PROJECT PERIODIC REPORT

PUBLISHABLE SUMMARY

Grant Agreement number: 257928

Project acronym: FIRST

Project title: Large scale information extraction and integration infrastructure for supporting financial decision making

Funding Scheme: STREP

Date of latest version of Annex I against which the assessment will be made:

Periodic report: **1st** ☐ **2nd** ☐ **3rd** ☐ **4th** ☒ **5th** ☐ **6th** ☐

Period covered: **from October 2011 to September 2012**

Name, title and organisation of the scientific representative of the project's coordinator:

Tomás Pariente Lobo, Coordinator, Atos Spain (ATOS)

Tel: +34 912148336

Fax: +34 917543252

E-mail: tomas.parientelobo@atosresearch.eu

Project website address: http://project-first.eu

Table of Contents

1. Publishable summary 3
1.1. Project context and objectives 3
1.2. Main results achieved so far 4
1.3. Expected final results 8
1.4. Project web site 9
1.5. FIRST Consortium 10

1. Publishable summary

1.1. Project context and objectives

The **overall objective** of this project is:

Provide a **large-scale information extraction and integration infrastructure for supporting financial decision-making in near real-time**.

FIRST responds to the challenge of managing, in near real time, the vast amount of unstructured data and information with respect to financial markets which increasingly circulates on the Internet. The different players present in the financial markets, as well as their regulating bodies, usually rely on their own ability to access, quickly extract, and interpret the information that they consider to be relevant for their decision making processes. However, the tools that are currently available are far from being able to offer the option of swiftly identifying situations of potential risk and opportunity automatically. This especially becomes evident when considering the billions of web pages that exist and are being created in an ad-hoc manner.

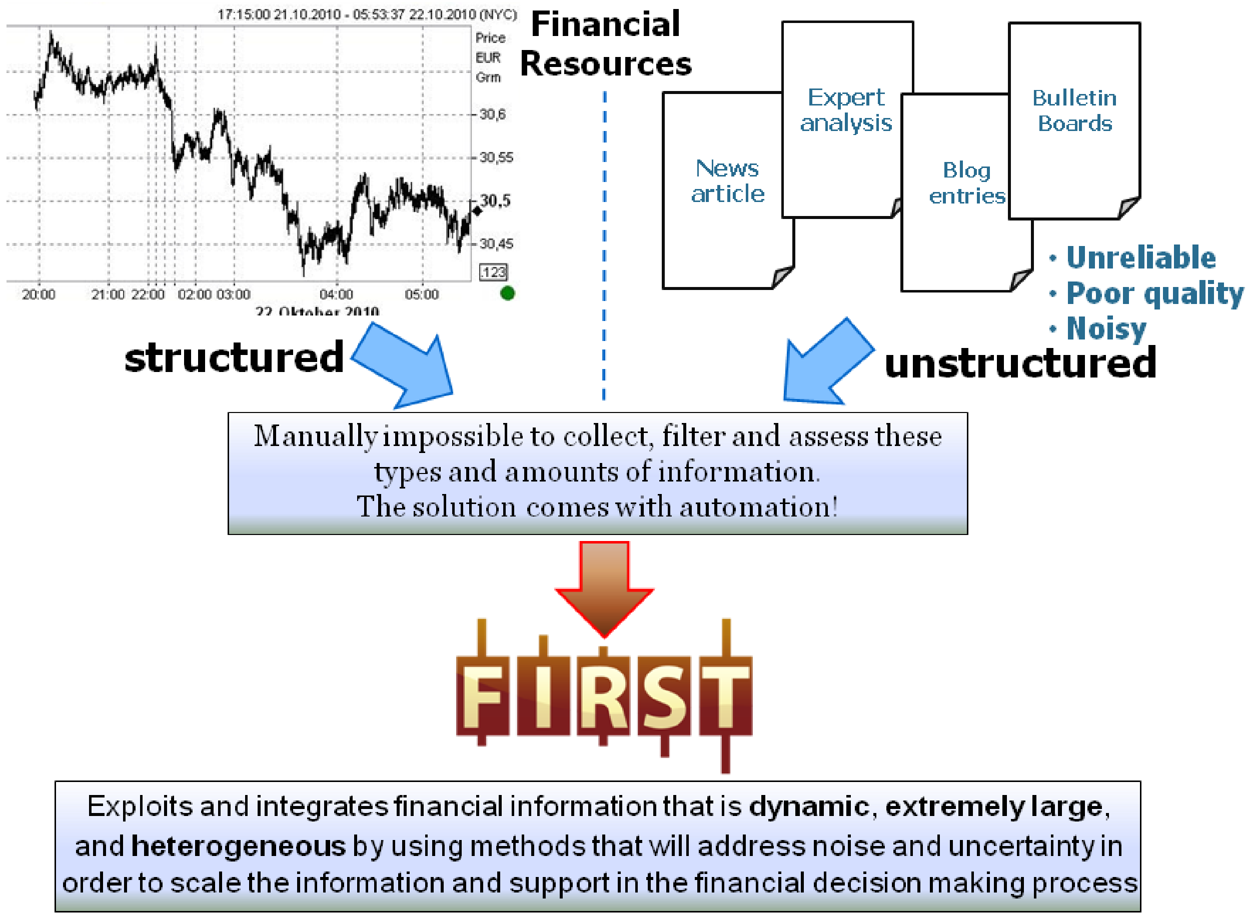
The main challenge of FIRST is to extract relevant sentiments and facts in near real time from textual information sources which are characterised by their large size, unstructuredness, their dynamism, and their heterogeneous nature, such as news articles, blogs, and online newsletters. Using stream-based online algorithms for natural language processing, information extraction, and visualization, the project will help to identify the relevant information which can be used in advanced financial decision models. Therefore, the main innovations that FIRTS offers are:

