeline used to extract articles provided by an API ................................... 26
9. Semantic data model for information retrieved from the Guardian News 
   API ............................................................................................................. 27
10. Breakdown of the news results obtained from APIs ................................... 27
11. Workflow for parsing LinkedIn profiles ..................................................... 29
12. Semantic data model for a user’s social media data .................................... 30
13. Number of keywords vs number of concepts, threshold=1 ....................... 32
14. Number of keywords vs number of concepts, threshold=2 ....................... 33
15. Number of keywords vs number of concepts, threshold=5 ....................... 33
16. Number of keywords vs number of concepts, threshold=10 ..................... 33
17. Sum of scores of keywords vs number of concepts, threshold=1 ............... 34
18. Sum of scores of keywords vs number of concepts, threshold=2 ............... 34
19. Sum of scores of keywords vs number of concepts, threshold=5 ............... 35
20. Sum of scores of keywords vs number of concepts, threshold=10 ............. 35
21. JSON-stat Unified Views Pipeline ............................................................... 40
22. JSON-stat data model .................................................................................. 40
23. Pipeline used to extract World Bank indices provided by an API ............... 43
24. RDF data model for indicator data ............................................................ 44
25. Breakdown of retrieved indicator data ....................................................... 44
List of Tables

1. Journal News .......................................................... 22
2. Agencies’ news ...................................................... 23
3. News API categories .................................................. 26
4. Semantically enriched user profile of a user ......................... 31
5. Example of one statistics retrieved from Eurostat .................. 40
6. Example of one statistics retrieved from Eurostat .................. 41
7. Indicator data that should be harvested in future work .............. 44
Table of Contents

Executive Summary 7

1 Introduction 8
  1.1 Scope of the Document ........................................ 8
  1.2 Relation to PROFIT Project .................................. 8
  1.3 Goals .......................................................... 9
  1.4 On the Selection of Data Sources and their Importance ...... 9

2 Background 11
  2.1 Unified Views .................................................. 11
  2.2 RDF triplestores ................................................. 11

3 Unstructured Data 12
  3.1 Description of the resources ................................. 13
  3.2 News Websites (Static) ......................................... 15
    3.2.1 Introduction ............................................... 15
    3.2.2 Retrieval Process ......................................... 15
    3.2.3 Persistence .................................................. 16
    3.2.4 Results and Statistics .................................... 17
  3.3 RSS Feeds (Dynamic) ........................................... 19
    3.3.1 Introduction ............................................... 19
    3.3.2 Retrieval Process ......................................... 19
    3.3.3 Persistence and Resources ................................. 21
    3.3.4 Results and Statistics .................................... 23
  3.4 News Service APIs (Dynamic) ................................ 25
    3.4.1 Introduction ............................................... 25
    3.4.2 Retrieval Process ......................................... 25
    3.4.3 Persistence and Resources ................................. 26
    3.4.4 Results and Statistics .................................... 27
  3.5 Social Media (On Demand) .................................... 28
    3.5.1 Introduction ............................................... 28
    3.5.2 Retrieval Process ......................................... 28
    3.5.3 Persistence .................................................. 30
    3.5.4 Results ...................................................... 30
  3.6 Analysis ......................................................... 32
    3.6.1 On the Quality of Annotations with Controlled Vocabularies .. 32

4 Numerical Sources 36
  4.1 Description of the resources .................................. 37
  4.2 Statistics (Dynamic) ........................................... 39
    4.2.1 Introduction ............................................... 39
    4.2.2 Retrieval Process ......................................... 40
4.2.3 Persistence and Resources ........................................ 40
4.2.4 Results and Statistics ............................................ 41
4.3 Indicator (Dynamic) .................................................. 42
  4.3.1 Introduction ...................................................... 42
  4.3.2 Retrieval Process ............................................... 42
  4.3.3 Persistence ...................................................... 43
  4.3.4 Results and Statistics .......................................... 44

Conclusion 45
Executive Summary

This document presents a detailed description of the data crawlers, adaptors and extractors that were developed in the first year of this work package.

The structure of this document is as follows:

1. We give a short overview on the scope of the document;
2. We introduce relevant background information about the tools we utilized and further enhanced;
3. We describe the harvesting of unstructured data, which in this case is textual information;
4. We outline the harvesting of structured, numerical information.

The main contributions of the deliverable may be found in:

Section 3 The tools we developed to extract news pages, RSS feeds and corresponding articles, APIs as well as social media;

Section 4 The tools we developed to extract statistics and indicator information.
1 Introduction

In this deliverable, we describe data crawlers, adaptors and extractors utilized to harvest information streams needed for future operations of the PROFIT platform. This information is further linked with the PROFIT thesaurus in order to publish and make harvested information accessible in the most structured and standard-based way possible.

1.1 Scope of the Document

This deliverable reports about our effort to build a holistic framework for the retrieval of unstructured as well as structured information. The information collected in the collaboration with the partners and considered in frames of this deliverable are of utmost importance in recent and further research activities as well as for satisfying the information needs of the users of the PROFIT platform.

In order to ensure the highest level of accessibility and semantic representation, we rely on the principles of Linked Data. Therefore, we utilized and extended Semantic Web technologies to harvest and persist the collected information. As Extract-Transform-Load (ETL) tool we adapted Unified Views (see Section 2.1) for the transformation of the collected data to be compliant with a semantic model. The data is persisted in the Viruoso (see Section 2.2) server of the project, and is already accessible for all members of the consortium.

Based on the expertise of all partners of the consortium the main input streams were identified to be textual inputs. We already collected a wide variety of information sources, such as news pages, RSS feeds and articles, news APIs, and social media, which are presented in Section 3.

We recently targeted also the extraction of financial statistics and indicator data; the results are presented in Section 4.

The document features a concrete description of each retrieval method together with an overview on the collected data.

1.2 Relation to PROFIT Project

The extracted data that are used in this project are used as input for current and future research activities as well as information for the users.

The main data that will be used can be classified in two categories:

1. structured data (numerical data) and
2. unstructured data (non-numerical data).

Structured together with unstructured data is utilized for different purposes, e.g. senti-
ment analysis, analysis of trends, forecasting, etc.

1.3 Goals

The goal of this work package is to build a holistic knowledge base for the PROFIT platform by developing a framework for the harvesting of information. This knowledge base is then used as an information source to keep users up-to-date, to facilitate further research in text and trend assessment, and to build forecasting models.

For these information sources, different requirements were formulated by the consortium:

- Quality of the extracted data;
- Static, dynamic, or on-demand retrieval of resources;
- Extendability of the harvesting framework;
- Semantic representation based on Linked Data principles.

The above requirements are addressed in the sections presenting the different methods for the retrieval of sources.

