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 2012 to September 2013**

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

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.
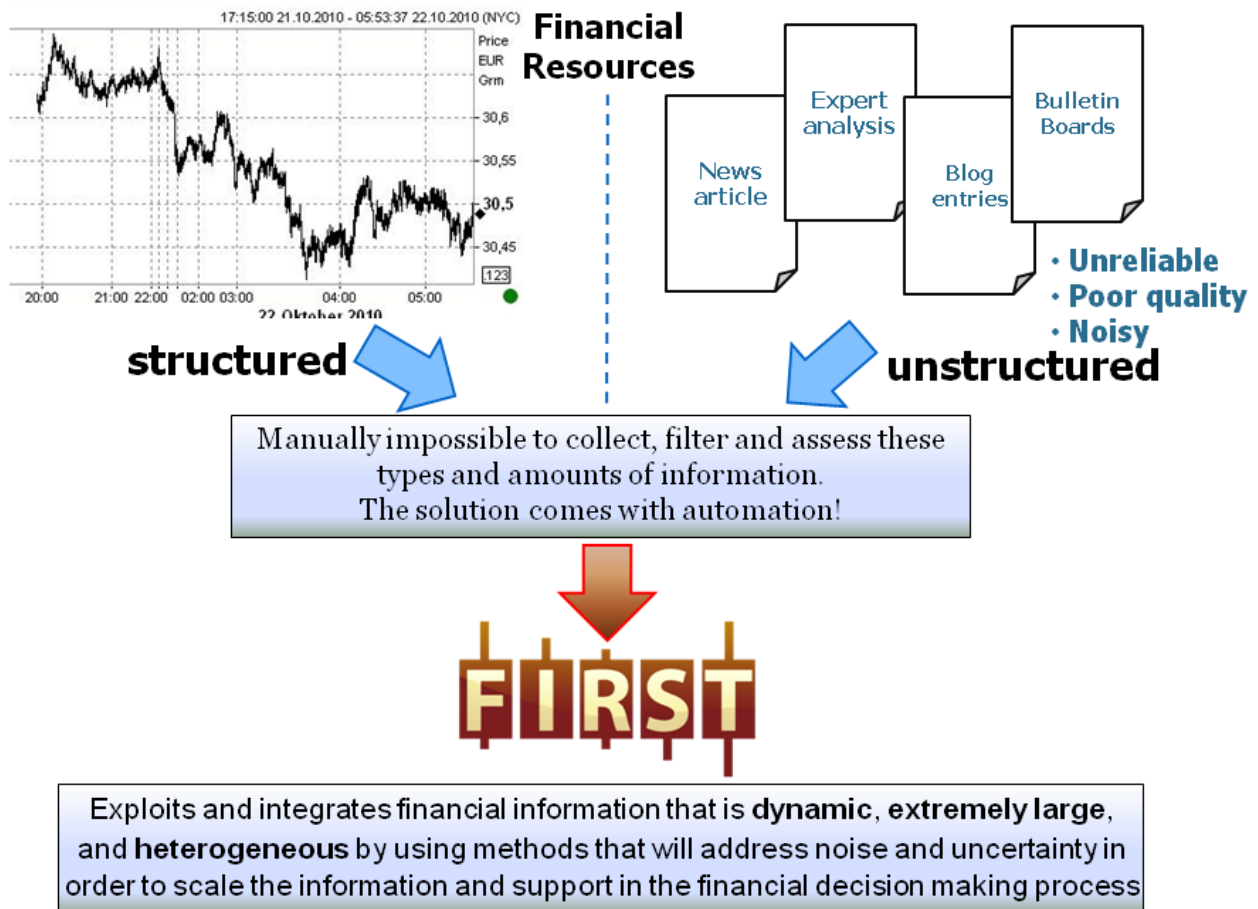
The main challenge of FIRST is to extract relevant sentiments and facts in near real time from textual information sources which are characterised by their large size, unstructuredness, their dynamism, and their heterogeneous nature, such as news articles, blogs, and online newsletters. Using stream-based 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 third year of the project (1.10.2012 - 30.09.2013).

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 artifacts 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 second year witnessed a major release of the FIRST tools. These tools included a set of integrated prototypes covering most of the areas of the project:

- Advanced and optimized prototypes of the information processing pipeline.
- Exploitation of the corpus created in the first year
- First version of Integrated Financial Market Information System services and the Sentify web portal, including services and infrastructure provided on top of the FIRST data. Acquired FRIST datasets curated and, cleansed.
- Development of different methods of sentiment analysis applied to reduced set of data, and will be extended in the future to cover the entire datasets.
- Development of quantitative and qualitative decision support models 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, provided as an open showcase giving access to FIRST functionality and visualization techniques.
- The dissemination and exploitation shifted from a scientific and technical dissemination towards being more exploitation and industrial-oriented.

Now we are reporting the first six months of the final year of the project. The highlights of this period can be summarized as follows:

- Ongoing effort towards the finalization of most of the technical work expected in the next reporting period.
- Major work on the use case prototypes prior to the final evaluation.
- Progress in the dissemination and exploitation prospects
- Engagement of potential customers with the presence of FIRST in a stand and two talks in the main track of the EFMA Customer Week industrial workshop.

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 no activity in the reporting period. Within this WP, three public documents had been delivered 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. The second year saw a resubmission of two deliverables keeping track of the evolution of the requirements. At the end of the project there will be a final resubmission of the same deliverables. WP8 is the natural continuation of the work done in WP1.