- Information extraction and sentiment analysis from unreliable semi-structured sources on a massive scale and in near real-time
- Automatic reuse of existing ontologies
- Large-scale ontology learning
- Advanced glass box decision models making use of semantic attributes

FIRST validates its innovations on three extremely challenging and complementary use cases:

1) Market surveillance
2) Reputational risk management
3) Online retail brokerage and investment management

This vision of the project is depicted in the following figure;



1.2. Main results achieved so far

The reporting period covers the second year of the project (1.10.2011-30.09.2012). The first year of the project was mainly concerned with requirements analysis, elicitation of the state of the art from the literature, design of software artefacts and the scaling strategy, set up of the infrastructure, creation of a large corpus of manually annotated web documents serving as golden standard, and implementation of early prototypes and demonstrators.

Building on the results of the first year, the first half of the second year saw the first stage of the "development, integration, and evaluation" phase of the project. This phase has been on advancing and optimizing the prototypes of the information processing pipeline by exploiting the corpus created in the first year and by introducing new methods. The golden standard corpus is also used for evaluation of the information extraction artefacts. Further, the integration infrastructure has been set up and the first Integrated Financial Market Information System has been released.

The second half of the second year has witnessed a major release of the FIRST tools. On the one hand, from the technical side the data acquired has been unified, cleansed and pre-processed providing now a FIRST dataset containing relevant textual annotated facts from financial instruments and objects for more than one year of continuous acquisition. Different methods of sentiment analysis have been developed and applied to reduced set of data, and will be extended in the future to cover the entire datasets. Quantitative and qualitative decision support models have been developed for the three use case scenarios along with a set of state-of-the-art visualization of near-real time data. The three use cases developed their early prototypes based on the data, the decision support models, and the services and infrastructure provided on top of the FIRST data. The first version of the Web Sentify portal has been also provided as an open showcase giving access to FIRST functionality and visualization techniques.

On the other hand the dissemination and exploitation strategy is starting to shift from a scientific and technical dissemination towards being more exploitation and industrial-oriented. In this sense the Milan industrial workshop marked an inflexion point in this strategy. It is worth noticing that most of the technical results are Open Source and can be accessed through the project web site (http://project-first.eu/).

The major results achieved so far with respect to the objectives of the different work packages that comprise the project are described in the following paragraphs:

WP1: Requirements analysis

- **Objectives:** WP1 defines the use cases and provides a detailed definition of requirements concerning different end-user aspects. WP1 lifespan correspond with the reporting period, so the main objectives were setting the three usecases, specify the detailed user requirements, and define the metrics for the aforementioned usecases.
- **Results:** WP1 has delivered three public documents in the first year of the project, explaining in detail the scenarios of usage for each of the three use cases, the requirements from the case study partners and the metrics and methods to be used for the validation of the project results from the usecase perspective. Besides, the process for generating an annotated corpus of text documents (golden standard) with respect to each use case has been agreed-upon and the corpus itself has been almost concluded in the first year. During the second year a resubmission of two deliverables keeping track of the evolution of the requirements has been done.