1.4 On the Selection of Data Sources and their Importance

The significance and readability of the results of this project is associated with the overall quality of both structured and unstructured data in use. The data selection is based on their explanatory power regarding the evolution and dynamics of economic activity, the perceived state of the economy and its expectations driven path. For example Gross Domestic Product (GDP) as measure of value of all final goods and services produced in a period in an economy or as a measure of income is used to capture the economic performance of a whole country and make possible international comparisons among countries regarding their prospects for future growth and development, investment, job creation, welfare etc.

Another interesting example on the importance of monitoring financial and macroeconomic data comes from the monetary policy strategies of central banks. Consider for instance the case of ECB and its transparency framework where the ECB perceives as being transparent "that the central bank provides the general public and the markets with all relevant information on its strategy, assessments and policy decisions as well as its procedures in an open, clear and timely manner". To achieve transparency the ECB attempts to be credible and predictable regarding its actions and policy reactions. In this line the ECB announces its monetary policy strategy and communicates its regular assessment of economic developments, providing access to a rich dataset in a timely manner. From the ECB’s point of view these operations "helps the markets to understand the systematic response pattern of monetary policy to economic developments and shocks. It makes policy moves more predictable for the markets over the medium term. Market expectations can thus be formed more efficiently and accurately". And
the resulting effect of this approach is that "market agents can broadly anticipate policy responses, this allows a rapid implementation of changes in monetary policy into financial variables. This in turn can shorten the process by which monetary policy is transmitted into investment and consumption decisions. It can accelerate any necessary economic adjustments and potentially enhance the effectiveness of monetary policy". Which finally facilitates the ECB’s mandate to maintain price stability, which is essential for economic growth and job creation, implying at the same time that minimizes social welfare losses occurring from various economic shocks. Consider for example a low inflation rate, which is also associated with low variability, resulting as the outcome of the effective monetary policy framework of the ECB. For consumers, the given rate of inflation with predictable dynamics, can help them allocate their assets in a more prudent manner by minimizing their risk and pursuing returns relative to their time preferences and risk taking behaviour. Consumers are also able to chose between different pension or savings regimes, anticipate a stable macroeconomic environment, allowing them to allocate their consumption and leisure choices in an optimal way at least in the medium term. Consumers are also more probable to decide more effectively their relation with the financial system acquiring loans with different characteristics and terms, since they can decide more effectively on the duration and the ex ante real interest rate. On the other hand investors can produce business plans more effectively, since the uncertainty of future returns is lower is predictable and stable macroeconomic environment, increasing the potential and opportunity of investment, increasing aggregate demand in the short term, increasing potential output in the medium to long term, creating new jobs at the same time, and allowing the economy to grow.

Thus the selection of data sources is a crucial point of the project and criteria of high standards have implemented in the selection process. The main criteria used to filter various sources are based on the following characteristics of the data providing institutes: accountability, vertical and horizontal comparability of data, reliability, accessibility of information to the public, well defined accounting standards, the diffusion of information to consumers business scholars and practitioners, the wide use of these data in research and economic analysis both by scholars and practitioners, the impartiality of the institutions, the history of the institution, openness to the public, and the institution’s relative special weight to the information of markets and market participants.

2 Background

In this section we give a short overview on the semantic technologies utilized in our work.

2.1 Unified Views

PoolParty UnifiedViews\(^1\) as part of the PoolParty Semantic Integrator provides a framework to develop, execute, monitor, debug, schedule, and share RDF data processing tasks. It can be seen as a Extract-Transform-Load (ETL) framework for RDF, although it doesn't strictly follow ETL processes as, for example one process can trigger the next one.

Data processing tasks are modeled as pipelines via a graphical interface and can consist of several Data Processing Units (DPUs). PoolParty UnifiedViews comes with a predefined set of DPUs and offers a well documented API to develop Custom DPUs (plugins) on demand.

Pipelines can be scheduled to run on a timely basis or can be triggered by other pipelines. The Execution Monitor gives detailed information on the execution of a pipeline. All data processed is stored in separate graphs and can be reviewed for debugging. The underlying triple store is configurable. Per default, a built-in memory store is used but basically any triple store supporting the SAIL API can be integrated. The Scheduler also includes a notification system that allows to send information about the outcome of scheduled data processing tasks.

2.2 RDF triplestores

A triplestore or Resource Description Framework (RDF) store is a database for the storage and retrieval of data through semantic queries. In this project, we utilize a Virtuoso server for the persistence of information in semantic representation.

Much like a relational database, one stores information in a triplestore and retrieves it via a query language. Unlike a relational database, a triplestore is optimized for the storage and retrieval of triples.

The retrieved data is accessible for all consortium members by using SPARQL queries. Each resource is stored in its own graph based on the origin of the collected data. In addition to queries, triples can be imported/exported following RDF paradigm and one of the serialization formats.

\(^1\)https://www.poolparty.biz/unifiedviews/
3 Unstructured Data

Unstructured data is primarily drawn from some of the most recognizable and highly circulated business financial and economic news agencies as Dow Jones, Financial Times, The Guardian, ECBs announcements, the Wall Street Journal and the site www.investing.com. Moreover the content of webpages such as Financial Times, Wall Street Journal, Forbes, Bloomberg, New York Times, CNN, BBC and The Times may serve as information set for users in the field of financial decisions and activities of an individual or household, including budgeting, insurance, mortgage planning, savings and retirement.

The fundamental goals of utilizing these resources are as follows:

1. the articles are shown to the users in order to keep them informed about the latest events,
2. the articles are used to analyze trends,
3. utilized in sentiment analysis, and
4. the users discuss the articles and leave feedback.

In this section we describe the data sources, the crawlers, adaptors and extractors, and the results of collecting the data. We divided our efforts in three different classes:

1. Static: These resources are harvested only once to build a static corpora, mainly used to create the baseline for the PROFIT thesaurus.
2. Dynamic: For dynamic resources, we scheduled our extraction pipelines to ensure a constant flow of relevant information.
3. On Demand: Personalized information is only collected with consent of the user and, therefore, only on demand.

Current and future work deals with analyzing huge corpora from Dow Jones and we are currently in the process of the refinement of the content for activities that need financial data with the highest quality standards.

We start this section with a short summary justifying the chosen data sources. In the following, we describe the harvesting process of news pages, RSS feeds, news service APIs, and Social Media. We conclude this section with an excerpt from a published paper that relies on the result of the extraction.
3.1 Description of the resources

Regarding the unstructured data sources we will not explicitly describe the criteria they met for brevity, but the brand name of news organizations such as Dow Jones, The Guardian, Financial Times, Wall Street Journal, Forbes, Bloomberg, New York Times, CNN, BBC, The Times, among others, that has built over the years based on their readership, reliability, accessibility of information to the public, the general accepted and recognized level of impartiality and their history in the industry, alongside with their special weight to the information shearing among of markets and market participants, render these media as important primary text content providers for our textual analysis.

Another important news feed introduced by most of the news providing organizations is the personal finance feature. This feature provided to users attempts to provide concentrated and in cases personal tailored information and advice, on current financial and economic developments.