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. This WP ended at the end of the first year. WP7 is the natural continuation of the work done in WP2.

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) developed the 4th version of the FIRST ontology; (ii) collected and consolidated a large amount of news and blogs from relevant financial sites (from Oct. 2011, over 8 mio documents, over

8.5TB of data); and (iii) developed the first version of an online, stream-based document clustering algorithm.

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 results are (i) integrated large-scale semantic information extraction and sentiment analysis with data acquisition near-real time pipeline; (ii) availability of financial instrument and company specific historic and contemporaneous sentiments; (iii) substantially enhanced corpus for enabling higher accuracy and better evaluation of sentiment analysis; (iv) availability of large amounts of web documents for validating the analysis and extraction methods; and (v) improved methods.

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: WP5 delivered a set of storage services and implemented a large-scale integrated knowledgebase. WP5 finished in the second year of the project, providing the 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 WP5 will deliver an update of the existing services and schemas at the end of the project.

WP6: Decision Support Infrastructure

Objectives: WP6 aims at designing and developing components for supporting financial decision making. The decision support components including models and visualization techniques will be subjected to use case requirements defined in WP1

Results: WP6 started in M7 and the work in the reporting period has focused on designing initial decision support models. (1) For qualitative modeling 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: (i) substantial improvements in the area of components integration, including first release of the fully operational FIRST processing pipeline comprising all devised components, (ii) advanced prototype of the Sentify web portal (formerly called "Integrated Financial Market Information system") with partial implementation of the document drill-down functionalities; and (iii)

improvements in the platform services, leading to the second stable release of the data access services.

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: WP8 started in M19. In order to demonstrate the project results, an early version of customized end-user prototypes was developed on the basis of the architecture provided by WP7 and provided in the second year of the project. During the reporting period the development of the three case study prototypes advanced and will be subject of evaluation 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 the uptake of the FIRST results. FIRST was present in a second industrial workshop addressing the user communities (EFMA Customer Week). The final dissemination and exploitation will be reported at the end of the project.

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 organize the kick-off meeting and guide the inception phase of the project
- to organize 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 management continued to be smooth. 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 <u>Integrated Financial Market Information System</u> 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 <u>GUI-based prototype</u>.

- **Objective 2:** Develop a <u>Data Acquisition Infrastructure</u> 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 <u>Ontology Infrastructure</u> with tools for manually and semi-automatically capturing and maintaining the domain knowledge models to be used for information extraction.

- **Objective 4:** Develop a <u>Semantic Information Extraction System</u> 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 <u>Information Integration Infrastructure</u> for integrating and consolidating information from heterogeneous (unstructured, semi-structured and structured) information sources.

- **Objective 6:** Develop a <u>Decision Support Infrastructure</u> 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 <u>three large-scale testbeds</u> 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**

FIRST

large scale inFormation extraction and Integration
infRastructure for SupporTing financial decision making

Home   News and Events   Dissemination Material   Who we are   Links   Rss   My account   Repository

Events Calendar

« May »

M	T	W	T	F	S	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

Project Description
Objectives
Technical Details
Project Results
Case Studies

Tech and financial Twits
FIRST Project

Dallas hasn't been this overwhelmed since the Savings & Loans crisis collapsed their real estate & oil dominated economy in the mid 1980s.
9 hours ago · reply · retweet · favorite

10+ recent retweets

A "two-state solution" is like a plane with First Class and Economy. First Class human rights for Israelis, Economy rights for Palestinians.
8 hours ago · reply · retweet ·

View   Edit   Outline

Home

How to focus on the relevant while making financial decisions in a world of information overkill?

Ask **FIRST** first

FIRST develops and provides a large scale information extraction and integration infrastructure which will assist in various ways during the process of financial decision making.

The FIRST project provides an **information extraction, information integration and decision making infrastructure for information management in the financial domain**. This area is extraordinarily faced with the challenges of extremely large, dynamic, and heterogeneous sources of information. The daily work and the business success of all decision makers in this industry depend on the availability of highly trustable, easily acquirable information.

Within the international financial industry not only individual market participants are negatively affected by information overflow and information asymmetries, but also financial flows and financial stability within entire industries of the global economy. This may also have severe consequences as has been observed in the current financial and economic crisis. Information is amongst the most valuable assets in the financial industry and reducing information asymmetries and increasing transparency by providing a fast, real-time, automatic and more comprehensive information base can help preventing false decisions.

FIRST: Opening up new before-the-fact information for earlier and better treatment of evolving conditions in advanced financial decision making

Co-funded by the European Union

Events To Come

No Events

Latest News

- FIRST public deliverables online 6 weeks 6 days
- FIRST in the EC CORDIS Web site 11 weeks 4 days
- FIRST 2nd Consortium Meeting in Frankfurt on February 14-15 on IDMS premises 14 weeks 1 day
- FIRST project now in Twitter and Facebook 10 weeks 4 days

Latest Files

- BSI_Fabijan_Bled_may2011.ppt At Presentations folder
- FIRST 3rd Meeting - Knowtator Documentation 2011-05-17.pptx At Presentations folder
- FIRST 3rd Meeting - Corpus Design Status 2011-05-17.pptx At Presentations folder
- 2011-05-17_FIRST_3rdmeeting_M1 At Presentations folder
- 2011-05-16_FIRST_3rdmeeting_WF At Presentations folder
- M12 prototype - Integration

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)