WP2: Technical analysis, scaling strategy, and architecture:

- **Objectives:** WP2 provides the foundation for a comprehensive, coherent, and integrated conceptual and technical architecture for the Integrated Financial Market Information System. As per WP1, WP2 was active only in the first year of the project achieving its objectives that were the gathering and analysis of the technical requirements, the analysis of the state of the art on relevant technologies and tools, the creation of an scalable and robust architecture, the setting of the integration approach as input for WP7, and finally the definition of the scaling strategy to be followed throughout the project.
- **Results:** Three public documents were submitted in the first year of the project: (i) technical requirements and the state-of-the-art of science and technology that will be used in FIRST; (ii) architectural analysis and system design, as well as integration middleware and GUI analysis, (iii) general overview on suitable scaling techniques to be applied within the project and the strategy of its implementation.

WP3: Data acquisition and Ontology infrastructure:

- **Objectives:** WP3 is aiming at setting up the infrastructure for data acquisition from unstructured resources and providing semi-automatically the ontology used for later information extraction.
- **Results:** WP3 started officially in M5. The main results during the reporting period have been (i) improved data acquisition software, (ii) an enlarged dataset of news and blog posts collected so far, collecting and consolidating a large amount of news and blogs from relevant financial sites (from Oct. 2011, over 5 mio documents, over 7TB of data), and (iii) improved version of the FIRST ontology. The FIRST information extraction ontology has been grounded and enhanced by instances based on analyzing the literature and enriching instances with real world textual representations obtained from the FIRST corpus.

WP4: Semantic information extraction system:

- **Objectives:** WP4 aims at providing semantic information extraction components that will, based on the ontology created in WP3, provide high-level features (i.e., relatively complex extracted facts and sentiments) for decision support models in WP6.
- **Results:** WP4 started officially in M5. During the reporting period, the main work has been on exploiting the manually annotated FIRST corpus for (1) improving quality of sentiment analysis, (2) automatic validation of sentiment analysis, (3) developing concepts and specifications for advanced information extraction and sentiment analysis, with methods for addressing noise and uncertainty, (3) implementing and evaluating a hybrid sentiment classification approach of knowledge-based, and machine-learning techniques, and (4) the provision of annotated sentiment data..

WP5: Information Integration Infrastructure:

- **Objectives:** WP5 provides a persistent storage for the information extracted in WP4. The information will be correlated with the existing information sources and made available to the project partners.
- **Results:** During the reporting period WP5 focused on gathering further requirements and technical details towards storage services and continuing the implementation of large-scale integrated knowledgebase. WP5 has formally finished in the second year of the project, provided services for Document Hash Storage, Document Metadata Storage and Ontology Archive. Also database schema and data storage/retrieval services have been defined for Sentiment Storage.

WP6: Decision Support Infrastructure

- **Objectives:** WP6 aims at designing and developing components for supporting financial decision making. The decision support components including models and visualisation techniques will be subjected to use case requirements defined in WP1
- **Results:** WP6 started in M7, and the work in the second year has focused on designing initial decision support models. (1) For qualitative modelling in UC#1 and UC#2; (2) dataset and high-level feature concepts have been collected and identified for a quantitative model in UC#1; (3), developing advanced visualization techniques over data streams that will be part of the Sentify portal.

WP7: Integrated financial market information system

- **Objectives:** WP7 aims at implementing an Integrated Financial Market Information System that comprises the features offered by the techniques and components developed within the other technical work packages. The main objective is subdivided in the following goals: (a) realization of the integration technologies and infrastructure dimensioned according to the data volume expected, (b) specification of the integration plan, (c) integration components produced, (d) realization of an integrated GUI, common access point to the features produced by the integrated components, (e) implementation of the scaling strategy devised in WP2
- **Results**: WP7 started in M13. The main results in the reporting period are: (1) release of the pipeline integration component (2) an early prototype of integrated components (mainly data acquisition and data analysis pipelines). (3) the specification and implementation of use case independent services that expose results of the information processing pipeline and of

the data stored; and (5) the development of the early version of the Sentify web portal as showcase of the project results, services and visualization techniques.

WP8: End-user Prototype and Evaluation

- **Objectives:** WP8 aims to develop demonstration facilities and end-user prototypes targeting the use case scenarios. The prototypes will be implemented by customizing the Integrated Financial Market Information System (WP7) for the respective demonstration scenarios. Furthermore,WP8 will conduct both quantitative and qualitative evaluations and collect and process end-user feedback
- **Results:** **I**n order to demonstrate the project results, the early version of customized end-user prototypes has been developed on the basis of the architecture provided by WP7. WP8 started in M19. Note that each use case prototype was released in a first version and do not yet cover the whole technical and functional infrastructure. They provide an insight in the functional logic, visualisation components and back-end infrastructure. The work will be continued in the next reporting period

WP9: Dissemination and exploitation:

- **Objectives:** WP9 defines and executes dissemination and exploitation plans so that the results of the project will be successfully adopted by scientific community and industry for the duration of the project and afterwards. The main objectives during the period were related to setting up the dissemination infrastructure and material and the initial discussions on the dissemination and exploitation plans.
- **Results:** WP9 main work in this period has been devoted on refining the dissemination and exploitation strategies. A major result has been the first industrial workshop addressing the user communities, co-located with the ABI Lab Forum 2012. Videos of presentations are available on VideoLectures website. The exploitation and open source plans have been also refined and reported.