Information about personal finance is provided by several broadsheet newspapers. In general, this is contained in a “personal finance” or “money” section of their websites (e.g., “La Republica” in Italy; “The Guardian” in the UK; “Die Welt” in Germany). Some news channels such as the BBC (British Broadcasting Corporation) and CNN (Cable News Network) also provide similar information on their websites.

The content of these webpages, comprises mostly news articles, analysis and reports on a variety of personal finance topics. In some cases, the content is classified into separate headings relating to an area of personal finance. For instance, the Telegraph newspaper in the UK classifies its personal finance-related content into separate pages as follows: investment, property, bills, insurance, pensions, banking, tax and financial services. Content and categories vary greatly across newspapers, with some being much more comprehensive than others.

The focus on news reports about personal finance provides a useful means for updating knowledge and raising awareness of changes to existing legislation and their impact on financial decisions. However, it may be argued that such benefit is limited by the extent of knowledge one already possesses about these topics. For instance, news reports on pension reforms may have little value for someone with little or no prior understanding of the pension system and the importance of retirement planning.

Furthermore, access to the content of these websites may be restricted to registered users with payment required, e.g. the Financial Times in the United Kingdom. The target audience is generally not specified but could be regarded as not suited for children, very young people and people with low levels of literacy who may struggle to understand the content.

Finally, newspapers’ personal finance sections also feature interactive tools to engage users and enhance financial knowledge, though this varies significantly across newspapers. Examples of such features include: an “Ask the expert” section where users can interact with financial experts by emailing personal finance-related questions, which are
answered and commented on by other users; interactive tools to aid financial decision-making such as mortgage repayment, savings, debt and stamp duty calculators; stock and fund screener tools; glossary of financial terms; newsletters to which users can subscribe to; tools to compare accounts, credit cards and mortgages across different financial services firms; and short videos and podcasts on personal finance.

Overall, personal finance information offered by newspapers on their websites provide practical and interactive means for individuals to enhance their financial knowledge. However, users may require prior financial knowledge to maximize such benefit.
3.2 News Websites (Static)

In this section, we describe the retrieval and processing of news articles of the website investing.com.

3.2.1 Introduction

News pages offer a vast amount of unstructured data that is of utmost importance for deriving high-quality information, automatically or aiding domain experts in their decision-making. In the scope of this project, this textual content has the potential to aid the identification of relevant concepts for thesauri and knowledge bases as well as it builds the baseline for sophisticated text mining tasks such as sentiment analysis. However, for the aforementioned tasks, it is absolutely required to ensure that the harvest data is exclusively gathered for the financial domain and fulfills all quality characteristics for further processing.

Based on the recommendations of our financial domain experts, we rely on three corpora obtained from the news page investing.com. Investing.com is a global financial portal that provides news, analysis, streaming quotes and charts, technical data and financial tools about the global financial markets. The portal provides a broad variety of financial data including Stocks, Bonds, Commodities, Currencies, Interest Rates, Futures and Options. The corpora contains high-quality financial articles and has been approved by financial experts as being representative corpora for the considered field.

We obtained a static collection of around 40,000 documents including the following subdomains:


The remainder of this section describes the implemented retrieval process and provides an overview on the obtained results.

3.2.2 Retrieval Process

The retrieval process implemented for this kind of sources is characterized of the following steps: (i) Crawling, (ii) Cleaning, and (iii) Persistence.

**Crawling** The basic structure of the selected subdomains, i.e. crude-oil-news, eur-usd-news, eu-stoxx50-news, is as follows: an overview page listing the title, link, publisher and teaser to the provided articles (see Figure 1) as well as a detail page providing the full-text to the article.
We implemented a generic parser, which iterates over the posted articles on each page and dives into the detail pages, collecting meta data as well as full-text.

**Cleaning**  Data cleaning is an essential step before the retrieved data is stored and provided for further processing. Although we believe that the editorial standards of *investing.com* meet high quality requirements, the nature of Web resources as unstructured data and mixed with styling features and advertisement requires the following steps:

- Escaping HTML characters
- Standard encoding format
- Removal of HTML styling features
- Removal of advertisement

For above listed steps, we rely on the Apache Tika\(^2\) framework, featuring content detection as well as basic text analysis.

**3.2.3 Persistence**

The retrieved results are persisted in a triplestore (i.e., Virtuoso) whereas the semantic model for this resource is shown in Figure 2.

By utilizing Virtuoso’s SPARQL endpoint, all consortium members can access and query the data obtained from *investing.com*. One exemplary SPARQL query is shown in Listing 1, which retrieves every article with metadata and full-text.

\(^2\)http://tika.apache.org
Furthermore, all three corpora are currently used in the PROFIT PoolParty server.

### 3.2.4 Results and Statistics

In total, the obtained data consists of 39,157 articles from the Website investing.com. The collected news articles span from 2009 till 2016. The following breakdown shows the size of the total corpus in ratio to the corpora exclusively created for the subdomains.
<table>
<thead>
<tr>
<th>Corpus</th>
<th>No. of articles</th>
<th>% of total</th>
</tr>
</thead>
<tbody>
<tr>
<td>crude-oil-news</td>
<td>14,204</td>
<td>36.27%</td>
</tr>
<tr>
<td>eur-usd-news</td>
<td>19,119</td>
<td>48.83%</td>
</tr>
<tr>
<td>eu-stoxx50-news</td>
<td>5,834</td>
<td>14.90%</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td><strong>39,157</strong></td>
<td><strong>100%</strong></td>
</tr>
</tbody>
</table>

Figure 3: Breakdown of the collected articles from investing.com
3.3 RSS Feeds (Dynamic)

In this section, we describe the continuous retrieval of news articles by exploiting financial RSS feeds.

3.3.1 Introduction

In the arising time of omnipresent information needs in our everyday life, we address the issue of harvesting up-to-date and relevant financial news by our own framework that enables the continuous input flow of news articles. Therefore, we implemented generic means allowing to easily define and schedule extraction processes with the following main functions:

1. Retrieval of RSS feeds;
2. Extraction of single items with meta data;
3. Retrieval of websites that are linked in the single RSS items;
4. Extraction of the websites’ main content;
5. Infer storage destination graph either from the source itself or allow the administrator to define it;
6. Persist the collected information.

For this we extended PoolParty Unified Views (see Section 2.1) with new functionality and defined a number of resources that are automatically and regularly collected.

3.3.2 Retrieval Process

For the retrieval process of meta data in RSS feeds in combination with corresponding full-text articles from Web pages, we extended PoolParty UnifiedViews with a new pipeline concept together with custom data processing unities (DPUs).

The conceptual pipelines designed for RSS feed collection for the PROFIT project are illustrated in Figure 4 and Figure 5. We distinguish between a generic pipeline (see Figure 4), which is suitable for multiple resources and infers the destination graph from the RSS feed itself, which is only possible if the atom:link tag is present, and single resource pipelines (see Figure 5) that needs to be configured manually.