WP10: Project management:

- **Objectives:** WP10 ensures the achievement of the project results through technical and administrative coordination as well as provide timely and efficient organizational and financial coordination with respect to contractual commitments. During this period the main work has been devoted to set up the foundation for an effective management of the project from the administrative and financial perspective, the quality assurance and the delivery of the official reporting to the EC. Therefore the objectives for the first year have been:
    - to carry out contractual, administrative and financial coordination and controlling
    - to cooperate and communicate between the Commission and the project
    - to provide overall RTD organisation, coordination, and control
    - to represent the project towards external parties
    - to ensure that preparation and provision of deliverables meet time and quality targets
    - to organise the kick-off meeting and guide the inception phase of the project
    - to organise Consortium Meetings
    - to provide quality assurance to the project by establishing a Project Manual
    - to monitor, supervise and manage the DOW amendment to include a new partner in the project

- **Results:** Within the reporting period the project is undertaking an amendment that affects mainly the work of IDMS in WP5 and WP8. Most of the deliverables have been delivered in due time according to the EC and internal quality procedures, and the few exceptions to that rule are available at the time of writing this report. No major issues on the horizon from the management perspective..

1.3. Expected final results

The main results of the project are tightly coupled to the major specific technological and scientific objectives listed below:

- **Objective 1:** Develop an Integrated Financial Market Information System based on a pluggable architecture framework for non-ICT skilled end-users for on-demand information access and scalable execution of financial market analyses. Furthermore, demonstrate in principle domain-independent applicability of the tools and infrastructures by developing an as generic as possible GUI-based prototype.
- **Objective 2:** Develop a Data Acquisition Infrastructure for acquiring and accessing massive amounts of historical heterogeneous information and live feeds of unstructured, semi-structured and structured information, leveraging on existing tools and infrastructure.
- **Objective 3:** Develop an Ontology Infrastructure with tools for manually and semi-automatically capturing and maintaining the domain knowledge models to be used for information extraction.
- **Objective 4:** Develop a Semantic Information Extraction System for scalable information extraction with addressing of uncertainty in the information sources, and storing extracted information fragments to the knowledge base.
- **Objective 5:** Develop an Information Integration Infrastructure for integrating and consolidating information from heterogeneous (unstructured, semi-structured and structured) information sources.
- **Objective 6:** Develop a Decision Support Infrastructure for integrating information from the knowledge base into financial event detection models, visualization models, decision models, and for scalable execution of these models.
- **Objective 7:** Based on the FIRST tools and infrastructure, develop end-user prototypes and validate them in three large-scale testbeds for the (1) market surveillance, (2) reputational risk management support and (3) online retail brokerage and investment management use case with representative groups of real end-users and massive historical weakly structured textual data and structured information, as well as live-feed data from European stock markets and news wire services.

As a summary of the major impacts of the project, FIRST aims to:

- increase the ability to exploit very large information spaces for citizens and professionals;
- advance significantly information extraction and integration methods;
- develop new information-based electronic service models for service providers;
- increase the competitiveness of European ICT solution providers targeted at financial information management;
- decrease information asymmetry and increase market transparency at the financial markets and thus contribute to making such markets more trustworthy, liquid, and dependable;
- also increase the competitiveness of European financial service providers and institutes.

1.4. Project web site

The project web page can be found at www.project-first.eu

1.5. FIRST Consortium

The FIRST Consortium consist of **9 partners** drawn across the European Union which covers the entire value chain of large-scale information management in the financial domain, with a broad spectrum of companies, financial institutions, and academia focusing in assessing the benefits about the productivity, competitiveness, and profitability of the FIRST solution.

- Atos Spain (Project coordinator – Spain)
- University of Hohenheim (Germany)
- Jozef Stefan Institute (Slovenia)
- Goethe University Frankfurt & E-Finance lab (Germany)
- B-Next Engineering GmbH (Germany)
- Banca Monte dei Paschi di Siena SpA (Italy)
- Interactive Data Managed Solutions AG (Germany)
- Boerse Stuttgart Holding GmbH (Germany)
- Georg-August-University Göttingen (Germany)