The main principle is as follows:

(1) RSS File Download: The first module, where the URL address of the RSS feed is specified, downloads the feed and stores it as file. Subsequently, two tasks are triggered that are executed concurrently.
Figure 4: Generic pipeline for RSS feed extraction

Figure 5: Single resource pipeline for RSS feed extraction without graph name inference
(2) **Meta-Data Extraction:** To extract the meta-data of a single source, e.g. date and publisher, the shorter subtask (see the four modules on the left side in Figure 4), processes each downloaded file and transforms it into RDF format. Special care was given to URL encoding and date conversions from RSS format (RFC 822 Date and Time syntax) to dateTime format of XMLSchema.

(3) **Full-Text Extraction:** The main component of the third part of the pipeline (see the six modules on the right side in Figure 4) is a new DPU for retrieving the full-text articles corresponding to the RSS feed items, i.e. a generic webcrawler module. In case of the generic pipeline, the crawler requires two input channels: one with downloaded Web page and second with the destination graph name which is inferred from the `atom:link` tag of the RSS feed. In case of absence of this tag, a copy of the Single Resource Pipeline (see Figure 5) is created and the destination graph is configured in the `uv-l-rdfToVirtuoso` module. The source code of the Web page is obtained by transforming the `link` tag of each RSS feed item to RDF statements and downloading the files. Subsequently, the webcrawler DPU uses boilerpipe\(^3\) to extract the main content of the web page.

### 3.3.3 Persistence and Resources

The results are automatically loaded in Profit’s Virtuoso whereas the semantic model for this resource is shown in Figure 6.

---

---

\(^3\)Boilerpipe provides algorithms to detect and remove the surplus "clutter" (boilerplate, templates) around the main textual content of a web page ([https://boilerpipe-web.appspot.com/](https://boilerpipe-web.appspot.com/))
Journal News  Based on the recommendation of our financial domain experts, we selected a number of journals news as data source of our crawling mechanism. Table 1 shows the resources that are included at the time this document was written.

Table 1: Journal News

<table>
<thead>
<tr>
<th>Resources</th>
<th>Sub-categories (ID)</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA today</td>
<td></td>
</tr>
<tr>
<td><a href="http://www.usatoday.com/money/personal-finance/">http://www.usatoday.com/money/personal-finance/</a></td>
<td>money_autos</td>
</tr>
<tr>
<td>CNN</td>
<td>money_funds</td>
</tr>
<tr>
<td><a href="http://rss.cnn.com/rss/%5Bsub-category">http://rss.cnn.com/rss/[sub-category</a>]</td>
<td>money_pf</td>
</tr>
<tr>
<td></td>
<td>money_pf_college</td>
</tr>
<tr>
<td></td>
<td>money_pf_insurance</td>
</tr>
<tr>
<td></td>
<td>money_pf_taxes</td>
</tr>
<tr>
<td></td>
<td>money_retirement</td>
</tr>
<tr>
<td>This is Money</td>
<td></td>
</tr>
<tr>
<td><a href="http://www.dailymail.co.uk/money/%5Bsub-category%5D/index.html">http://www.dailymail.co.uk/money/[sub-category]/index.html</a></td>
<td>cars</td>
</tr>
<tr>
<td></td>
<td>experts</td>
</tr>
<tr>
<td></td>
<td>holidays</td>
</tr>
<tr>
<td></td>
<td>money</td>
</tr>
<tr>
<td></td>
<td>investing</td>
</tr>
<tr>
<td></td>
<td>investingshow</td>
</tr>
<tr>
<td></td>
<td>markets</td>
</tr>
<tr>
<td></td>
<td>mortgageshome</td>
</tr>
<tr>
<td></td>
<td>pensions</td>
</tr>
<tr>
<td></td>
<td>saving</td>
</tr>
<tr>
<td>New York Times</td>
<td></td>
</tr>
<tr>
<td>Washington Post</td>
<td></td>
</tr>
<tr>
<td>BBC</td>
<td></td>
</tr>
<tr>
<td><a href="http://www.bbc.co.uk/news/business/your_money">http://www.bbc.co.uk/news/business/your_money</a></td>
<td></td>
</tr>
</tbody>
</table>

Although the general scope of the above listed news sources is based on a wide range of topics, we specifically focused on categories that include financial affairs.

Agencies’ news  Based on the recommendation of our financial domain experts, we included in a later stage of the project also three resources that specifically focus on bank and European Union affairs as well as financial indices (see Table 2).
### Table 2: Agencies’ news

<table>
<thead>
<tr>
<th>Resources</th>
<th>Sub-categories (ID)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>European Central Bank</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Statistical press releases (statpress)</td>
</tr>
<tr>
<td></td>
<td>Publications (pub)</td>
</tr>
<tr>
<td></td>
<td>Working Papers (wppub)</td>
</tr>
<tr>
<td></td>
<td>Legal acts, Opinions, Decisions (legalacts)</td>
</tr>
<tr>
<td></td>
<td>Recent open market operations (operations)</td>
</tr>
<tr>
<td></td>
<td>Open procurements, ... (procurements)</td>
</tr>
<tr>
<td></td>
<td>Euro area yield curve (yc)</td>
</tr>
<tr>
<td></td>
<td>Research bulletin (rbu)</td>
</tr>
<tr>
<td><strong>European Union Newsroom</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Economy, finance, tax and competition (39)</td>
</tr>
<tr>
<td></td>
<td>Employment and social rights (40)</td>
</tr>
<tr>
<td></td>
<td>Energy and natural resources (41)</td>
</tr>
<tr>
<td></td>
<td>Environment, consumers and health (42)</td>
</tr>
<tr>
<td></td>
<td>External relations and foreign affairs (43)</td>
</tr>
<tr>
<td><strong>S&amp;P Dow Jones Indices</strong></td>
<td></td>
</tr>
<tr>
<td><a href="https://us.spindices.com/rss/rss-details/?rssFeedName=">https://us.spindices.com/rss/rss-details/?rssFeedName=</a> [sub-categories]</td>
<td>us-equity</td>
</tr>
<tr>
<td></td>
<td>global-equity</td>
</tr>
<tr>
<td></td>
<td>fixed-income</td>
</tr>
<tr>
<td></td>
<td>commodities</td>
</tr>
<tr>
<td></td>
<td>real-estate</td>
</tr>
<tr>
<td></td>
<td>specialty</td>
</tr>
<tr>
<td></td>
<td>strategy</td>
</tr>
</tbody>
</table>

### 3.3.4 Results and Statistics

In the period from September 1, 2016 till November 30, 2016 we have collected 2,739 articles from RSS feeds by utilizing the above described framework.

The below listed breakdown shows that most of the resources from journal news have little impact except the feed “This is Money” provided by http://www.dailymail.co.uk. However, the breakdown of 59% of journal news to 41% agencies’ news is at a satisfactory level with a solid balance.

One important note is that the framework developed for this type of information has shown its potential to easily handle data sources on the fly. This flexibility and configurability will aid further development of the platform.
<table>
<thead>
<tr>
<th>Feed</th>
<th>No. of articles</th>
<th>% of total</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA today</td>
<td>33</td>
<td>1.20%</td>
</tr>
<tr>
<td>CNN</td>
<td>126</td>
<td>4.60%</td>
</tr>
<tr>
<td>This is Money</td>
<td>1,327</td>
<td>48.45%</td>
</tr>
<tr>
<td>New York Times</td>
<td>75</td>
<td>2.74%</td>
</tr>
<tr>
<td>Washington Post</td>
<td>10</td>
<td>0.37%</td>
</tr>
<tr>
<td>BBC</td>
<td>48</td>
<td>1.75%</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td><strong>1,619</strong></td>
<td><strong>59.11%</strong></td>
</tr>
<tr>
<td>European Central Bank</td>
<td>156</td>
<td>5.70%</td>
</tr>
<tr>
<td>European Union Newsroom</td>
<td>411</td>
<td>15.01%</td>
</tr>
<tr>
<td>S&amp;P Dow Jones Indices</td>
<td>553</td>
<td>20.19%</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td><strong>1,120</strong></td>
<td><strong>40.89%</strong></td>
</tr>
<tr>
<td><strong>total</strong></td>
<td><strong>2,739</strong></td>
<td><strong>100%</strong></td>
</tr>
</tbody>
</table>

Figure 7: Breakdown of retrieved RSS feeds
3.4 News Service APIs (Dynamic)

In this section, we describe the continuous retrieval of news articles by exploiting application programming interfaces (APIs).

3.4.1 Introduction

Some news pages have chosen to provide their content via application programming interfaces (APIs), which allows better control for developers to include the external content into their own applications. In order to provide such resources for the PROFIT project, we created a Unified Views extraction pipeline (cf. Section 2.1) tailored to the needs of one selected API, the Guardian News API. The main functions of this process are:

1. The pipeline creates file download statements, dynamically enriched with from-to date details;
2. The frameworks supports XML and JSON to RDF transformations;
3. Infer storage destination graph either from the source itself or allow the administrator to define it;
4. Persist the collected information.

In the following, the created pipeline is described in more detail.

3.4.2 Retrieval Process

For the retrieval process of information provided by the Guardian News API, we follow the following workflow, which is illustrated in Figure 8.

(1) Query Construction: The first part of the pipeline is concerned with the creation of API requests by dynamically creating from-date parameter and concatenating them to the base URL, which is shown in Listing 2. Subsequently, the files are automatically downloaded and stored temporarily on the server.

```sql
CONSTRUCT {...} WHERE {
BIND (NOW() AS ?now)
BIND (CONCAT("https://content.guardianapis.com/business?
  "from-date=" , str(year(?now)) , "," , str(month(?now)) , "," , str(day(?now)) , "," ,
  "show-fields=headline,trailText,body&
  "show-tags=keyword&
  "page-size=50&
  "format=xml&
  "api-key=..." ) AS ?newURI)}
```

Listing 2: CONSTRUCT query to dynamically create API requests
(2) XML/JSON to RDF Mapping: In the final steps, the downloaded files are parsed and the data is triplified by utilizing Unified Views basic functionality. In case that the data is already retrievable in XML format, it is possible to directly use a `uv-t-xslt` transformer. JSON format requires an additional `uv-t-jsonToXml` transformer, mapping it into standard XML format.

3.4.3 Persistence and Resources

The results are automatically loaded in PROFIT’s Virtuoso whereas the semantic model for this resource is shown in Figure 9.

The results of our harvesting resources from Guardian News API are shown in Table 3. One important note is that we have to create for each new API resource a new pipeline since the results are not standardized in contrast to RSS feeds. Consequently, the XML triplification has to be done for each new data source which also affects the resulting data model.

<table>
<thead>
<tr>
<th>Resources</th>
<th>Sub-categories (ID)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guardian News</td>
<td>money</td>
</tr>
<tr>
<td><a href="https://content.guardianapis.com/%5Bsub-category">https://content.guardianapis.com/[sub-category</a>]</td>
<td>business</td>
</tr>
</tbody>
</table>

Table 3: News API categories
3.4.4 Results and Statistics

At the time of writing this deliverable, we obtained 497 articles in total from the Guardian News API. The following breakdown shows the ratio of the selected subcategories, i.e. money (22%) and business (78%), further grouped by month of retrieval.

<table>
<thead>
<tr>
<th>Category</th>
<th>October</th>
<th>November</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No.</td>
<td>%</td>
<td>No.</td>
</tr>
<tr>
<td>business</td>
<td>138</td>
<td>27.77%</td>
<td>249</td>
</tr>
<tr>
<td>money</td>
<td>43</td>
<td>8.65%</td>
<td>67</td>
</tr>
<tr>
<td>total</td>
<td>181</td>
<td>36.42%</td>
<td>316</td>
</tr>
</tbody>
</table>

Figure 9: Semantic data model for information retrieved from the Guardian News API

Figure 10: Breakdown of the news results obtained from APIs
3.5 Social Media (On Demand)

In this section, we describe the retrieval of social media information, which is further used to personalize the PROFIT platform.

3.5.1 Introduction

While in previous sections, we described our framework for harvesting data for general analytic tasks, in this section we describe which libraries are built to retrieve user specific data for typical objectives including

1. personalizing the Profit platform, e.g., recommend literature based on user profile,
2. obtaining general user opinions for current trends,
3. and getting feedback on particular events.

Therefore, user specific data should be gathered on demand stemming from blogs and social media websites. However, due to ethical reasons and to obtain only relevant information, we refrain from uncontrolled crawling of individual-related data. In contrast, user data is only obtained along with the consent of the user and only for the sources he is willing to share.

3.5.2 Retrieval Process

In general, we target Twitter, Facebook, and Google+ by utilizing the mklab-stream-manager library as well as LinkedIn for which we developed our own customized parser.

**mklab-stream-manager** Stream Manager is designed to monitor a set of seven social streams: Twitter, Facebook, Instagram, Google+, Flickr, Tumblr and Youtube collecting content relevant to a keyword, a user or a location, using the corresponding API that is supported from each service. Twitter API works as a real-time service, whereas the other six act as polling consumers performing requests to the network periodically.

At the time of writing this deliverable, we investigated the retrieval of resources from Twitter, Facebook, and Google+, which seems to be the most promising ones for usage in the Profit platform.

**Twitter** is an online news and social networking service where users post and read short 140-character messages called "tweets". Also without registration, it is possible to read the tweets of a user. Setting up the Twitter stream manager of mklab requires to configure the API key and parameters obtained by registering a Developer Account.

---

4 http://mklab.iti.gr
5 https://about.twitter.com/company
Founded in 2004, Facebook’s mission is to give people the power to share and make the world more open and connected. People use Facebook to stay connected with friends and family, to discover what's going on in the world, and to share and express what matters to them.

Google+, launched in June 2011, features the ability to post photos and status updates to the stream or interest based communities, group different types of relationship, events, location tagging, and much more.

Setting up the Facebook and Google+ streams requires to register an API account similar to the Twitter Developer Account.

The remainder of the supported sources have no direct usage since the provided textual content is not significant. By using the mklab-stream-manager, the problem of different sources is reduced to setting up the parameters and tripling the incoming JSON items.

LinkedIn

Founded in 2003, LinkedIn connects the world’s professionals to make them more productive and successful. With more than 450 million members worldwide, including executives from every Fortune 500 company, LinkedIn is the world’s largest professional network on the Internet.

A typical LinkedIn user profile contains a short description of the user’s background (summary), projects, publications, links to the user’s own articles, and much more. For the retrieval of textual content, we concentrated specifically on the explicitly mentioned categories.

Figure 11: Workflow for parsing LinkedIn profiles

Figure 11 illustrates the conceptual workflow for extracting the articles. As input, the URL to the LinkedIn profile is given that is given by the user of the Profit platform. Subsequently, we are able to parse the user’s profile page, extracting the textual content

---

6https://www.facebook.com/pg/facebook/about/
7https://googleblog.blogspot.co.at/2011/06/introducing-google-project-real-life.html
8https://www.linkedin.com/company-beta/1337?pathWildcard=1337
of summary, projects and publications. Furthermore, we follow the link to the user’s published articles and for each article, we extract the full-text of the post.

### 3.5.3 Persistence

The data model of a user profile is illustrated in Figure 12. For each source, the user profile is enriched with the retrieved information, whereas for Twitter, Facebook, and Google+ only posts are retrieved while summary, publications, projects, and articles are obtained for LinkedIn.

![Figure 12: Semantic data model for a user’s social media data](image)

### 3.5.4 Results

Since the retrieval of individual and personal information is conducted on demand, and only with consent of the user, we have no statistics of this approach. However, in Table 4 we show an excerpt of information retrieved for an active user of LinkedIn.
<table>
<thead>
<tr>
<th>URL</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>social:hasSummary</td>
<td>… has studied business informatics at Vienna University of Economics and Business and at Vienna University of Technology and has obtained a Master’s degree in business studies...</td>
</tr>
<tr>
<td>social:hasPublication</td>
<td>Linked Data will to a high degree replace the added value of pure textual information...</td>
</tr>
<tr>
<td>social:hasPublication</td>
<td>Dieser Beitrag nhert sich den Einsatzmöglichkeiten von Semantic Web Technologien im unternehmerischen Umfeld aus zwei Perspektiven...</td>
</tr>
<tr>
<td>social:hasPublication</td>
<td>Semantic Web ist Vision, Konzept und Programm für die nächste Generation des Internets...</td>
</tr>
<tr>
<td>social:hasPublication</td>
<td>Semantische Techniken zur strukturierten Erschließung von Web 2.0-Content und kollaborative Anreicherung von Web Content mit maschinenlesbaren Metadaten wachsen zum Social Semantic Web zusammen...</td>
</tr>
<tr>
<td>…</td>
<td></td>
</tr>
<tr>
<td>social:hasProject</td>
<td>PoolParty is a world-class semantic technology suite that offers sharply focused solutions to your knowledge organization and content business...</td>
</tr>
<tr>
<td>social:hasProject</td>
<td>PoolParty Semantic Search (PPS) supports scenarios relevant for enterprise information management like semantic enterprise search...</td>
</tr>
<tr>
<td>social:hasProject</td>
<td>PoolParty Extractor (PPX) supports scenarios relevant for enterprise metadata and information management...</td>
</tr>
<tr>
<td>…</td>
<td></td>
</tr>
<tr>
<td>social:hasArticle</td>
<td>Industries that trust in Semantic Technologies In the last years, an increasing number of smart applications in the pharmaceutical industry, Media &amp; Publishing and E-Commerce were driven by Semantic Web technologies....</td>
</tr>
<tr>
<td>social:hasArticle</td>
<td>Things, not Strings Entity-centric views on enterprise information and all kinds of data sources provide means to get a more meaningful picture about all sorts of business objects...</td>
</tr>
<tr>
<td>social:hasArticle</td>
<td>Taxonomies are typically used to resolve challenges in knowledge organization....</td>
</tr>
<tr>
<td>…</td>
<td></td>
</tr>
</tbody>
</table>
3.6 Analysis

The described crawlers, adaptors, and extractors have already shown their potential to harvest useful information. One data source, investing.com has already been used to define measures for the quality of annotations with controlled vocabularies, which we present as preliminary results.

3.6.1 On the Quality of Annotations with Controlled Vocabularies

At the time of writing this deliverable, the retrieved investing.com corpora were used for creating, extending, and evaluating thesauri for the PROFIT project.

Already published work is concerned how one could measure how well a controlled vocabulary fits a corpus. For this purpose we find all the occurrences of the concepts from a controlled vocabulary (in form of a thesaurus) in each document of the corpus. After that we try to estimate the density of information in documents through the keywords and compare it with the number of concepts used for annotations. The introduced approach is tested with a financial thesaurus and the obtained corpora of financial news.

**Counting Keywords** We start with comparing the number of keywords with scores above some threshold and a number of extracted concepts. In Figures 13, 14, 15, 16. In the upper right corner of each figure the correlation coefficient $R^2$ and the slope of the line are presented. For all thresholds the overall picture remains the same, the correlation coefficient for the extended thesaurus is slightly larger than the same coefficient for the original thesaurus. The difference in the slope of the line is more significant.

![Figure 13: Number of keywords vs number of concepts, threshold=1](image)

(a) Original thesaurus  
(b) Extended thesaurus

**Counting Sums of Scores of Keywords** Next we investigate the dependency between the sum of the scores of the keywords above certain threshold and the number of ex-
Figure 14: Number of keywords vs number of concepts, threshold=2

Figure 15: Number of keywords vs number of concepts, threshold=5

Figure 16: Number of keywords vs number of concepts, threshold=10
tracted concepts. The correlation coefficients for the extended thesaurus are larger; this indicates that the extended thesaurus is better suitable for the annotation of the given corpora.

It is worth noting that the number of extracted keywords and, hence, the sums of scores are significantly smaller for the extended thesaurus as some of the terms became new concepts and are not counted anymore.

Figure 17: Sum of scores of keywords vs number of concepts, threshold=1

Figure 18: Sum of scores of keywords vs number of concepts, threshold=2
Figure 19: Sum of scores of keywords vs number of concepts, threshold=5

Figure 20: Sum of scores of keywords vs number of concepts, threshold=10
4 Numerical Sources

The primary sources of numerical data in use are drawn from five of the most commonly used and recognizable databases among scholars and practitioners, namely IMF, World Bank, OECD, ECB and AMECO/Eurostat. In the core of our analysis we focus on the use of financial variables, macroeconomic variables and officially published/announced forecasts, where available, of basic variables.

The main financial variables we use are long-term interest rates (Maastricht criterion), Euro/ECU exchange rates and short term interest rates. The main macroeconomic variables in use are GDP, inflation, unemployment and trade balance decomposed in exports of goods, exports of services, imports of goods and imports of services. The forecasts of the aforementioned variables are also utilized where applicable. Structured data as time series of economic sentiment indicators provided for proxies of consumer sentiment and business sentiment are also taken into account.

The fundamental goals of utilizing this kind of information are as follows:

1. numerical predictive analysis, especially economical prediction,
2. combine numbers with sentiments obtained by utilizing unstructured information sources, and
3. show numerical trends with visualization tools.

We start this section with a short description of potential data sources, followed by the presentation of two sources that were collected in the first phase of this deliverable, namely statistics obtained from JSON-stat resources and indicator data retrieved from World Bank. These two are dynamically harvested, which means that our harvesting pipelines are periodically executed to ensure up-to-date information although the update rate of the sources are monthly or at longer intervals.

In later stages of this work package, we will concentrate our efforts to extend this kind of information sources.
4.1 Description of the resources

Regarding the choice of the IMF data base, is based on the role and responsibilities of the IMF that states "The IMF’s primary purpose is to ensure the stability of the international monetary system – the system of exchange rates and international payments that enables countries (and their citizens) to transact with each other. The Fund’s mandate was updated in 2012 to include all macroeconomic and financial sector issues that bear on global stability". The IMF is one of the oldest global economic organizations that emerged from the UN conference in Bretton Woods, New Hampshire, United States, in July 1944. As one of the most fundamental missions of the Fund is to ensure the stability of the international monetary system apart from lending to countries with balance of payments difficulties, and giving practical help to members, it also keeps track of the global economy and the economies of member countries. Doing so, it monitors the economic and financial policies of its 189 member countries publishing a range of time series data on IMF lending, exchange rates and other economic and financial indicators. The Fund is a widely accepted and recognised organization since it is governed by and is accountable to the 189 countries that make up its near-global membership.

Also emerged from the UN conference in Bretton Woods, New Hampshire, United States, in July 1944, the World Bank is considered as one of the largest international financial institutions that provides loans to developing countries and aims to fight poverty and boost prosperity. Among others to fulfil its mission, the World Bank provides "low-interest loans, zero to low-interest credits, and grants to developing countries. These support a wide array of investments in such areas as education, health, public administration, infrastructure, financial and private sector development, agriculture, and environmental and natural resource management". To facilitate its mission the Word Bank offers free access to comprehensive, downloadable indicators about development in countries around the globe, and a large range of financial and economic indicators.

OECD, an international financial emerged after the WWII (1948), numbers today 35 Member countries, with primary goal to "use its wealth of information on a broad range of topics to help governments foster prosperity and fight poverty through economic growth and financial stability. We help ensure the environmental implications of economic and social development are taken into account.". Their mission is facilitated by an main open access data base covering a large area of financial, economic and structural indicators.

As a branch of the European Union, the AMECO/Eurostat data source are developed to facilitate the European Unions economic governance framework, which "aims to detect, prevent, and correct problematic economic trends such as excessive government deficits or public debt levels, which can stunt growth and put economies at risk.". This framework consists of three parts, monitoring, prevention and correction. The monitoring part, among others contains the task of analyzing economic data and produce forecasts. Most of these datasets used are also publically available.

The European Central Bank (ECB), established in 1998, consists of 19 EU member
states and is one of the largest currency areas in the world. At the mission statement the main objective is "to maintain price stability: safeguarding the value of the euro. The European Central Bank is responsible for the prudential supervision of credit institutions located in the euro area and participating non-euro area Member States, within the Single Supervisory Mechanism, which also comprises the national competent authorities. It thereby contributes to the safety and soundness of the banking system and the stability of the financial system within the EU and each participating Member State."

The European Central Bank declares its commitment "to performing all our tasks effectively. In so doing, we strive for the highest level of integrity, competence, efficiency and accountability. We respect the separation between our monetary policy and supervisory tasks. In performing our tasks we are transparent while fully observing the applicable confidentiality requirements.". As part of their mission and commitments the ECB makes publically available a large data set of economic and financial variables.
4.2 Statistics (Dynamic)

In this section, we describe how we harvest statistical resources that are provided in JSON-stat format.

4.2.1 Introduction

JSON-stat\(^9\) is a simple lightweight JSON dissemination format best suited for data visualization, mobile apps or open data initiatives, that has been designed for all kinds of disseminators.


In this section, we describe how to build our data collection pipeline optimized for periodically retrieving JSON-stat resources based on the use case of Eurostat statistics.

Eurostat is the statistical office of the European Union located in Luxembourg. Its mission is to provide high quality statistics for Europe. While fulfilling its mission, Eurostat promotes the following values: respect and trust, fostering excellence, promoting innovation, service orientation, professional independence.

Providing the European Union with statistics at European level that enable comparisons between countries and regions is a key task. Democratic societies do not function properly without a solid basis of reliable and objective statistics. On one hand, decision-makers at EU level, in Member States, in local government and in business need statistics to make those decisions. On the other hand, the public and media need statistics for an accurate picture of contemporary society and to evaluate the performance of politicians and others. Of course, national statistics are still important for national purposes in Member States whereas EU statistics are essential for decisions and evaluation at European level.

Statistics can answer many questions. Is society heading in the direction promised by politicians? Is unemployment up or down? Are there more CO2 emissions compared to ten years ago? How many women go to work? How is your country’s economy performing compared to other EU Member States?

International statistics are a way of getting to know your neighbours in Member States and countries outside the EU. They are an important, objective and down-to-earth way of measuring how we all live.

\(^9\)https://json-stat.org
4.2.2 Retrieval Process

For the retrieval process of JSON-Stat resources, we implemented a new transformer DPU that utilizes the mapping from this type of format to RDF. The basic pipeline is, therefore, reduced to one extractor, transformer, and loading unit as illustrated in Figure 21.

![Figure 21: JSON-stat Unified Views Pipeline](image)

4.2.3 Persistence and Resources

For each statistical resource, an own graph is created where the RDF data model is shown in Figure 22 and an exemplary entry is shown in Table 5.

```
PREFIX xs:<http://www.w3.org/2001/XMLSchema#>
PREFIX statistics:<http://schema.semantic-web.at#>
```

![Figure 22: JSON-stat data model](image)

<table>
<thead>
<tr>
<th>Resources</th>
<th>Sub-categories (ID)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>statistics:dimension EA</td>
</tr>
<tr>
<td></td>
<td>statistics:dimension Y1</td>
</tr>
<tr>
<td></td>
<td>statistics:dimension 2004</td>
</tr>
<tr>
<td></td>
<td>statistics:dimension GBAAA</td>
</tr>
<tr>
<td></td>
<td>statistics:dimension PYC_RT</td>
</tr>
</tbody>
</table>

Table 5: Example of one statistics retrieved from Eurostat

The statistics entry interprets as follows: The ID of the entry is based on a base url `http://schema.semantic-web.at/` and the path based on the dimensions of the statistics.
entry: PYC_RT/GBAAA/EA/Y1/2004/. These dimensions are also included as properties to facilitate searching. The value retrieved that is based on the current dimensions is given in the statistics:value property.

4.2.4 Results and Statistics

The early results contain three series that are listed in Table 6.

<table>
<thead>
<tr>
<th>Series</th>
<th>Specifics</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Maastricht criterion)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short term interest rates</td>
<td>Money market interest rates - monthly data</td>
<td><a href="http://ec.europa.eu/eurostat/web/interest-rates/database">http://ec.europa.eu/eurostat/web/interest-rates/database</a></td>
</tr>
</tbody>
</table>

Table 6: Example of one statistics retrieved from Eurostat
4.3 Indicator (Dynamic)

In this section, we describe how we exploit Unified Views to obtain indicator data for the financial domain.

4.3.1 Introduction

Indicator are statistical values that are taken together to show the condition or existence of something. For example, they are used trading and to compare countries base on different aspects such as the health situation or income level.

In this section, we focus on the retrieval of World Development Indicators, which is the primary World Bank\textsuperscript{10} collection of development indicators. It presents the most current and accurate global development data available.

The main categories for this collection of indicators are as follows:

- Population, total
- Surface area (sq. km)
- Population density (people per sq. km of land area)
- GNI, Atlas method (current US$)
- GNI per capita, Atlas method (current US$)
- GNI, PPP (current international $)
- GNI per capita, PPP (current international $)
- GDP growth (annual %)
- GDP per capita growth (annual %)

For the retrieval, we built a Unified Views (see Section 2.1) pipeline and customized the extraction modules. Although the world bank database is updated only thrice a year, we scheduled the pipeline monthly to keep updated since the update dates could vary.

4.3.2 Retrieval Process

The Unified Views pipeline for the World Bank API is shown in Figure 23 and works as follows:

\textbf{(1) API Request Construction: } The first module of the created pipeline downloads the first page of each indicator defined in the module itself. On every page we extract the total number of pages that are provided as meta data (see Listing 3), which we further use to create the API requests for downloading all pages.

\textsuperscript{10}http://data.worldbank.org
Figure 23: Pipeline used to extract World Bank indices provided by an API

(2) Indicator Extraction: The XML representation of an page of indicators retrieved from the Worldbank API is shown in Listing 3. Using Unified Views main functionality with a customized transformation script, the information is triplified and loaded in the project’s Virtuoso.

```xml
<wb:indicator id="EN.POP.DNST">
  Population density (people per sq. km of land area)
</wb:indicator>
<wb:country id="ZA">South Africa</wb:country>
<wb:date>2016</wb:date>
<wb:value/>
<wb:decimal>0</wb:decimal>
```

Listing 3: XML representation of the downloaded pages of the indicator API

4.3.3 Persistence

The RDF data model is based on the XML format provided by World Bank and shown in Figure 24. The URI of the single entries is composed out of the base domain name of resource, i.e. http://worldbank.semantic-web.at/, concatenated with the main attributes of each entry, i.e. name of the indicator, country, and year ({indicatorName}/{country}/{year}).
4.3.4 Results and Statistics

The following breakdown shows the early results obtained for the World Development Indicators provided by World Bank.

<table>
<thead>
<tr>
<th>Series</th>
<th>No. of entries</th>
<th>% of total</th>
</tr>
</thead>
<tbody>
<tr>
<td>World Development Indicators</td>
<td>33</td>
<td>100%</td>
</tr>
</tbody>
</table>

Figure 25: Breakdown of retrieved indicator data

Future work will deal with the extraction of Sentiment Indicators and Indicator Data provided by IMF Data, which is shown in Table 7.

Table 7: Indicator data that should be harvested in future work

<table>
<thead>
<tr>
<th>Type</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentiment Indicators</td>
<td><a href="http://ec.europa.eu/economy_finance/db_indicators/surveys/time_series/index_en.htm">http://ec.europa.eu/economy_finance/db_indicators/surveys/time_series/index_en.htm</a></td>
</tr>
<tr>
<td></td>
<td><a href="https://data.oecd.org/leadind/composite-leading-indicator-cli.htm#indicator-chart">https://data.oecd.org/leadind/composite-leading-indicator-cli.htm#indicator-chart</a></td>
</tr>
<tr>
<td>IMF Data</td>
<td><a href="https://www.imf.org/external/data.htm">https://www.imf.org/external/data.htm</a></td>
</tr>
</tbody>
</table>
Conclusion

The deliverable 2.3 “Data crawlers, adaptors and extractors” presents our methods to harvest data for the PROFIT project. Based on the expertise of the consortium partners, we defined a wide range of data sources that are targeted on financial topics or have a relation to the financial domain as it is highly versatile in respect to the information needs of end-users as well as analysts.

In the context of this deliverable the following was done:

1. We defined and built a framework for the retrieval of a wide range of unstructured data, consisting of the following sources:
   - **News Websites:** We collected large corpora with high quality standards, which was already used in D2.2. “Data and information streams – assessment tools”.
   - **RSS Feeds:** We extracted around 2700 news articles with meta data provided by RSS feeds.
   - **News Service APIs:** We harvested around 500 news articles provided by APIs.
   - **Social Media:** We analyzed social media collectors, which covers the four selected platforms for the PROFIT platform, and implemented a new dedicated parser for LinkedIn profiles.

2. We started with the collection of structured data, currently consisting of the following resources:
   - **Statistics:** We harvested three series of statistics provided by Eurostat.
   - **Indicator:** We extracted World Development Indicator provided by World Bank.

In order to develop a flexible framework for data harvesting, we extended Unified Views with the following pipeline templates, which enable to:

1. extract RSS feeds together with the full-text of linked articles,
2. harvest news articles provided by APIs,
3. transform statistics, and indicator data.

For above mentioned pipelines we had to develop two novel data processing units (DPUs) for Unified Views:

1. A DPU that allows to extract the main content of a website (e.g., remove advertisement) and
2. a JSON-stat DPU which enables the triplification of content in JSON-stat format to RDF.
Furthermore, we adapted existing modules of Unified Views to be more resistant in case of parsing and network issues.

In the further stages of the project and in the duration of the respective Task 2.2 (until the end of year 2) we aim at harvesting even more data sources, some of them are already identified and others will be identified when the partners advance further with their tasks.